HI intensity mapping with the MIGHTEE survey: power spectrum estimates

Sourabh Paul\textsuperscript{1} *, Mario G. Santos\textsuperscript{1,2}, Junaid Townsend\textsuperscript{1}, Matt J. Jarvis\textsuperscript{3}, Natasha Maddox\textsuperscript{4}, Jordan D. Collier\textsuperscript{5,6}, Bradley S. Frank\textsuperscript{2,5,7}, Russ Taylor\textsuperscript{8}

\textsuperscript{1} Department of Physics and Astronomy, University of the Western Cape, Robert Sobukwe Road, Bellville, 7535, South Africa
\textsuperscript{2} South African Radio Observatory (SARAO), 2 Fir Street, Observatory, Cape Town, 7925, South Africa
\textsuperscript{3} Astrophysics, Department of Physics, University of Oxford, Keble Road, Oxford, OX1 3RH, UK
\textsuperscript{4} Faculty of Physics, Ludwig-Maximilians-Universität, Schemestr. 1, 81679 Munich, Germany
\textsuperscript{5} Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa
\textsuperscript{6} School of Science, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia
\textsuperscript{7} The Inter-University Institute for Data Intensive Astronomy (IDIA), and University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa
\textsuperscript{8} Inter-University Institute for Data Intensive Astronomy, and Department of Physics and Astronomy, University of the Western Cape, Robert Sobukwe Road, Bellville, 7535, South Africa

ABSTRACT

Intensity mapping (IM) with neutral hydrogen is a promising avenue to probe the large scale structure of the Universe. With MeerKAT single-dish measurements, we are constrained to scales $> 1$ degree, and this will allow us to set important constraints on the Baryon acoustic oscillations and redshift space distortions. However, with MeerKAT’s interferometric observation, we can also probe relevant cosmological scales. In this paper, we establish that we can make a statistical detection of HI with one of MeerKAT’s existing large survey projects (MIGHTEE) on semi-linear scales, which will provide a useful complementarity to the single-dish IM. We present a purpose-built simulation pipeline that emulates the MIGHTEE observations and forecast the constraints that can be achieved on the HI power spectrum at $z = 0.27$ for $k > 0.3$ Mpc$^{-1}$ using the foreground avoidance method. We present the power spectrum estimates with the current simulation on the COSMOS field that includes contributions from HI, noise and point source models from the data itself. The results from our visibility based pipeline are in good agreement to the already available MIGHTEE data. This paper demonstrates that MeerKAT can achieve very high sensitivity to detect HI with the full MIGHTEE survey on semi-linear scales (signal-to-noise ratio $> 7$ at $k = 0.49$ Mpc$^{-1}$) which are instrumental in probing cosmological quantities such as the spectral index of fluctuation, constraints on warm dark matter, the quasi-linear redshift space distortions and the measurement of the HI content of the Universe up to $z \approx 0.5$.

Key words: cosmology: observations — large-scale structure of Universe — techniques: interferometric — radio lines: galaxies

1 INTRODUCTION

The spatial distribution of matter in the large scale structure of the Universe imprints intriguing details of many fundamental quantities imperative to our understanding of the Universe. However, this matter distribution is not directly observable to us and tracers such as galaxies are needed to map the cosmic web. On large scales where perturbations are small, the clustering properties of the tracers follow the fluctuations of the underlying matter field. Large galaxy surveys such as the Sloan Digital Sky Survey (SDSS, York et al. 2000) have mapped large areas of the sky at low-redshift and aided measurements of the cosmological baryon acoustic oscillation signal (BAO, Eisenstein et al. 2005). In particular, the anisotropic galaxy clustering measurements have put constraints on various cosmological parameters (Reid et al. 2012; Chuang et al. 2013; Sanchez et al. 2013, 2014, 2016;
An alternative and rather more promising tracer is the neutral atomic hydrogen (HI) which pervades the Universe from the recombination epoch through the Cosmic reionization to the present time. During reionization, the intergalactic medium (IGM) was ionized by the first sources. And post-reionization, neutral hydrogen exists only within clouds massive enough to shield themselves from ionizing ultra-violet (UV) photons. These structures are observed as Lyman-alpha absorbers. The stellar and galaxy evolution impacts the distribution of HI in these systems, and therefore, the detection of HI can provide much-needed insights of the galaxy and stellar evolution processes. The cosmic HI can be detected with line emission at the 21cm which arises due to the spin-flip transition of the electron in the atomic hydrogen ground state. The typical temperature of HI in the post reionization epoch ranges up to thousands of Kelvin, which is higher than that of the temperature between the hyperfine states responsible for the 21cm transition. Therefore, the 21cm line transition occurs as emission, which falls within the frequency coverage of many radio telescopes, e.g. MeerKAT (Jonas & MeerKAT Team 2016), ASKAP (Johnston et al. 2008), SKA (Braun et al. 2015). Furthermore, the measured redshift of the HI emission line provides an additional measure of cosmic distance. Thus, it is possible to construct a three-dimensional HI field and therefore measure the fluctuation in the underlying matter distribution. However, the inherent weakness of this signal along with the limited bandwidth of previous telescopes has restricted the detection of HI in individual galaxies to the local Universe ($z \approx 0.1$).

Fortunately, with the Intensity mapping (IM) technique, one can construct a low angular resolution 21cm map where the emission from many unresolved galaxies is combined into a single resolution element, boosting the signal (Bharadwaj & Sethi 2001; Wyithe & Loeb 2007; Bull et al. 2015; Santos et al. 2015a, 2017). This approach is analogous to a Cosmic Microwave Background (CMB) map, without the need to detect individual galaxies. The 21cm signal is intrinsically weak compared to the various astrophysical foregrounds which are few orders of magnitude stronger. Observations of very long duration are required to achieve the required sensitivity with the added complication of maintaining system stability for such long periods. The first tentative detection of the 21cm intensity mapping signal at $z \approx 0.8$ was reported from the Green Bank Telescope (GBT) observations. The cross-correlation signal of GBT observations with DEEP2 optical galaxy survey was first detected by Chang et al. (2010); whereas Masui et al. (2013) reported the cross-correlation signal with the WiggleZ Dark energy survey. In a first-ever attempt in auto-correlation, Switzer et al. (2013) used the 21cm intensity fluctuation auto-power spectrum to constrain the neutral hydrogen fluctuation at $z \approx 0.8$.

