

# Measuring the HI mass function below the detection threshold

Hengxing Pan,<sup>1,2,3★</sup> Matt J. Jarvis<sup>1b</sup>,<sup>3,4</sup> James R. Allison<sup>1b</sup>,<sup>3</sup> Ian Heywood,<sup>3,5,6</sup>  
 Mario G. Santos,<sup>4,6</sup> Natasha Maddox<sup>1b</sup>,<sup>7</sup> Bradley S. Frank<sup>6,8,9</sup> and Xi Kang<sup>1</sup>

<sup>1</sup>Purple Mountain Observatory, 2 West Beijing Road, Nanjing 210008, China

<sup>2</sup>University of the Chinese Academy of Sciences, 19A, Yuquan Road, Beijing 100049, China

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

<sup>4</sup>Department of Physics and Astronomy, University of Western Cape, Cape Town 7535, South Africa

<sup>5</sup>Department of Physics and Electronics, Rhodes University, P.O. Box 94, Grahamstown 6140, South Africa

<sup>6</sup>South African Radio Astronomy Observatory (SARAO), 2 Fir Street, Observatory 7925, South Africa

<sup>7</sup>Faculty of Physics, Ludwig–Maximilians Universität, Scheinerstr. 1, D-81679 Munich, Germany

<sup>8</sup>Department of Astronomy, University of Cape Town, Private Bag X3, 7701 Rondebosch, South Africa

<sup>9</sup>Department of Astronomy, Inter-University Institute for Data-Intensive Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa

Accepted 2019 October 24. Received 2019 October 1; in original form 2019 July 19

## ABSTRACT

We present a Bayesian stacking technique to directly measure the HI mass function (HIMF) and its evolution with redshift using galaxies formally below the nominal detection threshold. We generate galaxy samples over several sky areas given an assumed HIMF described by a Schechter function and simulate the HI emission lines with different levels of background noise to test the technique. We use MULTINEST to constrain the parameters of the HIMF in a broad redshift bin, demonstrating that the HIMF can be accurately reconstructed, using the simulated spectral cube far below the HI mass limit determined by the  $5\sigma$  flux-density limit, i.e. down to  $M_{\text{HI}} = 10^{7.5} M_{\odot}$  over the redshift range  $0 < z < 0.55$  for this particular simulation, with a noise level similar to that expected for the MIGHTEE survey. We also find that the constraints on the parameters of the Schechter function,  $\phi_*$ ,  $M_*$ , and  $\alpha$  can be reliably fit, becoming tighter as the background noise decreases as expected, although the constraints on the redshift evolution are not significantly affected. All the parameters become better constrained as the survey area increases. In summary, we provide an optimal method for estimating the HI mass at cosmological distances that allows us to constrain the HIMF below the detection threshold in forthcoming HI surveys. This study is a first step towards the measurement of the HIMF at high ( $z > 0.1$ ) redshifts.

**Key words:** methods: statistical – galaxies: mass function – radio lines: galaxies.

## 1 INTRODUCTION

Cold gas in the form of neutral hydrogen atoms (HI) provides the reservoir from which molecules are formed and which subsequently go on to fuel star formation in galaxies, from the distant to the nearby Universe. To understand the whole picture of galaxy formation and evolution, it is crucial to know how the HI gas evolves with cosmic time and as a function of environment.

The preferred approach for tracing the HI gas in the local Universe ( $z \approx 0$ ) is via the direct detection of the neutral hydrogen 21 cm hyperfine emission line. Two prime examples of this are the HI Parkes All-Sky (HIPASS) Survey (Barnes et al. 2001; Zwaan et al. 2005), conducted with the 64-m Parkes radio telescope in Australia and the Arecibo Legacy Fast ALFA (ALFALFA) survey (Giovanelli

et al. 2005; Martin et al. 2010; Jones et al. 2018) conducted by the Arecibo radio telescope with better sensitivity, angular and spectral resolution compared to the Parkes. However, both of these are limited in terms of redshift and sensitivity and have concentrated on surveying large swathes of the local Universe. Measuring the evolution of the HI content of galaxies with redshift requires receivers that work to lower frequencies, coupled with high sensitivity and preferably with relatively high spatial and spectral resolution to avoid confusion (e.g. Delhaize et al. 2013; Jones et al. 2016).

However, we are entering an era of high sensitivity HI surveys with the completion of a variety of new telescopes, in particular the wide-area surveys opened up with Phased Array Feed technology, principally the Australian Square Kilometre Array Pathfinder (ASKAP; Johnston et al. 2007; DeBoer et al. 2009) and the Netherlands Aperture Tile In Focus (Verheijen et al. 2009; Oosterloo et al. 2009), along with the extremely sensitive Meer Karoo Array Telescope (MeerKAT; Jonas 2009; Jonas & MeerKAT Team 2016),

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

which is conducting deeper but narrower blind surveys for HI: The MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) Survey (Jarvis et al. 2016) and the Looking At the Distant Universe with the MeerKAT Array (LADUMA) survey (Holwerda, Blyth & Baker 2012; Blyth et al. 2016). One of the most important goals of these surveys is to help us understand the cosmic evolution of HI in the Universe (e.g. Maddox, Jarvis & Oosterloo 2016; Meyer et al. 2015).

However, even with the high sensitivity of MeerKAT, the 21 cm emission signal is so weak that only the most HI-massive galaxies will be directly detected beyond the local ( $z > 0.1$ ) Universe, unless very long integration times are used (e.g. Fernández et al. 2016). A proposed method to go further than the planned survey limits is to rely on stacking techniques, which use the known positions of galaxies from surveys at other wavelengths and allows the measurement of simple quantities, such as the average HI mass (e.g. Rhee et al. 2018, and references therein).

Stacking in the context of HI studies usually involves co-adding the spectral line data at the known locations of many galaxies to improve the signal-to-noise ratio at the expense of information on individual galaxies contributing to the stack (e.g. Delhaize et al. 2013; Rhee et al. 2018). Here, we instead present a Bayesian technique to model the distribution of HI masses and not just co-add or average the data. This technique is based on a maximum-likelihood approach (Mitchell-Wynne et al. 2014), which was extended into a fully Bayesian framework by Zwart, Santos & Jarvis (2015), who used it to model the source counts of faint radio continuum sources below the nominal detection threshold. Henceforth we simply describe this as Bayesian stacking, although we note that it is not strictly ‘stacking’ in the normal sense of the word used in the literature.

Specifically, in this paper we extend this technique to determine the HI-mass function (HIMF) directly from the HI-line flux distribution constructed from extracting integrated fluxes across appropriate numbers of spectral channels at the redshift and position of galaxies selected in another waveband. We note that a similar technique has been used to measure the radio luminosity function of optically selected quasars below the noise (Malefahlo et al. 2019).

The idea behind this technique for HI studies is that assuming we can determine the behaviour of the noise, the individual line-integrated 21-cm fluxes measured at the positions and redshifts of known galaxies can provide more information than just the average HI mass. Accurate redshift information, in addition to positional information, is critical and so several spectroscopic surveys are planned to cover both the LADUMA and MIGHTEE fields (e.g. the Deep Extragalactic Visible Legacy Survey; Davies et al. 2018). Spectroscopic data from the Galaxy and Mass Assembly (Driver et al. 2011) survey will provide accurate redshift information for the ASKAP Deep Investigation of Neutral Gas Origins (Meyer 2009) survey.

We structure this paper as follows. In Section 2, we describe a source-count model for the HI flux distribution and introduce a Bayesian stacking technique for constraining the HIMF model below the detection threshold. We also describe how we generate the HI galaxy sample given an assumed HIMF model and how this is combined with realistic emission line profiles and input into a noisy spectral cube. In Section 3, we apply our method to the simulated data to test the robustness of our method in measuring the HIMF. In Section 4, we summarize our results, highlight some caveats in relation to simulated versus real data, and outline future work based on the real data from the MIGHTEE survey. We use the standard  $\Lambda$ -cold dark matter cosmology with a Hubble constant

**Table 1.** The input parameters of the HIMF model used for our simulation of an HI galaxy sample, along with the prior range used when running MULTINEST.  $\phi_*$  has units  $\text{Mpc}^{-3} \text{dex}^{-1}$ , while the  $M_*$ ,  $M_{\min}$ , and  $M_{\max}$  are in units of  $M_\odot$ . For the input value of the cosmological HI density we integrate the input HIMF over the range  $7.5 < \log_{10}(M_{\text{HI}}/M_\odot) < 12.5$  and divide by the critical density [see equation (4)].

Parameter	Input	Prior probability distribution
$\log_{10}(\phi_*)$	− 2.318	Uniform $\in [-5, 0]$
$\log_{10}(M_*)$	9.96	Uniform $\in [7, 12]$
$\alpha$	− 1.33	Uniform $\in [-5, 0]$
$\log_{10}(M_{\min})$	7.5	Uniform $\in [5, 10]$
$\log_{10}(M_{\max})$	12.5	Uniform $\in [8, 13]$
$\beta$	1.0	Uniform $\in [-1, 4]$
$\Omega_{\text{HI}} \times 10^4$	4.85	$0 < z < 0.1$
	5.24	$0.1 < z < 0.2$
	5.67	$0.2 < z < 0.3$
	6.12	$0.3 < z < 0.4$
	6.56	$0.4 < z < 0.5$
	6.89	$0.5 < z < 0.55$

$H_0 = 67.7 \text{ km s}^{-1} \text{ Mpc}^{-1}$ , total matter density  $\Omega_m = 0.308$ , and dark energy density  $\Omega_\Lambda = 0.692$  (Planck Collaboration et al. 2016).

## 2 METHOD

To simulate an HI cube, we first need to construct the HI flux distribution assuming a source-count model and the expected noise properties for a typical survey.

### 2.1 HIMF model

Under the usual assumptions (e.g. optically thin gas), the HI mass can be converted to the integrated flux (e.g. Meyer et al. 2017) via

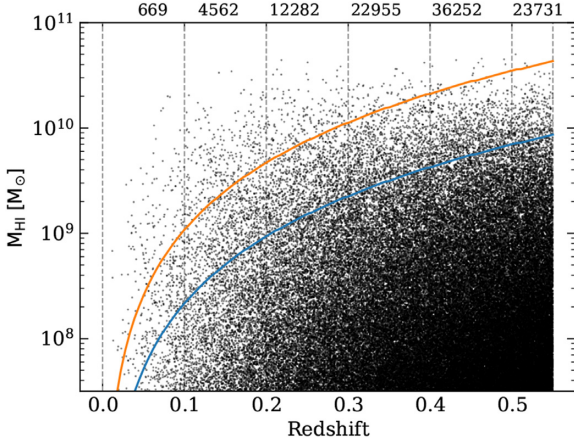
$$M_{\text{HI}} = 2.356 \times 10^5 D_L^2 (1+z)^{-1} S, \quad (1)$$

where the HI mass,  $M_{\text{HI}}$ , is in solar masses, the luminosity distance to the galaxy,  $D_L$ , is in Mpc, and the integrated flux  $S$  is in  $\text{Jy km s}^{-1}$ . The  $(1+z)$  factor is needed when  $S$  is expressed in units of  $\text{Jy km s}^{-1}$  rather than  $\text{Jy Hz}$ .

For the HIMF, we adopt a Schechter function model, which has been shown to fit the  $z \sim 0$  HIMF from both HIPASS (Zwaan et al. 2005) and ALFALFA (Jones et al. 2018), along with a pure density evolution term to characterize the evolution with redshift. Although a strong evolution of HI mass is not expected from the little information we have, including this term does verify the flexibility of our approach. The specific form we adopt is therefore given by

$$\phi(M_{\text{HI}}, z) = \ln(10)\phi_* \left(\frac{M_{\text{HI}}}{M_*}\right)^{\alpha+1} e^{-\frac{M_{\text{HI}}}{M_*}} (1+z)^\beta, \quad (2)$$

where  $\phi_*$ ,  $M_*$ ,  $\alpha$ , and  $\beta$  correspond to the normalization, characteristic mass, faint-end slope, and power of the redshift evolution, respectively. The HIMF represents the intrinsic number density of galaxies in the Universe as a function of their HI mass at a given redshift. The adopted values of these parameters for our models are given in Table 1.



**Figure 1.** Simulated HI mass versus redshift over a  $1 \text{ deg}^2$  area. The colour-coded lines indicate the  $\sigma_n = \sqrt{N_{\text{ch}}}\sigma_{\text{ch}}dv$  and  $5\sigma_n$  detection threshold. The numbers listed at the top of the figure indicate the number of galaxies in each redshift bin.

We then define the volume-weighted average HIMF over the redshift bin defined by  $z_1 < z < z_2$  by

$$\Phi(M_{\text{HI}}, z_1, z_2) = \frac{\int_{z_1}^{z_2} \frac{dV}{dz} \phi(M_{\text{HI}}, z) dz}{\int_{z_1}^{z_2} \frac{dV}{dz} dz}, \quad (3)$$

where  $dV/dz$  is the differential comoving volume at redshift  $z$ . The HI mass density of the Universe  $\Omega_{\text{HI}}$  from  $z_1$  to  $z_2$  can then be estimated by integrating the  $\Phi(M_{\text{HI}}, z_1, z_2)$ , which gives

$$\Omega_{\text{HI}} = \frac{1}{\rho_c} \int_{\log_{10}(M_{\text{min}})}^{\log_{10}(M_{\text{max}})} M_{\text{HI}} \Phi(M_{\text{HI}}, z_1, z_2) d \log_{10}(M_{\text{HI}}), \quad (4)$$

where  $\rho_c$  is the critical density of the Universe at  $z = 0$ .

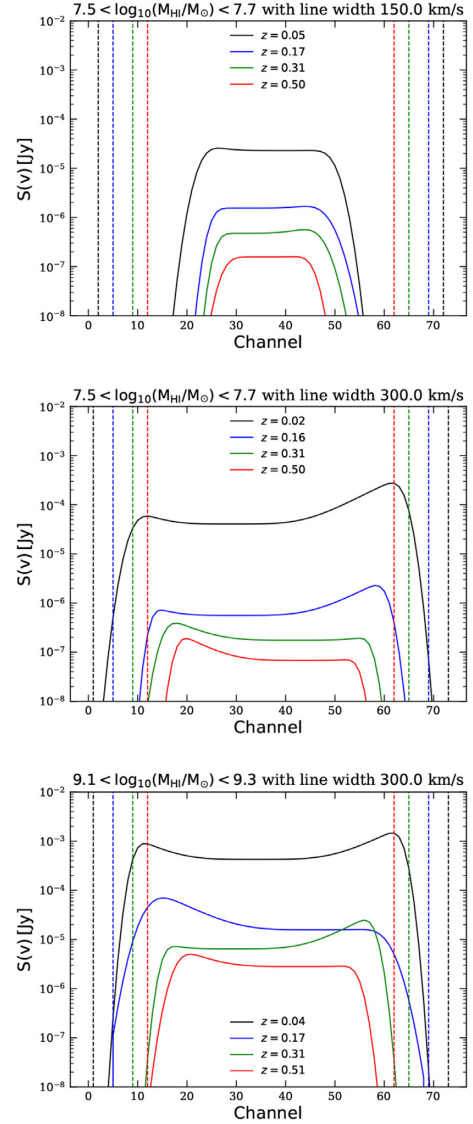
## 2.2 Generating the galaxy sample

Here, we simulate what we expect from the MIGHTEE survey as an example of a future deep HI survey. Using the HIMF described previously with the parameters shown in Table 1, we generate a mass distribution of simulated galaxies over a volume similar to that expected for the MIGHTEE survey (i.e.  $\approx 20 \text{ deg}^2$  over the redshift range  $0 < z < 0.55$ ). The cube has a spatial resolution of 5 arcsec and a total of 19 504 26 kHz channels from 913.4 to 1420.5 MHz.

