Neutral hydrogen intensity mapping on small scales using MeerKAT

A thesis submitted in partial fulfilment of the requirements for the degree of

Magister Scientae

in the

Department of Physics & Astronomy

The University of the Western Cape

Author: Mogamad-Junaid Townsend
(Student Number: 3542824)

Supervisor: Prof. Mario Santos

Co-Supervisor: Dr. Sourabh Paul

February 22, 2021

Declaration

I, ***Mogamad-Junaid Townsend***, declare that this thesis, ***Neutral hydrogen intensity mapping on small scales using MeerKAT***, and the work presented in it are my own, that it has not been submitted before for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged as complete references.

I also declare that I am a co-author on the paper, ***HI intensity mapping with the MIGHTEE survey: power spectrum estimates***, discussed and cited in this thesis as (Paul et al., 2020), and also from which many results and conclusions are derived.

Full Name: Mogamad-Junaid Townsend ....

Signature: ....

Date: 22/02/2021 ....

Acknowledgements

I would firstly like to thank Prof. Mario Santos, my supervisor and mentor, for giving me the opportunity to do meaningful research as a member of the Centre for Radio Cosmology (CRC). Your guidance and support has been invaluable.

I would also like to thank Dr. Sourabh Paul, my co-supervisor, for all the insights, useful discussions and encouragement every step of the way.

I wish to thank the South African Radio Astronomy Observatory (SARAO), the National Research Foundation (NRF), the Centre for Radio Cosmology and the Astronomy research group at the University of the Western Cape for providing financial support, as well the opportunity to do meaningful and exciting research.

I would like to thank the Inter-University Institute for Data Intensive Astronomy (IDIA) for the use of the ilifu cloud computing facility. Further, I would like to thank the support team at IDIA for all the assistance and technical support.

To my colleagues and friends - Xola Ndaliso, Narusha Isaacs, Eliab Malefahlo, Brandon Engelbrecht and the rest of the members of the postgraduate office - I wish to say thank you for making the time working with you all a privilege and a joy.

To my family I wish to offer special thanks for all the love and support. This especially to my mother, Shanaaz, my father, Waleed, and sister, Rehana, for always being there for me through the struggles and the triumphs.

Abstract

Neutral hydrogen intensity mapping on small scales using MeerKAT

M-J Townsend

M.Sc. Thesis

Department of Physics & Astronomy
The University of the Western Cape

In the post-reionisation universe, intensity mapping (IM) with the 21 cm line of neutral hydrogen (HI) provides a potential means of probing the large-scale structure of the universe. With such a probe, a wide variety of interesting phenomena such as the Baryon Acoustic Oscillations (BAO) and Redshift Space Distortions (RSD) can be studied. The MeerKAT telescope has the potential to make full use of this technique, especially in the single-dish mode, which will probe the scales relevant to BAO and RSD. A useful complementary of this is HI IM with MeerKAT in interferometer-mode, which will enable the extraction of cosmological information on semi-linear and small scales. In this study, full end-to-end simulations of interferometric observations with MeerKAT for HI IM were developed. With this, the power spectrum extraction was analysed using the foreground avoidance technique. This took into account the foreground wedge from point source contamination extracted from real MIGHTEE COSMOS data, as well as RFI flagging. The errors on the power spectrum estimator were then calculated through a Monte Carlo process using 1000s of realisations of both the thermal noise and HI signal. In doing so, precision constraints on the HI power spectrum are found at $z = 0.27$ on scales $0.4 < k < 10$ Mpc^{-1} for mock visibility data sets which contain the HI signal contaminated by noise, mimicking the MIGHTEE COSMOS field for total observation times $\gtrsim 20$ hours. These results illustrate the potential of doing precision cosmology with MeerKAT's MIGHTEE survey and interferometer-mode HI IM.

Keywords: Cosmology; MeerKAT; Interferometer; Intensity Mapping

February 22, 2021

Contents

List of Figures

List of Tables

1 Introduction

The last century has seen a significant growth in our understanding of the Universe, from the Cosmic Microwave Background (CMB) released during Recombination 400,000 years after the Big Bang up until the large-scale structure and accelerated expansion of the Universe due to dark energy in the late Universe. Despite this progress, we are yet to observe most of the volume of the Universe which is observable. The sheer abundance of hydrogen in the Universe is a useful means of filling out this observable volume. Its ubiquity, coupled with its 21 cm line allows us to probe the Universe from the period when the first structures started forming (Pritchard & Loeb, 2012). Along with this epoch, this spin-flip transition of hydrogen can be used to study the matter content of the late Universe and to probe important properties of the Universe such as its ionisation state and temperature (Liu & Shaw, 2020).

1.1 21 cm cosmology

The 21 cm spectral line enables the mapping of the Universe in three dimensions, with redshift providing line-of-sight distance information. In particular, the 21 cm line gives access to much higher redshift resolution than most other probes of cosmology due to the ease of obtaining higher spectral resolution using radio interferometers (Liu & Shaw, 2020).

Despite the relative infancy of 21 cm cosmology, there have already been exciting and promising results. These include the reported measurement at $z \sim 17$ which might be explained by 21 cm absorption (Bowman et al., 2018). This was an important and unexpected result, especially given the high redshift of the measurement as well as the unexpected depth of the absorption profile. In addition, low-frequency interferometers have placed ever tighter upper limits on the 21 cm power spectrum at redshifts in and around the Epoch of Reionisation (EoR). These include reported results from observations with the Murchison Widefield Array (MWA) (a recent example, Trott et al., 2020, at $z \sim 6.5$ - 8.7), Donald C. Backer Precision Array for Probing the Epoch of Reionisation (PAPER) (for example, Jacobs et al., 2014, in the range $7.5 < z < 10.5$), and the LOw Frequency Array (LOFAR) (for example, on Cosmic Dawn, Gehlot et al., 2019, at redshift $z = 19.8 - 25.2$), as well as an interesting result at $z \sim 18$ from the Owens Valley Long Wavelength Array (OVRO-LWA) (Eastwood et al., 2019).

While much of the research in 21 cm cosmology is focused on the study of the dark ages and epoch of reionisation (EoR), the post-reionisation era has seen many developments and exciting results. Using the Green Bank Telescope (GBT), a detection of fluctuations in the 21 cm signal was made through cross-correlations of 21 cm intensity maps and the DEEP2 optical galaxy survey in the redshift range $z \sim 0.53$ - 1.12 (Chang et al., 2010) and WiggleZ at $z \sim 0.8$ (Masui et al., 2013). Additional results include upper limits on the 21 cm auto-power spectrum and the use of the combination of auto- and cross-correlation results to constrain $\Omega_{\mathrm{HI}} b_{\mathrm{HI}}$ (Switzer et al., 2013). In a similar manner, the Parkes Radio Telescope has been used to measure 21 cm fluctuations in cross-correlation with the 2dF galaxy survey at $0.057 < z < 0.098$ (Anderson et al., 2018). While the post-reionisation results discussed here are from 21 cm intensity mapping with single-dish radio experiments, a possible complementary approach can be taken using radio interferometers such as MeerKAT. In doing so, the inherent Fourier properties of interferometric measurements can be taken advantage of and used to study the neutral hydrogen (HI) content, and therefore the distribution of matter, in the late Universe. It has the added complement to single-dish experiments in that it probes smaller scales due to better angular resolution and thus is a novel way of making a statistical detection of HI at small (non-linear to fully linear) cosmological scales.

1.1.1 21 cm line fundamentals

Predicted in 1942 by Hendrik C. van de Hulst (van de Hulst, 1945) and first detected by Ewen and Purcell in 1951 (Ewen & Purcell, 1951), the hyperfine splitting of the ground state of hydrogen occurs due to an interaction of the magnetic moments of the proton and electron. The parallel alignment of spin states in the hydrogen atom has a higher energy than the anti-parallel alignment, resulting in the emission (or absorption) of a photon with an energy, $\Delta E = 5.9 \times 10^{-6}$ eV. This energy corresponds to a frequency of 1420 MHz and wavelength of 21.1 cm, for which it is commonly referred to as the 21 cm line. This spin-flip transition is shown schematically in Figure 1.1, demonstrating the transition of the electron from spin-up to spin-down and the subsequent emission of a photon at a wavelength of 21 cm.

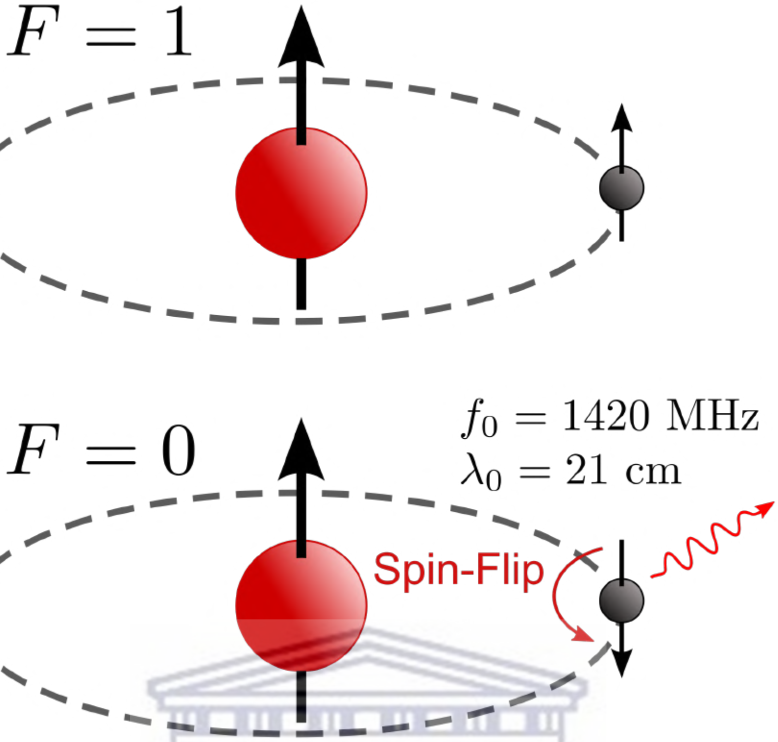

Figure 1.1: Schematic demonstrating the spin-flip transition of neutral hydrogen. As the electron transitions from spin-up to spin-down, a photon is emitted at a wavelength of 21 cm. Alternatively, the absorption of a photon could instead occur, resulting in the transition from spin-down to spin-up. Schematic taken from Wikipedia.

The 21 cm line can in principle be observed over a range of epochs, spanning from the Dark Ages (the period following recombination - when the CMB was released) and Cosmic Dawn (the epoch postulated as the phase of the Universe when the first stars and galaxies started forming) up until the Epoch of Reionization (EoR). However, the focus of this study is on the HI content of the Universe after reionisation, when most of the neutral hydrogen is contained in damped Lyα systems (Pritchard & Loeb, 2012). This is shown schematically along with the other epochs of the Universe's history which can be studied with the 21 cm line in Figure 1.2. At the end of reionisation, the mean signal is relatively low, but has a residual in emission which comes from the damped Lyα systems (usually HI galaxies). This is shown qualitatively at the right-hand end of Figure 1.3.

In order to study the 21 cm line, a quantity known as the spin temperature (and often referred to as the excitation temperature of the 21 cm line), T_S, which describes

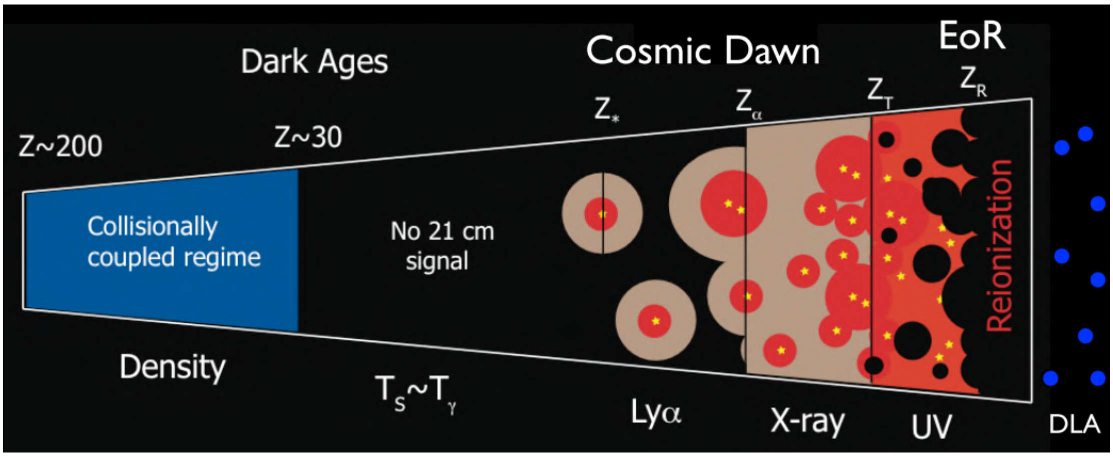

Figure 1.2: Schematic showcasing the different phases of the 21 cm signal. It is analogous to Figure 1.3 in that it shows the physical changes which occur in the signal from before the first stars and galaxies start forming up until the end of reionisation and post-reionisation epochs. Of particular importance in this study is the residual signal sources by neutral hydrogen in galaxies (the so-called damped Lyα systems). Schematic taken from Pritchard & Loeb (2012).

the number density of hydrogen atoms in the two spin states, is used. Following the prescription presented in Liu & Shaw (2020), the spin temperature can be defined as (Furlanetto et al., 2006; Pritchard & Loeb, 2012)

$$\frac{n_1}{n_0} = 3\exp\left(-\frac{T_\star}{T_S}\right) = 3\exp\left(-\frac{h\nu_{21}}{k_B T_S}\right), \tag{1.1}$$

where the factor of 3 arises from the relative degeneracy of the states, n_1 and n_0 denote the number of atoms in the excited and ground hyperfine states, respectively, h denotes Planck's constant, k_B Boltzmann's constant and $\nu_{21} = 1420$ MHz being the rest frequency of the 21 cm line.

The physics underlying the 21 cm signal depends on the radiative transfer through gas along the line-of-sight. Hence, it is important to note that the brightness temperature of the 21 cm line is observed through the contrast of a radiation background (usually the CMB) and the spin temperature of neutral hydrogen. We therefore see the 21 cm line in absorption when the spin temperature is lower than the CMB temperature, and in emission for the opposite. These two scenarios result in the observation of a deficit compared to what we expect.

After reionisation, the neutral fraction of hydrogen is relatively low at about $x_{HI} \approx$

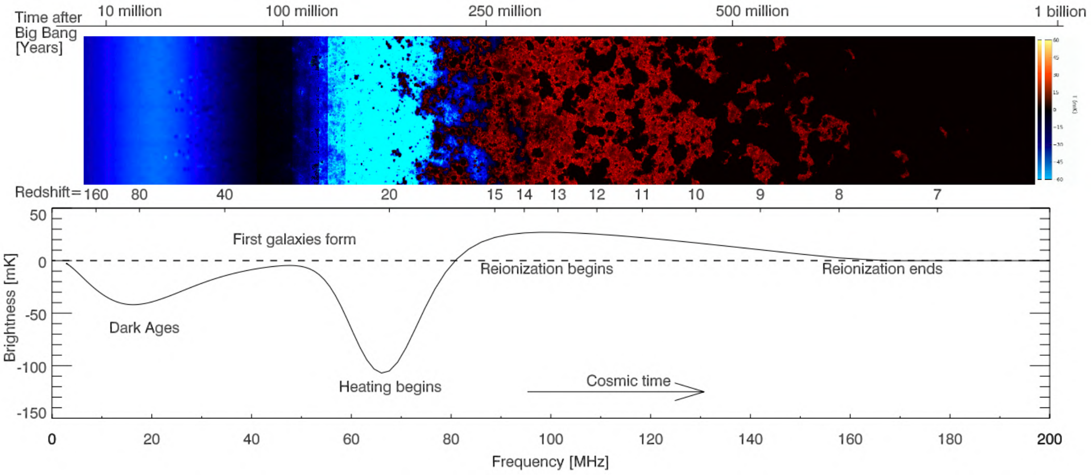

Figure 1.3: The early Universe evolution of the 21 cm signal. The top panel shows the time evolution of the 21 cm brightness fluctuations from shortly before the first luminous objects started forming in the Universe, up until when the reionisation epoch is over. The panel below shows the expected evolution of the global 21 cm signal corresponding to the fluctuations shown in the panel above. One can clearly see that the signal becomes small after the end of reionisation as most of the neutral hydrogen content is located in damped Lyα systems in the post-reionisation epoch. Figure taken from Pritchard & Loeb (2012).

0.02 (Villaescusa-Navarro et al., 2018). The remaining neutral content is contained in dense systems that have managed to shield against the ionising background which ionised most of the neutral hydrogen during the EoR. These dense systems are usually galaxies which contain neutral HI gas. This HI is either part of the cold ($T \lesssim 100$ K) or warm ($T \gtrsim 5000$ K) neutral medium (Liu & Shaw, 2020). In either case, the temperature of the gas is warmer than the CMB temperature and thus the 21 cm signal is seen in emission.

To model the physics that describes the 21 cm line in the post-reionisation epoch, the methodology discussed in the appendices of (Bull et al., 2015) is employed for the full description, since it considers the nature of the signal in the post-reionisation era from a phenomenological as well as observational standpoint.

Considering a clump of HI with number density $n_{\rm HI} = n_0 + n_1$ and assuming $T_S \gg T_\star$, Equation 1.1 becomes

$$n_1 \simeq 3n_0 = \frac{3}{4} n_{\rm HI}. \tag{1.2}$$

The emissivity of the clump can be expressed as

$$j_1 = \frac{A_{10} h \nu_{21}}{4\pi} n_1 \phi(\nu), \tag{1.3}$$

where $A_{10} \simeq 2.85 \times 10^{-15}$ s^{-1} (Pritchard & Loeb, 2012; Liu & Shaw, 2020) is the spontaneous emission Einstein coefficient and $\phi(\nu)$ is the line profile, assumed to be very narrow with a width of $d\nu$. The luminosity of the clump can then be expressed as

$$dL = \frac{3}{4} A_{10} h \nu_{21} n_{\mathrm{HI}} \phi(\nu) d\nu dA dr, \tag{1.4}$$

where dA dr is the volume of the clump, dr being specifically along the line-of-sight, ν evaluated in the clump's rest frame, all on condition that n_{HI} denotes the comoving number density of the clump. As mentioned earlier, the 21 cm line is seen in emission since the gas (and therefore spin) temperatures are well above the CMB temperature. Thus, absorption can be ignored and it follows that the total 21 cm intensity follows directly from Equation 1.4. The total flux against the background radiation (CMB) from an object at redshift z is then given as

$$dF = \frac{3 h \nu_{21} A_{10}}{16\pi \left(1+z\right)^2 r^2(z)} n_{\mathrm{HI}} \phi(\nu) d\nu dA dr, \tag{1.5}$$

with

$$(1+z) = \frac{\nu_{21}}{\nu}, \tag{1.6}$$

denoting the Doppler relation between frequency and redshift. The brightness, I, of the clump can be defined via the total flux, dF, as

$$dF \equiv I d\Omega d\nu_{\mathrm{o}}. \tag{1.7}$$

It is common practice to express the intensity in terms of the brightness temperature (Pritchard & Loeb, 2012). Using the Rayleigh-Jeans relation, $I = 2k_B T_b \nu_{21}^2 / c^2$, and combining equations 1.6 and 1.7, we have

$$\frac{3h\nu_{21}A_{10}}{16\pi\left(1+z\right)^{2}r^{2}(z)}n_{\mathrm{HI}}\phi(\nu)d\nu dAdr=\frac{2k_{B}T\nu^{2}}{c^{2}}d\Omega d\nu_{\mathrm{o}}, \tag{1.8}$$

which then finally becomes an expression for the brightness temperature:

$$T_{b}=\frac{3hc^{3}A_{10}}{32\pi k_{B}\nu_{21}^{2}}\frac{\left(1+z\right)^{2}}{H(z)}n_{\mathrm{HI}}. \tag{1.9}$$

To get to Equation 1.9, the line width $\frac{d\nu}{1+z}$ is assumed to be much smaller than the observed frequency interval $d\nu_{\mathrm{o}}$ and that $dA=r^{2}d\Omega$, and $dr=\lambda_{21}(1+z)/H(z)d\nu_{\mathrm{o}}$. It is further assumed that the line profile can be approximated as $\phi\simeq\frac{1}{d\nu}$. The comoving number density is given by

$$n_{\mathrm{HI}}=\Omega_{\mathrm{HI}}\frac{\rho_{c,0}}{m_{p}}\left(1+\delta_{\mathrm{HI}}\right), \tag{1.10}$$

where m_p is the mass of a proton, Ω_{HI} the comoving HI fraction, δ_{HI} the HI density contrast, and $\rho_{c,0}=3H_0^2/8\pi G$ is the critical density of the Universe today (i.e. at $z=0$).

It is important to understand the brightness temperature's evolution in redshift, and by extension the redshift evolution of the quantities on which it depends such as the HI density, Ω_{HI} and bias, b_{HI}. Assuming the HI luminosity in a given volume (with solid angle $\Delta\Omega$ and frequency interval $\Delta\nu$) is proportional to the HI mass in the volume, M_{HI}. The spin temperature will be much greater than the background temperature if all the HI in the volume contributes to the 21 cm signal, resulting in a brightness temperature from the volume:

$$T_{b}(\nu)=\frac{3.23\times10^{-4}}{\Delta\Omega\Delta\nu}\frac{M_{\mathrm{HI}}}{\left(1+z\right)^{2}D_{A}^{2}(z)}, \tag{1.11}$$

with proper volume given by

$$V=\Delta\Omega\Delta\nu\frac{\left(c/\nu\right)D_{A}^{2}}{H+dv/ds}, \tag{1.12}$$

where D_A is the angular diameter distance, H denotes the usual Hubble parameter and dv/ds the proper gradient of the peculiar velocity along the line-of-sight.

Since neutral hydrogen is primarily found inside of galaxies after reionisation, shielding it from ionising radiation, it is accurate to relate the HI mass to the underlying halo mass, and therefore, to relate the HI signal to the underlying matter density field, so that the HI emission acts as a tracer of not only the distribution of HI, but also as a biased tracer of the matter distribution. To do this, the assumption is made that a dark matter halo of mass M contains galaxies (at least one) with a total HI mass, M_{HI}. Further it is assumed that this HI mass is only a function of the halo mass and redshift, i.e. $M_{\mathrm{HI}} = M_{\mathrm{HI}}(M, z)$. Despite some expected level of fluctuation in the relation between the HI and halo mass, this deterministic relation is a good fit for 21 cm intensity mapping experiments, which will have low resolution pixels, and therefore a reasonable amount of HI galaxies per pixel. This would be able to average out any fluctuations and therefore enable the use of this deterministic relation (Santos et al., 2015; Bull et al., 2015). Further, Santos et al. (2015); Bull et al. (2015) highlight that the position-independence of the HI mass function can be accounted for by the averaging over many halos at the scales of interest in 21 cm intensity mapping experiments.

The signal can now be related to the underlying dark matter field using the mass function, M_{HI}. The number of halos of mass M in an observed volume element is given by $[1 + b(M,z)\delta_M(z)]\,\frac{dn}{dM}dM\,V$, where δ_M denotes the underlying dark matter fluctuation at redshift z, b denotes the halo bias and $\frac{dn}{dM}$ denotes the proper halo mass function. The observed brightness temperature is then obtained by integrating over all possible masses, which yields

$$T_b(\nu) = \frac{\alpha}{(1+z)} \frac{\rho_{\mathrm{HI}}(z)\,[1 + b_{\mathrm{HI}}\delta_M(z)]}{(H + dv/ds)\,(1 - v/c)}, \tag{1.13}$$

where $\alpha = 2.21 \times 10^{-27}$ (Bull et al., 2015). Further, the halo mass function and HI mass inside the halo can be used to calculate the proper HI density, ρ_{HI}, and HI bias:

$$\rho_{\mathrm{HI}}(z) = \int_{M_{\mathrm{min}}}^{M_{\mathrm{max}}} dM \frac{dn}{dM} M_{\mathrm{HI}}(M, z) \tag{1.14}$$

and

$$b_{\mathrm{HI}}(z) = \rho_{\mathrm{HI}}^{-1} \int_{M_{\mathrm{min}}}^{M_{\mathrm{max}}} dM \frac{dn}{dM} M_{\mathrm{HI}}(M,z) b(M,z). \tag{1.15}$$

Assuming that the peculiar velocity gradient as well as the v/c term are small for the large pixels considered as well as rewriting in terms of the fractional density,

$$\Omega_{\mathrm{HI}}(z) \equiv (1+z)^{-3} \frac{\rho_{\mathrm{HI}}(z)}{\rho_{c,0}}, \tag{1.16}$$

we get the expressions for the full and mean brightness temperatures of the 21 cm signal, respectively (Bull et al., 2015; Santos et al., 2015, 2017):

$$T_b(\nu, \Delta\Omega) \approx \overline{T}_b(z) \left[1 + b_{\mathrm{HI}}(z)\delta_M(z) - \frac{1}{H(z)} \frac{dv}{ds} \right] \tag{1.17}$$

$$\overline{T}_b(z) \approx 566h \left(\frac{H_0}{H(z)} \right) \left(\frac{\Omega_{\mathrm{HI}}(z)}{0.003} \right) (1+z)^2 \, \mu\mathrm{K}. \tag{1.18}$$

While there are many well described and parameterised fits for the HI mass function (see, for instance in, Villaescusa-Navarro et al., 2018; Padmanabhan et al., 2015, 2017; Camera & Padmanabhan, 2020, where this is discussed in detail), here a straightforward approach is taken which assumes that M_{HI} is related to the halo mass via a proportionality factor which can be fitted to data (Bull et al., 2015; Santos et al., 2015). Not all halos will contain galaxies with HI mass and therefore have to be accounted for. To do this, the assumption is made that only halos with circular velocities, $30 \leqslant v_c \leqslant 200$ kms^{-1} are able to host neutral hydrogen. The halo mass is then found through the relation to the circular velocity

$$v_c = 30\sqrt{1+z} \left(\frac{M}{10^{10} M_\odot} \right)^{1/3} \mathrm{kms}^{-1}. \tag{1.19}$$

This relation fails to fit well at high redshifts; a more accurate relation would be to take the proportionality of the mass as a function of redshift and then relate this to the HI mass (Bull et al., 2015). In this study, however, a power-law independent of redshift was chosen for the mass relation (motivated by, for instance, Bull et al., 2015; Santos et al., 2015, 2017):

$$M_{\rm HI} = AM^{\alpha}. \tag{1.20}$$

As was shown in Santos et al. (2015); Bull et al. (2015) and then used in Santos et al. (2017), $\alpha \simeq 0.6$ and $A \sim 220$ are the values chosen as motivated by constaints at $z = 0.8$ from Switzer et al. (2013). With all the relevant quantities describing the distribution of HI in the post-reionisation epoch discussed above, the theoretical framework of the HI power spectrum can now be delineated.

1.1.2 The HI power spectrum

Having described the halo model in relation to the distribution of HI in the post-reionisation Universe, the next step would be to describe the power spectrum of the 21 cm intensity fluctuations. This is especially important when it comes to understanding what can be learnt from intensity mapping surveys. In addition, since this study employs an approach using interferometry, it allows access to small scales which are semi- to fully non-linear. The power spectrum is then modelled using the halo model described above as well as some non-linear considerations. In order to do this, one requires knowledge of the linear matter power spectrum, $P_M(k, z)$, the halo mass function, $\frac{dn}{dM}$, and bias, $b(M, z)$, as well as the HI mass function, $M_{\rm HI}$, and the associated quantities such as the HI bias and density profile described above (Villaescusa-Navarro et al., 2018).

With these constituents, the full non-linear HI power spectrum can be written as the sum of the 1-halo and 2-halo terms (this is discussed thoroughly in Padmanabhan et al., 2015, 2017; Villaescusa-Navarro et al., 2018, where the halo model is also employed):

$$P_{\rm HI}(k, z) = P_{\rm HI,1h}(k, z) + P_{\rm HI,2h}(k, z). \tag{1.21}$$

The terms are defined as (see Villaescusa-Navarro et al., 2018)

$$P_{\mathrm{HI,1h}}(k,z) = \frac{1}{(\rho_{c,0}\Omega_{\mathrm{HI}}(z))^2}\int_0^\infty dM\frac{dn}{dM}M_{\mathrm{HI}}^2(M,z)|u_{\mathrm{HI}}(k|M,z)|^2 \tag{1.22}$$

and

$$\begin{aligned} P_{\mathrm{HI,2h}}(k,z) =& \frac{P_M(k,z)}{(\rho_{c,0}\Omega_{\mathrm{HI}}(z))^2} \\ & \times\left[\int_0^\infty dM\frac{dn}{dM}b(M,z)M_{\mathrm{HI}}(M,z)|u_{\mathrm{HI}}(k|M,z)|\right]^2, \end{aligned} \tag{1.23}$$

where $P_M(k,z)$ is the linear matter power spectrum and $u_{\mathrm{HI}}(k|M,z) = \frac{\tilde{\rho}_{\mathrm{HI}}(k|M,z)}{M_{\mathrm{HI}}(M,z)}$ is the normalised HI density profile in Fourier space, where the HI density profile, $\tilde{\rho}_{\mathrm{HI}}(k|M,z)$, forms an integral part of the description of the structure of the HI distribution and is distinct from the proper HI density given in Equation 1.14. Equation 1.21 essentially describes the power spectrum of neutral hydrogen at the scales which are relevant to understanding its distribution in galaxies and thus contains both cosmological and astrophysical information (Camera & Padmanabhan, 2020).