The purpose of this paper is to demonstrate a complementary approach to single-dish IM experiments, capable of the statistical detection of HI field and therefore the fluctuations in the underlying matter field by measuring the HI power spectrum with interferometric observations. Interferometers have inherent advantages over single-dish measurements. Besides providing high angular resolutions, they are less sensitive to systematics which poses a major problem to the auto-correlation power. However, the smallest $k$-modes accessible to an interferometer is determined by the shortest baselines which may hinder probing the BAO scales. Interferometers such as CHIME (Bandura et al. 2014), TIANLAI (Xu et al. 2014) and HIRAX (Newburgh et al. 2016) are custom designed to probe the BAO scales using the 21cm signal in the redshift range $z \approx 0.5 - 2$. In this paper, we study the feasibility of detecting the cosmological 21cm signal with MeerKAT and present forecasts on the statistical measurement of HI with MeerKAT L-band ($856 < \nu < 1712$ MHz) observations on semi-linear scales. We present the sensitivity estimates at $z \approx 0.27$ from our newly developed simulation pipeline, which is our first attempt towards the measurement of HI power spectrum with MeerKAT. Our pipeline is based on the methods being developed for similar statistical measurement of HI from the Epoch of Reionization from a series of experiments at lower radio frequencies such as LOFAR (Van Haarlem, M. P. et al. 2013), GMRT (Paciga et al. 2013), PAPER (Parsons et al. 2014), HERA (DeBoer et al. 2017) and MWA (Tingay et al. 2013). We show that with MIGHTEE (MeerKAT International GHz Tiered Extragalactic Exploration, Jarvis et al. 2016), one of MeerKAT’s large survey projects, we can achieve constraints on the HI power spectrum at $z = 0.27$. There are implicit advantages with such survey projects with a specific emission line. First, it provides a one to one correspondence between observed frequency and redshift, thereby delivering a very high redshift resolution. Secondly, these are generally less time consuming compared to a spectroscopic galaxy survey which requires very high sensitivity to detect individual galaxies.

The paper is structured as follows. In the next section, we provide a theoretical overview of the statistical detection of HI signal and how interferometer measurements enable us to estimate the HI power spectrum. In section 3, we give a brief outline of MIGHTEE observations and describe the simulation pipeline to extract the HI power spectrum from MIGHTEE data. The main simulation results, along with the sensitivity estimates, are discussed in detail in section 4. In section 5, we forecast the possibility of obtaining constraints on the HI power spectrum with MIGHTEE data and finally, section 6 contains the conclusion and scopes of future work. Throughout this paper, we have used the flat ΛCDM cosmological parameters $[\Omega_m, \Omega_b, h, n_s, \sigma_8] = [0.311, 0.040, 0.677, 0.967, 0.8102]$ from Planck Collaboration et al. (2018).

2 STATISTICAL MEASUREMENT OF HI

In this section, we formulate the basis for the HI power spectrum analysis through statistical measurements. Although a power spectrum lacks visual representation of the 21cm field like an image; there are some inherent advantages in the statistical approach where we can take the advantage of the Universe being statistically isotropic. Therefore, we can in principle coherently combine the various Fourier modes of the same amplitude although different in direction - which in turn aids in improving the sensitivity. Below, we define the power spectrum following the construction of the 3D Fourier transformation of sky temperature. The sky temper-
HI IM with MIGHTEE

atute can be decomposed as: $T(\theta, \nu) = \bar{T}(\nu)[1 + \Delta T(\theta, \nu)]$; where $\bar{T}(\nu)$ and $\Delta T(\theta, \nu)$ are the isotropic and fluctuating component of the temperature distribution; $\theta$ and $\nu$ denote the position vector on the sky plane and frequency of observation respectively. In an interferometric measurement, only the fluctuating component contributes and we define its Fourier transform as:

$$\Delta T(k) = \int_{-\infty}^{\infty} d^3r \Delta T(r) e^{-ik \cdot r},$$

(1)

where $r = \{\theta, \nu, \}$ specifies the 3D position of the emission, $\nu$ being the comoving distance to the point of observation and $k$ the comoving wave vector. With interferometers, one seeks to compute the two-point correlations of the cosmological signal and the most significant correlation function is the power spectrum $P(k)$, defined as:

$$\langle \Delta T^*(k) \Delta T(k') \rangle = (2\pi)^3 \delta^3(k - k') P(k).$$

(2)

Radio interferometers calculate the spatial correlation of electric fields from the sky with the measured visibility, obtained by correlating data from each antenna pair. Under the flat-sky approximation, the visibility can be expressed as:

$$V(b, \nu) = \int A(\theta, \nu) \Delta T(\theta, \nu) e^{-i2\pi b \theta/c} d\Omega.$$  

(3)

Here, $\theta$ refers to the position on the sky, $A(\theta, \nu)$ is the primary beam response of the telescope, $b$ denotes the baseline vector in physical units corresponding to each antenna pair and $d\Omega$ being the solid angle element. The cosmological HI power spectrum can be estimated from measured visibilities in the form of ‘delay spectrum’ by the following relation (Morales & Hewitt 2004; McQuinn et al. 2006; Parsons et al. 2012a; Parsons et al. 2014; Liu & Shaw 2020):

$$P_b(k_{\perp}, k_{\parallel}) = \frac{A_e}{\lambda^2 B^2} V(b, \tau)^2 \left( \frac{\lambda^2}{2k_B} \right) \cdot$$

