Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **517,** 4637–4655 (2022) https://doi.org/10.1093/mnras/stac2832
Advance Access publication 2022 October 8

The Tianlai dish array low-z surveys forecasts

Olivier Perdereau,[1]★ Réza Ansari,[1] Albert Stebbins,[2] Peter T. Timbie,[3] Xuelei Chen ,[4,5,6,7] Fengquan Wu,[4] Jixia Li,[4,5] John P. Marriner,[2] Gregory S. Tucker,[8] Yanping Cong,[4,5] Santanu Das ,[9] Yichao Li,[7] Yingfeng Liu ,[4,5] Christophe Magneville,[10] Jeffrey B. Peterson,[11] Anh Phan,[3] Lily Robinthal,[3] Shijie Sun,[4,5] Yougang Wang ,[4] Yanlin Wu,[3] Yidong Xu ,[4] Kaifeng Yu,[4,5] Zijie Yu,[4,5] Jiao Zhang,[12] Juyong Zhang[13] and Shifan Zuo[14]

[1]*Université Paris-Saclay, CNRS/IN2P3, IJCLab, F-91405 Orsay, France*
[2]*Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510-5011, USA*
[3]*Department of Physics, University of Wisconsin Madison, 1150 University Ave, Madison, WI 53703, USA*
[4]*National Astronomical Observatory, Chinese Academy of Sciences, 20A Datun Road, Beijing 100101, P. R. China*
[5]*University of Chinese Academy of Sciences, Beijing 100049, P. R. China*
[6]*Center of High Energy Physics, Peking University, Beijing 100871, P. R. China*
[7]*Department of Physics, College of Sciences, Northeastern University, Shenyang, Liaoning 110819, P. R. China*
[8]*Department of Physics, Brown University, 182 Hope St., Providence, RI 02912, USA*
[9]*Department of Physics, Imperial College London, Kensington, London SW7 2AZ, London, England*
[10]*CEA, DSM/IRFU, Centre d'Etudes de Saclay, F-91191 Gif-sur-Yvette, France*
[11]*Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*
[12]*College of Physics and Electronic Engineering, Shanxi University, Taiyuan, Shanxi 030006, P. R. China*
[13]*Hangzhou Dianzi University, 115 Wenyi Rd., Hangzhou 310018, P. R. China*
[14]*Department of Astronomy and Tsinghua Center for Astrophysics, Tsinghua University, Beijing 100084, P. R. China*

Accepted 2022 September 23. Received 2022 September 16; in original form 2022 May 11

ABSTRACT

We present the science case for surveys with the Tianlai dish array interferometer tuned to the [1300, 1400] MHz frequency range. Starting from a realistic generation of mock visibility data according to the survey strategy, we reconstruct maps of the sky and perform foreground subtraction. We estimate the level of residuals from imperfect subtraction, mostly due to mode mixing, i.e. distortions in the reconstructed 3D maps due to frequency-dependent instrument response. We show that a survey of the North Celestial Polar cap during a year of observations, covering an area of 150 deg^2, would reach a sensitivity of $1.5-2$ mK per 1 MHz $\times$ 0.25^2 deg^2 voxel and be marginally impacted by mode mixing. Tianlai would be able to detect $\sim$10 nearby massive HI clumps as well as a very strong cross-correlation signal of 21 cm intensity maps with the North Celestial Cap Survey optical galaxies. We also studied the performance of a mid-latitude survey, covering $\sim$ 1500 deg^2 overlapping the SDSS footprint. Despite a higher noise level for the mid-latitude survey, as well as significant distortions due to mode mixing, Tianlai would be able to detect a highly significant cross-correlation between the 21 cm signal and the Sloan spectroscopic galaxy sample. Using the extragalactic signals measured from either or both of these surveys, and comparing them with simulations such as those presented here will make it possible to assess the impact of various instrumental imperfections on the Tianlai dish array performance. This would pave the way for future intensity mapping surveys with higher sensitivity.

Key words: instrumentation: interferometers – large-scale structure of Universe – radio lines: galaxies.

1 INTRODUCTION

21 cm Intensity Mapping (IM) is a promising technique to map the cosmological large scale distribution of matter through the observation of 21 cm radio emission/absorption of neutral hydrogen gas (HI), while not requiring the detection of individual sources (Bharadwaj, Nath & Sethi 2001; Battye, Davies & Weller 2004) and has been largely explored in the context of the search for the EoR (Epoch of Reionization) signal (Pritchard & Loeb 2008; Morales & Wyithe 2010). Subsequently, it was suggested that post-EoR 21 cm Intensity Mapping surveys could be used to constrain dark energy through the measurement of the Baryon Acoustic Oscillations (BAO) scale (Chang et al. 2008; Seo et al. 2010; Ansari et al. 2012) in the large scale structure (LSS) distribution, over a broad redshift range ($z \lesssim 6$). Such surveys require instruments with large instantaneous bandwidth and field of view and several groups have built dense interferometric arrays to explore IM, such as CHIME (Bandura et al. 2014) or Tianlai (Chen 2012) in the last decade. Smaller instruments such as PAON4 (Ansari et al. 2020) and BMX (O'Connor et al. 2021) have also been built to explore specific technical aspects of these arrays, as well as transit mode operation and calibration.

★ E-mail: olivier.perdereau@ijclab.in2p3.fr

Table 1. Measured cross-correlations of redshifted H I emission with optical galaxy redshift surveys are listed in the first set of entries. The last two entries describes the proposed surveys analysed in this paper. GBT is the Green Bank Telescope. NCCSz is an ongoing redshift survey described in the text. The significance of the detection could not easily be determined from Pen et al. (2009). Wolz et al. (2022) uses an extended version of the GBT observations of Masui et al. (2013). Both Wolz et al. (2022) and Amiri et al. (2022) correlate separately the same radio data with three different redshift samples. The two Tianlai dish surveys require separate radio observations.

Redshift range	Significance	Radio telescope	Optical redshift survey	Reference
0–0.042	–	Parkes	6dF	Pen et al. (2009)
0.53–1.12	4σ	GBT	DEEP2	Chang et al. (2010)
0.58–1	6σ	GBT	WiggleZ	Masui et al. (2013)
0.057–0.098	5.7σ	Parkes	2df	Anderson et al. (2018)
0.6–1	4.8σ	GBT	WiggleZ	Wolz et al. (2022)
0.6–1	5σ	GBT	eBOSS-ELG	Wolz et al. (2022)
0.6–1	4.4σ	GBT	eBOSS-LRG	Wolz et al. (2022)
0.78–1.00	7.1σ	CHIME	eBOSS-LRG	Amiri et al. (2022)
0.78–1.10	5.7σ	CHIME	eBOSS-ELG	Amiri et al. (2022)
0.80–1.43	11.1σ	CHIME	eBOSS-QSO	Amiri et al. (2022)
0.400–0.459	7.7σ	MeerKAT	WiggleZ	Cunnington et al. (2022)
0–0.068	$\sim 40\sigma$	Tianlai dish	SDSS main sample	this paper (forecasts)
0–0.068	$\sim 15\sigma$	Tianlai dish	NCCSz	this paper (forecasts)

CHIME has proved to be a powerful fast radio burst and pulsar observation machine (The CHIME/FRB Collaboration 2021) and has motivated the design and construction of larger, dish-based, dense interferometric arrays such as HIRAX (Newburgh et al. 2016) and CHORD (Vanderlinde et al. 2019). Non-interferometric surveys have also been considered, such as BINGO (Battye et al. 2016; Wuensche et al. 2022), FAST (Hu et al. 2020) or using the SKA precursor MeerKAT (Wang et al. 2021). Intensity Mapping is also being used to search for signals from the EoR and the cosmic dawn, at redshifts above $z \gtrsim 10$, by several large radio-interferometers, such as LOFAR (van Haarlem et al. 2013), MWA (Tingay et al. 2013), HERA (DeBoer et al. 2017), and LWA (Eastwood et al. 2018). It is also planned to be used with SKA-low (Mondal et al. 2020).

Tianlai is an international collaboration, led by NAOC, which built and operated two radio-interferometers dedicated to 21 cm Intensity Mapping since 2016 (Das et al. 2018). The first instrument is composed of three cylindrical reflectors, equipped with a total of 96 dual-polarization feeds (Li et al. 2020) while the second instrument, the Tianlai Dish Pathfinder Array (hereafter T16DPA) features 16 on-axis dishes, 6 m in diameter, equipped with dual-polarization feeds, and arranged in a near-hexagonal configuration. The two instruments are located in a radio quiet site in Hongliuxia, Balikun county, in the Xinjiang autonomous region, in north-west China. The two arrays have been observing in the frequency band [700, 800] MHz, corresponding to the redshift range $z \sim [0.775, 1.029]$. We recently reported on the various aspects of the operation and performance of the Tianlai Dish Pathfinder Array (Wu et al. 2021).

Detection of the cosmological H I signal and the ability of large instruments to constrain the ΛCDM model, specifically the dark energy equation of state, through IM surveys covering the redshift range $z \lesssim 3$–6 has been extensively explored for large dedicated instruments (Bull et al. 2015; Cosmic Visions 21 cm Collaboration 2018), or with existing or planned general purpose instruments such as SKA (Villaescusa-Navarro, Alonso & Viel 2017; Bacon et al. 2020) and FAST (Smoot & Debono 2017).

Given the challenge of direct observation of the cosmological H I signal, several attempts have been made to detect 21 cm cross-correlations, i.e. 21 cm-induced correlations of low angular resolution radio maps with optical galaxy redshift surveys. Table 1 gives parameters of measurements of these 21 cm cross-correlations, as well as the corresponding references. Until recently cross-correlations have only been detected with large single dish radio telescopes. The first detection of cross-correlations with an interferometric array, an array specifically designed for hydrogen intensity mapping, has been made by CHIME (Amiri et al. 2022).

The two Tianlai pathfinder instruments are also interferometric arrays designed for hydrogen intensity mapping, but are smaller than CHIME. Both, especially the dish array, are too small to be sensitive to the cosmological 21 cm signal around $z \sim 1$. In this paper, we study the extragalactic H I signals that could be detected by the T16DPA by tuning its frequency band to very low redshifts ($z \lesssim 0.1$), through a detailed simulation of the reconstructed signal, taking into account the instrument response and survey strategy. Observation of these extragalactic H I signals can be used to precisely assess the instrument, survey, and data analysis performance, especially the foreground subtraction and mode mixing, as discussed below.

We present an overview of the science targets of the Tianlai Dish Array low-redshift surveys in Section 2, while the simulation and analysis methods common to the different science cases are discussed in Section 3, as well the expected survey sensitivities. Possible direct detection of nearby large H I overdensities, corresponding to galaxies or group of galaxies, referred to as H I clumps in this paper, is presented in Section 4. The prospects for detecting large scale structures at low redshifts ($z \lesssim 0.1$) in cross-correlation with the SDSS and NCCS optical galaxy surveys is discussed in Section 5. Our findings are summarized and further discussed in the last section (section 6).

2 LOW REDSHIFT SURVEYS WITH TIANLAI

The Tianlai dish array reflectors are equipped with feeds having a frequency bandwidth much larger than the instantaneous 100 MHz bandwidth of the digitization and correlator system. The instrument observation band is defined by the analogue RF filters and the local oscillator frequency, which can be easily modified. It is planned to tune the Tianlai Dish Pathfinder Array (T16DPA) frequency band to observe in the [1330, 1430] MHz band, corresponding to the redshift range $0 \lesssim z \lesssim 0.068$.

In addition, the T16DPA dishes are fully steerable, which allows targeted observations, although in drift-scan mode.

The North Celestial Polar cap (NCP), accessible to the Tianlai Dish Array, presents several advantages and is an optimal target to

carry out deep, high sensitivity observations, as suggested in Zhang et al. (2016). A preliminary analysis of long duration observations of the NCP at $z \sim 1$ with T16DPA has also been presented in Wu et al. (2021).

Detection of the cosmological H I signal in the presence of foregrounds, dominated by Galactic synchrotron radiation and radio sources which are several thousand times brighter than the signal, is one of the major challenges of Intensity Mapping surveys. The smooth frequency dependence of the foreground emissions, compared to the cosmological H I signal, is the key feature used to extract the 21 cm signal. However, radio instruments are diffraction-limited, resulting in frequency-dependent beams. In addition, for an interferometer, the set of transverse wave modes sampled by a given baseline varies with frequency. Mode mixing refers broadly to the impact of this frequency-dependent instrument angular response on foreground subtraction.

In addition, imperfect knowledge of key instrument parameters, such as the bandpass response, gain and phase calibration, instrument noise, and array configuration and pointing significantly degrades the overall survey performance. Low redshift surveys can be considered as a pathfinder to prove the effectiveness of a dense interferometric dish array using transit mode observations. As we shall discuss in detail below, there are low-z extragalactic H I signals, with structuring in redshift similar to the cosmological LSS signals expected to be seen at higher redshifts, that will be within the T16DPA sensitivity reach. The observation of these low-z extragalactic signals, and the comparison of their observed strength and spatial and spectral behaviour, with the predictions from simulations such as the ones discussed here, would be a reliable method to assess quantitatively the impact of the above-mentioned instrumental effects on the recovered signal, and the level of the residuals from foreground subtraction. Cross-correlations of 21 cm observations with optical surveys should be less sensitive to foregrounds than direct reconstruction of the extragalactic H I signal. However, as shown in Section 5, robust foreground removal techniques are still necessary to extract this cross-correlation signal.

Possible detection of individual extragalactic sources, galaxies or groups of galaxies is one advantage of low redshift observations with T16DPA, as the signal strength increases for nearby sources as $d_L^{-2}(z)$, where $d_L(z)$ is the luminosity distance to redshift z for a given source. Beyond a redshift of $z \gtrsim 0.2$, the impact of the increased volume on the aggregate emission of sources within a given reconstructed voxel is far from compensating for the decrease in signal strength. Indeed, the T16DPA angular resolution of $0.25^\circ - 0.5^\circ$, translates into a voxel transverse size ranging from ~ 2 Mpc at $z \sim 0.1$ to ~ 10 Mpc at $z \sim 0.5$. The voxel size thus exceeds even the cluster size at redshift 0.5, making direct detection of individual structures (galaxies, clusters) by T16PDA, quite unlikely beyond $z \gtrsim 0.1 - 0.2$, as will be shown in Section 4.