While this expression for the HI power spectrum is accurate and comprehensive, no analysis processes, such as running hydrodynamical simulations, are performed to model it. Instead, a model of the HI power spectrum taking the form (Pourtsidou, 2016):

$$P_{\mathrm{HI}}(k,z) = \overline{T}_b^2(z)b_{\mathrm{HI}}^2P_M(k,z), \tag{1.24}$$

with $\overline{T}_b$, given by Equation 1.17, and b_{HI}, given by Equation 1.15, is used in this study. The specific model of the matter power spectrum, $P_M(k,z)$, was generated in the CAMB software package (Lewis et al., 2000). In addition, all non-linear information which would have been incorporated through the 1-halo and 2-halo terms are included in this matter power spectrum model from CAMB. While the model discussed here is reasonable for dealing with dark matter non-linearity in real space, the non-linearity in redshift space is not adequately modelled by it. To do this, one has to consider the contribution from non-linear redshift space distortions (known as the 'Fingers-of-God' effect) which will introduce modifications to the theoretical HI signal (refer to Sarkar & Bharadwaj (2018) and Sarkar & Bharadwaj (2019) for detailed discussions on this). This contribution has, however, been left to future work to limit the focus of the discussion at hand.

Lastly, an important quantity for any cosmological survey is the shot noise. It is linked to the discrete nature of the observation (Spinelli et al., 2020), and in this case, arises due to Poisson fluctuations in halo number (Bull et al., 2015). It is crucial to have a good understanding of the shot noise as its amplitude sets the maximum scale at which cosmological information can be extracted, and also provides insights into the galaxies that contain HI (Villaescusa-Navarro et al., 2018). Here, the model for the shot noise power spectrum is chosen as (Bull et al., 2015)

$$P_{\mathrm{HI}}^{\mathrm{shot}}(z) = \left(\frac{\overline{T}_b(z)}{\rho_{\mathrm{HI}}(z)}\right)^2 \int_{M_{\min}}^{M_{\max}} dM \frac{dn}{dM} M_{\mathrm{HI}}^2(z). \tag{1.25}$$

In this case, the HI mass within a halo, M_{HI}, takes the functional form (Castorina & Villaescusa-Navarro, 2017):

$$M_{\mathrm{HI}}(M, z) = C\,(1 - Y_p)\,\frac{\Omega_b}{\Omega_M} \exp\left(-\frac{M_{\min}}{M}\right) M^{\alpha}, \tag{1.26}$$

where $Y_p = 0.24$ is the Helium fraction, α is a free parameter which regulates how rapidly HI is accreted onto haloes, and C denotes the normalisation constant, which is fixed using Equation 1.16 (Castorina & Villaescusa-Navarro, 2017). Further, $M_{\min}$ denotes the halo mass limit, below which the HI abundance in haloes is suppressed exponentially.

1.1.3 Neutral hydrogen intensity mapping

Neutral hydrogen intensity mapping (HI IM) is a novel technique which measures the fluctuations in the HI signal and can thus be used for a diverse range of studies in cosmology (Bharadwaj & Sethi, 2001; Bharadwaj et al., 2001; Battye et al., 2004;

McQuinn et al., 2006; Chang et al., 2008; Wyithe & Loeb, 2009; Bull et al., 2015; Santos et al., 2015). As described in sections 1.1.1 and 1.1.2, in the post-reionisation epoch, most of the neutral hydrogen is contained inside of galaxies (damped Lyα systems). With IM, the 21 cm signal can be probed with low angular resolution surveys (Villaescusa-Navarro et al., 2018; Liu & Shaw, 2020) in which the flux (including from sources that would otherwise be unresolved) is measured over large areas of the sky at various frequencies. This is demonstrated schematically in Figure 1.4, which shows the distribution of galaxies on the sky and the equivalent intensity map. Despite being at a lower resolution, the intensity map contains all the relevant information related to the distribution of HI. This is especially crucial as the HI contained in galaxies traces the underlying matter density field and therefore the large scale structure of the Universe, making it ideal for cosmological studies.

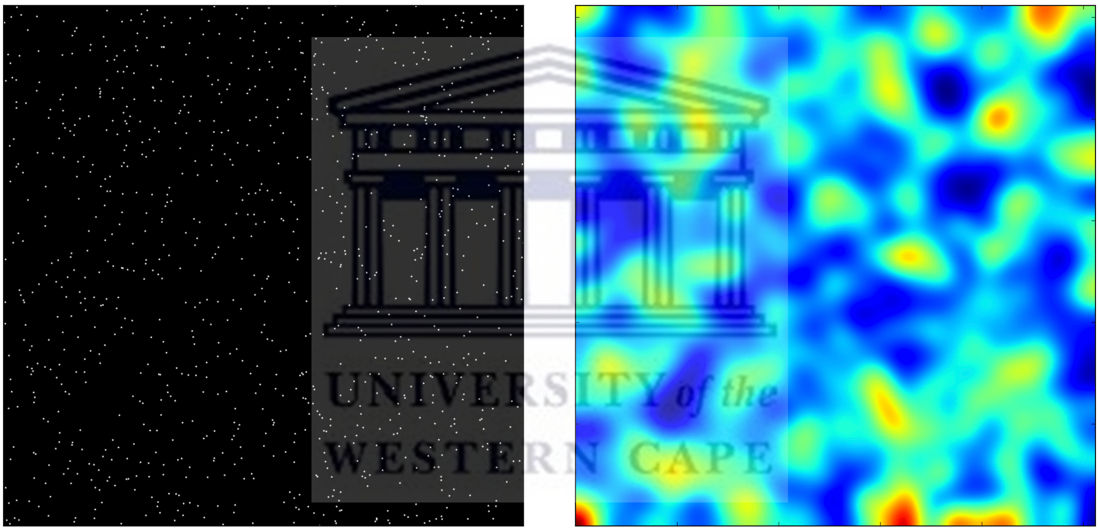

Figure 1.4: A region of space containing many galaxies alongside the intensity map of this galaxy field. With intensity mapping, the 21 cm signal is integrated in large angular pixels of the sky, foregoing the need to resolve the individual structures. This is especially beneficial as it would pick up traces of the signal from sources that are too faint to be resolved individually. Credit: Francisco Villaescusa-Navarro.

There are several advantages to intensity mapping over conventional approaches. Due to the well understood nature of the 21 cm line, HI IM is spectroscopic in nature, as it provides superlative redshift information. It is also more efficient, as it allows larger cosmological volumes to be surveyed as compared to, say, large galaxy surveys such as the Sloan Digital Sky Survey (SDSS) (York et al., 2000). Since the amplitude of the signal only depends on the amplitude and clustering of neutral hydrogen, the HI can be traced over a wide range of redshifts (Villaescusa-Navarro et al., 2018).

Results obtained in cross-correlation (Chang et al., 2010; Masui et al., 2013) and auto-correlation (Switzer et al., 2013) have been able to provide excellent constraints on neutral hydrogen fluctuations (discussed above) and as such have motivated further IM surveys to be undertaken (an example is discussed with the MeerKLASS survey in Santos et al., 2017). While most of the key results obtained so far have been with single-dish IM experiments, the focus here is on a complementary approach with the use of interferometers. Doing this will allow higher angular resolutions to be probed with IM as well as to help mitigate the arduous problems encountered with systematics in measuring the auto-correlation HI power spectrum (such as those encountered by Switzer et al., 2013). In addition, the complementary nature of using interferometers for IM means that an interesting range of important cosmological quantities can be probed alongside the BAO (and therefore enabling constraints on dark energy) which the single-dish method is sensitive to. There are exceptions to this, such as the HIRAX (Newburgh et al., 2016), CHIME (Newburgh et al., 2014) and Tianlai (Xu et al., 2015) interferometers, which are also designed with sensitivity to BAO scales in mind (see Section 1.4 for details on these experiments). Of the many cosmological insights which can be gained through this complementary approach, some are placing constraints on non-linear redshift space distortions (RSD), the spectral index of primordial fluctuations, n_s, as well as the distribution and content of HI in the late Universe.

1.1.4 Cosmological constraints with HI Intensity Mapping

As mentioned, two key probes of cosmology using HI intensity mapping are Baryonic Acoustic Oscillations (BAO) and Redshift Space Distortions (RSD). The BAO arises from the coupling of baryons and photons during the radiation era, in which this photon-baryon plasma undergoes acoustic oscillations. The radiation pressure and gravitation compete and thus sets up these oscillations. Indeed, the baryons oscillate in phase with the radiation due to the coupling of electrons, photons and baryons through Compton scattering. The whole plasma thus oscillates due to these sound waves. These waves are related to the characteristic scale that corresponds to the sound horizon and their imprint can be found on the CMB and the power spectrum of galaxies (Peebles, 1980).

Additionally, the BAO is a powerful probe of the angular diameter distance and Hubble rate as a function of redshift, and measurements of these can be used to constrain dark energy and the curvature of the Universe (Bull et al., 2015; Bacon et al., 2018), which can be achieved through surveys of the large-scale structure of the Universe through intensity mapping (Santos et al., 2015).

Redshift Space Distortions occur due to the motion of galaxies which host HI as well as the motion of this HI gas within the galaxies themselves. This affects the 21 cm intensity mapping signal and thus can be constrained in an intensity mapping survey Sarkar & Bharadwaj (2019). In particular, RSD is useful for measuring the growth rate, which is integral in constraining models of modified gravity (among others) (Bacon et al., 2018) and so is a crucial probe of cosmology using intensity mapping.

1.2 Astrophysical foregrounds

Astrophysical foregrounds are some of the most prominent contaminants hindering a measurement of the 21 cm signal. They consist of all radio emission within an observational band besides the 21 cm signal. As such, these foregrounds need to be mitigated in order to make a measurement of the 21 cm signal. This is due to how much brighter the foregrounds are compared to the signal itself. Experiments aiming to make a detection of the 21 cm signal, expect it to be of the order of ~ 0.1 mK, while the foregrounds themselves measure in the 10s to 100s of Kelvin. Hence, the foreground-to-signal ratio is around 10^5, which presents a formidable challenge (Liu & Shaw, 2020).

1.2.1 Types of astrophysical foregrounds

There are four main types of foregrounds that affect the measurement of the 21 cm signal (Alonso et al., 2015; Cunnington et al., 2019). This is due to them emitting radiation in the same frequency region as the redshifted HI signal, as well as being dominant over the HI which is inherently weak. They are:

(i) Galactic synchrotron emission, which occurs when high-energy electrons are accelerated through a magnetic field. Typically, these electrons are from relativistic cosmic rays which are accelerated by the galactic magnetic field.

(ii) Extragalactic point sources from beyond the Milky Way galaxy, including sources such as active galactic nuclei (AGN), which emit radiation at similar frequencies as the redshifted HI signal.

(iii) Galactic and extragalactic free-free emission (better known as Bremsstrahlung), caused by free electrons which scatter off ions without being captured. This interaction produces photons which have wavelengths which can be similar to that of the redshifted 21 cm line. These interactions occur within and outside the Milky Way galaxy.

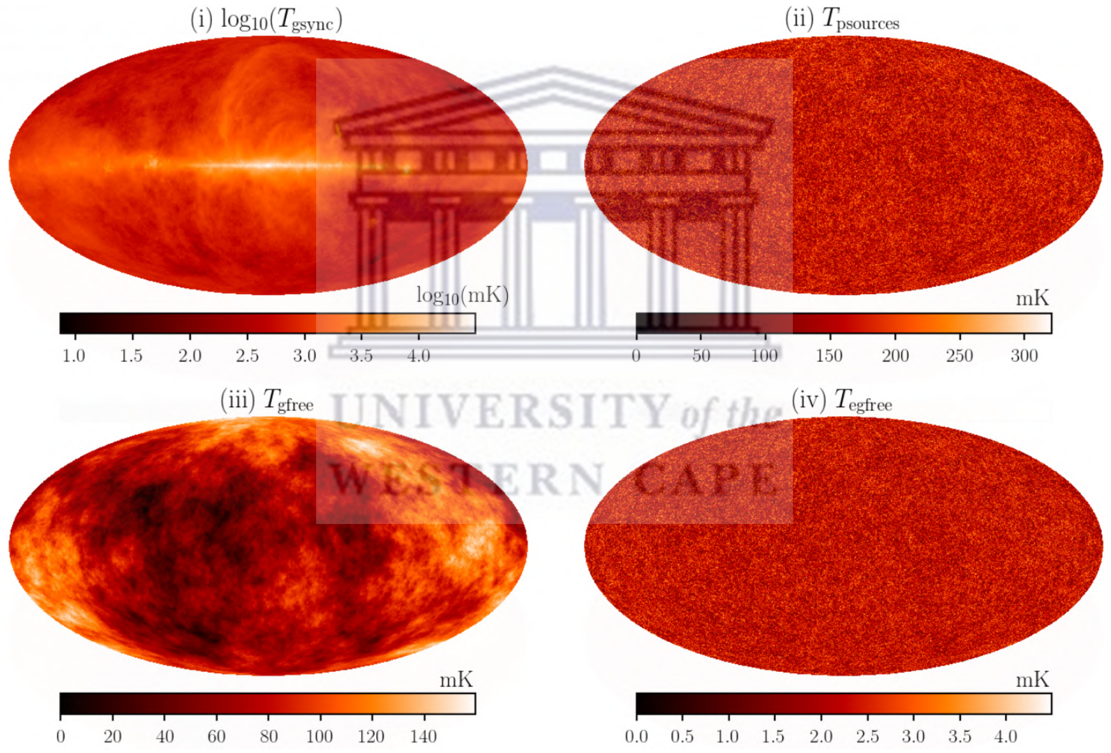

Figure 1.5: Simulated, full sky temperature maps of the four types of foregrounds discussed, at a frequency $\nu = 1136$ MHz, which corresponds to a redshift of $z = 0.25$. Each map has temperature given in mK, with the synchrotron map showing the logarithm of the temperature. Figure taken from Cunnington et al. (2019).

Figure 1.5 shows simulated, full sky temperature maps at $\nu = 1136$ MHz ($z = 0.25$) for each of the foregrounds discussed (Cunnington et al., 2019). It demonstrates that the synchrotron emission dominates over all the other foregrounds. Additionally, the point sources and galactic free-free temperatures are also much brighter than the expected 21 cm signal temperature, albeit fainter than the synchrotron radiation. The extragalactic free-free temperature is relatively low by comparison to the other foregrounds. However, for a precise measurement of the HI signal, it is crucial that all sources of foreground contamination be modelled accurately and removed.

1.2.2 Mitigation strategies

For HI intensity mapping, the observations are not aimed at observing galaxies. In other words, the entire signal at a given frequency is assumed to be HI. Hence, one needs to find a way to remove the foregrounds. The difference in spectral structure between the HI signal and foregrounds provides a means of doing this. The HI signal fluctuates in frequency, as each frequency corresponds to a given redshift, and therefore to different regions along the line-of-sight, thus they decorrelate (Alonso et al., 2015; Chapman et al., 2016). In contrast to this, the foregrounds are spectrally smooth. This has led to the development of numerous foreground cleaning techniques.

Parameterised fits are one such technique, which assume the spectral smoothness implicitly by fitting polynomials to data and then performing a subtraction to remove the foregrounds from said data (Santos et al., 2005; McQuinn et al., 2006). A caveat in this technique is that one needs to assume a specific model that describes the foregrounds, and thus requires a level of precision in the understanding of the foregrounds that is not currently possible. This is due to the fact that there is a lack of data for the foregrounds at the relevant frequencies (Chapman et al., 2016; Cunnington et al., 2019). Additionally, instrumental effects on observed foregrounds, such as polarization leakage, are unlikely to be smooth, which is at odds with one of the primary assumptions for parametric fits. These reasons have seen an increased focus on non-parametric or 'blind' foreground subtraction techniques, which avoid assuming specific foreground models in addition to making fewer assumptions on the form of the foregrounds. These properties of non-parametric techniques thus provide a better means of modelling non-smooth foreground components, offering a better means of removing them from data (Chapman et al., 2016).

There are several blind foreground removal techniques, some of which make use of what is known as mode projection. With mode projection, the data is expressed in some basis. The signal contribution from selected basis components that are foreground-dominated are then removed, which would in principle leave behind only the contribution from the HI signal (Liu & Shaw, 2020). Examples of techniques which make use of this (and which are widely utilised in the literature) include Principal Component Analysis (PCA; Liu et al., 2012), Independent Component Analysis (ICA; Chapman et al., 2012) and Generalized Morphological Component Analysis (GMCA; Chapman et al., 2013, 2016). Refer to Alonso et al. (2015) for a thorough outline of the differences between the PCA and ICA techniques, as well as to Bobin et al. (2008) for an extensive discussion of the GMCA technique.

An alternative to foreground removal is to avoid them. Foregrounds will in principle be spread across all the angular Fourier modes, but confined to the lowest line-of-sight modes, since the line-of-sight direction corresponds to the frequency (Morales & Hewitt, 2004; Morales, 2005). Since this will make the foregrounds compact in Fourier space, foreground avoidance simply implies ignoring those regions in Fourier space which are dominated by foregrounds. This technique has been used in many intensity mapping studies focused on the EoR and Cosmic Dawn (for example, Parsons et al., 2012a,b; Thyagarajan et al., 2013) and was specifically chosen as the means of overcoming the effects of foregrounds in this study. The method of foreground avoidance and how it is employed is discussed in broader detail in Section 2.2.

1.3 Radio Interferometry

1.3.1 Background

For almost the entire period of human civilisation, the only means of making observations of the Universe was restricted to measurements of visible light, i.e. that section of the electromagnetic spectrum which the human eye is sensitive to. With the discovery of extraterrestrial radiation emitted by an object that was not the sun by Jansky in 1931, the landscape of observational astronomy changed dramatically. Further observations were made over the course of the 20th century which eventually led to the development of radio telescopes dedicated to astronomical observations (Wilson et al., 2012).

The development of radio interferometry allowed finer angular resolutions to be resolved and thus allowed astronomers to not only study the Universe in the radio part of the spectrum, but also probe finer details. Additionally, radio interferometry allowed cross-matching work to be done between the optical and radio domains (Thompson et al., 2017).

With the advent of the 21st century, the development of radio interferometers which would allow measurements of cosmic hydrogen over large ranges of redshift, spanning from the Dark Ages, reionisation epoch and beyond, have taken centre stage in cosmology (Morales & Hewitt, 2004; Morales, 2005). The development of the intensity mapping technique for cosmological studies will allow researchers to take full advantage of the instruments at their disposal (see Section 1.4). In this study, the use of radio interferometry for intensity mapping of the 21 cm line is the key focus. Section 1.3.2 discusses the fundamentals of radio interferometry, which forms the basis of the analysis techniques used to develop the simulations and extract cosmological information as described in Section 2.

1.3.2 Fundamentals

At the fundamental level, the radio interferometer is a collection of dishes or dipoles. While a single dish or dipole element would generally measure the sky and map it out a single pixel at a time, the interferometer spreads out the collecting area into multiple receivers. In this way, electric field signals are received by pairs of antennas and then multiplied together and averaged over a short time interval. In the simplest case, this amounts to the two-element interferometer, shown in Figure 1.6. For this setup, assume that the interferometer consists of two antennas, A_1 and A_2, separated by the distance $\vec{b}$, known as the baseline, with its direction from A_2 to A_1. This baseline quantity can be defined as

$$b = \frac{\lambda}{\theta}, \tag{1.27}$$

where λ denotes the wavelength at which the interferometer is making an observation and θ its angular resolution. Further, assume that both of these antennas are only sensitive to radiation of the same polarisation state. Following the derivation outlined in Wilson et al. (2012), let V_1 be a voltage induced at the output of antenna

A_1 through an electromagnetic wave of amplitude E from a distant source,

$$V_1 \propto Ee^{j\omega t}, \tag{1.28}$$

and at A_2, we have

$$V_2 \propto Ee^{j\omega(t-\tau_g)}, \tag{1.29}$$

where τ_g is known as the geometrical delay which is caused by the orientation of the baseline $\vec{b}$ relative to the direction in which the electromagnetic wave is propagating, and the imaginary number defined as $j \equiv \sqrt{-1}$. The outputs from the two antennas will be correlated, due to the electromagnetic signals being inputted to a multiplying device, followed by an integrator, resulting in an output

$$R(\tau_g) \propto \frac{E^2}{T}\int_0^T e^{j\omega t}e^{-j\omega(t-\tau_g)}dt. \tag{1.30}$$

If $T \gg \frac{2\pi}{\omega}$, then the average over a time T will differ little from a the average over a full period, resulting in

$$\begin{aligned} R(\tau_g) &\propto \frac{\omega}{2\pi}E^2\int_o^{2\pi/\omega} e^{j\omega\tau_g}dt \\ &\propto \frac{\omega}{2\pi}E^2e^{j\omega\tau_g}\int_o^{2\pi/\omega} dt, \end{aligned} \tag{1.31}$$

which yields

$$R(\tau_g) \propto \frac{1}{2}E^2e^{j\omega\tau_g}. \tag{1.32}$$

Now, the output of the correlator followed by the integrator varies periodically with

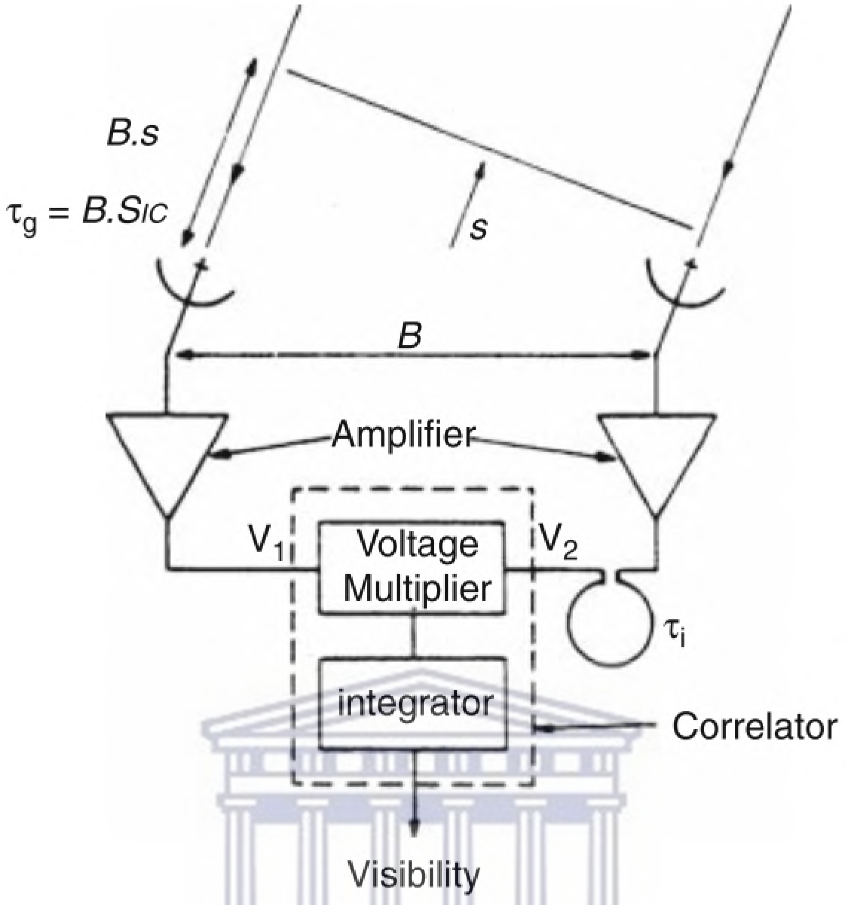

Figure 1.6: Schematic showing the two-element interferometer, with all the relevant components such as the amplifier, multiplier and integrator. Further, τ_g and τ_i denote the geometrical and instrumental delays. Additionally, while the figure uses the symbol "B" to denote the baseline, this has been replaced with "b" in the text. Schematic taken from Wilson et al. (2012).

τ_g, the geometrical delay, given as (Figure 1.6),

$$\tau_g = \frac{\vec{b} \cdot \vec{s}}{c}, \tag{1.33}$$

where $\vec{s}$ denotes the directional vector of the electromagnetic radiation from the source. If the relative orientation of $\vec{b}$ and $\vec{s}$ remain unchanged, then τ_g and $R(\tau_g)$ remain constant. In reality, $\vec{s}$ varies slowly due to the earth's rotation, causing variation in τ_g and therefore resulting in the measurement of interference fringes as a function of time.

If we have some radio brightness distribution, $I_\nu(\vec{s})$, the power per bandwidth, $d\nu$, and source element, $d\Omega$, will be

$$W^{\mathrm{P}}(\vec{s})I_{\nu}(\vec{s})d\Omega d\nu, \tag{1.34}$$

where $W^{\mathrm{P}}(\vec{s})$ denotes the effective collecting area in the direction $\vec{s}$. The correlator output is then

$$r_{12} = W^{\mathrm{P}}(\vec{s})I_{\nu}(\vec{s})e^{j\omega\tau}d\Omega d\nu, \tag{1.35}$$

where τ denotes the difference between the geometrical and instrumental delays (shown in Figure 1.6)

$$\tau = \tau_g - \tau_i = \frac{\vec{b}\cdot\vec{s}}{c} - \tau_i. \tag{1.36}$$

The total response is then given in terms of the baseline vector $\vec{b}$

$$R(\vec{b}) = \iint_{\Omega} W^{\mathrm{P}}(\vec{s})I_{\nu}(\vec{s})\exp\left[j2\pi\nu\left(\frac{\vec{b}\cdot\vec{s}}{c} - \tau_i\right)\right] d\Omega d\nu. \tag{1.37}$$

Equation 1.37 is known as the Visibility Equation (Wilson et al., 2012). To derive the complex visibility, a more appropriate coordinate system must be chosen for $\vec{b}$ and $\vec{s}$. This can be done by choosing a unit vector $\vec{s}$ pointing towards the center of the direction in which the antennas are pointed (as shown in Figure 1.6)

$$\vec{s} = \vec{s}_0 + \vec{\sigma}, \quad |\vec{\sigma}| = 1, \tag{1.38}$$

where $\vec{s}_0$ is a position chosen close to the center of the observed region. Substituting into Equation 1.37 yields

$$R(\vec{b}) = \exp\left[j\omega\left(\frac{\vec{b}\cdot\vec{s}_0}{c} - \tau_i\right)\right] d\nu \iint_{S} W^{\mathrm{P}}(\vec{\sigma})I(\vec{\sigma})\exp\left[j\frac{\omega}{c}\left(\vec{b}\cdot\vec{\sigma}\right)\right] d\vec{\sigma}. \tag{1.39}$$

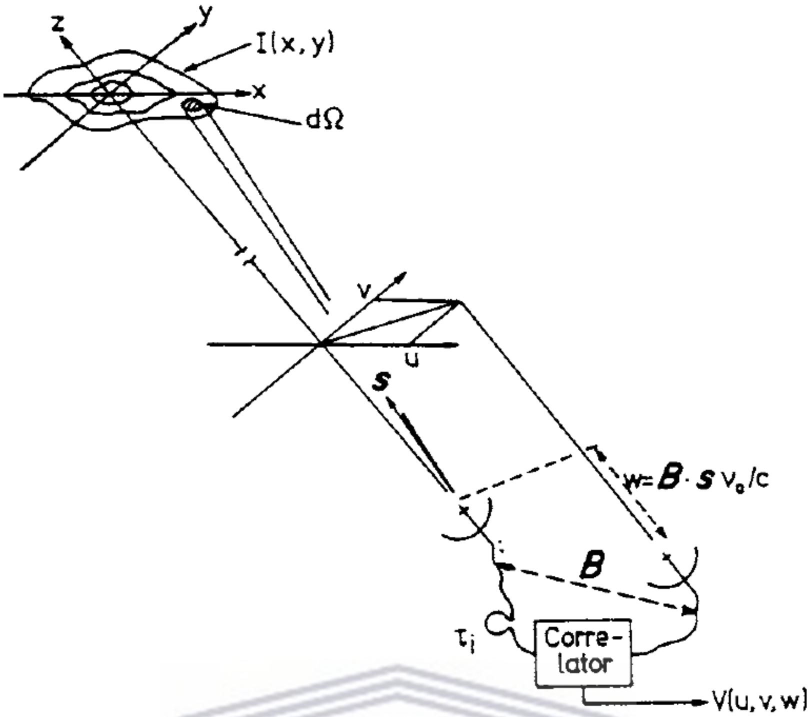

Figure 1.7: Schematic showing the end-to-end observation processes of an interferometer, emphasising the various coordinate systems relevant for the geometrical formulation of the interferometer equation and aperture synthesis. As mentioned in Figure 1.6, the baseline here denoted by "B" has been replaced with "b" in the text. Schematic taken from Wilson et al. (2012).

