Prepared for submission to JCAP

Mind the gap: the power of combining photometric surveys with intensity mapping

Chirag Modi,[a,b] **Martin White,**[a,c] **Emanuele Castorina,**[d,e] **Anže Slosar**[e]

[a]Berkeley Center for Cosmological Physics, Department of Physics, University of California, Berkeley, CA 94720

[b]Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Ave., New York, NY 10010, USA

[c]Department of Astronomy, University of California, Berkeley, CA 94720

[d]Dipartimento di Fisica 'Aldo Pontremoli', Università degli Studi di Milano, Milan, Italy

[e]Theoretical Physics Department, CERN, 1211 Geneva 23, Switzerland

[f]Department of Physics, Brookhaven National Laboratory, Upton, NY 11973

E-mail: cmodi@flatironinstitute.org, mwhite@berkeley.edu, emanuele.castorina@unimi.it, anze@bnl.gov

Abstract. The long wavelength modes lost to bright foregrounds in the interferometric 21-cm surveys can partially be recovered using a forward modeling approach that exploits the non-linear coupling between small and large scales induced by gravitational evolution. In this work, we build upon this approach by considering how adding external galaxy distribution data can help to fill in these modes. We consider supplementing the 21-cm data at two different redshifts with a spectroscopic sample (good radial resolution but low number density) loosely modeled on DESI-ELG at $z = 1$ and a photometric sample (high number density but poor radial resolution) similar to LSST sample at $z = 1$ and $z = 4$ respectively. We find that both the galaxy samples are able to reconstruct the largest modes better than only using 21-cm data, with the spectroscopic sample performing significantly better than the photometric sample despite much lower number density. We demonstrate the synergies between surveys by showing that the primordial initial density field is reconstructed better with the combination of surveys than using either of them individually. Methodologically, we also explore the importance of smoothing the density field when using bias models to forward model these tracers for reconstruction.

Keywords: cosmological parameters from LSS – 21 cm – galaxy clustering –bias model – forward modeling

Contents

1 Introduction

The study of large-scale structure in the high-redshift Universe is a promising tool for cosmology [1]. One means of mapping large-scale structure in the distant Universe is through the technique of intensity mapping (IM): performing a low resolution, spectroscopic survey to measure integrated flux from unresolved sources on large areas of sky at different frequencies. Such surveys capture the largest elements of the cosmic web and map out the distribution of matter in very large cosmological volumes in a fast and efficient manner, with good radial resolution [2, 3]. Since hydrogen is so abundant in the Universe, 21-cm emission from cosmic neutral hydrogen (HI) offers one tracer to map out the Universe in such a way. With its low energy and optical depth there is little chance of line confusions and it provides an efficient way to probe the spatial distribution of neutral hydrogen, and hence the underlying dark matter from the local Universe to the dark ages [3–8].

One issue with IM surveys is that foregrounds render the long-wavelength fluctuations along the line of sight unmeasureable, which can adversely affect the science that they can do [9–13]. In ref. [13] we studied one method for reconstructing long-wavelength fluctuations, using the distinct pattern of correlations imprinted by gravitational instability. In this paper we take another route, more similar to refs. [10, 12], and consider how adding additional data can help to fill in the modes that are lost to foregrounds in 21-cm observations.

2 Mock samples

To model our galaxy and IM surveys we use an extension of the Hidden Valley simulations [14], a set of 10240^3 particle N-body simulations performed in a $1024\,h^{-1}$Mpc box using the FastPM code [15]. Further details of this simulation and the manner in which HI is assigned to halos can be found in Appendix A. We consider the data in the redshift space. Anticipating that the HI would be observed by an instrument such as the Hydrogen Intensity and Real-time Analysis eXperiment (HIRAX; [16]) or the Packed Ultrawideband Mapping Array (PUMA [8, 17]) we assign the HI to a regular 512^3 grid and work with the Fourier transform of the density field. To the signal we add thermal noise and foregrounds as described in detail in refs. [13, 14]. For $z = 1$ we will present results for both PUMA and HIRAX thermal noise, while for $z = 4$ we will have observations only from PUMA and so restrict ourselves to that case.

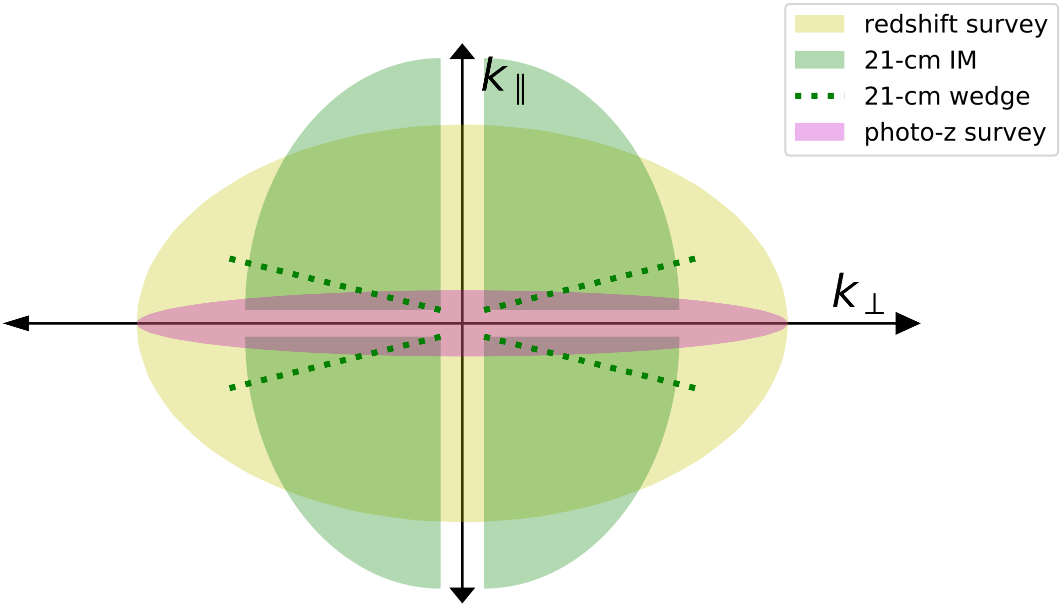

Figure 1. A schematic view of the modes probed by 21-cm surveys and optical surveys in the $k_\perp - k_\parallel$ plane. Photometric redshift surveys (purple) are capable of high angular resolution (i.e. probe to high $k_\perp$) but have limited radial resolution and thus only constrain well low $k_\parallel$ modes. A spectroscopic redshift survey (yellow) would enable access to the full $\vec{k}$ plane, but would be prohibitively expensive if done with high number density. An IM survey of 21-cm emission (green) has high number density and radial resolution but limited angular resolution and misses modes at low $k_\parallel$ (and in a region known as the foreground wedge).

Due to foregrounds and difficulties in instrument calibration, we assume the HI map has infinite noise for modes with $k_\parallel \approx 0$ and in a region of the $k_\perp - k_\parallel$ plane known as the "wedge" (see Fig. 1). For details of how these limitations arise, we refer the reader to refs. [13, 14] and the extensive references to the earlier literature. Our goal here is to ask to what extent a different survey, either a photometric or spectroscopic galaxy survey, can "fill in" these missing modes. Ideally the galaxy survey can take advantage of the high number density and excellent radial resolution of the HI survey while the HI survey can take advantage of the low-$k_\parallel$ sensitivity of the galaxy survey. We will present results for a 'pessimistic' choice of foreground wedge which removes information for angles less than three times the primary field of view (3×FOV). This corresponds to a wedge angle of $\theta_w = 6^\circ$ and 38° at $z = 1$ and $z = 4$ respectively [13]. In our notation, this means we exclude all modes with

$$k_\parallel < \sin(\theta_w) \frac{D(z)H(z)}{1+z} k_\perp \,. \tag{2.1}$$

where $D(z)$ is the line-of-sight comoving distance. We also remove all modes with $k_\parallel < 0.1\,\mathrm{Mpc}/h$. Obviously a better instrument calibration and foreground subtraction, which leads to a smaller wedge, will require less input from the auxilliary data.