Fig. 1 shows the HI mass against redshift for the generated galaxy sample. As we are aiming to explore the HIMF below the nominal noise threshold, the galaxy sample we use is volume-limited. However, we note that in reality we will require an input catalogue of galaxies in order to know where to extract fluxes from within the data cube, and the input catalogue itself will be flux limited. Therefore, any correlation that exists between the input catalogue selection and the HI mass of the galaxies will effect our ability to measure the HIMF and we return to this in Section 3.

## 2.3 Simulating HI emission line profiles

Once the galaxy samples are generated, we simulate the flux density,  $S_m(v)$ , of HI emission lines using the general form of the ‘busy function’ with a fourth-degree polynomial trough [i.e. equation (4) in Westmeier et al. 2014]. The parameters  $b_1$ ,  $b_2$ ,  $c$ ,

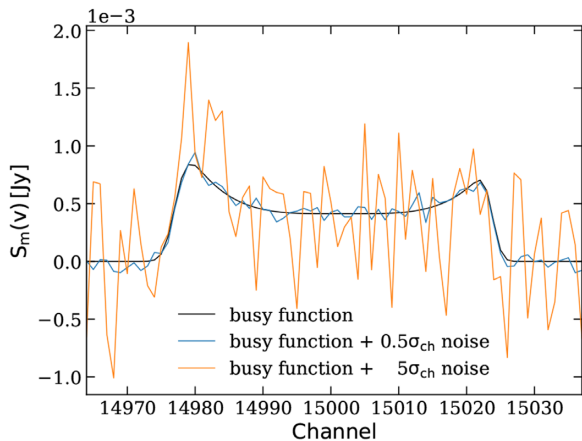


**Figure 2.** Flux density as a function of channel for galaxies with a mass of  $\log_{10}(M_{\text{HI}}/M_{\odot}) = 7.5$  (top and middle panels) and  $\log_{10}(M_{\text{HI}}/M_{\odot}) = 9.2$  (bottom panel). The colours indicate different redshifts. The vertical dashed lines show the limits of extracting the fluxes. Two different line widths are also shown  $2 \times w = 150 \text{ km s}^{-1}$  (top panel) and  $300 \text{ km s}^{-1}$  (middle and bottom panels). The central channel of the spectra is fixed at channel 37 in these examples.

$x_p$ , and  $w$  are randomly sampled to assign a variety of line profiles to the HI emission lines. In Fig. 2, we show the flux density as a function of channel for galaxies of a given mass. The major variations are the steepness of line flanks, amplitude of the troughs, centroid of polynomials, and widths of the profile. The difference between centroids of the error function and polynomial indicates the asymmetry of line profile, while the line widths approximate different galaxy inclinations. Although the form of lines changes, the integrated fluxes always keep constant for the sources with the same mass and redshift. The ranges used for these parameters are listed in Table 2. We note that these ranges are set to be fairly broad rather than be realistic in order to prove the robustness of our approach. In particular, a double-horn profile with the line width of  $150 \text{ km s}^{-1}$  for the low HI mass samples may not be expected

**Table 2.** The parameters of the busy function for simulating the H I lines. Note that only parameters  $b_1$ ,  $b_2$ ,  $c$ ,  $x_p$ , and  $w$  are randomly sampled, the centroid of the error function,  $x_e$ , is fixed at the redshift of the source.  $x_p$  and  $x_e$  are in units of channel, while the line width has units  $\text{km s}^{-1}$ .

Parameter	Meaning	Probability distribution
$b_1, b_2$	Steepness of line flanks	Uniform $\in [0.3, 0.7]$
$c \times 10^5$	Amplitude of the trough	Uniform $\in [0.3, 0.7]$
$ x_p - x_e $	Difference of centroids	Uniform $\in [0, 10]$
$2 \times w$	Line width	Uniform $\in [55, 330]$



**Figure 3.** A simulated flux density as a function of channel generated by the busy function with random parameters. The black line shows the flux density without noise. The blue and orange lines indicate the flux density with  $0.5\sigma_{\text{ch}}$ ,  $5\sigma_{\text{ch}}$  Gaussian noise.

from observations (e.g. Ianjamasimanana et al. 2012), nevertheless one can find many Gaussian profiles easily from our simulations if looking at the flux density with narrower line widths.

We then generate the Gaussian noise with standard deviation similar to that expected from the MIGHTEE survey. Specifically we adopt  $\sigma_{\text{ch}} = 90 \mu\text{Jy}$  per channel, which is the expected noise per channel with a channel width of 26 kHz for the MIGHTEE surveys and  $0.5\sigma_{\text{ch}}$  as an indication of the improvement possible with a deeper survey, such as LADUMA. We also test the case for a much higher noise of  $5\sigma_{\text{ch}}$ . To determine the total flux density per channel across the spectral range,  $S_m(v)$ , we simply add the noise to the profile of the busy function and write the total flux density into the simulated cube. We note that only point sources are simulated in this work since our approach is aimed towards the high-redshift regime where unresolved sources are expected to dominate given the spatial resolution of current telescopes.

In Fig. 3, we show the simulated emission line as a function of the channel with  $0.5\sigma_{\text{ch}}$  and  $5\sigma_{\text{ch}}$  Gaussian noise. We do not know a priori the shape of the line profile, we therefore fix the line limits that we measure the flux over at a velocity width  $N_{\text{ch}} \times dv = 406 \text{ km s}^{-1}$  which corresponds to  $N_{\text{ch}} = 74$  at  $z = 0$  and  $N_{\text{ch}} = 48$  at  $z = 0.55$ . This ensures that we fully cover the range of expected line profiles. The integrated flux is then given by  $S_m = \sum_{N_{\text{ch}}} S_m(v) dv$ , where  $dv = 5.5 \text{ km s}^{-1}$  at  $z = 0$ . We note that in our idealized case of uniform Gaussian noise  $\sigma_{\text{ch}}$ , the uncertainty associated with summing over all channels is given by  $\sigma_n = \sqrt{N_{\text{ch}}}\sigma_{\text{ch}}dv$  and is dependent on redshift (i.e. the  $dv$  increases as  $1+z$  and therefore the number of channels required to sample  $406 \text{ km s}^{-1}$  at the redshift of the source decreases as  $1+z$  as shown in Fig. 2, thus the noise increases at  $\sqrt{1+z}$ ). This is reasonable for our simulated cube as

we already know that the noise follows a Gaussian distribution, but it might be problematic when we deal with real data, where the noise is likely to vary as a function of spectral window, and in such cases we would have to fit the noise distribution as a function of redshift. Furthermore, we would also not expect the signal-to-noise ratio to be constant across the primary beam and return to this point in Section 4.

With the known positions and redshifts from deep optical and/or near-infrared spectroscopy we are able to extract the flux across the number of channels that we expect to encompass the full H I emission line and determine the source counts, which we subsequently model. Note that in this study we do not consider the effects of confusion since the new generation of deep H I surveys will achieve 5–15 arcsec spatial resolution and have a spectral resolution of a few  $10 \text{ km s}^{-1}$ . Indeed, if we consider how many ‘volumetric beams’ (or voxels; where in this case a voxel is the beam area multiplied by the number of channels that we integrate over for the H I line) per source we have given our choice of evolving mass function, then we estimate there are around 200 independent ‘voxels’ per source, which is well above the nominal confusion level. We also note that H I galaxies are not a strongly clustered population (e.g. Papastergis et al. 2013), and assume that lower mass H I galaxies are actually less clustered in general. As such our simulation of randomly positioned galaxies is probably not far from reality and would still be affected by confusion if it were an issue. This would result in not accurately recovering the parameters of the input HIMF. However, we do note that close pairs of galaxies in groups will inevitably lead to some small level of confusion, but given the numbers involved this will be an insignificant perturbation on the derived parameters.

We also note that our method could be extended to incorporate confusion noise, either by building it into the model itself (e.g. Chen, Zwart & Santos 2017), or via alternative methods of measuring the noise properties of the spectral cubes at the position of each source (Section 4).

## 2.4 The likelihood function

Based on Bayes’ theorem, the probability of the model given the data is proportional to the likelihood function (i.e. probability of the data given the model). With the assumption that the number of sources in different flux bins are independent, the likelihood for all the flux bins is given by

$$\mathcal{L} \propto \prod_{i=\text{bin}} P_i(k_i | dN/dS), \quad (5)$$

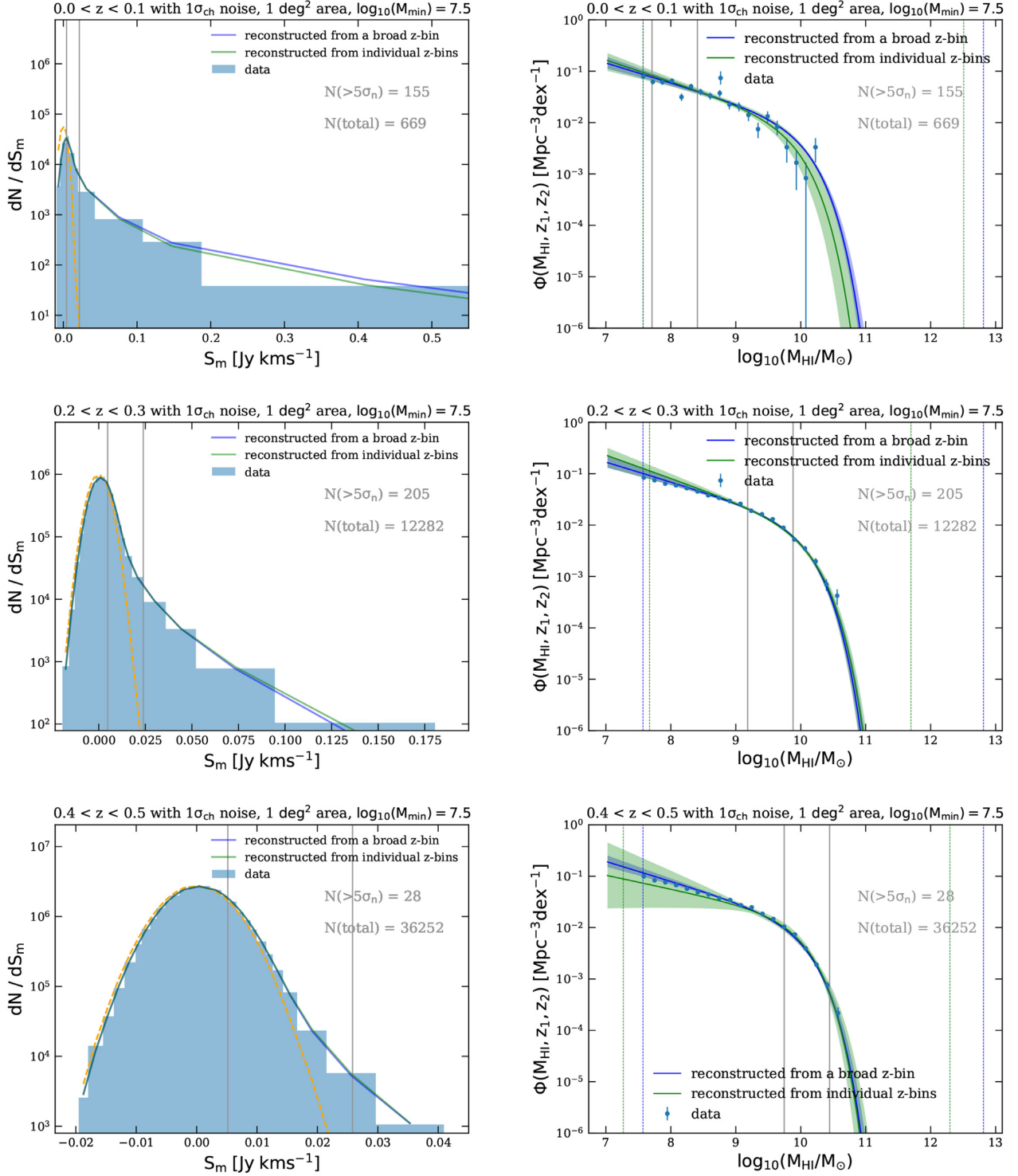
where  $P_i(k_i | dN/dS)$  is the probability of obtaining  $k_i$  objects in bin  $i$  given a source-count model for the H I flux distribution  $dN/dS$  (see e.g. Mitchell-Wynne et al. 2014; Zwart et al. 2015).

The measured flux from an H I emission line,  $S_m$ , is a combination of the intrinsic flux from the source,  $S$  plus the contribution from the noise,  $S_n$ , which can obviously be positive or negative.

With the assumption that the number of sources in a given patch of sky follows a Poisson distribution, the probability of finding  $k_i$  objects over the observed sky area in the  $i$ th flux bin  $[S_{m_i}, S_{m_i} + \Delta S_{m_i}]$  is given by

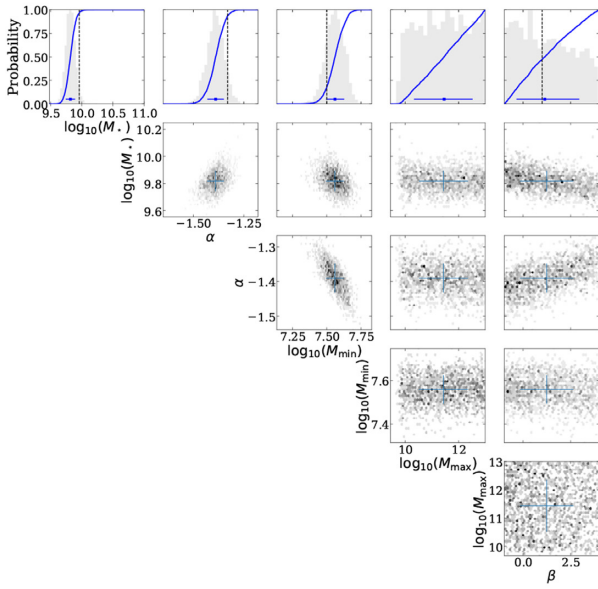
$$P_i(k_i | dN/dS) = \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!}, \quad (6)$$

where  $dN/dS$  is our source-count model for the H I galaxies in the observed area and  $\lambda_i$  is the theoretically expected number of H I sources in that bin.

