Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **495,** 1788–1806 (2020)
Advance Access publication 2020 May 5

doi:10.1093/mnras/staa1191

HIR4: cosmology from a simulated neutral hydrogen full sky using Horizon Run 4

Jacobo Asorey,[1★] David Parkinson,[1★] Feng Shi,[1] Yong-Seon Song,[1] Kyungjin Ahn,[2] Juhan Kim,[3] Jian Yao,[4] Le Zhang[4,5] and Shifan Zuo[6,7]

[1] *Korea Astronomy and Space Science Institute, Yuseong-gu, Daedeok-daero 776, Daejeon 34055, Korea*
[2] *Department of Earth Sciences, Chosun University, Gwangju 61452, Korea*
[3] *Center for Advanced Computation, Korea Institute for Advanced Study, 85 Heogiro, Dongdaemun-gu, Seoul 02455, Korea*
[4] *School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, P. R. China*
[5] *School of Physics and Astronomy, Sun Yat-Sen University, 2 Daxue Road, Tangjia, Zhuhai 519082, P. R. China*
[6] *Key Laboratory of Computational Astrophysics, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, P. R. China*
[7] *Department of Astronomy and Tsinghua Center for Astrophysics, Tsinghua University, Beijing 100084, P. R. China*

Accepted 2020 April 21. Received 2020 April 20; in original form 2020 January 10

ABSTRACT

The distribution of cosmological neutral hydrogen will provide a new window into the large-scale structure of the Universe with the next generation of radio telescopes and surveys. The observation of this material, through 21 cm line emission, will be confused by foreground emission in the same frequencies. Even after these foregrounds are removed, the reconstructed map may not exactly match the original cosmological signal, which will introduce systematic errors and offset into the measured correlations. In this paper, we simulate future surveys of neutral hydrogen using the Horizon Run 4 (HR4) cosmological *N*-body simulation. We generate H I intensity maps from the HR4 halo catalogue, and combine with foreground radio emission maps from the Global Sky Model, to create accurate simulations over the entire sky. We simulate the H I sky for the frequency range 700–800 MHz, matching the sensitivity of the Tianlai pathfinder. We test the accuracy of the fastICA, PCA, and log-polynomial fitting foreground removal methods to recover the input cosmological angular power spectrum and measure the parameters. We show the effect of survey noise levels and beam sizes on the recovered the cosmological constraints. We find that while the reconstruction removes power from the cosmological 21 cm distribution on large scales, we can correct for this and recover the input parameters in the noise-free case. However, the effect of noise and beam size of the Tianlai pathfinder prevents accurate recovery of the cosmological parameters when using only intensity mapping information.

Key words: cosmology: theory – dark energy – large-scale structure of Universe.

1 INTRODUCTION

The distribution of matter on large scales has provided an important cosmological probe, allowing for measurement of the cosmological parameters by probing both the initial conditions that generated the seeds of structure, and also the physics that causes the structures to grow and develop. This data has come from large area extragalactic surveys, mainly targeting galaxies in the optical, which act as tracers of the underlying dark matter density fluctuations. While the first surveys were initially at very low-redshifts (e.g. CfA survey, Geller & Huchra 1989; 2df, Colless et al. 2001), their reach has increased as the technology has improved. Results from these surveys are independent from, but complimentary to, that of the more distant cosmic microwave background (CMB).

These surveys have an advantage over the CMB, as the density distribution of the Universe can be measured from galaxies in three dimensions, rather than the 2D surface from which the CMB photons are emitted. This property allows for an increase in the sampling, as a particular scale can be measured in more directions, leading to a decrease in sample variance (e.g. Sefusatti & Komatsu 2007). It also allows for types of measurements impossible in 2D, for example the radial components of the BAO giving a direct probe of the Hubble rate $H(z)$ (Blake & Glazebrook 2003; Hu & Haiman 2003; Seo & Eisenstein 2003). However, there are disadvantages as well, since the structures at late time will have undergone some non-linear evolution under gravity which requires

★ E-mail: Jacobo.Asorey@ciemat.es (JA); DavidParkinson@kasi.re.kr (DP) more complex modelling. There is also the greater observational time and effort required by galaxy redshift survey, as opposed to a simpler photometric survey (Blake & Bridle 2005).

Most of these surveys of the matter distribution have so far been in the optical and near-infrared, as the combination of technological lead and targets available in the this wavelength range has made these sources most accessible. However, these surveys will soon reach a natural limit, as the expansion of the Universe will redshift the spectra of these objects out of the observable range, leading to a 'redshift desert' in the range $1.4 < z < 3$. The next generation of radio observatories offers an alternative in this region, through the measurement of 21 cm emission of neutral atomic hydrogen (H I). This line emission should be observable at greater redshifts, as radio telescopes can span a much larger range of frequencies than the optical wavelength band. This signal can be used either in the same manner as traditional optical galaxy spectroscopy, to identify the redshift of individual galaxies (H I galaxy surveys) (Haynes et al. 2018), or by measuring the total H I intensity in a larger sky area (pixel), known as H I intensity mapping (Bharadwaj & Sethi 2001). Intensity mapping in particular can be carried out much faster than optical spectroscopic surveys, giving the opportunity to conduct full-sky sky surveys and so measure the matter distribution on the largest scales (Camera et al. 2013).

However, since the spontaneous hydrogen spin-flip transition is highly forbidden, with a long mean lifetime, and the intergalactic density of cold neutral hydrogen is low, the 21 cm signal is relatively weak ($T \sim 1$ mK). This weak signal can be overwhelmed by foreground radio emission at the same observed frequency ($T \sim 1$ K or greater at around 800 MHz). These foreground radio emissions include the thermal radiation from the ionosphere of the Earth, the Galactic Synchrotron, free–free emission from ionized regions both Galactic and extragalactic, and radio point sources contaminants. Separating the cosmological H I signal from the non-H I foreground will be difficult, and a number of different methods have been proposed. All of these methods assume a smooth, large-scale frequency dependence of the foregrounds, that can be modelled and removed for each pixel, leaving the small-scale fluctuations that correspond to the 21 cm density fluctuation. These include: fastICA which removes the foregrounds using independent component analysis in frequency space (Chapman et al. 2012; Wolz et al. 2014), PCA (principal component analysis), and ICA (independent component analysis) (Alonso et al. 2015), the Correlated Component Analysis (CCA) method (Brown & Bonaldi 2015), Karhunen-Loeve Decomposition (Shaw et al. 2014, 2015), Generalized Morphological Component Analysis (GMCA) (Chapman et al. 2013).

The most important questions are then how well the cosmological parameters can be measured by future data sets (precision), and how much bias can be introduced into the measurements by the foreground removal process (accuracy). In order to settle both of these questions, forecasts need to move beyond the assumption of Gaussianity, and simple power spectrum distributions of the fluctuations, and consider accurate sky simulations, based on large-scale *N*-body simulations. In this paper we address both questions, with particular emphasis on accuracy, showing how biases or offsets in the posterior probability distribution of the cosmological parameters (offsets relative to the input parameter values of the simulations) are introduced by the effect of removing the foreground contamination, and can be corrected for using simulations.

Another method to increase the accuracy of the recovered cosmological parameters from the 21 cm radio sky would be to cross-correlate the intensity map with a galaxy catalogue that covers the same area of the sky and redshift range. Since the optical galaxy sample would not have the same systematic errors and uncertainties introduced by the radio foreground removal process, the cross-power spectrum should be a more accurate representation of the underlying density field. This would also be straightforward to demonstrate with simulations, if the haloes can be populated by a galaxy distribution that matches the planned survey. In a forthcoming paper we will simulate this cross-correlation using the same prescription and analysis approach as we have here (Shi et al., in preparation).

For the HIR4 (H I with Horizon Run 4) project, we have created a full simulation and analysis pipeline. Some alternative approaches in the literature consider hydrodynamical simulations (Villaescusa-Navarro et al. 2018) or fast simulations based on lognormal density field realizations (Alonso, Ferreira & Santos 2014). However, in our case, we have started with the dark matter-only particle Horizon Run 4 N-body simulation, and populated the haloes with clouds of neutral hydrogen. When using *N*-Body simulations, there has to be a necessary compromise between detailed hydrodynamical simulations and fast simulations in terms of volume and resolution. But the mass limit in large volume *N*-body simulations does not allow us to access all the hydrogen, as some will be located in haloes with masses below the limit. This can be solved by scaling the simulated maps with observational measurements of the neutral hydrogen density, $\Omega_{\rm HI}$. Having simulated the cosmological signal, we applied foreground radio emission at the appropriate wavelengths based on the best available current data. In addition, we included the estimated instrumental noise and array beaming effect for the Tianlai pathfinder (Chen 2011; Das et al. 2018) in our simulations. We then masked the sky and applied reconstruction techniques to remove the foregrounds and reconstruct the cosmological signal. Finally, we measured the angular power spectra of the 21 cm temperature maps, and measured the cosmological parameters from these. Since we have complete control over every step of the process, any mis-match or offset between the final cosmological results and the initial cosmological parameters set in the simulation will then provide a test of the steps we have taken and the assumptions we have made in the analysis process.

In Section 2 we describe the methodology we use to generate our simulations and analyse our data, including details about the Horizon Run 4 simulations, generating the simulated the cosmological and foreground sky, and reconstructing the cosmological signal, and measuring the angular power spectra. In Section 3 we show the sky maps that we generate, the measured angular power spectra for different cases, and the cosmological constraints. We summarize the forecast precision that these measurements will have in terms of the linear bias and growth rate of structure, and also address the accuracy at which the different reconstruction methods recover the input cosmology. In Section 4 we summarize our findings, and make recommendations for future analysis of real data.

2 METHODOLOGY

2.1 Cosmology theory

The intensity of the 21 cm brightness temperature field T_b, as a function of spatial position ($\boldsymbol{x}$) and cosmological time t, can be considered as a perturbation relative to the homogeneous mean temperature $\bar{T}_b$ evaluated at time t, such that

$$\Delta T_b(\boldsymbol{x}, t) = T_b(\boldsymbol{x}, t) - \bar{T}_b(t). \tag{1}$$

If we assume that the statistics of the density of neutral hydrogen track the statistics of the overall matter density, then we can make the usual assumption that the two are related through some bias parameter b, and rearrange equation (1) to give the 21 cm field temperature in terms of the matter density perturbation δ,

$$T_b(\boldsymbol{x}, t) = \bar{T}_b(t)[1 + b(\boldsymbol{x}, t)\delta(\boldsymbol{x}, t)]. \quad (2)$$

If we now transform from a real space position and homogeneous time coordinate to a measured sky direction $\boldsymbol{n}$ and redshift z, we need to include the effects due to redshift-space distortions, giving

$$T_b(\boldsymbol{n}, z) = \bar{T}_b(z)\left[1 + b(\boldsymbol{x}, t)\delta(\boldsymbol{x}, t) + \frac{1+z}{H(z)} n^i \partial_i(\boldsymbol{n}\cdot\boldsymbol{v})\right]. \quad (3)$$

The clustering of fluctuations is described by the anisotropic power spectrum in Fourier space that can therefore be written (assuming the Kaiser formula for redshift-space distortions (Kaiser 1987) and linear perturbation theory) as

$$P_{21\mathrm{cm}}(k, \mu, z) = \bar{T}_b(z)^2 P_{\delta\delta}(k, z)\left[b_{\mathrm{HI}}(z) + f(z)\mu^2\right]^2, \quad (4)$$

where $\mu = \cos\theta_{\mathrm{los}}$ is the ratio between a given mode and the radial modes given by the line of sight. We describe the linear growth rate $f(z)$ theoretically by parametrizing it with a growth index γ (Linder & Cahn 2007),

$$f(z) = \Omega_m(z)^\gamma, \quad (5)$$

and as we are assuming a ΛCDM model with Einstein gravity, the growth index is given by $\gamma = 0.545$ (Peebles 1984; Lahav et al. 1991).

2.2 Generating sky maps

We have created maps of 21 cm emission for different configurations. We start from an N-body simulation and then we proceed to create neutral hydrogen mass catalogues, that are finally converted in brightness temperature maps. We also add foregrounds and receiver noise to the combination.

2.2.1 *Horizon Run 4 simulations*

Since we are interested in simulating wide-field 21 cm intensity mapping surveys, we start by using a halo catalogue from a very large volume N-body simulation as the initial framework for the neutral hydrogen. We have used the Horizon Run 4 (Kim et al. 2015) simulations to create our intensity mapping maps. It is an N-body simulation run on a box of $L_{\mathrm{box}} = 3150\,h^{-1}$ Mpc. From the dark matter particles in the light cone a halo catalogue was built using a Friend-of-friends algorithm. The minimum halo mass is 2 $\times\ 10^{11} h^{-1}\mathrm{M}_\odot$ which correspond to 25 dark matter particles in the original light cone. It is based on a flat ΛCDM cosmology with a matter density of $\Omega_m = 0.26$, a Hubble parameter at redshift zero of $H_0 = 72\,\mathrm{km s^{-1} Mpc^{-1}}$, and an amplitude of fluctuations on scales of 8 h^{-1}Mpc of $\sigma_8 = 1/1.26$.

The total number of haloes in the catalogue is 1654 566 127. In Fig. 1, we show the mass halo function for haloes in the redshift range sampled by Tianlai and the range we consider in this paper. The simulation also include the velocities of the haloes. Using this velocities, we can create a catalogue that includes the linear redshift distortions. This redshift z_v is given by the combination of the cosmological redshift and the peculiar velocity one (Li et al. 2016; Davis et al. 2019)

$$(1 + z_v) = (1 + z_{\mathrm{true}})(1 + v_z/c), \quad (6)$$

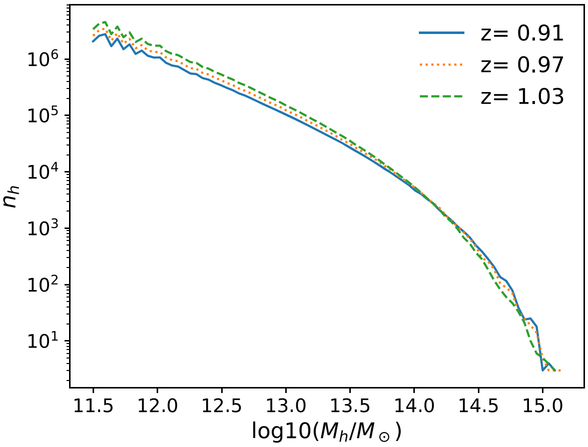

Figure 1. Distribution of halo masses for haloes selected in the expected redshift range for the Tianlai pathfinder. The sharp cut-off for the small masses is due to the resolution limit of the HR4 simulation.

where z_{true} is the true redshift given by the Hubble–Lemaître flow and z_v is the redshift that includes the effect of the peculiar velocity in the radial direction v_z.

