Synergies Between 21cm & Optical Surveys for Probing Large Scale Cosmic Structure

PhD Thesis

By

Steven D. CUNNINGTON

Institute of Cosmology & Gravitation
UNIVERSITY OF PORTSMOUTH

The thesis is submitted in partial fulfilment of the requirements
for the award of the degree of Doctor of Cosmology of the
University of Portsmouth.

Submission: SEPTEMBER 2019

Supervisors:
Prof. David Bacon
Dr. Alkistis Pourtsidou

ABSTRACT

We are currently living through an era of precision cosmology where we have gathered a substantial amount of data with the aim of understanding our Universe. However, our current understanding is far from complete as our most successful cosmological model relies on the Universe's energy-matter content being vastly dominated by components that are not yet detected and not currently compatible with our wider general model of physics. This leaves plenty more investigation to be done and new techniques for probing our Universe are highly sought after.

Mapping unresolved neutral hydrogen within galaxies is one of these novel techniques and has been gaining momentum over the last decade. By using the 21cm signal from neutral hydrogen, which traces the underlying large scale cosmic structure, we can map and statistically analyse 3D density distributions and compare these to theoretical models. I provide a detailed introduction to this novel HI intensity mapping technique in Chapter 2.

This thesis also explores what gains can be made by combining HI intensity mapping data with more conventional optical galaxy redshift surveys. There are many reasons why a cross-correlation such as this will be beneficial. While the intensity mapping technique is developed and refined, optical data can boost the inherently weak HI signal allowing detection and a deeper understanding of the intensity mapping process. Also in the future, when we have dedicated intensity mapping instruments gathering data, cross-correlations will see reductions in the different systematics which could otherwise dominate the uncertainty in any auto-correlations. With the use of computer simulations, I look to forecast benefits to be gained from this synergy and in Chapter 3 I provide an example of how HI intensity maps can be used to constrain photometric redshifts on optical imaging surveys.

The largest problem preventing the success of HI intensity mapping comes from 21cm foregrounds whose signals dominate by several orders of magnitude over the weak HI cosmological signal. While we have several methods for cleaning these foregrounds, understanding the impact these reconstructions have on the data is crucial and is the key theme in Chapters 4 and 5. Again using computer simulations of cosmological HI intensity mapping signals and their foreground contamination, I show how foregrounds can be removed and with some additional treatment, successfully used in cross-correlation with an optical photometric galaxy survey.

This indicates a promising future for cosmology and suggests the next-generation of optical telescopes such as LSST and Euclid, should benefit hugely from synergies with intensity mapping data provided by a next-generation radio telescope such as the SKA.

DEDICATION & ACKNOWLEDGEMENTS

"Like a cow that every day in 10 years sees the train cross in front at the same time. If you ask the cow, 'what time is the train going to come', it is not going to know the right answer."

— Mauricio Pochettino

I would like to begin by thanking my brilliant supervisors, David and Alkistis, both of whom I am in an infinite state of debt to. During my PhD I have seen both of them receive promotions and they are thoroughly deserving of this success. I have been truly fortunate to have two academics with a seemingly endless supply of knowledge and patience and a large part of my development is owed to them.

I would also like to thank all of my (past and present) collaborators Ian Harrison, Laura Wolz, Chris Blake and Julian Bautista, each of whom it has been a pleasure to work with and learn from. Furthermore, I thank the many talented colleagues and acquaintances within the cosmology community who have helped with useful discussions and advice such as David Alonso, Phil Bull, Stefano Camera, Emma Chapman, Benjamin Joachimi, Mario Santos, Anna Zoldan, and no doubt many more.

Throughout my PhD I have been fortunate enough to be offered the chance to speak at a number of local astronomical societies or science groups. The enthusiasm I find at every one of these volunteer-run groups is astounding and I am grateful for the excellent engagement I receive when I discuss my research with these groups. Some of the most challenging questions I have received after providing a talk have come from these sessions and they have motivated me to broaden my knowledge as far as possible.