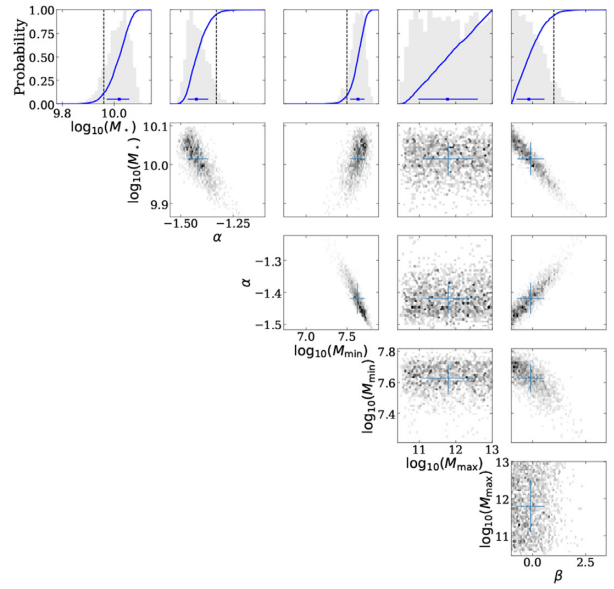


**Figure 4.** Reconstructed flux  $S_m$  distribution (left) and HIMF (right) in different redshift bins from our baseline survey. The blue lines show the results from a broad redshift bin (i.e.  $0 < z < 0.55$ ), while the green lines show the results from an individual redshift bin. The grey vertical lines indicate the  $\sigma_n = \sqrt{N_{\text{ch}}}\sigma_{\text{ch}}\Delta\nu$  and formal  $5\sigma_n$  detection threshold at the centre of each redshift bin. Only three redshift bins are shown for simplicity. For the left-hand panels, we apply the Bayesian Blocks method (Scargle et al. 2013) that allows varying bin widths to the data. The dashed orange line denotes a Gaussian distribution with a width equal to  $\sigma_n$ . For the right-hand panels, the colour-coded regions are the 68 per cent credible intervals in the HIMF estimated from the posterior samples. The colour-coded vertical lines are  $\log_{10}(M_{\text{min}})$  and  $\log_{10}(M_{\text{max}})$  of the best fitting from a broad redshift bin and an individual redshift bin. The errors on the simulated data points are just Poisson errors.

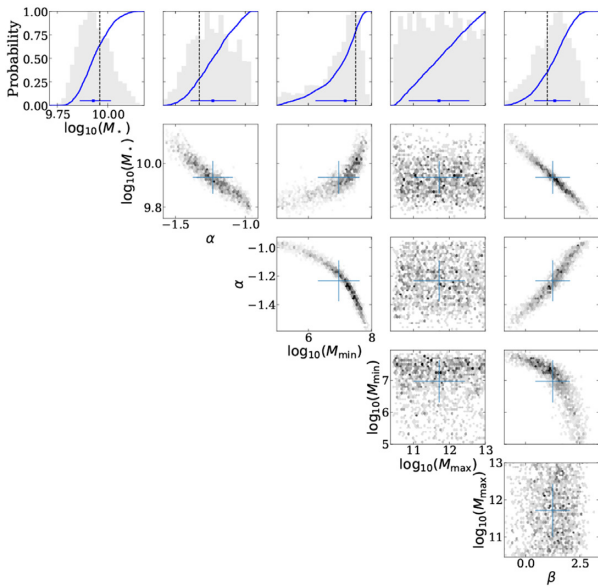
0.0 < z < 0.1 with  $1\sigma_{\text{ch}}$  noise, 1 deg<sup>2</sup> area,  $\log_{10}(M_{\text{min}}) = 7.5$



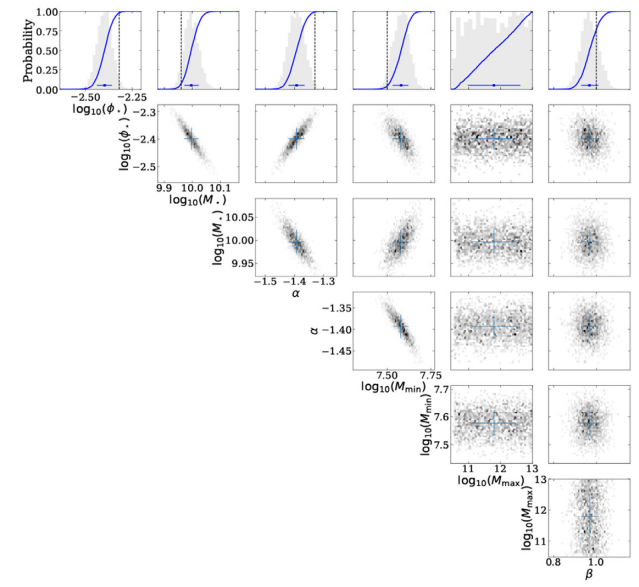
0.2 < z < 0.3 with  $1\sigma_{\text{ch}}$  noise, 1 deg<sup>2</sup> area,  $\log_{10}(M_{\text{min}}) = 7.5$



0.4 < z < 0.5 with  $1\sigma_{\text{ch}}$  noise, 1 deg<sup>2</sup> area,  $\log_{10}(M_{\text{min}}) = 7.5$



0.0 < z < 0.55 with  $1\sigma_{\text{ch}}$  noise, 1 deg<sup>2</sup> area,  $\log_{10}(M_{\text{min}}) = 7.5$



**Figure 5.** The posterior distributions for HIMF model parameters for three individual redshift bins for our baseline survey (top and bottom left panels) and for the single broad redshift bin (bottom right panel). The grey histograms are the (1D or 2D marginal posterior probability densities). The blue curves are the cumulative distribution (integrating over the grey histograms from left to right). The best-fitting parameters are shown by the blue dots with  $1\sigma$  error bars, while the input parameter values are the vertical black lines. Please refer to Fig. 9 for bigger labels.

If the noise follows a given distribution  $P_n(S_n)$ , we can write  $\lambda_i$  as the convolution of  $dN/dS$  and  $P_n(S_n)$

$$\lambda_i = \int_{S_{\text{min}}}^{S_{\text{max}}} dS \frac{dN}{dS} \int_{S_{m_i}}^{S_{m_i} + \Delta S_{m_i}} dS_m P_n(S_m - S). \quad (7)$$

where  $S_{\text{min}}$  and  $S_{\text{max}}$  are the minimum and maximum flux that can be measured from the data. These correspond to  $M_{\text{min}}$  and  $M_{\text{max}}$  in our model via equation (1).

For the idealized case of where the noise follows a Gaussian distribution centred at zero with a variance  $\sigma_n^2$ , equation (7) becomes

$$\begin{aligned} \lambda_i &= \int_{S_{\text{min}}}^{S_{\text{max}}} dS \frac{dN}{dS} \int_{S_{m_i}}^{S_{m_i} + \Delta S_{m_i}} dS_m \frac{1}{\sigma_n \sqrt{2\pi}} e^{-\frac{(S_m - S)^2}{2\sigma_n^2}} \\ &= \int_{S_{\text{min}}}^{S_{\text{max}}} dS \frac{dN}{dS} \frac{1}{2} \left[ \text{erf} \left( \frac{S_{m_i} - S}{\sqrt{2}\sigma_n} \right) - \text{erf} \left( \frac{S_{m_i} + \Delta S_{m_i} - S}{\sqrt{2}\sigma_n} \right) \right], \quad (8) \end{aligned}$$

**Table 3.** Reconstructed parameters of the HIMF model from our baseline survey with changing background noise, where we fix  $\log_{10}(\phi_*) = -2.318$  for individual redshift bins.

Parameter	$\times \sigma_{\text{ch}}$	$0 < z < 0.1$	$0.1 < z < 0.2$	$0.2 < z < 0.3$	$0.3 < z < 0.4$	$0.4 < z < 0.5$	$0.5 < z < 0.55$	$0 < z < 0.55$
$\log_{10}(\phi_*)$	0.5	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	-2.338 <sup>+0.026</sup> <sub>-0.027</sub>
$\log_{10}(\phi_*)$	1.0	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	-2.397 <sup>+0.039</sup> <sub>-0.042</sub>
$\log_{10}(\phi_*)$	5.0	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	-2.378 <sup>+0.131</sup> <sub>-0.152</sub>
$\log_{10}(M_*)$	0.5	9.819 <sup>+0.077</sup> <sub>-0.071</sub>	9.977 <sup>+0.04</sup> <sub>-0.051</sub>	10.014 <sup>+0.038</sup> <sub>-0.044</sub>	9.955 <sup>+0.047</sup> <sub>-0.042</sub>	9.914 <sup>+0.043</sup> <sub>-0.042</sub>	9.905 <sup>+0.064</sup> <sub>-0.066</sub>	9.966 <sup>+0.019</sup> <sub>-0.018</sub>
$\log_{10}(M_*)$	1.0	9.819 <sup>+0.075</sup> <sub>-0.075</sub>	9.955 <sup>+0.051</sup> <sub>-0.061</sub>	10.02 <sup>+0.039</sup> <sub>-0.048</sub>	9.962 <sup>+0.071</sup> <sub>-0.068</sub>	9.928 <sup>+0.088</sup> <sub>-0.067</sub>	9.918 <sup>+0.089</sup> <sub>-0.098</sub>	9.996 <sup>+0.027</sup> <sub>-0.025</sub>
$\log_{10}(M_*)$	5.0	9.887 <sup>+0.178</sup> <sub>-0.091</sub>	9.932 <sup>+0.053</sup> <sub>-0.066</sub>	9.955 <sup>+0.073</sup> <sub>-0.081</sub>	9.765 <sup>+0.441</sup> <sub>-0.407</sub>	9.814 <sup>+1.237</sup> <sub>-0.391</sub>	9.611 <sup>+0.821</sup> <sub>-0.499</sub>	9.978 <sup>+0.095</sup> <sub>-0.08</sub>
$\alpha$	0.5	-1.379 <sup>+0.036</sup> <sub>-0.037</sub>	-1.38 <sup>+0.034</sup> <sub>-0.024</sub>	-1.413 <sup>+0.04</sup> <sub>-0.034</sub>	-1.355 <sup>+0.05</sup> <sub>-0.053</sub>	-1.271 <sup>+0.068</sup> <sub>-0.066</sub>	-1.264 <sup>+0.132</sup> <sub>-0.107</sub>	-1.356 <sup>+0.018</sup> <sub>-0.017</sub>
$\alpha$	1.0	-1.389 <sup>+0.041</sup> <sub>-0.041</sub>	-1.358 <sup>+0.053</sup> <sub>-0.042</sub>	-1.425 <sup>+0.056</sup> <sub>-0.041</sub>	-1.347 <sup>+0.108</sup> <sub>-0.098</sub>	-1.233 <sup>+0.166</sup> <sub>-0.159</sub>	-1.293 <sup>+0.245</sup> <sub>-0.181</sub>	-1.393 <sup>+0.028</sup> <sub>-0.028</sub>
$\alpha$	5.0	-1.291 <sup>+0.09</sup> <sub>-0.073</sub>	-1.388 <sup>+0.111</sup> <sub>-0.085</sub>	-1.207 <sup>+0.139</sup> <sub>-0.193</sub>	-1.111 <sup>+0.107</sup> <sub>-1.639</sub>	-1.532 <sup>+0.534</sup> <sub>-2.203</sub>	-1.197 <sup>+0.18</sup> <sub>-1.783</sub>	-1.365 <sup>+0.105</sup> <sub>-0.111</sub>
$\log_{10}(M_{\text{min}})$	0.5	7.539 <sup>+0.041</sup> <sub>-0.052</sub>	7.557 <sup>+0.038</sup> <sub>-0.044</sub>	7.626 <sup>+0.05</sup> <sub>-0.06</sub>	7.569 <sup>+0.086</sup> <sub>-0.095</sub>	7.401 <sup>+0.149</sup> <sub>-0.198</sub>	7.356 <sup>+0.249</sup> <sub>-0.455</sub>	7.547 <sup>+0.028</sup> <sub>-0.029</sub>
$\log_{10}(M_{\text{min}})$	1.0	7.56 <sup>+0.068</sup> <sub>-0.066</sub>	7.544 <sup>+0.069</sup> <sub>-0.082</sub>	7.636 <sup>+0.078</sup> <sub>-0.096</sub>	7.487 <sup>+0.18</sup> <sub>-0.255</sub>	7.169 <sup>+0.383</sup> <sub>-0.943</sub>	7.337 <sup>+0.406</sup> <sub>-1.189</sub>	7.58 <sup>+0.042</sup> <sub>-0.049</sub>
$\log_{10}(M_{\text{min}})$	5.0	7.306 <sup>+0.197</sup> <sub>-0.32</sub>	7.676 <sup>+0.167</sup> <sub>-0.246</sub>	6.922 <sup>+0.641</sup> <sub>-0.981</sub>	6.476 <sup>+2.002</sup> <sub>-0.786</sub>	8.148 <sup>+0.677</sup> <sub>-0.621</sub>	7.004 <sup>+1.416</sup> <sub>-1.053</sub>	7.502 <sup>+0.174</sup> <sub>-0.22</sub>
$\log_{10}(M_{\text{max}})$	0.5	11.449 <sup>+1.04</sup> <sub>-1.124</sub>	11.649 <sup>+0.855</sup> <sub>-0.961</sub>	11.825 <sup>+0.82</sup> <sub>-0.868</sub>	11.686 <sup>+0.874</sup> <sub>-0.915</sub>	11.79 <sup>+0.809</sup> <sub>-0.914</sub>	11.803 <sup>+0.839</sup> <sub>-0.881</sub>	11.693 <sup>+0.894</sup> <sub>-0.87</sub>
$\log_{10}(M_{\text{max}})$	1.0	11.46 <sup>+1.066</sup> <sub>-1.129</sub>	11.666 <sup>+0.93</sup> <sub>-0.932</sub>	11.773 <sup>+0.829</sup> <sub>-0.8</sub>	11.709 <sup>+0.831</sup> <sub>-0.861</sub>	11.712 <sup>+0.839</sup> <sub>-0.84</sub>	11.783 <sup>+0.757</sup> <sub>-0.852</sub>	11.794 <sup>+0.829</sup> <sub>-0.819</sub>
$\log_{10}(M_{\text{max}})$	5.0	11.234 <sup>+1.199</sup> <sub>-1.307</sub>	11.733 <sup>+0.852</sup> <sub>-0.965</sub>	11.783 <sup>+0.867</sup> <sub>-0.856</sub>	10.866 <sup>+1.441</sup> <sub>-1.017</sub>	10.288 <sup>+1.541</sup> <sub>-0.49</sub>	10.652 <sup>+1.433</sup> <sub>-1.061</sub>	11.747 <sup>+0.854</sup> <sub>-0.903</sub>
$\beta$	0.5	1.103 <sup>+1.848</sup> <sub>-1.51</sub>	-0.02 <sup>+0.984</sup> <sub>-0.683</sub>	0.002 <sup>+0.606</sup> <sub>-0.561</sub>	0.947 <sup>+0.463</sup> <sub>-0.529</sub>	1.467 <sup>+0.416</sup> <sub>-0.426</sub>	1.418 <sup>+0.602</sup> <sub>-0.614</sub>	0.972 <sup>+0.039</sup> <sub>-0.042</sub>
$\beta$	1.0	1.142 <sup>+1.816</sup> <sub>-1.493</sub>	0.591 <sup>+1.258</sup> <sub>-1.019</sub>	-0.173 <sup>+0.733</sup> <sub>-0.579</sub>	0.75 <sup>+0.859</sup> <sub>-0.892</sub>	1.346 <sup>+0.73</sup> <sub>-0.951</sub>	1.098 <sup>+0.999</sup> <sub>-0.938</sub>	0.968 <sup>+0.042</sup> <sub>-0.039</sub>
$\beta$	5.0	1.596 <sup>+1.62</sup> <sub>-1.693</sub>	1.013 <sup>+1.792</sup> <sub>-1.395</sub>	1.163 <sup>+1.377</sup> <sub>-1.45</sub>	2.333 <sup>+1.11</sup> <sub>-1.776</sub>	1.876 <sup>+1.42</sup> <sub>-1.698</sub>	1.966 <sup>+1.268</sup> <sub>-1.519</sub>	0.967 <sup>+0.042</sup> <sub>-0.04</sub>
$\Omega_{\text{H I}} \times 10^4$ (individual)	0.5	3.76 <sup>+1.13</sup> <sub>-0.89</sub>	4.79 <sup>+0.67</sup> <sub>-0.3</sub>	5.7 <sup>+0.24</sup> <sub>-0.35</sub>	6.02 <sup>+0.36</sup> <sub>-0.24</sub>	6.63 <sup>+0.28</sup> <sub>-0.35</sub>	6.71 <sup>+0.58</sup> <sub>-0.44</sub>	
$\Omega_{\text{H I}} \times 10^4$ (individual)	1.0	3.68 <sup>+1.08</sup> <sub>-0.81</sub>	5.0 <sup>+0.54</sup> <sub>-0.4</sub>	5.58 <sup>+0.41</sup> <sub>-0.31</sub>	5.85 <sup>+0.43</sup> <sub>-0.37</sub>	6.29 <sup>+0.6</sup> <sub>-0.52</sub>	6.39 <sup>+0.98</sup> <sub>-0.81</sub>	
$\Omega_{\text{H I}} \times 10^4$ (individual)	5.0	4.05 <sup>+1.1</sup> <sub>-1.29</sub>	5.24 <sup>+0.65</sup> <sub>-0.59</sub>	5.28 <sup>+1.09</sup> <sub>-0.87</sub>	4.94 <sup>+1.35</sup> <sub>-1.38</sub>	8.12 <sup>+2.13</sup> <sub>-2.09</sub>	6.07 <sup>+3.38</sup> <sub>-3.84</sub>	
$\Omega_{\text{H I}} \times 10^4$ (broad)	0.5	4.84 <sup>+0.17</sup> <sub>-0.16</sub>	5.21 <sup>+0.16</sup> <sub>-0.16</sub>	5.62 <sup>+0.16</sup> <sub>-0.16</sub>	6.05 <sup>+0.16</sup> <sub>-0.16</sub>	6.47 <sup>+0.17</sup> <sub>-0.16</sub>	6.78 <sup>+0.19</sup> <sub>-0.17</sub>	
$\Omega_{\text{H I}} \times 10^4$ (broad)	1.0	4.72 <sup>+0.2</sup> <sub>-0.17</sub>	5.09 <sup>+0.19</sup> <sub>-0.18</sub>	5.5 <sup>+0.2</sup> <sub>-0.18</sub>	5.91 <sup>+0.2</sup> <sub>-0.18</sub>	6.34 <sup>+0.21</sup> <sub>-0.2</sub>	6.64 <sup>+0.22</sup> <sub>-0.22</sub>	
$\Omega_{\text{H I}} \times 10^4$ (broad)	5.0	4.64 <sup>+0.4</sup> <sub>-0.48</sub>	5.0 <sup>+0.43</sup> <sub>-0.5</sub>	5.4 <sup>+0.47</sup> <sub>-0.53</sub>	5.8 <sup>+0.5</sup> <sub>-0.57</sub>	6.21 <sup>+0.54</sup> <sub>-0.61</sub>	6.5 <sup>+0.57</sup> <sub>-0.64</sub>	