We will consider two populations of mock galaxies, loosely modeled on samples that might be returned by upcoming surveys. Both the galaxy samples will again be considered in the redshift space. At $z = 1$ we consider a galaxy sample with good redshift measurements and $\bar{n} = 10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$. Such a sample is similar to the emission line galaxy (ELG) sample to be targeted by the Dark Energy Spectroscopic Instrument (DESI) survey [18], and henceforth we will refer it so. For this exploratory calculation, rather than model these galaxies in great detail, we simply choose a mass-limited sample of halos in our N-body simulation with $\bar{n} = 10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$. We assume the redshifts of these halos are known precisely. Such a halo sample has a complex, scale-dependent bias and about the right level of shot noise while being very easy to model. We do not expect our results to depend upon the details of this choice.

The second sample, appropriate for $z \geq 1$, is a photometric sample of galaxies such as will be observed by the Vera Rubin Observatory - Legacy Survey of Space and Time (LSST; [19]) and thus we will refer to it as LSST sample. We follow ref. [20] and consider an analogue of Lyman break dropout galaxies (LBGs), which is fairly typical of proposed future surveys [1, 21, 22]. This sample has high number density but poor radial resolution, leading to a density field with lower noise at low $k_\parallel$ but high noise for large $k_\parallel$. As above we model these galaxies as a mass-limited halo sample. We introduce the photometric redshift scatter simply by enhancing the noise for this sample as an exponential in $k_\parallel$ (see 3.3). We will consider the LSST sample at two redshifts, $z = 1$ and $z = 4$ with number densities $\bar{n} = 5 \times 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$ and $\bar{n} = 3.5 \times 10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$ respectively. We expect our results to qualitatively remain the same for other photometric samples as long as they can be described with a bias model and redshift scatter, even if the specific gains may vary depending on the number density and photometric smoothing.

3 Method

Our reconstruction method largely follows the steps outlined in Ref. [13]. We reconstruct the initial density field by optimizing its posterior, conditioned on the observed data (HI and galaxy density in different regions of k-space), assuming Gaussian initial conditions. Evolving this initial field allows us to reconstruct the observed data on all scales. The initial conditions are reconstructed in the manner described in refs. [13, 23, 24] (for alternative approaches to reconstruction see also refs. [25–29]).

For the forward model $[\mathcal{F}(\mathbf{s})]$ connecting the observed data (δ) with the Gaussian initial conditions (ICs; $\mathbf{s}$) we use a second order Lagrangian bias model coupled to the non-linear dynamics of the simulation. We explore different non-linear dynamics for the gravitational evolution and find the best results when using Zeldovich dynamics to evolve particles from their Lagrangian to Eulerian position. The forward-modeled final density fields (of galaxies and HI) are obtained by assigning particles to a grid at their final redshift-space positions with a weight that is a function of the density and shear at their positions in the initial conditions [13]. Our Lagrangian bias model [13, 30–37] connects the matter field to the dark matter halos and includes terms up to quadratic order (δ_L, $\delta^2_{L,R}$ and $s^2_{L,R} \equiv \sum_{ij} s^2_{ij}$ the scalar shear field[1] where $s^2_{ij} = (\partial_i\partial_j\partial^{-2} - [1/3]\delta^D_{ij})\delta_L$) computed from the ICs of the simulation, evolved to $z = 0$ using linear theory. We subtract the zero-lag terms to make these fields have zero mean.

To construct the bias terms of our model, and in particular to estimate the quadratic operators, we smooth the linear density field with a Gaussian kernel with smoothing scale R. The dependence on the smoothing scale R of reconstruction algorithms is an open problem

[1] Since δ^2_L and s^2 are correlated, we actually define a new field $g^2_L = \delta^2_L - s^2$ which does not correlate with δ^2_L on large scales and use this instead of shear field.

in the field [38, 39], since a formal understanding of the renormalization of the bias expansion at the field level has not been obtained yet [36, 39]. It should also be kept in mind that, even without any additional smoothing, the size of the FFT grid provides an unavoidable cut off of the power on small scales. We thus explore different smoothing scales for reconstruction and empirically motivate our choices. For $z = 4$, any smoothing leads to lower reconstruction performance, implying that the optimal smoothing is likely smaller than the grid resolution. For $z = 1$, we find that when reconstructing with only HI data, in which case the only large scale information is provided by non-linear coupling of gravitational evolution, the large scales are best reconstructed when smoothing the linear field in the forward model of HI data with $R = 6\,h^{-1}$Mpc. However when combining this with additional data of galaxy field, we find that no smoothing is required since the information on large scales is dominated by the galaxy field. Thus for $z = 1$, we smooth the linear field in the forward model of HI data with $R = 6\,h^{-1}$Mpc and do not smooth the linear field for the galaxy data. We intend to the return to the issue of the smoothing scale in a forthcoming publication.

Our modeled tracer field is then [13, 37, 40]:

$$\delta^b_{\rm HI}(\mathbf{x}) = \delta_{[1]}(\mathbf{x}) + b_1\delta_{[\delta_L]}(\mathbf{x}) + b_2\delta_{[\delta^2_{L,R}]}(\mathbf{x}) + b_g\delta_{[g_{L,R}]}(\mathbf{x}) \qquad . \tag{3.1}$$

where $\delta_{[W]}(\mathbf{x})$ refers to the field generated by weighting the particles with the 'W' field.

To fit[2] for the bias parameters, we minimize the mean square model error between the data and the model fields which is equivalent to minimizing the error power spectrum i.e. the power spectrum of the residuals between the bias model and true (clean) data, $\mathbf{r}(k) = \delta^b(\mathbf{k}) - \delta^{data}(\mathbf{k})$, in Fourier space. We do this separately for the HI and galaxy fields and thus have two sets of bias parameters. The smoothing scale affects the forward modeling and reconstruction differently. The accuracy of forward modeling the observation from true initial conditions is not as sensitive but the accuracy of the reconstructed field from observed data gets impacted more significantly. Thus ideally one would like to keep both the bias parameters and the smoothing scale as a free-parameters to be fit at the time of reconstruction instead of fitting them in advance. We plan to explore this in the future.

Once the bias parameters are known, we reconstruct the initial (and final density) field by maximizing the posterior as a function of the IC amplitudes, $\mathbf{s}(\mathbf{k})$, using L-BFGS[3] [41]. The negative log-likelihood for the Gaussian prior can be combined with the negative log-likelihood of the data to get the posterior (see also [36, 38])

$$\begin{aligned}\mathcal{P} = &\sum_k \frac{1}{\mathrm{N_{modes}(k)}}\left(\sum_{\substack{\mathbf{k},|\mathbf{k}|\sim k,\\ \mathbf{k}\notin w}} \frac{|\delta^b_{\rm HI}(\mathbf{k}) - \delta^{\rm obs}_{\rm HI}(\mathbf{k})|^2}{P_{\rm err-HI}(k,\mu)}\right)\\ &+\sum_k \frac{1}{\mathrm{N_{modes}(k)}}\left(\sum_{\mathbf{k},|\mathbf{k}|\sim k} \frac{|\delta^b_{\rm g}(\mathbf{k}) - \delta^{\rm obs}_{\rm g}(\mathbf{k})|^2}{P_{\rm err-g}(k)}\right)\\ &+\sum_k \frac{1}{\mathrm{N_{modes}(k)}}\left(\sum_{\mathbf{k},|\mathbf{k}|\sim k} \frac{|\mathbf{s}(\mathbf{k}))|^2}{P_{\rm s}(k)}\right)\end{aligned} \tag{3.2}$$

[2] In principle one could fit for the bias parameters using summary statistics, such as the power spectrum, as in e.g. refs. [37, 40], though we anticipate that the constraints from the field itself would be tighter.