Were it not for my regular opportunities to play football, I feel this PhD would have tested my sanity far more. So therefore I am grateful for all those around the University of Portsmouth I have played with. In particular I would like to thank our 5-a-side team ICG Intergalácticos, whose rapid progression in the University Staff League I hope will continue. Also, a thank you to all those involved in our series of 11-a-side matches against the Physics undergraduate students, especially since the staff are still leading the series 3-1!

I am grateful to my parents and a wide supporting family who provide the foundations to my life. I would like to end by thanking the most important person in my life. You have been with me since the start of my higher education in Physics and your love and companionship has helped more than you will ever realise or I'll ever be able to express. My partner Justine, I dedicate this to you.

My work is supported by the University of Portsmouth. Numerical computations for this thesis were done on the Sciama High Performance Compute (HPC) cluster which is supported by the ICG, SEPNet, and the University of Portsmouth.

TABLE OF CONTENTS

AUTHOR'S DECLARATION

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Word Count: 47,527

DISSEMINATION

Research Papers

1. **Cunnington, S. D.**, Wolz, L., Pourtsidou, A., & Bacon, D. (2019). Impact of Foregrounds on HI Intensity Mapping Cross-Correlations with Optical Surveys. Submitted to: Mon. Not. Roy. Astron. Soc. arXiv:1904.01479 [astro-ph.CO]

2. SKA Cosmology Science Working Group: Bacon, D. J., Battye, R. A.,Bull, P., Camera, S., Ferreira, P. G., ..., **Cunnington, S. D.** et al. (2018). Cosmology with Phase 1 of the Square Kilometre Array: Red Book 2018: Technical specifications and performance forecasts. Submitted to: Publ. Astron. Soc. Austral. arXiv:1811.02743 [astro-ph.CO]

3. **Cunnington, S. D.**, Harrison, I., Pourtsidou, A., & Bacon, D. (2018). HI Intensity Mapping for Clustering-Based Redshift Estimation. Mon. Not. Roy. Astron. Soc. 482 (2019), 3341–3355. doi:10.1093/mnras/sty2928. arXiv:1805.04498 [astro-ph.CO]

   Not discussed in this thesis:

   Collett, T. E. & **Cunnington, S. D.** (2016). Observational selection biases in time-delay strong lensing and their impact on cosmography. Mon. Not. Roy. Astron. Soc. 462(3), 3255–3264. doi:10.1093/mnras/stw1856. arXiv:1605.08341 [astro-ph.CO]

PREFACE

Observing sources of light from the night sky with the aim of understanding their origins has been something humans have done for millennia. Historical evidence exists which suggests ancient civilizations methodically recorded celestial objects. As our technology developed and our wider understanding of natural sciences blossomed, we have been able to devise models to explain the observations and make predictions which test these theories.

However, until recently in our history, this scientific discipline has been based on observations limited to within our own cosmic neighbourhood. Observing beyond our own Galaxy, the Milky Way, presents a huge technological challenge and requires detecting photons from other galaxies millions of light-years away which will therefore appear extremely faint. These complications mean that regular extra-galactic observations have only been a possible scientific discipline within the last 100 years when our telescopes have become sufficiently capable of detecting distant galaxies.

These distant observations present the possibility for us to study the Universe as one entity. We can now aim to build a representative map of large-scale cosmic structure and determine the matter from which this structure is built. Through detecting faraway photons which originated billions of years ago, we also peer deeper into our Universe's past. Cosmologists then look to pursue questions about our Universe's origins, evolution and predict its fate. These are the principal aims for the topic of cosmology.

While the topic of cosmology is a relatively young one by scientific standards, we have succeeded in developing a standard model to explain our Universe. On one hand, this model still has huge unanswered questions regarding the exact nature of 95% of the Universe's energy and matter content, the so called *dark sector* which includes *dark energy* and *dark matter*. However, on the other hand, with the assumption that we broadly understand how this dark sector behaves, our standard model proves successful at explaining our observations and replicating them with theoretical simulations.