The exponential factor extracted from the integral in Equation 1.39 defines the phase of $R(\vec{b})$, while the integral over the intensity distribution gives what is known as the visibility

$$V(\vec{b}) = \iint\limits_{S} W^{\mathrm{P}}(\vec{\sigma}) I(\vec{\sigma}) \exp\left[j\frac{\omega}{c}\left(\vec{b}\cdot\vec{\sigma}\right)\right] d\vec{\sigma}. \tag{1.40}$$

If one chooses coordinates of the form

$$\frac{\omega}{2\pi c}\vec{b} = \{u, v, w\}, \quad \frac{\omega \pm \Delta\omega}{2\pi c} = \frac{\nu}{c}\left(1 \pm \frac{\Delta\nu}{\nu}\right), \tag{1.41}$$

where $\{u, v, w\}$ are measured in units of wavelength, $\lambda = \frac{2\pi c}{\omega}$, and further chooses the vector $\vec{\sigma} = \{x, y, z\}$, such that x and y are direction cosines with respect to the

u and v axes, then the visibility becomes

$$\begin{aligned} V(u,v,w) = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W^{\mathrm{P}}(x,y) I(x,y) \\ &\times \exp\left[j2\pi\left(ux+vy+w\sqrt{1-x^2-y^2}\right)\right]\frac{dxdy}{\sqrt{1-x^2-y^2}}. \end{aligned} \tag{1.42}$$

From Figure 1.7, it should be noted that the xy tangent plane is a projection of the celestial sphere with the tangent point at position $\vec{s}_0$. Moreover, the u axis is directed towards local east, while the v axis has its direction toward local north. The limits of integration in Equation 1.42 are such that they demand that $W^{\mathrm{P}}(x,y) = 0$ for $x^2+y^2 > l^2$, where l denotes the full width of the telescope primary beams. If only a small patch of the sky is observed, then $\sqrt{1-x^2-y^2} \cong 1$, and Equation 1.42 becomes

$$\begin{aligned} V(u,v,w=0) \cong V(u,v,w)e^{-j2\pi w} = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W^{\mathrm{P}}(x,y) I(x,y) \\ &\times e^{j2\pi(ux+vy)}dxdy. \end{aligned} \tag{1.43}$$

The exponential, $e^{-j2\pi w}$ is the conversion factor that would approximately change the observed phase of the visibility to that which would be measured in the uv plane. In reality, an interferometer would make observations of the sky over a range of frequencies that form its bandwidth. Hence, Equation 1.43 can be expressed with its frequency dependence explicitly given as

$$V(u,v,\nu) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W^{\mathrm{P}}(x,y,\nu) I(x,y,\nu) e^{j2\pi(ux+vy)}dxdy. \tag{1.44}$$

To further simplify the expression of the visibility, one can make the notational change $\vec{\theta} = \{\theta_x, \theta_y\} = \{x, y\}$ such that $d^2\theta = dxdy$. A baseline is generated by the distance between pairs of antennas and thus forms a vector, $\vec{u}_\nu$, defined as

$$\vec{u}_\nu = \{u, v\} = \frac{\vec{b}}{\lambda}. \tag{1.45}$$

The brightness temperature of the sky can be defined in terms of the flat-sky coordinates $\vec{\theta} = \{\theta_x, \theta_y\}$ introduced above. The Fourier pair for the brightness temperature can then be expressed as (Liu & Shaw, 2020)

$$\tilde{T}(u, v, \nu) \equiv \int_{-\infty}^{\infty} T(\theta_x, \theta_y, \nu) e^{-j2\pi\vec{u}_\nu \cdot \vec{\theta}} d^2\theta \tag{1.46}$$

and

$$T(\theta_x, \theta_y, \nu) \equiv \int_{-\infty}^{\infty} \tilde{T}(u, v, \nu) e^{j2\pi\vec{u}_\nu \cdot \vec{\theta}} d^2u_\nu, \tag{1.47}$$

where $T(\theta_x, \theta_y, \nu)$ is the brightness temperature of the sky and $\tilde{T}(u, v, \nu)$ its Fourier conjugate. Using the Rayleigh-Jeans relation, one can replace the intensity distribution in Equation 1.44 with the brightness temperature. Doing so, as well as using the introduced coordinates, the visibility, V, can be written as (Liu & Shaw, 2020)

$$V(u, v, \nu) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T(\theta_x, \theta_y, \nu) W^{\mathrm{P}}(\theta_x, \theta_y, \nu) e^{-j2\pi\vec{u}_\nu \cdot \vec{\theta}} d^2\theta, \tag{1.48}$$

where $W^{\mathrm{P}}(\theta_x, \theta_y, \nu)$ denotes the primary beam which accounts for the fact that the antennas making up the interferometer do not have equal sensitivity at all areas of the sky (Liu & Shaw, 2020). The visibility essentially forms the fundamental observable of an interferometer. The geometrical aspects involved in Equation 1.48 are shown in Figure 1.7, which together with Figure 1.6 shows the setup used in the derivation of the visibility. Here, a region of the sky plane is observed via its brightness temperature by the two antennas making up a baseline. The signals captured are then correlated which generate the visibility.

In an interferometric observation, each baseline probes a particular Fourier mode. With information from each baseline in an array, multiple Fourier modes are probed. These modes can then be Fourier transformed to obtain an image of the observed sky. This is the basic principle behind synthesis imaging with an interferometer. Appendix A gives a brief discussion on the basic properties of the Fourier Transform as employed in the above derivation as well as in later sections.

Another thing to note is that with the more unique modes available, the closer the image will be to the true sky, with the two being related by what is known as the synthesised beam. This beam is formed via the Fourier transform of the uv plane, which is a map of all the uv points generated during the interferometric observation. Another phenomena which plays a role in the generation of uv points is the Earth's rotation. As the Earth rotates, the baseline vectors rotate relative to this and therefore rotate through the uv plane. This is known as rotation synthesis as is shown in Figure 1.8.

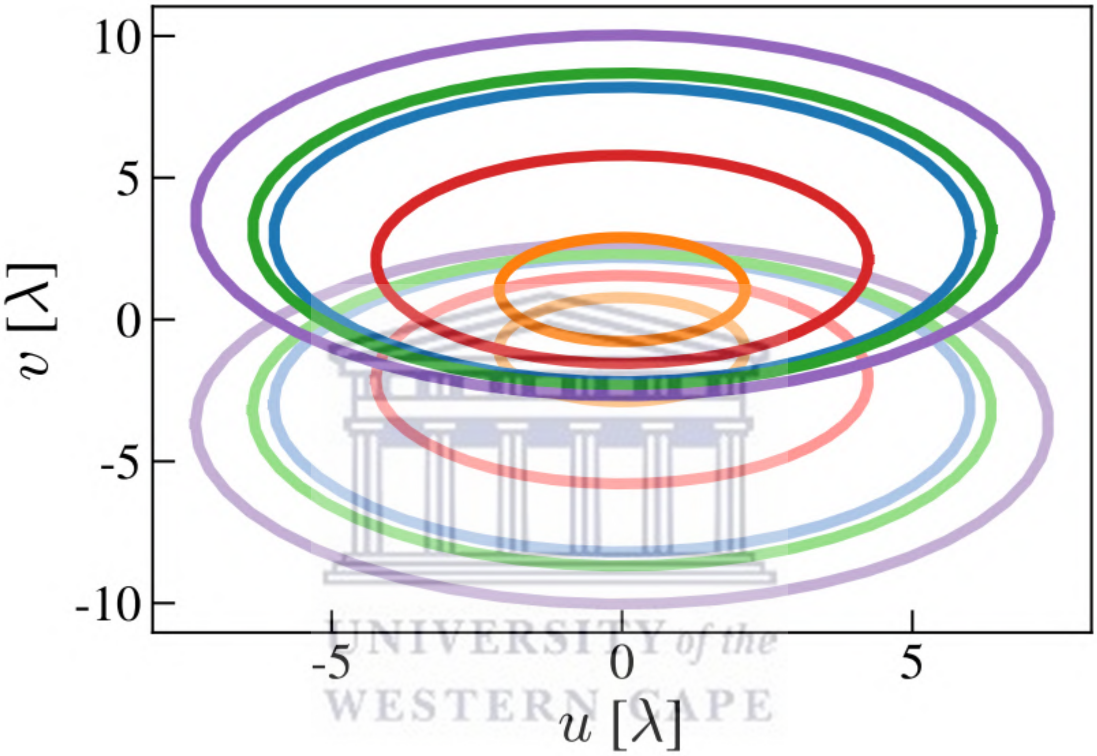

Figure 1.8: A schematic showcasing the principle of rotation synthesis. In particular, it shows the movement of baselines through the uv plane over the course of a day of observation with an array of 6 antennas. Here each sampled (u, v) point is accompanied by its corresponding $(-u, -v)$ point. Schematic taken from Liu & Shaw (2020)

In general, there are two distinct classes of interferometric observation modes. These are known as the tracking and drift-scan modes. With tracking, individual antennas are steered and thus track particular fields in the sky, while the drift-scan mode fixes the antennas, resulting in the observation of those regions of the sky which are observable from the Earth as it rotates with the telescope. While some radio telescopes have been build specifically for drift-scan observations, it is generally true that most can observe in either mode.

Further, with radio telescopes, there are two observation modes. These are the single-dish and interferometric modes and instruments are usually purposely built for one or the other. Naturally, the interferometric instruments are considered more ideal when considering angular resolution, but are hindered by gaps in the synthesised beams generated by observations which is a problem not faced by single-dish observations. Generally, the mode of observation depends on the type of science being done. An advantage, then, is that for cosmological studies, particularly HI IM, both modes are suitable, as has already been shown by the vast amount of effort put into the study of the plausibility of using interferometers for this purpose (for example Morales & Hewitt, 2004; Morales, 2005; Parsons et al., 2010, 2012a, as well as many others) as well as results from single-dish efforts (as noted above in studies such as Chang et al., 2010; Masui et al., 2013). For comprehensive reviews of the principles of interferometry, refer to books such as Wilson et al. (2012) and Thompson et al. (2017), as well as papers such as Liu & Shaw (2020), which deals with interferometry in the context of the data analysis techniques employed for 21 cm cosmology.

1.4 21 cm experiments

In the design and construction of experiments aimed at 21 cm cosmology there are a number of general considerations. These include high sensitivity due to the faintness of the 21 cm signal across all redshifts of relevance. Additionally, these experiments have to be sensitive to a broad range of scales relevant to doing cosmology with the 21 cm line. An important example would be the BAO scales. Stability is another aspect of key importance as the influence of contaminants and systematics can easily hinder the science conducted on data sets collected with these experiments. Lastly, these experiments need to be broadband as a goal of the science done with them is to map the Universe in three dimensions. Thus, efficiency over a broad range of frequencies is crucial.

Recently, a number of instruments have been planned for the purpose of 21 cm cosmology. In particular, intensity mapping measurements in the post-reionisation epoch have already been made with a number of single-dish experiments (shown in Figure 1.9) that were not necessarily designed specifically for 21 cm cosmology. These include:

- The Green Bank Telescope (GBT), which is a 100 m single-dish telescope located in West Virginia with an observational frequency range from 110 MHz to 115 GHz (Prestage et al., 2009).
- The Parkes Radio Telescope, which is a 64 m single-dish telescope located in Australia and which is able to observe over the frequency range 1230 MHz $\leqslant \nu \leqslant$ 1530 MHz (Staveley-Smith et al., 1996).

Figure 1.9: The Green Bank Telescope (left) and the Parkes Radio Telescope (right). Photo credits are linked in each experiments name.

Both of these single-dish experiments have already made measurements relevant to 21 cm cosmology in cross-correlation using the intensity mapping technique, despite both being general purpose radio telescopes. Experiments that were specifically designed for the purpose of intensity mapping in the post-reionisation Universe, and that are yet to make measurements or which are still relatively new or being built include:

- The Canadian Hydrogen Intensity Mapping Experiment (CHIME), located in British Columbia, Canada, is an instrument comprised of four large cylindrical reflectors (the full experiment is planned to have five), each with dimensions 20 m × 100 m. It operates as a drift-scan telescope over a frequency ranging spanning 400 to 800 MHz (Newburgh et al., 2014). It has been designed with the goal of BAO detection over the redshift range, $z = 0.8 - 2.5$, but can also be used for other scientific endeavours such as the study of pulsars and Fast Radio Bursts (FRBs).
- The Hydrogen Intensity and Real-time Analysis eXperiment (HIRAX) is located in South Africa and shares the same scientific goals as CHIME. At

completion, HIRAX will comprise of a square 32 × 32 grid of 1024 dishes operating between 400 and 800 MHz (Newburgh et al., 2016). HIRAX differs from CHIME in that its dishes can be manually repointed.

- The Tianlai experiment is located in the Xinjiang Autonomous Region of China. It is an 21 cm intensity mapping experiment with a frequency range of 400 to 1420 MHz, with similar science goals as CHIME and HIRAX (Xu et al., 2015). Its final design is yet to be decided on, but there are already two pathfinder experiments, with one being a set of three cylindrical reflectors and the other a set of 16 dishes.
- The BAO from Integrated Neutral Gas Observations (BINGO) experiment is located in South America and will aim to make detections over a frequency range spanning 960 to 1260 MHz (Battye et al., 2013). It complements CHIME, HIRAX and Tianlai due to this frequency range (which sets it at a different redshift range) as well as the fact that it will be a (proposed 40 m) single-dish experiment operating in drift-scan mode.

While the list discussed above is not exhaustive, it gives a general picture of the planned experiments focused on 21 cm cosmology science goals. All four instruments are shown in Figure 1.10.

1.4.1 The MeerKAT interferometer

The MeerKAT radio telescope is a 64-dish precursor to the Square Kilometer Array mid-frequency telescope (SKA1-MID). It is located in the Karoo region of South Africa (shown in Figure 1.11). Currently, it is the most sensitive decimetre-wavelength radio interferometer in the world (Jonas, 2016). As a precursor, MeerKAT will be incorporated into the full array towards the end of the construction of SKA1-MID. MeerKAT is able to operate in both single-dish and interferometer mode and operates over two frequency bands, namely, the L-band which spans a frequency range of $900\text{MHz} < \nu < 1420\text{MHz}$, corresponding to a redshift range $0 < z < 0.58$, as well as the UHF-band with ranges $580\text{MHz} < \nu < 1000\text{MHz}$ and $0.4 < z < 1.45$.

There are a number of large survey projects (LSPs) underway with MeerKAT. These include LADUMA (Baker et al., 2018), MHONGOOSE (de Blok et al., 2016), ThunderKAT (Fender et al., 2016), MALS (Gupta et al., 2016), the MeerKAT Fornax

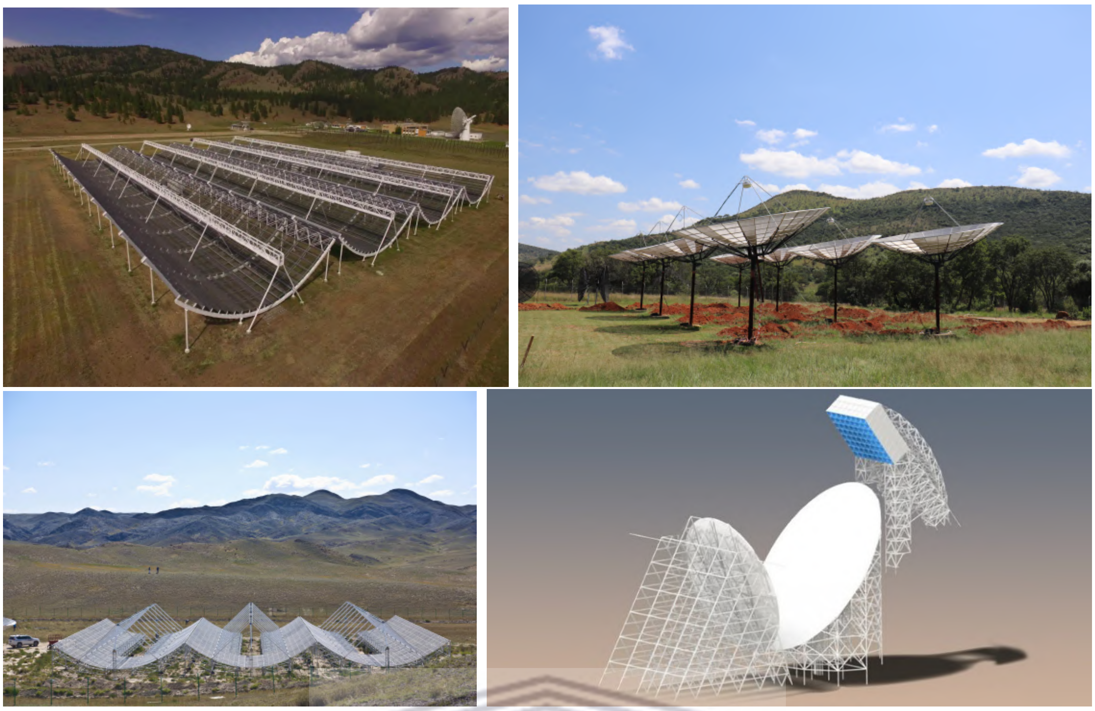

Figure 1.10: The CHIME (top-left), HIRAX (top-right), Tianlai (bottom-left) and BINGO (bottom-right) instruments. While CHIME and HIRAX already have sites and set instrument types, Tianlai is still being tested via pathfinders, while BINGO is in the concepting phase of its construction. Photo credits are linked in each experiments name.

Survey (Serra et al., 2016), MeerTime (Bailes et al., 2016), TRAPUM, MeerTRAP (Sanidas et al., 2018), and specifically important for this study, MIGHTEE (Jarvis et al., 2016). Whereas these surveys all have various scientific goals related to the distribution, dynamics, and evolution of HI in the Universe (MHONGOOSE, LADUMA, MALS and the MeerKAT Fornax Survey), transient and pulsar science (ThunderKAT, MeerTime, TRAPUM and MeerTRAP), and many other astrophysical and extra-galactic prospects, the possibility of utilising the MeerKAT interferometer for 21 cm intensity mapping has been considered as well (as discussed in Santos et al., 2017; Pourtsidou, 2016).

In particular, Santos et al. (2017) notes that with an single-dish mode HI intensity mapping survey in one of the two frequency bands available over a period of five months of observation and an observed area of around 4000 deg^2 precise measurements of the HI can be made through the cross-correlation of MeerKAT data and optical galaxy surveys, as well as with the auto-correlation approach. Further, they note the complementary and synergistic nature of using these two approaches. Lastly, since such a survey will also generate interferometric data, it will allow the

Figure 1.11: The MeerKAT interferometer, located at the Karoo site of the eventual Square Kilometer Array (SKA) telescope, for which it is a precursor experiment. Image credit: SARAO.

consideration of various other science cases such as galaxy evolution and polarisation.

This wide area survey with the MeerKAT telescope known as the MeerKAT Large Area Synoptic Survey (MeerKLASS) (Santos et al., 2017) would cover the observation time and area mentioned above and provide a potential detection of the BAO feature, besides other interesting science goals it could well achieve. Due to this, it would play a crucial role in laying the groundwork towards cosmology with the eventual SKA1-MID.

1.4.2 The MIGHTEE survey

While MeerKLASS has been proposed for cosmology, the MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) survey (Jarvis et al., 2016), one of the large surveys currently being undertaken by MeerKAT could potentially provide a means of doing pioneering studies of, and providing a complementary approach with, interferometric data for the purpose of HI intensity mapping.

In practice, it is an extra-galactic continuum survey over four of the most well studied regions in the southern hemisphere with the aim of investigating the evolution of

AGN, star-forming galaxies and galaxy clusters over a wide range of redshifts and observational fields. These fields are the COSMOS, XMM-LSS, ECDFS and ELAIS-S1 regions. The observational strategy with its various pointings are shown in Figure 1.12, which showcases the abundance of observational fields as well as the large area that will be observed in the full survey (about 20 deg^2).

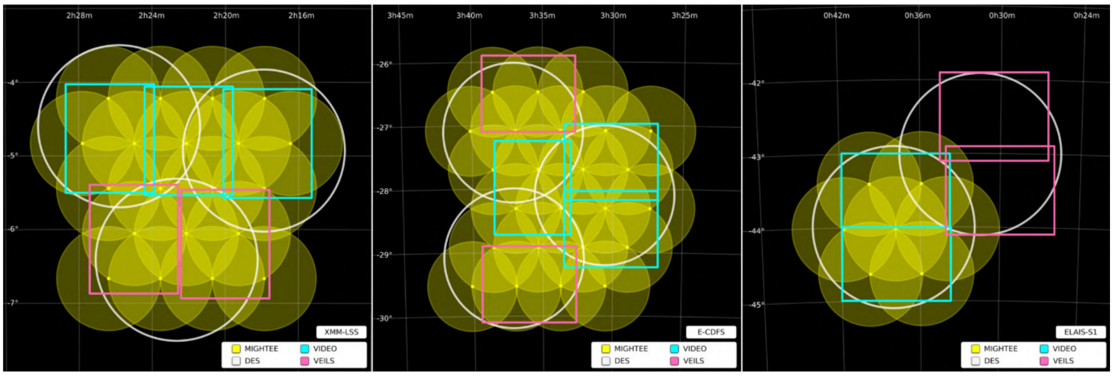

Figure 1.12: The pointing strategies of the MIGHTEE survey for three of the four regions it will observe through its full duration. In particular, it shows XMM-LSS (left), ECDFS (middle), as well as ELAIS-S1 (right), which together will cover approximately 17 deg^2. Not shown is the COSMOS field, which will contribute to bringing the full observational area to around 20 deg^2. Figure taken from Jarvis et al. (2016).

Despite the MIGHTEE survey's primary goals, the data available could provide a means of testing 21 cm intensity mapping in interferometer mode and is used as a motivation in the development of the simulation pipeline designed to mimic the MIGHTEE observations. The specifications of the survey are further used to test the plausibility of measuring the power spectrum from the interferometric data sets that MIGHTEE will produce.

1.5 Thesis outline

In this study, a purpose-built, visibility-based pipeline that emulates an interferometric observation with the MeerKAT instrument and MIGHTEE survey is utilised to study effects such as thermal noise and foreground contamination and the impact they have on calculating the power spectrum from simulated observations of the HI signal. Various cases are considered to not only estimate the HI power spectrum, but also test the simulation code in general.

The remainder of the thesis is structured as follows: Section 2 details the theoretical aspects of the delay spectrum method used to estimate the HI power spectrum and describes the simulation codes developed for this purpose. Section 3 discusses the simulation outputs and power spectrum estimates. Finally, the conclusions and outlook to future research prospects are presented in Section 4. Throughout this thesis, cosmological parameters from the Planck 2018 results are used for a flat, ΛCDM universe (Planck Collaboration et al., 2018). In particular, $\{\Omega_M, \Omega_b, h, n_s, \sigma_8\} = \{0.311, 0.049, 0.677, 0.967, 0.8102\}$.

2 HI power spectrum estimates: Simulations

2.1 Motivation

In this study, a purpose-built simulation pipeline was utilised to investigate the use of MIGHTEE data for the purpose of HI intensity mapping and constraining the HI power spectrum. In particular, the *uv* distributions, survey parameters and sky models produced from MIGHTEE COSMOS observations were used as input to help understand various effects that would hinder a detection of the HI power spectrum using HI intensity mapping in interferometry mode (for more details see: Paul et al., 2020, on which the author of this thesis is a co-author). While intensity mapping has been extensively studied using interferometers (for example: Morales & Hewitt, 2004; Morales, 2005; Parsons et al., 2012a,b; Thyagarajan et al., 2013), these studies were restricted to low frequency and high redshift with the purpose of studying the spatial information of the Epoch of Reionization. In this study, a similar analysis is performed, but with the focus being on the late Universe i.e., high frequency and low redshifts. An example of such a study is Bull et al. (2015), which focuses on developing a framework for forecasting cosmological constraints for HI intensity mapping experiments at low and intermediate redshifts. For thorough reviews of the use of intensity mapping with interferometers for post-EoR science, refer to Kovetz et al. (2017) and Liu & Shaw (2020).

Two sets of simulations are discussed here, with the first focused on studying the foregrounds individually (Section 2.3) and the second being an end-to-end purpose-built observation and power spectrum calculation pipeline developed with the goal of testing the effectiveness of applying a delay spectrum framework to high frequency interferometric data products (Section 2.4), such as those produced by the MIGHTEE survey. To narrow things down, both simulations study the COSMOS field, which is well studied in the literature (as with the VLA COSMOS project, including the 1.4 GHz and 3 GHz Large projects with Schinnerer et al., 2004, 2007; Smolcic et al., 2017) and acts as a first test field for running the simulations. This was done by using both a *uv* distribution as well as sky model from an actual MIGHTEE COSMOS dataset.

The compact core of MeerKAT allows for the probing of non-linear scales in interferometry mode. The techniques employed here are complementary to single dish intensity mapping which focuses on much larger scales. These particular scales are useful for studying, for example, Baryonic Acoustic Oscillations (BAO) (see Eisenstein et al., 2005, for the first detection of this phenomenon). These two complementary regions of interest for each type of experimental setup is shown in Figure 2.1, where different scales that can be probed by either the single-dish or interferometer setup are shown on the HI power spectrum. In particular, the single-dish range is bounded by the survey area on the left and the primary beam of the telescope on the right, while the interferometer range is bounded by the minimum and maximum baselines.

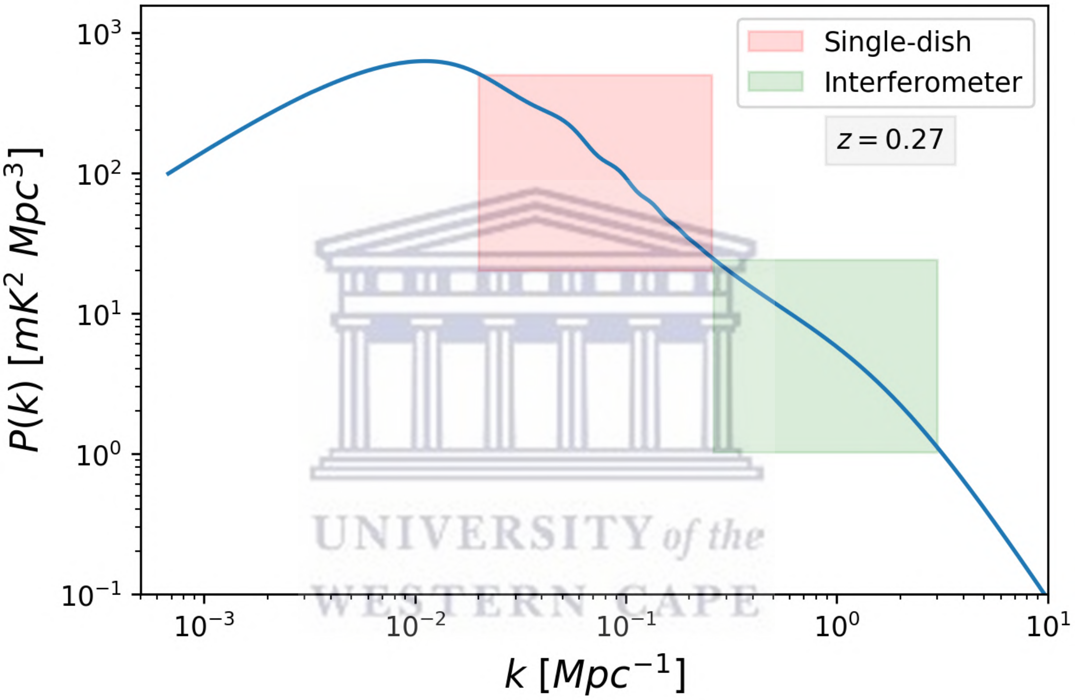

Figure 2.1: The scales of interest on the HI power spectrum for single-dish and interferometer setups at $z = 0.27$. In particular, the figure highlights the sensitivity of single-dish experiments to larger scales, which are relevant to studies of the BAO and RSD, while also showing the smaller scales which the interferometric setup could potentially probe.