[3] https://en.wikipedia.org/wiki/Limited-memory_BFGS

where $P_{\rm s}$ is the prior power spectrum of the initial conditions and the sum is over modes that are measured by each survey (i.e. modes not in the foreground wedge for HI and low $k_{\parallel}$ modes for the galaxies). For the HI field the error power spectrum, $P_{\rm err-HI}$, is a combination of the modeling error estimated from the simulations after fitting the bias parameter, and the noise power spectrum. The noise changes the amplitude of $P_{\rm err-HI}$, especially on small scales, and also introduces an angular dependence. We have indicated this by the μ dependence in $P_{\rm err-HI}$. Note the data automatically include shot-noise, since we have a single realization of the halo field in the simulation.

For the galaxies the error power spectrum, $P_{\rm err-g}$, is due to a combination of shot noise and modeling error estimated from the simulations with fitted bias parameters. For the photometric sample we also need to include the smearing of the density field along the line of sight. We include this by damping the signal in the likelihood term with a Gaussian smoothing kernel:

$$\sum_{\mathbf{k},|\mathbf{k}|\sim k} |\delta_{\rm g}^{b}(\mathbf{k}) - \delta_{\rm g}^{\rm obs}(\mathbf{k})| \to \sum_{\mathbf{k},|\mathbf{k}|\sim k} |\delta_{\rm g}^{b}(\mathbf{k}) - \delta_{\rm g}^{\rm obs}(\mathbf{k})| \exp\left(-\frac{k^2\mu^2}{2\sigma_{\rm ph}^2}\right) \quad . \tag{3.3}$$

Here $\sigma_{\rm ph}$ is the photometric smoothing scale along the line of sight. It is equal to $\simeq 180\,h^{-1}$Mpc at $z=1$ and $\simeq 100\,h^{-1}$Mpc at $z=4$.

4 Results

In this section we present the results for our reconstructions. Our primary metrics to gauge the performance of our model and reconstruction are the cross correlation function, $r_{cc}(k)$, and the transfer function, $T_f(k)$, defined as

$$r_{cc}(k) = \frac{P_{XY}(k)}{\sqrt{P_X(k)P_Y(k)}} \quad , \quad T_f(k) = \sqrt{\frac{P_Y(k)}{P_X(k)}} \quad , \tag{4.1}$$

The cross correlation, r_{cc}, measures how faithfully the reconstructed map describes the input map, up to rescalings of the output map amplitude. For better visual clarity, we instead show the error power spectrum $(1-r_{cc}^2)$ with lower values indicating better reconstruction. The transfer function, on the other hand, tells us about the amplitude of the output map as a function of scale, with $r\,T_f = P_{XY}/P_X$. These metrics will always be defined between either the model or the reconstructed fields as Y, and the corresponding true field as X unless explicitly specified otherwise.

We have to consider 2 sets of bias parameters, one for HI and the other for the galaxies. We keep them fixed under the assumption that they have been estimated prior to reconstruction. For the HI field at $z=1$, we also smooth the initial density field with $R=6\,h^{-1}$Mpc to estimate quadratic terms as described earlier. The dynamics in the forward model is taken to be the Zeldovich approximation (ZA). All of the data are in redshift space, with the forward model using the ZA velocities to perform the translation. The reconstruction procedure is outlined in detail in ref. [13] but, briefly, it is done in a series of optimization steps. We begin on a 256^3 grid and smooth the likelihood term in Eq. 3.2 for both galaxies and HI on small scales i.e. multiply the residuals with Gaussian kernel to fit the large scales first. This smoothing is reduced in 5 steps (16, 8, 4, 2, $0\,h^{-1}$Mpc), each with 100 iterations. More details on this annealing scheme can be found in [14, 24]. Note that this smoothing is different from

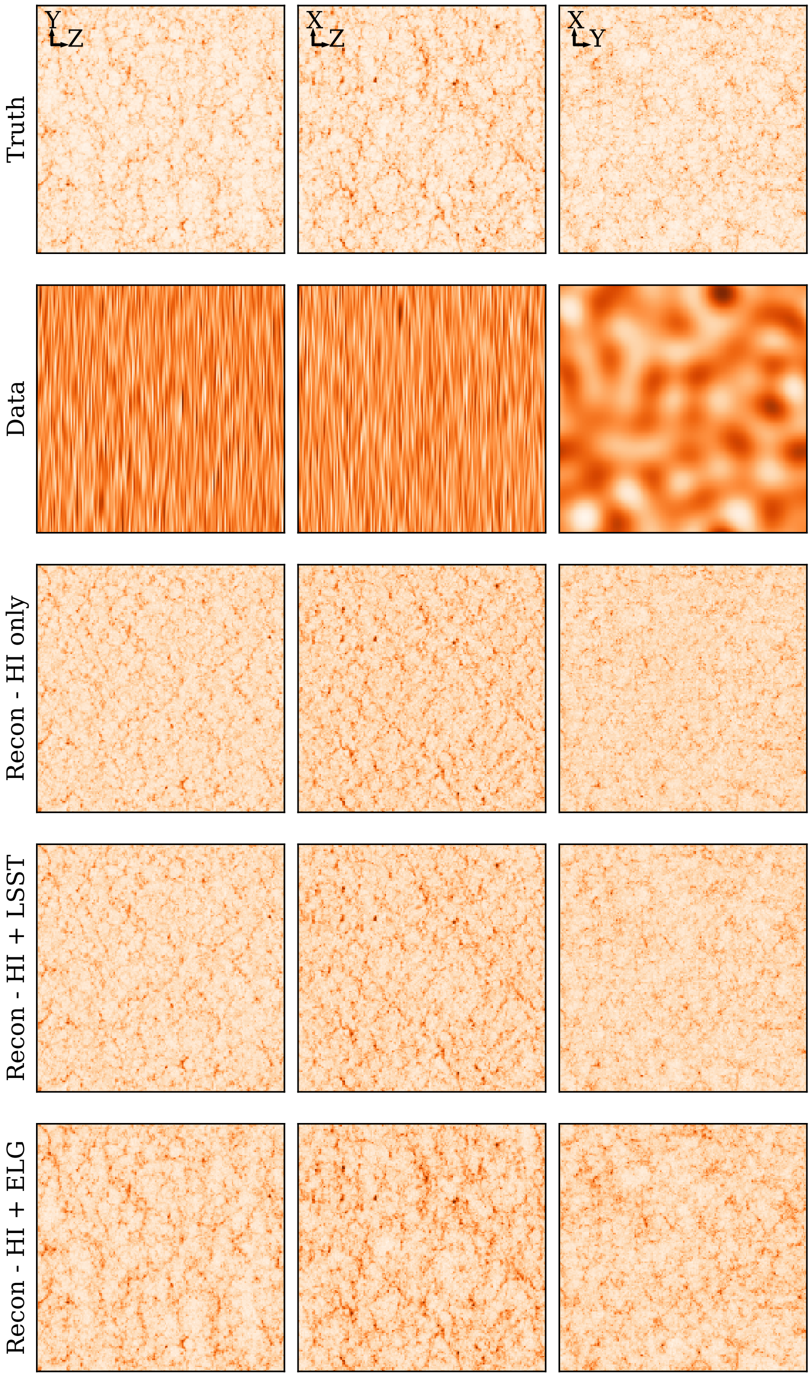

Figure 2. Slices perpendicular to the 3 box axes of the true HI field; the data with thermal noise and wedge; and the reconstructed HI field with HI data only, HI+LSST and HI+ELG data at $z = 1$. We project over $80\,h^{-1}$Mpc slices, transverse directions show the full box ($1024\,h^{-1}$Mpc). The color scale is the same for all rows and columns except the data row.

the smoothing of the linear field to estimate quadratic components of the bias model discussed earlier and is part of the optimization scheme, not the forward model. The reconstructed field is then upsampled to 512^3 grid and smoothing is reduced in 2 steps (2 and $0\,h^{-1}$Mpc) each with 100 iterations.