where  $\sigma_n = \sqrt{N_{\text{ch}}}\sigma_{\text{ch}}dv$ , and  $N_{\text{ch}}$  is the number of channels over the flux-density profile of the H I line for each galaxy,  $\sigma_{\text{ch}}$  is the rms per channel, and  $dv$  is the velocity width of each channel.

We then relate the source-count model with the redshift-dependent HIMF,  $\phi(M_{\text{H I}}, z)$ , by

$$\begin{aligned} \frac{dN}{dS} &= \frac{dN}{dM_{\text{H I}}} \frac{dM_{\text{H I}}}{dS} \\ &= \frac{dN}{d \log_{10}(M_{\text{H I}})} \frac{d \log_{10}(M_{\text{H I}})}{dM_{\text{H I}}} \frac{dM_{\text{H I}}}{dS} \\ &= \int_{z_1}^{z_2} \frac{dV}{dz} \phi(M_{\text{H I}}, z) \frac{\log_{10}(e)}{M_{\text{H I}}} 2.356 \times 10^5 D_L^2 (1+z)^{-1} dz, \quad (9) \end{aligned}$$

where  $z_1$  and  $z_2$  indicate the redshift limits for a given redshift bin and  $dV = r^2 \Omega dr$  where  $r$  is the comoving distance and  $\Omega$  is the solid angle of a given survey.

## 2.5 Parameter estimation

Here, we describe our approach to constrain the parameters of the input mass function so that we compare the best-fitting parameters with the mass function parameters in order to test if we could recover them.

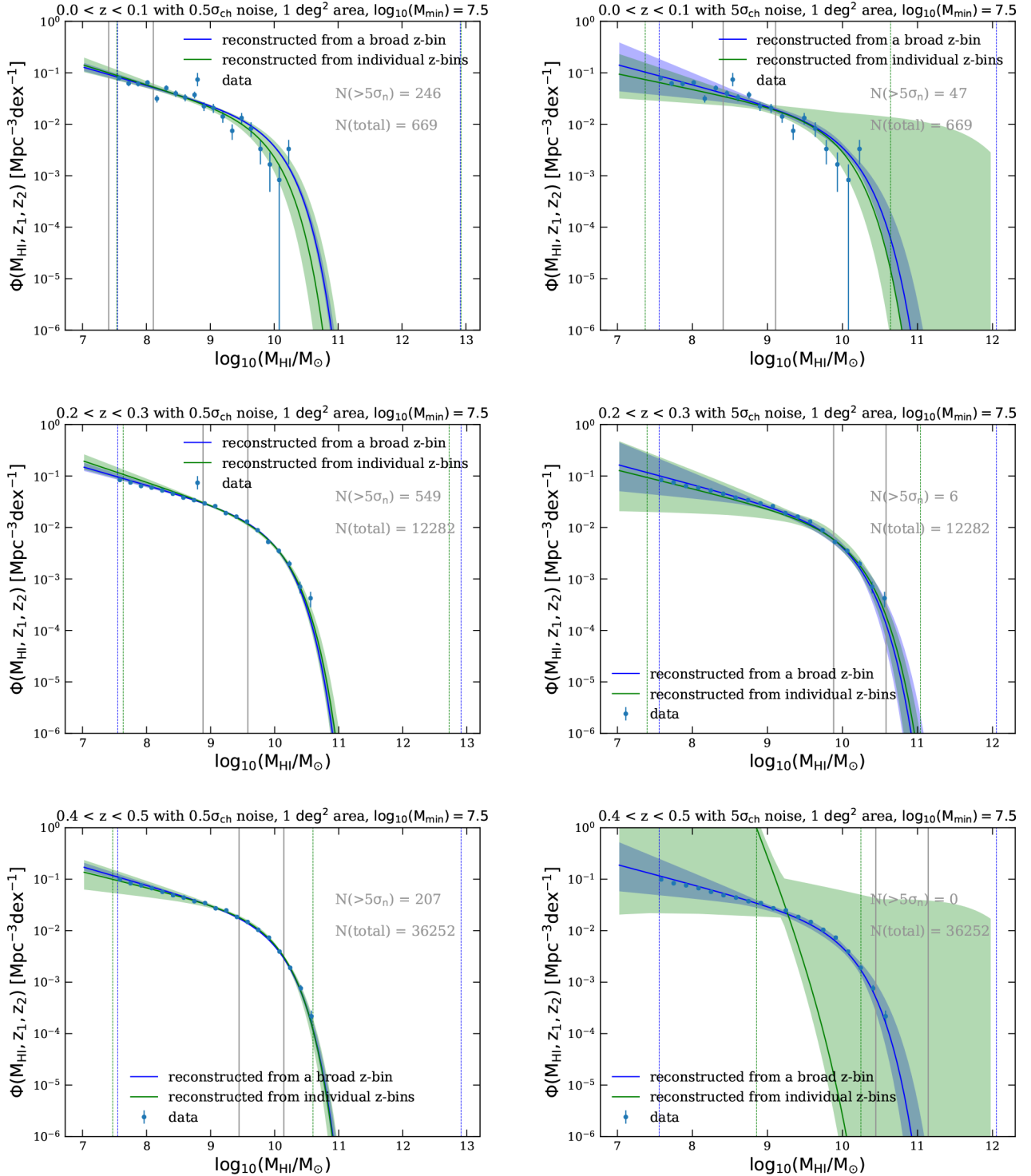
### 2.5.1 Priors

Priors are the prior knowledge or limits of a parameter before any relevant evidence is taken into account. In Bayesian statistical inference, they provide the sampling parameter space for computing the posterior probability. We assign a uniform prior probability distribution to the power terms  $\alpha$  and  $\beta$  and adopt uniform logarithmic priors for  $\phi_*$ ,  $M_*$ ,  $M_{\text{min}}$ , and  $M_{\text{max}}$ . These are listed in Table 1.

### 2.5.2 MULTINEST

In order to determine the posterior probability distribution and fit the model, we use MULTINEST<sup>1</sup> to sample the prior parameter space (Feroz, Hobson & Bridges 2009; Buchner et al. 2014). MULTINEST is an efficient and robust Bayesian inference tool based on a nested sampling technique (Skilling 2004), which produces the posterior samples from distributions with an associated error estimate and the Bayesian evidence term allowing model selection. In this paper we accept the median of posterior samples as the best estimate of each parameter considering the complex nature of the posterior, and the set of parameters with the maximum likelihood as the best-fitting

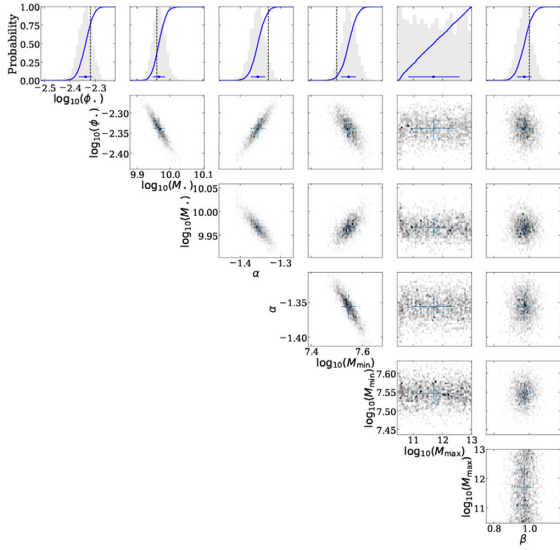
<sup>1</sup><https://github.com/JohannesBuchner/PyMultiNest>



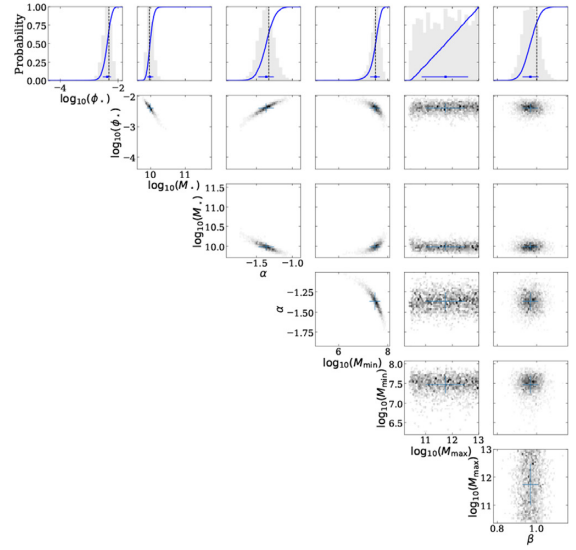
**Figure 6.** Reconstructed HIMF from our simulated survey with  $0.5\sigma_{\text{ch}}$  (left) and  $5\sigma_{\text{ch}}$  (right) Gaussian background noises in different redshift bins. The blue lines show results for a single broad redshift bin (i.e.  $0 < z < 0.55$ ), while the green lines show the results for individual redshift bins. The colour-coded regions are the 68 per cent credible intervals in the HIMF estimated from the posterior samples. The colour-coded vertical lines are  $\log_{10}(M_{\text{min}})$  and  $\log_{10}(M_{\text{max}})$  of the best fitting from a broad redshift bin and an individual redshift bin. The grey vertical lines on the left-hand panel indicate the  $\sigma_n = \sqrt{N_{\text{ch}}} \times 0.5\sigma_{\text{ch}}dv$  and the formal  $5\sigma_n$  detection threshold at the centre of each redshift bin, while those on the right-hand panel indicate the  $\sigma_n = \sqrt{N_{\text{ch}}} \times 5\sigma_{\text{ch}}dv$  and  $5\sigma_n$  detection threshold. The errors on the simulated data points are just Poisson errors. Only three redshift bins are shown for simplicity.