(4)

Here, $A_e$ and $B$ are the effective antenna area and bandwidth respectively, $\lambda$ is the wavelength at the centre of the band, $k_B$ is the Boltzmann constant, $x$ denotes the comoving distance to the redshift $z$ corresponding to $\lambda$, whereas $y$ signifies the comoving width along the reshift axis corresponding to $B$. In Equation 4, we have decomposed the wave vector $k$ into the components on the plane of the sky $k_{\perp}$ and along the line of sight $k_{\parallel}$; and they are related to the interferometric variables as:

$$k_{\perp} = \frac{2\pi b}{\lambda x}, \quad k_{\parallel} = \frac{2\pi \nu_2 h}{\nu_1} E(z) / \nu(1+z)^2$$

(5)

where, $\nu_2$ is the rest frame frequency of the $21$cm line; $H_0$ and $E(z) = [\Omega_M (1+z)^3 + \Omega_k (1+z)^2 + \Omega_{\Lambda}]^{1/2}$ are the standard cosmological parameters. $V(b, \tau)$ is the Visibility function in the delay space ($\tau = b \cdot \hat{s}/c$) obtained by delay transforming the measured visibilities with an FFT. The primary advantage of the ‘delay space’ approach (Parsons et al. 2012a, Vedantham et al. 2012; Liu et al. 2014a, Paul et al. 2016) is that for foregrounds, the HI frequency domain is smooth as they originate from continuum emissions such as the Synchrotron emission from both our galaxy and other extragalactic sources. Therefore in the Fourier space, they are restricted to fewer Fourier modes. The HI signal, on the other hand, has different characteristics in the frequency domain as the frequency is a measure of cosmological distance and therefore has significant structures in the frequency space. This feature potentially allows us to separate the HI from foregrounds. In this approach, the visibilities observed by each antenna pair are Fourier transformed along the frequency axis, which isolates the foreground contribution in the ‘delay spectrum’. The Fourier conjugate variable can be associated with the line of sight cosmological distance, and the ‘delay spectrum’ constructed from this method is capable of recovering the cosmological $3D$ HI power spectrum. One caveat in this approach is that the Fourier mode on sky plane ($k_{\perp}$) is calculated at the centre of the frequency band for each baseline. However, in the actual scenario, each physical baseline corresponds to a range of $k_{\perp}$ modes across the bandwidth. The span of the $k_{\perp}$ modes is higher with increasing baseline length and bandwidth. However, if one restricts to shorter baselines (where the sensitivity is higher due to a large number of uv points) and small frequency range, this effect is not severe and the ‘delay power spectra’ is a good approximation to the actual cosmological $3D$ HI power spectrum (Parsons et al. 2012b; Liu et al. 2014a). Also, this approach comes with an added advantage that one can work with the data from individual baselines (which is the regular format of the primary data output of a radio interferometer, i.e. visibilities).

Equation 4 assumes that the change in $k_{\perp}$ is minimal across the bandwidth for baselines considered for power spectrum calculation to justify the conversion from $\tau$ to $k_{\parallel}$ in Equation 5 (Liu et al. 2014a). This approximation holds well for short baselines and small frequency range over which the delay transform is performed.

2.1 HI signal

The prime observable in the HI intensity mapping experiments is the $21$cm emission line from neutral hydrogen. The mean brightness temperature of the HI $21$cm emission can be expressed as (Santos et al. 2015a, 2017):

$$T_h(z) \approx 566 h \left( \frac{H_0}{H(z)} \right) \left( \frac{\Omega_M}{0.003} \right) (1+z)^2 K.$$  

(6)

Here, $\Omega_H(z)$ is the neutral hydrogen density function:

$$\Omega_H(z) = \frac{\rho_{HI}(z)}{\rho_c(1+z)^3},$$

(7)

where $\rho_{HI}(z)$ and $\rho_c$ are the proper HI density and critical density of the Universe at $z = 0$ respectively. $\Omega_H(z)$ is a crucial quantity in determining the hydrogen content of the Universe at various redshifts and therefore plays a significant role in the calculation of the $21$cm brightness temperature. Several experiments have measured $\Omega_H(z)$ over a range of redshifts. Direct $21$cm observations from galaxies have measured this quantity at low redshifts (Zwaan et al. 2005; Jones et al. 2018); whereas the quasar absorption spectra in the damped Lya systems have put constraints on $\Omega_H(z)$ at higher redshifts ($z > 2$) (e.g., Rao et al. 2006; Prochaska & Wolfe 2009; Noterdaeme, P. et al. 2012; Font-Ribera et al. 2012; Zafar et al. 2013; Crighton et al. 2015; Necleman et al. 2016; Sánchez-Ramírez et al. 2016; Bird et al. 2017). The HI spectral stacking has been used to constrain the HI abundance at the intermediate redshift range 0.2 < $z$ < 2 (Lah et al. 2007; Delhaize et al. 2013; Rhee et al. 2013, 2016; Kanekar et al. 2016; Rhee et al. 2018); and
HI emission studies with ASKAP, MeerKAT and SKA are expected to explore this range in more detail.

The HI signal follows the underlying dark matter fluctuation and therefore the brightness temperature as a function of position and frequency is given by:

$$T_b(\nu, \Omega) \approx \bar{T}_b(z) \left[ 1 + b_{\mathrm{HI}} \delta_m(z) - \frac{1}{H(z)} \frac{dv}{dr} \right],$$

where $v$ is the peculiar velocity of emitters. The HI density function $\rho_{\mathrm{HI}}(z)$ and bias function $b_{\mathrm{HI}}(z)$ can be computed using the halo mass function $(dM/dN)$ and the HI mass content inside a dark matter halo of mass $M_{\mathrm{HI}}$:

$$\rho_{\mathrm{HI}}(z) = \int_{M_{\mathrm{min}}}^{M_{\mathrm{max}}} dM \frac{dn}{dM}(M, z) M_{\mathrm{HI}}(M, z),$$

$$b_{\mathrm{HI}}(z) = \frac{1}{\rho_{\mathrm{HI}}} \int_{M_{\mathrm{min}}}^{M_{\mathrm{max}}} dM \frac{dn}{dM}(M, z) M_{\mathrm{HI}}(M, z) b(M, z),$$

where $b(M, z)$ is the halo bias.