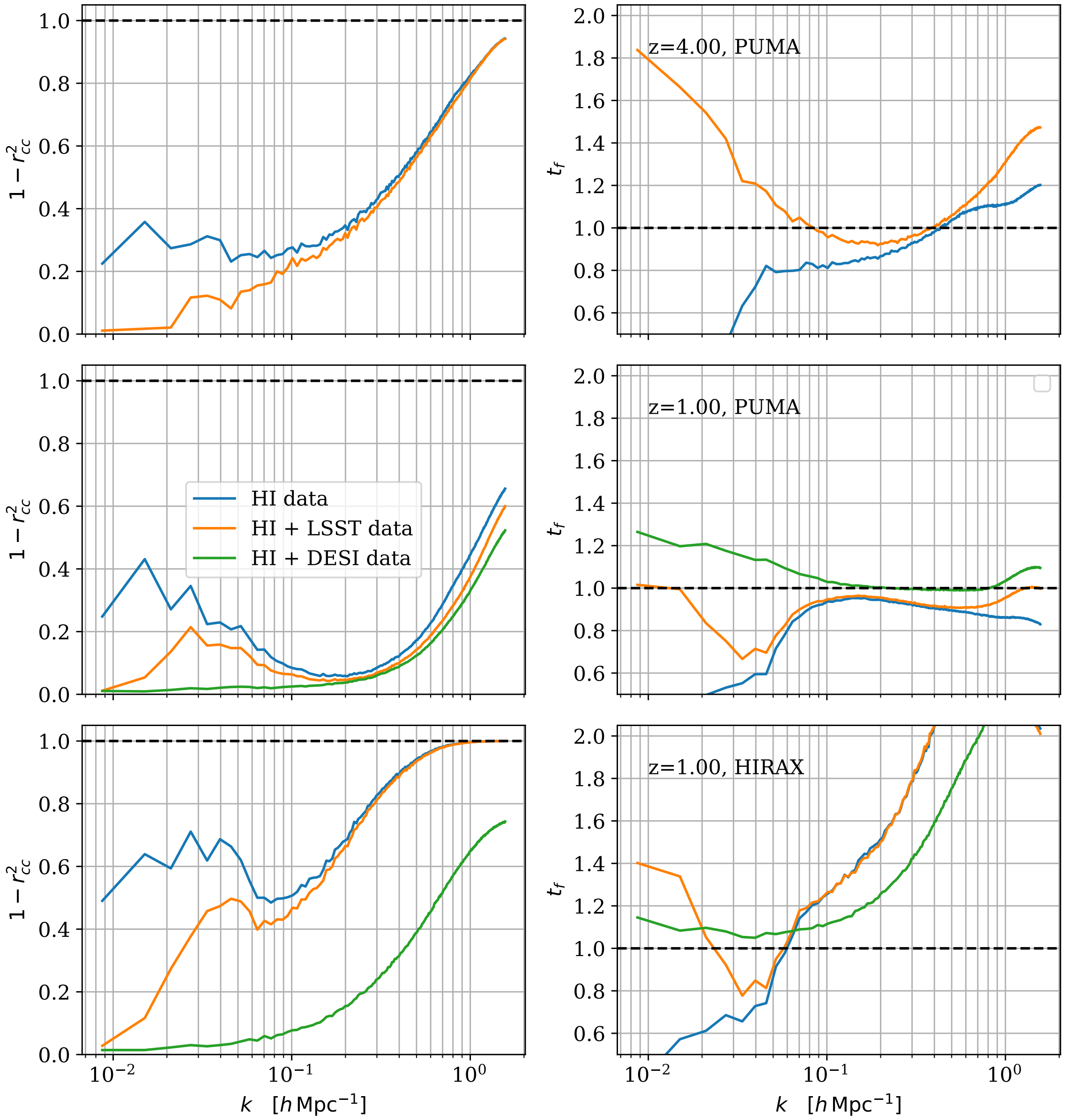

Figure 3. Cross correlation (left) and transfer function (right) of the reconstructed HI field for PUMA noise level at $z = 4$ (top) and $z = 1$ (middle), and with HIRAX noise level at $z = 1$ (bottom). We show results for reconstruction without galaxies (blue), with a spectroscoic DESI-ELG like sample of $\bar{n} = 10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$ (green), and a photometric LSST-like sample (orange) of $\bar{n} = 5 \times 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$ and $\bar{n} = 3.5 \times 10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$ at $z = 1$ and $z = 4$ respectively.

We begin by showing the data and reconstruction at the level of the fields in Figure 2 for $z = 1$. The first two rows show the true HI field and the HI data with thermal noise and fore-

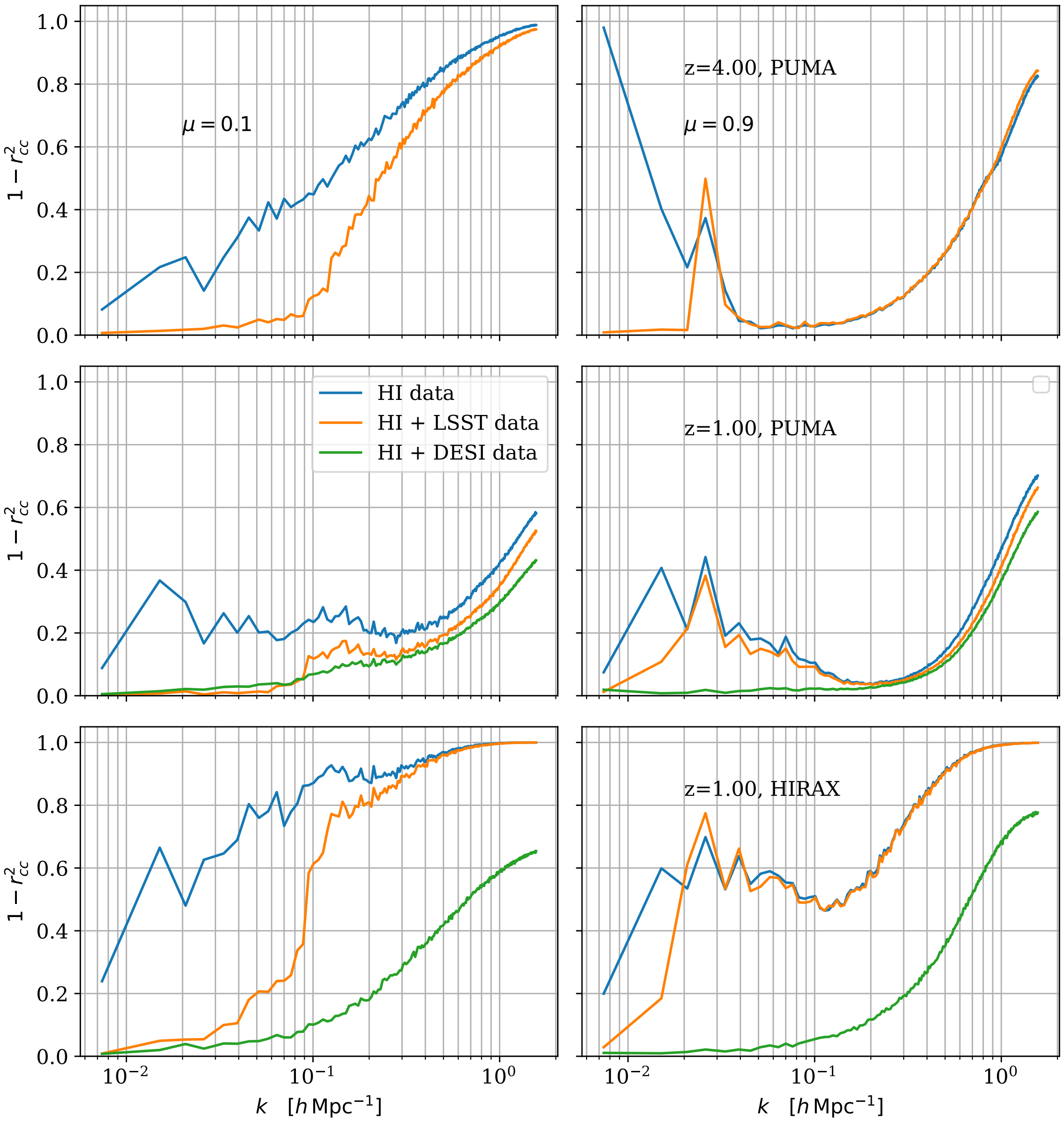

Figure 4. Same as Figure 3 but showing the cross-correlation for the corresponding fields in wedges perpendicular to the line of sight (left, $\mu = [0, 0.2]$) and along the line of sight (right, $\mu \in [0.8, 1]$).

ground wedge. The next three rows show the HI reconstructed field when the reconstruction is done with only HI data, HI and LSST data as well as HI and ELG data. Reconstruction is closest to the underlying truth when we combine HI with ELG data, but improvement with LSST data is also apparent in X-Z and Y-Z projections.