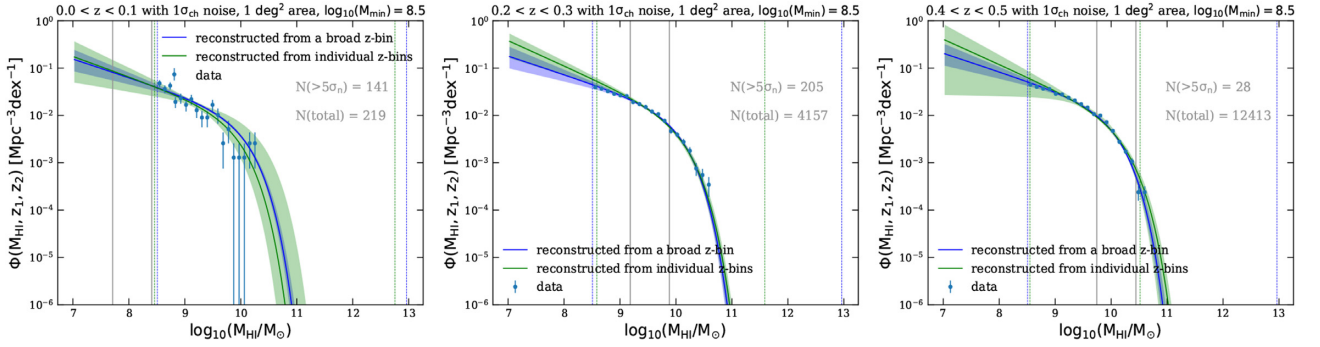
0.0 < z < 0.55 with 0.5 $\sigma_{\text{ch}}$  noise, 1 deg<sup>2</sup> area, log<sub>10</sub>(M<sub>min</sub>) = 7.5



0.0 < z < 0.55 with 5 $\sigma_{\text{ch}}$  noise, 1 deg<sup>2</sup> area, log<sub>10</sub>(M<sub>min</sub>) = 7.5



**Figure 7.** The posterior distributions for the HIMF model parameters, for a single broad redshift bin with 0.5  $\sigma_{\text{ch}}$  (left) and 5  $\sigma_{\text{ch}}$  (right) background noise levels. All other properties of this figure are the same as this in Fig. 5.



**Figure 8.** Reconstructed HIMF from our simulated survey with increasing minimum mass  $\log_{10}(M_{\text{min}}) = 8.5$  of the H I galaxy sample in different redshift bins. The blue lines show the results from a single broad redshift bin (i.e.  $0 < z < 0.55$ ), while the green lines show the results for individual redshift bins. The grey vertical lines indicate the  $\sigma_n = \sqrt{N_{\text{ch}}}\sigma_{\text{ch}}dv$  and  $5\sigma_n$  detection threshold at the centre of each redshift bin. All other properties of this figure are the same as that of Fig. 6.

HIMF. We do not compare different models, we therefore do not use the evidence term. However, we note that an obvious use of the evidence would be to determine the model which best describes the form of the evolution of the HIMF, and we leave this to future work using real data.

## 2.6 Survey parameters

For our baseline survey we assume a 1 deg<sup>2</sup> survey with noise of  $\sigma_{\text{ch}} = 90 \mu\text{Jy}$  per channel over the redshift range  $0 < z < 0.55$ . We then consider the effects of reducing and increasing the rms of the spectral line cube to investigate how the noise properties affect our ability to measure the HIMF accurately. Furthermore, increasing the background noise with a fixed extraction window is also similar to the situation where only photometric redshifts are available, as one has to extract a larger window for fully covering the line profile and this will inevitably integrate more noise into the measured flux. In real observations we are likely to be limited by an optical flux limit for the spectroscopic sample that is required in order to know

where to extract fluxes in the spectral line cube. In reality, the supporting spectroscopy will be limited by a complex optical flux limit that is dependent on galaxy colour, morphology, and emission line properties. For our purposes, we assume that this flux limit roughly corresponds to a stellar mass limit (although we note that we do not expect this to be a clean one-to-one relation) and many studies have shown that stellar mass and H I mass are correlated, particularly at  $M_{\text{HI}} < 10^9 M_{\odot}$  (e.g. fig. 1 in Maddox et al. 2015). We therefore also investigate how a higher minimum H I mass for our simulated galaxy sample may affect the recovered HIMF. Finally, we also investigate how the survey area influences the constraints by considering a range of surveys from 0.3 to 20 deg<sup>2</sup>.

## 3 RESULTS

We start by looking at the reconstructed parameters of our baseline survey and then investigate the effects of increasing noise, minimum mass, and survey area. Note that we include all simulated samples (i.e.  $M_{\text{HI}} > 10^{7.5} M_{\odot}$ ) across the redshift range  $0 < z < 0.55$  in the

**Table 4.** Reconstructed parameters of the HIMF model from our baseline survey with increasing minimum mass of the H I galaxy sample, where we fix  $\log_{10}(\phi_{\star}) = -2.318$  for individual redshift bins. We note that  $\Omega_{\text{H I}}$  is estimated by integrating the HIMF down to  $M_{\text{H I}} = 10^{7.5} M_{\odot}$  for both cases.

Parameter	$\log_{10}(M_{\text{min}})$ (input)	$0 < z < 0.1$	$0.1 < z < 0.2$	$0.2 < z < 0.3$	$0.3 < z < 0.4$	$0.4 < z < 0.5$	$0.5 < z < 0.55$	$0 < z < 0.55$
$\log_{10}(\phi_{\star})$	7.5	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	$-2.397^{+0.039}_{-0.042}$
$\log_{10}(\phi_{\star})$	8.5	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	$-2.37^{+0.055}_{-0.054}$
$\log_{10}(M_{\star})$	7.5	$9.819^{+0.075}_{-0.075}$	$9.955^{+0.051}_{-0.061}$	$10.02^{+0.039}_{-0.048}$	$9.962^{+0.071}_{-0.068}$	$9.928^{+0.088}_{-0.067}$	$9.918^{+0.089}_{-0.098}$	$9.996^{+0.027}_{-0.025}$
$\log_{10}(M_{\star})$	8.5	$9.851^{+0.089}_{-0.085}$	$9.932^{+0.062}_{-0.068}$	$10.029^{+0.029}_{-0.042}$	$9.941^{+0.08}_{-0.064}$	$9.99^{+0.088}_{-0.086}$	$9.957^{+0.107}_{-0.142}$	$9.987^{+0.032}_{-0.034}$
$\alpha$	7.5	$-1.389^{+0.041}_{-0.041}$	$-1.358^{+0.053}_{-0.042}$	$-1.425^{+0.056}_{-0.041}$	$-1.347^{+0.108}_{-0.098}$	$-1.233^{+0.166}_{-0.159}$	$-1.293^{+0.245}_{-0.181}$	$-1.393^{+0.028}_{-0.028}$
$\alpha$	8.5	$-1.353^{+0.087}_{-0.079}$	$-1.304^{+0.09}_{-0.076}$	$-1.489^{+0.059}_{-0.041}$	$-1.319^{+0.127}_{-0.14}$	$-1.347^{+0.177}_{-0.164}$	$-1.479^{+0.423}_{-0.248}$	$-1.386^{+0.051}_{-0.047}$
$\log_{10}(M_{\text{min}})$	7.5	$7.56^{+0.068}_{-0.066}$	$7.544^{+0.069}_{-0.082}$	$7.636^{+0.078}_{-0.096}$	$7.487^{+0.18}_{-0.255}$	$7.169^{+0.383}_{-0.943}$	$7.337^{+0.406}_{-1.189}$	$7.58^{+0.042}_{-0.049}$
$\log_{10}(M_{\text{min}})$	8.5	$8.446^{+0.046}_{-0.056}$	$8.471^{+0.035}_{-0.039}$	$8.567^{+0.032}_{-0.034}$	$8.469^{+0.076}_{-0.094}$	$8.454^{+0.108}_{-0.146}$	$8.643^{+0.121}_{-0.276}$	$8.502^{+0.023}_{-0.025}$
$\log_{10}(M_{\text{max}})$	7.5	$11.46^{+1.066}_{-1.129}$	$11.666^{+0.93}_{-0.932}$	$11.773^{+0.829}_{-0.8}$	$11.709^{+0.831}_{-0.861}$	$11.712^{+0.839}_{-0.84}$	$11.783^{+0.757}_{-0.852}$	$11.794^{+0.829}_{-0.819}$
$\log_{10}(M_{\text{max}})$	8.5	$11.455^{+1.056}_{-1.072}$	$11.702^{+0.904}_{-0.967}$	$11.822^{+0.799}_{-0.83}$	$11.654^{+0.886}_{-0.817}$	$11.693^{+0.883}_{-0.845}$	$11.842^{+0.799}_{-0.88}$	$11.796^{+0.845}_{-0.854}$
$\beta$	7.5	$1.142^{+1.816}_{-1.493}$	$0.591^{+1.258}_{-1.019}$	$-0.173^{+0.733}_{-0.579}$	$0.75^{+0.859}_{-0.892}$	$1.346^{+0.73}_{-0.951}$	$1.098^{+0.999}_{-0.938}$	$0.968^{+0.042}_{-0.039}$
$\beta$	8.5	$1.289^{+1.734}_{-1.536}$	$1.255^{+1.529}_{-1.455}$	$-0.446^{+0.71}_{-0.381}$	$1.091^{+0.794}_{-0.998}$	$0.704^{+0.86}_{-0.962}$	$0.793^{+1.429}_{-1.208}$	$0.928^{+0.074}_{-0.067}$
$\Omega_{\text{H I}} \times 10^4$ (individual)	7.5	$3.7^{+1.08}_{-0.79}$	$5.01^{+0.55}_{-0.39}$	$5.64^{+0.44}_{-0.34}$	$5.85^{+0.52}_{-0.41}$	$6.25^{+0.66}_{-0.53}$	$6.57^{+1.11}_{-1.04}$	
$\Omega_{\text{H I}} \times 10^4$ (individual)	8.5	$3.87^{+1.26}_{-0.95}$	$5.0^{+0.45}_{-0.48}$	$6.1^{+0.52}_{-0.5}$	$6.0^{+0.86}_{-0.57}$	$6.87^{+1.07}_{-1.12}$	$8.68^{+4.72}_{-2.66}$	
$\Omega_{\text{H I}} \times 10^4$ (broad)	7.5	$4.74^{+0.2}_{-0.18}$	$5.11^{+0.21}_{-0.19}$	$5.52^{+0.22}_{-0.19}$	$5.94^{+0.23}_{-0.19}$	$6.36^{+0.24}_{-0.2}$	$6.67^{+0.25}_{-0.22}$	
$\Omega_{\text{H I}} \times 10^4$ (broad)	8.5	$4.9^{+0.3}_{-0.31}$	$5.26^{+0.3}_{-0.28}$	$5.66^{+0.29}_{-0.26}$	$6.07^{+0.3}_{-0.27}$	$6.49^{+0.31}_{-0.3}$	$6.79^{+0.35}_{-0.32}$	

following analysis for exploring the effects of increasing noise and survey area, except for the effect of a higher minimum mass cut.

For analysing the redshift dependence of our approach, we simply split the samples into six individual redshift bins ( $z$ -bins) as shown on the top of Fig. 1. Although there should be an optimal way of setting the bins (for instance, the same way as we do for binning the flux below) our current results will only be improved by a better redshift binning scheme.

To capture the shape of the flux distribution in an individual redshift bin accurately, it is sensible to set the flux bins to be uniform in log-flux space since the low flux sources dominate. However, considering that nearly half of low flux sources have negative measured fluxes (due to the noise), we instead apply the Bayesian Blocks method (Scargle et al. 2013) to determine suitable flux bins. The Bayesian Blocks method requires the minimization of a cost function across the data set and allows varying bin widths while being robust to statistical fluctuations, which is ideal for dynamic distributions of the fluxes. The user can determine how many bins are generated through the correct detection rate  $p_0$  defined by equation (11) of Scargle et al. (2013). A small  $p_0$  will be more robust to statistical fluctuations in the data, but could be overly coarse. The optimal value of  $p_0$  is typically determined empirically. At  $z > 0.1$  we accept the default value of  $p_0 = 0.05$ , but at  $z < 0.1$  we use  $p_0 = 0.5$  to avoid coarseness of the binning. In general, our results are insensitive to a large range of reasonable values for  $p_0$ .

### 3.1 The reconstructed HIMF from our baseline survey

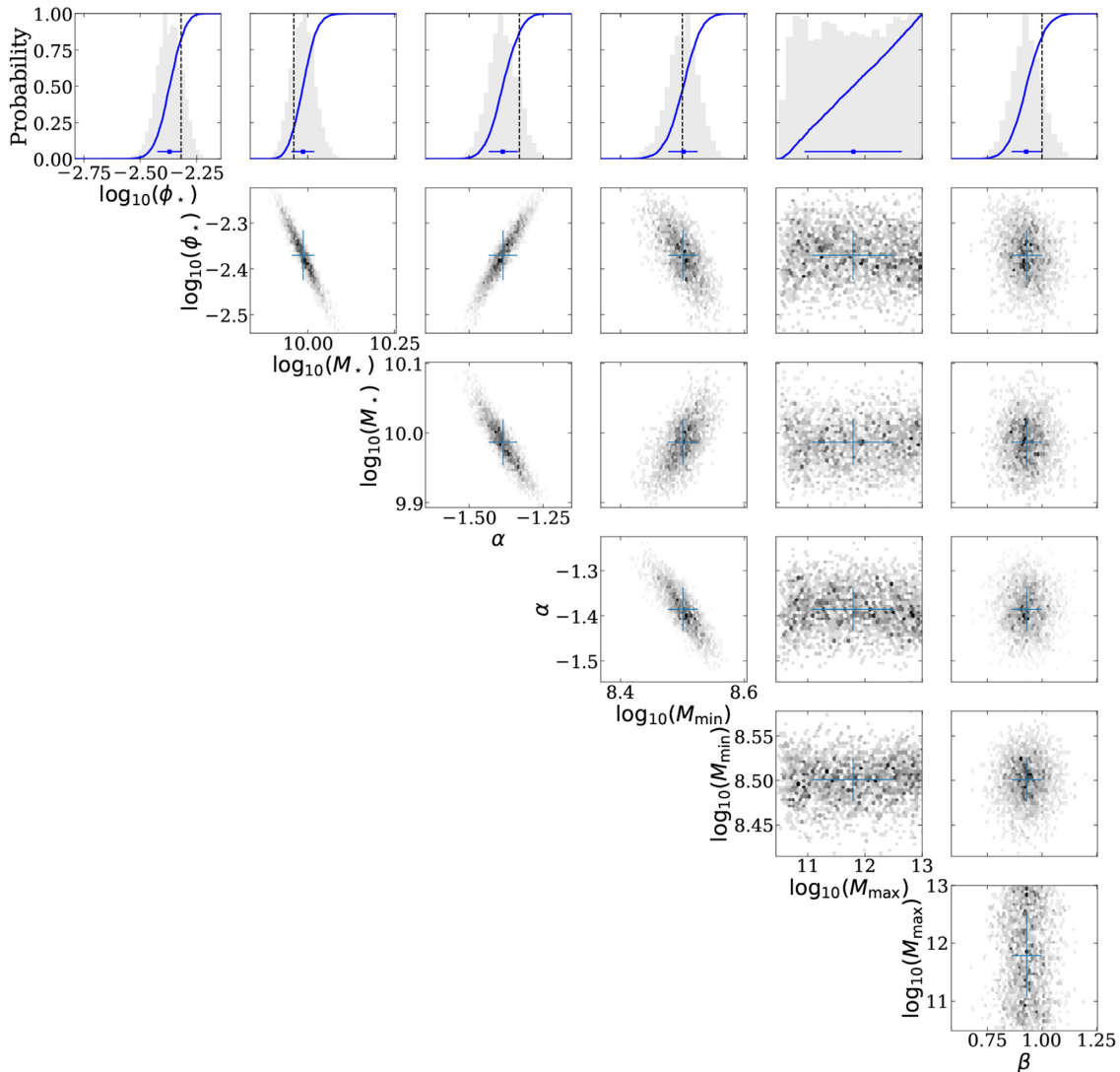
In the HIMF parametrization, the combined  $\alpha$  and  $\phi_{\star}$  parameters determine the slope and normalization of the low-mass end of the HIMF, while  $M_{\text{min}}$  provides the H I mass at which the HIMF can be

measured down to. The combination of  $M_{\star}$  and  $\phi_{\star}$  determines where the knee of the mass function occurs. The  $M_{\text{max}}$  parameter indicates the upper mass limit that can be constrained by the data. In all cases considered here  $M_{\text{max}}$  is not well constrained, but this has very little effect on the fits due to the exponential cut-off at high masses in the Schechter function always giving very few galaxies beyond  $M_{\star}$ . The parameter defining redshift dependence,  $\beta$ , is sensitive to the number of sources in individual redshift bins.