A 2D analogue to Figure 2.1 is shown in Figure 2.2. It also shows the regions of Fourier space which the single-dish and interferometric experiments are sensitive to. Specifically, it highlights some important boundaries along the line-of-sight such as bandwidth- ($k_{\rm BW}$), foreground- ($k_{\rm FG}$) and frequency channel- ($k_{\rm channel}$) and non-linearity-limited ($k_{\rm NL}$) Fourier modes, while also highlighting key angular boundaries such as those related to area ($k_{\rm area}$), the field-of-view ($k_{\rm FOV}$) and the longest baseline ($k_{\rm D_{max}}$).

While being sensitive to the same line-of-sight ($k_\parallel$) scales, the two are complementary in terms of angular sensitivity ($k_\perp$). It is for this reason that interferometric experiments have the capability of probing smaller (non-linear) scales compared to their single dish counterparts. This distinction is due to interferometers being able to probe larger $k_\perp$ through their higher angular resolution.

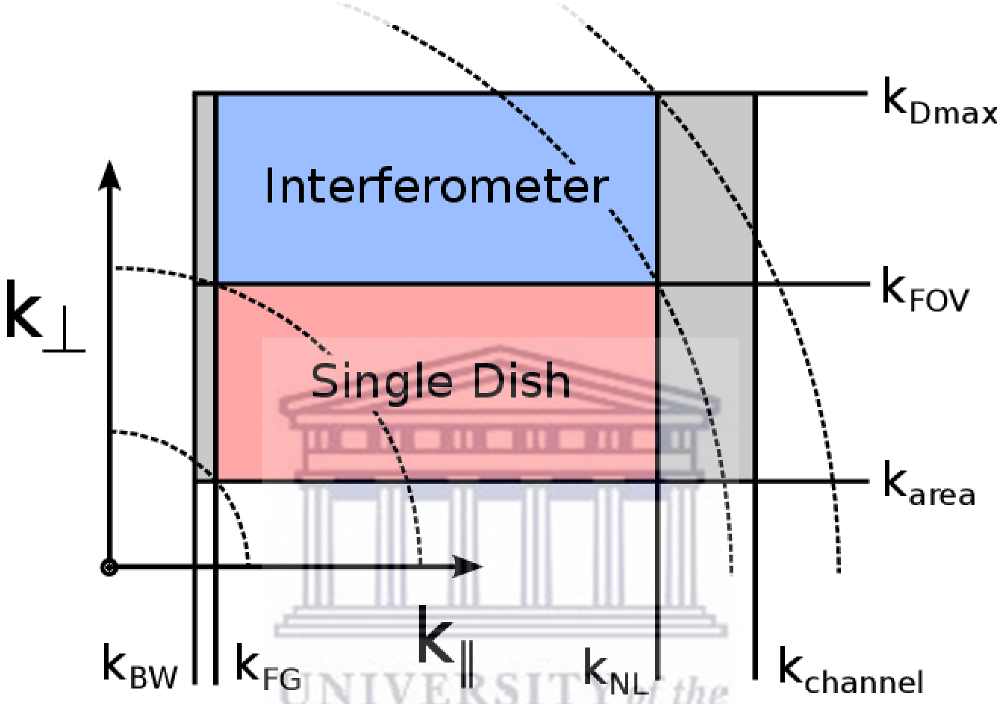

Figure 2.2: A schematic highlighting the regions in k space which are sensitive to both single dish and interferometric experiments. As seen in the schematic, the two types of experiment are subject to the same spectral constraints (along $k_\parallel$), while being complementary in terms of angular sensitivity ($k_\perp$). This schematic was taken from Bull et al. (2015).

Overall, the goal of the simulations are to emulate MIGHTEE observations as well as test effects that might arise in the data analysis and compare this to the actual MIGHTEE data. In this study, however, the effects were restricted to thermal noise contamination, foregrounds, as well as instrumental effects that might arise in the region of Fourier space, which is assumed to be relatively clean of contamination and thus enabling a detection of the HI power spectrum, known as the HI window.

2.2 Delay spectrum methodology

The principle of the delay spectrum approximation is to use an estimator for the HI power spectrum that is very close to the raw data without the need to go to imaging space. This is shown schematically in Figure 2.3, where the visibilities can either be Fourier transformed along uv to obtain the image cube or alternatively Fourier transformed along the frequency axis to obtain the Fourier representation, or the representation of the raw data in delay space. Further, Figure 2.4 shows the structure of both the HI signal and foregrounds in delay space. In this Fourier representation, the HI signal is spherically symmetric, while the foregrounds show strong separable-axial symmetry. This symmetry difference allows one to use foreground avoidance, since the foregrounds are localised well in delay space. While one could potentially utilising foreground cleaning techniques in the delay space directly, the choice of foreground avoidance over cleaning avoids possible signal loss.

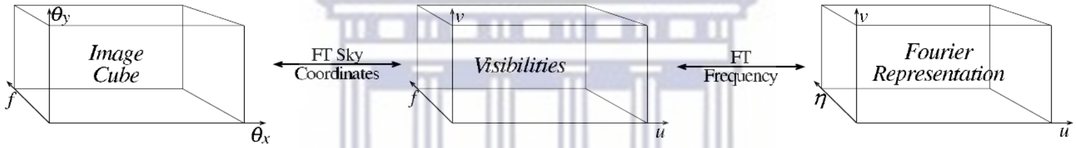

Figure 2.3: Schematic showing the relationship between the image cube, measured visibilities and the Fourier representation. The measured visibilities or visibility-frequency cube is the fundamental observable of an interferometer and can either be Fourier transformed along the spatial coordinates to yield an image cube or Fourier transformed along frequency to show the spatial structure in the visibilities in a full Fourier representation. The diagram was taken from Morales & Hewitt (2004).

Figure 2.5 shows the practical distinction between the two methods which can be employed to estimate the HI power spectrum from interferometric data, namely the *delay-style* (measured) and *imaging-style* (reconstructed) methods. Particularly emphasised are the ways in which the Fourier transforms are taken to produce the two types of power spectra as well as the way flat spectrum foregrounds create oscillatory structures in visibilities. The left-hand diagram of Figure 2.5 shows the real and imaginary components of the emission from a flat-spectrum source as corrugated shading, while the diagonal black lines show the baseline separations at which visibilities are measured from the emission. These diagonal lines also demonstrate how the foreground source creates oscillatory visibilities and how these oscillations get faster for longer baselines. In the right-hand diagram, the oscillations

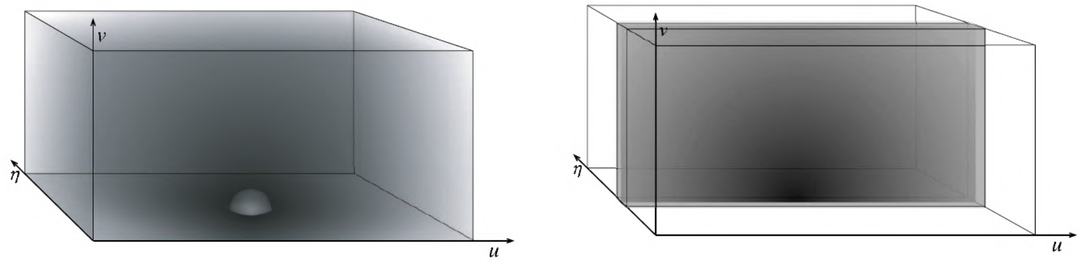

Figure 2.4: The spatial structure of the HI signal (left) and a single residual foreground source (right). While the HI signal shows spherical symmetry in Fourier space, the foreground source shows strong separable-axial symmetry. The symmetry observed for the foreground is due to its assumed spectral smoothness, hence leading to almost all the power being concentrated at small values of η (or τ) and therefore $k_{\|}$. This symmetry difference helps one separate the HI signal and foregrounds in Fourier space. The schematic was taken from Morales & Hewitt (2004).

are shown for the baselines in the left-hand diagram, while also showcasing how the Fourier transform is taken for the reconstructed (the thick dashed vertical line along a fixed angular scale) and measured (thin dash-dot line along the direction of the baselines) power spectrum estimators (Morales et al., 2012).

The *imaging-style* or reconstructed sky power spectrum estimator is essentially a 3D Fourier transform and square of the reconstructed or imaged sky mapped to cosmological coordinates. If this reconstructed sky was equivalent to the true sky, then the foreground power would be confined to the lowest line-of-sight Fourier modes. However, due to errors in this reconstruction, power is leaked to higher modes creating a wedge-like structure (which is explored below). In the *delay-style* estimator, the visibilities are Fourier transformed along frequency and squared to form the power spectrum. Hence, no reconstruction (or imaging) of the sky is attempted. For short baselines, the Fourier transform is almost parallel to the line-of-sight, but not entirely. From the right-hand diagram of Figure 2.5, it is clear that the estimator is along k_τ instead of $k_{\|}$. It is for this reason that the delay spectrum method has its name. Despite this discrepancy in the way the Fourier transform is taken, here it is assumed that this transform is done parallel to the line-of-sight, i.e. $k_\tau \approx k_{\|}$, since the discrepancy has a negligible effect on the cosmological power spectrum (Morales et al., 2012).

The delay spectrum methodology was primarily developed in Morales & Hewitt (2004) and Morales (2005) and subsequently improved and altered to accommodate

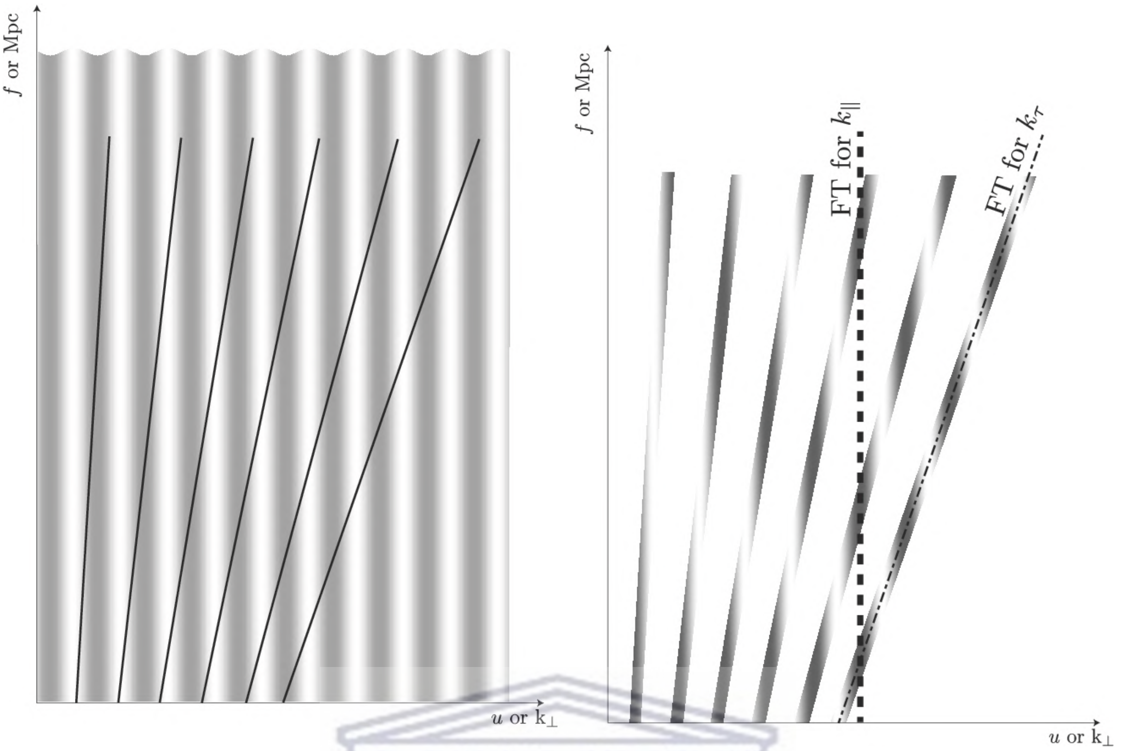

Figure 2.5: The diagram on the left shows how flat spectrum foreground sources generate oscillatory visibilities and shows how these oscillations become faster for longer baselines. The diagram on the right then shows the oscillations for different baselines. In the reconstructed power spectrum case, the Fourier transform is taken along frequency for fixed angular scale (baseline) and is shown as the thick dashed line, while the measured sky power spectrum case takes the Fourier transform along baselines, shown as the thin dash-dot line. This is one of the key distinctions between the reconstructed (imaging-style) and measured (delay-style) power spectrum estimators. Both schematics are from Morales et al. (2019).

its usage for the data or simulation analyses conducted for specific experiments. Primary examples are in the case of PAPER as was done in Parsons et al. (2012a,b) and Pober et al. (2013) as well as for the MWA in Vedantham et al. (2012); Thyagarajan et al. (2013); Paul et al. (2016). Following the formalism presented in Thyagarajan et al. (2013), the observed visibilities are assumed to take the form:

$$\begin{aligned} V^{\mathrm{obs}}(u,v,\nu) =& \Big\{ \left[V^{\mathrm{HI}}(u,v,\nu) + V^{\mathrm{FG}}(u,v,\nu)\right] \otimes W^{\mathrm{P}}(u,v,\nu) \\ & + V^{\mathrm{TN}}(u,v,\nu)\Big\} S(u,v,\nu), \end{aligned} \tag{2.1}$$

where V^{HI} denotes the HI signal visibilities and V^{FG} the foreground visibilities which are then convolved with the spatial frequency response of the power pattern, $W^{\mathrm{P}}(u,v,\nu)$. $W^{\mathrm{P}}(u,v,\nu)$ forms a Fourier pair with $W^{\mathrm{P}}(\theta_x,\theta_y,\nu)$, which denotes the primary beam power pattern. These visibilities are then further contaminated by thermal noise, given in the form of their contribution to the observed visibilities as V^{TN} and are sampled accordingly at baselines in the interferometric array given by a sampling function, S. While the effects of this sampling function are not insignificant, it was accounted for in the simulation pipeline described below by various gridding and averaging processes which are performed to factor in the structure of the uv distribution used. Further, one can Fourier transform these observed visibilities along frequency to obtain (Thyagarajan et al., 2015)

$$V^{\mathrm{obs}}(u,v,\tau) = \int V^{\mathrm{obs}}(u,v,\nu)W^{\mathrm{B}}_{\nu}(\nu)e^{-j2\pi\tau\nu}d\nu, \tag{2.2}$$

where τ denotes the delay and $W^{\mathrm{B}}_{\nu}(\nu)$ is a spectral weighting window function. Equation 2.2 then represents the observable in Fourier space and thus contains the spatial information of the sky that was observed (Morales & Hewitt, 2004). Now, in the case of small bandwidths and uv coverage, $\tau \approx \eta$, where η is the Fourier conjugate of the frequency, ν, or more specifically, the line-of-sight. Hence, the delay, τ, has units of seconds, but physically represents the spatial structure and therefore should be thought of as an inverse distance. This is due to the fact that the frequency of the redshifted 21 cm line maps to the line-of-sight distance of the source (Hogg, 1999), i.e.

$$D(z) = \int_0^z \frac{cd\tilde{z}}{H(\tilde{z})}, \tag{2.3}$$

with $z = \frac{\nu_{21}}{\nu_{\mathrm{o}}} - 1$, $H(\tilde{z})$ denotes the Hubble parameter at redshift $\tilde{z}$ and $\nu_{21} \approx 1420$ MHz is the rest frame frequency of the 21 cm spin-flip transition of Hydrogen. Further, the interferometric coordinates (u,v,τ) can be related to the cosmological coordinates (k_x,k_y,k_z) (spatial wave vectors) by (Morales & Hewitt, 2004; Morales, 2005)

$$\begin{aligned}\vec{u} &= \{u, v\} \\ &= \{\frac{k_x D(z)}{2\pi}, \frac{k_y D(z)}{2\pi}\} \\ &= \frac{\vec{k}_\perp D(z)}{2\pi}\end{aligned} \tag{2.4}$$

and

$$\tau \approx \frac{c(1+z)^2}{2\pi H_0 \nu_{21} E(z)} k_\parallel, \tag{2.5}$$

where H_0, the value of the Hubble parameter today, and

$$E(z) = \left[\Omega_M (1+z)^3 + \Omega_\Lambda\right]^{1/2} \tag{2.6}$$

are standard cosmological terms in a flat ΛCDM universe while $D(z)$ is given by Equation 2.3. Note that here the convention was chosen such that $\{k_x, k_y\} = \vec{k}_\perp$ and $k_z \equiv k_\parallel$, with $\vec{k} = \{\vec{k}_\perp, k_\parallel\}$ and the magnitudes are given by

$$k_\perp = |\vec{k}_\perp| = \left(k_x^2 + k_y^2\right)^{1/2} \text{ and } k = |\vec{k}| = \left(k_\perp^2 + k_\parallel^2\right)^{1/2}. \tag{2.7}$$

The pair of relations in Equation 2.7 gives the cylindrical and spherical averaging of k space.

Figure 2.3 summarises the relationship between the observed visibilities (essentially given by Equation 2.1) and the possible Fourier transforms. Equation 2.2 corresponds to going from the middle cube in Figure 2.3 to the cube on the right, showing the Fourier representation. As discussed above, Figure 2.4 shows the spatial structure of the two primary quantities of interest in the observed visibilities. Due to the different symmetries of each, they can effectively be separated in Fourier space, with the foregrounds being isolated at low τ (and therefore low $k_\parallel$) due to their spectral smoothness and the HI signal spread over the entire Fourier space.

However, the assumption that spectral smoothness should isolate foreground power at low $k_{\parallel}$ does not hold entirely. Due to the inherent chromaticity of an interferometer, foreground contamination spreads to higher values of $k_{\parallel}$, effectively forming a wedge-like structure in Fourier space (as was shown to be the case in, for example, Datta et al., 2010; Morales et al., 2012; Parsons et al., 2012a,b; Thyagarajan et al., 2013, 2015).

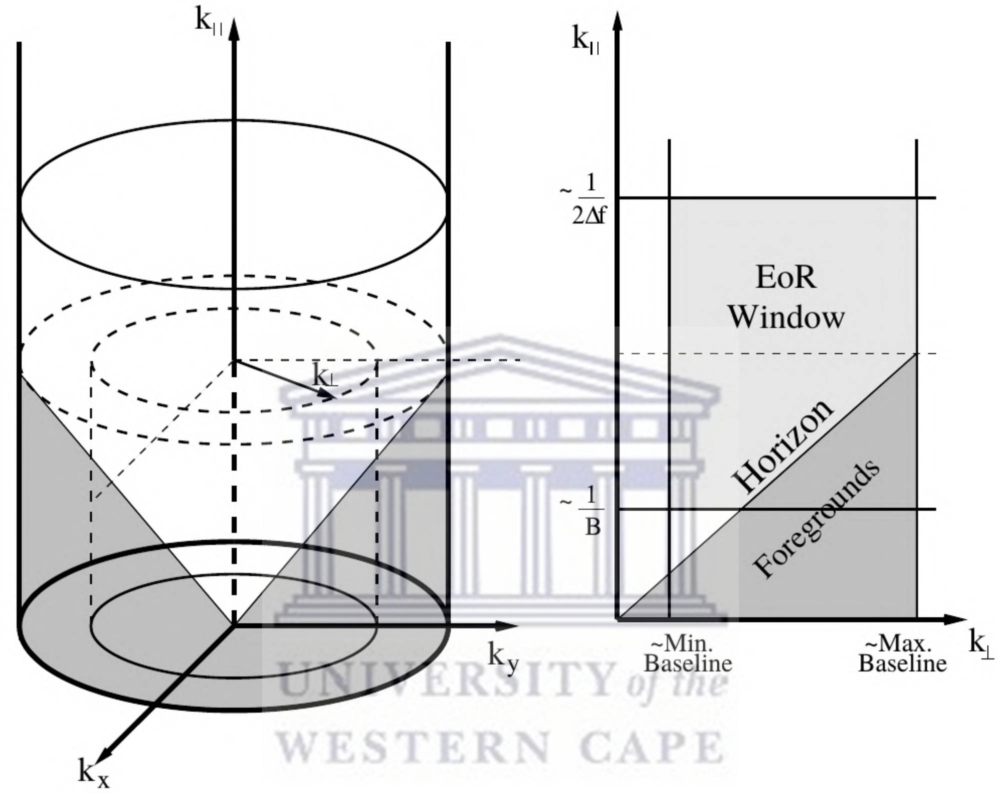

Figure 2.6: Diagram showing the separate regions in k space, including the three-dimensional (left) and two-dimensional (right) cases. In both, the regions that are dominated by foregrounds and dominated by the HI signal are clearly shown. In particular, the shaded region in the left panel is collapsed into the foreground wedge shown on the right. The region known as the HI window above this is believed to be relatively free of contamination and is a candidate region for making an estimate of the HI power spectrum. The transition from three to two dimensions is done by collapsing k_x and k_y into $k_{\perp}$. Also shown are the regions in k space at which instrumental and systematic effects play a role, such as the bandwidth and baselines. The diagram is from Thyagarajan et al. (2013).

In particular, the geometric representation of k space in Figure 2.6 shows the wedge region in both three and two dimensions. The shaded region shown is dominated by foreground sources as well as their sidelobes due to the frequency dependence of the interferometric instrument and is therefore referred to as the foreground wedge.

Since most of the studies that focus on this sort of separation techniques are directed at studying the Epoch of Reionization, the region where the HI signal is dominant is commonly referred to as the EoR window. For the purposes of this study, this region will be referred to as the HI window.

Also shown in Figure 2.6 and further emphasised in Figure 2.7 are the various boundaries of the HI window in the cylindrical k space. The $k_\perp$ boundaries are as a result of the thermal noise increasing where there are sparser amounts of baselines (Chapman et al., 2016).

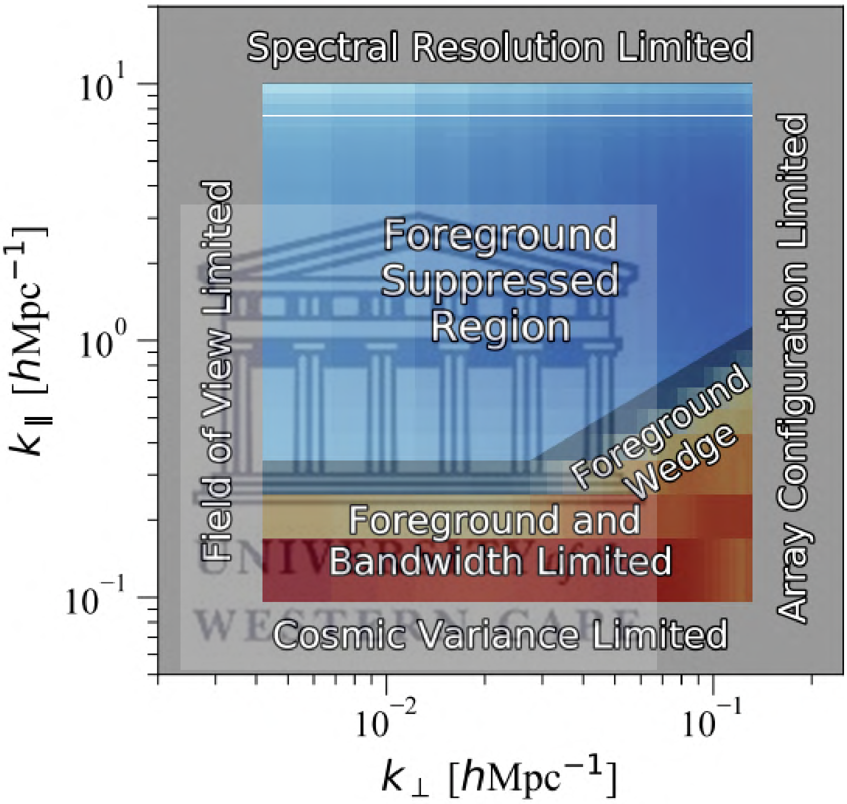
Spectral Resolution Limited
Foreground Suppressed Region
Field of View Limited
Array Configuration Limited
Foreground Wedge
Foreground and Bandwidth Limited
Cosmic Variance Limited
10^1
10^0
10^{-1}
$k_\parallel$ [hMpc^{-1}]
10^{-2}
10^{-1}
$k_\perp$ [hMpc^{-1}]

Figure 2.7: Qualitative diagram showing the two-dimensional regions of k space. The regions shown are specific to radio interferometers, highlighting key things such as the angular k modes that are limited by the extent of the survey area and the array configuration as well as the limits on the highest line-of-sight k modes due to the spectral resolution. For $k_\parallel$, the lowest modes are also limited by cosmic variance, the bandwidth and foregrounds, all highlighted in the diagram. Additionally, the inherent chromaticity of interferometers cause foreground leakage at higher $k_\parallel$ modes for increasing $k_\perp$ modes, leading to the structure known as the foreground wedge, also shown in Figure 2.6. This diagram was taken from Liu & Shaw (2020).

At low $k_\perp$, the HI window is bounded by the shortest baseline, b_{min}. At higher $k_\perp$, this boundary depends on the array configuration and coincides with the angular resolution of the instrument which is effectively given by the longest baseline, b_{max}. Further, the $k_{\parallel}$ boundaries are frequency dependent, where the lower boundary depends on the chosen bandwidth (this can be smaller than the total observational bandwidth) while the higher boundary depends on the frequency resolution of said observation, given by B and $\Delta\nu$, respectively. Mathematically, these boundaries can be expressed as (Vedantham et al., 2012; Chapman et al., 2016)

$$k_{\perp,\text{max}} = \frac{2\pi b_{\text{max}} \nu_{21}}{c\,(1+z)\,D(z)} \;;\; k_{\perp,\text{min}} = \frac{2\pi b_{\text{min}} \nu_{21}}{c\,(1+z)\,D(z)} \tag{2.8}$$

and

$$k_{\parallel,\text{max}} = \frac{2\pi H_0 \nu_{21} E(z)}{c\,(1+z)^2\,\Delta\nu} \;;\; k_{\parallel,\text{min}} = \frac{2\pi H_0 \nu_{21} E(z)}{c\,(1+z)^2\,B}. \tag{2.9}$$

The foreground wedge itself is thought to extend no further than the boundary known as the horizon limit (Parsons et al., 2012b; Thyagarajan et al., 2013):

$$k_{\parallel} = \frac{H_0 E(z) D(z)}{c\,(1+z)} k_{\perp}. \tag{2.10}$$

To understand the power spectrum from a statistical standpoint, the three-dimensional Fourier transform of the sky temperature is defined as (distinct from the representation in Section 1.3.2 in that this represents a means of computing the fluctuations in the sky temperature in k space)

$$\tilde{T}(\vec{k}) = \int_{-\infty}^{\infty} T(\vec{r}) e^{-j\vec{k}\cdot\vec{r}} d^3 r, \tag{2.11}$$

where $\vec{r} = \{\theta_x, \theta_y, D(z)\}$ is the comoving position vector. The inverse is then given as

$$T(\vec{r}) = \frac{1}{(2\pi)^3} \int_{-\infty}^{\infty} \tilde{T}(\vec{k}) e^{j\vec{k}\cdot\vec{r}} d^3 k. \tag{2.12}$$

The power spectrum is then defined through the relation:

$$\langle \tilde{T}(\vec{k})\tilde{T}(\vec{k}')^{*}\rangle \equiv (2\pi)^3\,\delta_D(\vec{k}-\vec{k}')P(\vec{k}), \tag{2.13}$$

where the power spectrum $P(\vec{k})$ is essentially the quantity we seek in using the delay spectrum method. The delay-space power spectrum itself can then be calculated from the correlation of the visibilities since the visibilities are themselves Fourier transforms of the three-dimensional brightness temperature and therefore contain information about the power spectrum in their correlation given by (Morales & Hewitt, 2004; Parsons et al., 2012a; Thyagarajan et al., 2013)

$$\begin{aligned} P_{\mathrm{d}}(\vec{u}) &= \left\langle V^{\mathrm{obs}}(\vec{u}_i)^{*}V^{\mathrm{obs}}(\vec{u}_j)\right\rangle\delta_{ij} \\ &= \left\langle \left|V^{\mathrm{obs}}(\vec{u})\right|^2\right\rangle, \end{aligned} \tag{2.14}$$

where $\vec{u} = \{u, v, \tau\}$. The delay power spectrum in k space is then obtained by performing a coordinate transform $(u, v, \tau) \longrightarrow (k_x, k_y, k_z)$ via a Jacobian matrix which gives (Thyagarajan et al., 2013, 2015)

$$\begin{aligned} P_{\mathrm{d}}(\vec{k}) &= P_{\mathrm{d}}(\vec{u})|J(\vec{u})| \\ &= P_{\mathrm{d}}(\vec{u})\begin{vmatrix} \frac{\partial k_x}{\partial u} & \frac{\partial k_x}{\partial v} & \frac{\partial k_x}{\partial \tau} \\ \frac{\partial k_y}{\partial u} & \frac{\partial k_y}{\partial v} & \frac{\partial k_y}{\partial \tau} \\ \frac{\partial k_z}{\partial u} & \frac{\partial k_z}{\partial v} & \frac{\partial k_z}{\partial \tau} \end{vmatrix} \\ &= \left\langle \left|V^{\mathrm{obs}}(\vec{u})\right|^2\right\rangle \left(\frac{A_e}{\lambda^2 B}\right)\left(\frac{D^2\Delta D}{B}\right)\left(\frac{\lambda^2}{2k_B}\right)^2, \end{aligned} \tag{2.15}$$

where the visibilities given in Equation 2.14 have been substituted into Equation 2.15. A_e and λ denote the effective antenna area and wavelength corresponding to the central frequency of the observational band, respectively and k_B denotes the Boltzmann constant. $D \equiv D(z)$ here denotes the transverse comoving distance, given in Equation 2.3, while ΔD is the comoving depth along the line-of-sight corresponding to the bandwidth, B (Thyagarajan et al., 2015), given by (Parsons et al., 2012a) as

$$\Delta D(z) = \frac{c\,(1+z)^2}{H_0 E(z) \nu_{21}} \tag{2.16}$$

where it is denoted by Y and given indirectly through the relation, $Yk_z = 2\pi\tau$. The Jacobian transformation acts to convert the units of the power spectrum from the Jy^2, obtained by squaring the visibilities, to the cosmological units of power, $\mathrm{mK}^2\mathrm{Mpc}^3$, in addition to normalising the k space power spectrum. Importantly, it should be noted that the delay power spectrum is an approximation to the cosmological HI power spectrum, due to it being the convolution of the HI power spectrum with the window functions (for a full expression of this relation, see (Thyagarajan et al., 2015; Paul et al., 2016)). Due to this, the delay spectrum is the quantity used to estimate the input HI power spectrum (given in Section 1.1.2 as Equation 1.24) from the simulations.