Figure 3 shows the results at the level of the two point function for the following three cases: $z = 4$ with PUMA and $z = 1$ with both PUMA and HIRAX. For $z = 1$ we show results with both photometric LSST and spectroscopic DESI-ELG data, while at $z = 4$ we only have the photometric galaxy sample. For comparison, we also show the reconstruction with HI data only. In each case, we see an improvement in the cross-correlation between the reconstructed HI field and the true HI field as compared to the case when reconstruction is done only with HI data. At $z = 1$, we find that the reconstruction with data from a spectroscopic survey far outperforms the reconstruction with a photometric survey, even though the latter has 50 times higher number density. At $z = 4$, as the photometric smoothing scale decreases for LSST, the reconstruction improves and we can recover the largest scales almost perfectly. Additionally, in every case we find that more power is reconstructed at the largest scales for HI than the original case, as shown in the transfer function. Thus we are recovering more structure when we add information from different tracers, however at the same time it consistently biased high. If uncorrected, this can lead to potential biases. Previous work has suggested correcting for this using simulations [23, 24], but implementing such a correction is beyond the scope of this work.

In Figure 4, we show the same results for cross-correlation but in μ bins along and perpendicular to the line of sight $\mu \in [0 - 0.2]$ and $\mu \in [0.8 - 1.0]$. We remind the reader that our goal was to supplement the large scale modes lost in the foreground wedge, especially those perpendicular to the line of sight, with modes in galaxy clustering surveys. Both spectroscopic and photometric surveys probe the perpendicular modes, but the latter loses the line of sight ones. Figure 4 clearly shows that the large-scale modes perpendicular to the line of sight are reconstructed very well. Furthermore, when reconstructed with HI data, the LSST field also recovers the modes along the line of sight that are otherwise missing due to photometric uncertainties. The combination of 21-cm data and LSST is therefore greater than the sum of its parts. A translation of these metrics into performance gains for particular science goals can be found in §6 of ref. [13].

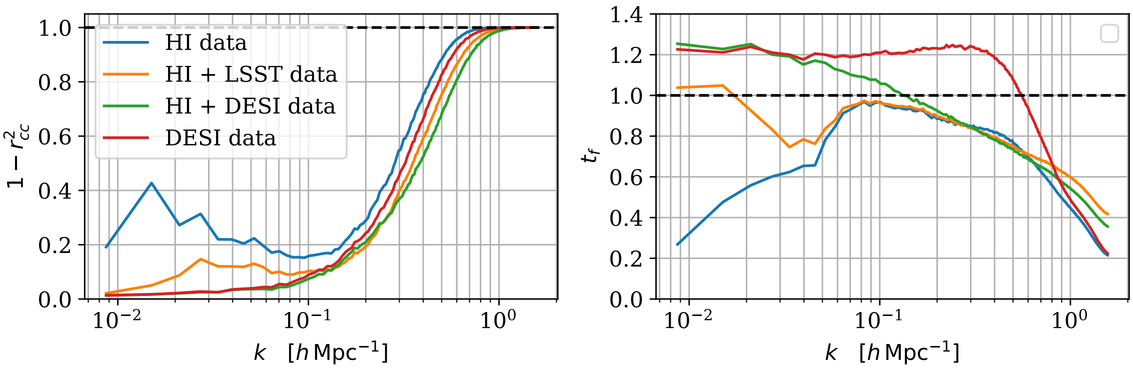

Figure 5. Angle average cross correlation and transfer function for the reconstructed initial field with the true initial field at $z = 1$ for PUMA noise-levels. We show the results when reconstruction is done with only HI data, only DESI ELG-like data, $\bar{n} = 10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$, the combination of the two and HI with LSST data.

In addition to filling in the foreground wedge of HI data, we can also use the combination of the different surveys to explore other synergies between the different cosmological tracers. For instance we can ask how well our method reconstructs the primordial initial conditions that are shared by, and give rise to, both the observed tracer fields. In Figure 5, we show the cross-correlation coefficient and transfer function for reconstructed initial conditions from data at $z = 1$ with the true initial conditions. We compare the cases when reconstruction is done only with a single tracer, such as HI data or ELG-like data with the combination of surveys i.e. HI and ELG or HI and LSST data. On large scales, DESI and LSST data dominate, improving the reconstruction over using only HI field. However on small scales, HI allows us to improve over the galaxy fields and push further into the non-linear regime. The combination of the different probes yields higher returns than each tracer considered independently.

5 Conclusions

Interferometric 21-cm surveys have the potential to map out the distribution of matter in the largest cosmological volumes with good radial resolution. However they must contend with bright foregrounds that, when coupled to instrument imperfections, can lead to a loss of large scale modes in the foreground wedge. In this work we build upon the forward modeling approach of Ref. [13], that reconstructed these long-wavelength fluctuations by exploiting the non-linear coupling between small and large scales induced by gravitational evolution, by adding external galaxy distribution data that can help to fill in the missing modes. Specifically, we consider supplementing the 21-cm data from mock HIRAX and PUMA surveys at different redshifts with a spectroscopic sample (good radial resolution but low number density) and a photometric sample (high number density but poor radial resolution). The mock galaxies for these datasets are loosely modeled on DESI-ELG sample at $z = 1$ and LSST sample at $z = 1$ and $z = 4$ respectively.

We find that the spectroscopic sample reconstructs the modes significantly better than the photometric sample, despite having much lower number density. Both galaxy samples are able to reconstruct the largest modes ($k < 0.1\,h\,\mathrm{Mpc}^{-1}$) transverse to the line of sight very well with $r_c > 95\%$ for PUMA noise levels at both the redshifts. However the contribution of an LSST-like photometric sample to scales smaller than $k \simeq 0.1\,h\,\mathrm{Mpc}^{-1}$ is not significant, especially for the thermal noise of HIRAX survey. At the same time, we find that 21-cm data also reconstructs the correlations in the LSST data along the line of sight on small scales that are otherwise lost due to photometric smoothing. The spectroscopic sample, on the other hand, improves reconstruction across all the scales, especially for a noise-dominated survey like HIRAX where the reconstruction with only 21-cm data is poor ($r_c = 60\%$ even at $k = 1\,h\,\mathrm{Mpc}^{-1}$). We also explore the synergies of different surveys in reconstructing the initial density field and find that the combination of surveys performs better ($r_c = 90\%$ at $k = 0.1\,h\,\mathrm{Mpc}^{-1}$) than using surveys individually (best $r_c = 84\%$ for a single spectroscopic survey).

With regards to forward modeling approaches, we find that the smoothing scale plays an important role in quadratic bias model when the data itself lacks any direct information on large scales, such as the HI field at low redshifts. In this case, using a large smoothing scale to suppress small scales non-linearities when estimating quadratic fields improves reconstruction of large scale modes in HI data. Interestingly, this does not seem to be the case at higher redshifts. The appropriate numerical procedure for the implementation of an effective field theory when modeling large-scale structure at the field level (that would remove the dependence on the smoothing scale) remains an area of active research, and our results show the importance of understanding this issues at a more fundamental level. We plan on pursuing these directions in future work.

Acknowledgments

M.W. is supported by the U.S. Department of Energy and by NSF grant number 1713791. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. This work made extensive use of the NASA Astrophysics Data System and of the `astro-ph` preprint archive at `arXiv.org`.