Fig. 4 shows the source counts for the measured H I flux  $S_m$  alongside the reconstructed HIMFs model for our baseline survey. The green line shows the results recovered from individual redshift bins, where we fix  $\log_{10}(\phi_{\star}) = -2.318$  and then constrain the other five parameters. This breaks the degeneracy between  $\phi_{\star}$  and  $\beta$  in a given individual redshift bin. Since we are assuming a pure density evolution of the HIMF  $\phi_{\star}$  and  $\beta$  are completely degenerate in the sense one can compensate for a high  $\phi_{\star}$  with a low  $\beta$  and vice versa. The blue line in Fig. 4 shows the results recovered using the evolving HIMF model in a single broad redshift bin spanning  $0 < z < 0.55$ . As shown, the model distributions of  $S_m$  and the resulting HIMF, using both methods, are in good agreement with the input simulated data across all redshift bins.

Taking a closer look at the recovered HIMF from the individual redshift bins, in Fig. 5 we show the posterior probability distributions for the parameters of the HIMF model for three bins in redshift and our fits to the parameters are given in Table 3. We find that the constraints on the low- and high-mass ends of the HIMF evolve differently. The signal-to-noise ratio is significantly reduced at higher redshifts due to increasing luminosity distance and this leads to a significantly weaker constraint on the low-mass end with increasing redshift. However, for the high-mass end, the sources are sufficiently bright that the noise does not significantly affect the measured line fluxes out to intermediate ( $z \sim 0.3$ ) redshift. In

$0.0 < z < 0.55$  with  $1\sigma_{\text{ch}}$  noise,  $1 \text{ deg}^2$  area,  $\log_{10}(M_{\text{min}}) = 8.5$



**Figure 9.** The posterior distributions for the HIMF model parameters for a single broad redshift bin from our simulated survey with increasing minimum mass  $\log_{10}(M_{\text{min}}) = 8.5$  of the H I galaxy sample. All other properties of this figure are the same as that of Fig. 5.

this regime, as the number of bright sources increases with the increasing volume they work to strengthen the constraints on the position of the knee of the mass function. However, eventually this effect is offset by the reduction in signal-to-noise ratio due to the increasing luminosity distance at  $z > 0.3$ . We therefore find the tightest constraints on  $M_*$  in the middle of the redshift range considered, with other redshift bins biased to low values of  $M_*$ , due to the lack of volume to sample the most massive galaxies at low redshifts and the strong effect of the noise on the faint end slope at high redshifts for this particular simulation. This trend is also apparent for the average H I mass density  $\Omega_{\text{H I}}$  (individual), as shown in the bottom rows of Table 3, where the accuracy of  $\Omega_{\text{H I}}$  is very much tied to the accuracy of  $M_*$  as this is where the bulk of the mass density resides. The evolution parameter is not well constrained from individual redshift bins in general, due to the limited number of sources in each bin.

One method of overcoming the deficiencies highlighted by considering smaller individual redshift slices is to model the evolution across the full range in redshift with a single model (blue line/region in Fig. 4). Fig. 5 (bottom right panel) shows the posterior probability distributions for the parameters in the broad redshift bin and the last column of Table 3 gives a summary of the parameter fits. We see in this case that the fits are much closer to the input values. This is because the faint end slope is highly constrained in the lowest redshift bins and the non-evolving value of  $M_*$  is well constrained at the intermediate redshifts. The high-mass end then just follows the fixed exponential fall-off. Given that these parameters are highly constrained, then the evolution parameter is also well constrained and within the uncertainties of the input value. Obviously, this is an idealized case, but does show where having a large observed volume can help constrain the mass function at low redshift, under the assumption that the fundamental shape does not

**Table 5.** Reconstructed parameters of the HIMF model from our baseline survey with various survey areas, where we fix  $\log_{10}(\phi_*) = -2.318$  for individual redshift bins.

Parameter	Area (deg <sup>2</sup> )	$0 < z < 0.1$	$0.1 < z < 0.2$	$0.2 < z < 0.3$	$0.3 < z < 0.4$	$0.4 < z < 0.5$	$0.5 < z < 0.55$	$0 < z < 0.55$
$\log_{10}(\phi_*)$	0.3	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	-2.274 <sup>+0.068</sup> <sub>-0.073</sub>
$\log_{10}(\phi_*)$	1.0	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	-2.397 <sup>+0.039</sup> <sub>-0.042</sub>
$\log_{10}(\phi_*)$	5.0	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	-2.343 <sup>+0.016</sup> <sub>-0.017</sub>
$\log_{10}(\phi_*)$	20.0	-2.318	-2.318	-2.318	-2.318	-2.318	-2.318	-2.335 <sup>+0.008</sup> <sub>-0.008</sub>
$\log_{10}(M_*)$	0.3	10.08 <sup>+0.562</sup> <sub>-0.162</sub>	9.994 <sup>+0.083</sup> <sub>-0.08</sub>	9.908 <sup>+0.1</sup> <sub>-0.093</sub>	9.787 <sup>+0.068</sup> <sub>-0.048</sub>	9.947 <sup>+0.11</sup> <sub>-0.082</sub>	9.917 <sup>+0.135</sup> <sub>-0.123</sub>	9.944 <sup>+0.046</sup> <sub>-0.045</sub>
$\log_{10}(M_*)$	1.0	9.819 <sup>+0.075</sup> <sub>-0.075</sub>	9.955 <sup>+0.051</sup> <sub>-0.061</sub>	10.02 <sup>+0.039</sup> <sub>-0.048</sub>	9.962 <sup>+0.071</sup> <sub>-0.068</sub>	9.928 <sup>+0.088</sup> <sub>-0.067</sub>	9.918 <sup>+0.089</sup> <sub>-0.098</sub>	9.996 <sup>+0.027</sup> <sub>-0.021</sub>
$\log_{10}(M_*)$	5.0	9.973 <sup>+0.037</sup> <sub>-0.038</sub>	9.97 <sup>+0.033</sup> <sub>-0.029</sub>	9.986 <sup>+0.024</sup> <sub>-0.025</sub>	9.962 <sup>+0.025</sup> <sub>-0.026</sub>	9.967 <sup>+0.03</sup> <sub>-0.031</sub>	9.98 <sup>+0.056</sup> <sub>-0.055</sub>	9.966 <sup>+0.01</sup> <sub>-0.01</sub>
$\log_{10}(M_*)$	20.0	9.972 <sup>+0.021</sup> <sub>-0.025</sub>	9.96 <sup>+0.015</sup> <sub>-0.014</sub>	9.967 <sup>+0.011</sup> <sub>-0.011</sub>	9.978 <sup>+0.012</sup> <sub>-0.012</sub>	9.972 <sup>+0.014</sup> <sub>-0.014</sub>	9.963 <sup>+0.026</sup> <sub>-0.024</sub>	9.967 <sup>+0.005</sup> <sub>-0.005</sub>
$\alpha$	0.3	-1.278 <sup>+0.085</sup> <sub>-0.064</sub>	-1.356 <sup>+0.072</sup> <sub>-0.061</sub>	-1.272 <sup>+0.136</sup> <sub>-0.119</sub>	-1.012 <sup>+0.06</sup> <sub>-0.116</sub>	-1.181 <sup>+0.141</sup> <sub>-0.102</sub>	-1.148 <sup>+0.117</sup> <sub>-0.139</sub>	-1.321 <sup>+0.054</sup> <sub>-0.053</sub>
$\alpha$	1.0	-1.389 <sup>+0.041</sup> <sub>-0.041</sub>	-1.358 <sup>+0.053</sup> <sub>-0.042</sub>	-1.425 <sup>+0.056</sup> <sub>-0.041</sub>	-1.347 <sup>+0.108</sup> <sub>-0.098</sub>	-1.233 <sup>+0.166</sup> <sub>-0.159</sub>	-1.293 <sup>+0.245</sup> <sub>-0.181</sub>	-1.393 <sup>+0.028</sup> <sub>-0.028</sub>
$\alpha$	5.0	-1.342 <sup>+0.026</sup> <sub>-0.021</sub>	-1.345 <sup>+0.025</sup> <sub>-0.025</sub>	-1.398 <sup>+0.031</sup> <sub>-0.029</sub>	-1.376 <sup>+0.039</sup> <sub>-0.039</sub>	-1.332 <sup>+0.059</sup> <sub>-0.059</sub>	-1.375 <sup>+0.119</sup> <sub>-0.113</sub>	-1.356 <sup>+0.012</sup> <sub>-0.012</sub>
$\alpha$	20.0	-1.345 <sup>+0.017</sup> <sub>-0.012</sub>	-1.34 <sup>+0.013</sup> <sub>-0.013</sub>	-1.342 <sup>+0.014</sup> <sub>-0.014</sub>	-1.375 <sup>+0.019</sup> <sub>-0.019</sub>	-1.342 <sup>+0.028</sup> <sub>-0.025</sub>	-1.307 <sup>+0.06</sup> <sub>-0.053</sub>	-1.343 <sup>+0.006</sup> <sub>-0.006</sub>
$\log_{10}(M_{\min})$	0.3	7.445 <sup>+0.13</sup> <sub>-0.169</sub>	7.631 <sup>+0.102</sup> <sub>-0.132</sub>	7.328 <sup>+0.241</sup> <sub>-0.453</sub>	6.12 <sup>+0.802</sup> <sub>-0.738</sub>	6.895 <sup>+0.626</sup> <sub>-1.138</sub>	6.341 <sup>+0.732</sup> <sub>-0.794</sub>	7.504 <sup>+0.092</sup> <sub>-0.109</sub>
$\log_{10}(M_{\min})$	1.0	7.56 <sup>+0.068</sup> <sub>-0.066</sub>	7.544 <sup>+0.069</sup> <sub>-0.092</sub>	7.636 <sup>+0.078</sup> <sub>-0.096</sub>	7.487 <sup>+0.18</sup> <sub>-0.255</sub>	7.160 <sup>+0.383</sup> <sub>-0.943</sub>	7.337 <sup>+0.406</sup> <sub>-1.189</sub>	7.58 <sup>+0.042</sup> <sub>-0.049</sub>
$\log_{10}(M_{\min})$	5.0	7.522 <sup>+0.031</sup> <sub>-0.03</sub>	7.513 <sup>+0.036</sup> <sub>-0.038</sub>	7.639 <sup>+0.045</sup> <sub>-0.051</sub>	7.595 <sup>+0.069</sup> <sub>-0.077</sub>	7.487 <sup>+0.124</sup> <sub>-0.149</sub>	7.552 <sup>+0.2</sup> <sub>-0.296</sub>	7.546 <sup>+0.021</sup> <sub>-0.021</sub>
$\log_{10}(M_{\min})$	20.0	7.502 <sup>+0.017</sup> <sub>-0.016</sub>	7.529 <sup>+0.018</sup> <sub>-0.02</sub>	7.531 <sup>+0.026</sup> <sub>-0.027</sub>	7.584 <sup>+0.037</sup> <sub>-0.039</sub>	7.511 <sup>+0.055</sup> <sub>-0.067</sub>	7.377 <sup>+0.132</sup> <sub>-0.177</sub>	7.524 <sup>+0.01</sup> <sub>-0.011</sub>
$\log_{10}(M_{\max})$	0.3	10.75 <sup>+1.49</sup> <sub>-0.97</sub>	11.658 <sup>+0.898</sup> <sub>-1.058</sub>	11.676 <sup>+0.873</sup> <sub>-0.919</sub>	11.654 <sup>+0.879</sup> <sub>-0.878</sub>	11.814 <sup>+0.775</sup> <sub>-0.865</sub>	11.525 <sup>+0.978</sup> <sub>-0.786</sub>	11.793 <sup>+0.824</sup> <sub>-0.888</sub>
$\log_{10}(M_{\max})$	1.0	11.46 <sup>+1.066</sup> <sub>-1.129</sub>	11.666 <sup>+0.93</sup> <sub>-0.932</sub>	11.773 <sup>+0.829</sup> <sub>-0.8</sub>	11.709 <sup>+0.831</sup> <sub>-0.861</sub>	11.712 <sup>+0.839</sup> <sub>-0.84</sub>	11.783 <sup>+0.757</sup> <sub>-0.852</sub>	11.794 <sup>+0.829</sup> <sub>-0.819</sub>
$\log_{10}(M_{\max})$	5.0	11.642 <sup>+0.912</sup> <sub>-0.936</sub>	11.754 <sup>+0.811</sup> <sub>-0.902</sub>	11.796 <sup>+0.845</sup> <sub>-0.802</sub>	11.882 <sup>+0.757</sup> <sub>-0.902</sub>	11.821 <sup>+0.796</sup> <sub>-0.824</sub>	11.895 <sup>+0.726</sup> <sub>-0.769</sub>	11.855 <sup>+0.77</sup> <sub>-0.777</sub>
$\log_{10}(M_{\max})$	20.0	11.633 <sup>+0.871</sup> <sub>-0.939</sub>	11.739 <sup>+0.864</sup> <sub>-0.859</sub>	11.835 <sup>+0.765</sup> <sub>-0.739</sub>	11.808 <sup>+0.792</sup> <sub>-0.721</sub>	11.924 <sup>+0.749</sup> <sub>-0.777</sub>	11.932 <sup>+0.732</sup> <sub>-0.776</sub>	11.9 <sup>+0.701</sup> <sub>-0.732</sub>
$\beta$	0.3	1.746 <sup>+1.552</sup> <sub>-1.778</sub>	1.271 <sup>+1.663</sup> <sub>-1.516</sub>	1.636 <sup>+1.361</sup> <sub>-1.467</sub>	3.077 <sup>+0.521</sup> <sub>-0.749</sub>	1.358 <sup>+0.791</sup> <sub>-1.195</sub>	1.029 <sup>+1.054</sup> <sub>-1.12</sub>	0.87 <sup>+0.071</sup> <sub>-0.075</sub>
$\beta$	1.0	1.142 <sup>+1.816</sup> <sub>-1.493</sub>	0.591 <sup>+1.258</sup> <sub>-1.019</sub>	-0.173 <sup>+0.733</sup> <sub>-0.579</sub>	0.75 <sup>+0.859</sup> <sub>-0.892</sub>	1.346 <sup>+0.73</sup> <sub>-0.951</sub>	1.098 <sup>+0.999</sup> <sub>-0.938</sub>	0.968 <sup>+0.042</sup> <sub>-0.039</sub>
$\beta$	5.0	0.638 <sup>+1.545</sup> <sub>-1.154</sub>	0.651 <sup>+0.605</sup> <sub>-0.676</sub>	0.389 <sup>+0.386</sup> <sub>-0.383</sub>	0.747 <sup>+0.329</sup> <sub>-0.324</sub>	0.925 <sup>+0.347</sup> <sub>-0.359</sub>	0.684 <sup>+0.582</sup> <sub>-0.615</sub>	0.997 <sup>+0.019</sup> <sub>-0.019</sub>
$\beta$	20.0	0.026 <sup>+1.046</sup> <sub>-0.713</sub>	0.908 <sup>+0.293</sup> <sub>-0.32</sub>	0.899 <sup>+0.17</sup> <sub>-0.172</sub>	0.663 <sup>+0.156</sup> <sub>-0.164</sub>	0.877 <sup>+0.164</sup> <sub>-0.15</sub>	0.94 <sup>+0.264</sup> <sub>-0.272</sub>	1.01 <sup>+0.009</sup> <sub>-0.009</sub>
$\Omega_{\text{H}1} \times 10^4$ (individual)	0.3	4.21 <sup>+4.14</sup> <sub>-1.1</sub>	5.76 <sup>+1.31</sup> <sub>-0.76</sub>	5.38 <sup>+0.75</sup> <sub>-0.5</sub>	6.01 <sup>+0.66</sup> <sub>-0.66</sub>	6.54 <sup>+0.74</sup> <sub>-1.04</sub>	5.47 <sup>+1.45</sup> <sub>-1.3</sub>	
$\Omega_{\text{H}1} \times 10^4$ (individual)	1.0	3.68 <sup>+1.08</sup> <sub>-0.81</sub>	5.0 <sup>+0.54</sup> <sub>-0.4</sub>	5.58 <sup>+0.41</sup> <sub>-0.31</sub>	5.85 <sup>+0.43</sup> <sub>-0.37</sub>	6.29 <sup>+0.6</sup> <sub>-0.52</sub>	6.39 <sup>+0.98</sup> <sub>-0.81</sub>	
$\Omega_{\text{H}1} \times 10^4$ (individual)	5.0	4.96 <sup>+0.65</sup> <sub>-0.57</sub>	5.2 <sup>+0.21</sup> <sub>-0.21</sub>	5.66 <sup>+0.14</sup> <sub>-0.15</sub>	5.98 <sup>+0.18</sup> <sub>-0.16</sub>	6.52 <sup>+0.22</sup> <sub>-0.26</sub>	6.71 <sup>+0.35</sup> <sub>-0.46</sub>	
$\Omega_{\text{H}1} \times 10^4$ (individual)	20.0	4.72 <sup>+0.35</sup> <sub>-0.31</sub>	5.22 <sup>+0.1</sup> <sub>-0.1</sub>	5.71 <sup>+0.08</sup> <sub>-0.08</sub>	6.05 <sup>+0.09</sup> <sub>-0.09</sub>	6.53 <sup>+0.12</sup> <sub>-0.12</sub>	6.62 <sup>+0.19</sup> <sub>-0.21</sub>	
$\Omega_{\text{H}1} \times 10^4$ (broad)	0.3	5.15 <sup>+0.3</sup> <sub>-0.44</sub>	5.5 <sup>+0.3</sup> <sub>-0.43</sub>	5.88 <sup>+0.31</sup> <sub>-0.42</sub>	6.27 <sup>+0.32</sup> <sub>-0.42</sub>	6.65 <sup>+0.37</sup> <sub>-0.44</sub>	6.93 <sup>+0.4</sup> <sub>-0.47</sub>	
$\Omega_{\text{H}1} \times 10^4$ (broad)	1.0	4.72 <sup>+0.2</sup> <sub>-0.17</sub>	5.09 <sup>+0.19</sup> <sub>-0.18</sub>	5.5 <sup>+0.2</sup> <sub>-0.18</sub>	5.91 <sup>+0.2</sup> <sub>-0.18</sub>	6.34 <sup>+0.21</sup> <sub>-0.2</sub>	6.64 <sup>+0.22</sup> <sub>-0.22</sub>	
$\Omega_{\text{H}1} \times 10^4$ (broad)	5.0	4.78 <sup>+0.11</sup> <sub>-0.07</sub>	5.16 <sup>+0.11</sup> <sub>-0.07</sub>	5.58 <sup>+0.11</sup> <sub>-0.06</sub>	6.01 <sup>+0.11</sup> <sub>-0.06</sub>	6.45 <sup>+0.12</sup> <sub>-0.06</sub>	6.77 <sup>+0.13</sup> <sub>-0.06</sub>	
$\Omega_{\text{H}1} \times 10^4$ (broad)	20.0	4.83 <sup>+0.04</sup> <sub>-0.04</sub>	5.21 <sup>+0.04</sup> <sub>-0.04</sub>	5.65 <sup>+0.04</sup> <sub>-0.05</sub>	6.1 <sup>+0.04</sup> <sub>-0.05</sub>	6.54 <sup>+0.05</sup> <sub>-0.05</sub>	6.88 <sup>+0.05</sup> <sub>-0.05</sub>	