In this paper, we assume a simple power-law model of the halo mass following the prescription of Santos et al. (2015a): $M_{\mathrm{HI}}(M) = AM^\alpha$ with $\alpha = 0.6$ and $A \sim 220$ that fits both low and high redshift observations within reasonable accuracy. The scaling relations for all relevant quantities to compute the HI signal are obtained with the above formulation and are used throughout this paper (see Santos et al. (2017) for details). With all these parameters in place, the HI power spectrum in redshift space can be computed in terms of the matter power spectrum $P_M(k, z)$ and the bias function as:

$$P_{\mathrm{HI}}(k, z) = \bar{T}_b(z)^2 b_{\mathrm{HI}}(z)^2 P_M(k, z).$$

3 SENSITIVITY FOR ESTIMATING HI POWER SPECTRUM

Santos et al. (2015b) showed that SKA1-mid will have the required sensitivity for a reasonable amount of integration time to constrain the cosmological parameters; however precursor telescopes like MeerKAT should be able to integrate down to such sensitivities on deep single pointings. The MeerKAT radio telescope is located in the Karoo region of South Africa. The telescope array consists of 64 dish antennas of 13.5 meter diameter. The central core region of 1 km diameter houses 48 antennas, whereas the other 16 antennas are distributed up to a radius of 4 km from the centre. The dense core of MeerKAT facilitates higher sensitivity at low $k_L$ modes which can aid the statistical detection of HI at relevant cosmological scales using the interferometer data.

The simulation pipeline outlined in this paper aims to present realistic outcomes that can be compared with the real MIGHTEE data. Considering that, the pipeline needs to incorporate contributions that are present in the real data, which includes HI, noise and foregrounds. Along with an input HI model (described in section 2), the thermal noise can be modelled as random processes with the help of various system parameters. For foreground modelling, we choose to adopt a continuum model which we create by imaging the MIGHTEE field of interest as described in the following subsection.

3.1 MIGHTEE

In this paper, we use data from MIGHTEE survey for the sensitivity estimation. In particular, we process the single pointing COSMOS field observation with an on-source integration time of $\sim 11.2$ hours. To estimate the Noise power spectrum and therefore, the sensitivity level, one only requires the uv distribution and telescope information such as system temperature, effective area, time and frequency resolution. Also, in theory, one can compute the delay power spectrum from the calibrated visibility data itself without going to the image domain, assuming that foreground isolation is reasonably accurate in the final power spectrum. However, we perform flagging, calibration on the raw data with the purpose-built processMeerKAT$^1$ pipeline for MeerKAT data calibration; and initial processing in the image domain to obtain a continuum model to replicate the foreground contribution in our simulation pipeline. The processMeerKAT uses CASA (McMullin et al. 2007) based algorithms to flag RFI contaminated components and bad data; compute phase and flux gains from the reference calibrator observations. The raw data had an integration period of 4 seconds and spanned a total 4096 channels of 208.984 kHz channel-width in L-band. Post flagging and calibration, the data was split to a sub-band of 950 – 1150 MHz and time-averaged to 8 seconds, which was further processed for continuum imaging. As we are not using the full available band for the continuum imaging, the resulting model will lack contributions from fainter sources. However, as we will see later, the detected point sources in this band are good enough for an accurate model of the foreground contamination.

After flagging and calibration by the processMeerKAT pipeline, the following steps are used for the continuum imaging process. We apply the CASA $tclean$ and $gaincal$ tasks on the MS file for deconvolution and self-calibration respectively. We have used the multi-scale multi-frequency (MS-MFS) synthesis algorithm (Rau, U. & Cornwell, T. J. 2011) for the imaging with $nterms=2$ to estimate the intensity along with its spectral variation. To account for the non-coplanarity of the MeerKAT baselines, we also used the W-projection algorithm (Cornwell et al. 2008). To reduce the error emanating from the temporal variations in the system gain, we perform a few rounds of self-calibration with $gaincal$ (both phase and amplitude+phase) and $tclean$ loop. For the first few iterations, the number of clean iterations is set to a low number to avoid cleaning deep and detect false source components from noise pixels. With each self-calibration, the number of clean iterations is increased gradually. At the final $tclean$ task, the threshold for the clean is set at $5\sigma$ level where $\sigma$ is the standard deviation in the residual image. We create an image of size 1024 x 1024 pixels with 8 arcseconds cell size. The image size has been kept larger compared to the main field of view to include the bright sources from the sidelobes. We do not perform any primary beam correction to the image. This is actually an advantage as it means that primary beam effects will be included in our point source

$^1$ The processMeerKAT pipeline has been developed and maintained by the Inter-University Institute for Data Intensive Astronomy (IDIA) Pipelines team. For further details see https://idia-pipelines.github.io/docs/processMeerKAT
foreground model. In Figure 1, we present the central region of the image of size $1.14^\circ \times 1.14^\circ$. From the 11.2 hours of sub-band data, we obtain an rms $\sim 10\mu$Jy beam$^{-1}$. The model obtained from the final imaging step is further used as the foreground model in the power spectrum pipeline as described in the following section. As the foreground model is generated directly from the CLEAN components, the effects of primary beam are implicitly included in the model.