What about statistical detection of LSS through the 3D map autocorrelation power spectrum? The LSS power spectrum changes slowly with redshift, contrary to distances. One might then expect that an IM instrument's ability to measure the LSS power spectrum would not change significantly with redshift. Unfortunately, the sensitivity to observe the cosmological LSS power spectrum decreases sharply as redshift increases, due to the way the radio interferometer's noise projects on sky. Indeed, the noise power spectrum recorded by a radio instrument (either a single dish or an interferometer) projected on the sky, noted $P_{\rm noise}(k)$, scales as $(1+z)^2 d_M^2(z)$, where $d_M(z)$ stands for the transverse comoving distance at redshift z, as shown by Ansari et al. (2012), Bull et al. (2015), and others. This trend is due to the mapping from instrument coordinates, the two angles

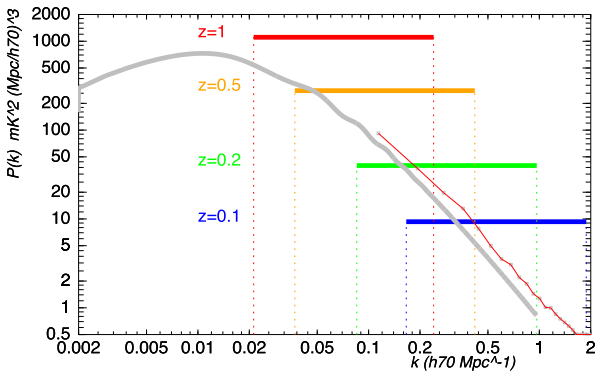

Figure 1. Projected noise power spectrum $P_{\rm noise}(k)$ and the accessible transverse $k_\perp$ range for a survey of the NCP region by T16DPA. The grey line show the expected linear 21 cm power spectrum at redshift $z = 1$, while the red line shows the non-linear power spectrum from simulations (Villaescusa-Navarro et al. 2018) assuming a mean 21 cm brightness temperature $\bar{T}_{21} = 0.136$ mK.

defining a direction on sky and the frequency to a 3D position in a cosmological volume, but does not include any intrinsic noise level variation with frequency. Another derivation of the projected noise redshift dependence is presented in appendix A, and we obtain:

$$P_{\rm noise}(z) \simeq (1+z)^2 d_M^2(z) \frac{c}{H(z)} \frac{\delta\nu}{\nu_{21}} (\delta\theta_0)^2 \times \left(\sigma_0^T\right)^2, \quad (1)$$

where σ_0^T denotes the per pixel noise level and $\delta\theta_0$ the pixel angular resolution for reconstructed maps at $z = 0$.

Despite this steep increase of the noise level with redshift, an intensity mapping survey at $z \gtrsim 1$ would be feasible using an array with several hundred dishes, thanks to the decrease of σ_T^2 or the noise power as the inverse of the number of antenna in the array. A survey of the NCP by T16DPA would be sensitive to spherical harmonics $Y_{\ell,m}$ of the order of ℓ in the range $75 \lesssim \ell \lesssim 850$ at $\nu \sim 1400$ MHz (see Section 3), corresponding to angular scales $2\pi/\ell$.

Taking into account evolution of the instrument angular scale range with redshift ($\ell \propto 1/(1+z)$), we obtain the survey transverse wavenumber sensitivity range:

$$k_\perp(z) = \frac{\ell(z)}{d_M(z)} \quad (2)$$

$$\ell^{\rm min}(z=0) \simeq 75 \qquad \ell^{\rm max}(z=0) \simeq 850 \quad (3)$$

$$k_\perp^{\rm min,max} = \frac{1}{(1+z)\, d_M(z)} \times \ell^{\rm min,max}(z=0) \quad (4)$$

We have gathered in Table 2 the voxel dimensions, and the accessible transverse $k_\perp$ range for a survey with angular scale sensitivities similar to T16DPA, map pixels with angular size 0.2° at $\nu \sim 1400$ MHz and frequency resolution 1 MHz. The projected noise level as a function of redshift is shown in Fig. 1 as well as the accessible transverse $k_\perp$ range for a T16DPA survey. The maximum value of the radial wavenumber $k_\parallel^{\rm max}$ is also listed assuming voxels with $\delta\nu = 1$ MHz resolution. However, the T16DPA correlator computes visibilities with $\simeq 244$ kHz frequency resolution, so the survey could reach a maximum $k_\parallel$ four times higher than the values listed in the table. Many foreground subtraction methods rely on the smoothness of synchrotron emission with frequency and thus remove the signal modes with low $k_\parallel$. This is also true for e.g. PCA (Liu & Shaw 2020), where the original smoothness in synchrotron radiation is broken in the data by the non-smooth instrument response, but a correlation over different frequencies is retained. The simulations

Table 2. Listed as a function of redshift are the comoving distance d_M, voxel size (radial ($a_\parallel$), and transverse ($a_\perp$) sizes, range of wavenumbers sampled and per voxel noise assuming 1 MHz pixels and per pixel noise of $\sigma_T^2 = 1\,\mathrm{mK}^2$. Comoving lengths are in units of $\mathrm{Mpc/h_{70}}$, comoving wavenumbers in units of $\mathrm{h_{70}Mpc^{-1}}$ and white noise power $P_{\rm noise}$ in units of $\mathrm{mK^2/(Mpc/h_{70})^3}$. See the text for lower cut-off on $k_\parallel$ induced by foreground subtraction.

z	d_M	$a_\perp$	$a_\parallel$	$k_\perp^{\rm min}$	$k_\perp^{\rm max}$	$k_\parallel^{\rm max}$	$P_{\rm noise}$
0.1	451	1.7	3.7	0.16	1.9	0.85	10
0.2	880	3.7	4.2	0.08	0.96	0.75	40
0.5	2028	10.6	5.5	0.037	0.42	0.57	280
1.0	3536	24.7	7.3	0.021	0.24	0.43	1100
2.0	5521	57.8	9.7	0.013	0.15	0.32	3600

we have carried out here suggest a low cut-off value $k_\parallel^{\rm min} \sim 0.15 k_\parallel^{\rm max}$ (see Section 3.4).

The left-hand panel of Fig. 2 shows the radio sky near the NCP (North Celestial Pole) at 1350 MHz, as synthesized by our foreground model, described in Section 3.1, which includes diffuse Galactic synchrotron and radio sources. The map has been smoothed by a Gaussian filter to enhance rendering and ease the comparison with the reconstructed map. We highlighted the brightest radio sources in this field. Their characteristics, retrieved from on-line archives such as NED[1] or with Simbad[2] and VizieR[3] are summarized in Table 3. This field has also been observed by large scale radio interferometric arrays such as 21CMA (Zheng et al. 2016) and LOFAR (Yatawatta et al. 2013).

The visibility simulation and 3D map reconstruction is briefly described in Section 3, as well as the foreground subtraction methods we have used. Two surveys are studied here as described in Section 3.1: a deep survey covering $\sim 150\,\mathrm{deg}^2$ around the NCP, and a mid latitude survey covering a $\sim 12\,\mathrm{deg}$ band in declination around $\delta = 50^\circ$. We will show in Section 4 that it is possible to detect individual galaxies or groups of galaxies at very low redshifts ($z \lesssim 0.05$) in the NCP region. We have also studied the statistical detection of the LSS through cross-correlation with optical surveys, as discussed in Section 5. A mid latitude survey, covering a larger area, would be less sensitive due to higher noise level, but even more so due to much larger residuals from imperfect foreground subtraction, as discussed in Section 3. However, thanks to the larger sky area, it would be possible to detect the cross-correlation signal with high statistical significance in both surveys.

3 PLANNED SURVEYS, SIMULATION, AND ANALYSIS

In this section, we first give an overview of the simulations performed in this work, and evaluate the performance of the critical steps of the analysis. First, in Section 3.1, we describe the simulation parameters and tools that have been used. Then, the map reconstruction method and its response in the spherical harmonic domain is presented in the Section 3.2. In subsequent analyses we have used two foreground removal techniques, which are compared in Section 3.3. Finally, the noise levels of the simulated maps are analysed in Section 3.4.

3.1 Overview of the simulations

The JSKYMAP[4] package has been used for computing visibilities for the Tianlai dish array and the survey strategies studied in this paper. The package also provides several tools for reconstructing maps from transit visibilities. Here, we have used the m-mode visibility computation and map making tools, which operate in the spherical harmonics space $Y_{\ell,m}$ as described in Zhang et al. (2016) and Shaw et al. (2015). The simulation and analysis pipeline includes several other C++ or python software modules, which handle the preparation of the input data, such as the generation of H I sources from optical catalogues, foreground subtraction, source detection, power spectrum computation, and optical radio cross-correlation computation.

The study presented here uses only intensity maps, ignoring polarization. Foregrounds have been modelled through the co-addition of the diffuse synchrotron emission, represented by the reprocessed Haslam map at 408 MHz (Remazeilles et al. 2015) and the radio sources from the NVSS catalogue (Condon et al. 1998). In practice, for each simulated observation frequency, diffuse synchrotron emission and radio-sources have been extrapolated from their reference frequencies, (408 and 1400 MHz) using a constant spectral index β for the diffuse component. Changing the spectral index in the range $\beta \sim -2. \ldots -2.5$ has negligible impact on the the results discussed in this paper. We have used the rather pessimistic $\beta = -2$ value which maximizes the foreground level. All sources with flux larger than 0.05 Jy and $\delta > 15^\circ$ have been included in the simulation. Other more complete foreground simulation approaches, using models as the GSM (Shaw et al. 2015) or the Planck Sky Model (Delabrouille et al. 2013) as in Alonso, Ferreira & Santos (2014) are beyond the scope of this work.

The array configuration used in the simulations corresponds to the actual positions of the Tianlai antennas. We have used a frequency dependent single dish beam pattern $B(\theta)$, with azimuthal symmetry and modelled as an Airy disc with an effective dish diameter $D_{\rm eff} = 5.6\,\mathrm{m}$.

$$B(\theta) \propto \left(\frac{2J_1(x)}{x}\right)^2 \qquad x = 2\pi \frac{D_{\rm eff}}{2\lambda} \sin\theta, \tag{5}$$

where J_1 is the order one Bessel function of the first kind, λ is the wavelength, and θ is the angle with respect to the dish axis.

There are 120 different baselines, excluding autocorrelations and ignoring polarization. Visibilities have been computed with a right ascension or time sampling of $\delta\alpha = 30\,\mathrm{s}$, well below the array angular resolution $0.25^\circ - 0.5^\circ$, and we have used a $\delta\nu = 1\,\mathrm{MHz}$ frequency resolution. Two surveys have been studied here, spanning a total duration of several months, up to a year. Their footprints are highlighted in Fig. 3.

(i) A survey of the NCP region with four constant declination scans $\delta = 90^\circ, 88^\circ, 86^\circ, 84^\circ$, and covering an area of about $100\,\mathrm{deg}^2$ around the north pole. We have used a fiducial area within 7 degree from the north pole, $\delta > 83^\circ$, which would yield a surveyed area $\sim 150\,\mathrm{deg}^2$.

(ii) A survey in a mid-latitude area, covering a much larger portion of sky, using six constant declination scans at $\delta = 49^\circ$, 51°, 53°, 55°, 57°, 59°, covering a 12° band in declination $48^\circ \leq \delta \leq 60^\circ$, representing about ~ 12 per cent of the sky or $\sim 2500\,\mathrm{deg}^2$. However, we have excluded a region in right ascension contaminated by the galactic plane and bright sources such as Cas A and Cyg A

[1] https://ned.ipac.caltech.edu/
[2] http://simbad.u-strasbg.fr/simbad/
[3] https://vizier.u-strasbg.fr/

[4] https://gitlab.in2p3.fr/SCosmoTools/JSkyMap (also check the wiki pages)

Table 3. Main characteristics of the brightest radio sources in the NCP field from Fig. 2. (†: at 15 arcsec distance).

NVSS id	Other id	RA (hms)	Dec (dms)	21cm flux (Jy)	Object type	Redshift
J011732+892848	6C 004713+891245 †	1h17m32.82s	89d28'48.7"	2.1	–	–
J022248+861851	3C 61.1	2h22m35.046s	86d19'6.17"	6	Seyfert 2 gal.	0.18781
J093923+831526	3C 220.3	9h39m23.40s	83d15'26.2"	2.95	AGN	0.685
J213008+835730	3C 345.1	21h30m8.60s	83d57'30.5"	1.8	radio gal.	0.865

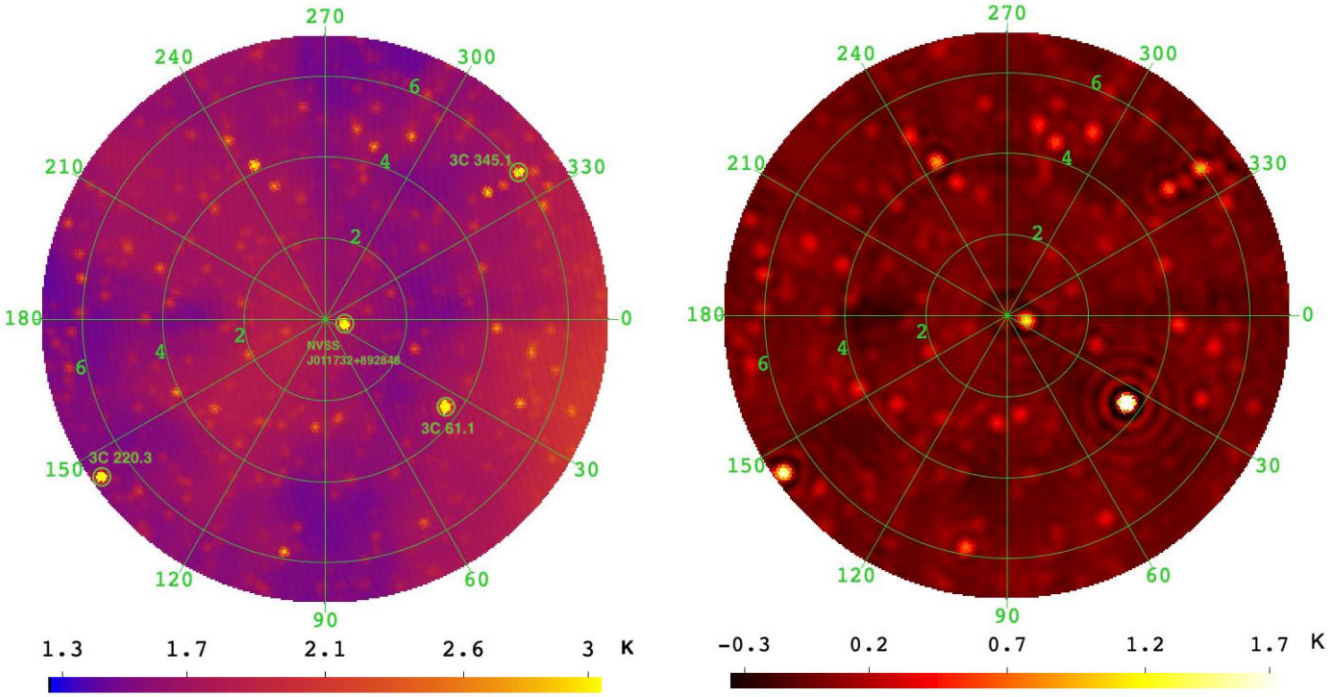

Figure 2. Left-hand panel: foreground map of a circular region of 7 degree radius around the NCP at 1350 MHz, smoothed with a 15 arcmin resolution Gaussian beam. The Haslam map of diffuse emission at 408 MHz as well as NVSS radio sources, extrapolated to 1350 MHz with a spectral index $\beta = -2$, have been co-added. The four brightest of these sources sources, reported in Table 3, have been identified. Right-hand panel: reconstructed map of the NCP region, as observed by T16DPA at $f = 1350$ MHz. This is the 7° radius area around $\delta = 90°$, extracted from the reconstructed spherical map using m-mode map making and after (ℓ, m) space filtering. The colour scale corresponds to temperature in Kelvin.

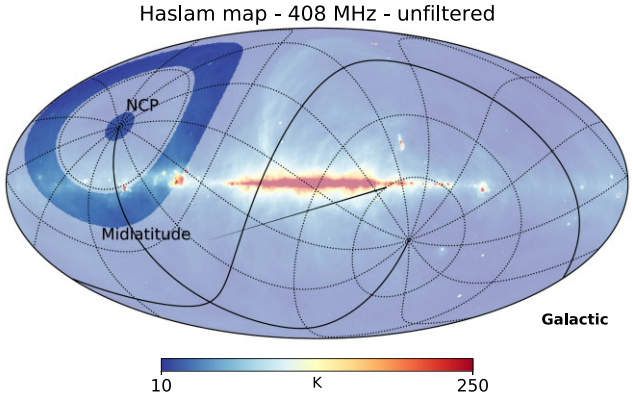

Figure 3. Footprints of the two surveys examined in this work, one around the NCP and the other in a mid-latitude band, are highlighted over the Haslam 408 MHz unfiltered map from Remazeilles et al. (2015), plotted in galactic coordinates with equatorial coordinate lines.

when computing noise power spectrum and mode mixing residuals. The fiducial area used, $40° < \alpha < 260°$, represents about 1500 deg^2.