The discussed theoretical framework forms the basis of the simulations detailed in the sections that follow (particularly Equation 2.15 and the related quantities). Section 2.3 discusses the foreground simulations that use the delay spectrum methods to study the nature of foregrounds in delay space, while Section 2.4 then details out a full simulation pipeline that aims to mimic the observational and power spectrum estimation processes while including the key components - the HI signal, thermal noise and foregrounds - as they are given in Equation 2.1. Also discussed are the power spectrum estimation techniques employed to test the pipeline as well as the various cases considered.

2.3 Foreground simulations

Analysing the nature and structure of foregrounds underpins many of the key cosmological probes such as the statistical detection of the HI signal. The delay spectrum method allows one to probe modes that are not contaminated by foregrounds through avoidance. Despite this, it is useful to study foregrounds to understand how they behave beyond the foreground wedge, which is thought to be the limit of foreground contamination in Fourier space (as discussed in, for example: Vedantham et al., 2012; Pober et al., 2013; Chapman et al., 2016). The objective of the foreground simulations were then to use an input model of the sky and calculate power spectra from said models as well as the relevant k modes corresponding to the spectral and angular properties of the simulated interferometer.

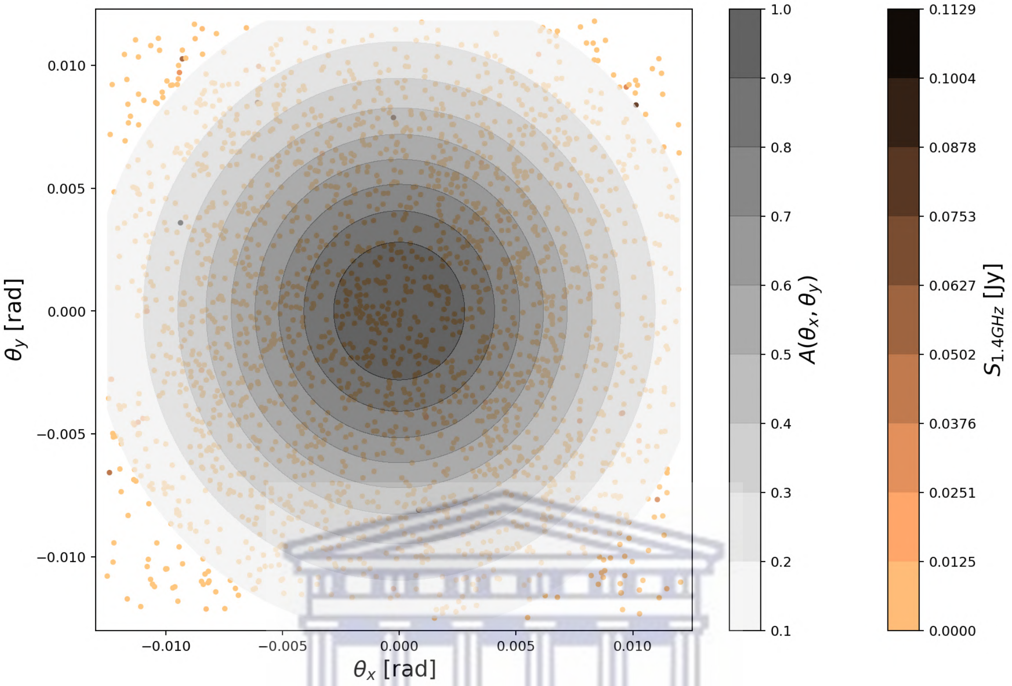

Figure 2.8: The sky model produced from the VLA COSMOS Large project (Schinnerer et al., 2007). The area of the model spans 2 deg^2 and contains 2417 sources. The primary beam overlaid on the sky model is a normalised Gaussian.

The specific mathematical model representing the sky chosen for the foreground simulations can be expressed as (Paul et al., 2016)

$$\vec{T}_{\text{sky}}(\vec{\theta}, \nu) = \sum_i S_{\nu,i} \; \delta_D^2(\vec{\theta} - \vec{\theta}_i), \tag{2.17}$$

where ν specifies the frequency corresponding to the flux, $S_{\nu,i}$, and the position of each source is specified by the 2D Dirac delta function, $\delta_{\text{D}}^2(\vec{\theta} - \vec{\theta}_i)$. Note that this represents a sky with only a discrete set of extragalactic point sources. For this study, a catalogue of sources from the VLA COSMOS 1.4 GHz Large Project (Schinnerer et al., 2007), with 2417 of the original total of 3643 sources over an area of sky spanning 2 deg^2 was chosen as the sky model as the focus of the full simulation pipeline is also on the COSMOS field. The simulated bandwidth for the model was chosen to be 75 MHz. Further, the central frequency for the VLA

COSMOS sky model was scaled down from 1.4 GHz to 1 GHz using a spectral index of α = -0.78 (Ishwara-Chandra et al., 2010), on the assumption that the sources in the catalogue were all extragalactic point sources. This central frequency corresponds to a wavelength of $\lambda = 30$ cm and well as a redshift of $z = 0.42$ for the VLA COSMOS sky model, and therefore, for the overall simulated observation. This was chosen to test how foregrounds would behave in the L-band of the MeerKAT interferometer, which spans 900 - 1670 MHz (Booth et al., 2009) and thus enable the investigation of foregrounds in the MIGHTEE survey, for which observations are being made in this band.

The sky model for the VLA COSMOS catalogue is shown in Figure 2.8. It has been centred on zero and also shows the response of a normalised Gaussian primary beam on the field of view. In addition to the sky model, a model for the uv distribution was chosen as input to the foreground simulations. In particular, an actual uv distribution generated for about 11.2 hours of observation time on the COSMOS field with the MeerKAT interferometer as part of the MIGHTEE survey was chosen for this purpose. This distribution is shown in Figure 2.9 and shows the tracks generated over time as the uv points map out uv space. This particular figure shows the distribution at 1115.14 MHz (which is the central frequency for the simulation pipeline described in Section 2.4). In the foreground simulations, the same uv distribution was used, but set at a frequency of 1 GHz (the central frequency of the simulated observation).

Generally, the foreground simulations consist of performing a Fourier transform of the sky model multiplied by the primary beam, $W^{\mathrm{P}}(\theta_x, \theta_y, \nu)$, along the spatial axes to obtain the visibilities, $V^{\mathrm{FG}}(u, v, \nu)$. Since these simulations only consider the foreground component, they represent a special case of Equation 2.1, in which $V^{\mathrm{obs}} = V^{\mathrm{FG}}$. Specifically, a foreground visibility is generated via the Fourier transform of Equation 2.17, given as (Paul et al., 2016)

$$V^{\mathrm{FG}}(u, v, \nu) = \sum_i W^{\mathrm{P}}(\vec{\theta}_i, \nu) S_{\nu,i} \exp(-j2\pi\vec{u} \cdot \vec{\theta}_i). \tag{2.18}$$

Using Equation 2.18, the visibilities are generated per baseline instead of a uv grid, over the entire frequency range corresponding to the selected sky model. In addition, the frequency channel width can also be set by choosing the number of frequency

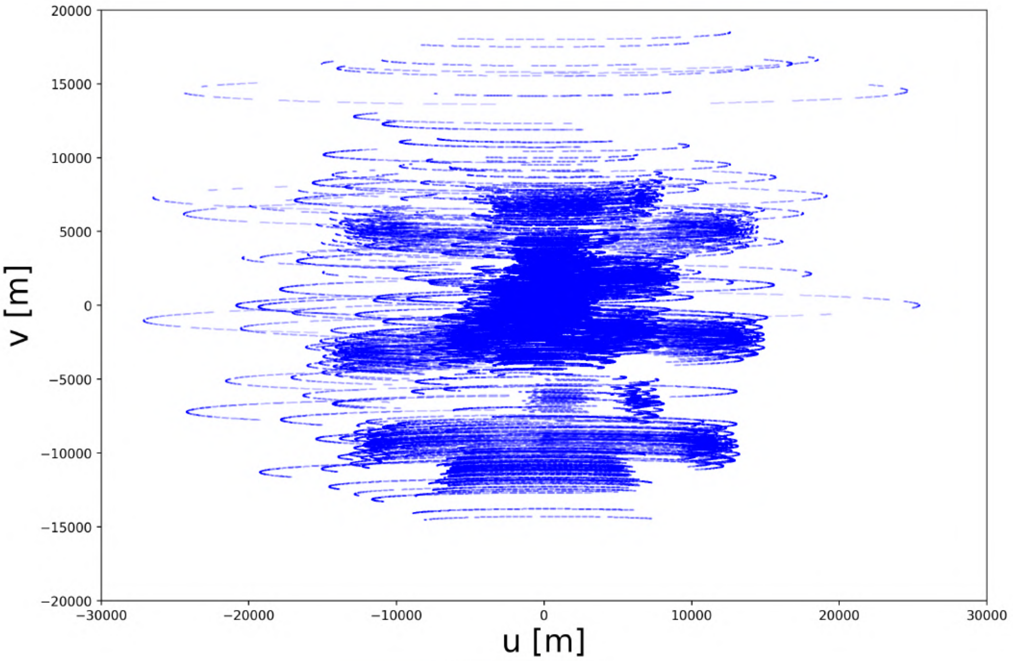

Figure 2.9: The *uv* distribution of an observation of the COSMOS field for a period of 11.2 hours with the MeerKAT interferometer as part of the MIGHTEE survey, at a frequency of 1115.14 MHz.

channels over the selected bandwidth for the VLA COSMOS sky model. Although this simulation mimics an interferometric observation, no flagging or effects related to the flagging of channels were included, since the focus was on analysing an ideal case where all the visibilities are uncontaminated and only contain information about the foregrounds in the sky model so as to investigate whether or not the foreground wedge would be generated.

The model used for the primary beam power pattern was specifically set to a normalised Gaussian in the foreground simulation. This was done primarily for simplicity, since modelling the Fourier response of the primary beam of a real interferometer can become exceedingly challenging (Paul et al., 2016). However, since the primary beam, $W^{\mathrm{P}}(\theta_x, \theta_y, \nu)$ is Gaussian, its Fourier transform, $W^{\mathrm{P}}(u, v, \nu)$, is also Gaussian, thus reducing the complication considerably. The intricate effects of the primary beam were not studied extensively, and so are just included as a standard in the interferometric calculations.

The visibilities, V^{FG}, were then Fourier transformed along the frequency axis and multiplied by a spectral weighting window function, W^{B}_{ν}, to obtain its representation in delay space, $V^{\mathrm{FG}}(u, v, \tau)$. This can be expressed mathematically as

$$V^{\mathrm{FG}}(u, v, \tau) = \sum_k V(u, v, \nu_k) \exp(-j2\pi\nu_k\tau) W^{\mathrm{B}}_{\nu}(\nu_k). \tag{2.19}$$

The inclusion of a window function serves the purpose of minimising foreground spillover to regions beyond the foreground wedge. While any tapering function could in principle be used for this purpose, two specific functions were chosen:

(i) Case where $W^{\mathrm{B}}_{\nu} = 1 \; \forall \; n$, i.e. the Tophat or Boxcar window function

(ii) Case where W^{B}_{ν} is given by the Blackman-Harris window function (Harris, 1978)

The Blackman-Harris window function is given by,

$$\begin{aligned} w(n) = a_0 - a_1 \cos\left(\frac{2\pi}{N} n\right) + a_2 \cos\left(\frac{2\pi}{N} 2n\right) \\ - a_3 \cos\left(\frac{2\pi}{N} 3n\right), \end{aligned} \tag{2.20}$$

where $n = 0, 1, ..., N-1$.

In particular, the window function presented here is the four-term Blackman-Harris with sidelobes at a -92 dB level, corresponding to $a_0 = 0.35875$, $a_1 = 0.48829$, $a_2 = 0.14128$ and $a_3 = 0.01168$. The Boxcar window function is only defined for completeness, as it is assumed to always be applied in the Fourier transform along the frequency where no other window function is applied. Figure 2.10 shows both windows normalised to unity alongside their Fourier responses, which clearly show the main lobe and sidelobes of both window functions.

The choice of the Blackman-Harris window was made due to it having a high level performance in signal analysis, even when compared to functions such as the Kaiser-Bessel, Dolph-Chebyshev and Barcilon-Temes windows, for instance (Harris, 1978). Lastly, it should also be noted that the definitions of the two chosen window functions are discrete in nature, as they are assumed to be applied to data sets which are discrete in nature (which is always the case in realistic scenarios).

In addition to the transformation of the visibilities to delay space, the τ values were sampled from the frequency range with the minimum value of τ corresponding to the reciprocal of the bandwidth, $\frac{1}{B}$, while the maximum value was set to correspond to the Nyquist limit, $\frac{1}{2\Delta\nu}$, where $\Delta\nu$ denotes the frequency channel width.

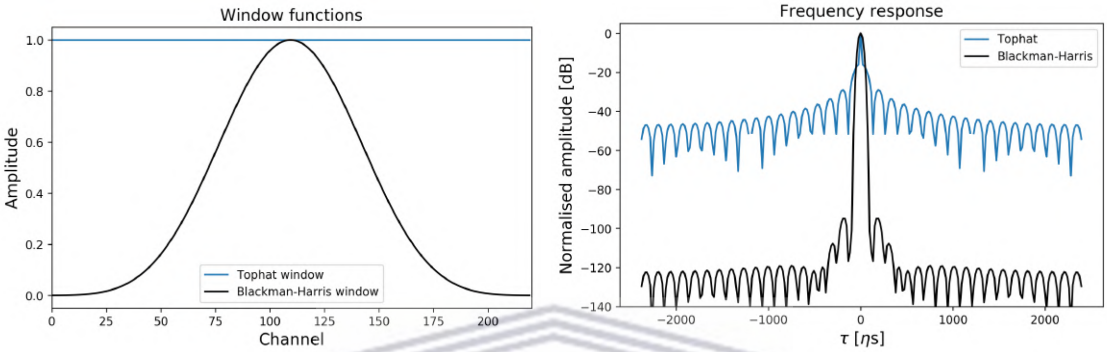

Figure 2.10: The Blackman-Harris and Boxcar/Tophat window functions (left) alongside their Fourier transforms (right). Here, one clearly sees the loss of signal that results in using the Blackman-Harris window function. While this window function does reduce spillover from the foreground wedge into the HI window, it also suppresses the signal and has to be accounted for in the simulations in order to recover the full cosmological signal. As is clear from the frequency response plot, the sidelobe level of the Blackman-Harris window function is at -92 dB. The channel number (in the left-hand plot) and delay range (right-hand plot) are specific to the full simulation pipeline, while the y-axes of both plots show the general ranges associated with both windows.

The power spectrum was then calculated by taking the square of the visibility set in delay space and multiplying by the relevant normalisation factors, using Equation 2.15. This yields the simulated power spectrum in cosmological units. However, it should be noted that no structural information would be lost if the normalisation factor were to be neglected in Equation 2.15, since it only acts to change the units and alter the amplitude of the power spectrum by said factor.

The power spectrum calculated here is three-dimensional and can be cylindrically averaged to obtain a 2D power spectrum, $P(k_\perp, k_\parallel)$, by collapsing the k_x and k_y axes to form $k_\perp$, as demonstrated in Figure 2.6 and using Equation 2.7. In addition, a spherically averaged power spectrum, $P(k)$, can be obtained by averaging the 3D power spectrum in spherical shells and again using Equation 2.7 to obtain the relevant k modes. For the purpose of studying the foregrounds, only the cylindrically averaged power spectrum, $P(k_{\perp}, k_{\parallel})$, was calculated in the foreground power spectrum simulations since the primary objective here was to investigate the structure and location of the foregrounds in delay space with the purpose of its avoidance in eventual estimations of the HI power spectrum from an input model (as is described in Section 3.2.1 after delineating the processes of the full simulation pipeline in Section 2.4).

The $k_{\perp}$ modes were sampled from the uv distribution using Equation 2.4 with a bin size of $\Delta k_{\perp} = 0.45$. One thing to note here is that the wavelength used to perform this sampling corresponds to the central frequency of the selected bandwidth, since the delay spectrum approximation is being utilised. Likewise, the $k_{\parallel}$ modes were sampled from the non-negative τ values using Equation 2.5. Additionally, the central wavelength (frequency) of the selected frequency range corresponds to a redshift that was used to calculate all the cosmological parameters, such as the comoving distance and Hubble parameter, which were used to sample the $(k_{\perp}, k_{\parallel})$ modes. Together, these quantities form the primary outputs of the foreground simulations. Table 2.1 essentially summarises the key parameters used in the foreground simulations as well as in the calculations carried out to eventually produce power spectra.

Simulation parameter	Value
Sky model	VLA COSMOS
Bandwidth	75 MHz
ν_c	1 GHz
λ_c	30 cm
z_c	0.42
Source count	2417
uv distribution	MeerKAT (11.2 hours)
Window function	Blackman-Harris
Primary beam model	Gaussian

Table 2.1: Table summarising the key parameters used in the foreground simulations on the VLA COSMOS sky model. Particularly, it highlights the key input components such as the window function used, source count in the sky model, and bandwidth.

Using the foreground simulations with the methodology and specific parameters described, the objective was then to see if the output 2D power spectrum would showcase the delay style analysis signature (Morales et al., 2019). The result of this is shown in Figure 2.11.

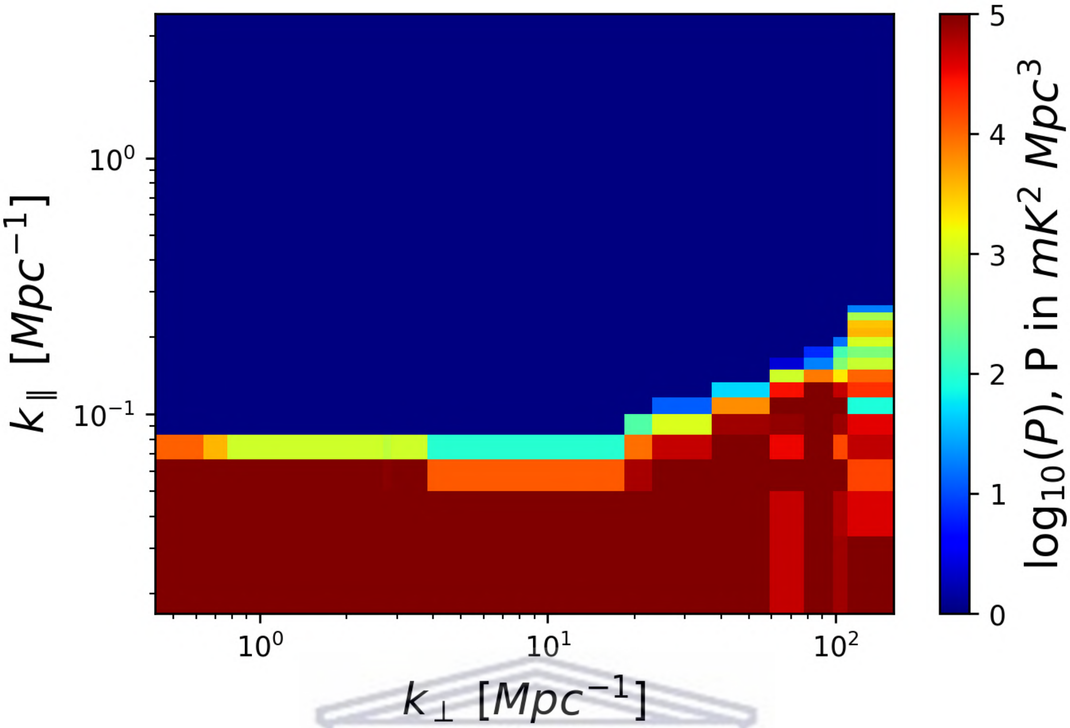

Figure 2.11: The 2D foreground power spectrum for the VLA COSMOS source model and MeerKAT *uv* distribution over 11.2 hours of observation, generated using the foreground specific simulations. Here, the Blackman-Harris window function was again applied to reduce spillover. In addition, this power spectrum is normalised to cosmological units [mK^2 Mpc^3]. Further, the minimum $k_\perp$ mode sampled in the simulation is $k_\perp \sim 0.43$ Mpc^{-1}, thus motivating a choice to set the transverse k mode bin size to $\Delta k_\perp = 0.45$ Mpc^{-1}.

Specifically, Figure 2.11 shows the power contribution from foregrounds situated in the expected lower $k_\parallel$ modes. Most of the power occupies the region below the foreground wedge (defined by Equation 2.10), indicating that the processes applied in the foreground simulations are accurate. Besides some spillover of power above the wedge, there are no structures present in the power spectrum above the foreground dominated region, which is expected since the simulated observations contained no information other than that from foregrounds. Lastly, a point worth noting is that the foreground contribution is solely from point sources in the catalogue and so this 2D power spectrum represents the expected power for a sky containing only point sources.

This result demonstrates the advantage of the delay style (Morales et al., 2019) power spectrum analysis, which is essentially isolating the foregrounds to a wedge-like structure in Fourier space which comes about due to mode mixing as well as the smooth frequency structure of the foregrounds. The objective was to check if one is able to utilise foreground avoidance to extract an estimate of the HI power spectrum from an observation which contains the HI signal, while being contaminated by thermal noise, thus motivating the investigation of foregrounds in their own right. For a diagrammatic outline of the foreground simulations discussed here, refer to Appendix B.

2.4 Simulation Pipeline: Power spectrum from MIGHTEE

Whereas the previous section discussed simulations focused solely on generating power spectra on foreground models, the full simulation pipeline takes into account every component in Equation 2.1. It therefore includes HI signal, foreground and thermal noise information in the observed visibilities. The simulation pipeline essentially generates visibilities for each baseline, and therefore each point on the uv distribution generated by the interferometer during an observation. Hence, the simulated visibilities take the form,

$$\begin{aligned} V(u_i, v_i, \nu) = &\, V^{\mathrm{TN}}(u_i, v_i, \nu) \\ &+ \left[V^{\mathrm{HI}}(u_i, v_i, \nu) + V^{\mathrm{FG}}(u_i, v_i, \nu)\right] \otimes W^{\mathrm{P}}(u_i, v_i, \nu), \end{aligned} \tag{2.21}$$

added to the full visibility set in the pipeline. As mentioned, the observation by the interferometer itself was simulated by the input uv distribution. As the full simulations aimed to mimic an observation from the MIGHTEE survey, the uv distribution used was therefore from the MIGHTEE COSMOS field for 11.2 hours (Figure 2.9).

Further, the subset frequency band used in this work has a central frequency value of 1115.14 MHz and spans a bandwidth of $\sim$ 46 MHz. Additionally, the time resolution, Δt, is 8 seconds with a frequency channel width, $\Delta\nu$, of 0.208984 MHz over 220 frequency channels. The fixed central frequency corresponds to a redshift of approximately 0.27, therefore setting all outputted power spectra at this particular redshift. Other important input parameters are listed in Table 2.2. Together these parameters form the basic inputs to the pipeline from a measurement set generated

Parameter	Value
Δu	$60\ \lambda$
Δv	$60\ \lambda$
$N_{\text{gridpoints}}$	1500
N_{chan}	220
Δt	8 s
$\Delta \nu$	0.208984 MHz
ΔB	45.97648 MHz
ν_c	1115.14 MHz
λ_c	26.9 cm
z	~ 0.27
$\frac{A_e}{T_{\text{sys}}}$	6.22 m^2/K
t_{obs}	11.2 hours
$\Delta k_{\perp}$	0.35 Mpc^{-1}
$\Delta k_{\parallel}$	0.031 Mpc^{-1}
Δk	0.45 Mpc^{-1}

Table 2.2: Table summarising the key parameters used in the full simulation pipeline. The parameters included in the table are those that were fixed for all the science cases considered in this study. The observation time, t_{obs} is the primary parameter altered in the science cases considered, as the objective was to see how well the input HI power spectrum model can be recovered from the simulation pipeline, in which the rest of the parameters mimic those of an actual MIGHTEE COSMOS observation.

from a MIGHTEE COSMOS observation and were used to calculate all relevant quantities used in the pipeline such as the power spectrum normalisation factor (such as those given in Equation 2.15).

2.4.1 Simulation processes

Below, the step-by-step processes involved in the simulation pipeline are outlined in detail:

(i) Each baseline corresponds to a (u, v) coordinate which changes after every integration interval over the duration of the tracking. The uv coordinates are extracted from the measurement set of the MIGHTEE COSMOS observation over a time period of 11.2 hours. These uv points are calculated from the baseline distribution at the central frequency of the chosen bandwidth (1115.14 MHz), so that for a given time the baseline gives the same uv point at all frequencies. The contributions from the thermal noise, HI signal and foregrounds (as in Equation 2.21) are then calculated at each (u, v) point, and therefore per baseline.

(ii) To model the foregrounds, data from a single pointing of the COSMOS field over an on-source integration time of 11.2 hours is processed. The data is put through flagging and calibration using the processMeerKAT pipeline and then split into a sub-band spanning 950 - 1150 MHz. This data was then further processed for continuum imaging using deconvolution and self-calibration. The total intensity image generated from this is shown in Figure 2.12. Further, the CLEAN components obtained in this data processing are then used as the foreground model in the simulation pipeline. These components form the model visibilities from the measurement set. Due to no beam corrections being applied, the foreground visibilities will include some beam effects (hence the beam's inclusion in Equation 2.21). Due to this measurement set being part of a MIGHTEE COSMOS observation, there is minimal contribution from diffuse emission (Paul et al., 2020). Therefore, one can make the assumption that the only significant contribution to the foreground model comes from extragalactic point sources. For more details on the data processing and modelling of the foreground contribution, refer to Paul et al. (2020).

(iii) Each visibility generated from the input uv distribution has a contribution from thermal noise. Now, in order to simulate this contribution to the visibilities, at each (u, v) point (or baseline) in the dataset, the real and imaginary components of V^{TN} were randomly sampled from a Gaussian distribution with a mean of zero and standard deviation given by (Morales, 2005)

$$\sigma_{\mathrm{TN}} = \frac{2k_{\mathrm{B}}T_{\mathrm{sys}}}{A_{\mathrm{e}}\sqrt{\Delta\nu\Delta t}}, \tag{2.22}$$

where k_{B} denotes Boltzmann's constant, while T_{sys} gives the system temperature of the interferometer (MeerKAT in this case). A_{e} then gives the effective area of an individual antenna of said interferometer with Δt and $\Delta\nu$ denoting the time and frequency resolution, respectively. The thermal noise contribution is approximately constant across the frequency channel, and only vary across uv space since the thermal noise level will depend on the density of baselines in the uv distribution (Paul et al., 2020).