### 3.2 Simulation Pipeline

Each visibility measurement by the interferometer receives a contribution from system noise. This, along with the cosmological signal itself, adds to the uncertainty in $k$ space. Therefore, it is important to have a measure of thermal noise contribution to estimate the power spectrum sensitivity level. In this section, we describe the details of our simulation pipeline. For this part, we use only a small subset of the available data by splitting the data further on a narrower band of $B = 220 \times 208.984$ kHz centred at 1115.14 MHz which corresponds to $z = 0.27$. Below, we delineate the steps of our simulation pipeline in detail:

(i) Each physical baseline corresponds to a $(u,v)$ coordinate which changes after every integration interval ($t_{\text{int}} = 8$ seconds) over the period of tracking. The uv coordinates are extracted for the entire duration of the data ($t_{\text{total}} \sim 11.2$ hours) from the visibility file. Note that the $(u,v)$ points are calculated at the band center. Before flagging, its number should be: $N_{uv} = \frac{N_{\text{ant}}(N_{\text{ant}}-1)}{t_{\text{total}}}$, where $N_{\text{ant}}$ is the number of antennas used in the observation. However, we do not include the flagged baselines in our simulation. Each baseline receives contributions from the HI, foreground and noise. Therefore, for the $i$th baseline, the visibility can be expressed as: $V_i = V_{\text{HI},i} + V_{\text{FG},i} + V_{\text{N},i}$. $V_{\text{HI}}$ is the contribution from an input model HI signal which we want to recover from the final power spectrum; whereas $V_{\text{FG}}$ and $V_{\text{N}}$ are the foreground and noise components respectively.

(ii) In Figure 1, we observe that the discrete extragalactic sources dominate the radio sky in the frequency range of our interest. Other than bright radio sources, the Galactic synchrotron emission originated from the interactions of cosmic-ray electrons and the interstellar magnetic field, is expected to contribute to the total foreground budget. However, this diffuse emission has a strong presence along the Galactic plane. As the COSMOS target field is far from the Galactic disc, the contribution from the Galactic synchrotron emission is significantly low, and we do not see any considerable diffuse structure in the continuum image (Figure 1). This suggests that the diffuse Galactic component has a weak contribution on these angular scales, probably only detectable after subtracting the extragalactic point sources. Therefore in our foreground model, we only consider the bright extragalactic sources which are extracted as CLEAN components of the continuum image during the deconvolution process. Application of the source detection algorithm PyBDSF (Mohan & Rafferty 2015) on the continuum image reveals a total of 3391 extragalactic foreground sources ($\sim$ mJy radio flux density) in our continuum model.

(iii) For the system noise contribution, we generate $V_{\text{N}}$ per baseline, which we model as a Gaussian random variable. Therefore at each baseline and channel, the real and imaginary part of $V_{\text{N}}$ are generated from a random process with rms calculated by the following relation (Taylor et al. 1999):

$$\sigma_{\text{N}} = \frac{2k_B T_{\text{sys}}}{A_{\text{e}} \sqrt{\delta \nu \delta t}}$$

Here, $T_{\text{sys}}$ and $A_{\text{e}}$ are the system temperature and effective area of each antenna respectively, $k_B$ is the Boltzmann constant, $\delta \nu$ the channel width and $\delta t$ the time resolution. The corresponding values used to calculate $\sigma_{\text{N}}$ are: $\delta \nu = 208.984$ kHz, $\delta t = 8$ sec and $A_{\text{e}}/T_{\text{sys}} = 6.22$ m$^2$/K. The contribution of system noise is approximately constant across the frequency channels (and therefore along $k_\parallel$); it only varies across $k_\perp$ as the noise level depends on the baseline density (Figure 2).

(iv) Next, we generate a $uv \nu$ cube by segmenting the uv plane onto a discrete grid with cell-size $\Delta u \times \Delta v = 60\mu$; this choice of grid resolution is motivated by the primary beam size in the Fourier domain. The third axis of the $uv \nu$ cube is in frequency which is already discretized into channels of width 208.984 kHz. In Figure 2, we show the distribution of uv points. On the left panel of Figure 2, each data point corresponds to a uv pixel, and the color shows the number of uv points within each cell. On the right panel, the mean number of uv points is plotted as a function of $uv$ distance from the origin $(u = v = 0)$. On smaller $uv$ distance, the uv points are densely populated; and as $|k_\perp|$ is proportional to the $uv$ distance, it translates to higher sensitivity at small $k_\perp$ modes. It is worth mentioning here that the calibrated visibility measurements have some channels flagged due to excessive RFI presence. These gaps are reflected in the foreground model visibility as well. If we consider visibilities filled with too many flagged channels for the power spectrum calculation, it will cause a spurious flow of foreground power to higher $k_\parallel$ modes. Therefore, we only consider those $uv$ points for which the visibility measurements have at least 80% unflagged channels (out of 220).

Moreover, to minimize the contamination from the remaining flagged channels, we substitute each flagged foreground component with the foreground visibility from the nearest neighbour unflagged channel. This is to make sure that the simulated visibilities have no channels with zeros while performing the delay transformation along the frequency axis. For the real data, the same selection rule applies for assigning the baselines on the grid; and the flagged channels are replaced with the visibility entry from the nearest neighbour unflagged data channel.

(v) The visibilities within a uv pixel are then averaged assuming the sky signal to be the same across all baselines contributing to that grid point. At this stage, we also add the contribution from the input model cosmological signal $V_{\text{HI}}$ to the averaged visibility per grid point, which we generate as a random process with variance calculated from Equation 11. This ensures that the resulting ‘delay spectrum’ (Equation 4), estimated from the complex HI visibilities, agrees with the input HI power spectrum. Next, we perform the delay transformation with an FFT along the frequency axis. During the FFT, each visibility per grid-point is multiplied with a spectral weighting function (Blackman-Harris Window) to suppress the leakage of foreground power to higher $k_\parallel$ modes. The resulting $V(u,v,\tau)$ are then used in Equation 4 to compute the 3D power spectrum in $(k_\perp, k_\parallel)$. 