The T16DPA system noise temperature has been determined to be $T_{\rm sys} \sim 80$ K (Wu et al. 2021). The simulations performed in this study have been carried out with a fiducial noise level of 5 mK per $\delta t = 30$ s visibility sample, and for a $\delta\nu = 1$ MHz frequency band. Such a noise level should indeed be reached for a single linear polarization after 8.5 d spent on each constant declination scan, corresponding to a total integration time $t_{\rm int} = 8.5 \times 30 = 255$ s, per $\delta\alpha = 30$ s RA sampled visibilities, leading to a per polarization noise level :

$$\sigma_{V_{ij}} = \frac{T_{\rm sys}}{\sqrt{t_{\rm int}\delta\nu}} = \frac{80\,{\rm K}}{\sqrt{255 \times 10^3}} \simeq 5\,{\rm mK}.$$

As shown in Wu et al. (2021), the Tianlai dish array daytime data are contaminated by the Sun signal leaking into the far side lobes. It is therefore planned to use only nighttime data. Taking into account the $\sqrt{2}$ gain expected in the noise level when combining the two orthogonal linear polarization components, T16DPA should be able to reach a noise level of 5 mK per RA visibility sample by surveying the NCP region during two periods of one month each, separated by 6 months, to get nighttime coverage of the full right ascension range. A noise level of 2.5 mK would also be reachable by observing the NCP area over a year, spending a month on each declination. A similar noise level reduction could be achieved for the mid-latitude survey.

However, as will be shown in Section 3.4, the survey sensitivity is limited by larger mode mixing residuals in this case.

3.2 Map reconstruction

For each frequency, a spherical map is reconstructed through m-mode map making. In this method, the linear system of equations relating observed visibilities to the unknown sky is solved in spherical harmonics $a_{\ell,m}$ space (Shaw et al. 2015; Zhang et al. 2016). Writing the map making equations in spherical geometry handles naturally many complications which arise in the flat space approximation used usually in radio interferometry and is suited for instruments with large field of view. One would have to deal with very large systems, with more than 10^6 unknowns and $10^6 - 10^8$ equations for typical IM instruments. However, for instruments observing in transit mode and covering the full 24 h in RA, it can be shown that by projecting the equations over the Fourier modes corresponding to the right ascension direction, called m-modes, this large linear system can be decomposed into $m_{\max} \gtrsim 1000$ smaller independent systems, making the numerical solution tractable.

The resulting system is most often underdetermined and a regularization scheme has to be used to solve for the unknown sky. A pseudo-inverse method is used in the JSKYMAP package, for which the numerical stability as well as the noise level are controlled through two parameters: $r_{\rm PSI}$, which determines the ratio of the smallest to largest eigenvalue retained for each inversion, and $\lambda_{\rm PSI}$, the absolute threshold on the minimal eigenvalue. The values of these parameters have been set to the medium level of $r_{\rm PSI} = 0.02$ and $\lambda_{\rm PSI} = 0.001$ for the analysis presented here.

We have used spherical maps with a resolution of 5 arcmin, although the array angular resolution is closer to $10 - 15$ arcmin, as stated in Section 2. The reconstructed maps' pixels thus have a certain degree of redundancy, with pixel-to-pixel noise values being correlated with each other for neighbouring pixels. However these higher resolution maps presented a slight advantage for source detection and foreground removal. We have used the `SphereThetaPhi` pixelization scheme, which features almost square and equal area pixels along θ, ϕ directions, implemented in the SOPHYA[5] library, instead of the more frequently used HEALPIX scheme.[6] This `SphereThetaPhi` scheme, belonging to the IGLOO pixelizations (Crittenden 2000), preserves to some extent the symmetry around a pixel located exactly at the pole $\theta = 0$ and has also the advantage of being fully flexible in terms of angular resolution or pixel size. We also apply a filter in the spherical harmonics space $a_{\ell,m}$, before map reconstruction and foreground subtraction.

The quality of the reconstruction degrades at the two ends of the T16DPA ℓ sensitivity range. At low ℓ, this is explained by the absence of the autocorrelation signal, which is not used in map reconstruction, and the minimal baseline length, about 8.8 m, which limits the sensitivity below $\ell \lesssim 75$ for the NCP survey. At the other end, the noise level increases for $\ell \gtrsim 850$, which corresponds to the angular resolution of the array size for the NCP survey. We have therefore smoothly damped $a_{\ell m}$ coefficients for $\ell \lesssim 75$ and $\ell \gtrsim 875$. A Gaussian filter with $\sigma_\ell = 750$ has also been applied. Moreover, all $m = 0$ modes have been set to zero; this is intended to remove wiggles with near perfect azimuthal symmetry which appears due to the partial sky coverage combined with limited sensitivity range in ℓ. Fig. 2, right-hand panel shows an example of a reconstructed map,

[5]SOPHYA C++ class library http://www.sophya.org
[6]https://healpix.sourceforge.io/

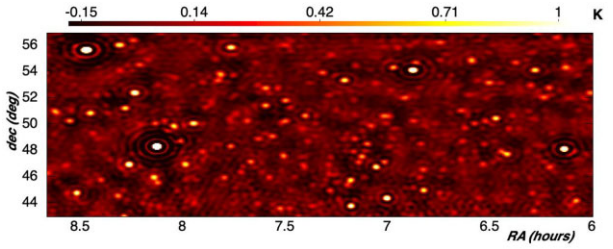

Figure 4. Reconstructed map of the mid-latitude region, after (ℓ, m) filtering, as observed by T16DPA at $f = 1350$ MHz. The patch of sky shown covers the declination range $43^\circ < \delta < 57^\circ$ and the right ascension range $90^\circ < \alpha < 130^\circ$ (6 to 8.66 h), with 5 arcmin pixel size.

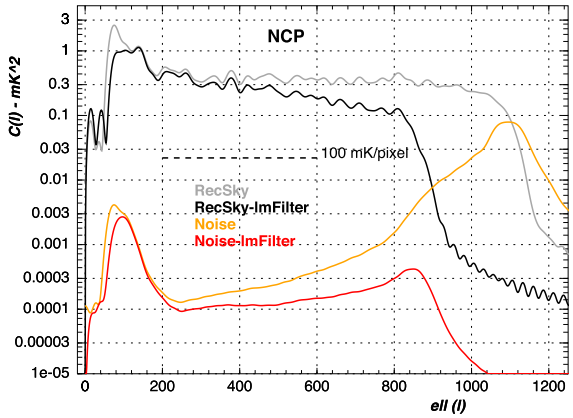

Figure 5. Average angular power spectrum $C(\ell)$ from a data cube of 100 reconstructed maps covering an area with 7° radius around the NCP and frequency range 1300–1400 MHz. The reconstructed sky power spectrum, as well as the noise power spectrum, are shown before (lighter colours) and after (ℓ, m) domain filtering (darker colours). The grey and black curves represent the reconstructed sky power spectrum, before and after filtering, respectively, while the orange and red curves show the power spectrum of maps reconstructed from noise only visibilities.

after (ℓ, m) space filtering at a frequency of $f = 1350$ MHz. Sources present in the true sky map (Fig. 2), as well as larger structures, are clearly visible, while the noise level $(2 - 4$ mK) is too low to be noticeable. Some artefacts, such as rings around bright sources, can easily be seen and are due to incomplete (ℓ, m) plane coverage and filtering. Similarly, a patch of reconstructed sky from a T16DPA survey of the mid-latitude area is shown in Fig. 4. This survey has a significantly higher noise level, compared to the NCP case, which is however not noticeable on this reconstructed map, as the brightest sources reach a few K.

The angular power spectra and noise of the reconstructed sky of the NCP region are shown in Fig. 5. The sky power spectrum is higher at larger angular scales, with an overall level of about ~ 1 K. The effects of the instrument and map-making response in the ℓ space are visible at the two ends, $\ell \lesssim 50$ and $\ell \gtrsim 1150$, of the unfiltered spectrum (grey curve). The additional effect of the (ℓ, m) domain filtering can clearly be seen by comparing the sky power spectrum before (grey curve) and after (black curve) this filtering. The projected noise angular power spectrum $C_{\rm noise}(\ell)$ is also shown in this figure. These $C_{\rm noise}(\ell)$ curves have been computed from maps reconstructed from white noise-only visibilities, with an RMS fluctuation level of 5 mK per $\delta\alpha = 30$ s visibility sample. As expected, the noise spectrum increases significantly towards the high-

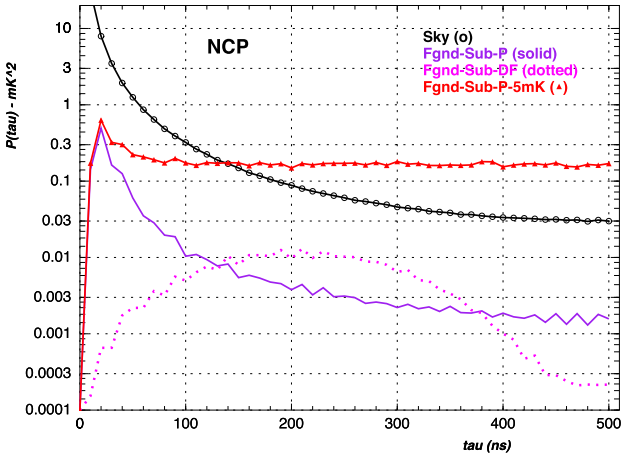

Figure 6. Average power spectrum along the frequency axis $P(\tau)$ from the a data cube of 100 reconstructed maps, with (ℓ, m) filtering of the NCP region. The horizontal axis correspond to the delay parameter τ, ranging from 0 to 500 ns. The reconstructed sky power spectrum is shown in black, while the purple (solid) and magenta (dotted) curves show the residuals after foreground subtraction, for the polynomial fit (P) and difference along the frequency (DF) methods, respectively. The red curve correspond to the foreground subtracted map residuals, including noise and using polynomial fit (P).

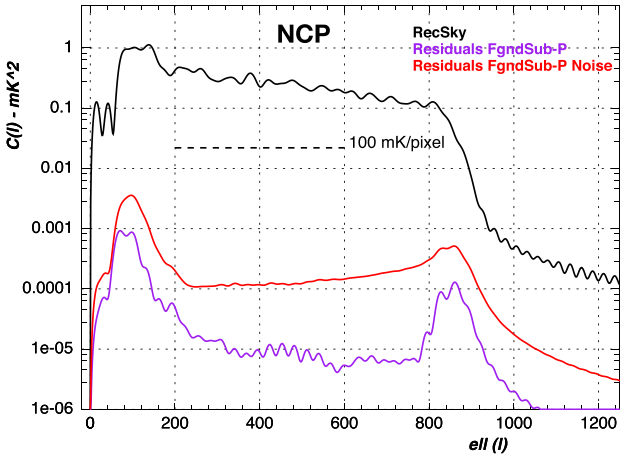

Figure 7. Spectra similar as those shown in Fig. 5, but now with foreground cleaning. The reconstructed sky power spectrum is shown in black, the residual after foreground subtraction (polynomial fit-P) in purple without noise, and in red, with noise.

ℓ end of the spectral sensitivity range, above $\ell \gtrsim 800$. This is due first to the decrease of the baselines' redundancy with ℓ, and second, to the incomplete coverage of wave modes in the (ℓ, m) domain at the high-ℓ end. The effect of (ℓ, m) filtering on the noise power spectrum $C_{\rm noise}(\ell)$ can be seen by comparing the orange curve, obtained before filtering, with the red curve, after filtering.

3.3 Foreground subtraction

Two simple foreground subtraction methods have been used here which exploit the smooth frequency dependence of the synchrotron-dominated foreground. The first method (**P**-Polynomial), similar to the one in Ansari et al. (2012), represents the synchrotron emission frequency dependence as a second degree polynomial in frequency. The coefficients are determined for each direction through a linear χ^2 fit to the measured temperatures, and the resulting fitted foreground $T^{\rm fgnd-P}_{\alpha,\delta}(\nu)$ is then subtracted from the 3D temperature map:

$$T^{\rm fgnd-P}_{\alpha,\delta}(\nu) = A_{(\alpha,\delta)}\,\nu^2 + B_{(\alpha,\delta)}\,\nu + C_{(\alpha,\delta)} \tag{6}$$

$$T^{\mathbf{P}}(\alpha,\delta,\nu) = T(\alpha,\delta,\nu) - T^{\rm fgnd-P}_{\alpha,\delta}(\nu) \tag{7}$$

The second method (**DF**) is a **D**ifference filter along **F**requency. For each frequency ν, we subtract the average of two nearby frequencies ν_-, ν_+, with a specified frequency gap $\Delta\nu$; $\nu_- = \nu - \Delta\nu$ and $\nu_+ = \nu + \Delta\nu$:

$$T^{\rm fgnd-DF}_{\alpha,\delta}(\nu) = \frac{1}{2}\left(T(\alpha,\delta,\nu_-) + T(\alpha,\delta,\nu_+)\right) \tag{8}$$

$$T^{\mathbf{DF}}(\alpha,\delta,\nu) = T(\alpha,\delta,\nu) - T^{\rm fgnd-DF}_{\alpha,\delta}(\nu) \tag{9}$$

We have used $\Delta\nu = 2$ MHz throughout this paper.

Fig. 6 shows the average delay power spectra (see e.g. Parsons & Backer 2009) for the sky and residual after foreground subtraction $P(\tau)$ for the NCP survey. They correspond to the average of the spectra along the frequency direction, obtained through a radial Fast Fourier Transform (FFT), for each direction of the sky. The Fourier modes along the frequency are labelled as τ and correspond to a lag or delay time. Given the 100 MHz bandwidth, with 100 frequency planes, the frequency modes or delay cover the range from 10 ns $\leq \tau \leq$ 500 ns. The black curve represents the average reconstructed sky $P_{\rm sky}(\tau)$, with the power highly concentrated at very low delay modes ($\tau \leq 20 - 30$ ns), but with still significant power up to $\tau \lesssim 100$ ns. The effect of the two foreground subtraction methods and their τ-response can be understood by looking at the shape of the average delay-spectrum of the residual maps without noise (purple and magenta curves). It should be noted that no H I signal is included in the simulations for figures discussed in this subsection, which illustrate the foreground subtraction performance and the corresponding level of residuals.

It can be seen that the polynomial foreground subtraction (P) suppresses delay-modes below $\tau \lesssim 50 - 60$ ns, while the differential filter along frequency (DF) can be considered as a band pass filter, removing $\tau \lesssim 100$ ns and $\tau \gtrsim 400$ ns. The (DF) method is more effective at removing foreground modes at low delay, but leads to noisier maps. In addition to removing low-delay modes, the polynomial subtraction method (P) damps the power $P(\tau)$ by a factor about 30 for all modes above 100 ns. The red curve correspond to the power spectrum $P(\tau)$ from maps after foreground subtraction (P method) with noise included, and it is noise dominated.