(iv) A visibility-frequency cube was generated with uniform uv bin spacing, $\Delta u =$

$\Delta v = 60\lambda$ and the frequency axis consisting of 220 channels with the spacing given by the frequency resolution, $\Delta\nu$. The λ value given for the bin size is specifically fixed to the wavelength corresponding to the central observational frequency. Moreover, the bin size was motivated by the size of the primary beam in Fourier space (Paul et al., 2020). Now, due to the presence of radio frequency interference (RFI) in the calibrated visibilites from which the foreground model and uv distribution are taken, a criterion was set such that only those baselines for which the visibility measurements have 80% of their channels unflagged were considered. To further reduce the contamination, the remaining flagged channels were filled out with the foreground visibilities from the nearest neighbour unflagged channel. These measures were taken to minimize the spillover of foreground power to higher $k_{\parallel}$ modes in the calculated power spectrum, with the latter also specifically ensuring that there are no zero-valued frequency channels in the delay transformation done along the frequency axis.

(v) The foreground and thermal noise visibilities within a uv pixel are then averaged using the assumption that the sky signal is the same across every baseline contributing to that particular grid point. After the thermal noise and foreground contributions have been added and averaged on the grid, the HI signal visibility contribution, V^{HI}, was added. To model the HI signal visibilities, the input model HI power spectrum given by Equation 1.24 was used as the variance of a Gaussian distribution with a mean of zero ($\mu_{\mathrm{HI}} = 0$). Mathematically, the variance can be expressed as

$$\sigma^2_{\mathrm{HI}} = \frac{P_{\mathrm{HI}}(k)}{\left(\frac{A_e}{\lambda^2 B}\right)\left(\frac{D^2 \Delta D}{B}\right)\left(\frac{\lambda^2}{2k_B}\right)^2}, \tag{2.23}$$

so that it is actually the normalised input model HI power spectrum. From this distribution, the real and imaginary parts of each HI visibility were randomly sampled at each (u, v) point. However, before these visibilities were added to the grid, they were in the coordinates, $V^{\mathrm{HI}}(u, v, \tau)$, due to the fact that the real and imaginary components were sampled from a distribution which had a Fourier space variance, i.e., the input HI power spectrum, and so had to be inverse fast Fourier transformed (IFFT) along frequency for the representation in the standard coordinates, $V^{\mathrm{HI}}(u, v, \nu)$. Mathematically, this inverse Fourier transform is given by

$$V^{\mathrm{HI}}(u,v,\nu) = \int V^{\mathrm{HI}}(u,v,\tau)e^{j2\pi\tau\nu}d\tau. \tag{2.24}$$

This quantity, $V^{\mathrm{HI}}(u,v,\nu)$, was then added to the averaged visibility per grid point. To do this, two crucial assumptions were made. Firstly, that the HI signal is the same for all points in each uv pixel, which is true if the eventual $uv\tau$ bins after delay transforming the $uv\nu$ cube are small enough. Secondly, that the values in different pixels are uncorrelated, which is only true if a $uv\tau$ bin is large enough compared to the primary beam and bandwidth. Hence, there is a tension in the choice of the eventual $uv\tau$ bin size, or equivalently, $\Delta k_\perp$ and $\Delta k_\parallel$, since $\Delta k_\perp$ should be set by the telescope primary beam, and $\Delta k_\parallel$ by the bandwidth used. The only alternative would be to include the correlation between Fourier modes using Equations 2.1 and 2.2 (since Equation 2.15 is an approximation).

(vi) With the observed visibilities generated by the pipeline on a three-dimensional $uv\nu$ cube, the next step is to proceed with calculating the power spectrum using the delay space methodology. Firstly, the grid is multiplied along the frequency axis by a Blackman-Harris spectral weighting window function (given by Equation 2.20). As discussed in the section on foreground simulations, this is done to minimise foreground leakage into the region of Fourier space dominated by the HI signal and thermal noise (everything above the foreground wedge). The product of the grid and window function is then fast Fourier transformed (FFT) along the frequency axis. This is essentially given by Equation 2.2. The delay power spectrum is then calculated using Equation 2.15, with the normalisation factors calculated from the input parameters. From this three-dimensional power spectrum, $P_{\mathrm{d}}(k_x,k_y,k_z) = P_{\mathrm{d}}(\vec{k}_\perp,k_\parallel)$, the cylindrically (2D) and spherically averaged (1D) power spectra can be calculated. These are denoted by $P(k_\perp,k_\parallel)$ and $P(k)$, respectively.

(vii) To obtain these 2D and 1D power spectra, an inverse noise weighting is applied to the 3D delay power spectrum during the averaging processes involved (Paul et al., 2020). In order to calculate the 1D power spectrum, the k modes which lie in the foreground dominated region were excluded using the boundary given by Equation 2.10. A polynomial relating $k_\parallel$ to $k_\perp$ was used to approximate this theoretical expression. This was then used as a criterion for calculating the

1D power spectrum by averaging and weighting above the foreground region (Paul et al., 2020). Due to foreground avoidance, the modes which were known to be contaminated by foregrounds were then excluded from the process of calculating the 1D power spectrum.

(viii) The centre of each bin in the 3D power spectrum was chosen as the (u, v) values from which the $k_{\perp}$ modes were calculated. A bin size, $\Delta k_{\perp} = 0.35$ Mpc^{-1}, which is approximately the same as the lowest mode, $k_{\perp,\mathrm{min}} \sim 0.33$ Mpc^{-1}, was chosen to perform this calculation. Additionally, the $k_{\|}$ modes were calculated on the sampled τ values, with a bin size of $\Delta k_{\|} = 0.031$ Mpc^{-1}. To perform both these samplings, Equation 2.7 was used. Since the 1D power spectrum is represented in k space, these modes were calculated using logarithmic bins in k, Δk, which increase as $(\vec{k}_{\perp}, k_{\|})$ increases in 3D k space. The smallest bin, Δk, as well as the $k_{\perp}$ and $k_{\|}$ bin sizes are shown in Table 2.2.

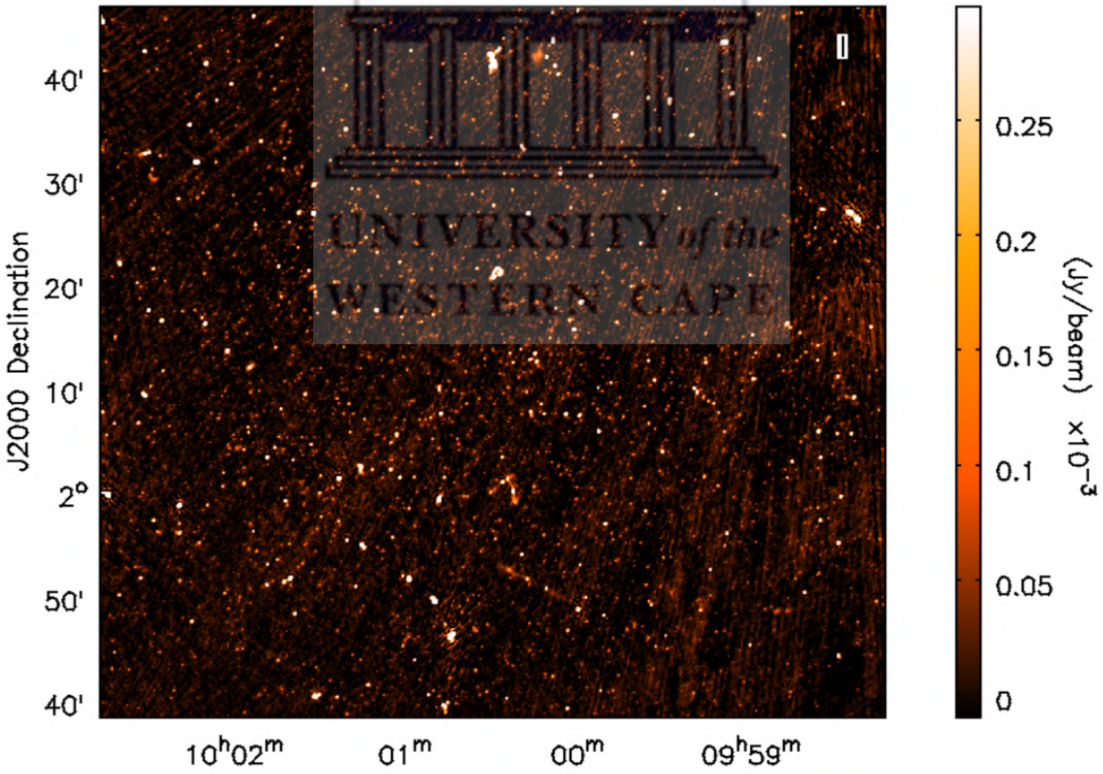

Figure 2.12: The total intensity image of the COSMOS field at 1115.14 MHz, generated from 11.2 hours of data from the MIGHTEE survey. The continuum model generated in the imaging process is utilised in the simulation pipeline to generate the foregrounds. Figure taken from Paul et al. (2020).

Overall, the 1D and 2D power spectra, along with the sampled k modes make up the main outputs from the pipeline, with the 1D power spectra being the main simulation product utilised in the estimator analysis discussed in the next section.

3 Results

In this section, the output power spectra from the simulation pipeline are presented. These include both the 1D and 2D power spectra of the thermal noise, foregrounds and HI signal, as well as the power spectrum from a simulated observation containing information from each of these components. Additionally, estimates of the HI power spectrum are presented with their errors. These results were generated through a statistical analysis employing Monte Carlo processes to calculate the final errors. Further, this estimator analysis was performed for various observation time cases, as well as for the case of the auto- and cross-correlation of visibilities in the pipeline.

3.1 Simulation pipeline

Since the simulation pipeline allows one to include each of the key components in different combinations (HI signal, thermal noise, foregrounds), it also allows one to generate the power spectra of these combinations and therefore of each individual component. To do this, the simulation generates visibilities, $V(u, v, \nu)$, for a given component, then delay transforms them along the frequency axis to obtain $V(u, v, \tau)$. These delay space visibilities are then squared and multiplied by a normalisation factor (Equation 2.15) before being used to generate power spectra through cylindrical (2D, $P(k_\perp, k_\parallel)$) and spherical (1D, $P(k)$) averaging, as described in Section 2.4.

This is shown in the case of the HI signal power spectrum in both 2D and 1D in Figure 3.1. Here, only the contribution of the HI signal was included in the simulated visibilities which were then used to generate the power spectra. The 2D power spectrum demonstrates the spherical symmetry inherent in the HI signal in frequency space with the power producing almost constant contours along annuli in k space. The 1D power spectrum shows the mean values over 1000 realisations produced by the simulation pipeline plotted alongside the input model HI power spectrum as a continuous curve and points averaged to correspond to the same k modes as those sampled by the simulation pipeline.

As the simulation pipeline used this input model to produce the HI visibilities in the pipeline, one expects the realisations to match this input model to some degree. In

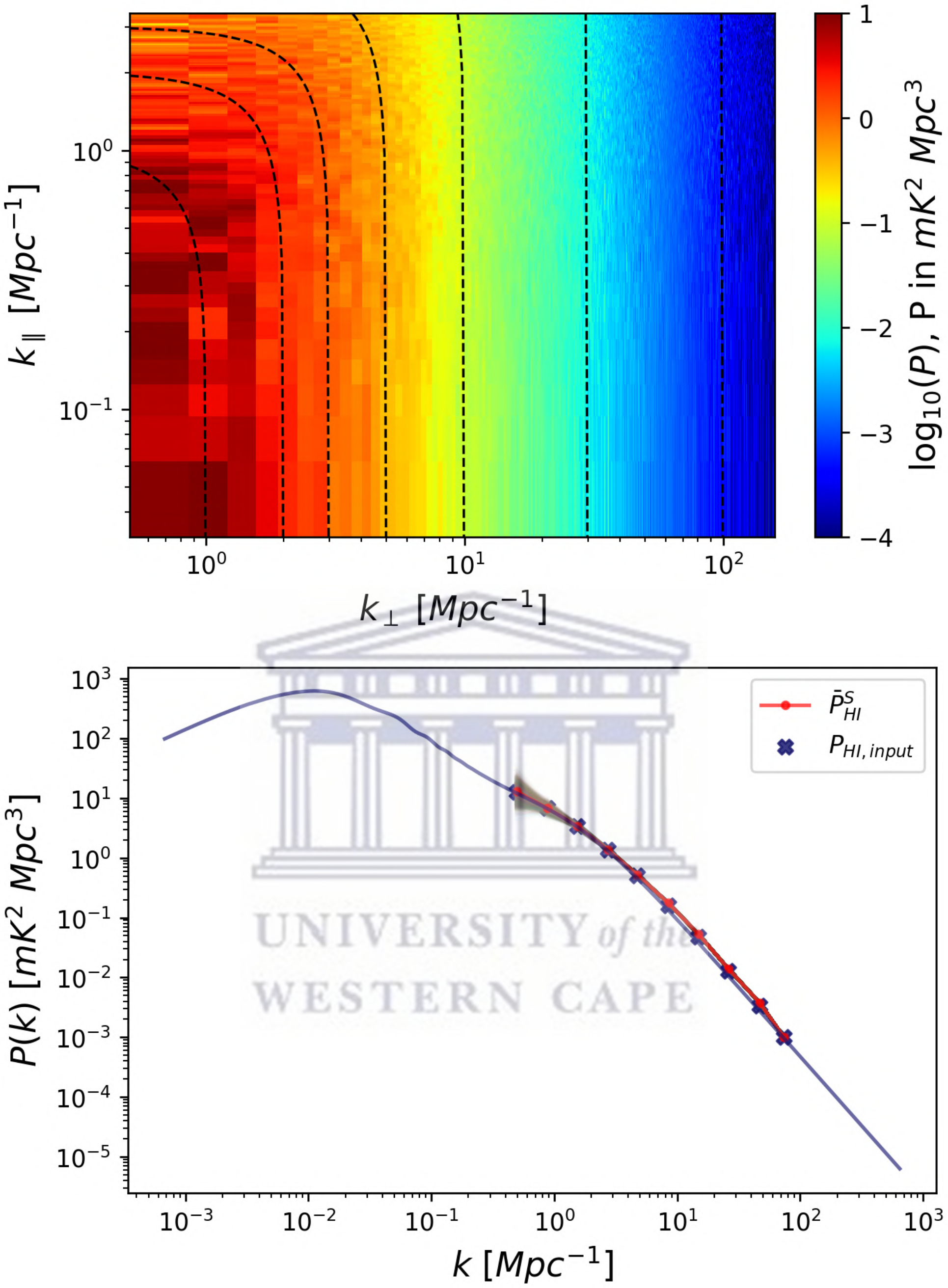

Figure 3.1: 2D (top) and 1D (bottom) HI signal power spectrum outputted from the simulation pipeline for one pointing of the COSMOS field with an area of about 1 deg^2. In the 2D power spectrum, the spherical symmetry of the HI signal is clearly visible through the constant contours of power along k. In the 1D case, the input HI power spectrum is shown as both a curve and points at the k values sampled by the simulation pipeline along with 1000 realisations of the HI power spectrum outputted from the simulation pipeline. Also shown is the average over the 1000 realisations at each of the sampled k modes.

particular, the average over the realisations returns the input model quite well, with some deviations observed at low k due to cosmic variance. This comes from the fact that the visibilities generated in the pipeline to calculate the power spectrum are sampled from a Gaussian distribution with the number of points in the average at low k being small. At higher k, there is also a deviation observed between the input and output power spectra, but which is observed to be consistent with the scatter seen over multiple realisations and is noted to be unbiased on average.

Figure 3.2 shows the thermal noise component of the power spectrum in both 1D and 2D. It should be noted that the thermal noise does not change along frequency, since it is assumed that the system temperature, T_{sys}, remains constant in the simulation. This results in the noise RMS being constant in frequency (the noise itself is a random Gaussian). This particular effect can be seen in the 2D power spectrum of the thermal noise shown in Figure 3.2, where the power would eventually converge to a constant along $k_{\|}$ for each value of $k_{\perp}$ after averaging multiple realisations of the thermal noise power spectrum. In addition to this, the thermal noise increases along $k_{\perp}$ in the 2D power spectrum due to the number of uv points becoming sparser at longer baselines. This is also seen in the 1D power spectrum as it increases for increasing k. Due to this effect, the higher k modes sampled and outputted by the simulation pipeline will be noise dominated. Hence, when extracting an estimate of the HI power spectrum from the simulations, the k modes of interest will be restricted to $k \leqslant 10\text{Mpc}^{-1}$ (refer to the 1D power spectrum plot in Paul et al. (2020) for information on how shot noise also factors into this).

Figure 3.3 essentially shows the foreground power spectrum produced from an input model of radio sources from a MIGHTEE COSMOS field observation of 11.2 hours (total intensity image of this shown in Figure 2.12). In this power spectrum, the wedge is shown prominently, separating the region of interest from the rest of the simulated k space. In this analysis, the foreground component of the simulation pipeline has simply been used to fit a function separating the foreground wedge from the HI and thermal noise dominated regions. This curve is an approximation of the theoretical horizon limit given in Equation 2.10. It is situated well above the wedge to reduce any possible spillover affecting the resulting estimates of the HI power spectrum. While this does significantly reduce the risk of foreground contamination in the estimate of the HI power spectrum extracted from the simulations, it also leads to the loss of some of the lower k modes. This is important to note as the

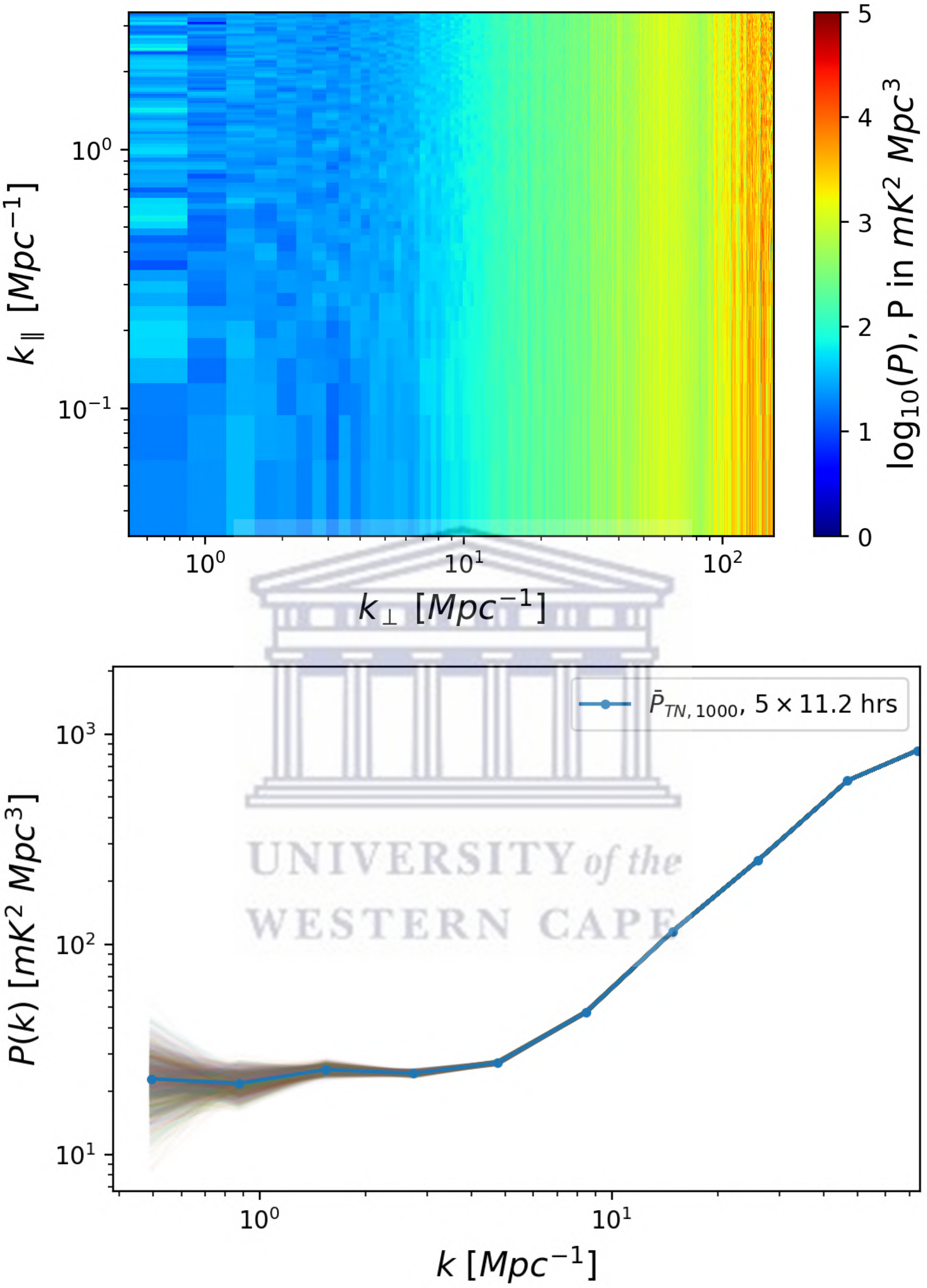

Figure 3.2: 2D (top) and 1D (bottom) thermal noise power spectra outputted from the simulation pipeline. In both the 2D and 1D cases, an increase is the noise power is observed for increasing k, despite the 2D noise power spectrum being constant along $k_\parallel$ due to the assumption that the noise RMS is frequency independent.

lower k modes contain information on larger physical scales, which is then lost in the process of forming the estimator.

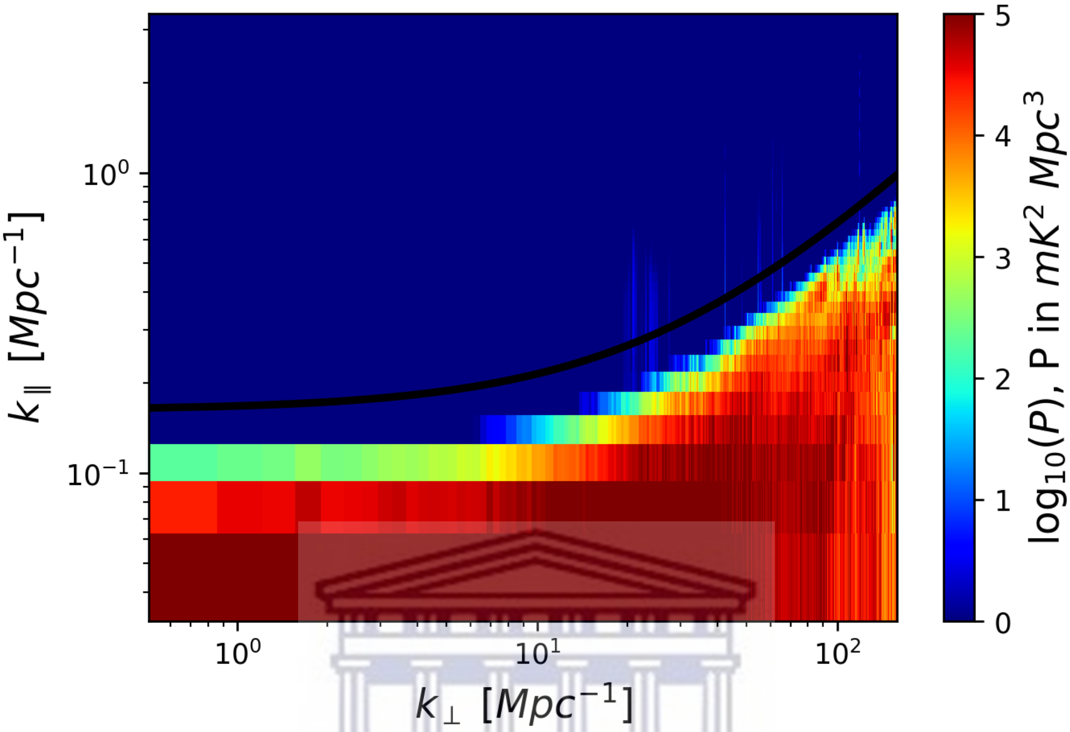

Figure 3.3: The 2D foreground power spectrum outputted by the simulation pipeline with the horizon limit curve fitted to the edge of the foreground wedge. This curve is an approximation to Equation 2.10. In the analysis, the foregrounds were excluded from the simulated visibilities. However, due to foreground contamination in real data sets, the region below the wedge was removed entirely. This would restrict the extracted power spectra to only those k modes in the HI and thermal noise dominated regions of k space. Hence, the methods employed here rely on foreground avoidance instead of foreground cleaning, limiting the number of k modes the analysis technique is sensitive to.

Lastly, shown in Figure 3.4 is the full 2D power spectrum generated by the simulation pipeline, containing the contributions from the foregrounds, thermal noise and HI signal. This power spectrum would be closest to that calculated on real interferometric data, as it contains the contributions from all the key components. While the power is clearly dominated by the foregrounds, there are structures above the wedge that show the presence of the HI signal and thermal noise (albeit very faintly compared to the foregrounds).

Although the foregrounds are not included in the main estimator analysis (due to foreground avoidance), it is useful to understand which components dominate which regions in the 2D power spectrum. From Figure 3.4, it is clear that there is a clear distinction between the region which is foreground dominated and that which is dominated by the thermal noise and signal, providing further motivation for the foreground avoidance technique, despite the loss of some of the lower k modes.

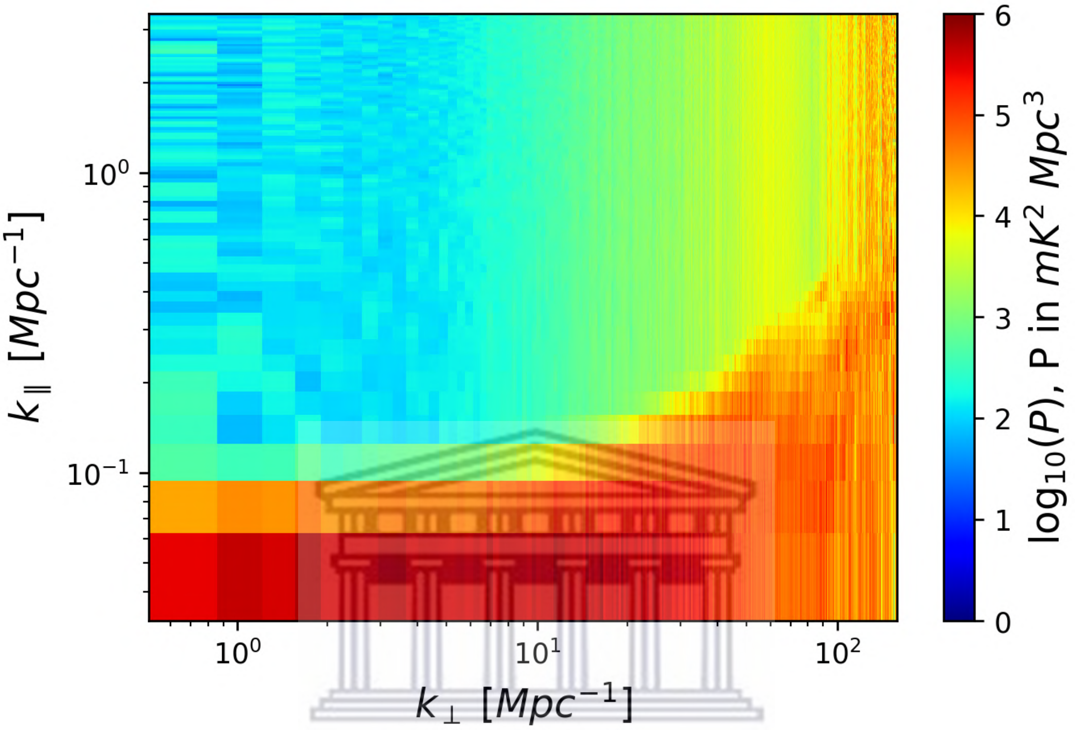

Figure 3.4: The full simulation 2D power spectrum, containing the contributions from the foregrounds, noise and HI signal. While the wedge is dominated by the foregrounds, it is clear that the HI window is noise and signal dominated, leaving a region in k space which allows one to employ foreground avoidance to estimate the power spectrum.