change. Alternatively, we could incorporate the evolutionary term into the  $M_*$  parameter and adopt a pure-mass evolution model. In this case we could then use the Bayesian evidence to distinguish between pure mass and pure density evolution, and we leave this to future work using real data.

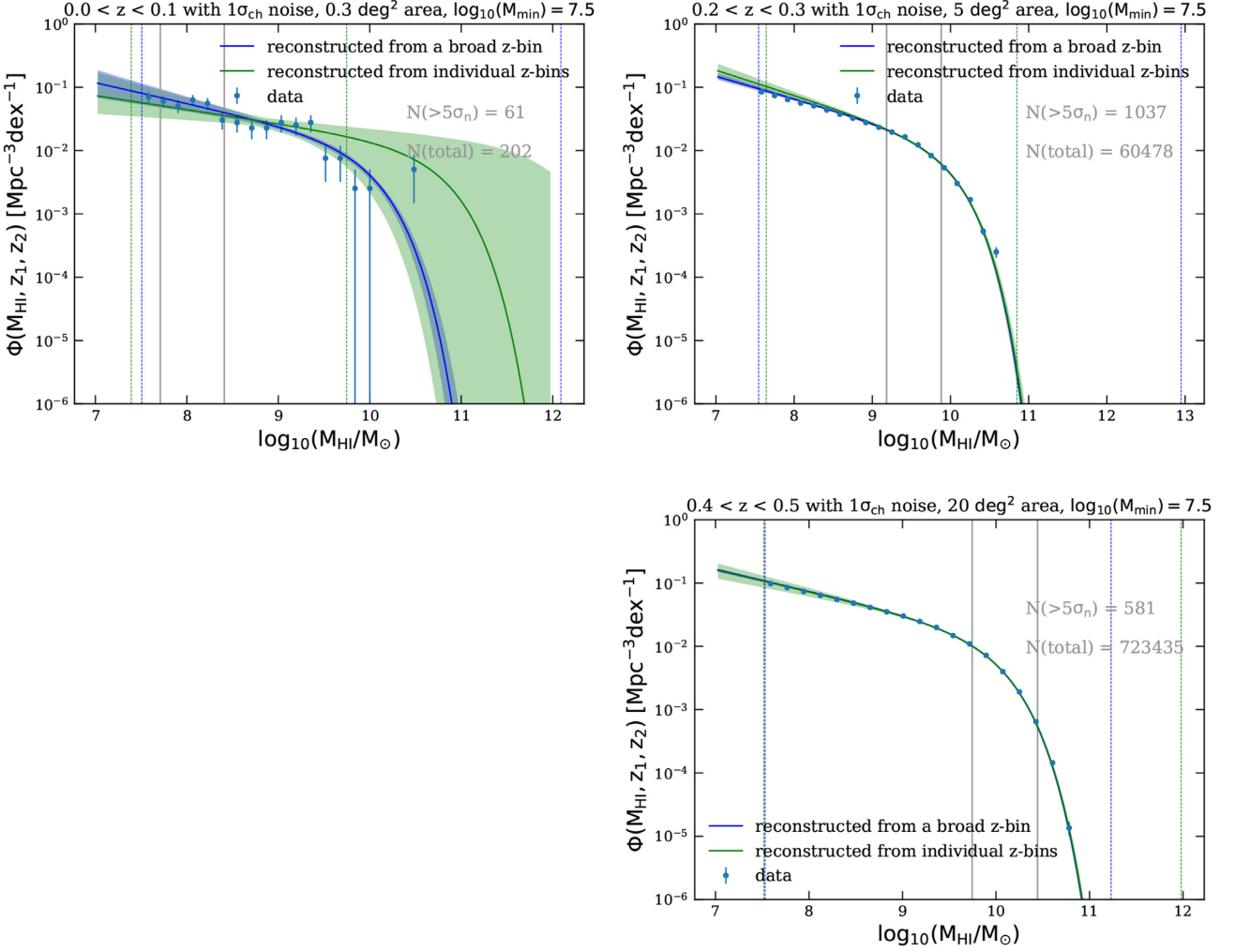
The overall benefits of using a broad redshift bin can also be appreciated by considering the constraints on  $\Omega_{\text{H}1}$ (broad), where there is much closer agreement between the input and derived values.

### 3.2 The effect of changing the background noise

We now look at the effect of changing the noise on the reconstructed HIMF from our baseline survey. We show the resulting parameters for 0.5 and  $5\sigma_{\text{ch}}$  noise in Table 3 for the individual redshift bins

and the whole redshift range, respectively. Fig. 6 shows the HIMF, and its evolution, reconstructed from the simulated data for these different levels of noise. We only show the relevant figures for the model fit over the full redshift range in Fig. 7, rather than for the individual bins as the general trends remain the same as discussed in Section 3.1. A subtle difference is that the tightest constraints on  $M_*$  and  $\Omega_{\text{H}1}$  for the  $0.5\sigma_{\text{ch}}$  noise has drifted away from the middle of the redshift range towards higher range considered due to the increased signal-to-noise ratio for all sources.

As expected, increasing the noise by a factor of five leads to significant increase in the uncertainties for the parameters which define the Schechter function and determine  $\Omega_{\text{H}1}$ , particularly for the single redshift bin. The absolute values for the Schechter function parameters (Table 3) show that increasing the noise leads to  $\phi_*$  increasing and then being compensated for with a flatter faint end



**Figure 10.** Clockwise from top left to bottom right panels, the reconstructed HIMF from a single broad redshift bin (blue) and a smaller redshift bin (green) from survey areas equal to 0.3, 5, and 20 deg<sup>2</sup> in different redshift bins. The grey vertical lines indicate the  $\sigma_n = \sqrt{N_{\text{ch}}}\sigma_{\text{ch}}dv$  and  $5\sigma_n$  detection threshold at the centre of each redshift bin. All other properties of this figure are the same as that of Fig. 6.

slope and a lower  $M_*$ . Again this can be understood by considering the redshift range where the tightest constraints are coming from for this noisier data set. The constraints on the low-mass slope are most affected as our ability to correctly assign the correct flux to the low-mass sources is significantly inhibited by the increased noise in all redshift bins. This in turn leads to a more significant degeneracy between  $M_*$  and  $\phi_*$ , where the 2D posterior distribution changes from a linear degeneracy towards a more ‘banana-shape’ (Fig. 7).

On the other hand, the values for the  $\beta$  are not affected significantly by increasing the noise, due to the fact that once the Schechter function parameters are set in place then the evolution term mainly depends on the number of sources in the most constrained mass bins at all redshifts, which are generally slightly lower than  $M_*$  and therefore are not significantly degenerate with  $M_*$  or  $\alpha$ .

### 3.3 The effect of increasing the minimum mass of the H I galaxy sample

Any parent spectroscopic redshift catalogue for galaxies used to conduct an H I stacking experiment will inevitably be flux limited. This naturally leads to a limit on the H I mass that any model function can be fit down to. Furthermore, this may also change over

the redshift range of interest. In this section, we therefore investigate this by simply increasing the minimum mass from  $M_{\text{min}} = 10^{7.5}$  to  $10^{8.5} M_{\odot}$  on the reconstructed HIMF.

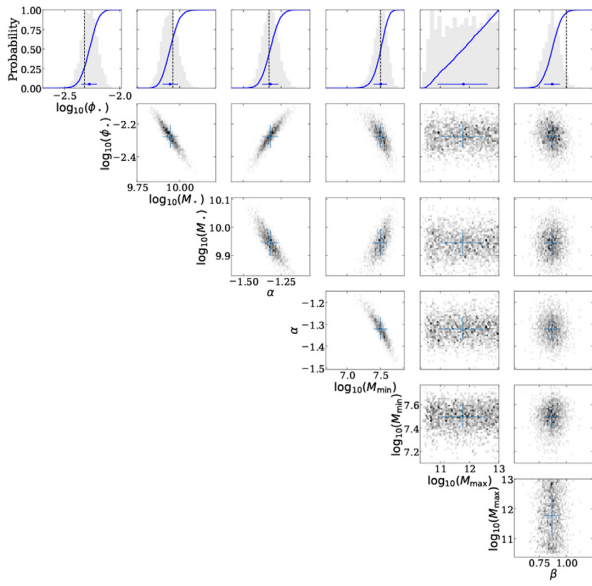
In Fig. 8, we show the resulting HIMF model and its evolution with redshift, with the input catalogue limited to  $M_{\text{HI}} > 10^{8.5} M_{\odot}$ . Posterior values for the parameters are given in Table 4 and Fig. 9. As before we do not present the posterior distributions for the individual redshift bins.