Fig. 7 presents the average angular power spectrum of the residual signal $C_{\rm res}(\ell)$ of a set of 100 sky maps after foreground subtraction by the polynomial fit (P) method for the NCP survey. The maps are reconstructed from mock visibilities, for four constant declination scans at or near the NCP, and the (ℓ, m) plane filtering have been applied. Compared to the input sky angular power spectrum $C_{\rm sky}(\ell)$, shown as the black curve, one can see that the foreground angular power spectrum is suppressed by a factor $\gtrsim 20\,000$ for the polynomial subtracted foreground (P). This suppression factor reaches $\gtrsim 60\,000$ for the DF method (not shown). These values correspond to a factor $\sim$150 (P) and $\sim$250 (DF) damping in the amplitude for temperature fluctuations due to foregrounds. While this might not be sufficient for the direct detection of the cosmological 21 cm signal, the foreground residuals due to mode mixing and imperfect subtraction would be well below the instrumental noise level for the NCP survey by Tianlai.

However, T16DPA becomes less efficient for mitigating mode mixing for a mid-latitude survey. The angular power spectrum of the residual after foreground subtraction using the DF method, for a mid-

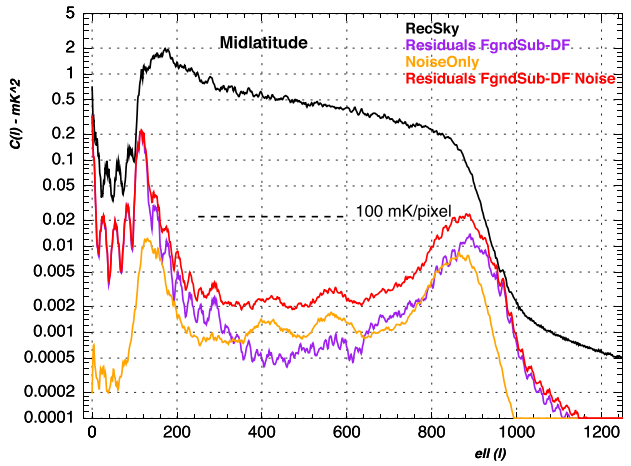

Figure 8. Angular power spectra $C(\ell)$, for sky, noise, and residuals after foreground subtraction for the mid-latitude survey, computed from a set of three frequency maps at 1348, 1350, 1352 MHz. The differential filter along the frequency (DF) foreground subtraction method has been used here on (ℓ, m)-space filtered maps. The black curve correspond to the reconstructed sky power spectrum, while the purple and red curves show the power spectra after foreground subtraction, without and with noise, respectively. The orange curve corresponds to the power spectrum of the maps reconstructed from noise-only visibilities.

latitude survey, is shown in Fig. 8. A set of three reconstructed maps at $f_{-2} = 1348$ MHz, $f_0 = 1350$ MHz, and $f_{+2} = 1352$ MHz have been used to compute these power spectra. A fiducial area, representing $\sim 1500\,\mathrm{deg}^2$, has also been used to exclude the right ascension ranges contaminated by the Galactic plane and Cas A and Cyg A. Comparing the black curve, which represents the reconstructed $C_{\rm sky}(\ell)$ of the diffuse synchrotron and radio sources, and the purple curve $C_{\rm res}(\ell)$, corresponding to the residuals after foreground subtraction, we see that the $C(\ell)$ power spectrum has been damped by a factor $\sim$1200, or about $\sim$35 for the amplitude of the temperature fluctuations. The residual after foreground subtraction reaches a level of ~ 15 mK, similar to the noise contribution.

The fact that this damping factor is seven times lower in amplitude (about 50 times in the power spectrum) for the mid-latitude case compared to the NCP case is explained by a higher level of mode mixing for a mid-latitude survey by Tianlai, compared to the NCP case. Indeed, for observations toward the NCP, the sky orientation of the projected baselines changes with the sidereal rotation, improving the map making performance in terms of individual mode reconstruction. The circular configuration of T16DPA was optimized for a good coverage of the angular sky modes or the (u, v) plane, minimizing the number of redundant baselines compared to a regular rectangular grid configuration, for example. Although arrays with redundant baselines offer advantages for the gain and phase calibration, they would exhibit higher levels of mode mixing. For very large arrays, with several hundred or several thousand elements, a combination of redundant and non redundant baselines should be used to mitigate both mode mixing and calibration issues.

3.4 Noise level and survey sensitivity

The left-hand panel of Fig. 9 shows an example of a noise map for the Tianlai NCP survey, while the histogram of the corresponding pixel value distribution is shown in the right-hand panel. The $((5\,\mathrm{arcmin})^2 \times 1\,\mathrm{MHz})$ pixel-to-pixel temperature fluctuation is close to 4 mK, and even 2.2 mK for the central 3° radius area, assuming a 5 mK noise level per $\delta\alpha = 30$ s visibility sample. Fig. 10 shows a similar noise map for the mid-latitude survey, with an RMS pixel fluctuation level of ~ 16 mK. The noise level scales slightly faster than the square root of the ratio of the surveyed sky area, $(2500\,\mathrm{deg}^2/150\,\mathrm{deg}^2 \simeq 16)$, as the mid latitude survey discussed here requires six constant declination scans, hence 50 per cent more observing time.

As mentioned already, the maps with 5 arcmin pixels used here have a higher resolution than the effective instrument and reconstruction angular resolution, which is limited to $\ell^{\rm max} \sim 850$ or 12 arcmin. The noise correlation between neighbouring pixels is visible on the noise maps. The RMS fluctuation level decreases by a factor 1.5 if maps with 0.25° or 15 arcmin pixels are used. For the NCP survey the noise power spectrum $C_{\rm noise}(\ell)$ is nearly flat for $200 < \ell < 800$, and 15–40 times higher than the foreground subtraction residuals, as shown in Fig. 7. For the NCP region, Tianlai should be able to get down to the regime where the foreground residue is suppressed to a level below that of the noise even for a deep survey that would reach ~ 2 mK per pixel noise level.

We have assumed a perfect knowledge of the instrument response, individual antenna beams pattern, and perfect gain and phase calibration. Discussion of the impact of an imperfect knowledge of the instrument response on the survey performance is beyond the scope of this paper. Preliminary studies suggest however that a phase calibration error between baselines, following a zero-mean normal distribution with an RMS of 7°, would result in a substantial increase of the foreground residual power spectrum, reaching a level 10 times higher than the one due to the instrument noise for the NCP case.

4 H I CLUMP DETECTION

The aim of this analysis is to assess the number of direct detections of H I clumps in a low-z survey of either a mid-latitude band or a circular region around the North Celestial Pole with T16DPA. We estimate the H I clumps detection efficiency as a function of their flux for the NCP and mid-latitude surveys in 4.1, simulating an artificial clump population with a fixed given flux and placed at random positions. In 4.2 we combine these detection efficiencies with the 21 cm flux distribution derived from the ALFALFA[7] H I clump mass function (Jones et al. 2018) to determine the expected number of detected clumps, assuming a spatially uniform random distribution, for the NCP and mid-latitude cases.

4.1 H I clumps detection efficiencies

To assess the detection efficiency for point-like H I sources we have used a pipeline sharing most of the components described in Section 3.1. We simulate observations of the NCP and mid-latitude surveys as described there, for only three frequencies: 1348, 1350, and 1342 MHz.

To the generic astrophysical components (diffuse synchrotron Galactic emission, and continuum NVSS sources) we add, for the central frequency only, a set of spatially uniformly distributed point-like sources of a given fixed flux (in Jy). For each frequency we compute simulated visibilities, and then reconstruct sky maps and apply an angular mode filter, as explained in 3.1. The noise level per visibility sample (30 s integration time) used in the following is 5 mK. In order to account for the impact of foregrounds, we

[7] http://egg.astro.cornell.edu/alfalfa/index.php

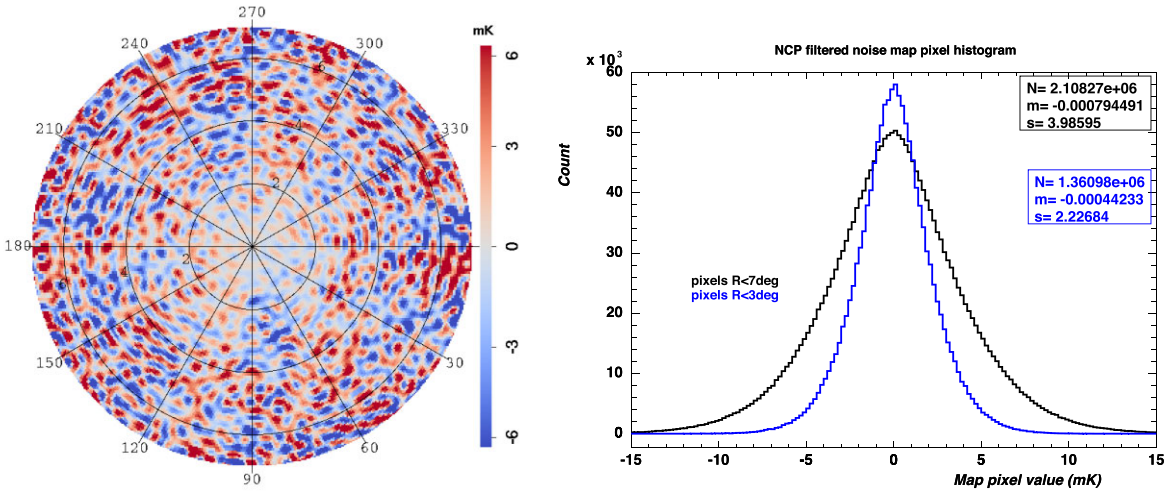

Figure 9. Left-hand panel: Noise map after reconstruction with (ℓ, m) filtering of the NCP region covering an area with 7 degree radius at $f = 1350$MHz (map scale in mK). Right-hand panel: Noise map pixel value distribution, in black for the full 7 degree radius map around NCP, and blue, restricted to the central 3 degree radius, covering $\sim 30\text{deg}^2$. Note that restricted area corresponding to the blue histogram correspond to $\sim$ 18 per cent of the full area; The blue histogram has been rescaled to enhance the figure readability.

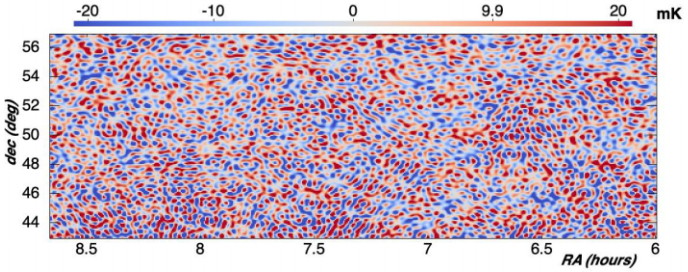

Figure 10. Noise map of the mid-latitude region, corresponding to the reconstructed map shown in Fig. 4.

have used the difference filter (**DF**) along the frequency direction described in Section 3.4, which corresponds to subtracting from the central frequency map the average of the two outer frequency ones. The foreground removal residuals analysed in Section 3.3 are therefore one of the limitations to the detection of additional H I sources. Finally, we reproject the resulting difference sky maps into rectangular (mid-latitude case) or square (NCP case) maps. An additional high-pass filter in the angular modes domain has been applied in the mid-latitude case to reduce foreground subtraction residuals.

The final step of the pipeline is the source detection. We use a basic scheme based on the `DAOStarFinder` class from the `photutils` Python package (Bradley et al. 2021). Loose sphericity criteria for this source detector have been set, to compensate for remaining artefacts due to map reconstruction and foreground subtraction. We set the detection threshold, expressed as multiples of the map pixel-to-pixel RMS fluctuation level, to 7 and 10 for the NCP and mid-latitude cases, respectively, to avoid spurious detections. For the same reason, the detection efficiency is determined by the number of sources detected within 2 pixels of the corresponding simulated positions. In the NCP case, we simulated for each flux value 5 sources over the 7 degree diameter circular observed region, but repeated this operation 20 times to reach a statistical accuracy of a few per cent. In the mid-latitude case, the rectangular surveyed area is much larger, so that more sources are simulated in each run and this iteration is needed only 2 or 3 times.

The detection efficiencies we measure in the NCP and mid-latitude simulations are reported in Fig. 11. Thanks to the higher integration time per map pixel in the NCP case, these results show that the detection threshold $S_*^{\rm th}$, defined as the flux limit with a detection efficiency ≥ 50 per cent, is much lower in the NCP case than the mid-latitude one: $S_*^{\rm th} \simeq 0.08$ Jy for the NCP case, compared to $S_*^{\rm th} \simeq 0.9$ Jy for the mid-latitude case. With the detection algorithm's settings and selection described above, less than 5 per cent of the detected sources are considered as accidentals, as shown by the green dots in Fig. 11. We have in each case fitted the source detection efficiency as a function of the flux to an error function $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \text{d}t$. These fitted functions have then been used in the computation of the expected number of H I clump direct detections in the low-z Tianlai surveys. We recall that the noise per visibility sample level of 5 mK used throughout this paper is a conservative estimate, as it could be achieved in only 3 months of observations, split into two halves separated by 6 month to avoid the sun-contaminated daytime, as explained in Section 3.1. We also indicate in Fig. 11 an estimation of the efficiency curve in this lower noise hypothesis in the NCP case, for which instrumental noise is the main limitation.

4.2 Number of expected H I clumps observations

As shown in Villaescusa-Navarro et al. (2018) (their section 4), in the redshift range considered in our analysis, most of the H I mass lies

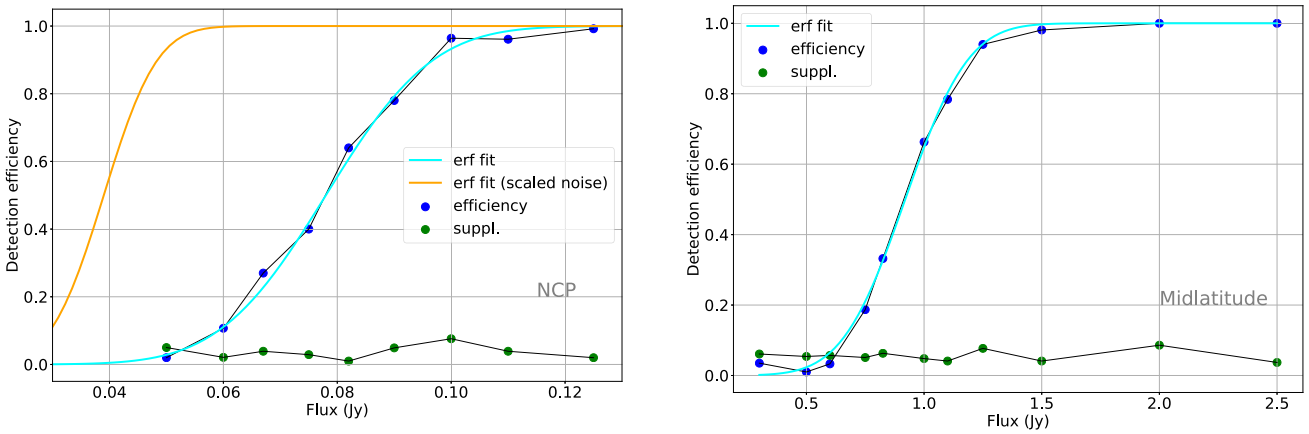

Figure 11. H I clump detection efficiency as a function of flux, for the NCP (left-hand panel) and mid-latitude (right-hand side) measured by our simulations. On each part we represent as blue dots the efficiency measured at each simulated flux. The green dots correspond to the number of spurious detection (detections located farther than 2 pixels from the simulated clumps positions). The cyan curve is a fit of the efficiencies values with an error function. In the NCP case (left-hand panel) we also indicate, in orange, what the efficiency function would become if the noise per visibility sample was decreased by a factor ~ 2 e.g. thanks to a longer integration time per declination.