3.2 HI power spectrum estimation

3.2.1 Auto-correlation of visibilities

The estimator analysis was performed with the goal of extracting an estimate of the HI power spectrum, which was used as the input model of the HI signal in the simulation pipeline. In order to do this, the observed power spectrum, $P_{\rm o}(k)$, was assumed to contain information about the signal with contamination from thermal noise. Mathematically,

$$P_{\rm o}(k) = P_{\rm HI+TN}(k). \tag{3.1}$$

To then extract an estimate of the input HI power spectrum, a model thermal noise power spectrum is subtracted from the observed power spectrum. However, due to the random processes involved in generating the thermal noise visibilities, a single realisation of the thermal noise power spectrum would be an inadequate model and would not represent the contamination in the observed power spectrum very well. To compensate for this, an average over N realisations of the thermal noise power spectrum is chosen to represent this model, i.e.,

$$\overline{P_{\rm TN}}(k) = \frac{\sum\limits_{i}^{N} P^{i}_{\rm TN}(k)}{N}, \tag{3.2}$$

where i specifies the $i^{\rm th}$ realisation of the thermal noise power spectrum generated in the pipeline. Therefore, the estimator takes the form,

$$\tilde{P}(k) = P_{\rm o}(k) - \overline{P_{\rm TN}}(k), \tag{3.3}$$

which is over a single realisation of the observed power spectrum, $P_{\rm o}(k)$. To improve the statistics of the estimator, N realisations of the observed power spectrum were generated as well and therefore N estimator values were likewise generated for each k when one subtracts the same thermal noise power model from each observed power spectrum realisation,

$$\tilde{P}^{i}(k) = P^{i}_{\rm o}(k) - \overline{P_{\rm TN}}(k). \tag{3.4}$$

At each k mode sampled by the simulation pipeline, the input HI power spectrum is then estimated by the mean over the N calculated estimator values,

$$\overline{\tilde{P}}(k) = \frac{\sum\limits_{i}^{N} \tilde{P}^{i}(k)}{N}, \tag{3.5}$$

along with the standard deviation of the sample of N estimator values, $\sigma_{\tilde{P}(k)}$, which can be expressed mathematically as

$$\begin{aligned} \sigma_{\tilde{P}(k)} &= \sqrt{\frac{\sum\limits_{i}^{N} \left[\tilde{P}^{i}(k) - \overline{\tilde{P}}(k)\right]^{2}}{N}} \\ &= \sqrt{\frac{\sum\limits_{i}^{N} \left[P_{\mathrm{o}}^{i}(k) - \overline{P}_{\mathrm{TN}}(k) - \overline{\tilde{P}}(k)\right]^{2}}{N}}. \end{aligned} \tag{3.6}$$

The estimation is then compared to the input HI power spectrum to test how well the pipeline is able to recover the HI signal from the simulated observations.

The *uv* distribution used in the pipeline is extracted from a measurement set of fixed observation time. However, it is possible to improve the sensitivity by integrating data from multiple observations coherently. This is advantageous as it results in no loss of signal. To model this, scenarios in which the observation time is increased are considered. In particular, observations on the same field as the existing data and under uniform observational conditions are assumed. In particular, cases of 2 and 5 times the fiducial observation time (11.2 hours) are considered in the estimator analysis.

Using the two above mentioned observation times, the simulation pipeline was run for the auto-correlation of visibilities. Hence, the same simulated visibility set was essentially multiplied with its complex conjugate to obtain the power spectrum as opposed to having two separately generated visibility sets and multiplying one with the complex conjugate of the other as in the case of a cross-correlation.

The noise model was generated by running $N = 1000$ realisations of the simulation pipeline in the case where the visibilities only contain thermal noise. These realisations were then all averaged at each of the sampled k modes in the case of the 1D power spectrum. This mean noise power spectrum as well as the 1000 realisations are shown for both observation time cases in Figure 3.6. Figure 3.5 shows the histogram of the noise power spectrum realisations at each k mode.

As expected, in both Figure 3.6 and Figure 3.5, we see that the noise level for the 5 × 11.2 hours case is consistently lower than that of the 2 × 11.2 hours case. Unfortunately, the observed power spectrum at k modes higher than 10 Mpc^{-1} will still be noise dominated and so is unreliable for the purpose of estimating the HI power spectrum. Due to this, the estimator results are only shown for the sampled k modes lower than 10 Mpc^{-1}.

Figure 3.7 shows the histogram of the estimator values calculated for both observation time cases as well as the value of the input HI power spectrum as a dark vertical line for the first six k modes sampled in the pipeline. Since both the observed power spectrum and thermal noise model were run for 1000 realisations, there are effectively 1000 estimator values at each value of k.

The histogram of the estimator values for both cases lie almost centered on the HI power spectrum value, with the differences between the two distributions being their spread, thus making the estimator unbiased. This difference is due to the variation in the noise levels of each, with the longer observation time expectantly producing a lower error. Further, the variance in the histograms also include cosmic variance, making the difference between the two observation time cases smaller at low k, due to the power spectrum estimates at these low modes being volume-limited. Figure 3.8 shows the mean values of the estimator sample for each $k \leqslant 10$ Mpc^{-1}, as well as the errors on each mean value.

In both cases, the means of the estimator results are shown to coincide quite well with the input HI power spectrum at each given k, with the clear difference being the error on each. This error includes both noise and cosmic variance. While the 5 × 11.2 hours case is more reliable due to the lower error, both cases seem to perform quite well in getting back the input HI power spectrum. Hence, in the presence of noise, one is able to recover the HI power spectrum from an observation of at least 5 × 11.2 hours.

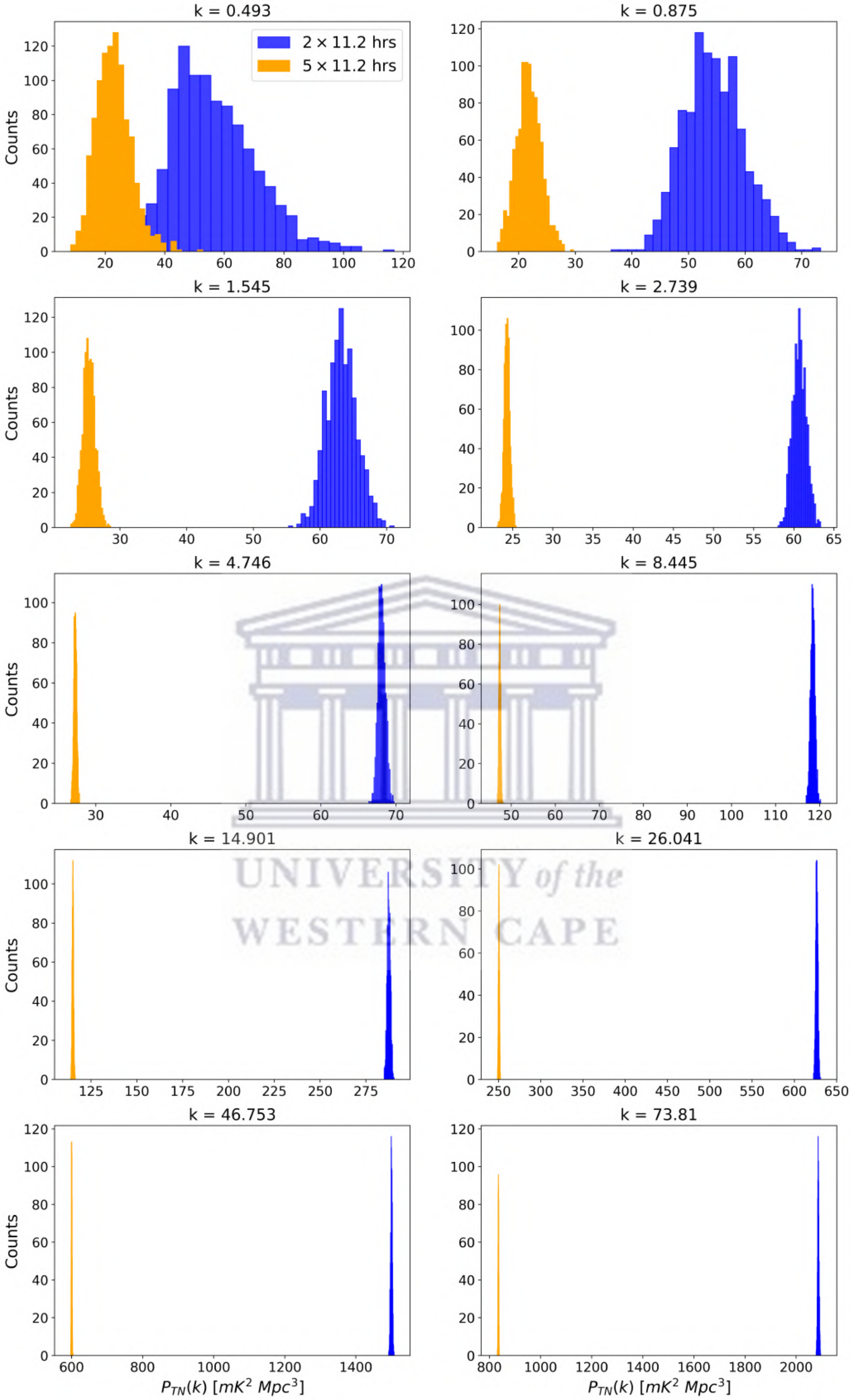

Figure 3.5: Histograms of the noise power spectra realisations for the observation time cases, 2 × 11.2 hours and 5 × 11.2 hours, at each of the sampled k modes. As k increases, the distributions move further away from each other, due to the noise power spectrum being higher for the lower observation time case. This is further demonstrated graphically in Figure 3.6.

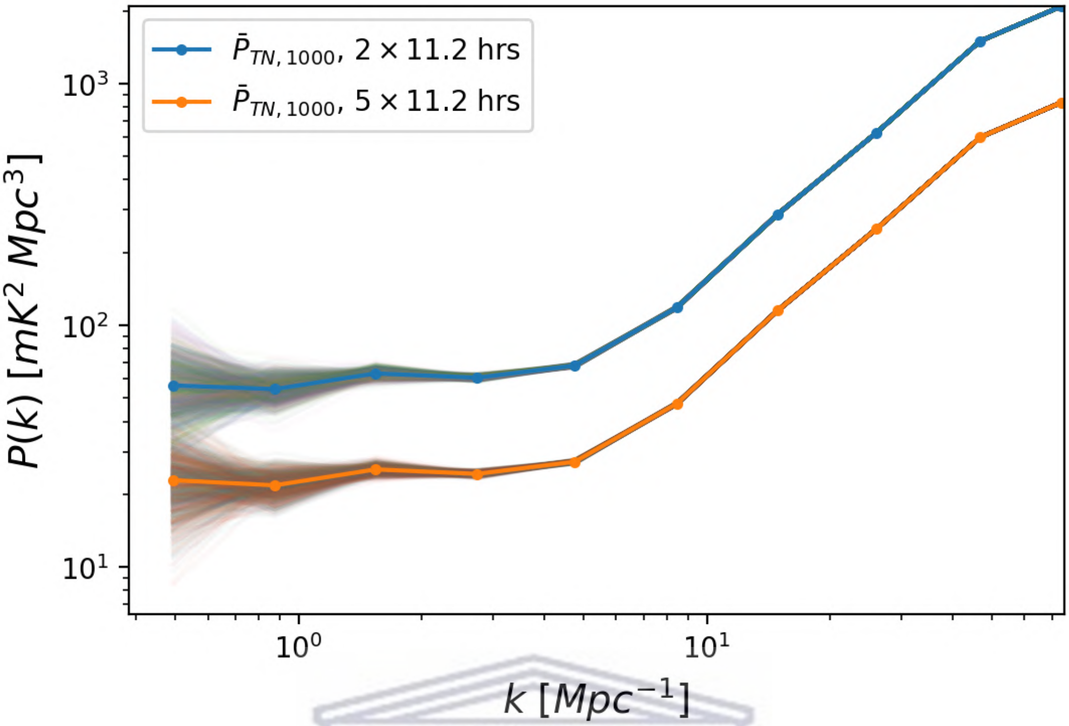

Figure 3.6: Model thermal noise power spectra generated for 2 × 11.2 hours and 5 × 11.2 hours over 1000 realisations each. The plot shows each set of 1000 realisations as well as the mean of the total number of realisations in each case. One thing to note is that with increasing observation time, the noise level drops by a factor of $\frac{1}{\sqrt{N_{\text{multiple}}}}$, where N_{multiple} denotes the constant multiple used to achieve higher observation times from the fiducial 11.2 hours set in the simulation pipeline.

3.2.2 Cross-correlation of visibilities

As an alternative approach to the auto-correlation of visibilities, this work also considers the cross-correlation. This cross-correlation is simply the average of one visibility set (V_{A}), multiplied by the conjugate of another (V_{B}), i.e.,

$$P_{A,B}(\vec{u}) = \Big\langle V_{\text{A}}(\vec{u}_i)V_{\text{B}}(\vec{u}_j)^* \Big\rangle \delta_{ij}, \tag{3.7}$$

in a similar manner shown in Equation 2.14, with the difference being the correlation of two different sets of visibilities, instead of one with itself. Doing so carries some inherent advantages, such as the removal of the noise bias, as well as the possible mitigation of systematics. It is worth considering the cross-correlation for these reasons as these effects are known to hinder the calculation of the power spectrum

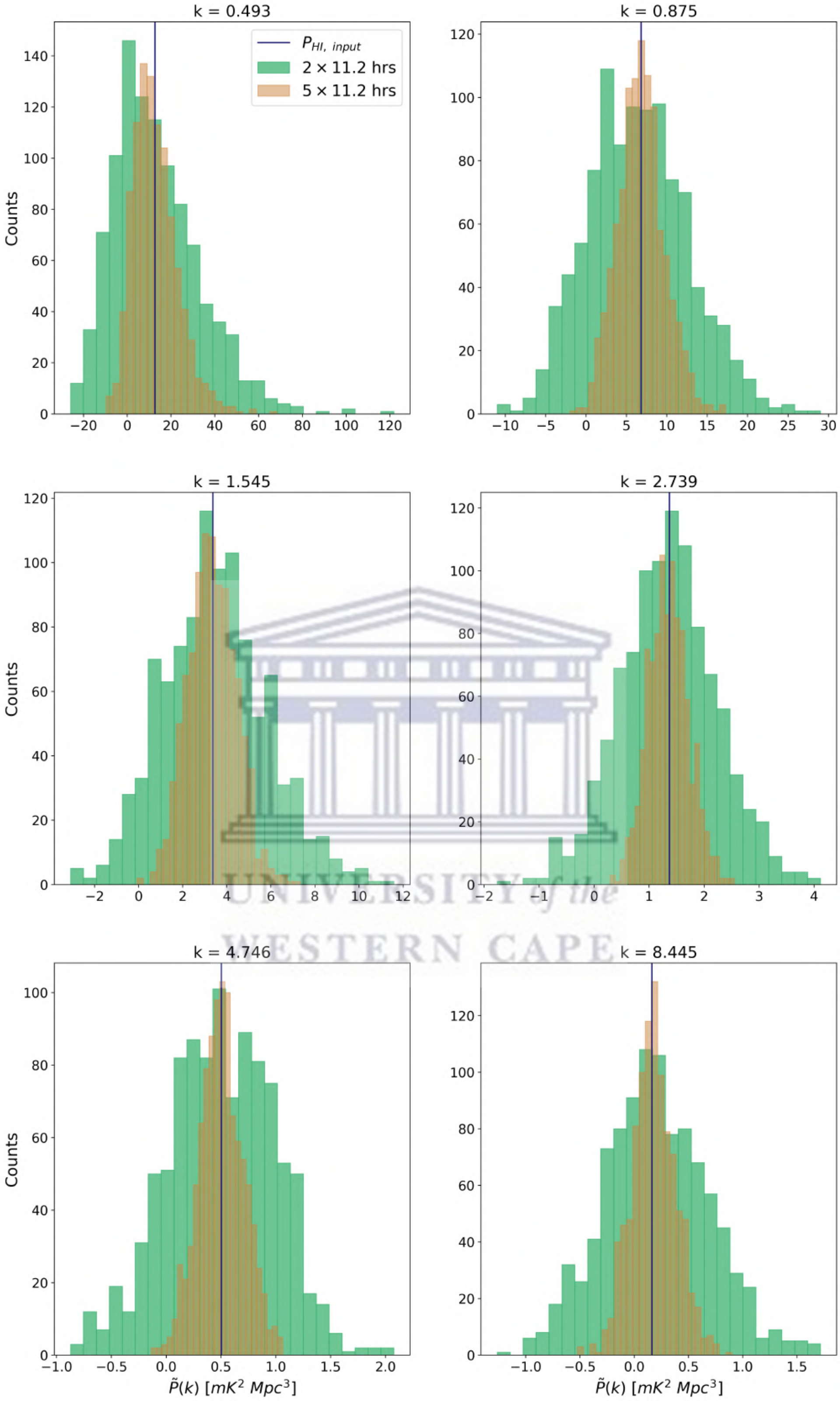

Figure 3.7: Histograms of the estimator values for the auto-correlation case and observation times, 2 × 11.2 hours and 5 × 11.2 hours. Only the histograms for the first six k modes are shown as these were the estimator values of interest. Further, the input HI power spectrum values at each of the six k modes is plotted as a solid vertical line.

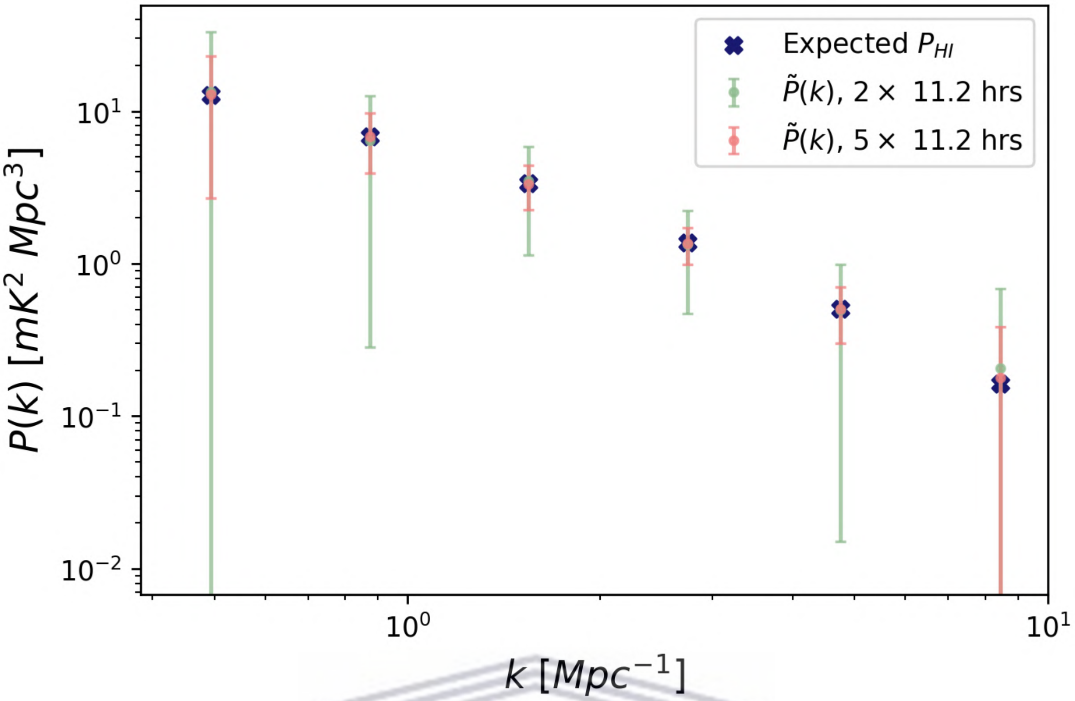

Figure 3.8: The result of taking the mean of the estimator samples for both the 2 × 11.2 hours and 5 × 11.2 hours observation time cases. Also shown is the error on these values, which was taken as the standard deviations of the samples at each k. It should be noted that this error includes both noise and cosmic variance. Further, the input HI power spectrum is plotted as points for comparison with the estimator results.

in auto-correlation.

In the case of cross-correlation, two sets of visibilities were simulated with different realisations of the thermal noise. Hence, the thermal noise visibilities will be uncorrelated and will effectively make the thermal noise power spectrum consistent with zero for all values of k sampled.

Due to this, the simple difference estimator used in the case of auto-correlation will not be needed, as the model thermal noise power spectrum in this case will essentially be zero as shown in Figure 3.9. This figure shows the mean calculated over 1000 realisations of the thermal noise cross power spectrum, which averages out to zero. The small fluctuations present in the realisations are due to the thermal noise visibility components being randomly sampled from a Gaussian distribution. This is further demonstrated in Figure 3.10, which shows the histogram of the thermal noise power spectrum values at each k, where there is a spread around zero.

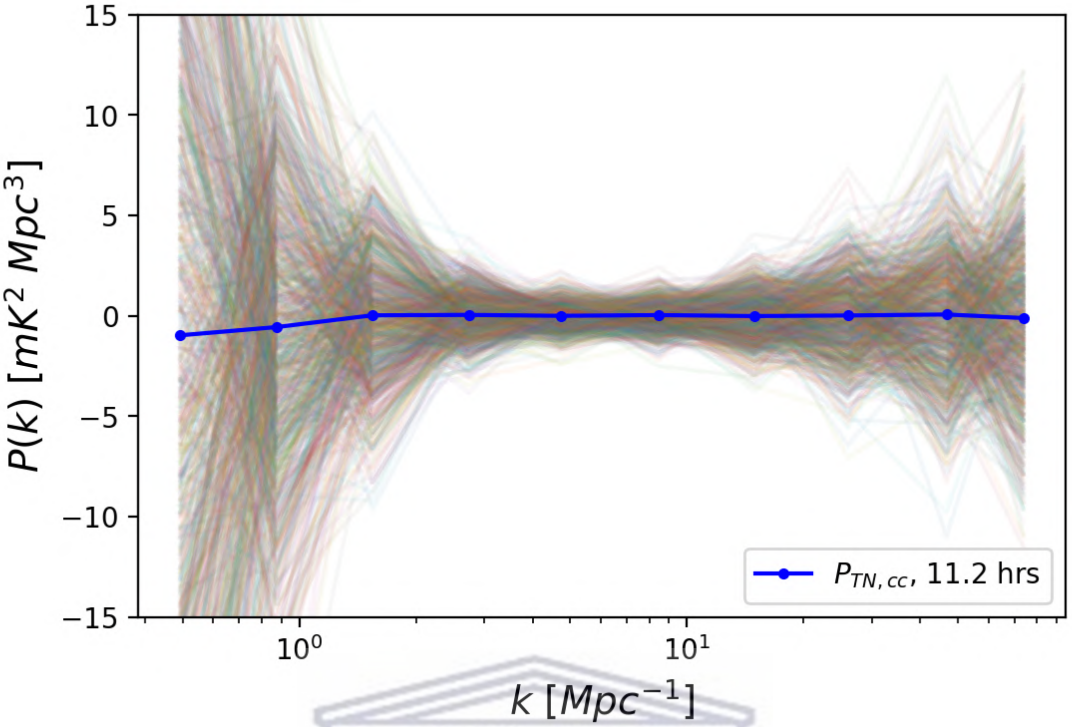

Figure 3.9: The 1D thermal noise cross-power spectrum for the case of cross-correlation. Here, the average over 1000 realisations is shown to be consistent with zero at most k modes probed, with slight oscillations about zero. This result is expected, as the two visibility sets used to generate the noise cross-power spectra, which is then used to obtain the average, are uncorrelated and should converge to zero as the number of realisations increase.

The estimator in this case is equivalent to the observed power spectrum since the thermal noise will be zero, on average, due to there being no noise bias, in the eventual power spectrum after performing a cross-correlation. In this case, a 1000 realisations of the observed power spectrum were run for an observation time of 11.2 hours so as to compare this result with that obtained via auto-correlation.

The two sets of 1000 realisations generated here effectively equals the auto-correlation set of realisations that were generated with an observation time of 2 × 11.2 hours, since this is equivalent to splitting this data set and cross-correlating. Figure 3.12 shows the histograms of the auto-correlation estimator values as well as the cross-correlation observed power spectrum values for $k \leqslant 10$ Mpc^{-1}. Also shown is the input HI power spectrum value at each of the k modes sampled.

While the distributions are quite similar, the cross-correlation case has a wider

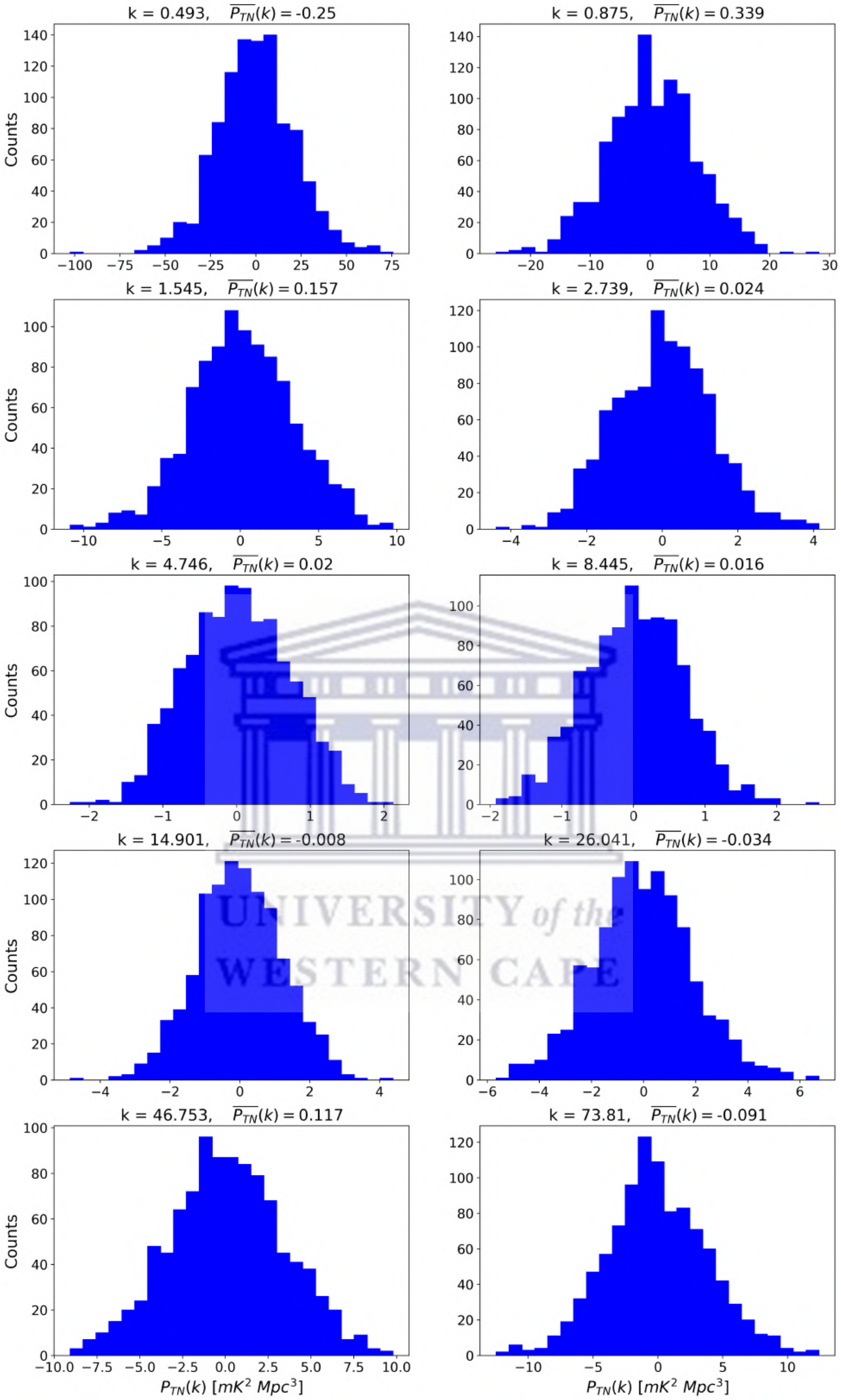

Figure 3.10: Histograms of the thermal noise cross-power spectrum at the sampled k modes for the cross-correlation case. Here, it is clear that each of the histograms are centered on zero, as observed in graphically in Figure 3.9. The spread around zero is due to the random nature of the sampling of noise visibilities, but as shown in Figure 3.9, this is averaged to zero over the 1000 realisation sample.