MNARAS 000, 000–000 (0000)
domain following the conversions listed in Equation 5 where \((u, v)\) are calculated at the grid-center. The power estimates \(P(k_\perp, k_\parallel)\) in 3D \(k\) space are further combined with an inverse variance (thermal noise) weighting to compute the 2D and 1D power spectrum \(P(k_\perp, k_\parallel)\) and \(P(k)\) respectively, where \(k_\perp = |k_\perp|\). The lowest \(k_\perp\) mode probed by the \(uv\) distribution is indicated by the smallest baseline; and for the case in hand, \((k_\perp)_{\text{min}} \approx 0.33 \text{ Mpc}^{-1}\). Therefore, we choose the bin width \(\Delta k_\perp = 0.35 \text{ Mpc}^{-1}\) for estimating \(P(k_\perp, k_\parallel)\). The lowest mode along \(k_\parallel\) and the bin width follows the bandwidth \(B\) used for FFT, and is given by \(\Delta k_\parallel \approx 0.031 \text{ Mpc}^{-1}\). To compute the 1D power spectrum \(P(k)\), we use a logarithmic binning across \(k\), i.e. the bin size gets larger with increasing \(k\); and the smallest bin width is \(\Delta k \sim 0.4 \text{ Mpc}^{-1}\).

4 RESULTS

The target 21cm signal can provide new insights into cosmology as well as the properties of low mass HI galaxies which are difficult to detect directly otherwise. However, since the signal is buried underneath strong foregrounds and noise, it is imperative to check the detected signal is indeed cosmological in nature and the processes and techniques used in the measurement are lossless. The prime objective of our simulation pipeline is, therefore, to demonstrate that we can recover a cosmological input signal in the presence of strong foregrounds and system noise. In Figure 3a, we present the 2D power spectrum \(P_{\text{HI}}(k_\perp, k_\parallel)\) for a single realization where we have considered only HI and ignored the foreground and noise contributions. This shows that the HI signal is maximum at low \(k\) and exhibits an isotropic nature.

The presence of strong foregrounds curtails the observability of the 21cm signal. Two possible approaches can be undertaken to deal with the extragalactic foregrounds. One can try to model the foregrounds with great accuracy and subtract in the image domain leaving behind only the 21cm signal and Gaussian noise. This can be particularly successful with point sources for which we already have information such as position, flux and spectral index. Another approach is to rely on the fact that foregrounds are expected to be smooth in frequency and therefore potentially distinguishable from the cosmological signal that has significant spectral structures. For the 21cm signal, the frequency is a measure of the redshift, and thereby of the line of sight.
HI IM with MIGHTEE

Figure 2. Left panel: Distribution of baselines on a 2d uv plane for 11.2 hours tracking of the COSMOS field. The uv plane is segmented onto a discrete grid with cell-size $\Delta u = \Delta v = 60\lambda$. The color represents the number of uv points on the grid, which is maximum at the centre and falls with increasing baseline length. Right panel: Baseline population as a function of uv distance. The solid line represents the average number of baselines as a function of uv distance (bin size: $\Delta uv = 100\lambda$). Also, each cell is plotted as a function of its uv distance on the lower x-axis, whereas the upper x-axis shows the corresponding $k_{\perp}$ estimates. The y-axis denotes the number of uv points. The number is high at small uv distance, implying increased sensitivity at low $k_{\perp}$ and it falls with increasing $k_{\perp}$.

Figure 3. 2d power spectrum in the unit of mK$^2$Mpc$^3$ generated from the simulation pipeline for a single realization, the dashed lines denote contours of constant radius $k$. (a) shows the case for only HI; whereas in (b), the contribution from both HI and thermal noise are considered.

distance to the emitters. Foregrounds do not have any such cosmological significance, and hence in the delay space, the foreground contribution can be restricted to the first few Fourier modes. One can take advantage of this difference and separate the signal using foreground cleaning methods (Alonso et al. 2015; Wolz et al. 2015; Switzer et al. 2015; Wolz et al. 2017) or use the so called foreground avoidance approach (Datta et al. 2010; Morales et al. 2012; Parsons et al. 2012a,b; Vedantham et al. 2012; Pober et al. 2013; Thyagarajan et al. 2013; Paul et al. 2016).

The other component, thermal noise, is inversely proportional to baseline density which is maximum at short uv values (for MeerKAT) that correspond to the lowest $k_{\perp}$ modes. In Figure 4a, we present the 2d power spectrum $P_{\text{full}}(k_{\perp}, k_{\parallel})$ for a single realization which includes all three visibility components. The thermal noise contribution is simulated from the uv distribution for the 11.2 hours track-
Figure 4. (a) Simulated 2d power spectrum (single realization) for 11.2 hours tracking of the COSMOS field, it includes contributions from the model cosmological signal, thermal noise and model foreground (Figure 1). The foreground contribution is well isolated at lower $k_\parallel$ values and the region beyond the foreground wedge is dominated by noise and the 21cm signal. The dashed black line denotes the boundary of the foreground wedge (Equation 13) and the pixels above the line are used to compute the 1d power spectra. (b) 2d power spectrum generated from the calibrated visibility data (Stokes I) using the same pipeline. Both the plots are in good agreement in terms of foreground isolation and overall power level (in units of mK^2 Mpc^{-3}).

Figure 5. Stokes V from data and simulated thermal noise comparison; the thermal noise model is calculated as the average of 1000 realizations of noise power spectrum estimated from the simulation pipeline where the noise visibility $V_N$ is used as the only input.
in Figure 4 show striking similarities; in particular, this comparison allows us to assess the efficacy of the foreground isolation approach, and compare the power levels within and beyond the foreground wedge. It also indicates the range of scales that can be probed with the current setup. However, we do not correct for noise-bias and possible instrumental systematics while calculating the power spectrum in Figure 4b. A comprehensive direction-dependent calibration strategy might be required to mitigate the errors due to sources away from the center of the field and ionospheric effects, but we do not see such effect at these levels.