Table 4. H I mass function parameters determined using either the whole ALFALFA data set and its near subset.

Data set	α	m_*	ϕ_*
Full	-1.25 ± 0.02	9.94 ± 0.01	0.0045 ±0.0002
Near	-1.22 ± 0.02	9.76 ± 0.04	0.0062 ±0.0005

Table 5. Number of expected H I clump discoveries per square degree for the NCP and mid-latitude surveys, for the two parametrizations of the H I mass function given in Table 4, estimated with the fitted efficiency curves shown in Fig. 11. The sky area covered for each survey is given, as well as the total number of expected detections. In the NCP case, we also indicate our estimations in the hypothesis of a lower noise per visibility sample (or longer integration time), using the corresponding efficiency curve.

Data set	NCP	NCP (low noise)	mid-lat.
Surface (deg^2)	150	150	1500
clumps deg^{-2} (full MF)	0.048	0.113	0.0012
clumps deg^{-2} (near MF)	0.035	0.083	0.0009
N-clumps (full MF)	7.2	16.95	1.8
N-clumps (near MF)	5.25	12.45	1.35

Table 6. Frequency bins used for the mid-latitude Tianlai-SDSS cross-correlation analyses.

Bin	$\nu_{\rm min}$ (GHz)	$\nu_{\rm max}$ (GHz)	$z_{\rm centre}$
1	1260	1310	0.096
2	1310	1360	0.061
3	1360	1410	0.025

Table 7. Number of standard deviations of the cross-correlation between the optical data cubes and the simulated observations towards NCP for various optical selections, in the two highest frequency intervals in our analysis.

Frequency range	All	$r \leq 16$	$\delta \geq 86$ deg	$r \leq 16$ and $\delta \geq 86$ deg
1360–1410 MHz	23.7	19.3	17.5	14.7
1310–1360 MHz	7.6	5.3	16.6	14.7

in galaxies. We assume the H I clumps population to have a random spatial distribution and to follow the characteristics measured using ALFALFA survey data by Jones et al. (2018). As shown in that paper, the H I galaxy mass function – the number density of H I galaxies in a logarithmic mass bins – is well described by a Schechter function:

$$\Phi(M) = \frac{{\rm d}N_{\rm HI}}{{\rm d}V\,{\rm d}\log_{10}(M)} = \log(10)\Phi^* \left(\frac{M}{M^*}\right)^{\alpha+1} \exp\left(-\frac{M}{M_*}\right) \quad (10)$$

where Φ^* corresponds to the normalization, M^* the knee mass and α, the low mass slope. Jones et al. (2018) fit these parameters using several subsets of the ALFALFA H I source catalogue; these results show some spatial dependence. We will retain here the parameters fitted with the whole data set (ALFALFA 100 per cent) and its 'near' subset ($v_{\rm CMB} < 4000\,{\rm km\,s^{-1}}$), listed in Table 4. The difference between the 'full' and 'near' parameter may give an indication of the systematics linked to the H I mass function. ALFALFA also observed some variation of these parameters in different regions of the sky but we do consider different areas in this study, therefore we stick to this global variation with observed distance in the following.

We use the following expression to relate the H I mass and the total 21 cm flux S_{21}, also quoted in Jones et al. (2018) :

$$\frac{M_{HI}}{M_\odot} = 2.356 \times 10^5 \left(\frac{D}{1{\rm Mpc}}\right)^2 S_{21}, \quad (11)$$

where D is the source distance in Mpc and S_{21} the integrated 21 cm flux in Jy km s^{-1}. For each redshift value, we compute D using fiducial cosmological parameters from Planck Collaboration XIII (2016). Using this distance and assuming a 210 km s^{-1} velocity width (corresponding to 1 MHz in the frequency domain) we can translate the flux limits or detection efficiencies in Jy into H I mass limits or detection efficiencies at each redshift. The integral of the H I mass function convolved with the detection efficiency gives the expected number density of clumps detectable by Tianlai at any given redshift. Integrating over the redshifts and taking into account volume element evolution with redshift, we obtain the expected total number of H I clump detections.

Fig. 12 shows the expected number of H I clump detections per square degree and per redshift bin ($\delta z = 0.001$) for the NCP and

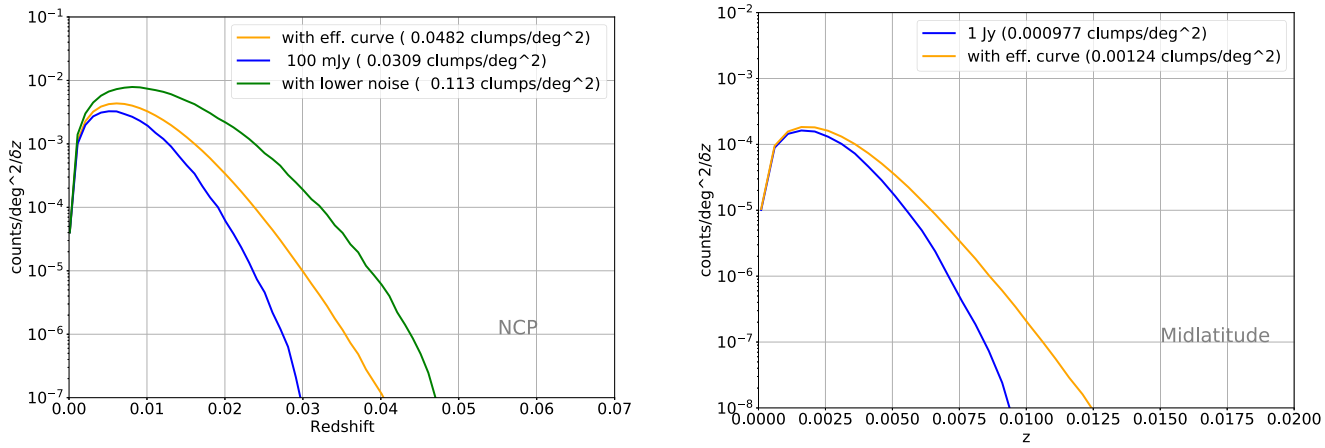

Figure 12. Expected number of detectable H I clumps as a function redshit, per square degree and per redshift bin $\delta z = 0.001$, for the NCP (left-hand panel) and mid-latitude (right-hand panel) surveys. The total numbers of detections per square degree, integrated over redshift are reported in the caption of the figures. Detection counts are shown using either a sharp detection threshold on the flux, in Jy (blue curves), or using the full shape of the detection efficiency curve (orange). The green curve in the left-hand panel (NCP) shows the number of detectable clumps with a lower noise (~ 2.5 mK) on visibilities.

mid-latitude cases as a function of redshift, assuming the H I mass function parameters from ALFALFA full sample. The total number of detections per square degree are shown in this figure and also listed in Table 5. As can be seen from Fig. 12, Tianlai would only be able to detect very nearby H I galaxies, below $z \lesssim .02$ for the NCP survey, and $z \lesssim 0.005$ for the mid-latitude survey. The numbers vary slightly with the H I mass function parameters used; if we use the near H I mass function parameters from ALFALFA at these very low redshifts, as listed in Table 4, the expected number of detection is even lower as reported in Table 5, due to the lower knee mass.

In the mid-latitude case, the detection threshold is higher, hence a number of detection per redshift interval lower than in the NCP case is expected. This higher threshold is not totally compensated by the larger surface area covered by this survey, making the NCP survey the most promising in terms of expected H I clump discovery rate. As mentioned in Section 3 a lower noise per visibility sample may be achieved by observing over longer period, less than a year for the full survey. In addition, the **DF** foreground subtraction method used for determining the source detection efficiency increases the noise level of the resulting difference map. The noise level which is the major limitation of the detection efficiency for the NCP case can be reduced by a factor about 2–3, combining lower visibility noise from the long survey duration and lower noise impact using the (**P**) foreground subtraction. A detection threshold $S_*^{\rm th} \lesssim 0.05$Jy could then be reached for the NCP survey, as indicated in Fig. 11, leading to the expected number of detectable clumps indicated in Table 5 (third columns), between 12 and 17 in total, depending on the H I MF used.

5 CROSS-CORRELATION WITH OPTICAL GALAXY CATALOGUES

In this section, we assess the prospects of detecting the cross-correlation signal between the intensity maps from the Tianlai low redshift surveys with optical galaxy catalogues: SDSS (Ahumada et al. 2020) for the mid-latitude and NCCS (Gorbikov & Brosch 2014) for the polar cap survey. The SDSS catalogue does not cover the NCP, but we are carrying out a spectroscopic survey using the WIYN telescope to obtain the spectroscopic redshifts for the brightest galaxies of the NCCS catalogue. To evaluate the cross-correlation signal for the NCP case, we have used an artificial catalogue built by rotating the coordinates of the objects in the SDSS catalogue to overlap with our radio observations of the NCP. The respective footprints of the SDSS and NCCS are shown in Fig. 13.

The selection criteria used to retrieve data from the SDSS DR16 server[8] are given in Appendix B1. Starting from the optical galaxy catalogue, we create a catalogue of H I sources using a two step procedure. We first derive a stellar mass from the optical galaxy properties, following Taylor et al. (2011). The stellar mass is then converted into an H I mass using the relations derived by Brown et al. (2015), from the study of a combined ALFALFA-SDSS catalogue. H I emission parameters (flux and linewidth) are then derived from the H I mass as explained in Section 4.2. A more detailed description of the procedure for converting the optical catalogue of galaxies into a list of 21 cm source properties can be found in Appendix B2.

5.1 Cross-correlation extraction pipeline

The procedure used to determine the cross-correlation of Tianlai low-z observations with the SDSS or NCCS optical catalogues uses the pipeline described in Section 3.1, with a few additional components:

(i) A 21 cm source catalogue is created from the SDSS galaxies with their redshifts or from the rotated SDSS catalogue for the NCP case.

(ii) Simulated visibilities that would be observed with the T16DPA setup are computed, combining signals from the different sky components: diffuse synchrotron emission, radio sources, noise, and redshifted 21 cm sources. Instrument noise is added to visibility samples as white noise.

(iii) Sky maps are reconstructed, independently for each frequency, using the m-mode decomposition method described in Section 3.1. A linear filter in spherical harmonic (ℓ, m) space is applied to compute spherical sky maps used in the next pipeline stages.

(iv) The contribution of foreground emissions due to the Milky Way and radio sources is estimated and subtracted using either of the two approaches presented in Section 3.3.

[8] https://skyserver.sdss.org/dr16/en/tools/search/sql.aspx

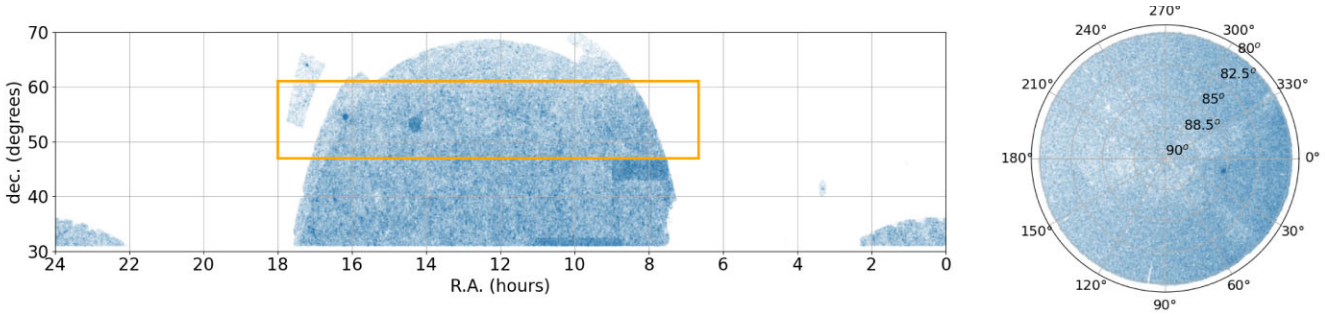

Figure 13. Footprints of the SDSS (left-hand panel) and NCCS (right-hand panel) catalogues used in this paper. We selected galaxies above $\delta = 30$ deg in the SDSS catalogues. The rectangular area outlined on the SDSS footprint is the area where cross-correlation with Tianlai low-z simulated observations have been computed. We only plotted the positions of the NCCS sources with $V \leq 17$.

(v) For the mid-latitude case, we project the filtered maps in an equatorial band around the central latitude of the simulated observations, and select the relevant portion of this band for cross-correlation studies, as indicated in the left-hand panel of Fig. 13. For the NCP case, the spherical maps are projected into square maps, using a Gnomonic or tangent plane projection centred on the North Celestial Pole. These re-projected maps are denoted RecSky(ν) for each frequency ν in the following.

(vi) An optical source sky cube is also constructed from the optical catalogue, using only the angular positions and the redshifts of the galaxies, ignoring the photometric information. All galaxies at the proper position and redshift interval are included, with their 21 cm brightness set to one in an arbitrary unit. This strategy was adopted in order to minimize the impact of the model relating H I properties to the optical magnitudes on the estimated cross-correlation signal. The source cube is then projected into frequency planes Src(ν), sharing the corresponding RecSky(ν) geometry.

(vii) For each frequency, we quantify the cross-correlation between the reconstructed radio sky maps after reprojection and the corresponding plane from the optical source cube RecSky(ν) $\times$ Src(ν). A correlation coefficient is computed at each frequency, as well as cross-correlation power spectra, in the spherical harmonics domain, $C_\nu^\times(\ell)$, for the mid-latitude case, and in the Fourier domain $P_\nu^\times(k_\perp)$for the NCP analysis.

It would of course be possible to characterize the cross-correlation through 3D power spectra $P^\times(k_\perp, k_\parallel)$. However, given that the signal fades away quickly with decreasing frequency at the very low redshift range considered here, and the different systematic effects affecting the radial $k_\parallel$ and transverse $k_\perp$ directions, we do not expect any advantage in using $P^\times(k_\perp, k_\parallel)$ instead of the set of $P_\nu^\times(k_\perp)$.

The optical sources cube to be correlated with the reconstructed sky cube has the same angular (5 arcmin pixels) and radial resolutions (corresponding to 1 MHz frequency in the HI cube) as the sky cube. Each galaxy in the optical catalogue is assigned to a pixel in the cube. The frequency is determined from the source redshift, and the position in the plane from the angular coordinates of the galaxy. All galaxies have the same weight, equal to one, regardless of their photometric magnitudes. A Gaussian smearing using the expected velocity width that we estimated as explained in Appendix B2 as FWHM along the frequency direction is then applied, as well as a 2D Gaussian filter to each plane, with a fixed angular width $\sigma_\perp =$ 10 arcmin.

We also build a series of randomized optical sources cubes, called shuffles to determine the level of residual cross-correlation signal, due to imperfect foreground subtraction and instrumental noise. The source angular positions and redshifts are shuffled independently from each other to generate these. We expect a null correlation of the reconstructed sky maps with these shuffled cubes. A hundred such cubes have been built and correlated with the reconstructed sky cube to estimate the cross-correlation signal dispersion. The model described in Appendix B1, used to derive H I emission parameters for the SDSS galaxies, captures only the average trend of broad distributions. Using the derived 21 cm brightness to build the optical source cube would have yielded an artificially enhanced cross-correlation signal, or costly simulations would be required to take into account the large dispersion in the $M_{\rm HI} = f$(mag-u, g, r, i, z) relation.