The standard model of cosmology is not something that can predict the exact locations of every galaxy, cluster and void. Instead it predicts statistical properties of fields and how the matter within them is distributed. Cosmological constraints and parameters are therefore inherently statistical and their precision is driven in large part by access to extensive data sets. The aim of modern, precision cosmology is therefore to record as large volumes of the Universe as possible to maximise statistical precision, placing tighter bounds on our statistical constraints. This in turn can potentially reinforce or rule out certain hypothese on the nature of the dark sector. It also can reveal any tensions in independent measurements of the same parameters, thus highlighting areas of our standard model which might not be as robust an explanation of reality as we had hoped.

With new telescopes promising significant improvements in depth, breadth and sensitivity over their previous generations, now is the time to plan, forecast and conduct further tests on cosmology using these future surveys. This thesis therefore looks to focus on the future for cosmology surveys and their aims for probing large-scale cosmic structure. There are telescope surveys which look to perform this objective with completely differing methods, which

is hugely beneficial in cross-correlation approaches. Synergies between telescope surveys will be a central theme of this thesis. In particular it will explore how cosmology can benefit from cross-correlations between two types of surveys. The first being conventional optical surveys which look to detect and resolve individual galaxies and predict a radial distance to these galaxies based on their redshifted spectral features. The second being surveys of neutral hydrogen emission which can be detected from the signature 21cm signal emitted from such sources.

This 21cm cosmology approach is seen as a novel way of conducting observations and has the potential to map relevant scales with unprecedented precision and volume. As with any novel technique, we need to understand it in detail along with the relevant systematics before any confidence can be placed on the conclusions gleaned from it. This thesis will introduce the 21cm cosmological signal and the radio telescopes we use to survey it.

The excitement surrounding future surveys is justified but so is the emphasis on the various challenges which need to be overcome. Cross-correlations between the surveys offers an excellent opportunity to alleviate a large amount of systematics which threaten to be the largest source of error. A thorough understanding and forecasting of this cross-correlation potential is required to assure we can maximise what we can learn from these cosmological surreys. Furthermore, synergies with optical will help lay the foundations of 21cm observations which has the potential to lead the way for the future of precision cosmology.

CHAPTER 1

INTRODUCTION

Constructing a model which describes the Universe requires agreement with the most fundamental of observations. The most basic observable that can be made in cosmology is that the night sky is predominantly dark. Known as Olbers' paradox, this simple observation seems to contradict any infinite and static model of the Universe. If we lived in a Universe that had an infinite amount of stars, eternally stationary with respect to us, we would be bathed in ever compounding starlight and darkness would be impossible.

However, in 1929 Edwin Hubble concluded that Doppler shifts of light from galaxies, caused by their relative recessional motion away from us, appear to increase the more distant the galaxy is [106]. This suggests that in general galaxies are moving away from us and the further away a galaxy, the faster it is moving. To avoid the unsettling conclusion that we are at the exact centre of the Universe, we explain this phenomenon by instead concluding that space in the Universe is expanding and therefore all galaxies embedded within it are moving further apart from each other, where they are not gravitationally bound. This appears to quash any theory which suggests the Universe is static. Furthermore, Belgian priest Georges Lemaître noticed that if time is reversed in an expanding Universe, eventually a point is reached where all space and matter is on top of each other in a point-like singularity [123]. Lemaître concluded that this was how the Universe must have begun, in a 'Primeval Atom' (or *big bang* as its now more popularly known). If true, this also quashes any theory which suggests the Universe's existence is infinite[1].

Building on the work of Hubble and Lemaître, the early pioneers of cosmology, we now have a model of the Universe which begins in a 'big bang' followed by phases of space expansion. This means we are no longer troubled by a dark night sky as Olbers and colleagues were in the early 19$^{\text{th}}$ century. The Big Bang theory, whilst strictly speaking not proven, appears very consistent with the standard model [126]. Suggesting that the Universe is finite in age and began with a rapid expansion of space, ultimately makes photons from all sources in an infinite Universe incapable of travelling the distances required to reach us. The continuing phases of

[1] Assuming our Universe is not periodically 'bouncing'

space expansion within the Universe, first noticed by Hubble, also mean that distant photons, including relic radiation from the Big Bang, will be *redshifted* into microwave wavelengths hence why we see very little visible light in the night sky.