2.2.2 *Neutral Hydrogen mass modelling*

As we are focusing in the post-reionization Universe, we can assume almost all the neutral hydrogen content around redshift $z = 1$ is inside dark matter haloes (Villaescusa-Navarro et al. 2018; Spinelli et al. 2020). We assign masses of neutral hydrogen to dark matter haloes based on the halo mass M_{h} and virial velocity v_c, following the halo model for neutral hydrogen developed in successive improvements in Barnes & Haehnelt (2015), Padmanabhan, Choudhury & Refregier (2016), Padmanabhan & Refregier (2017). There are other approaches (e.g. Modi et al. (2019)), but since we expect our signal to be dominated by the most massive neutral hydrogen haloes, we find this prescription to be sufficient for our use.

In Padmanabhan & Refregier (2017) the mass of neutral hydrogen hosted in a dark matter halo of mass M_{h} is given by:

$$M_{\mathrm{HI}}(M_{\mathrm{h}}) = f_{\mathrm{HI}} f_{H,c} M_{\mathrm{h}} \left(\frac{M_{\mathrm{h}}}{10^{11} h^{-1} M_\odot}\right)^\beta \times \exp\left[-\left(\frac{v_c^{\mathrm{min}}}{v_c(M_{\mathrm{h}})}\right)^3\right] \exp\left[-\left(\frac{v_c(M_{\mathrm{h}})}{v_c^{\mathrm{max}}}\right)^3\right], \quad (7)$$

where $f_{\mathrm{H\,I}}$ is a multiplicative constant that corresponds to the amount of neutral Hydrogen with respect to the fraction of cosmic Hydrogen, $f_{H,c} = (1 - Y_p)\Omega_{b0}/\Omega_{m0}$ where $Y_p = 0.24$ is the primordial Helium abundance. The model includes a logarithmic slope β and two velocity cut-offs v_c^{min} and v_c^{max}. The reason for the velocity cut-offs is that low-mass haloes are not capable of keeping the neutral hydrogen while massive haloes heat the gas and it stops being neutral. The values used for our simulation, which are partially based on Padmanabhan & Refregier (2017) best fit to data, are $f_{\mathrm{HI}} = 0.17$, $\beta = -0.55$, $v_c^{\mathrm{min}} = 30\,\mathrm{km\,s^{-1}}$ and $v_c^{\mathrm{max}} = 200\,\mathrm{km\,s^{-1}}$. We have used different values for the cut-off velocity parameters in order to find similar values of the hydrogen bias, as given in observations. In order to define the virial velocity, we have used the

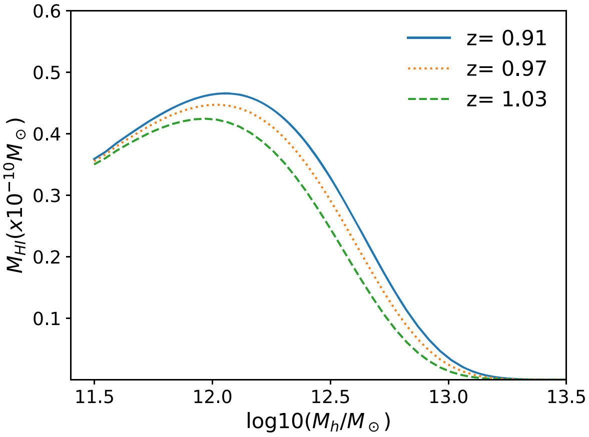

Figure 2. Distribution of hydrogen masses following equation (7) at three redshifts within the Tianlai redshift range and the HR4 simulation.

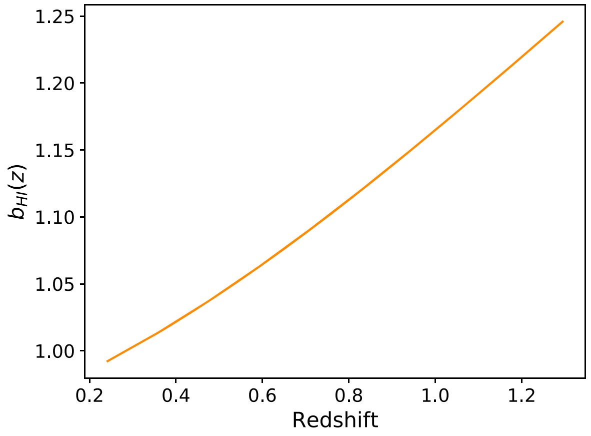

Figure 3. Bias of neutral hydrogen for the haloes in our sample for the the full redshift range of the Horizon Run simulation. The bias have been measured using equation (11) and we measured the halo mass function and the neutral hydrogen mass distribution from the simulation.

spherical collapse model as in

$$v_c(M_\mathrm{h}) = \sqrt{\frac{G_N M_\mathrm{h}}{R_c}}, \tag{8}$$

where $R_c(M_\mathrm{h})$ is the virial radius. Following Padmanabhan et al. (2016), we determine the virial radius by

$$R_c(M_\mathrm{h}) = 46.1\mathrm{kpc}\left(\frac{\Delta_v \Omega_m h^2}{24.4}\right)^{-1/3}\left(\frac{1+z}{3.3}\right)^{-1}\left(\frac{M_\mathrm{h}}{10^{11}M_\odot}\right)^{1/3}. \tag{9}$$

We use Δ_v given by the solution to a spherical top hat perturbation collapse for a virialized halo for the flat ΛCDM Universe, $\Omega_k = 0$ (Peebles 1980; Eke, Cole & Frenk 1996; Bryan & Norman 1998) where

$$\Delta_v = 18\pi^2 + 82x - 39x^2, \tag{10}$$

and $x = \Omega_m(z) - 1$. We show in Fig. 2, the average neutral hydrogen mass for a given halo mass at three different redshifts. We can observe that the mass distribution decreases with the massive haloes cut-off while we do not reach the cutt-off in the least massive haloes as the resolution of the Horizon Run 4 is not enough to reach the quenching scale.

We can define the bias of neutral hydrogen, $b_\mathrm{HI}(z)$ as

$$b_\mathrm{HI}(z) = \frac{\int \mathrm{d}Mn(M,z)M_\mathrm{HI}(M,z)b(M,z)}{\int \mathrm{d}Mn(M,z)M_\mathrm{HI}(M,z)}, \tag{11}$$

while the neutral hydrogen density parameter is:

$$\Omega_\mathrm{HI}(z) = \frac{\rho_\mathrm{HI}}{\rho_{c,0}} = \frac{1}{\rho_{c,0}}\int_0^\infty \mathrm{d}Mn(M,z)M_\mathrm{HI}(M,z). \tag{12}$$

A more in-depth discussion of the neutral hydrogen bias, using hydrodynamical simulations, can be found in Ando et al. (2019), Wang et al. (2019). In Fig. 3 we show the bias estimated using equation (11). The values are consistent with previous studies in the literature (Marín et al. 2010). We use this parameter as a benchmark parameter to test our simulated catalogues.

2.2.3 *Brightness temperature maps*

Once we have assigned hydrogen masses to the haloes in our simulation, we can continue to the next step, the creation of intensity maps. In order to do so, we define a redshift bins configuration and a given pixelization resolution. Then, we stack all the hydrogen masses for all the haloes in each cube defined by an angular pixel and a redshift bin. The corresponding mass M_HI is what we use to create the temperature maps. Following Battye et al. (2013), Bull et al. (2015) we define the 21 cm brightness temperature as:

$$T_b(\hat{\mathbf{n}}, z) = \frac{3h_\mathrm{Pl}c^3A_{12}}{32\pi k_b m_h \nu_{21}^2}\frac{(1+z)^2}{H(z)}\rho_\mathrm{HI}(\hat{\mathbf{n}}, z), \tag{13}$$

where k_b is the Boltzmann constant, h_Pl is the Planck constant, m_h is the mass of the neutral hydrogen atom, A_{12} is the quantum efficiency, c is the speed of light, and ρ_HI is the density of neutral hydrogen in the volume given by the frequency and area, $d\nu$ and $d\Omega$, which correspond the frequency (redshift) and pixel bins, respectively.

One caveat of our method is that we cannot access all the halo masses that host neutral hydrogen as the HR4 simulation has a lower limit for the mass of the haloes. As we do not have access to all the halo masses, we do not completely sample the full ρ_HI in a given volume cell. This lack of mass will produce a smaller brightness amplitude T_b than the expected one in nature. As the 21 cm cosmological signal has a low amplitude, reconstruction from a foreground dominated map becomes more difficult as the amplitude of the cosmological signal from neutral hydrogen becomes smaller. Therefore, we need to take into account the shortfall of neutral hydrogen in the simulation by scaling the average brightness temperature according to observations.

From equation (12), we see that ρ_HI is proportional to the density parameter of neutral hydrogen, Ω_HI. We can use measurements of this parameter in order to calibrate the mean temperature of our 21 cm maps. We have decided to follow the definition given in Square Kilometre Array Cosmology Science Working Group (2018), Cunnington et al. (2019). The approach is based on a polynomial fit to the Ω_HI data compiled in Crighton et al. (2015).

Both analysis (Square Kilometre Array Cosmology Science Working Group 2018; Cunnington et al. 2019) define this fit as:

$$\Omega_\mathrm{HI}(z) = 0.00048 + 0.00039z - 0.000065z^2. \tag{14}$$

We can compare this approach with other models in the literature. Using the same data compilation Crighton et al. (2015) but a different model for the redshift dependence given by $\Omega_\mathrm{HI} \propto (1$

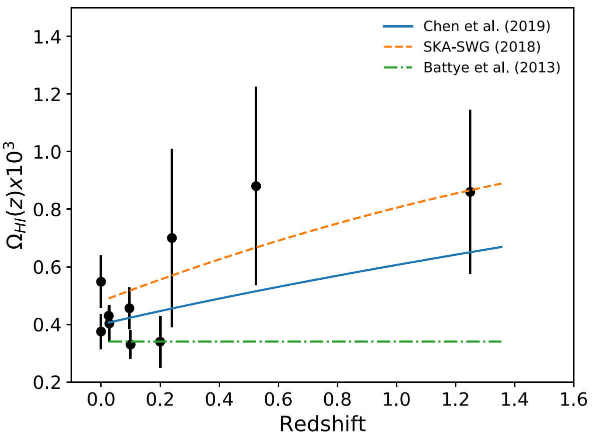

Figure 4. Comparison between theoretical descriptions of $\Omega_{\rm HI}$ given by lines and measurements compiled in Crighton et al. (2015) (black circles).

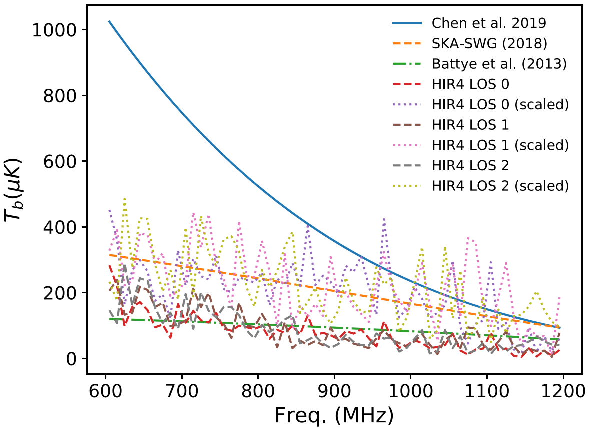

Figure 5. Neutral hydrogen brightness temperature evolution with frequency. We show three different theoretical models in the literature for the mean temperature and the evolution in redshift of three different lines of sight, from the mock temperature maps for both the temperature given by the hydrogen mass from the HR4 haloes and the one rescaled following equation (15).

$+ z)^{0.6}$. In Battye et al. (2013), the density parameter is assumed constant and fitted to low redshift data where $\Omega_{\rm HI} = 2.45e - 4$. We can see the comparison in Fig. 4.

Then, we re-scale the brightness temperature $\bar{T}_b^{\rm old}$ given by equation (13) in each map and pixel

$$T_b^{\rm new}(\hat{\mathbf{n}}, z) = \frac{T_b^{\rm old}(\hat{\mathbf{n}}, z)}{\bar{T}_b^{\rm old}} \bar{T}_b^{\rm new}, \qquad (15)$$

so that the new mean temperature, $\bar{T}_b^{\rm new}$ is given by:

$$\bar{T}_b^{\rm new}(z) = 180 h_0 \frac{(1+z)^2}{E(z)} \Omega_{\rm HI}(z) mK, \qquad (16)$$

where $\Omega_{\rm HI}(z)$ is given by equation (14).

We can see in Fig. 5 the effect of this rescaling in the mean temperature and in the temperature fluctuations $\Delta T_b = T_b(x) - \bar{T}_b$. The change in the fluctuations is needed in order to recover the right amplitude of the power spectra and the input bias from the simulation from the pure 21 cm signal. We also compare our chosen approach with other definitions of brightness temperature such as Battye et al. (2013) where $\Omega_{\rm HI}$ was estimated assuming a constant mean value and only low redshift data, which tends to indicate a lower $\Omega_{\rm HI}$ or Chen et al. (2019) which also consider the data compilation from Crighton et al. (2015) but assumes a different model based on a power law of $(1 + z)$ to define the redshift evolution of $\Omega_{\rm HI}$ therefore deviating from the polynomial model that we consider in equation (14) and that is defined in Square Kilometre Array Cosmology Science Working Group (2018).

2.2.4 *Foreground maps*

Radio measurements are dominated by foreground radio emissions. In order to study the influence of foregrounds on the measured signal with our radio telescopes, we need to add or model them in our simulated catalogues. In order to do so, we have created a suite of foreground maps for each frequency bin considered in our mock catalogues. In particular, we use the updated Global Sky Model (GSM) (de Oliveira-Costa et al. 2008; Zheng et al. 2017) to generate foreground maps as this method produces a good approximation to the Galactic Diffuse emission.