We have thus chosen to ignore the photometric information when building the optical source cubes. The amplitude of the optical source cube data is thus arbitrary – hence the normalization of correlation coefficients and spectra. However, their comparison with each other, when including or not different sky components, remains relevant.

We ran the pipeline and analysed the results for two noise configurations, with or without 5 mK noise added to visibility samples, and different combinations of sky components:

(i) continuum sources only (Haslam-based synchrotron map and NVSS sources)
(ii) H I simulated sources only
(iii) all components, i.e. combining diffuse synchrotron, continuum radio sources, H I sources and noise

5.2 Mid-latitude survey cross-correlation with the SDSS catalogue

We ran the pipeline described in Section 5.1 for all 1 MHz frequency shells between 1260 and 1410 MHz. We computed auto- and cross-correlation C_ℓ between the reconstructed sky cube and the sources cube, built from the optical catalogue for each frequency plane. The cross-correlation is also computed with each of the shuffled sources cubes. One expects the cross power spectra between the randomized sources cube and the simulated maps to be null in average, and that the dispersion around their average will give an estimate of the uncertainty in the computed cross-correlation coefficient or power spectrum. The significance of the correlation in a single frequency bin is low, so we combine the results into three arbitrary frequency (or redshift) intervals as described in Table 6. The resulting more significant combined results will also give an indication of the variation of correlation amplitude with redshift.

We present in Fig. 14 the angular autocorrelation power spectra, C_ℓ, of the maps we reconstruct without noise, averaged over frequen-

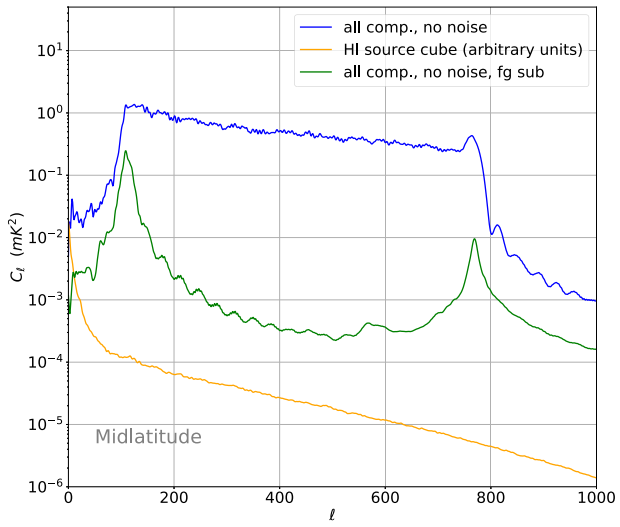

Figure 14. The averages of the autocorrelation spectra (C_ℓ) for frequencies between 1350 and 1375 MHz. The blue and green curves correspond to the sky cubes with all components, with no noise added to the visibilities, before (blue) and after (green) foreground subtraction using polynomial fit. The orange curve is the autocorrelation spectrum from the corresponding source cube. Note that its normalization is arbitrary, so its absolute level cannot be directly compared to the blue and green curves.

cies between 1350 and 1375 MHz. The spectrum prior to foreground subtraction is truncated below $\ell \sim 100$ and above $\ell \gtrsim 750$, mainly as a result of the map-making and filtering procedure described in Section 3.1. In between these two ℓ values, the effect of the foreground subtraction (polynomial fit) decreases the autocorrelation spectrum by 3 to 4 orders of magnitude. For comparison we also show the autocorrelation power spectrum computed from the sources cube. Due to the incomplete sky coverage of the survey, the different ℓ-modes are correlated with each other. We smoothed the power spectra with a $\Delta\ell = 15$ Gaussian to mitigate this effect.

We present in Fig. 15 a set of cross-correlation power spectra $C_\ell^\times$, between simulated maps and the source cube, starting with the ideal situation of a simulation including only H I sources and without noise (in orange), then adding noise (in magenta), then other components (in green), and finally the complete simulation of all components with noise (in blue).

We can observe the impact of adding noise or including more components which leaves systematic residuals in the maps after foreground subtraction, due e.g. to mode mixing. We observe that these two effects have roughly similar impacts in terms of the cross-power spectra between simulated planes and the data cube, as could be expected from the analysis reported in Section 3.4. We also note that in a broad ℓ range, the averaged cross-power spectrum for the complete simulation (in blue) stays positive, and well outside the dispersion from the 100 shuffled cubes. This reinforces the indication that Tianlai could observe a significant cross-correlation with the SDDSS catalogue, when performing a mid-latitude, low-z survey.

One might wonder if the foreground subtraction is a necessary step for cross-correlation detection. We present in Fig. 16 the averaged cross-power spectra between the 'reference planes' or sources planes and sky cube reconstructed from visibilities, for the frequency interval 1350 − 1375 MHz for three cases. The green and black curves show the cross-correlation from visibilities including all astrophysical components but no noise, with and without foreground

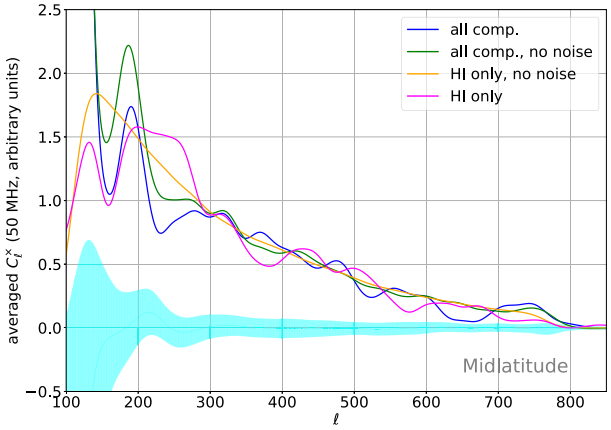

Figure 15. Smoothed cross-power spectra $C_\ell^\times$ (computed in each case between simulated maps and the source cube) averaged over the highest frequency bin (lowest redshift, bin 3), for different simulation cases, after foreground subtraction. The cyan band outlines the dispersion around central values of the average of the cross power spectra between maps from the simulation combining all components and noise and the 100 shuffled data cubes.

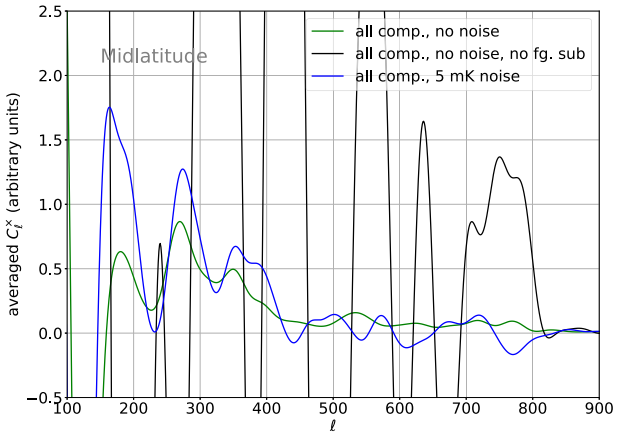

Figure 16. Effect of the foreground removal on the cross-correlation spectra $C_\ell^\times$ averaged for frequencies between 1350 and 1375 MHz. We present the cross-spectra obtained between the sources cube and data simulated with all astrophysical components but no noise, respectively before (black) and after (blue) foreground subtraction (polynomial fit). The green curve shows the result obtained with 5 mK noise added, after foreground subtraction. All spectra from this figure have been smoothed with a width of 15, to damp ℓ-to-ℓ correlations in the raw cross-spectra.

subtraction. The blue curve corresponds to the foreground subtracted maps, computed from visibilities with noise. The improvement brought by the foreground subtraction for extracting a significant cross-correlation signal is clearly visible. The cross-power spectrum amplitudes of the blue and green curves, after foreground subtraction, stay positive for a broad range multipoles, unlike the black curve. This ℓ-range ($250 \lesssim \ell \lesssim 500$) is indeed less affected by map-making, filtering, and foreground subtraction imperfections.

To summarize the cross-correlation detection perspective, we compute the average of the cross-spectra amplitudes $C_\ell^\times$ for $\ell \in [250, 500]$, well outside the ℓ range affected by map-making and filtering artefacts, after foreground subtraction. We present results of this averaging procedure for the three redshift bins in Fig. 17.

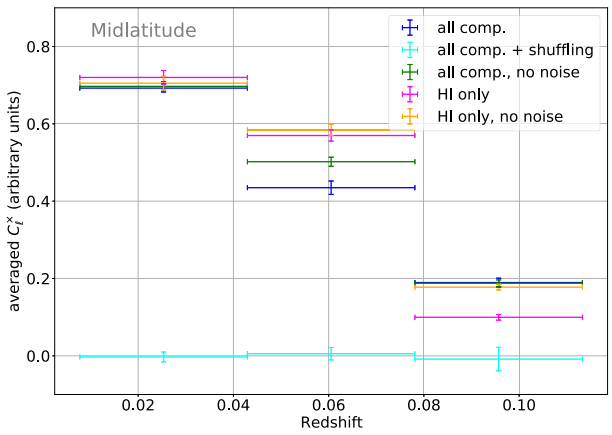

Figure 17. Average of the cross-correlation ($C_\ell^\times$) in the interval $\ell \in [250, 500]$ for each of the redshift bins defined in Table 6. We compare these for different components and noise combinations. All cross-correlation spectra are in arbitrary units.

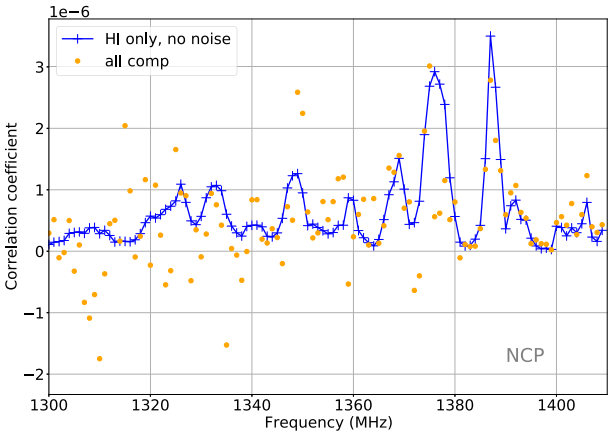

Figure 18. Raw correlation coefficients $\langle \mathrm{RecSky}(\nu) \times \mathrm{Src}(\nu) \rangle$, computed for each frequency plane. The blue curve (and crosses) correspond to the case where only H I sources were contributing to the simulated visibilities, while orange circles correspond to the case where visibilities were computed from all sky components (foregrounds and H I sources) and include 5 mK noise.

We observe that in the two lowest redshift bins the averaged cross-spectra of the simulation case including foregrounds, noise and H I sources is positive, significantly different from zero, when compared to the dispersion computed from the shuffled source cubes. These differences amount to 43 and 18 standard deviations, respectively.

Although this statistical significance estimate is very crude, it still shows that a low-redshift mid-latitude survey operated in the conditions described here with the Tianlai dish array would show a very significant correlation when matched with SDSS low redshift galaxies. For the highest redshift bin (the first one from Table 6), the differences in the average $C_\ell^\times$ computed for the different simulation configurations are not statistically significant when compared to the statistical uncertainty derived from shuffled cubes.

5.3 Cross-correlation forecast from the NCP survey

No comprehensive redshift galaxy survey covering the NCP area is available. We thus first forecast in Section 5.3.1 the sensitivity of a Tianlai survey toward the NCP for 21 cm-optical cross-correlation detection using a rotated SDSS catalogue, providing an artificial spectroscopic coverage of the north celestial pole. In a second step, in Section 5.3.2, we use the foreseen characteristics of the ongoing spectroscopic NCP survey to assess the level of its relative incompleteness with respect to the SDSS, and obtain a forecast of its level of cross-correlation with Tianlai data we will obtain.

5.3.1 With a rotated SDSS catalogue

We start with the catalogue described in Appendix B2 and we rotate the sky coordinates in order to align $(\alpha, \delta) = (180, 45)$ deg with the NCP, thus getting an artificial coverage of the NCP area. We use a pipeline similar to that used in the mid-latitude cross-correlation study presented in Section 5.2 with some changes in the last stages. Given the range of declinations (83 deg $\leq \delta \leq$ 90 deg) studied here, we restrict the projection to a circular area around the NCP, within a radius of 7 degrees. The spherical maps are projected into small square flat maps (169 $\times$ 169 maps with 5 arcmin pixels) through a gnomonic projection. A source cube with identical resolution is constructed from the rotated catalogue, with the same prescription as for the mid-latitude case, with smearing in angular direction, and in frequency, according to velocity dispersion. We also create 100 shuffled source cubes to assess statistical and systematic effects. For the NCP analysis we have use the same frequency intervals as those listed in Table 6. Given the small angular extent of the maps, the cross-correlation power spectra are computed using a standard Fast Fourier Transform (FFT), rather than a spherical harmonics transform.

As a first test of the presence of a correlation between the simulated sky and the source cube, we computed the correlation coefficient for the two maps for each frequency plane $\langle \mathrm{RecSky}(\nu) \times \mathrm{Src}(\nu) \rangle$. Fig. 18 represents the evolution of these raw correlation coefficients, as a function of frequency. Large values of the correlation coefficients computed for the ideal case (no foreground, no noise, in blue), appearing as peaks in the distribution, near 1387 or 1370 MHz for example, might be interpreted as the sign of presence of non-linear clustering, maybe sheets or filaments appearing in the galaxy distribution. Most features seen above 1370 MHz in the ideal case stay visible in the realistic case where all sky components and noise are included (orange circles), for example the peaks around 1375 and 1386 MHz.

To evaluate more thoroughly the cross-correlation between the simulated radio maps and the maps derived from the optical catalogue, and its angular scale dependence, we compute 2D FFTs for each pair of maps of a given frequency plane. We average the amplitudes of the FFT modes in bins of angular wave modes (azimuthal average), which results in 1D amplitude vectors (one per frequency plane). We present in Fig. 19 averages of such spectra, in the second and third frequency intervals, 1310–1360 MHz, in the left-hand panel and 1360–1410 MHz, in the right-hand panel. We restrict the range of angular frequencies in these plots to $0.5 - 2\,\mathrm{deg}^{-1}$ since, on the one hand, at low angular frequencies, i.e. large angular scales, maps are dominated by reconstruction and foreground subtraction artefacts, and by noise at high angular frequencies.

Fig. 19 presents a comparison of the power spectra reconstructed for different simulation cases with the ideal case including only H I sources and no noise, shown as orange stars. In that case, we get a smooth and positive spectrum which exhibits a very significant positive cross-correlation, as judged from the dispersion obtained in the 100 shuffled realizations, shown in cyan.