A HiddenValley2

In this work we have used a second run of the Hidden Valley simulations, first reported in ref. [14]. The HiddenValley2 simulation employs the same code, box size, particle loading and initial conditions but has been run to $z = 0.5$ with outputs at $z = 1.5$, 1.0 and 0.5. In addition to lower redshift outputs we have also updated the model used to populate the dark matter halos in the simulation with HI, with parameters recalibrated to the wider redshift range and updated measurements of the abundance and clustering of HI.

We make use of an $M_{HI} - M_{\rm halo}$ relation in order to populate our dark matter only simulation with HI, much as we did in our earlier work [14]. We assume the total HI mass in a halo of mass M_h is [42, 43]

$$M_{HI}(M_h) = A(z)\left(\frac{M_h}{M_{\rm cut}}\right)^{\alpha}\ \exp\left[-\frac{M_{\rm cut}}{M_h}\right] \tag{A.1}$$

This HI mass is split into a central component and a component that moves with the virial velocity of the halo. Specifically a fraction $f_{\rm cen}$ of the HI mass is taken to reside at the halo center and move with the halo center-of-mass velocity. The remaining $f_{\rm sat} = 1 - f_{\rm cen}$ of the HI mass has an additional, Gaussian line-of-sight velocity distribution with dispersion equal to the virial velocity dispersion of the halo. For numerical convenience we implemented this by dividing this HI into $N_{\rm sat} = \lfloor 1 + (0.1M_h/M_{\rm cut})^{0.5}\rfloor$ equal parts and for each drawing an additional line-of-sight velocity component from a Gaussian. As we found in refs. [13, 14] that the details of how we treated such fingers of god were largely unimportant for our science, we have opted to take the simple modeling approach here. Following Figure 7 of ref. [44], we model the fraction of HI in satellites as:

$$f_{\rm sat} = \min\left[0.8,\ \frac{0.5\times(\log M_h - 9.5)^2}{12.8 - 9.5}\right] \quad \forall M_h > 10^{9.5}\, h^{-1}M_\odot \tag{A.2}$$

otherwise $f_{\rm sat} = 0$.

There is unfortunately very little data with which to tune the parameters of Eq. (A.1). The amplitude, $A(z)$, is largely constrained by the abundance of HI, $\tilde{\Omega}_{HI}(z)$ (Fig. 6). We have followed the common convention in absorption line studies and HI intensity mapping and quoted the abundance as a comoving HI density divided by the (physical) $z = 0$ critical density. However we have used a tilde to distinguish this quantity from the more common usage of Ω as a ratio of (physical or comoving) density at z to critical density at z. The agreement with the data above $z \approx 0$ is quite good. The values of α and $M_{\rm cut}$ are less constrained. Both physical intuition and numerical simulations suggest that there is a minimum halo mass ($M_{\rm cut} \sim 10^9 - 10^{10} M_\odot$) below which neutral hydrogen will not be self-shielded from UV photons. Above this mass the amount of HI should increase as the halo mass increases, though not necessarily linearly (however, simulations suggest $\alpha \approx 1$ at $z > 2$). The characteristic halo

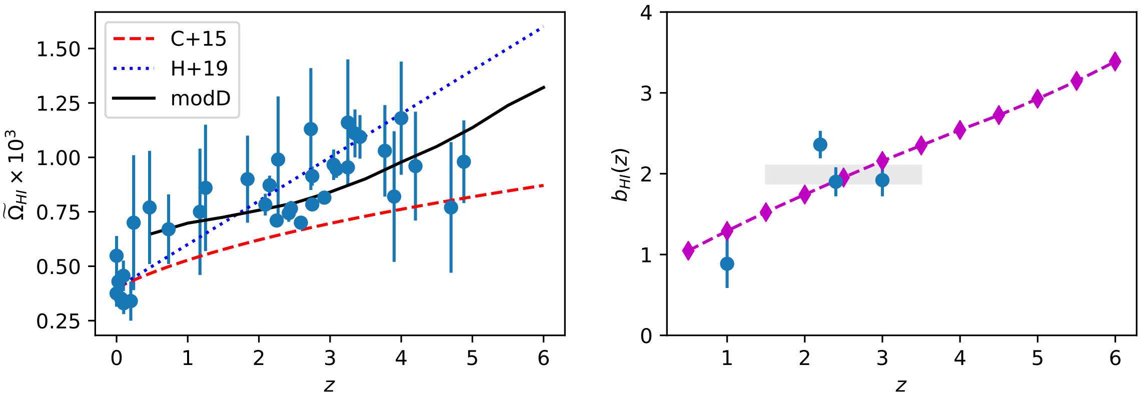

Figure 6. (Left) The abundance of HI as a function of redshift (points; [45–58]). The dotted line shows a linear fit, $\tilde{\Omega}_{HI} = (0.4 + 0.2\,z) \times 10^{-3}$, from Fig. 14 of ref. [58] while the dashed line shows the power-law fit, $\tilde{\Omega}_{HI} = 0.4(1+z)^{0.4}$, of ref. [56] and the solid black line shows our fiducial model. (Right) The HI bias as a function of redshift. The error on the $z \approx 1$ point is dominated by the uncertainty in $\tilde{\Omega}_{HI}$.

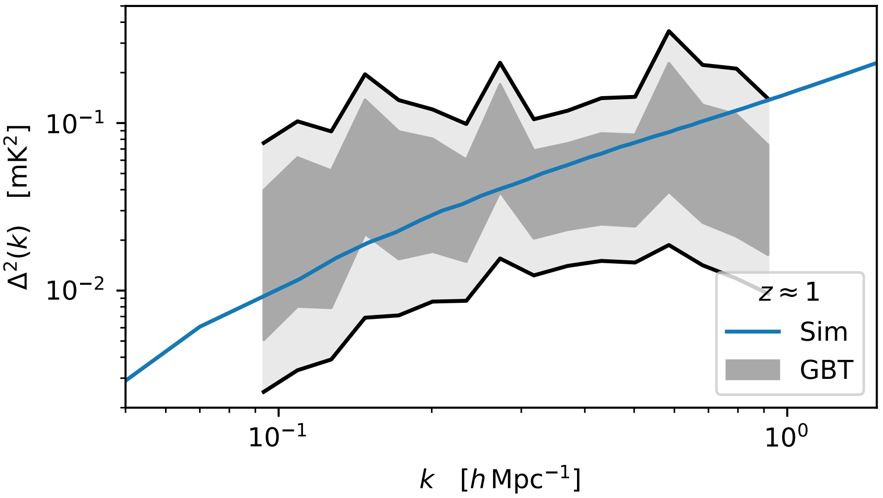

Figure 7. The monopole redshift-space clustering of HI at $z = 1$ from our model, compared to the GBT measurements of refs. [59, 60] at $z \simeq 0.8$ (grey bands showing 68% and 95% confidence regions). The upper limits from the HI auto power are approximately independent in each bin but the lower limits are perfectly correlated (see ref. [60]).

mass scale ($M_{\rm cut}$) is determined by the clustering of HI. At $z \simeq 0$ an analysis of HI-selected galaxies in the Sloan Digital Sky Survey constrains the HOD [61]. At $z \simeq 1$ the large-scale bias is known approximately from cross-correlation with optical galaxies [59, 60, 62, 63], however at higher z there are no direct measurements. The clustering of Damped Lyα systems (DLAs) has been measured at $z \simeq 2-3$ by ref. [64], and since the DLAs contain the majority of the HI at those redshifts this can be used as a proxy for the HI clustering amplitude. The numbers inferred agree reasonably well with an analysis of the most recent hydrodynamical simulations (see Table 6 of ref. [44]). Based on these considerations we take $\alpha = (1+2z)/(2+2z)$,

$$A(z) = 1.7\times 10^{9}\,(1+z)^{-5/3}\ \ h^{-1}M_{\odot}\quad ,\quad M_{\rm cut} = 6\times 10^{10}\exp\left(-\frac{3z}{4}\right)\ h^{-1}M_{\odot}\quad . \tag{A.3}$$

Figure 6 shows the abundance and large-scale bias of HI predicted from our model ('Model D') compared to observations. Our model has sufficient flexibility to fit the available data, while also being in broad agreement with the current generation of hydrodynamic simulations. From Fig. 6 it appears our model overpredicts the clustering at $z \approx 1$, however this is partly due to the way the large-scale bias is estimated in the observations. We take a closer look at the agreement at $z \simeq 0.8$ (we use the $z = 1$ output of our simulation) in Fig. 7. Here we compare the product, $\bar{T}^2\ \Delta_0^2(k)$, predicted by our simulation to the range allowed by the HI auto-correlation and HI-WiggleZ cross-correlation [59, 60]. While there may be some evidence for more small-scale power in the model than the observations, the level of agreement is quite good for most of the range and within the errors on the observation for all scales.