GSM minimizes the cost function given by a matrix decomposition and the data of 29 frequency smoothed maps with frequencies between 10 MHz and 5 THz using an iterative algorithm in which the initial guess is made using a PCA decomposition of six components of the data matrix. The GSM model maps include mostly information from five different physical mechanisms: synchrotron, free–free, CMB, warm dust, and cold dust. Using the first six components of the PCA decomposition, we can produce a foreground temperature map, $T_b^{\rm foreground}(\hat{n})$, at a any given frequency within the range of the algorithm.

2.2.5 *Masking*

We can create full-sky simulations for the cosmological 21 cm signal. But much of the emission from Galactic Foregrounds is coming from close to the Galactic Centre and the Galactic Plane. Therefore, the first step to remove the signal from foregrounds is to mask the highest intensity emission from the Milky Way. To do so, we have considered a simple procedure in which we apply a brightness temperature cut of $T_{b,\,\rm mask} = 8$ K, and so every pixel with $T_b > T_{b,\,\rm mask}$ is removed from the analysis.

We show in Fig. 6 the masks that we use regarding the Galactic Emission. On the left we show in magenta the area removed from the hydrogen maps in order to estimate the observed angular power spectra, assuming that all areas of the sky are accessible, and in yellow the area that is used. On the right-hand panel, we show the mask used for the Tianlai survey, with the same colour scheme. The yellow area is now also restricted to the footprint of the Tianlai survey, since in considering this survey, we need to apply a declination mask as Tianlai cannot access the southern ecliptic hemisphere. Therefore, we only include the region for declinations above $\delta > -40$. We consider this mask only when using the noise maps regarding Tianlai survey that we describe in Section 2.3.

2.3 Instrumental effects and Tianlai cylinder array

The last ingredient that we consider in our maps is the instrumental noise. While the 21 cm cosmological signal and the foreground signal are produced by astrophysical processes, the telescopes that measure this signal have also an intrinsic thermal noise plus the extraterrestrial signal is convolved with the instrument beaming.

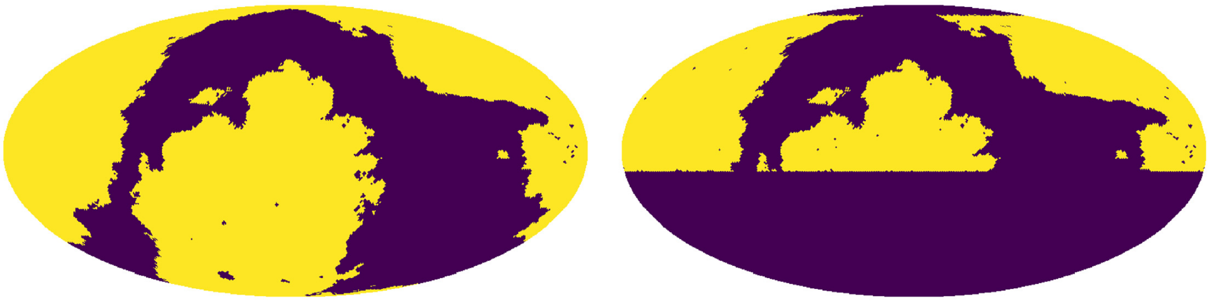

Figure 6. On the left, we show the masked region in blue when removing the galactic centre to reduce the foreground signal in the maps. On the right, we show the mask used when considering the Tianlai pathfinder survey noise maps as this mask includes the information of the Tianlai footprint.

2.3.1 Map making for transit radio telescopes

Unlike a traditional radio interferometer that usually observes a small patch of sky and exploit the Fourier transform mapping between the sky and the *uv*-plane by assuming the flat-sky approximation, the Tianlai cylinder array is wide-field transit interferometers. The observable, *visibility*, measures different parts of the curved sky as a function of time (i.e. the azimuthal angle ϕ). At any instant, a visibility for any particular frequency is given by

$$V_{ij}(\phi) = \int d^2\hat{\mathbf{n}} B_{ij}(\hat{\mathbf{n}};\phi) T(\hat{\mathbf{n}}) + n_{ij}(\phi), \qquad (17)$$

where $n_{ij}(\phi)$ represents the noise term and the beam transfer function $B_{ij}(\hat{\mathbf{n}};\phi)$ is given by

$$B_{ij}(\hat{\mathbf{n}};\phi) = A_i(\hat{\mathbf{n}}) A_j^*(\hat{\mathbf{n}}) e^{2\pi i \hat{\mathbf{n}}\cdot\mathbf{u}_{ij}}. \qquad (18)$$

Here, the visibility is measured by correlating the signals from a pair of feeds i and j, located at positions r_i and r_j with $\mathbf{u}_{ij} \equiv (r_i - r_j)/\lambda$, where λ is the observed wavelength, $\hat{\mathbf{n}}$ is the sky direction, and $A_i(\hat{\mathbf{n}})$ denotes the primary beam of feed i.

Recently, a novel 'm-mode' formalism for the analysis of transit radio telescopes was proposed by Shaw et al. (2014). It provides an easy way to linearly map wide-field interferometric data on the full sky. The measured visibilities can be written as a summation of spherical harmonic modes. By taking into account the fact that the measured visibilities change periodically with the sidereal day (i.e. the periodicity in ϕ), one can find a simple relation between the so-called m-mode visibilities V_m^{ij} and the sky by

$$V_m^{ij} = \sum_{\ell} B_{\ell m}^{ij} a_{\ell m} + n_m^{ij}. \qquad (19)$$

Here, $a_{\ell m}$ and $B_{\ell m}^{ij}$ denote the coefficients in the spherical harmonic expansions of the sky $T(\hat{\mathbf{n}})$ and the beam transfer function $B_{\ell m}^{ij}(\phi)$, respectively, which read

$$T(\hat{\mathbf{n}}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\hat{\mathbf{n}}), \qquad (20)$$

$$B_{ij}(\hat{\mathbf{n}};\phi) = \sum_{\ell m} B_{\ell m}^{ij}(\phi) Y_{\ell m}^*(\hat{\mathbf{n}}). \qquad (21)$$

Following Shaw et al. (2014) and Zhang et al. (2016), one can then group m-mode visibilities from all baselines of the array together into a vector **v**, and similarly group m-mode harmonic coefficients of the sky and m-mode noises for all baselines into vectors of **a** and **n**, respectively. The measurement equation in equation (19) for each m-mode thus can be further simply rewritten in matrix form as

$$\mathbf{v} = \mathbf{B}\mathbf{a} + \mathbf{n}, \qquad (22)$$

where we have expressed the beam transfer matrices in an explicit matrix notation **B**. This equation is valid for any particular m and frequency ν.

Using the maximum likelihood method, the best estimate of the sky spherical harmonics coefficients from a given set of sampled visibilities for each individual m and frequency ν is solved by

$$\hat{\mathbf{a}} = (\mathbf{B}^\dagger \mathbf{N}^{-1} \mathbf{B})^{+} \mathbf{B}^\dagger \mathbf{N}^{-1} \mathbf{v}, \qquad (23)$$

where the superscript + represents the pseudo-inverse. Here, we assume that the instrumental noises at two different frequencies are uncorrelated and the noise follows a complex Gaussian distribution with zero mean and covariance $<\mathbf{n}\mathbf{n}^\dagger> = \mathbf{N}$.

2.3.2 Configuration of the Tianlai cylinder array

The Tianlai Pathfinder presently consists of an array of three adjacent cylinder telescopes, located in Hongliuxia, a radio-quiet site in north-west China (44°9′9.66″ N 91°48′24.72″ E). Each of the cylinders is 15 m wide and 40 m long. With wide field of view radio interferometers, the Tianlai Pathfinder is dedicated to measure the 3D maps of neutral hydrogen (the so-called 21 cm intensity mapping) of the northern sky in the Universe by surveying neutral hydrogen over large areas of the sky at low redshifts in the range of $1.03 > z > 0.78$ (700–800 MHz). Currently, the three cylindrical reflectors oriented in the North–South direction, each having 33, 32, and 31 feed antennas, respectively (see Fig. 7).

For the Tianlai cylinder array, by assuming uncorrelated thermal noises across all baselines and frequencies, the noise level (RMS) in units of brightness temperature is given by (Thompson, Moran & Swenson 2001):

$$\sigma_{ij}^N = \left(\mathbf{N}_m^{ij}\right)^{1/2} = \frac{T_{\rm sys}}{\sqrt{\Delta\nu \Delta t_{ij}}} \left(\frac{\lambda^2}{A_e}\right), \qquad (24)$$

where Δt_{ij} is the total integration time of baseline *ij*, $T_{\rm sys}$ is the system temperature, A_e is the effective area of antenna, λ is the observing frequency, and $\Delta\nu$ is the width of the frequency channel. The system temperature is the sum of the sky brightness and the analogue receiver noise temperature, $T_{\rm sys} = T_{\rm sky} + T_{\rm rec}$. At the frequency of interest (700–800 MHz), the Tianlai array would be expected to achieve a total system temperature of 50–100 K, and thus we assume $T_{\rm sys} = 50$ K in this study. We also assume two full years of observation for the Tianlai pathfinder survey. The effective antenna area A_e is calculated by $A_e\Omega = \lambda^2$, where the beam solid angle Ω is well approximated by $\Omega \simeq 0.1$ for the current Tianlai cylinder array.

By realistically simulating the noise visibilities for the Tianlai instrumental configuration, the noise sky maps, $T_b^{\rm noise}(\hat{n})$, for all

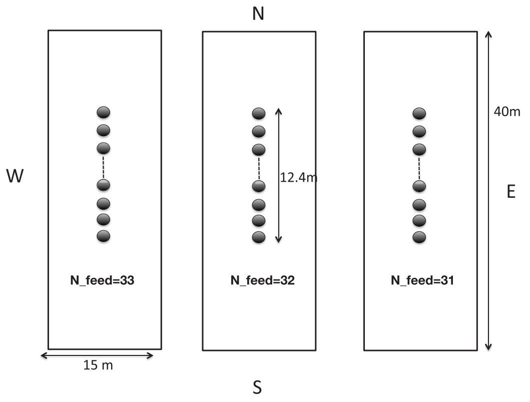

Figure 7. Configuration of the Tianlai pathfinder cylinder array. At present, the array has three adjacent cylindrical reflectors oriented in the North–South direction, each with 15 m wide and 40 m long. The three cylinders are equipped with a total of 96 dual polarization receivers, which are irregularly spaced on the cylinders. The number of feeds on each cylinder is 31, 32, and 33, respectively, spanning the same distance of 12.4 m along the North–South direction on each cylinder. The feed spacing is thus 0.388, 0.4, and 0.413 m, respectively.

frequencies from these visibilities are reconstructed by the above map-making process with using the maximum-likelihood solution in equation (23).

In the main analysis of this paper, we only consider Tianlai-like noise when building the HIR4 maps, but for reference we show here a comparison with a case based on SKA-MID phase 1 noise. We can estimate the beam of the SKA survey by only considering the baseline of the telescope. The intensity mapping beam size is $\Sigma_{\rm beam} = 1.133\theta^2$. As defined in Cunnington et al. (2019), the beam resolution is giving by

$$\theta_{\rm beam} = \frac{1.22c}{\nu D_{\rm base}}, \qquad (25)$$

where $D_{\rm base} = 15$ m for the SKA configuration.

We can also simulate the effect of a Gaussian beam with the pure 21 cm maps. In order to do so, we have smoothed the 21 cm intensity mapping field in each frequency beam with a Gaussian filter with a smoothing scale given by equation (25) (Cunnington et al. 2019).

2.4 Observed temperature map

We define the observed temperature map as the one that combines the cosmological signal from the simulation, the foregrounds and the observational noise. We define the observed temperature as:

$$T_b^{\rm obs}(\hat{n}) = T_b^{\rm HI}(\hat{n}) + T_b^{\rm foreground}(\hat{n}) + T_b^{\rm noise}(\hat{n}), \qquad (26)$$

where $T_b^{\rm HI}(\hat{n})$ is the brightness temperature of cosmological neutral hydrogen at pixel given by position $(\hat{n})$ given by equation (13) and with the mean temperature rescaled as given by equation (15). Then we add the brightness temperature, $T_b^{\rm foreground}(\hat{n})$, from the same pixel in the foreground map, created using the GSM, at the mean frequency of the corresponding frequency bin, following prescription in Section 2.2.4. Finally, we can add the value of the noise temperature at that angular position given by the Tianlai noise maps explained in Section 2.3.

In Fig. 8, we show how the beaming affects the expected signal when we use all the information in the frequency bin 790–800 MHz. The SKA noise assume 10 000 h of integration, leading to a noise variance of $\sigma = 4.4 \times 10^{-5}$ mK, whereas the expected average noise for Tianlai given one year of integration is $\sigma = 2.6 \times 10^{-4}$ mK. Note that in this paper we generate full-sky simulated noise maps for Tianlai, assuming two full years of observations, to use in the analysis pipeline, but show that this assumption of one year on the sky gives a similar noise value to the simulated maps at large angular scales. We can see that the effect of the noise on the power spectrum for SKA-MID phase 1 is much smaller than the effect of a Gaussian beam, given by 25, over the range of multipole values that would be considered for cosmological analysis.

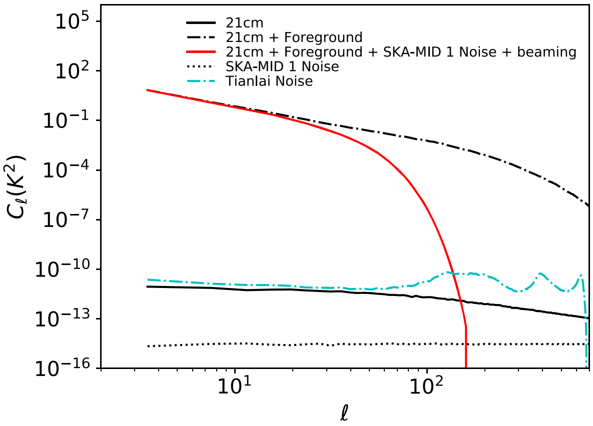

Figure 8. Example of angular power spectrum for the cosmological 21 cm temperature maps (black solid), maps with foreground signal (black dot-dashed), maps with foregrounds and the expected dish beam and noise maps for SKA-MID phase 1 receivers (red solid). We also show the power spectra of the noise alone for Tianlai (cyan dot-dashed) and SKA-MID phase 1 (black dotted). The frequency band in this case is 790–800 MHz. The SKA-MID phase 1 noise assume 10 000 h of integration, leading to a noise variance of $\sigma = 4.4 \times 10^{-5}$ mK, whereas the expected average noise for Tianlai is $\sigma = 2.6 \times 10^{-4}$ mK.