We can observe that adding noise (magenta circles) significantly degrades this at large angular scales ($\lesssim 2^\circ$). Nevertheless, these

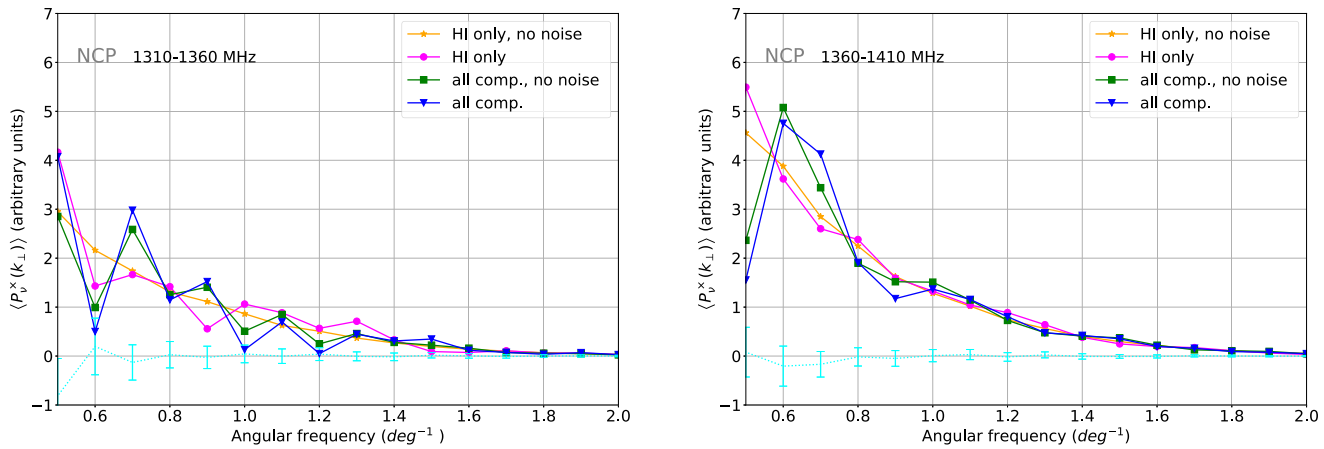

Figure 19. Reconstructed cross-power spectra $P_\nu^\times(k_\perp)$ averaged in the two frequency intervals corresponding to bins 2 (left-hand panel) and 3 (right-hand panel) from Table 6. We compare results obtained from various simulation parameters, from the ideal (only H I sources, no noise) to the complete case. On both sides, the cyan results represent the averages and dispersions from the power spectra from the cross-correlations with a set of 100 shuffled catalogues. As we use only the sources' positions to build their cube, units are arbitrary.

spectra amplitudes stay positive, with higher statistical significance in the highest frequency interval (1360–1410 MHz). The foreground residuals, observed when including all astrophysical components, but without adding noise to visibilities, also affect the cross-correlation spectra, shown as the green squares in Fig. 19. The impact of the foreground subtraction residuals appears subdominant, compared to that of the noise, in agreement with survey sensitivity analysis presented in 3.4. Finally, with noise added, the cross-spectra (shown as blue triangles) in both frequency intervals stay significantly positive.

Using the dispersions extracted from the shuffled simulations we can compute the statistical significance of the distance from a null value of obtained cross-power spectra. Doing this in each of the three frequency intervals of Table 6, and averaging over the 0.6−1.3 deg^{-1} angular frequencies, we find a significance of 23.8, 7.5, and 5.8 standard deviations for the cross-correlations at 1385, 1335, and 1285 MHz, respectively. This angular frequency interval is somewhat arbitrary but was chosen because below 0.6 deg^{-1} the cross-power spectra $P_\nu^\times(k_\perp)$ become distorted by large angular scale foreground removal residuals and large statistical fluctuations (due to incomplete sky coverage), and the small scales are diluted by the instrumental beam at the other end.

We can therefore conclude that we would detect with a high significance a cross-correlation between an optical galaxy catalogue with similar characteristics, completeness, and sensitivity, as the SDSS covering the region surrounding the NCP, at least for the lowest redshift interval. The NCP cross-correlation is significant, but lower than in our forecast for a mid-latitude survey shown in Section 5.2, which amount to 43 and 18 standard deviations in the two highest frequency bins, respectively. However, we note that, as shown in Section 3.4, instrument noise is the main limitation for the NCP survey and the noise per visibility sample we have used here is conservative, hence so is our forecast. On the other hand, the sky area covered by the NCP survey is much smaller than the mid-latitude case, which does explain part, if not all, of the difference between the mid-latitude and NCP cross-correlation significance. We also note that this evaluation is done under simple assumptions for instrumental performance, and should therefore be considered as an upper limit of what the real data will show.

5.3.2 What can we expect with the future Tianlai NCCS-based spectroscopic catalogue?

The Tianlai collaboration is currently carrying out a spectroscopic survey based on the NCCS photometric catalogue, performed with the WIYN telescope[9] and its HYDRA spectrograph.[10] A 4 degree radius disc around the NCP has been targeted by the WIYN observations up to now. The cleaned version of the NCCS catalogue provides, in addition to purely photometric information, an indication of the point-like or extended nature of each object, called the `PESS` score. They advertise that a score greater than 2 indicates an extended source used to select objects for spectroscopy.

In order to get a realistic evaluation of the cross-correlation signal strength which can be achieved with our NCCS-based spectroscopic catalogue, we need to determine the completeness level of this sample w.r.t. that of the SDSS sample used in the previous section. The photometry in NCCS is provided in the Johnson–Cousins system whereas in SDSS the `ugriz` system is used. According to https://www.sdss.org/dr12/algorithms/sdssubvritransform/, specifically their Lupton (2005) relations, we find an approximate magnitude conversion equation $R = r - 0.25$ using the average colour $r - i \simeq 0.35$ of our SDSS galaxies for $z < 0.1$. We compared the number of objects in our SDSS rotated catalogues, with $z < 0.1$ and $r < 17.5$, within 7 degrees of the NCP with the number of extended objects (`PESS` $\geq$2) in the same area in the NCCS, with $R < 17.75$. We find 1131 and 1027 objects in those samples, respectively. We conclude that the object density in the NCP area (7 degree radius) of the two catalogues agree within 10 per cent. Note that we assume that galaxies further than $z = 0.1$ will most probably be flagged as point-like sources in the NCCS. Spectroscopic data reduction from this WIYN survey has only obtained reliable redshifts for 40 per cent of extended objects. Although this efficiency might be increased in the future, we will use this incompleteness factor, translated into a r magnitude cut in the SDSS catalogue.

Looking at the SDSS r-magnitude distribution, we find that requiring $r \leq 16$ will result in a similar galaxy count over the NCP area as the one in the NCCS spectroscopic catalogue. We therefore

[9] https://www.wiyn.org/
[10] https://www.wiyn.org/Instruments/wiynhydra.html

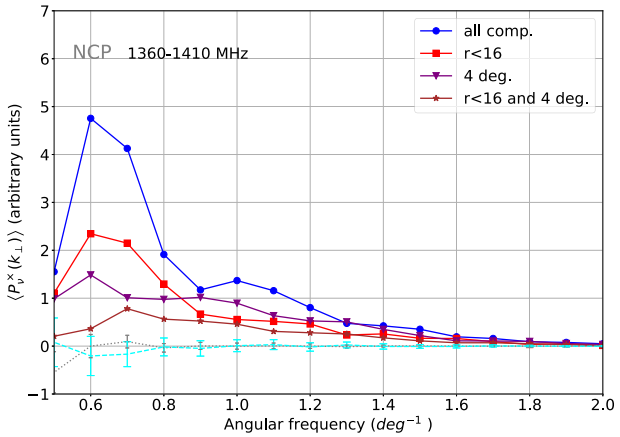

Figure 20. Effect of excluding a fraction of optical galaxies when computing the cross-correlation for the NCP case for bin 3 from Table 6, i.e. the lowest redshift bin. The 1D cross-power spectra $P_\nu^\times(k_\perp)$ as a function of the angular frequency are shown, as blue dots, when all sources are included, as red squares when only sources with $r < 16$ are included, as purple triangles for sources with $\delta \geq 86$ deg, and as brown stars, when requiring both $r < 16$ and $\delta \geq 86$ deg. The spectra obtained from the shuffled source cubes, are also shown, in cyan (larger ticks) for the larger area, in grey (smaller ticks)for the smaller area. The vertical axis is in arbitrary units as the source cube is built from optical object positions only.

estimated the cross-correlation signal level between our simulated maps and the corresponding optical data cubes built while imposing $r \leq 16$, and compared these with the estimation obtained with the full optical catalogue. We also performed this analysis using the central 4 degree radius disc only, which is the area currently covered by the on-going WIYN observations. We show the cross-correlation power spectra we reconstruct in these four cases, for the highest frequency interval (lower redshift) in Fig. 20.

Following the same method outlined in Section 5.3.1 we compute the statistical significance of the cross-correlation spectra using the dispersion of the null spectra obtained from the shuffled cubes. We report these results in Table 7 for the two highest frequency intervals. Any selection w.r.t. the full sample changes the signification of the cross-correlation signal, to various extents. From the values reported in Table 7 we observe that the brightness cut has a larger effect for the highest redshift bin: as the sources' brightness decrease with redshift, the fraction of those rejected by the cut gets larger with redshift. We also note that selecting a narrower region around the NCP has a larger effect for the lowest redshift bin. We may benefit of the lower noise per map pixel in this central area. Also, as the solid angle included in the analysis increases with redshift, more sources are included in the optical catalogue at higher redshift.

We conclude that given the efficiency and sky coverage of our ongoing NCCS spectroscopic survey, a significant cross-correlation between the Tianlai radio observations and the enriched NCCS catalogue should be observed, with more than 14 standard deviations statistical significance, in the two highest frequency intervals. However, this signal will reach a higher statistical signification if the area covered by our spectroscopic observations can be extended beyond $\delta \geq 86$ deg. This would also be true if we succeed in increasing our efficiency at getting redshifts for fainter sources. Finally, the cross-correlation signal will be stronger if we achieve a noise per visibility sample lower than the conservative value of 5 mK used in our simulation, which is feasible by increasing the observation time.

6 CONCLUSIONS, PERSPECTIVES

We have shown that low-z surveys carried out by the Tianlai dish array, by tuning the analogue electronic to the 1300–1400 MHz band would be sensitive enough to see extragalactic 21 cm signal, either through direct detection of a few nearby massive HI galaxies or in cross-correlation with optical surveys.

The Tianlai instrument is designed to observe in transit mode, with sky coverage obtained through constant declination scans. By pointing the antennae toward the North Celestial Pole (NCP), a small sky area is covered, leading to increased sensitivity thanks to long integration time. We have thus studied two complementary survey strategies in this paper, a deep survey covering $\sim$150 deg^2 around the NCP, and a shallower survey, corresponding to a $\sim$10 deg declination band at mid-latitudes, covering $\sim$1500 deg^2 useful area around $\delta = 55^\circ$ declination and overlapping the SDSS footprint.

A noise level of $\lesssim$ 2–4mK per 1MHz $\times$ 0.25^{2}deg^2 pixels, should be reached for a 3 months survey of the NCP area, leading to a point source detection threshold of $\sim$0.08Jy. Our study shows that the residuals from the mode mixing for the NCP survey should be negligible compared to the noise fluctuations. Extending the survey duration to a year would thus decrease the noise to $\sim 1-2$ mK pixel^{-1} and reach a source detection threshold of $\sim$0.05Jy. Tianlai should then be able to detect $\gtrsim$ 10 nearby HI clumps or galaxies through their 21 cm emission, up to redshifts $z \lesssim 0.02$. We have also studied the possibility to detect statistically the extragalactic 21 cm emission, dominated by the one from galaxies at these low redshifts, in cross-correlation with optical surveys. We show that the cross-correlation signal between Tianlai NCP intensity maps and the NCCS optical galaxy survey could be detected with very high significance, $\sim$15 standard deviations, provided we get redshifts for most of the NCCS galaxies with magnitude $R < 16.25$. A spectroscopic survey of NCCS galaxies is indeed being carried out with the WIYN telescope, and it would be helpful to pursue this effort to extend the surveyed area to the full 7 degree radius disc foreseen for the planned Tianlai low-z NCP survey.

We have adopted a conservative approach throughout this paper, using the 5 mK noise level for visibilities, corresponding to a 3 months survey and not the 2.5 mK value corresponding to the nominal 1-yr NCP survey. This leaves some room to mitigate a fraction of increased fluctuations associated with calibration errors, partial beam knowledge and possibly correlated noise. Further studies are needed to assess the impact of these instrument imperfections on the survey performance. As an example, our preliminary investigation shows that a 7 deg RMS phase calibration errors between baselines would lead to a 3-fold increase of the residual fluctuations at the map level for the NCP survey.

The mid-latitude survey suffers larger residuals from mode mixing, as well as higher noise level due to the larger sky area, but has the advantage of having a large overlap with the SDSS spectroscopic survey at low redshifts. The lower residuals due to mode mixing (frequency dependent synthesized beam) for the NCP case can be explained by the larger number of effective baselines created by the rotated array, projected on the same sky area due to the earth rotation, while this rotation makes different parts of the sky to drift in front of the instrument at mid-latitudes. The fluctuations due to the noise reaches about $\sim$15mK pixel^{-1} for the mid-latitude survey, with residuals from foreground subtraction at a similar level, leading to a direct source detection threshold about 0.9 Jy, about ten times higher than that of the NCP case. Despite this significantly higher level of residuals and noise in the foreground subtracted maps, our study shows that the cross-correlation of the Tianlai foreground subtracted

maps with the SDSS optical galaxies could be detected with high significance ($\gtrsim 40$ standard deviations).

The results presented in this paper can be consolidated in the future by studying the impact of instrumental effects not yet taken into account, phase and amplitude calibration errors and drifts, correlated noise, as well as the impact of poorly known individual antenna beams, specially far side lobes. There is also room for optimizing further the observation strategy, for example, the question of the optimal area to be covered in the NCP region for a given survey duration. Another important parameter would be the exact frequency coverage. However, Tianlai already has three sets of filters, covering 1170–1270, 1250–1350, and 1330–1430 MHz, respectively, in addition to the currently used 700–800 MHz filters. The contribution of different frequency slices vary due to the decrease of the sensitivity with the redshift, partially compensated by the volume increasing with the redshift. One has also to take into account the frequency intervals which would be impacted by RFI to determine the optimal frequency band.

ACKNOWLEDGEMENTS

This research made use of PHOTUTILS, an Astropy package for detection and photometry of astronomical sources (Bradley et al. 2021). The Tianlai Dish Array was built with the support of the CAS special fund for repair and purchase, and operated with the support of NAOC Astronomical Technology Center. We are supported by the Ministry of Science and Technology (MOST) of China 2018YFE0120800, National Key R&D Program 2017YFA0402603, the National Natural Science Foundation of China (NSFC) grants 11633004 and 11473044, and the Chinese Academy of Sciences (CAS) grants QYZDJ-SSW-SLH017. Two of the authors (RA, OP) acknowledge financial support from the Programme National de Cosmologie et Galaxies (PNCG) and from the FCPPL (France-China Particle Physics Laboratory) of CNRS, France.

DATA AVAILABILITY

The simulated data generated for and analysed in this article will be shared on reasonable request to the corresponding author. Public data sets used in this article may be retrieved from the references indicated in the text, namely: SDSS DR16 server, Lambda archive and Vizier at CDS.