Obviously, imposing a higher lower-mass limit to the H I sample will automatically lead to a smaller sample overall and naturally lead to poorer constraints on the low-mass slope of the HIMF and we see this in our results. However, the constraints on  $M_*$  and  $\phi_*$  are not affected as much. Reassuringly, our fit does find where the HIMF model is reliably fit to, where we find  $M_{\text{min}}$  increases from  $10^{7.5}$  to  $10^{8.5} M_{\odot}$  when we alter the lower mass limit of the sample. The  $\Omega_{\text{HI}}$  is less affected at higher redshifts since  $M_*$  and  $\phi_*$  are still well constrained.

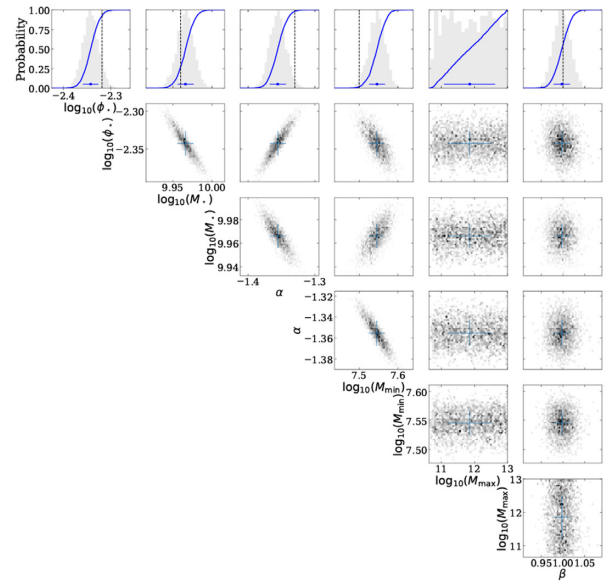
### 3.4 The effect of changing the survey area

Finally, we look at the effect of changing the survey area on the reconstructed model HIMF for our baseline survey. We show the

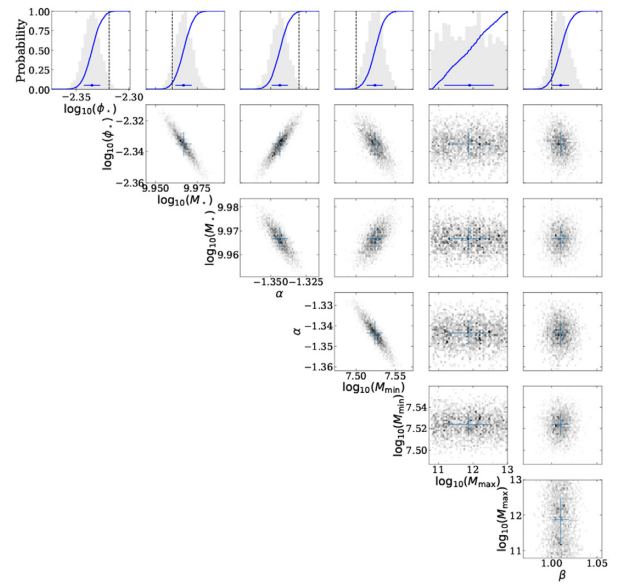
$0.0 < z < 0.55$  with  $1\sigma_{\text{ch}}$  noise,  $0.3 \text{ deg}^2$  area,  $\log_{10}(M_{\text{min}}) = 7.5$



$0.0 < z < 0.55$  with  $1\sigma_{\text{ch}}$  noise,  $5 \text{ deg}^2$  area,  $\log_{10}(M_{\text{min}}) = 7.5$



$0.0 < z < 0.55$  with  $1\sigma_{\text{ch}}$  noise,  $20 \text{ deg}^2$  area,  $\log_{10}(M_{\text{min}}) = 7.5$



**Figure 11.** The posterior distributions for the HIMF model parameters for a single broad redshift bin for survey areas equal to 0.3, 5, and  $20 \text{ deg}^2$  clockwise from top left to bottom right panels. All other properties of this figure are the same as that of Fig. 5.

posterior values for the reconstructed HIMF model parameters over a range of areas from 0.3 to  $20 \text{ deg}^2$  in Table 5. Fig. 10 shows the reconstructed HIMF over these different survey areas and its evolution with redshift.

Fig. 11 shows the posterior distributions for the parameters of the HIMF model measured in a broad  $z$ -bin over  $0.3\text{--}20 \text{ deg}^2$  areas. It indicates that all the parameters  $\phi_*$ ,  $M_*$ ,  $\alpha$ ,  $M_{\text{min}}$ ,  $M_{\text{max}}$ ,  $\beta$ , and the HI mass density  $\Omega_{\text{HI}}$  become better constrained as the survey area increases as expected.

The key advantage in moving to larger area is the increase in volume at lower redshift, which enables one to fully constrain

the mass function in smaller redshift bins, as there are enough galaxies at all masses. This is shown most clearly in Table 5 where the parameters that define the Schechter function are most closely aligned with their input values as area increases, even at  $z < 0.1$ .

## 4 CONCLUSIONS

We present a source-count model for the HI flux distribution in galaxies and a Bayesian stacking technique to constrain the HIMF, and its evolution over the redshift range  $0 < z < 0.55$ , below the detection threshold and down to  $M_{\text{HI}} = 10^{7.5} M_{\odot}$  for this particular

simulation. We generate galaxy samples using an assumed HIMF model and simulate the HI emission lines with different levels of background noise and different survey areas to investigate the robustness of this technique. We use the Monte Carlo sampling algorithm MULTINEST to reconstruct the HIMF model parameters, both for a single broad redshift bin and for individual redshift bins. In both cases, we demonstrate that the HIMF can be accurately reconstructed from the simulated data using our method, which allows us to push the boundary of the HIMF far below the expected HI mass detection limit of  $M_{\text{HI}} \sim 10^9 M_{\odot}$  at  $z \sim 0.1$  for the MIGHTEE and LADUMA HI surveys. More specifically, we find that:

(i) As expected, constraints on the Schechter function parameters  $\phi_*$ ,  $M_*$ , and  $\alpha$  become tighter as the background noise decreases, but the term that parametrizes the evolution remains unchanged. Thus, if one has confidence that the shape of the HIMF does not change but only evolves as mass evolution, or in our case, pure density evolution, then this evolution can be measured robustly with relatively poor data. This is because once the low-mass slope has been well fit at low redshift, and the characteristic mass is well defined at intermediate redshift, then the evolution parameter cannot vary from its true value. As such, the HI mass density as a function of redshift can be measured robustly to high redshift, as this depends most strongly on the position of the knee in the mass function ( $M_*$ ).

(ii) The constraints on  $\phi_*$ ,  $M_*$ , and  $\alpha$ , along with  $\beta$  are weaker if the minimum mass is increased from  $M_{\text{min}} = 10^{7.5}$  to  $10^{8.5} M_{\odot}$ . However, our parametrization means that we naturally find the minimum HI mass ( $M_{\text{min}}$ ) that the data allow the model to fit to.

(iii) The constraints on the parameters  $\phi_*$ ,  $M_*$ ,  $\alpha$ , and  $\beta$  become stronger as the survey area increases. However, increasing the area naturally leads to a more robust determination of the HIMF at the lowest redshifts due to the significant increase in volume that will hence contain the higher mass galaxies.

This work is a first step towards the measurement of the HIMF in the high-redshift regime. In future work, we will apply this method to real data from the MIGHTEE survey. However, there are some caveats one needs to be aware of in doing so:

(i) In practice, the noise may not necessarily follow a Gaussian distribution, and thus we would need to determine the real noise distribution to be used in place of the Gaussian distribution in equation (8). This may lead to a significant increase in the computing time for equation (8), and especially if the  $\sigma_n$  is dependent on redshift.

(ii) We have also not considered the effect of confusion on our results, since we expect future deep HI surveys to achieve 5–15 arcsec spatial resolution and a spectral resolution of at least a few  $10 \text{ km s}^{-1}$ . However confusion will need to be considered if our approach is used on lower resolution data (e.g. HIPASS and ALFALFA).

(iii) We also assume that the galaxy sample is complete for a given mass threshold, in this case  $M_{\text{min}} = 10^{7.5}$  or  $10^{8.5} M_{\odot}$ . However, in practice a complete mass-limited sample of galaxies is difficult to construct, especially at the higher redshifts. Thus, as with any other experiment that involves stacking based on a sample defined at a different wavelength, one always needs to be aware that the derived HIMF is that of the galaxies in the parent sample, and will always miss galaxies that could have relatively large HI masses but low stellar masses.

(iv) Other potential issues are instrumental effects that generate spectral features below the  $5\sigma$  noise level and might contaminate the

flux in our Bayesian stacking signal, e.g. strong continuum sources and their sidelobes that are not subtracted accurately enough. Such contamination should not correlate with the galaxy catalogue used for stacking and so the stacked flux should be unbiased on average. However, the residuals will still contribute to the noise level and ultimately we may need to test different cleaning techniques or use independent noise estimates for each galaxy.

(v) If we were interested in extracting the fluxes of sources which we expect to be extended based on their optical properties, we could enlarge the aperture over which we extract the flux. However, we note that one could also simply image the cube with a lower resolution restoring beam at the high-frequency (low-redshift) end of the data cube, ensuring that all sources are unresolved.

## ACKNOWLEDGEMENTS

We thank the anonymous referee for helpful comments that have contributed to this paper. This work was supported by the China Scholarship Council and the Science and Technology Facilities Council (STFC) of the United Kingdom. MJJ and IH are supported by the South African Radio Astronomy Observatory and STFC under the consolidated grant ST/S000488/1. JRA acknowledges support from a Christ Church Career Development Fellowship. MGS acknowledges support from the South African Square Kilometre Array Project and National Research Foundation (grant no. 84156). IH acknowledges support from the Oxford Hintze Centre for Astrophysical Surveys which is funded through generous support from the Hintze Family Charitable Foundation.

## REFERENCES

- Barnes D. G. et al., 2001, *MNRAS*, 322, 486  
 Blyth S. et al., 2016, in Proc. Sci., MeerKAT Science: On the Pathway to the SKA, 25–27 May. Stellenbosch, South Africa, p. 4  
 Buchner J. et al., 2014, *A&A*, 564, A125  
 Chen S., Zwart J. T. L., Santos M. G., 2017, preprint (arXiv:1709.04045)  
 Davies L. J. M. et al., 2018, *MNRAS*, 480, 768  
 DeBoer D. R. et al., 2009, *IEEE Proc.*, 97, 1507  
 Delhaize J., Meyer M. J., Staveley-Smith L., Boyle B. J., 2013, *MNRAS*, 433, 1398  
 Driver S. P. et al., 2011, *MNRAS*, 413, 971  
 Fernández X. et al., 2016, *ApJ*, 824, L1  
 Feroz F., Hobson M. P., Bridges M., 2009, *MNRAS*, 398, 1601  
 Giovanelli R. et al., 2005, *AJ*, 130, 2598  
 Holwerda B. W., Blyth S. L., Baker A. J., 2012, in Tuffs R. J., Popescu C. C., eds, IAU Symp. Vol. 284, The Spectral Energy Distribution of Galaxies (SED’11). Kluwer, Dordrecht, p. 496  
 Ianjamasimanana R., de Blok W. J. G., Walter F., Heald G. H., 2012, *AJ*, 144, 96  
 Jarvis M. et al., 2016, in Proc. Sci., MeerKAT Science: On the Pathway to the SKA, 25–27 May. SISSA, Trieste, p. 6  
 Johnston S. et al., 2007, *PASA*, 24, 174  
 Jonas J., MeerKAT Team, 2016, in Proc. Sci., MeerKAT Science: On the Pathway to the SKA, 25–27 May. SISSA, Trieste, p. 1  
 Jonas J. L., 2009, *IEEE Proc.*, 97, 1522  
 Jones M. G., Haynes M. P., Giovanelli R., Papastergis E., 2016, *MNRAS*, 455, 1574  
 Jones M. G., Haynes M. P., Giovanelli R., Moorman C., 2018, *MNRAS*, 477, 2  
 Maddox N., Hess K. M., Obreschko D., Jarvis M. J., Blyth S. L., 2015, *MNRAS*, 447, 1610  
 Maddox N., Jarvis M. J., Oosterloo T. A., 2016, *MNRAS*, 460, 3419  
 Malefahlo E., Santos M. G., Jarvis M. J., White S. V., Zwart J. T. L., 2019, preprint (arXiv:1908.05316)

- Martin A. M., Papastergis E., Giovanelli R., Haynes M. P., Springob C. M., Stierwalt S., 2010, *ApJ*, 723, 1359
- Meyer M., 2009, in Proc. Sci., Panoramic Radio Astronomy: Wide-Field 1-2 GHz Research on Galaxy Evolution. Groningen, the Netherlands, p. 15
- Meyer M., Robotham A., Obreschkow D., Driver S., Staveley-Smith L., Zwaan M., 2015, SISSA. Trieste, PoS, p. 131
- Meyer M., Robotham A., Obreschkow D., Westmeier T., Duffy A. R., Staveley-Smith L., 2017, PASA, 34
- Mitchell-Wynne K., Santos M. G., Afonso J., Jarvis M. J., 2014, *MNRAS*, 437, 2270
- Oosterloo T., Verheijen M. A. W., van Cappellen W., Bakker L., Heald G., Ivashina M., 2009, in Proc. Sci., Wide Field Astronomy and Technology for the Square Kilometre Array. Chateau de Limelette, Belgium, p. 70
- Papastergis E., Giovanelli R., Haynes M. P., Rodríguez-Puebla A., Jones M. G., 2013, *ApJ*, 776, 43
- Planck Collaboration et al., 2016, *A&A*, 594, A13
- Rhee J., Lah P., Briggs F. H., Chengalur J. N., Colless M., Willner S. P., Ashby M. L. N., Le Fèvre O., 2018, *MNRAS*, 473, 1879
- Scargle J. D., Norris J. P., Jackson B., Chiang J., 2013, *ApJ*, 764, 167
- Skilling J., 2004, in Fischer R., Preuss R., Toussaint U. V., eds, AIP Conf. Ser. Vol. 735, Nested Sampling. Am. Inst. Phys., New York, p. 395
- Verheijen M., Oosterloo T., Heald G., van Cappellen W., 2009, in Proc. Sci., Panoramic Radio Astronomy: Wide-field 1-2 GHz Research on Galaxy Evolution. Groningen, the Netherlands, p. 10
- Westmeier T., Jurek R., Obreschkow D., Koribalski B. S., Staveley-Smith L., 2014, *MNRAS*, 438, 1176
- Zwaan M. A., Meyer M. J., Staveley-Smith L., Webster R. L., 2005, *MNRAS*, 359, L30
- Zwart J. T. L., Santos M., Jarvis M. J., 2015, *MNRAS*, 453, 1740

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