2.5 Analysis methods

2.5.1 *Reconstruction methods*

The goal of any reconstruction method is to decompose the map into a set of signals with some different qualities or attributes. In the case of 21cm, we use reconstruction methods to split the map into the foreground part (generated locally to our Galaxy) and the cosmological part, based on the assumption that the frequency dependence of the two will be very different. In this section we describe three such methods, fast independent component analysis (fastICA), PCA, and log-polynomial fitting.

fastICA: The fastICA method assumes that the maps can decomposed into a set of signals with some non-Gaussian distribution (the foregrounds) and some Gaussian noise (the cosmological signal). This is given by

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}. \qquad (27)$$

Here $\mathbf{x}$ is the final map, which will be split up by pixel and frequency bin, $\mathbf{s}$ is vector of components, with an amplitude for each component that depends on the position in the map, $\mathbf{A}$ is the mixing matrix that defines how the components evolve with frequency, and $\mathbf{n}$ is the distribution of Gaussian noise. Note that the data is 'pre-whitened', such that the mean in each frequency bin is removed, and only replaced during the reconstruction phase. This means that the mean temperature of each bin, which is a sum of the mean of the

21 cm cosmological signal and the mean of the foreground emission at that frequency, cannot be reconstructed by this process, and only the distribution of fluctuations can be separated.

In our implementation, the non-Gaussian components should correspond to the foregrounds, which should be well behaved and continuous in frequency space. In contrast, the intensity of the cosmological 21 cm emission depends on the mass of neutral hydrogen present in the ‘voxel’, which is a stochastic quantity with a Gaussian distribution, and so resembles the noise in a fastICA reconstruction process. As such, the reconstructed map will be given by

$$\mathbf{x}_{\rm recon} = \mathbf{x}_{\rm data} - \mathbf{A}\mathbf{s}, \tag{28}$$

where $\mathbf{x}_{\rm recon}$ is the reconstructed map, ordered by pixel and frequency bin, which includes both the cosmological 21 cm distribution and the Gaussian noise from the instrument.

The method used to estimate $\mathbf{s}$ the vector of components, and $\mathbf{A}$ the mixing matrix, is very similar to that described in Chapman et al. (2012), Wolz et al. (2014) and other, previous works. Using the implementation of fastICA as part of the `SCIKIT-LEARN` python machine learning package (Pedregosa et al. 2011) we maximize the negentropy, defined by $J(y) = H(y_{\rm gauss}) - H(y)$, assuming the negentropy is approximated by a $\log\cosh(y)$ function. The negenetropy functions as a measure of distance from Gaussianity, and so maximizing it with respect to the components should remove the foreground signal, leaving behind the Gaussian cosmological signal.

PCA: For the PCA method, we again work with ‘pre-whitened’ data, where the mean of the simulated data has been subtracted. From this we compute the data covariance matrix, looking at the covariance between different bins in frequency space. We then compute the eigenvectors and eigenvalues of this frequency–frequency covariance matrix. Since the foregrounds dominate the power in maps at all frequencies, they will dominate the eigenmodes with the highest eigenvalues. Also, the foregrounds are expected to have smooth frequency structure, so that they could be described by just a few smooth frequency eigenvectors. With these reasons, finally we project out the principal components with the largest eigenvalues in frequency space from every spatial pixels to obtain foreground cleaned maps. However, correlations in frequency space can also be slightly generated cosmologically, and so again this foreground removal approach may also have the effect of removing the 21 cm signal from the power spectrum over all scales.

Specifically, following de Oliveira-Costa et al. (2008), we reshape the 3D observed data into an $N_\nu \times N_\theta$ matrix $\mathbf{x}$, where N_θ contains all 2D spatial pixels (the same as in the fastICA approach). The empirical $\nu - \nu$ covariance of the data is

$$\mathbf{C} = \frac{\mathbf{x}\mathbf{x}^T}{N_\theta}. \tag{29}$$

By using the PCA analysis, the matrix $\mathbf{C}$ can be decomposed into $\mathbf{C} = \mathbf{U}\boldsymbol{\Lambda}\,\mathbf{U}^T$, where $\boldsymbol{\Lambda}$ is diagonal and contains the eigenvalues in descending order and $\mathbf{U}$ is an orthogonal matrix whose columns are the eigenvectors (i.e. the principal components). Now we define the deprojection matrix, $\boldsymbol{\Pi} = \mathbf{I} - \mathbf{U}\mathbf{S}\mathbf{U}^T$. Here $\mathbf{I}$ is the identity matrix and $\mathbf{S}$ is a selection matrix with 1 along the diagonal for modes to be removed and 0 elsewhere. We can then apply $\boldsymbol{\Pi}$ to our map along each line of sight

$$\mathbf{x}_{\rm recon} = \mathbf{x}_{\rm data}\left[\mathbf{I} - \mathbf{U}\mathbf{S}\mathbf{U}^T\right], \tag{30}$$

in order to project out the selected principal components which are significantly dominated by foregrounds.

log-polynomial fitting: For the log-polynomial fitting, we do not use the pre-whitened field but take the raw combined map $\mathbf{x}$ and try the linear least-square fitting with an n-th order polynomial,

$$\log T(\hat{n},\ \nu) = \sum_{j=0}^{n} s_j(\hat{n})\,(\log \nu)^j\,, \tag{31}$$

at every direction $\hat{n}$. We do not consider the noise covariance matrix at this point, such that the fitting becomes equivalent to

$$\mathbf{y} = \mathbf{A}\mathbf{s}, \tag{32}$$

at every direction $\hat{n}$, with $\mathbf{y} \equiv \{\log T(\hat{n},\ \nu_i)\}$, $A_{ij} \equiv (\log \nu_i)^j$ and $\mathbf{s} \equiv \{s_j(\hat{n})\}$. The best-fitting parameter set is then given by the estimator

$$\hat{\mathbf{s}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}, \tag{33}$$

where the superscript T denotes the transpose. The reconstructed map is then given by $\mathbf{x}_{\rm recon} = \mathbf{x}_{\rm data} - \mathbf{A}\hat{\mathbf{s}}$ just as in equation (28). This is obviously not the best practice that considers the property of the noise, but is one that just relies on the smoothness of the foreground. One can of course follow the procedure by de Oliveira-Costa et al. (2008) with the noise covariance matrix $\mathbf{N} \equiv \langle \mathbf{n}\mathbf{n}^T\rangle$ for a better estimator,

$$\hat{s} = (\mathbf{A}^T\mathbf{N}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{N}^{-1}\mathbf{y}, \tag{34}$$

once $\mathbf{N}$ properly reflects the instrumental noise and the cosmic signal. Because the log-polynomial fitting is already found to be out-performed by PCA in a wide range of frequency (de Oliveira-Costa et al. 2008), we simply use equation (31–33) in this work to illustrate its relative power.

2.5.2 *Angular power spectrum*

In order to study the cosmological information encoded in our maps, we decompose the distribution of intensity in a certain basis set (in this case spherical Bessel functions). If the continuous intensity field in a particular direction $T(\boldsymbol{\theta})$ is Gaussian and randomly distributed, then it can be decomposed into its multiple moment using spherical harmonics $Y_{\ell m}$

$$a_{\ell m} = \int {\rm d}\boldsymbol{\theta}\, Y^*_{\ell m}\, T(\boldsymbol{\theta}). \tag{35}$$

Assuming an isotropic universe, we get the power spectrum from the autocorrelation function,

$$\left\langle a^*_{\ell m} a_{\ell' m'}\right\rangle = \delta_{\ell\ell'}\delta_{mm'}C_\ell. \tag{36}$$

Since the spherical Bessel functions are dimensionless, the spherical harmonic coefficients $a_{\ell m}$ must have units of intensity or temperature per unit area, and the power spectrum $C_\ell \propto T^2$.

To measure the angular power spectrum we used the `NaMaster`[1] code (Alonso, Sanchez & Slosar 2019), which uses the pseudo-Cl (aka MASTER) approach, including the effect of the sky mask.

2.5.3 *Covariance matrix*

Finally, we need to include the measurement errors on the angular power spectra in order to constrain the cosmological parameters. As we are focusing on the linear scales in this paper, we assume that the density field is linear and described by a Gaussian distribution, in order to define the covariance matrix. When considering a full

[1] Downloaded from https://github.com/LSSTDESC/NaMaster.

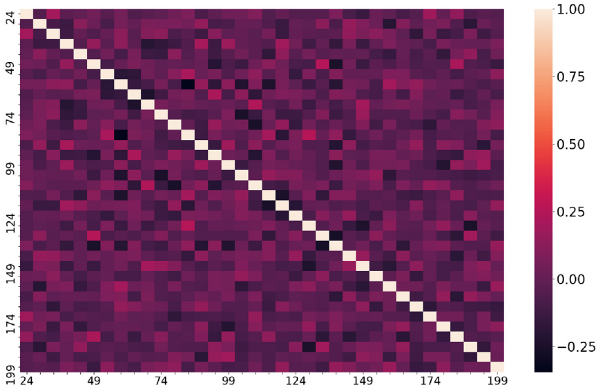

Figure 9. Correlation matrix $r = C_{ij}/(\sqrt{C_{ii}C_{jj}}$ for the different ℓ-values considered in our cosmological analysis for the $f = 790$–800 MHz frequency bin.

sky map, it might be possible to assume that the different modes are not correlated, giving the standard relation that

$$\mathrm{Cov}(C_\ell, C_{\ell'}) = \mathrm{Var}(C_\ell)\delta_{\ell,\ell'}, \tag{37}$$

where the cosmic variance component of the angular power spectra variance is given by

$$\mathrm{Var}(C_\ell) = \frac{2C_\ell^2}{N(\ell)}, \tag{38}$$

where $N(\ell) = (2\ell + 1)\Delta\ell f_{\mathrm{sky}}$ is the number of modes for each multipole ℓ (Cabré et al. 2007; Crocce, Cabré & Gaztañaga 2011; Asorey et al. 2012). The variance depends therefore on the fraction of sky used for the cosmological analysis, $f_{\mathrm{sky}} = A_{\mathrm{survey}}/A_{\mathrm{sphere}}$, and on the amplitude of the multipole bins $\Delta\ell$ used to measure the angular power spectra. As we include the effect of the angular mask by drawing theoretical realizations of the density field, we assume $f_{\mathrm{sky}} = 1$ when using equation (38) to get the first estimates of the cosmological parameters before re-doing the analysis with a more realistic covariance matrix.

However, we will be removing part of the sky to reduce the effect of foregrounds, and so mode–mode coupling in the cut sky can lead to non-zero off-diagonal elements in the covariance matrix (e.g. Brown, Castro & Taylor 2005). To estimate the full covariance matrix for the intensity map, including cut-sky effects, we would need a large number of simulated skies, each a different realization of an n-body simulation. Horizon Run 4 is only a single realization, and we do not have multiple simulations of equal volume and resolution available. To get around this problem, we create a large number of Gaussian simulated skies, $n_{\mathrm{realizations}} = 100$, with the same measured power spectrum as the HR4 intensity map. We then used NaMaster to measure the power spectrum and cross-correlations between ℓ-values across the ensemble of Gaussian realizations to estimate the co-variance. We show an example of the correlation matrix in Fig. 9 of the $f = 790$–800 MHz frequency bin angular power spectrum.

2.5.4 *Cosmological parameter estimation*

Once we have the measured angular power spectrum and estimated the covariance matrix, we can use this to estimate values of the cosmological parameters. The results from this parameter fitting can be used to indicate if there is any bias/offset that has been introduced by the reconstruction methods. In this case we choose the amplitude scaling parameters, the linear bias b, the growth rate f and the background 21 cm temperature $\bar{T}_{21}$. We fix the other cosmological parameters to the values specified in Section 2.2.1.

To speed up the analysis, we decompose the angular power spectrum into the term that depends only on the bias, the term that depends only on the RSD effect, and the cross-term between the two. Following the approach of Asorey et al. (2012), Asorey, Crocce & Gaztañaga (2014), we can reconstruct any angular power spectrum in the parameter space of $\{b, f, \bar{T}_b\}$ through the following equation

$$C_\ell = \bar{T}_{21}^2\left[b^2 C_\ell^{\mathrm{no\ rsd}} + f^2 C_\ell^{\mathrm{no\ bias}} + 2bf C_\ell^{\mathrm{cross}}\right]. \tag{39}$$

This approach assumes no growth or bias evolution through the redshift bin, which is justified given the very thin redshift slicing that can be performed. It also fixes the overall amplitude of the cosmological density field σ_8.

We see from equation (39) that the background 21 cm temperature parameter is completely degenerate with a combination of the growth and bias parameters. Of these three parameters then, only two can be independently measured by the 21 cm autocorrelation angular power spectrum. (This degeneracy can be broken with cross-correlations with other tracers, but we leave that discussion for the future.) We fix $\bar{T}_b$, and measure the combinations $\bar{T}_b b$ and $\bar{T}_b f$.

We use Bayes' theorem to estimate the posterior distribution of the free parameters θ of our model, given the mock data generated from HR4 D. Bayes' theorem is given by

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}. \tag{40}$$

Here $P(\theta|D)$ is the posterior, $P(\theta)$ is the prior, and $P(D|\theta)$ is the likelihood. $P(D)$ is the evidence, which here is an unimportant overall normalization factor.

For the likelihood we use the χ^2 with Gaussian errors, such that

$$\chi^2 = R^T C^{-1} R, \tag{41}$$

where R is the array of the residual, the difference between the theoretical value from equation (39) and the data. We weight this difference with the inverse of the covariance matrix $C = \mathrm{Cov}(C_\ell, C_{\ell'})$, defined in Section 2.5.3 and it is constant when sampling the space of parameters. Finally, the prior is set such that both b and $f > 0$, with some large upper limit.

We use the affine-invariant ensemble sampler, known as MCMC hammer and described in Foreman-Mackey et al. (2013),[2] to sample the parameter space.