We further compare the noise model in the simulation pipeline to the Stokes V power spectrum. The extragalactic foreground sources have negligible circular polarization, and therefore, the Stokes V mode is a good estimator of the thermal noise in the system. We compute the Stokes V visibilities from the calibrated dataset, and estimate the corresponding 2D and 1D power spectra. In Figure 5a, we show the power ratio between Stokes V and our thermal noise model (generated from the simulation pipeline with $V_N$ as the only input). The ratio plot shows excess power in Stokes V at low $k$, which indicates large scale polarization leakage in the calibrated data. At higher $k$, the Stokes V power spectrum exhibits noise-like features, and it is in good agreement with our thermal noise model (Figure 5b). Note that, Figure 4b and Figure 5 are presented to compare the accuracy of our simulation pipeline in estimating the noise power spectrum, we do not present any data-results further in this paper.

Next, we compute the 1d power spectra along shells of constant $|k|$. While doing so, we exclude the $(k_L,k_\parallel)$ cells contaminated by the foreground wedge. More precisely, we use the following selection criteria

$$k_\parallel < \frac{\pi H_0 E(z) \sin(\theta)}{c(1+z)} k_L$$

where $\theta$ refers to the angular extent of the MeerKAT beam. To obtain good statistics, we generate 1000 realizations from the simulation pipeline. In these cases, we generate different realizations of the noise and 21cm signal with the same foreground model.

5 DETECTABILITY

To assess the prospect of detection in detail, we investigate scenarios where we increase the duration of observation in integer multiples of the existing 11.2 hours case. We implement the increased integration hours by assuming observation on the same field taken at the same time of the existing data on different days, i.e. we do not gain any additional $uv$ points. Then the data can be coherently added on the same field without any loss of signal. We consider three hypothetical cases where we consider 2, 5 and 10 times of the existing 11.2 hours of tracking observation on the COSMOS field. For these three cases, we also generate 1000 realizations of power spectrum from the simulation pipeline. In these 1000 estimates of power spectrum, the input HI and noise components are independent realizations with the same foreground contribution. As the foreground wedge is excluded from calculating the 1d power, the foreground contribution is suppressed, and it includes dominant contributions from both the cosmological signal and system noise.

To extract the cosmological 21cm signal from the 1d power spectra, we can subtract a good thermal noise model from it. However, a single realization of thermal noise may not capture the thermal noise properties well and therefore, we generate the thermal noise power spectra from 1000 realizations and obtain a thermal noise model by calculating the mean for all three cases. The mean thermal noise power can then be subtracted from the 1d realizations, and we can have estimates of the noise-free 21cm signal. In Figure 6, we present the distribution of 1d power after subtracting the average noise power spectra from full simulations as histograms for the first four $k$ values. If the noise model is accurate, Figure 6 should manifest the distribution of the extracted cosmological signal which exhibits a Gaussian profile for all the cases considered here. The histogram curve gets narrower with increasing integration time as the variance decreases due to decrement in thermal noise power. In Figure 7, we present the mean of the distributions shown in Figure 6 with error bars which are calculated as the corresponding standard deviations. Figure 7 clearly indicates that we are able to extract the input cosmological HI signal with reasonable accuracy with our pipeline (no bias).

So far, we have considered deep integration on a single pointing by increasing our fiducial observation case of 11.2 hours. We have multiplied the integration time by $N_{\text{mult}} = 2, 5, 10$ and observed that the thermal noise power spectra drops as $P_k(11.2\text{hours})/N_{\text{mult}}$. However, in the absence of long observation on a single pointing, sensitivity can be improved by averaging power spectrum estimates from multiple independent fields. The MIGHTEE survey aims to study four well-known fields from the southern hemisphere: COSMOS, XMM-LSS, ECDFS and ELAIS-S1 (Jarvis et al. 2016) with multiple pointings to cover 20 degree$^2$ sky area. When the power spectrum measurements from independent fields are combined, the uncertainty in the mean power spectrum reduces as $\sqrt{N_{\text{field}}}$, where $N_{\text{field}}$ is the number of independent fields. The MIGHTEE survey aims to achieve $\sim 1\mu$Jy sensitivity over the L-band. This can be achieved with approximately 1000 hours of on source observation duration across all four fields. In Table 1, we provide the power spectrum constraints that can be achieved with single COSMOS pointing of 22.4 hours, as well as the full MIGHTEE survey (Figure 8). For the full survey, we have assumed uniform distribution of independent pointings across the full survey area. Please note that the actual sensitivity will depend on the observation strategy such as number of independent pointing and integration time per pointing. In Table 1 and Figure 8, we provide ballpark estimates of the full MIGHTEE survey capability. These results give a total signal-to-noise ratio (SNR) $\approx 4.92$ on the COSMOS field (22.4 hours integration) in the range $0.3 < k < 11$ Mpc$^{-1}$; whereas the full MIGHTEE survey is capable to achieve a SNR $\approx 49$ for the same $k$ range.

6 CONCLUSION AND FUTURE SCOPES

Intensity mapping of the neutral Hydrogen line is a promising avenue to probe the large scale structure of the Universe and provide precision cosmological constraints. MeerKAT
Figure 6. Distribution of 1d power from 1000 realizations for the first four $k$ values after subtracting the average noise spectrum as histograms. The expected values of the 21cm power spectrum are shown as vertical dashed lines in each subplot. The power on the x axis has units of mK$^2$Mpc$^3$).