References

[1] S. Ferraro and M. J. Wilson, *Inflation and Dark Energy from spectroscopy at $z > 2$*, *BAAS* **51** (2019) 72 [1903.09208].

[2] E. D. Kovetz, M. P. Viero, A. Lidz, L. Newburgh, M. Rahman, E. Switzer et al., *Line-Intensity Mapping: 2017 Status Report*, *ArXiv e-prints* (2017) [1709.09066].

[3] E. Kovetz, P. C. Breysse, A. Lidz, J. Bock, C. M. Bradford, T.-C. Chang et al., *Astrophysics and Cosmology with Line-Intensity Mapping*, in *BAAS*, vol. 51, p. 101, May, 2019, 1903.04496.

[4] S. R. Furlanetto, S. P. Oh and F. H. Briggs, *Cosmology at low frequencies: The 21 cm transition and the high-redshift Universe*, *PhysRep* **433** (2006) 181 [astro-ph/0608032].

[5] F. B. Abdalla, P. Bull, S. Camera, A. Benoit-Lévy, B. Joachimi, D. Kirk et al., *Cosmology from HI galaxy surveys with the SKA*, *Advancing Astrophysics with the Square Kilometre Array (AASKA14)* (2015) 17 [1501.04035].

[6] M. Santos, P. Bull, D. Alonso, S. Camera, P. Ferreira, G. Bernardi et al., *Cosmology from a SKA HI intensity mapping survey*, *Advancing Astrophysics with the Square Kilometre Array (AASKA14)* (2015) 19 [1501.03989].

[7] Cosmic Visions 21 cm Collaboration, R. Ansari, E. J. Arena, K. Bandura, P. Bull, E. Castorina et al., *Inflation and Early Dark Energy with a Stage II Hydrogen Intensity Mapping Experiment*, *arXiv e-prints* (2018) arXiv:1810.09572 [1810.09572].

[8] E. Castorina, S. Foreman, D. Karagiannis, A. Liu, K. W. Masui, P. D. Meerburg et al., *Packed Ultra-wideband Mapping Array (PUMA): Astro2020 RFI Response*, *arXiv e-prints* (2020) arXiv:2002.05072 [2002.05072].

[9] H.-J. Seo and C. M. Hirata, *The foreground wedge and 21-cm BAO surveys*, *MNRAS* **456** (2016) 3142 [1508.06503].

[10] J. D. Cohn, M. White, T.-C. Chang, G. Holder, N. Padmanabhan and O. Doré, *Combining galaxy and 21-cm surveys*, *MNRAS* **457** (2016) 2068 [1511.07377].

[11] A. Obuljen, E. Castorina, F. Villaescusa-Navarro and M. Viel, *High-redshift post-reionization cosmology with 21cm intensity mapping*, *JCAP* **05** (2018) 004 [1709.07893].

[12] S.-F. Chen, E. Castorina, M. White and A. Slosar, *Synergies between radio, optical and microwave observations at high redshift*, *JCAP* **2019** (2019) 023 [1810.00911].

[13] C. Modi, M. White, A. Slosar and E. Castorina, *Reconstructing large-scale structure with neutral hydrogen surveys*, *JCAP* **2019** (2019) 023 [1907.02330].

[14] C. Modi, E. Castorina, Y. Feng and M. White, *Intensity mapping with neutral hydrogen and the Hidden Valley simulations*, *JCAP* **2019** (2019) 024 [1904.11923].

[15] Y. Feng, M.-Y. Chu, U. Seljak and P. McDonald, *FASTPM: a new scheme for fast simulations of dark matter and haloes*, *MNRAS* **463** (2016) 2273 [1603.00476].

[16] L. B. Newburgh, K. Bandura, M. A. Bucher, T. C. Chang, H. C. Chiang, J. F. Cliche et al., *HIRAX: a probe of dark energy and radio transients*, in *Proc. SPIE* , vol. 9906 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, p. 99065X, Aug., 2016, 1607.02059, DOI.

[17] A. Slosar, Z. Ahmed, D. Alonso, M. A. Amin, E. J. Arena, K. Bandura et al., *Packed Ultra-wideband Mapping Array (PUMA): A Radio Telescope for Cosmology and Transients*, in *Bulletin of the AAS*, vol. 51, p. 53, Sep, 2019, 1907.12559.

[18] DESI Collaboration, A. Aghamousa, J. Aguilar, S. Ahlen, S. Alam, L. E. Allen et al., *The DESI Experiment Part I: Science,Targeting, and Survey Design*, *ArXiv e-prints* (2016) [1611.00036].

[19] LSST Science Collaboration, P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel et al., *LSST Science Book, Version 2.0*, *ArXiv e-prints* (2009) [0912.0201].

[20] M. J. Wilson and M. White, *Cosmology with dropout selection: straw-man surveys &amp; CMB lensing*, *JCAP* **2019** (2019) 015 [1904.13378].

[21] C. Modi, M. White and Z. Vlah, *Modeling CMB lensing cross correlations with CLEFT*, *JCAP* **8** (2017) 009 [1706.03173].

[22] D. Schlegel, J. A. Kollmeier and S. Ferraro, *The MegaMapper: a z>2 spectroscopic instrument for the study of Inflation and Dark Energy*, in *Bulletin of the American Astronomical Society*, vol. 51, p. 229, Sept., 2019, 1907.11171.

[23] U. Seljak, G. Aslanyan, Y. Feng and C. Modi, *Towards optimal extraction of cosmological information from nonlinear data*, *Journal of Cosmology and Astro-Particle Physics* **2017** (2017) 009 [1706.06645].

[24] C. Modi, Y. Feng and U. Seljak, *Cosmological reconstruction from galaxy light: neural network based light-matter connection*, *JCAP* **10** (2018) 028 [1805.02247].

[25] J. Jasche and B. D. Wandelt, *Bayesian physical reconstruction of initial conditions from large-scale structure surveys*, *MNRAS* **432** (2013) 894 [1203.3639].

[26] H. Wang, H. J. Mo, X. Yang, Y. P. Jing and W. P. Lin, *ELUCID—Exploring the Local Universe with the Reconstructed Initial Density Field. I. Hamiltonian Markov Chain Monte Carlo Method with Particle Mesh Dynamics*, *ApJ* **794** (2014) 94 [1407.3451].

[27] H.-M. Zhu, Y. Yu, U.-L. Pen, X. Chen and H.-R. Yu, *Nonlinear reconstruction*, *Phys. Rev. D* **96** (2017) 123502 [1611.09638].

[28] N. G. Karaçaylı and N. Padmanabhan, *Anatomy of cosmic tidal reconstruction*, *MNRAS* **486** (2019) 3864.

[29] M. Schmittfull, T. Baldauf and M. Zaldarriaga, *Iterative initial condition reconstruction*, *PRD* **96** (2017) 023505 [1704.06634].

[30] T. Matsubara, *Resumming cosmological perturbations via the Lagrangian picture: One-loop results in real space and in redshift space*, *PRD* **77** (2008) 063530 [0711.2521].

[31] T. Matsubara, *Nonlinear perturbation theory with halo bias and redshift-space distortions via the Lagrangian picture*, *PRD* **78** (2008) 083519 [0807.1733].