3 RESULTS

3.1 Maps

We have created full sky maps in the 700–800 MHz frequency range. We have created pure 21 cm maps for three different bandwidths, df = 2.5, 5, and 10 MHz. This matches the expected frequency range of the Tianlai survey. Although we can go up to frequency bands of df = 1 MHz, we have decided to test this three different configurations to test our simulated maps, our pipeline, and the growth rate of structure test with different layers of systematic errors.

In Fig. 10 we show a 21 cm map for the frequency bin of $f = 790$–800 MHz as an example of our mock pure 21 cm maps. This

[2] The code emcee can be found at https://github.com/dfm/emcee.

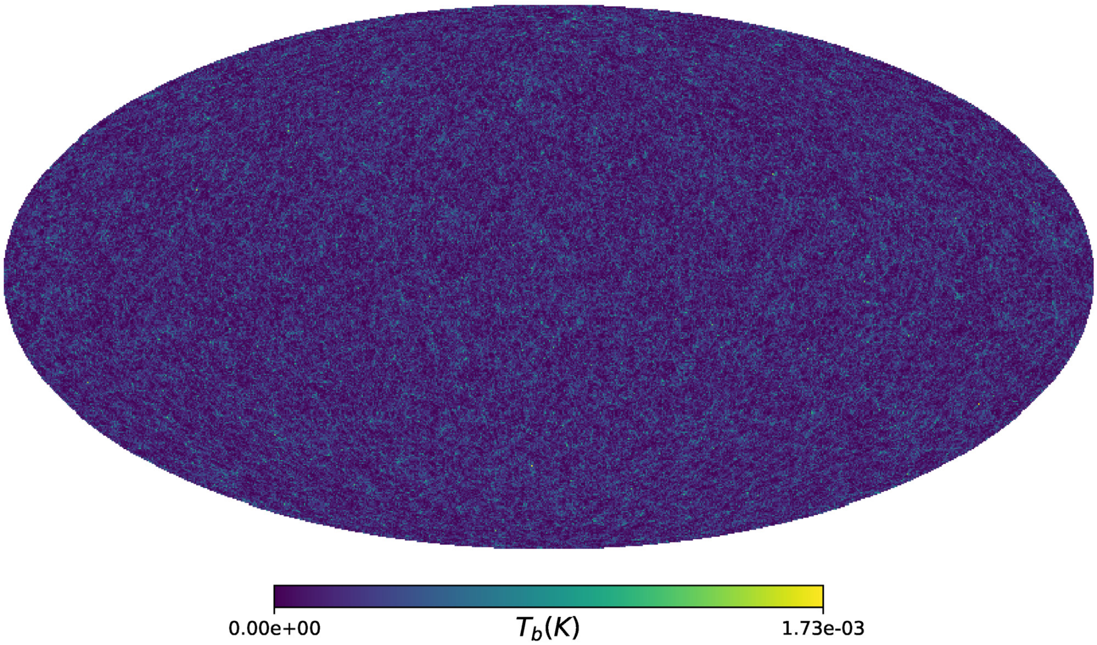

Figure 10. Neutral hydrogen map for frequency $f = 790$–800 MHz and $n_{\rm side} = 512$ generated from the full sky HR4 halo catalogue.

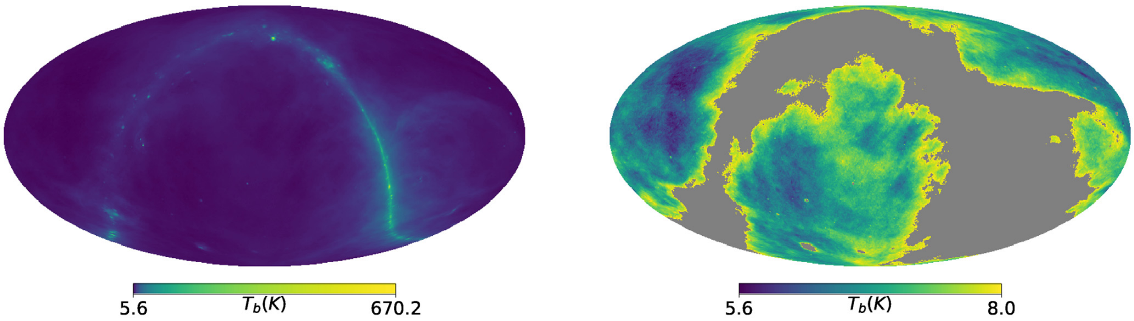

Figure 11. On the left, neutral hydrogen map for frequency $f = 790$–800 MHz and $n_{\rm side} = 512$ generated from the full sky HR4 halo catalogue and combined with a foreground map generated using the GSM model at $f = 795$ MHz. On the right, same map is masked with the MW mask shown in the left-hand panel of Fig. 6.

map contains the cosmological information encoded in the neutral hydrogen as tracer of matter in the redshift range $0.775 < z < 0.797$ given by the frequency band in which we have chosen to select galaxies when splitting the total frequency range in 10 bins with bandwidths of d$f = 10$ MHz. We must notice that we are capable of producing full sky maps because the HR4 halo catalogue is a full sky simulation up to redshift $z = 1.5$.

However, we show the contrast with a map that includes the foreground emission in Fig. 11. We see that the amplitude of the foreground signal is significantly higher than the cosmological signal. The first attempt of foreground removal we applied consisted on applying the Milky Way cut to remove most of the Galactic Diffuse emission. This is shown in the right-hand panel of Fig. 11.

In Fig. 12, we show a comparison between the different components that we consider on our maps. On the right-hand panel, we show a patch of the map shown in Fig. 10 which corresponds to the brightness temperature of neutral hydrogen in the frequency bin 790–800 MHz. On the left-hand panel, we show what corresponds to an observed map, without considering receiver noise. This map only includes the information from the cosmological neutral hydrogen brightness temperature and the foreground signal in the same redshift range given by the GSM. As can be seen, no structure can be distinguished when the foregrounds are added. In the central panel, we show the reconstructed brightness temperature map when fastICA has been applied to the map on the left-hand panel and considering two components. We can recognize the structure on the right-hand panel in the middle panel.

3.2 Reconstructed angular power spectra

We define the reconstructed angular power spectra as the ones given by the maps obtained after removing foregrounds, first with a Milky Way mask, and secondly applying a foreground removal algorithm.

3.2.1 Transfer functions

The main goal of foreground removal is to reconstruct the original cosmological information. When we compare the reconstructed

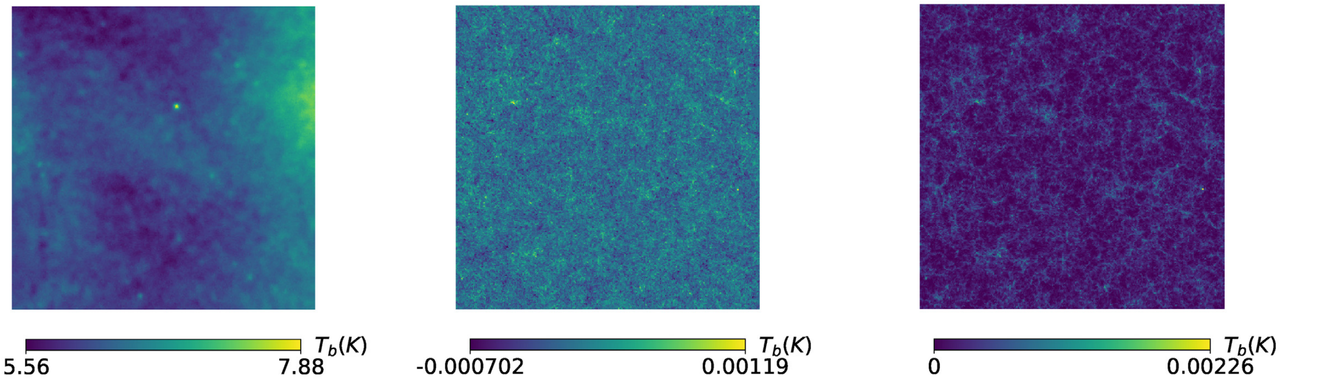

Figure 12. On the left, we show the observed map that includes the cosmological signal and the radio foreground signal. On the centre, the fastICA recovered map is shown after removing first two non-Gaussian components. On the right-hand panel, we show the original cosmological map. We can see how the central and the right-hand panels are similar.

angular power spectra C_ℓ after foreground removal with the angular power spectra from the cosmological maps, we find that the foreground removal process removes partially cosmological information as seen in Fig. 12.

The approach we have used in order to try to reconstruct the removed cosmological information is to define a transfer function, T_ℓ:

$$C_\ell^{21} = T_\ell C_\ell^{21,fr} \tag{42}$$

that is calibrated using the pure cosmological 21 cm map, C_ℓ^{21} and the same map once we run our foreground removal (*fr*) method on it, $C_\ell^{21,fr}$. The procedure is simple, we run the foreground removal technique on the simulated cosmological 21 cm maps in order to measure $C_\ell^{21,fr}$. Then, we fit the ratio $C_\ell^{21}/C_\ell^{21,fr}$ to the following functional form of a transfer function:

$$T_\ell = \exp^{-\ell_\star \log \ell} + C \tag{43}$$

given by a power law in the large scales and a constant in the small scales. The reason for this is caused by the fact that the foreground removal techniques tend to remove information from the long-wavelength modes. To calibrate this effect in the large scales, we consider a normalization scale in the power-law term, $\ell_\star$. while we just calibrate the transfer function in the small scales by the best-fitted constant C. If the foreground removal technique is working nicely, this constant should be close to 1.

We repeat this approach for each map used in this paper. When we run the foreground removal algorithm on a map with foregrounds, we apply this fitted transfer functions T_ℓ to each measure angular power spectra to correct from the effect of foreground removal on cosmological information.

In Fig. 13 we show the ratio between the cosmological power spectra and the power spectra of the foreground removed maps. We also show the transfer function best fit for the df = 10 bin configuration.

We show a table with all the best-fitting values for each frequency bin in Appendix A.

3.2.2 *Frequency bin width of 10 MHz*

We need to understand the physics of the different maps included in this analysis by first measuring the auto-angular power spectra. In Fig. 14, we show some examples in order to understand and test our simulated maps. We show the angular power spectra for the f = 700–710 MHz bin, which corresponds to a redshift range z = 1–1.03 and it is the highest redshift bin we consider in this bin configuration. In the top left-hand panel of Fig. 14, we show the angular power spectra C_ℓ for the pure cosmological signal (black circles). In red circles, we can see the signal given by the map produced after removing two components of the fastICA decomposition, while we see the signal given by the map after removing three components is given by the blue squares. We can see how we are removing cosmological information, as the amplitude of the power spectrum is smaller. We have considered the highest redshift bin because it is on the edges of our frequency range where the fastICA reconstruction performance is worse.

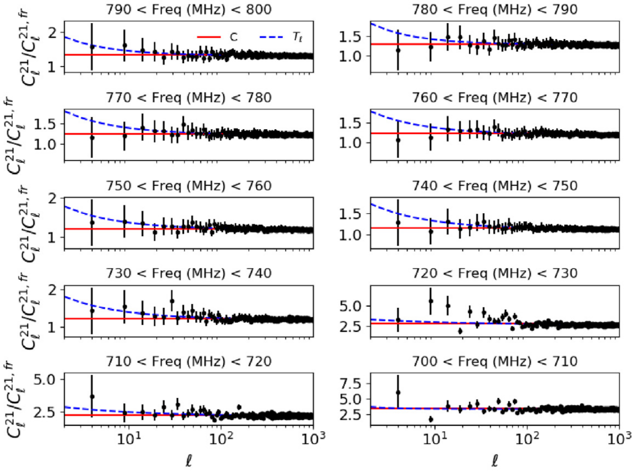

Figure 13. We show the transfer function for the 10 bins in the df = 10 MHz configuration, from the lower redshift bin on the top left to the higher redshift bin in the bottom right. The black circles correspond to the ratio between $C_\ell^{21}/C_\ell^{21,fr}$ while the dashed line corresponds to the function given by equation (43) and the best-fitting values of $\ell_\star$ and C for each frequency bin. The solid line corresponds to the best-fitting value of C in each redshift bin.

We can see the effect of the transfer functions in the top right-hand panel. In this case, we have applied the transfer function correction to the angular power spectra of the top left panel and as can be seen, we recover the original information. But this only happens when we run the foreground removal algorithm on the pure cosmology maps. When we include the foreground maps from the GSM model, we do not completely remove the information from

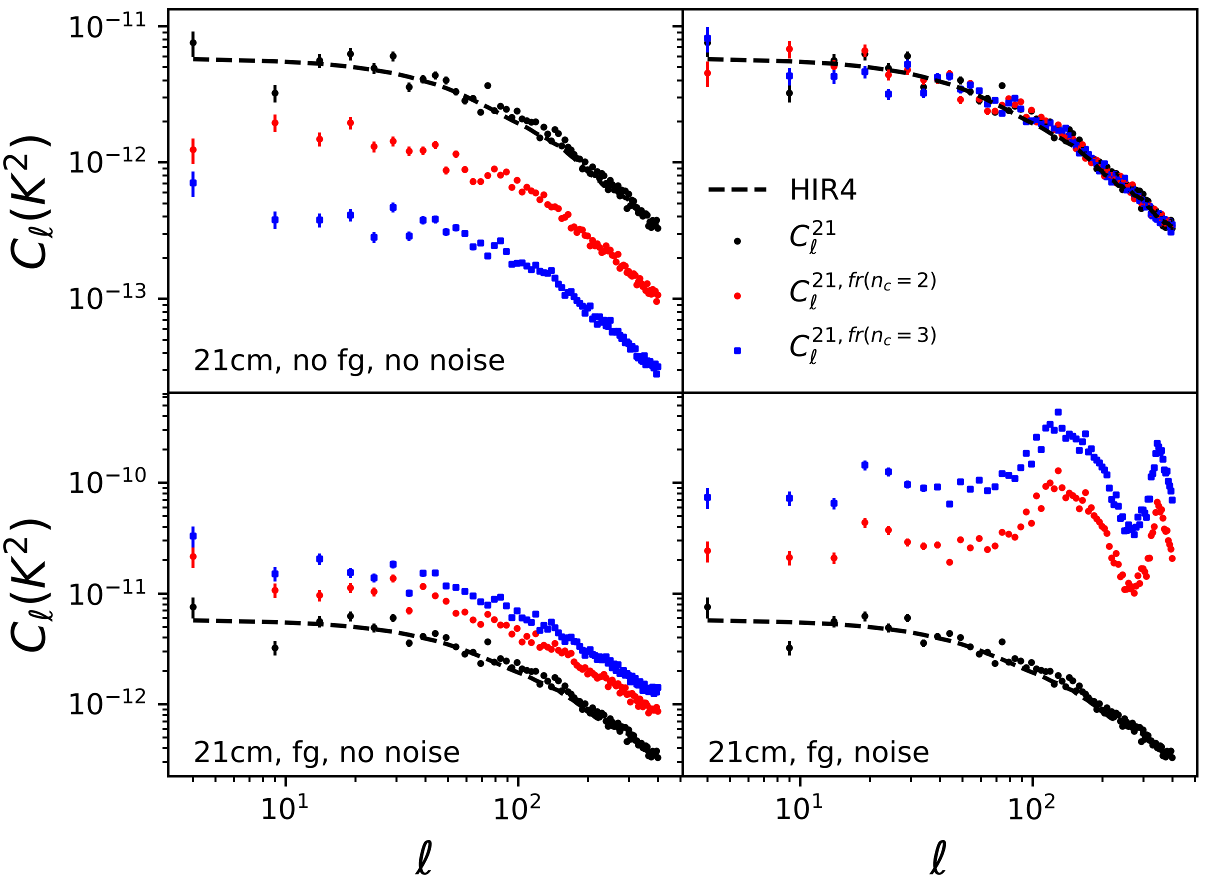

Figure 14. We show the angular power spectra for the $f = 700$–710 MHz frequency bin. On the top left, we show a comparison between the pure 21 cm signal and the angular power spectra corresponding to fastICA foreground removal maps for two and three components. On the top right, we apply the transfer functions to recover the original cosmological information. On the bottom left, we apply the transfer function to the maps with foregrounds, after foreground removal. Finally in the bottom right we include the Tianlai pathfinder noise map. The black dashed line corresponds to the simulation input.