Figure 7. Extracted HI power spectrum with 1σ errorbars from the distribution of average noise subtracted 1d power spectra (Figure 6). For comparison, the model cosmological signal with shot noise (appendix A) are included too.
with its single dish capabilities can probe scales > 1 degree. As an interferometric array, MeerKAT has a dense core and therefore can also probe the semi-linear cosmological scales, complementary to the single-dish intensity mapping. MIGHTEE, one of MeerKAT’s large survey projects, aims to provide radio continuum, spectral line and polarisation information. In this paper, we have shown that MeerKAT has the unique capability to make a statistical detection of HI on semi-linear scales, the first of its kind, with the existing and planned MIGHTEE observations on well-known fields.

We have developed a simulation pipeline that can emulate the MIGHTEE observations. The current simulation includes contributions from the HI signal, noise and extragalactic point sources on the COSMOS field. With the current setup, we verify the possibility to detect the HI power spectrum at \( z = 0.27 \) for \( k > 0.3 \text{ Mpc}^{-1} \). We have shown the constraints that can be achieved on the HI power spectrum with a single pointing deep integration on the COSMOS field. As MIGHTEE includes observations from multiple fields, the power spectrum can also be estimated independently from those fields and combined to further reduce the uncertainty.

The next step will be to try a detection with the actual data. As we go deeper in sensitivity, further improvements might be necessary to the simulation in order to validate the signal extraction (e.g. the effects of the MeerKAT primary beam on the HI signal contamination). We therefore plan to continuously improve this simulation pipeline using insights from the calibrated data.

### ACKNOWLEDGEMENT

MGS and SP acknowledge support from the South African Square Kilometre Array Project and National Research Foundation (Grant No. 84156). We acknowledge the use of the Ilifu cloud computing facility - www.ilifu.ac.za, through the Inter-University Institute for Data Intensive Astronomy (IDIA). This work was carried out using the data processing pipelines developed at the IDIA and available at https://idia-pipelines.github.io. NM acknowledges support from the Bundesministerium für Bildung und Forschung (BMBF) award 05A20WM4. The MeerKAT telescope is operated by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation.

### DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding authors (SP, MGS).

### REFERENCES

Alam S., et al., 2017, Monthly Notices of the Royal Astronomical Society, 470, 26172652
Alonso D., Bull P., Ferreira P. G., Santos M. G., 2015, MNRAS, 447, 400
Anderson L., et al., 2014a, Monthly Notices of the Royal Astronomical Society, 439, 83
Anderson L., et al., 2014b, Monthly Notices of the Royal Astronomical Society, 441, 24
Bandura K., et al., 2014, in Stepp L. M., Gilmozzi R., Hall H. J., eds, Vol. 9145, Ground-based and Airborne Telescopes V. SPIE, pp 738 – 757, doi:10.1117/12.2054950, https://doi.org/10.1117/12.2054950
Beutler F., et al., 2016a, Monthly Notices of the Royal Astronomical Society, 464, 34093430
Beutler F., et al., 2016b, Monthly Notices of the Royal Astronomical Society, 466, 22422260
Bharadwaj S., Sethi S. K., 2001, Journal of Astrophysics and Astronomy, 22, 293
Bird S., Garnett R., Ho S., 2017, MNRAS, 466, 2111
Braun R., Bourke T., Green J. A., Keane E., Wagg J., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 174
Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, ApJ, 803, 21
Castorina E., Villaescusa-Navarro F., 2017, Monthly Notices of the Royal Astronomical Society, 471, 17881796
Chang T.-C., Pen U.-L., Bandura K., Peterson J. B., 2010, Nature, 466, 463
Chuang C.-H., et al., 2013, MNRAS, 433, 3559
Cornwell T. J., Golap K., Bhatnagar S., 2008, IEEE Journal of Selected Topics in Signal Processing, 2, 647
Crighton N. H. M., et al., 2015, MNRAS, 452, 217
Datta A., Bowman J. D., Carilli C. L., 2010, The Astrophysical Journal, 724, 526
DeBoer D. R., et al., 2017, Publications of the Astronomical Society of the Pacific, 129, 045001

**Table 1.** Table summarizing the estimated power spectrum constraints with the MIGHTEE survey.

| \( k \) (Mpc\(^{-1}\)) | \( P_{HI}(k) \) (mK\(^2\)Mpc\(^{-3}\)) | Power spectrum (1σ estimates) | \( P_{HI}(k) \) (mK\(^2\)Mpc\(^{-3}\)) | \( P_{HI}(k) \) (mK\(^2\)Mpc\(^{-3}\)) |
|---|---|---|---|---|
| 0.49 | 13.55 | 18.428 | 1.846 |
| 0.87 | 7.68 | 5.945 | 0.596 |
| 1.54 | 4.25 | 2.508 | 0.251 |
| 2.74 | 2.25 | 0.856 | 0.086 |
| 4.75 | 1.38 | 0.489 | 0.049 |
| 8.45 | 1.04 | 0.507 | 0.051 |

**Figure 8.** Expected constraints with full MIGHTEE survey
APPENDIX A: SHOT NOISE

The shot noise power due to Poisson fluctuation in halo number is given by (Bull et al. 2015):

\[ P^{\text{shot}}(z) = \left( \frac{T_b(z)}{\rho HI(z)} \right)^2 \int_{M_{\text{min}}}^{M_{\text{max}}} dM \frac{dn}{dM} M^2 HI(M) \]  

(A1)

We model the HI mass within a halo of mass M for the calculation of shot noise as in Castorina & Villaescusa-Navarro (2017): Lepori et al. (2018):

\[ M HI(M, z) = C (1 - Y_p) \frac{\Omega_b}{\Omega_m} \exp \left[-\frac{M_{\text{min}}}{M} \right] M^\alpha \]  

(A2)

where \( Y_p = 0.24 \) is the helium fraction, \( \alpha \) is a free parameter and C is the normalization constant. \( M_{\text{min}} \) denotes the halo mass limit below which HI abundance is suppressed due to processes like the photoionization by UV background, galactic winds, etc. The theoretical halo mass function uses the N-body simulation best fit from Tinker et al. (2008).

MNRAS 000, 000–000 (0000)