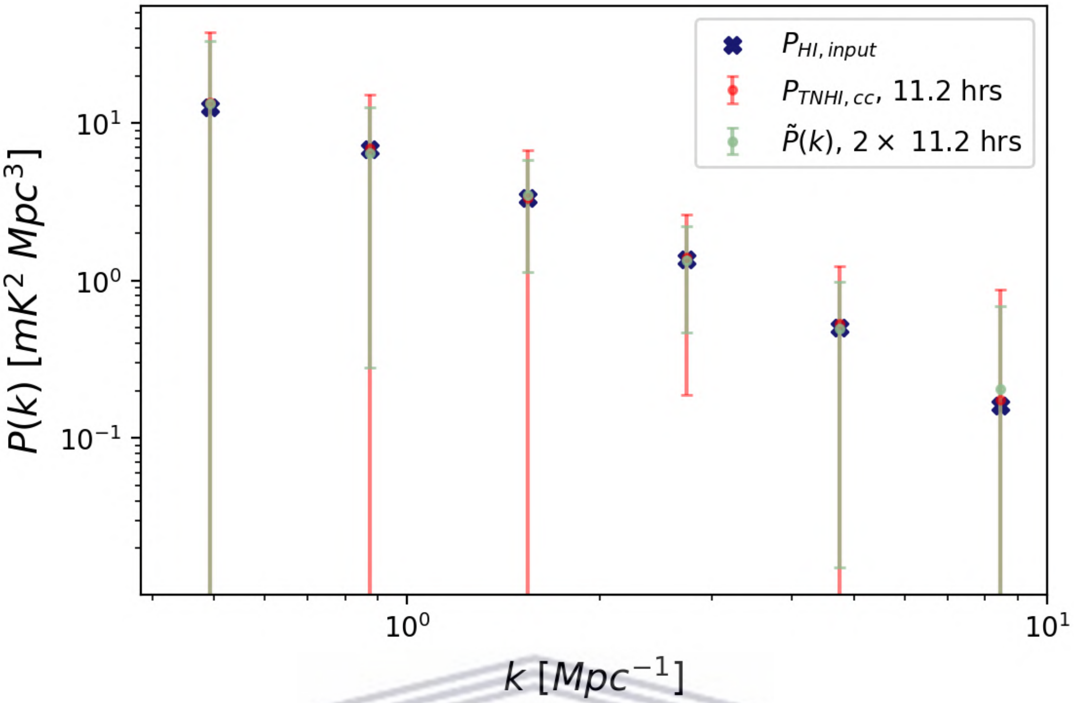

Figure 3.11: The estimator of the power spectrum in the case of auto-correlation and cross-correlation for observation time, 2 × 11.2 hours. Also shown is the input HI power spectrum for comparison with these estimator results. In the case of cross-correlation, the thermal noise is assumed to be consistent with zero (on average) and therefore does not contribute to the observed power spectrum. Hence, the estimator in the case of cross-correlation is simply the observed power spectrum generated by the simulation pipeline, which is then averaged and plotted here with its error at each $k \leqslant 10$ Mpc^{-1}.

spread than that of the auto-correlation case. This is observed in the error on both estimators in Figure 3.11, which demonstrates that for the same observation time, the case of one visibility set (or observation) performs better in constraining the power spectrum than the case of two visibility sets (observations), each observed at half the full observation time, 2 × 11.2 hours. This would similarly be true for the observation time case of 5 × 11.2 hours, as the only marked difference between the two is the thermal noise level in the case of auto-correlation. This is simply a result of having half the observation time in the noise for the cross-correlation case.

In summary, table 3.1 shows the average thermal noise power spectrum values alongside the corresponding k value at which it was calculated from the 1000 realisations generated by the simulation pipeline in the auto-correlation case. The thermal noise values for the cross-correlation case are not shown as they are consistent with zero

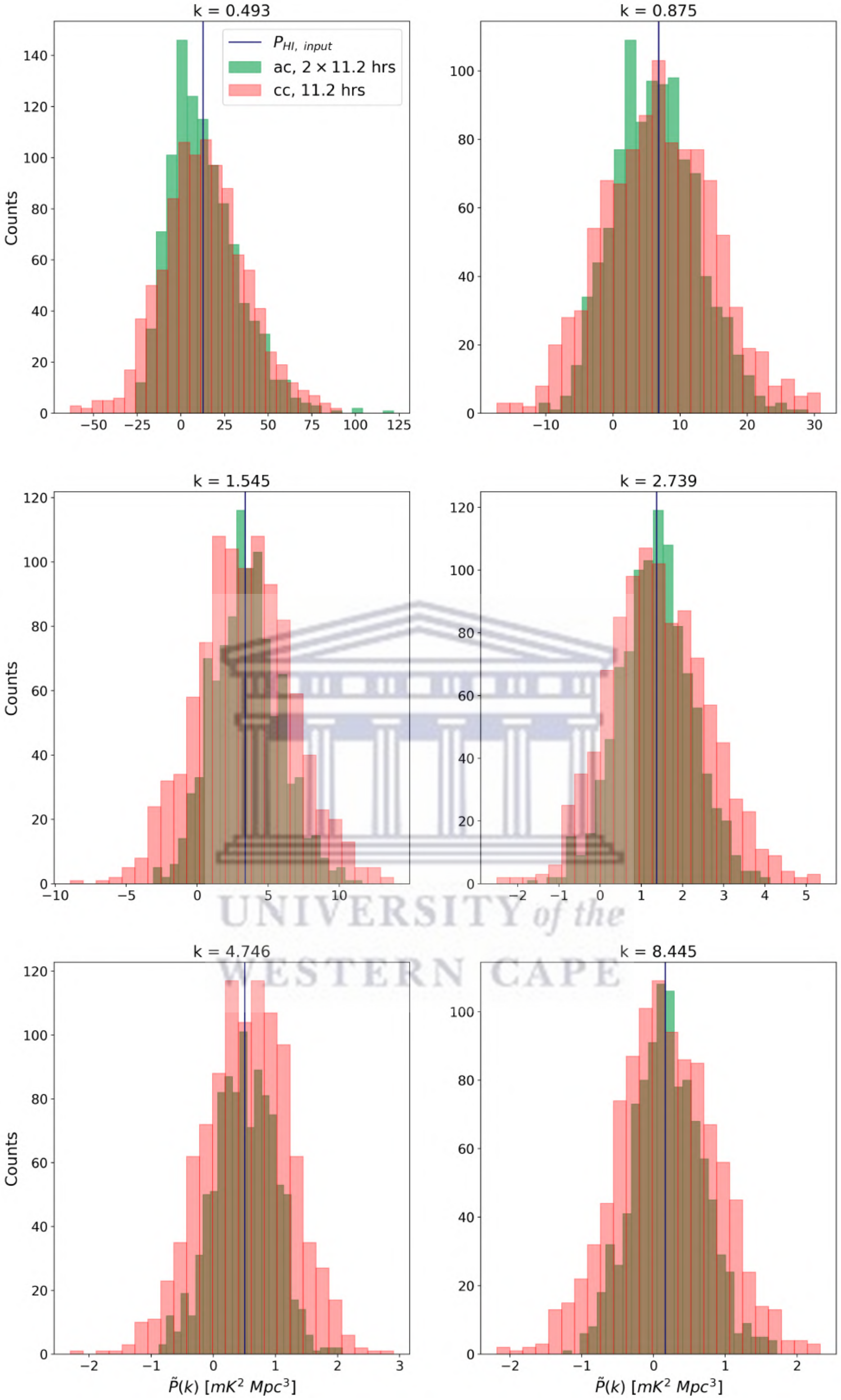

Figure 3.12: Comparison of the histograms of the estimator values from the cross- and auto-correlation for the first six k modes sampled in the simulation pipeline. The vertical straight line in each histogram plot shows the input HI power spectrum at the specific value of k. While the histograms of both cases lie centered on the input HI power spectrum value, the spread in the cross-correlation case is wider, analogous to the error observed in Figure 3.11.

k [Mpc^{-1}]	$\overline{P_{\mathrm{TN}}}(k)$ [mK^2 Mpc^3]	
	2 × 11.2 hours	5 × 11.2 hours
0.493	56.384	22.861
0.875	54.436	21.771
1.545	63.213	25.338
2.739	60.693	24.281
4.746	68.101	27.247
8.445	118.442	47.384
14.901	287.474	114.999
26.041	626.453	250.551
46.753	1499.872	599.943
73.810	2086.446	834.653

Table 3.1: Table showing average thermal noise power spectrum values at each k mode sampled by the simulation pipeline for both auto-correlation cases considered. The average thermal noise power spectrum values for the cross-correlation case are not shown. This is done, since, on average, the thermal noise cross-power spectrum is consistent with zero at each value of k.

k [Mpc^{-1}]	$P^i_{\mathrm{HI}}(k)$ [mK^2 Mpc^3]	$\overline{\tilde{P}}(k)$ [mK^2 Mpc^3]		
		Auto-correlation		Cross-correlation
		2 × 11.2 hours	5 × 11.2 hours	11.2 hours
0.493	12.67	13.283 ± 19.857	12.827 ± 10.144	13.588 ± 24.209
0.875	6.8	6.445 ± 6.163	6.803 ± 2.879	6.925 ± 8.275
1.545	3.37	3.486 ± 2.351	3.339 ± 1.088	3.357 ± 3.364
2.739	1.37	1.348 ± 0.878	1.351 ± 0.366	1.401 ± 1.213
4.746	0.503	0.499 ± 0.484	0.503 ± 0.202	0.535 ± 0.703
8.445	0.162	0.205 ± 0.484	0.178 ± 0.204	0.175 ± 0.699

Table 3.2: Table summarising the estimator mean and error results for all auto- and cross-correlation cases considered. Also shown are the relevant k modes as well as the input model HI power spectrum value at each k.

to some degree. Although the average over 1000 realisations did not result in all of the values being zero, an increase in realisations would show this to be the case. As discussed earlier, the thermal noise power spectrum increases rapidly after $k =$ 10 Mpc^{-1}, which is shown to be the case for both observation times. After $k =$ 8.445 Mpc^{-1}, the power more than doubles for both observation time cases and then increases rapidly for higher k. This is the reason the k modes of interest for the estimator results are restricted to $k \leqslant 10$ Mpc^{-1} for both the auto- and cross-correlation cases.

Table 3.2 shows the results from the estimator analysis for both the auto-correlation and cross-correlation cases. Alongside the mean values are the error on each. Also shown are the relevant k modes as well as the input model HI power spectrum at each value of k for comparison with the estimator results. Clearly, the auto-correlation for an observation time of 5 × 11.2 hours is the closest to the input HI power spectrum. By comparison, the 2 × 11.2 hours case in auto-correlation has an error on each estimator value that is almost consistently double that found for the 5 × 11.2 hours case. Despite this, the auto-correlation case for 2 × 11.2 hours does perform better than its counterpart in cross-correlation.

4 Conclusions

In this thesis, the potential use of the MIGHTEE survey for the purpose of HI intensity mapping was explored. With the interferometric data, cosmological information could be extracted to constrain parameters at small (non-linear) scales due to the dense core of the MeerKAT interferometer, providing a complementary approach to the single-dish experiments which generally probe larger scales such as those relevant for BAO studies. As such, the objective was to investigate what can be done with MIGHTEE in terms of cosmology with intensity mapping and the novel delay-spectrum techniques commonly employed in interferometric data analysis cases for the EoR (such as in the case of Parsons et al., 2012a; Thyagarajan et al., 2013).

In order to do this, a purpose-built, visibility-based, simulation pipeline which not only mimics the MIGHTEE observations, but also calculates the observed power spectrum, was analysed. Additionally, to test how well the simulation outputs were able to mimic real interferometric data sets, cosmological information was extracted from the simulation outputs using a power spectrum estimator for various test cases relevant to the MIGHTEE survey.

In the presence of noise contaminants, the test was to see how well the HI power spectrum could be constrained. In order to do this, two test observation time cases were considered: $t_{\text{obs}} = \{2 \times 11.2, 5 \times 11.2\}$ hours. These cases were from a standard data set from an actual MIGHTEE COSMOS observation with total observed time of 11.2 hours, which was artificially increased in the pipeline. Further, foreground contamination was taken into consideration by utilising the foreground avoidance technique to exclude the k modes in the regions of Fourier space that were foreground dominated. The effects of RFI were also taken into consideration by setting a criterion in which only those baselines which have 80% of their frequency channels unflagged are used for the eventual power spectrum estimation.

For each observation time case, 1000 realisations of the power spectrum from observations containing thermal noise and the HI signal, in addition to 1000 realisations of the thermal noise for the model noise power spectrum, were generated. These realisations were then used to produce final power spectrum estimates as well as their associated errors.

The results for the power spectrum estimates are promising. At $z = 0.27$, and both observation time cases, the estimators are able to recover the input HI power spectrum quite well in auto-correlations of visibilities over the range $0.493 \leqslant k \leqslant 8.445$ Mpc^{-1}. The same was seen for the case of cross-correlation of different sets of visibilities, despite the auto-correlation cases outperforming it, even for comparable observation times. Overall, the results seem to favour a deep integration over a single field and using this data to estimate the HI power spectrum. Additionally, the fact that MIGHTEE will observe multiple fields over the course of the full survey provides a means of reducing the cosmic variance (Paul et al., 2020). Both these insights provide a hopeful picture for the prospect of using MIGHTEE data to study the universe at low redshifts via the HI signal, as the full survey will be over four well studied extra-galactic fields over a total observation time of around 1000 hours.

Despite these positive results, there are limitations which need to be accounted for. Firstly, due to the relative simplicity of the power spectra generated in the simulation pipeline, the results obtained are for the best scenario in which the noise is the primary contaminant, since foreground avoidance excludes the modes contaminated by said foregrounds. In this case, it is a relatively simple matter to model and account for it. In reality, there are various other contaminants which need to be taken into account, such as systematics. In future, an aim would be to factor in systematic effects which would enable their effects on the eventual power spectrum estimates to be studied in more detail.

Further improvements would include modelling of the primary beam, which has largely been ignored in this study. This could be added via a more complex foreground model which includes instrumental effects such as the primary beam in order to study how this affects the foreground wedge. Despite the analysis relying on foreground avoidance, it remains crucial to accurately model and account for foregrounds when trying to make precision measurements of the HI signal, especially if one wants to recover modes usually lost in this process (Chapman et al., 2016). In that case, another possible extension to the analysis could include comparing the method of foreground avoidance to foreground cleaning.

Moreover, the use of sophisticated simulations (such as hydrodynamical or N-body simulations) could also potentially be used to better model the input HI signal and thus allow a more accurate study of the power spectrum to be performed. This is especially important on the non-linear scales which the MIGHTEE survey probes due to the need to accurately model the structures at these scales. Interesting studies in which this was done can be found in Villaescusa-Navarro et al. (2018) and Spinelli et al. (2020), where quantities important in 21 cm intensity mapping such as the HI mass function and density were investigated using sophisticated simulations. An interesting expansion on this would be to test how accurately one would be able to measure several parameters, such as the standard ΛCDM cosmological parameters, as well as the HI mass function parameters. In terms of the pipeline and results obtained at present, possible extensions would include running test cases with higher observation times (such as the full MIGHTEE observation time of 1000 hours) and looking at the power spectrum estimates from this.

Overall, as the MIGHTEE survey continues making observations, continuous improvements to the simulations and the integration of more sophisticated models of all phenomena involved in the observations and power spectrum estimation work will be undertaken. All the insights attained through this process would aid in the goal of eventually making an actual measurement of the HI signal through interferometric HI intensity mapping.

Bibliography

Alonso D., Bull P., Ferreira P. G., Santos M. G., 2015, Monthly Notices of the Royal Astronomical Society, 447, 400

Anderson C. J., et al., 2018, Monthly Notices of the Royal Astronomical Society, 476, 3382

Bacon D., et al., 2018, Publications of the Astronomical Society of Australia, 37

Bailes M., et al., 2016, in MeerKAT Science: On the Pathway to the SKA. p. 11 (arXiv:1803.07424)

Baker A. J., Blyth S., Holwerda B. W., LADUMA Team 2018, in American Astronomical Society Meeting Abstracts #231. p. 231.07

Battye R. A., Davies R. D., Weller J., 2004, Monthly Notices of the Royal Astronomical Society, 355, 1339

Battye R. A., Browne I. W., Dickinson C., Heron G., Maffei B., Pourtsidou A., 2013, Monthly Notices of the Royal Astronomical Society, 434, 1239

Bharadwaj S., Sethi S. K., 2001, Journal of Astrophysics and Astronomy, 22, 293

Bharadwaj S., Nath B. B., Sethi S. K., 2001, Journal of Astrophysics and Astronomy, 22, 21

Bobin J., Starck J.-L., Moudden Y., Fadili M., 2008, Advances in Imaging and Electron Physics, 152

Booth R., de Blok W., Jonas J., Fanaroff B., 2009, arXiv e-prints, p. arXiv:0910.2935

Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018, Nature

Bracewell R. N., 1965, The Fourier Transform and Its Applications. New York: McGraw-Hill Book Company, New York

Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, Astrophysical Journal, 803, 1

Camera S., Padmanabhan H., 2020, Monthly Notices of the Royal Astronomical Society, 13, 1

Castorina E., Villaescusa-Navarro F., 2017, Monthly Notices of the Royal Astronomical Society, 471, 1788

Chang T. C., Pen U. L., Peterson J. B., McDonald P., 2008, Physical Review Letters, 100, 1

Chang T., Pen U., et al. Bandura K., 2010, Nature, pp 463–465

Chapman E., et al., 2012, Monthly Notices of the Royal Astronomical Society, 423

Chapman E., et al., 2013, Monthly Notices of the Royal Astronomical Society, 429, 165

Chapman E., Zaroubi S., Abdalla F. B., Dulwich F., Jelić V., Mort B., 2016, Monthly Notices of the Royal Astronomical Society, 458, 2928

Cunnington S., Wolz L., Pourtsidou A., Bacon D., 2019, arXiv e-prints, 23, 1

Datta A., Bowman J. D., Carilli C. L., 2010, Astrophysical Journal, 724, 526

Eastwood M., et al., 2019, The Astronomical Journal, 158, 84

Eisenstein D. J., et al., 2005, The Astrophysical Journal, 633, 560

Ewen H., Purcell E., 1951, Nature, 168

Fender R., et al., 2016, Proceedings of Science

Furlanetto S. R., Peng Oh S., Briggs F. H., 2006, Physics Reports, 433, 181

Gehlot B. K., et al., 2019, Monthly Notices of the Royal Astronomical Society, 488, 4271

Gupta N., et al., 2016, in MeerKAT Science: On the Pathway to the SKA. p. 14 (arXiv:1708.07371)

Harris F. J., 1978, Proceedings of the IEEE, 66, 51

Hogg D. W., 1999, arXiv e-prints, pp astro–ph/9905116

Ishwara-Chandra C. H., Sirothia S. K., Wadadekar Y., Pal S., Windhorst R., 2010, Monthly Notices of the Royal Astronomical Society, 405, 436

Jacobs D., et al., 2014, The Astrophysical Journal, 801

Jarvis M. J., et al., 2016, Proceedings of Science, pp 25–27

Jonas J. L., 2016, Proceedings of Science, pp 25–27

Kovetz E. D., et al., 2017, arXiv e-prints

Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473

Liu A., Shaw J. R., 2020, Publications of the Astronomical Society of the Pacific, 132

Liu A., Pritchard J. R., Tegmark M., Loeb A., 2012, | 10.1103/PhysRevD.87.043002

Masui K. W., et al., 2013, Astrophysical Journal Letters, 763, 1

McQuinn M., Zahn O., Zaldarriaga M., Hernquist L., Furlanetto S. R., 2006, The Astrophysical Journal, 653, 815

Morales M. F., 2005, The Astrophysical Journal, 619, 678

Morales M. F., Hewitt J., 2004, The Astrophysical Journal, 615, 7

Morales M. F., Hazelton B., Sullivan I., Beardsley A., 2012, Astrophysical Journal, 752

Morales M. F., Beardsley A., Pober J., Barry N., Hazelton B., Jacobs D., Sullivan I., 2019, Monthly Notices of the Royal Astronomical Society, 483, 2207

Newburgh L., et al., 2014, Proceedings of SPIE - The International Society for Optical Engineering, 9145

Newburgh L. B., et al., 2016, Ground-based and Airborne Telescopes VI, 9906, 99065X

Padmanabhan H., Choudhury T. R., Refregier A., 2015, Monthly Notices of the Royal Astronomical Society, 447, 3745

Padmanabhan H., Refregier A., Amara A., 2017, Monthly Notices of the Royal Astronomical Society, 469, 2323

Parsons A. R., et al., 2010, Astronomical Journal, 139, 1468

Parsons A., Pober J., McQuinn M., Jacobs D., Aguirre J., 2012a, Astrophysical Journal, 753

Parsons A. R., Pober J. C., Aguirre J. E., Carilli C. L., Jacobs D. C., Moore D. F., 2012b, Astrophysical Journal, 756

Paul S., et al., 2016, Astrophysical Journal, 833, 213

Paul S., Santos M. G., Townsend J., Jarvis M. J., Maddox N., Collier J. D., Frank B. S., Taylor R., 2020, arXiv e-prints

Peebles P. J. E., 1980, The large-scale structure of the universe. Princeton University Press Princeton, N.J

Planck Collaboration et al., 2018, Planck 2018 results. VI. Cosmological parameters (arXiv:1807.06209)

Pober J. C., et al., 2013, Astrophysical Journal Letters, 768

Pourtsidou A., 2016, Proceedings of Science, pp 25–27

Prestage R., Constantikes K., Hunter T., King L., Lacasse R., Lockman F., Norrod R., 2009, Proceedings of the IEEE, 97, 1382

Pritchard J. R., Loeb A., 2012, Rept. Prog. Phys., 75, 86901

Sanidas S., Caleb M., Driessen L., Morello V., Rajwade K., Stappers B., 2018, in Weltevrede P., Perera B., Preston L., Sanidas S., eds, Vol. 337, Pulsar Astrophysics the Next Fifty Years. pp 406–407, doi:10.1017/S1743921317009310

Santos M. G., Cooray A., Knox L., 2005, The Astrophysical Journal, 625, 575

Santos M. G., et al., 2015, in Proceedings of Science. pp 1–27 (arXiv:1501.03989)

Santos M. G., et al., 2017, arXiv e-prints, pp 1–22

Sarkar D., Bharadwaj S., 2018, Monthly Notices of the Royal Astronomical Society, 476, 96

Sarkar D., Bharadwaj S., 2019, Monthly Notices of the Royal Astronomical Society, 487, 5666

Schinnerer E., et al., 2004, The Astronomical Journal, 128, 1974

Schinnerer E., et al., 2007, The Astrophysical Journal Supplement Series, 172, 46

Serra P., et al., 2016, in MeerKAT Science: On the Pathway to the SKA. p. 8 (arXiv:1709.01289)

Smolcic V., et al., 2017, Astronomische Nachrichten, 602, A1

Spinelli M., Zoldan A., De Lucia G., Xie L., Viel M., 2020, Monthly Notices of the Royal Astronomical Society, 493, 5434

Staveley-Smith L., et al., 1996, Publications of the Astronomical Society of Australia, 13, 243

Switzer E., et al., 2013, Monthly Notices of the Royal Astronomical Society, 434

Thompson A. R., Moran J. M., Swenson Jr. G. W., 2017, Interferometry and Synthesis in Radio Astronomy, third edn. Springer International Publishing, doi:10.1007/978-3-319-44431-4

Thyagarajan N., et al., 2013, Astrophysical Journal, 776

Thyagarajan N., et al., 2015, Astrophysical Journal, 804

Trott C. M., et al., 2020, Monthly Notices of the Royal Astronomical Society, 493, 4711

Vedantham H., Udaya Shankar N., Subrahmanyan R., 2012, Astrophysical Journal, 745

Villaescusa-Navarro F., et al., 2018, The Astrophysical Journal, 866, 135

Wilson T., Rohlfs K., Huttemeister S., 2012, Tools of Radio Astronomy, fifth edn. Springer International Publishing, doi:10.1007/978-3-540-85122-6

Wyithe J. S. B., Loeb A., 2009, Monthly Notices of the Royal Astronomical Society, 397, 1926

Xu Y., Wang X., Chen X., 2015, Astrophysical Journal, 798

York D. G., et al., 2000, The Astronomical Journal, 120, 1579

de Blok W. J., et al., 2016, Proceedings of Science

van de Hulst H., 1945, Ned. Tijdschr. Natuurk., pp 210–221

Appendix A: Fourier Transforms

The Fourier transform is a mathematical transform which decomposes a function into its constituent frequencies. Formally, it can be viewed as a mapping from the real numbers to the complex numbers, $f : \mathbb{R} \to \mathbb{C}$. It can be defined as

$$F(s) = \int_{-\infty}^{\infty} f(x) e^{-j2\pi sx} dx, \tag{A1}$$

with its inverse given by

$$f(x) = \int_{-\infty}^{\infty} F(s) e^{j2\pi xs} ds. \tag{A2}$$

$f(x)$ and $F(s)$ then form a Fourier pair and can be written symbolically as

$$f(x) \leftrightarrow F(s). \tag{A3}$$

The pair of variables x and s are known as the conjugates of each other. For example, if x represents frequency in Hz, then its conjugate, s, would represent its inverse, i.e. s would then denote time in seconds. An important result in Fourier theory is the Convolution Theorem, given mathematically as

$$f(y) \otimes g(y) \leftrightarrow F(s) \times G(s), \tag{A4}$$

which essentially says that the convolution of two functions is equivalent to the product of their Fourier transforms. Further, the convolution of the two functions, $f(x)$ and $g(x)$, is defined as

$$h(y) = \int_{-\infty}^{\infty} f(x) g(y - x) dx = f(y) \otimes g(y). \tag{A5}$$

These results essentially sum up the most important properties of the Fourier transform used in this thesis. For a more thorough discussion of Fourier transforms, refer to the appendices in Wilson et al. (2012); Thompson et al. (2017), or the general text of Bracewell (1965).

Appendix B: Foreground simulation processes

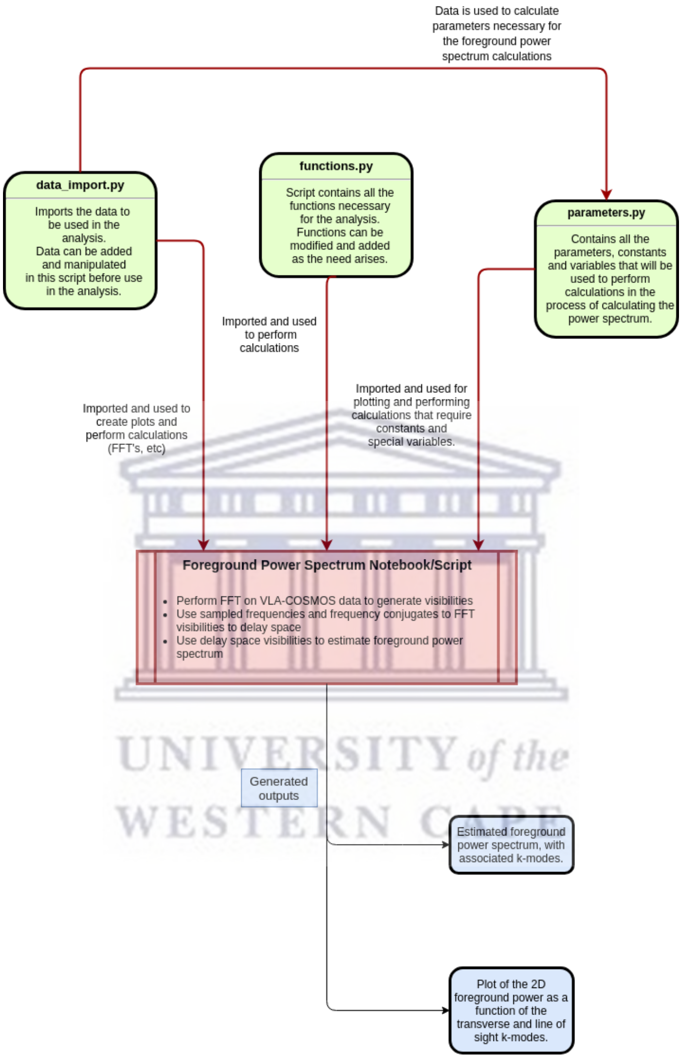
Data is used to calculate parameters necessary for the foreground power spectrum calculations
data_import.py
Imports the data to be used in the analysis. Data can be added and manipulated in this script before use in the analysis.
functions.py
Script contains all the functions necessary for the analysis. Functions can be modified and added as the need arises.
parameters.py
Contains all the parameters, constants and variables that will be used to perform calculations in the process of calculating the power spectrum.
Imported and used to perform calculations
Imported and used to create plots and perform calculations (FFT's, etc)
Imported and used for plotting and performing calculations that require constants and special variables.
Foreground Power Spectrum Notebook/Script
Perform FFT on VLA-COSMOS data to generate visibilities
Use sampled frequencies and frequency conjugates to FFT visibilities to delay space
Use delay space visibilities to estimate foreground power spectrum
Generated outputs
Estimated foreground power spectrum, with associated k-modes.
Plot of the 2D foreground power as a function of the transverse and line of sight k-modes.

Figure B1: Flow chart showing the code scripts and notebooks in which the processes involved in generating the foreground power spectrum of the VLA COSMOS point source catalogue were performed. Scripts for data importing, the functions used as well as the various cosmological and observational parameters are called in a notebook to run the processes such as FFT's and output the 2D foreground power spectrum with the relevant $k_\perp$ and $k_\parallel$ modes.

Figure B1 shows a flow diagram summarising the processes involved in the calculation of the 2D foreground power spectrum of the VLA COSMOS data set, as described in Section 2.3. It gives a simple outline of the processes which one would need to follow in order to go from generating visibilities on a sky model to calculating the power spectrum and the associated k modes from these visibilities. It includes the scripts used for each task that forms part of the calculation process (the importing of the uv distribution and point source catalogue, the primary beam and other relevant functions, and the setting of parameters such as the cosmological parameters, bandwidth, number of frequency channels, etc.) as well as the notebook in which the final calculation of the power spectrum was performed.