the foregrounds, and therefore the amplitude of the angular power spectra of the corrected maps is higher than the one coming only from the large-scale structure temperature fluctuations. This can be seen in the right-hand bottom panel. Finally, if we include the noise maps predicted for Tianlai, we can see how the noise signal is the predominant one, even after foreground removal, in the left-hand bottom panel. The wiggles that can be seen in the bottom left-hand panel in Fig. 14 are due to the effect of the Tianlai baseline on the measurement beam. This beam presence is seen in any single measurement in maps that include our simulated Tianlai noise maps.

3.2.3 *Frequency bin width of 5 MHz*

We studied the effect of including more frequency bins in the foreground removal recovery of cosmological information. In Fig. 15, we show the angular power spectra for the lowest frequency bin 700–705 MHz. We can see that even the maps produced directly by the fastICA foreground removal, with no transfer function correction, are closer to the cosmological power spectrum than the maps for the d$f = 10$ MHz case, as shown in top left-hand panel of Fig. 15. This is due to the fact that the bigger the number of frequency bins, the better we trace the smooth evolution with frequency of the foregrounds. We can see in this same panel that when the foreground removal is more efficient, we also remove cosmological information as the amplitude of the angular power spectra of the maps produced by fastICA foreground removal is smaller than the original signal. This is due to the fact that the algorithm is unable to distinguish between the cosmological signal and the radio foregrounds.

By definition, we show in the top right-hand panel that the transfer function allow us to recover the original input. On the bottom left-hand panel, we show the effect of applying foreground removal and transfer functions to the map with foregrounds and we see that we almost recover the original information. Therefore, the increase on the number of frequency bins is significantly improving our reconstruction of the cosmological information encoded in the observed maps.

Finally, we still see that the noise is the main signal on the recovered maps in the bottom right-hand panel, as the amplitude of the noise map is significantly higher than the cosmological signal. This implies that receiver noise and the smoothing caused by the survey strategy is the main systematic regarding the recovery of the original simulation information.

We also notice that there is almost no difference between both foreground removal maps, either if we remove only two components or three components. If this is the case, then using only the two component maps is a more reasonable option.

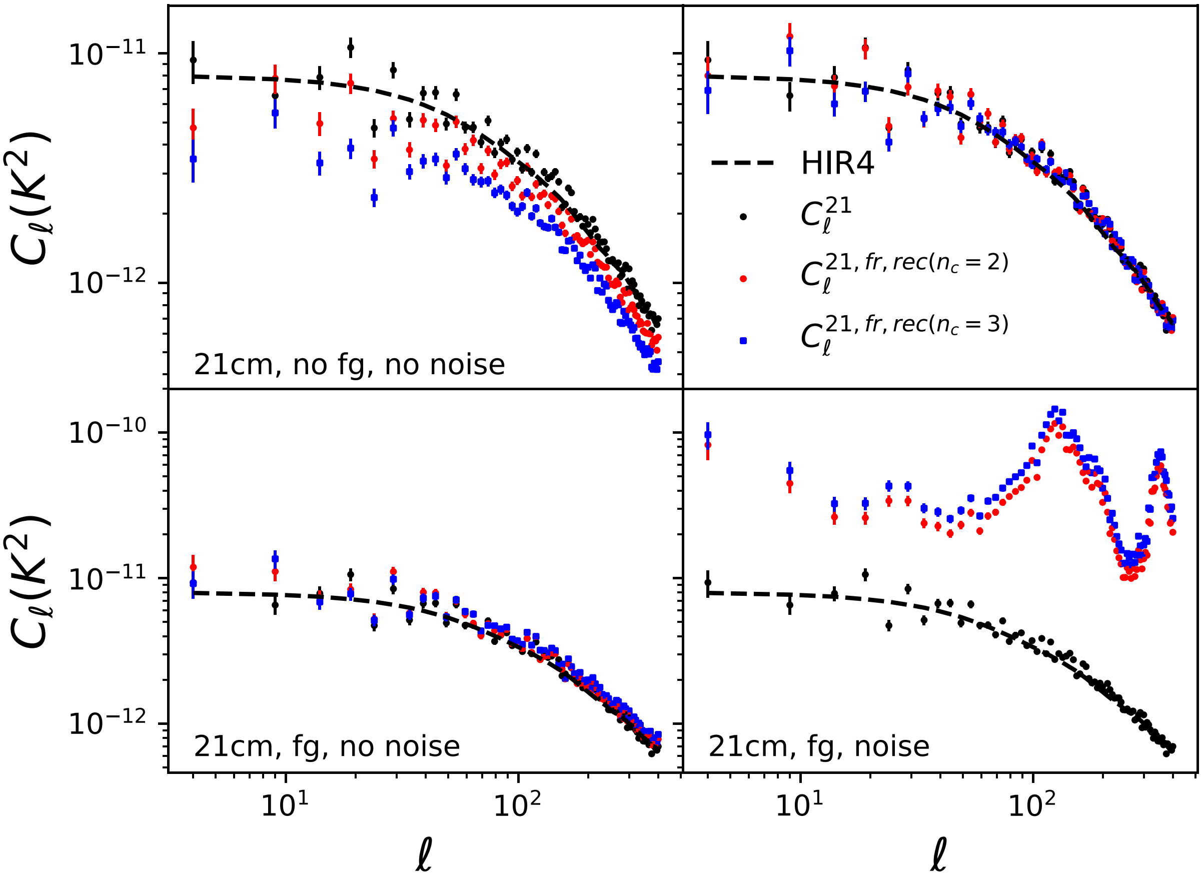

Figure 15. In this figure, we show the angular power spectra for the $f = 700$–705 MHz frequency bin. The top left-hand panel shows a comparison between the pure 21 cm signal and the angular power spectra corresponding to fastICA foreground removal maps for two and three components. On the top right, we show the reconstruction with the transfer functions to recover the original cosmological information. On the bottom left, we show the reconstruction with the transfer function for the maps with foregrounds. Bottom right-hand panel includes the Tianlai pathfinder noise map and the simulation input (dashed line).

3.2.4 Frequency bin width of 2.5 MHz

We finally test the smallest frequency configuration we are considering, the one in which the frequency bins have a bandwidth of df = 2.5 MHz, which corresponds to 40 bins in redshift. This is the best sampling of the frequency evolution that we use in this paper. By evaluating Fig. 16, we see almost no difference with respect to the df = 5 MHz case shown in Fig. 15. The most noticeable difference is the fact that the reconstructed map given by two or three foreground removal components are even more indistinguishable than before. This happens because the more frequency bins we consider, the better the foreground removal technique works and the foregrounds are better removed with less components. Therefore, the conclusions extracted from the previous case are the same here.

3.3 Cosmological constraints

In order to understand the cosmological information encoded in the simulation, to benchmark our H1R4 catalogues and to estimate how the foregrounds and the receiver noise affect the constraints on the growth rate of structure. Our procedure consisted on fitting the individual angular power spectrum of each redshift bin, for each bin configuration in the different catalogues introduced in Section 2.4. When fitting the cosmological maps given by $T_b^{\rm obs} = T_b^{\rm HI}$ or the maps that include foreground signal, $T_b^{\rm obs} = T_b^{\rm HI} + T_b^{\rm foreground}$ we restrict the fitting of the angular power spectrum to the scales $\ell = 20$–200. When we include the noise, the information on the small scales is meaningless for cosmology purposes and we limit our analysis to the scales between $\ell = 20$–60.

In order to include the information from the covariance matrix, we have done a preliminary fit assuming a linear Gaussian errors. Then, once we have a best-fitting values on the bias $b(z)$ and the growth $f(z)$, we created Gaussian realizations using a theoretical power spectrum for each bin given by the best-fitting parameters. When obtaining a most realistic covariance matrix, we have also considered the effect of the mask in which can imply the apparition of off-diagonal elements in the covariance matrix, although they may be small as shown in Fig. 9. The cosmological constraints shown here are the ones that were obtained using the full covariance matrix information.

3.3.1 Frequency bin width of 10 MHz

We have constrained the values of the neutral hydrogen bias and the growth rate of structure for different bin configurations and foreground and noise levels. For each bin configuration, we have increased the layers of complexity by adding foreground signal and receiver noise. In addition, we tested the effect of foreground removal techniques in the cosmological constraints.

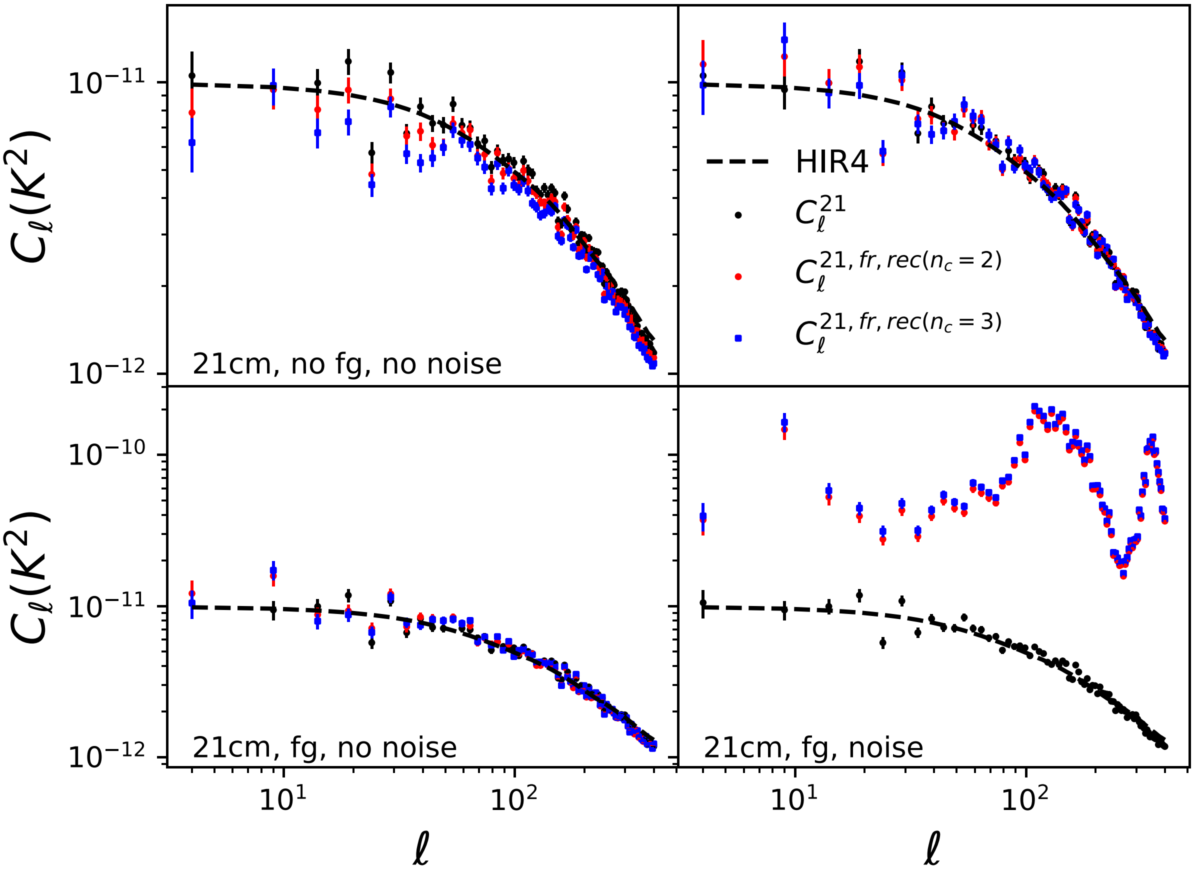

Figure 16. Same as Figs 14 and 15 for the $f = 700$–702.5 MHz frequency bin. On the top left-hand panel, we show a comparison between the pure 21 cm signal and the angular power spectra corresponding to fastICA foreground removal maps for two and three components with the correction function applied in the top right-hand panel. We show the same on the bottom left but when the maps include the radio foregrounds. Finally in the bottom right we include the Tianlai pathfinder noise and all the plots include the input value from the simulation (dashed line).

We see on the top panel of Fig. 17 the constraints on both the hydrogen bias and the growth rate of structure when we only consider the cosmological information in our simulated maps. In this case, we recover the cosmological signal that was used to generate the mock hydrogen catalogues. But we also show what happens if we run the foreground removal algorithm on the 21 cm cosmological maps. On the same panel we can see how the best-fitting values on the bias are smaller than the theoretical expected values. This is due to the fact that with only 10 bins the reconstruction, which is based on segregating the smooth signal from the foreground from the density fluctuations in the temperature field. Let us remind that there is much more fitting bias when measuring the hydrogen bias $b(z)$ than when measuring the growth rate $f(z)$. This happens because the hydrogen bias is mostly constrained by the small scales as it affects all the scales, while the growth rate only affects the larger scales as the linear redshift-space distortions only add a boost in the amplitude at large scales. This is the reason why the bias on the growth rate is much smaller.