REFERENCES

Ahumada R. et al., 2020, ApJS, 249, 3
Alonso D., Ferreira P. G., Santos M. G., 2014, MNRAS, 444, 3183
Amiri M. et al., 2022, preprint (arXiv:2202.01242)
Anderson C. J. et al., 2018, MNRAS, 476, 3382
Ansari R. et al., 2012, A&A, 540, A129
Ansari R. et al., 2020, MNRAS, 493, 2965
Bacon D. J. et al., 2020, Publ. Astron. Soc. Austr., 37, e007
Bandura K. et al., 2014, in Stepp L. M., Gilmozzi R., Hall H. J., eds, Proc. SPIE Conf. Ser. Vol. 9145, Ground-based and Airborne Telescopes V. SPIE, Bellingham, p. 914522
Battye R. A., Davies R. D., Weller J., 2004, MNRAS, 355, 1339
Battye R. et al., 2016, preprint (arXiv:1610.06826)
Bharadwaj S., Nath B. B., Sethi S. K., 2001, J. Astrophys. Astron., 22, 21
Bradley L. et al., 2021, astropy/photutils: 1.2.0
Brown T., Catinella B., Cortese L., Kilborn V., Haynes M. P., Giovanelli R., 2015, MNRAS, 452, 2479
Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, ApJ, 803, 21
Chang T.-C., Pen U.-L., Peterson J. B., McDonald P., 2008, Phys. Rev. Lett., 100, 091303
Chang T.-C., Pen U.-L., Bandura K., Peterson J. B., 2010, Nature, 466, 463
Chen X., 2012, Int. J. Mod. Phys. Conf. Ser., 12, 256
Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693
Cosmic Visions 21 cm Collaboration 2018, preprint (arXiv:1810.09572)
Crittenden R. G., 2000, Astrophys. Lett. Commun., 37, 377
Cunnington S. et al., 2022, MNRAS, preprint (arXiv:2206.01579)
Das S. et al., 2018, in Zmuidzinas J., Gao J.-R., eds, Proc. SPIE Conf. Ser. Vol. 10708, Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy IX. SPIE, Bellingham, p. 1070836
DeBoer D. R. et al., 2017, PASP, 129, 045001
Delabrouille J. et al., 2013, A&A, 553, A96
Eastwood M. W. et al., 2018, AJ, 156, 32
Gorbikov E., Brosch N., 2014, MNRAS, 443, 725
Hogg D. W., 1999, preprint (astro–ph/9905116)
Hu W., Wang X., Wu F., Wang Y., Zhang P., Chen X., 2020, MNRAS, 493, 5854
Jones M. G., Haynes M. P., Giovanelli R., Moorman C., 2018, MNRAS, 477, 2
Li J. et al., 2020, Sci. China Phys. Mech. Astron., 63, 129862
Liu A., Shaw J. R., 2020, PASP, 132, 062001
Masui K. W. et al., 2013, ApJ, 763, L20
Mondal R., Shaw A. K., Iliev I. T., Bharadwaj S., Datta K. K., Majumdar S., Sarkar A. K., Dixon K. L., 2020, MNRAS, 494, 4043
Moorman C. M., Vogeley M. S., Hoyle F., Pan D. C., Haynes M. P., Giovanelli R., 2014, MNRAS, 444, 3559
Morales M. F., Wyithe J. S. B., 2010, ARA&A, 48, 127
Newburgh L. et al., 2016, in Hall H. J., Gilmozzi R., Marshall H. K., eds, Proc. SPIE Conf. Ser. Vol. 9906, Ground-based and Airborne Telescopes VI. SPIE, Bellingham, p. 99065X
O'Connor P. et al., 2021, in Marshall H. K., Spyromilio J., Usuda T., eds, Proc. SPIE Conf. Ser. Vol. 11445, Ground-based and Airborne Telescopes VIII. SPIE, Bellingham, p. 114457C
Parsons A. R., Backer D. C., 2009, AJ, 138, 219
Pen U.-L., Staveley-Smith L., Peterson J. B., Chang T.-C., 2009, MNRAS, 394, L6
Planck Collaboration XIII, 2016, A&A, 594, A13
Pritchard J. R., Loeb A., 2008, Phys. Rev. D, 78, 103511
Remazeilles M., Dickinson C., Banday A. J., Bigot-Sazy M. A., Ghosh T., 2015, MNRAS, 451, 4311
Seo H.-J., Dodelson S., Marriner J., Mcginnis D., Stebbins A., Stoughton C., Vallinotto A., 2010, ApJ, 721, 164
Shaw J. R., Sigurdson K., Sitwell M., Stebbins A., Pen U.-L., 2015, Phys. Rev., D91, 083514
Smoot G. F., Debono I., 2017, A&A, 597, A136
Taylor E. N. et al., 2011, MNRAS, 418, 1587
The CHIME/FRB Collaboration, Amiri M. et al., 2021, Astrophys. J. Supp., 257, 59
Tingay S. J. et al., 2013, PASA, 30, e007
van Haarlem M. P. et al., 2013, A&A, 556, A2
Vanderlinde K. et al., 2019, in Canadian Long Range Plan for Astronomy and Astrophysics White Papers. p. 28
Villaescusa-Navarro F., Alonso D., Viel M., 2017, MNRAS, 466, 2736
Villaescusa-Navarro F. et al., 2018, ApJ, 866, 135
Wang J. et al., 2021, MNRAS, 505, 3698
Wolz L. et al., 2022, MNRAS, 510, 3495
Wu F. et al., 2021, MNRAS, 506, 3455
Wuensche C. A. et al., 2022, A&A, 664, A15
Yatawatta S. et al., 2013, A&A, 550, A136

Zhang J., Ansari R., Chen X., Campagne J.-E., Magneville C., Wu F., 2016, MNRAS, 461, 1950
Zheng Q., Wu X.-P., Johnston-Hollitt M., Gu J.-h., Xu H., 2016, ApJ, 832, 190

APPENDIX A: NOISE POWER EVOLUTION WITH REDSHIFT

We derive here the redshift dependence of the noise power $P_{\rm noise}(k, z)$ projected on sky for a radio survey, using a simple and hopefully pedagogical approach. We follow notation in Hogg (1999) for cosmological distances; d_M stands for the transverse comoving distance, while $d_L = (1+z)d_M$ is the luminosity distance and $d_A = d_M/(1+z)$ the angular diameter distance.

Let's consider brightness temperature sky maps $T_b(\alpha, \delta)$ with angular resolution $\delta\theta$ and frequency resolution $\delta\nu$. Instrument angular resolution $\delta\theta$ varies with wavelength as $\delta\theta \propto \lambda/D_{\rm array}$, where $D_{\rm array}$ is the array spatial extent and $\lambda = c/\nu$ the observation wavelength. Projecting such a map on a cosmological volume at redshift z, determined by the observation frequency ν, we obtain voxels with transverse ($a_\perp$) and radial ($a_\parallel$) comoving dimensions, corresponding to a comoving volume $\delta V = a_\perp^2 \times a_\parallel$:

$$
\begin{aligned}
\nu &= \frac{\nu_{21}}{1+z} \qquad \nu_{21} = 1420.4\,\mathrm{MHz}\\
\delta\theta &= (1+z)\delta\theta_0 \qquad \delta\theta_0 = \delta\theta\,(\nu_{21})\\
a_\parallel &= (1+z)\frac{c}{H(z)}\frac{\delta\nu}{\nu} = (1+z)^2\frac{c}{H(z)}\frac{\delta\nu}{\nu_{21}}\\
a_\perp &= (1+z)d_A(z)\delta\theta = d_M(z)\delta\theta
\end{aligned}
$$

The fluctuations of the map pixel's value due to instrumental noise, denoted σ^T and characterized by the system temperature $T_{\rm sys}$, can be easily related to the noise power $P_{\rm noise}$. The voxel dimensions $(a_\perp, a_\parallel)$ determine the maximum accessible wave numbers $(k_\perp, k_\parallel)$. Assuming white noise and ignoring damping due to averaging over the voxel, we can write the Plancherel-Parseval identity, relating the variance of the map pixels' values to the map Fourier coefficients $F(k_x, k_y, k_z)$:

$$
\begin{aligned}
\left(\sigma^T\right)^2 &= \sum_{k_x,k_y,k_z} |F(k_x,k_y,k_z)|^2\\
\left(\sigma^T\right)^2 &= P_{\rm noise}\iiint_{-k^{\rm max}}^{k^{\rm max}}\left(\frac{1}{2\pi}\right)^3 dk_x dk_y dk_z\\
k_{\perp,\parallel}^{\rm max} &= \frac{\pi}{a_{\perp,\parallel}} \rightarrow P_{\rm noise} = \left(\sigma^T\right)^2\left(a_\perp^2 a_\parallel\right)
\end{aligned}
$$

We would find a pessimistic $P_{\rm noise}$ redshift dependence if we applied the above formulae directly. Indeed, the array instantaneous field of view (FOV) increases with redshift like $(\delta\theta)^2$ as $(1+z)^2$, increasing the mapping speed. The per pixel noise would then decrease with redshift as $(1+z)^{-1}$. Taking this effect into account, we find a redshift dependence for the noise power spectrum identical to the one derived in section 3.2 of Ansari et al. (2012).

$$
P_{\rm noise}(z) \simeq (1+z)^2 d_M^2(z)\frac{c}{H(z)}\frac{\delta\nu}{\nu_{21}}(\delta\theta_0)^2\left(\sigma_0^T\right)^2, \tag{A1}
$$

where σ_0^T denotes the per pixel noise level at $z = 0$.

APPENDIX B: SIMULATION OF EXTRAGALACTIC H I 21CM SIGNAL

B1 Preparation of the input catalogue from SDSS

We extract SDSS data through their SQL server with a geometric selection of the intersection of the `PhotoObj`, `SpecObj`, and `Photoz` catalogues (the latter providing absolute magnitudes). From this initial catalogue we select objects satisfying the empirical fiducial criteria:

(i) sources belonging to the category `GALAXY`
(ii) a spectroscopic redshift in the interval [0.005,1.0]
(iii) selection of objects with ordinary colours: $-0.5 < r - i < 2.5$ and $-0.2 < g - i < 1.65$ to avoid a few rare outliers

B2 From optical photometry to radio parameters

In order to be able to simulate observations of a sky composed of diffuse and point-like continuum radio sources as well as 21cm emission from galaxies, we need to determine the 21 cm emission properties of the latter starting from an optical galaxy catalogue. We follow a two-step procedure to achieve this. First, following Taylor et al. (2011) we estimate the stellar mass of each galaxy from their photometric properties using their equation (8)[11] :

$$
\log(M_\star/M_\odot) = 1.15 + 0.70(M_g - M_i) - 0.4M_i \tag{B1}
$$

This equation can easily be applied for SDSS objects, using their Photoz table that includes absolute magnitudes. From a cross-match between the ALFALFA and SDSS, Brown et al. (2015) investigated the relation between the H I and stellar masses of galaxies. From their Fig. 3, where results of stacking ALFALFA observations for the full SDSS sample are reported, we extract a simple relation between the H I and stellar masses :

$$
\log(M_{HI}/M_\star) = 0.179 - 0.66(\log M_\star - 9.21) \tag{B2}
$$

Combining equations (B1) and (B2) we can now estimate the H I mass for each galaxy. Moorman et al. (2014) (their Fig. 7) show the relation between the H I mass and W_{50}, the gas velocity dispersion width at half-maximum flux, from which we infer a simple linear relation between $\log W_{50}$ and $\log M_{\rm HI}$.

Using this equation and (11) we can finally estimate the 21 cm peak flux and frequency width for each galaxy, after accounting for their distance as determined by their redshift within a fiducial cosmology (Planck ΛCDM).

We prepared a input catalogue for our simulations from the SDSS photo- and spectroscopic catalogue, choosing galaxies with declination above $\delta = 30$ deg. The redshift and the derived stellar and H I mass distributions for these sources are shown in Fig. B1. The computed H I flux and frequency dispersion width as a function of redshift are shown in Fig. B2.

[11] We use this equation as found in the MNRAS published version of Taylor et al. (2011), where it seems to have been corrected w.r.t. the corresponding arXiv preprint (arXiv:1108.0635).

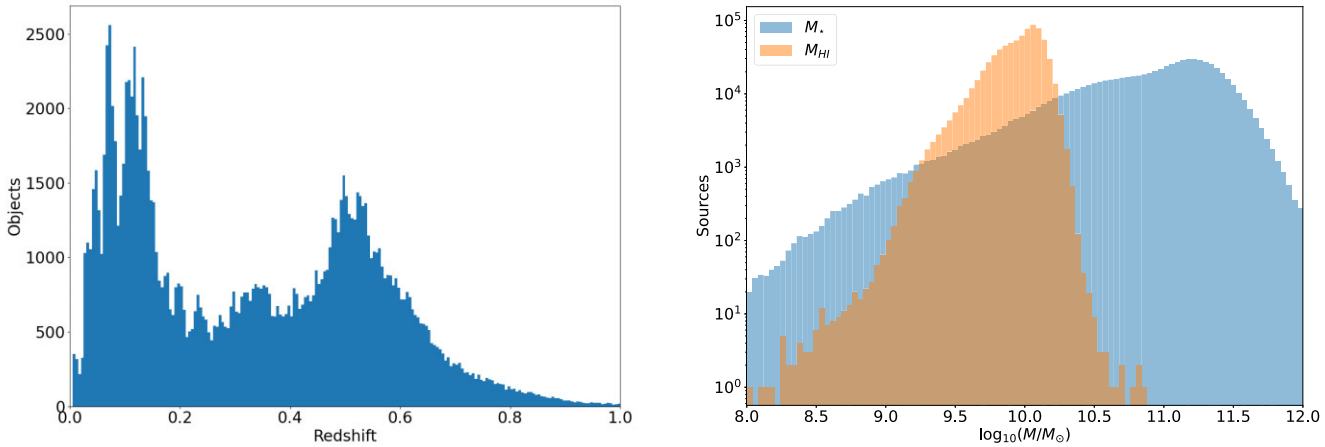

Figure B1. Left-hand panel: Redshift distribution of the SDSS sources after additional selection criteria described in paragraph B1. Right-hand panel: Derived stellar mass distribution in blue, and corresponding H I mass distribution in orange.

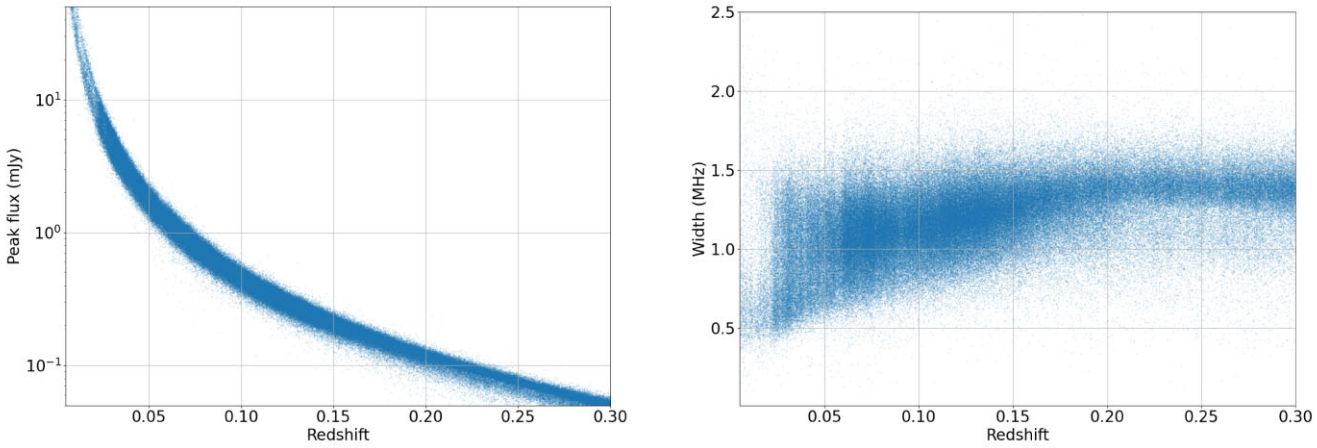

Figure B2. 21 cm peak flux (in mJy) and frequency dispersion width (in MHz) computed from W_{50} as a function of redshift for our simulated H I source catalogue, derived from the SDSS catalogues.

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