[32] J. Carlson, B. Reid and M. White, *Convolution Lagrangian perturbation theory for biased tracers*, *MNRAS* **429** (2013) 1674 [1209.0780].

[33] M. White, *The Zel'dovich approximation*, *MNRAS* **439** (2014) 3630 [1401.5466].

[34] Z. Vlah, E. Castorina and M. White, *The Gaussian streaming model and convolution Lagrangian effective field theory*, *JCAP* **12** (2016) 007 [1609.02908].

[35] C. Modi, E. Castorina and U. Seljak, *Halo bias in Lagrangian space: estimators and theoretical predictions*, *MNRAS* **472** (2017) 3959 [1612.01621].

[36] M. Schmittfull, M. Simonović, V. Assassi and M. Zaldarriaga, *Modeling Biased Tracers at the Field Level*, *arXiv e-prints* (2018) [1811.10640].

[37] C. Modi, S.-F. Chen and M. White, *Simulations and symmetries*, *MNRAS* **492** (2020) 5754 [1910.07097].

[38] F. Schmidt, F. Elsner, J. Jasche, N. M. Nguyen and G. Lavaux, *A rigorous EFT-based forward model for large-scale structure*, *JCAP* **01** (2019) 042 [1808.02002].

[39] G. Cabass and F. Schmidt, *The Likelihood for LSS: Stochasticity of Bias Coefficients at All Orders*, *JCAP* **07** (2020) 051 [2004.00617].

[40] N. Kokron, J. DeRose, S.-F. Chen, M. White and R. H. Wechsler, *The cosmology dependence of galaxy clustering and lensing from a hybrid N-body-perturbation theory model*, *arXiv e-prints* (2021) arXiv:2101.11014 [2101.11014].

[41] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, New York, NY, USA, second ed., 2006.

[42] H. Padmanabhan, A. Refregier and A. Amara, *A halo model for cosmological neutral hydrogen : abundances and clustering*, *MNRAS* **469** (2017) 2323 [1611.06235].

[43] E. Castorina and F. Villaescusa-Navarro, *On the spatial distribution of neutral hydrogen in the Universe: bias and shot-noise of the H I power spectrum*, *MNRAS* **471** (2017) 1788 [1609.05157].

[44] F. Villaescusa-Navarro, S. Genel, E. Castorina, A. Obuljen, D. N. Spergel, L. Hernquist et al., *Ingredients for 21 cm Intensity Mapping*, *ApJ* **866** (2018) 135 [1804.09180].

[45] M. A. Zwaan, M. J. Meyer, L. Staveley-Smith and R. L. Webster, *The HIPASS catalogue: Ω_{HI} and environmental effects on the HI mass function of galaxies*, *MNRAS* **359** (2005) L30 [astro-ph/0502257].

[46] R. Braun, *Cosmological Evolution of Atomic Gas and Implications for 21 cm H I Absorption*, *ApJ* **749** (2012) 87 [1202.1840].

[47] A. M. Martin, E. Papastergis, R. Giovanelli, M. P. Haynes, C. M. Springob and S. Stierwalt, *The Arecibo Legacy Fast ALFA Survey. X. The H I Mass Function and Ω_H I from the 40% ALFALFA Survey*, *ApJ* **723** (2010) 1359 [1008.5107].

[48] J. Delhaize, M. J. Meyer, L. Staveley-Smith and B. J. Boyle, *Detection of H I in distant galaxies using spectral stacking*, *MNRAS* **433** (2013) 1398 [1305.1968].

[49] J. Rhee, M. A. Zwaan, F. H. Briggs, J. N. Chengalur, P. Lah, T. Oosterloo et al., *Neutral atomic hydrogen (H I) gas evolution in field galaxies at $z \sim 0.1$ and ~ 0.2*, *MNRAS* **435** (2013) 2693 [1308.1462].

[50] P. Lah, J. N. Chengalur, F. H. Briggs, M. Colless, R. de Propris, M. B. Pracy et al., *The HI content of star-forming galaxies at $z = 0.24$*, *MNRAS* **376** (2007) 1357 [astro-ph/0701668].

[51] S. M. Rao, D. A. Turnshek and D. B. Nestor, *Damped Lyα Systems at z&lt;1.65: The Expanded Sloan Digital Sky Survey Hubble Space Telescope Sample*, *ApJ* **636** (2006) 610 [astro-ph/0509469].

[52] S. M. Rao, D. A. Turnshek, G. M. Sardane and E. M. Monier, *The statistical properties of neutral gas at z &lt; 1.65 from UV measurements of Damped Lyman Alpha systems*, *MNRAS* **471** (2017) 3428 [1704.01634].

[53] P. Noterdaeme, P. Petitjean, W. C. Carithers, I. Paris, A. Font-Ribera, S. Bailey et al., *VizieR Online Data Catalog: SDSS-III DR9 DLA catalogue (Noterdaeme+, 2012)*, *VizieR Online Data Catalog* (2012) J/A+A/547/L1.

[54] A. Songaila and L. L. Cowie, *The Evolution of Lyman Limit Absorption Systems to Redshift Six*, *ApJ* **721** (2010) 1448 [1007.3262].

[55] T. Zafar, C. Péroux, A. Popping, B. Milliard, J. M. Deharveng and S. Frank, *The ESO UVES advanced data products quasar sample. II. Cosmological evolution of the neutral gas mass density*, *A&A* **556** (2013) A141 [1307.0602].

[56] N. H. M. Crighton, M. T. Murphy, J. X. Prochaska, G. Worseck, M. Rafelski, G. D. Becker et al., *The neutral hydrogen cosmological mass density at z = 5*, *MNRAS* **452** (2015) 217 [1506.02037].

[57] S. Bird, R. Garnett and S. Ho, *Statistical properties of damped Lyman-alpha systems from Sloan Digital Sky Survey DR12*, *MNRAS* **466** (2017) 2111 [1610.01165].

[58] W. Hu, L. Hoppmann, L. Staveley-Smith, K. Geréb, T. Oosterloo, R. Morganti et al., *An accurate low-redshift measurement of the cosmic neutral hydrogen density*, *MNRAS* **489** (2019) 1619 [1907.10375].

[59] K. W. Masui, E. R. Switzer, N. Banavar, K. Bandura, C. Blake, L.-M. Calin et al., *Measurement of 21 cm Brightness Fluctuations at* $z \sim 0.8$ *in Cross-correlation*, *ApJL* **763** (2013) L20 [1208.0331].

[60] E. R. Switzer, K. W. Masui, K. Bandura, L.-M. Calin, T.-C. Chang, X.-L. Chen et al., *Determination of* $z \sim 0.8$ *neutral hydrogen fluctuations using the 21 cm intensity mapping autocorrelation*, *MNRAS* **434** (2013) L46 [1304.3712].

[61] A. Obuljen, D. Alonso, F. Villaescusa-Navarro, I. Yoon and M. Jones, *The H I content of dark matter haloes at z ≈ 0 from ALFALFA*, *MNRAS* **486** (2019) 5124 [1805.00934].

[62] C. J. Anderson, N. J. Luciw, Y.-C. Li, C. Y. Kuo, J. Yadav, K. W. Masui et al., *Low-amplitude clustering in low-redshift 21-cm intensity maps cross-correlated with 2dF galaxy densities*, *MNRAS* **476** (2018) 3382 [1710.00424].

[63] L. Wolz, A. Pourtsidou, K. W. Masui, T.-C. Chang, J. E. Bautista, E.-M. Mueller et al., *HI constraints from the cross-correlation of eBOSS galaxies and Green Bank Telescope intensity maps*, *arXiv e-prints* (2021) arXiv:2102.04946 [2102.04946].

[64] I. Pérez-Ràfols, A. Font-Ribera, J. Miralda-Escudé, M. Blomqvist, S. Bird, N. Busca et al., *The SDSS-DR12 large-scale cross-correlation of damped Lyman alpha systems with the Lyman alpha forest*, *MNRAS* **473** (2018) 3019 [1709.00889].