When we include the foregrounds and repeat the same fit to the same cosmological parameters, we obtain the plot shown in the middle of Fig. 17. In this case, the best-fitting values for the hydrogen bias continue to be biased with respect to the input theoretical values used to generate the catalogues. Again, the growth rate values are recovered for the same reasons stated above. The only main difference with respect to the previous plot is that we need higher values on the bias in order to fit the observed angular power spectra show in the bottom left-hand panel of Fig. 14 as the addition of foregrounds in the map makes the foreground removal less efficient.

Finally, in the bottom panel we see that when considering the noise maps, it is impossible to recover the bias information because all the small scale information is destroyed by the beaming. As the bias is not constrained, we obtain best fits on the growth rate consistent with no redshift-space distortions as the noise avoids this possibility.

3.3.2 *Frequency bin width of 5 MHz*

We checked how the narrowing of the binning affects the reconstruction of the cosmological information after foreground removal. We show in the top panel of Fig. 18 how the addition of more frequency bins affects the performance of fastICA. Comparing with the top panel of Fig. 17, where there were only ten frequency bins, we find that the best-fitting values for the hydrogen bias and growth rate are closer to the input values. This is explained by the better sampling of the evolution of the foreground temperature maps with frequency, allowing for a better reconstruction of the cosmological signal and a more accurate fit.

We also notice a similar pattern in both cases ($df = 5$ and $df = 10$) in which the constraints for both the bias and the growth rate of structure at higher redshifts are more biased (less accurate)

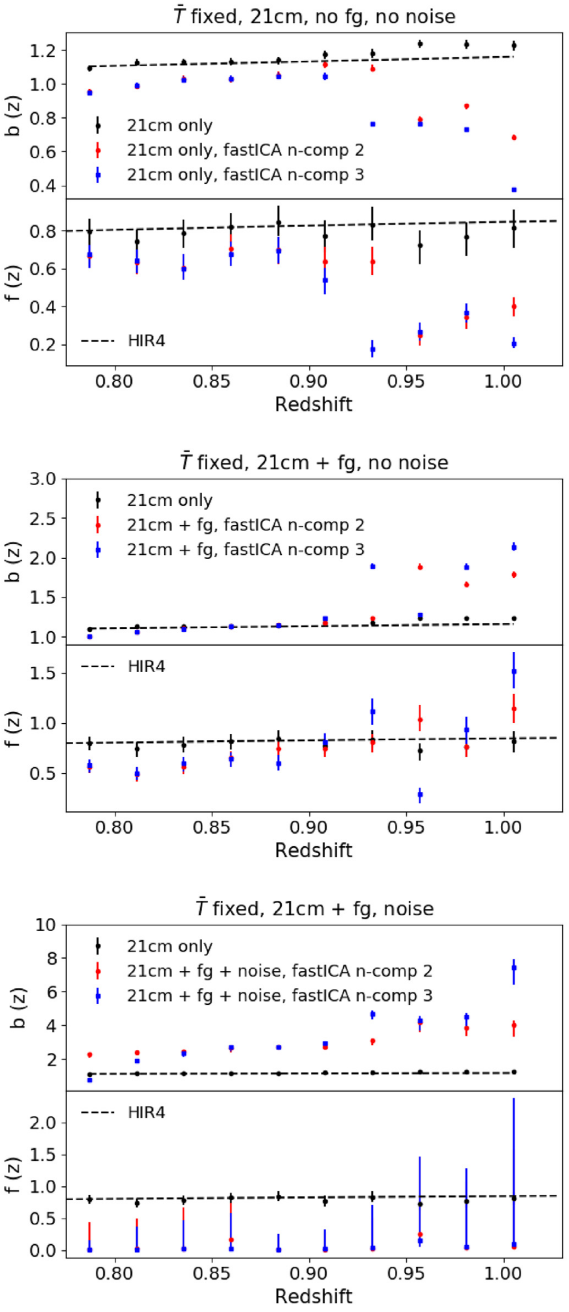

Figure 17. The 68 per cent confidence limit on the linear bias $b(z)$ and the growth rate of structure $f(z)$ as a function of redshift bin for the df = 10 MHz case. Top panel shows the case in which only 21 cm pure maps are considered (and no transfer function applied), middle panel considers the case in with foreground signal while the bottom panel shows the constraints when including the noise map. In both middle and lower panel the fit is done by applying a transfer function. We include the HR4 theoretical values (dashed lines).

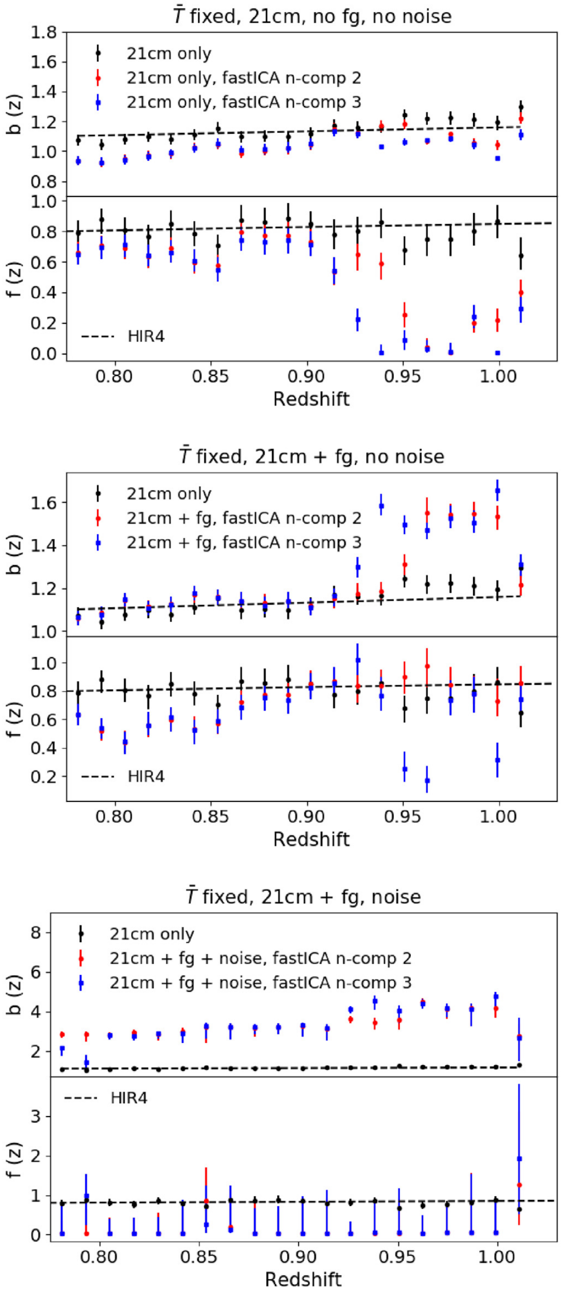

Figure 18. The 68 per cent confidence limit on the linear bias $b(z)$ and the growth rate of structure $f(z)$ as a function of redshift bin for the df = 5 MHz case. Top panel shows the case with 21 cm cosmological maps and two different fastICA reconstructions (without correction). Middle panel considers the case with foreground signal and the bottom panel shows the constraints when including the noise map. We include the simulation input with the dashed lines).

than at lower redshifts. (Compare the top and middle panels of Figs 17 and 18). This happens because the reconstruction method removes more power on all scales at higher redshift than at lower redshift, as can be seen on Fig. 13 for the df = 10 MHz case. This removed power cannot be easily corrected for, even when using the transfer function, as the reduction in cosmological power is not matched when reconstructing a foreground contaminated sky. In particular, for fastICA, there is the possibility is that the small number of angular modes present in the sky map on very large scales means that the power on these scales is not Gaussian distributed (owing to cosmic variance). The fastICA reconstruction method is then flexible enough to 'fit' and remove this non-Gaussianity, though it is not clear why this only happens at high redshift. This problem regarding non-Gaussianities is not as large in PCA and log-polynomial fit algorithms, but both of these reconstruction methods also suppress cosmological information on all scales, as they are not able to segregate all foreground components from the cosmological signal and to model completely the foreground signal, respectively.

In the middle panel, we show the best-fitting parameters for the case in which we added the foreground signal to the cosmological maps. Again, the recovery of the growth rate is much better than the recovery of the bias values. But we also notice that we can measure the growth rate of structure with a 10 per cent precision.

When we introduce the noise we do not recover the theoretical value of the bias. As we restrict our fitting at $\ell_{\max} = 60$, we obtain less offset (or more accurate) fits on the growth rate of structures by degrading the precision of the fit.

3.3.3 *Frequency bin width of 2.5 MHz*

The configuration with the narrowest bin configuration, df = 2.5 MHz, should allow for a better reconstruction of the cosmological information as it allows us to sample better the smooth components from the foregrounds. We can see this in the different panels of Fig. 19. On the top panel we see that for a conservative foreground removal approach we almost do not remove part of the cosmological signal as in the previous cases. This is an improvement with respect to the previous cases due to the better sampling of the frequency range. The reason that removing only two components of the fastICA decomposition works better than for three is due to the fact that when we remove more components the risk of removing cosmological information increases, as it happens in this case.

When we also include the foreground signal in the maps, we also do better foreground removal if only considering two components of the fastICA decomposition. The recovering of the input parameters is also good at higher redshifts as the narrower redshift bin improves the foreground removal. We can measure the growth rate with a precision of 10 per cent again in this case.

Finally, we see the same pattern than in the df = 5 case when introducing the noise on the mock maps. By reducing the scales included in the fitting, $\ell = 20$–60, we cannot measure the hydrogen bias as we do not include the small scales but we can measure the growth rate of structures with a 100 per cent precision.

3.4 Foreground removal comparison

In the previous results we only considered the maps after foreground removal reconstruction given by fastICA. In this section we explore the alternative reconstruction methods.

In terms of the PCA reconstruction, we found that the results to be very similar to those from fastICA, with the same number of modes. Once again, increasing the number of PCA modes from $n = 1$ to $n = 2$ had a noticeable impact on the cosmology recovered, significantly reducing the offset between the measured posterior and the posterior for the 21 cm only case. Changing from $n = 2$ to $n = 3$ introduced no significant change in the size of the measured offset, considering all of the bins as a whole.

In terms of the in log-polynomial fitting, we first experimented with various n (maximum order of polynomial), with equations (31)–(33). We found that $n = 3$ gives the best result in the frequency range [700–800] MHz in removing the foreground. This result seems to be due to the smallness of n that is just optimal to mimic the smoothness of the foreground. We also found that the goodness of reconstruction, in terms of the recovered angular power spectrum, depends on the frequency ν_i. This reflects the relative weakness in the polynomial fitting, as is also demonstrated by de Oliveira-Costa et al. (2008, fig. 3). As shown in that paper, the log-polynomial model is not as good as describing the physics of the foregrounds as the log-polynomial model is simpler than the real physical model.

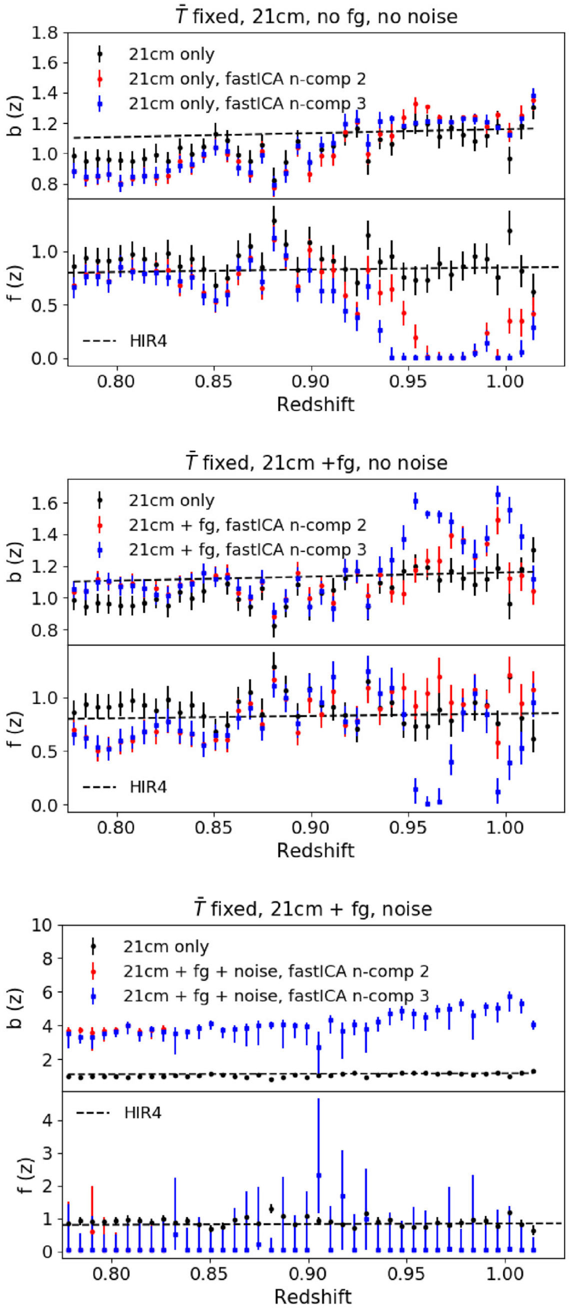

Figure 19. Same as Figs 17 and 18 for the df = 2.5 MHz case. In the top panel we show the best fit for the case with pure simulated cosmological maps (without corrections). The dashed line represents the simulation input information. Middle panel includes the GSM foreground signal. We include the results for two (red circles) and three (blue squares) fastICA components in the reconstructed maps. The bottom panel shows the best fit when including Tianlai noise.

In Fig. 20 we show a comparison between the fits for the hydrogen bias, b and the growth rate $f(z)$ when considering the fastICA and the PCA reconstructed maps with $n_{\rm comp} = 2$ and when removing components using a polynomial of third order. From the previous results we learned that this number of components is enough in terms

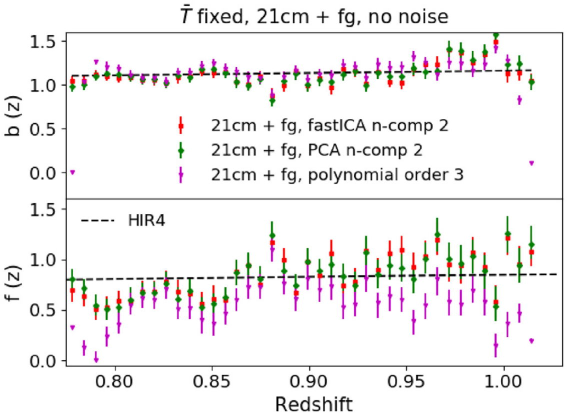

Figure 20. Comparison between foreground removal techniques. We show the constraints on the hydrogen bias and the growth rate for the fastICA, PCA, and log-polynomial fitting reconstructed maps, including the transfer functions, for two components and order three for the polynomial.

of foreground removal and we also consider the case with frequency bins of df = 2.5 MHz as this configuration produces the best reconstruction. By inspecting the figure, we see that both fastICA and PCA methods provide equally good results while polynomial method does not recover that well the growth rate information, especially on the first and last redshift bins. Therefore, we decided to use fastICA in this study arbitrarily as there is no method that performs better than the other. However, we may consider to use PCA for future analysis as its reliability is also quite good.

4 SUMMARY

With the rise of radio cosmology, hydrogen intensity mapping has been raised as a promising cosmological probe. Currently there are a number of different SKA pathfinders that are producing the first wide-field intensity mapping surveys. In order to understand the systematic errors, which we expect from radio foregrounds and receiver noise and beaming, and the possibilities of improving the cosmological constraints in the near future with this type of survey, we need to develop sophisticated cosmological and observational simulations.

We have created the first simulations of 21 cm intensity map signal across the entire sky produced using the HR4 simulations. The simulated catalogues can cover the full sky up to redshift z = 1.5. To match the frequency range of the Tianlai pathfinder experiment, in this paper we have focused on the range 700–800 MHz. Starting from the Friend-of-friends halo catalogue, we have applied a halo model in which the neutral hydrogen mass contained in each halo is given by the dark matter halo mass and the virial velocity of the halo, obtained assuming a spherical collapse model, following the prescription of Padmanabhan & Refregier (2017). In particular, neutral hydrogen populations are suppressed in low-mass haloes, as the gas is not bound to the halo, and in more massive haloes as the neutral hydrogen gas is heated and becomes excited. Once we have a sample of haloes with neutral hydrogen, we convert the mass in a given redshift bin and an angular pixel to a brightness temperature, and generate our set of maps for each frequency band.

To test the consistency of our analysis process, and also to forecast the effectiveness of the maps as a cosmological probe, we measure the angular power spectra and covariance matrix of the 21 cm intensity at different redshifts. We use these data products to constrain the hydrogen bias b_{HI} and the growth rate of structure f. We show that from the pure 21 cm cosmological maps we obtain the same values for these parameters as those predicted assuming the cosmology that generated the original HR4 simulation.

We have also created maps that include the foreground signal as well as the cosmological contribution. The foreground maps are created using the Global Sky Model. This method uses the information from 29 maps at different frequencies and performs a PCA decomposition of six components in order to produce foreground maps at any frequency. In particular, for the frequency range we are considering (700–800 MHz), the main foregrounds are synchrotron emission, Galactic neutral hydrogen and thermal free–free emission. We have not considered adding any ad hoc information from extragalactic point source emission, part of which should be in the GSM maps.

Once we included the foreground signal, we first masked the Galactic Centre, as the foreground emission is unavoidable here. With the remaining unmasked part, we applied the reconstruction techniques in order to remove the foregrounds and recover the cosmological signal, which were independent component analysis fastICA and PCA. We created recovered maps by removing two or three components.

We show that when we apply the foreground removal algorithm to the data, we are removing part of the cosmological information, even if we apply it in the case where no foreground is present. Since a strategy is needed to account for the missing power, our chosen option is to define a transfer function that corrects for this. The parameters of the transfer function are fixed by the best fit to the ratio between the original cosmological signal (pure 21 cm simulations) and the maps that are produced by foreground removal when we apply it directly to the original maps (reconstructed maps). This correction technique becomes more successful as we increase the number of bins, i.e. it works better when the foreground removal is optimal.

We found that in all cases without instrument noise, but where the transfer function correction to the angular power spectrum has been applied, we still recover the input values for the hydrogen bias growth rate. There is a small degree of offset between the input and recovered values of b and f, but this decreases as the number of frequency bins increases, as the foreground reconstruction process becomes more effective for a larger number of bins.

Finally, we considered the effect of noise maps produced for the Tianlai survey. In this case it was impossible to use the small scale part of the angular power spectra for cosmological parameter estimation. When constraining the angular power spectrum, the bias information is set from the amplitude while the information on the growth rate comes from the boost in the low multipole-part of the spectra. If we are unable to recover any cosmological signal on small scales, this then removes our ability to constrain the hydrogen bias $b(z)$. This in turn diminishes our ability to see any relative change between the large scale and the small scale power due to redshift-space distortions and weakens our constraints on the linear growth rate of structure $f(z)$.

We have shown in this paper when considering the predicted noise present in the Tianlai instrument, we are not able to recover any information on the hydrogen bias, and can only partially recover the information on the growth of structure through truncating to the large scale information. Enhanced noise removal techniques should be considered in the future in order to fully recover the cosmological information in an unbiased manner.

The presence of noise and foreground residuals can also be mitigated by cross-correlation of the radio intensity map with some optical galaxy catalogue. We will extend this work to use the HR4 simulation to generate a galaxy redshift survey over the same region of sky and redshift, and demonstrate the utility of cross-correlation in accurately recovering the input cosmological parameters (Shi et al., in preparation).

ACKNOWLEDGEMENTS

We like to thank Sungwook Hong, Benjamin L'Huillier, and Changbom Park for multiple discussions and for providing the HR4 simulations. We thank Glen Rees, Syed Faisal ur Rahman, and Song Chen for comments on early software development relating to this project. JA, DP, FS, and Y-SS are supported by the project 우주거대구조를 이용한 암흑우주 연구 ('Understanding Dark Universe Using Large Scale Structure of the Universe'), funded by the Ministry of Science. KA is supported by NRF-2016R1D1A1B04935414 and NRF-2016R1A5A1013277. LZ is supported by the National Science Foundation of China (11621303, 11653003, 11773021, 11890691), the National Key R&D Program of China (2018YFA0404601, 2018YFA0404504), the 111 project, and the CAS Interdisciplinary Innovation Team (JCTD- 2019-05). This work has been produced using CosKASI Seondeok cluster and the KISTI Nurion machine with project 'HIR4: simulated neutral hydrogen intensity maps from Horizon Run 4 for Future Galaxy Surveys'.

This research made use of ASTROPY,[3] a community-developed core Python package for Astronomy (Astropy Collaboration 2013; Price-Whelan et al. 2018), the HEALPIX and HEALPY package (Górski et al. 2005; Zonca et al. 2019), the NUMPY package Oliphant (2006), the SCIPY package (Virtanen et al. 2019), and MATPLOTLIB package (Hunter 2007).

REFERENCES

Alonso D., Ferreira P. G., Santos M. G., 2014, MNRAS, 444, 3183
Alonso D., Bull P., Ferreira P. G., Santos M. G., 2015, MNRAS, 447, 400
Alonso D., Sanchez J., Slosar A., 2019, MNRAS, 484, 4127
Ando R., Nishizawa A. J., Hasegawa K., Shimizu I., Nagamine K., 2019, MNRAS, 484, 5389
Asorey J., Crocce M., Gaztañaga E., Lewis A., 2012, MNRAS, 427, 1891
Asorey J., Crocce M., Gaztañaga E., 2014, MNRAS, 445, 2825
Astropy Collaboration et al., 2013, A&A, 558, A33
Barnes L. A., Haehnelt M. G., 2015, MNRAS, 454, 218
Battye R. A., Browne I. W. A., Dickinson C., Heron G., Maffei B., Pourtsidou A., 2013, MNRAS, 434, 1239
Bharadwaj S., Sethi S. K., 2001, J. Astrophys. Astron., 22, 293
Blake C., Bridle S., 2005, MNRAS, 363, 1329
Blake C., Glazebrook K., 2003, ApJ, 594, 665
Brown M. L., Bonaldi A., 2015, MNRAS, 447, 1973
Brown M. L., Castro P. G., Taylor A. N., 2005, MNRAS, 360, 1262
Bryan G. L., Norman M. L., 1998, ApJ, 495, 80
Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, ApJ, 803, 21
Cabré A., Fosalba P., Gaztañaga E., Manera M., 2007, MNRAS, 381, 1347
Camera S., Santos M. G., Ferreira P. G., Ferramacho L., 2013, Phys. Rev. Lett., 111, 171302
Chapman E. et al., 2012, MNRAS, 423, 2518
Chapman E. et al., 2013, MNRAS, 429, 165
Chen X., 2011, Sci. Sinica Phys. Mech. Astron., 41, 1358
Chen S.-F., Castorina E., White M., Slosar A., 2019, J. Cosmol. Astropart. Phys., 2019, 023
Colless M. et al., 2001, MNRAS, 328, 1039
Crighton N. H. M. et al., 2015, MNRAS, 452, 217
Crocce M., Cabré A., Gaztañaga E., 2011, MNRAS, 414, 329
Cunnington S., Wolz L., Pourtsidou A., Bacon D., 2019, , MNRAS, 488, 5452
Das S. et al., 2018, in Proc. SPIE, Vol. 10708. p. 1070836
Davis T. M., Hinton S. R., Howlett C., Calcino J., 2019, MNRAS, 490, 2948
de Oliveira-Costa A., Tegmark M., Gaensler B. M., Jonas J., Landecker T. L., Reich P., 2008, MNRAS, 388, 247
Eke V. R., Cole S., Frenk C. S., 1996, MNRAS, 282, 263
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Geller M. J., Huchra J. P., 1989, Science, 246, 897
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Haynes M. P. et al., 2018, ApJ, 861, 49
Hu W., Haiman Z., 2003, Phys. Rev. D, 68, 063004
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Kaiser N., 1987, MNRAS, 227, 1
Kim J., Park C., L'Huillier B., Hong S. E., 2015, J. Korean Astron. Soc., 48, 213
Lahav O., Lilje P. B., Primack J. R., Rees M. J., 1991, MNRAS, 251, 128
Li X.-D., Park C., Sabiu C. G., Park H., Weinberg D. H., Schneider D. P., Kim J., Hong S. E., 2016, ApJ, 832, 103
Linder E. V., Cahn R. N., 2007, Astropart. Phys., 28, 481
Marín F. A., Gnedin N. Y., Seo H.-J., Vallinotto A., 2010, ApJ, 718, 972
Modi C., Castorina E., Feng Y., White M., 2019, J. Cosmol. Astropart. Phys., 2019, 024
Oliphant T., , 2006, Guide to NumPy, Trelgol Publishing, USA
Padmanabhan H., Refregier A., 2017, MNRAS, 464, 4008
Padmanabhan H., Choudhury T. R., Refregier A., 2016, MNRAS, 458, 781
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Peebles P. J. E., 1980, The large-scale structure of the universe, Princeton University Press, United States
Peebles P. J. E., 1984, ApJ, 284, 439
Price-Whelan A. M. et al., 2018, AJ, 156, 123
Sefusatti E., Komatsu E., 2007, Phys. Rev. D, 76, 083004
Seo H.-J., Eisenstein D. J., 2003, ApJ, 598, 720
Shaw J. R., Sigurdson K., Pen U.-L., Stebbins A., Sitwell M., 2014, ApJ, 781, 57
Shaw J. R., Sigurdson K., Sitwell M., Stebbins A., Pen U.-L., 2015, Phys. Rev. D, 91, 083514
Spinelli M., Zoldan A., De Lucia G., Xie L., Viel M., 2020, MNRAS, 493, 5434
Square Kilometre Array Cosmology Science Working Group et al., 2020, Publ. Astron. Soc. Austr., 37, e007
Thompson A. R., Moran J. M., Swenson George W. J., 2001, Interferometry and Synthesis in Radio Astronomy, 2nd Edition. Springer, Berlin
Villaescusa-Navarro F. et al., 2018, ApJ, 866, 135
Virtanen P. et al., 2019, preprint (arXiv:1907.10121)
Wang Z. et al., 2019, preprint (arXiv:1901.02724)
Wolz L., Abdalla F. B., Blake C., Shaw J. R., Chapman E., Rawlings S., 2014, MNRAS, 441, 3271
Zhang J., Zuo S.-F., Ansari R., Chen X., Li Y.-C., Wu F.-Q., Campagne J.-E., Magneville C., 2016, Res. Astron. Astrophys., 16, 158
Zheng H. et al., 2017, MNRAS, 464, 3486
Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, J. Open Source Softw., 4, 1298

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Table A1. Best-fitting parameters of the transfer function for the different bin configuration and the different number of components of the fastICA decomposition used to reconstruct the hydrogen information.

[3] http://www.astropy.org

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: TRANSFER FUNCTION FITTING

As described in Section 3.2.1, we show in Table A1 the best-fitting values for the transfer function parameters $\ell_\star$ and C for the different bin configurations and fastICA $n_{\rm comp} = 2$ and $n_{\rm comp} = 3$.

Table A1. Best-fitting parameters of the transfer function for the different bin configuration and the different number of components of the fastICA decomposition used to reconstruct the hydrogen information. Full table can be found as supplementary material online.

	$n_c = 2$		$n_c = 3$		$n_c = 2$		$n_c = 3$		$n_c = 2$		$n_c = 3$	
Freq. (MHz)	$\ell_\star$	C	$\ell_\star$	C	$\ell_\star$	C	$\ell_\star$	C	$\ell_\star$	C	$\ell_\star$	C
797.5–800	1.196	1.327	1.193	1.327	1.104	1.338	1.101	1.338	0.821	1.34	0.82	1.34
795–797.5	–	–	–	–	–	–	–	–	1.237	1.33	1.228	1.33
792.5–795	–	–	–	–	0.937	1.317	0.942	1.318	1.136	1.327	1.137	1.327
790–792.5	–	–	–	–	–	–	–	–	1.048	1.326	1.045	1.326

This paper has been typeset from a TEX/LATEX file prepared by the author.