Central to these theories involving space expansion is the assumption that we do not exist in a particularly special place in the Universe. We make a symmetry argument and assume that on large enough scales (around $62h^{-1}$Mpc [149]), matter in the Universe is uniformly distributed. This is known as the *cosmological principle*, which put simply states that:

The Universe is homogeneous and isotropic on sufficiently large scales.

Here *homogeneous* means the matter in the Universe is uniformly distributed on large scales throughout space and hence its distribution is independent of spatial coordinates ($\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$). *Isotropy* means the large scale distribution of matter in the Universe appears the same whichever direction an observer looks and is thus independent of $\hat{\mathbf{n}}$ the line-of-sight. The cosmological principle allows us to greatly simplify and solve the laws of gravity which on cosmological scales is the dominant force. The best description of gravity comes from Albert Einstein's general relativity field equations, which are given as [71]

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu} - \left[g_{\mu\nu} \Lambda \right]. \tag{1.1}$$

A derivation of these field equations is beyond the scope of this thesis but a brief explanation of them is warranted since it is these equations which describe the evolution of the universe. Firstly, this is described as a set of equations since each index μ and ν can be one of the four coordinates of space-time where the convention $[0,1,2,3] \equiv [ct, x, y, z]$ is used. There are only 10 possible distinct permutations of these components hence (1.1) represents a set of 10 equations.

$G_{\mu\nu}$ is the Einstein tensor which describes the geometry of a universe. This tensor can be expressed using $R_{\mu\nu}$ and R, the Ricci tensor and Ricci scalar respectively. $T_{\mu\nu}$ is the energy-momentum tensor which characterises the matter distribution in a universe and has the form $T_{00} = \varepsilon(t)$ (where ε is the energy density parameter[2]) and $T_{ij} = p(t)\delta_{ij}$ (where p is the pressure parameter and δ_{ij} is the *Kronecker delta*) with i and j representing spatial indices only. G represents Newton's gravitational constant and c is the speed of light, both often set equal to one for simplicity (but not in this thesis). $g_{\mu\nu}$ is the metric tensor and the simplest example in flat space is referred to as the *Minkowski metric* where $\eta = \text{diag}[-1,1,1,1]$. Lastly within the bracketed term we have the cosmological constant Λ, an 'optional' inclusion for describing a universe undergoing accelerated expansion.

Solving these equations is the starting step for the standard model of cosmology which primarily becomes a description of the underlying cosmological matter density field. Coupled with our best theories of smaller scale physics i.e. Quantum Field Theory (QFT) [65][75], we understand that structure within these cosmic matter fields originates in initial seed perturbations. Evidence for these initial perturbations exits within Cosmic Microwave Background

[2]Often energy density ε and mass density ρ are used interchangeably since $\varepsilon = \rho c^2$ and in cosmology natural units of $c = 1$ are often used.

(CMB) radiation [163][64] but the exact origin of them is not entirely clear and often attributed to 'quantum fluctuations' connected with early universe physics. These tiny seeds of density fluctuation lead to non-uniform gravitational potentials which grow causing structure to evolve. The model is therefore consistent with the existence of complex galaxies, stars and planetary systems which are forged from these non-linear gravitational collapses.

This chapter outlines the theoretical framework from which the standard model of cosmology is built. Basic cosmological observables are also introduced which form the complimentary observational framework to support the model.

1.1 The Space-Time Metric

Since Einstein's relativity [70] suggests that space and time act as one entity, an excellent framework from which to describe the physics of our Universe is a 4-dimensional manifold. In order to describe physical distances in a 4-dimensional space-time, an infinitesimal line element can be defined, which we call the *metric*. This will map coordinate distances to physical distances turning observer dependent measurements into invariant ones. As an example, a separation on a 2-dimensional mountain map which one measures to be 1cm will have different physical distances depending on where the separation is on the map. It will require a metric which considers the local gradients of the mountain to obtain an accurate physical distance. What is very useful about metrics in cosmology is that they incorporate the curvature of space-time, which according to general relativity, is equivalent to gravity.

For completely flat space-time with no curvature (referred to as *Euclidean* geometry) we can use the Minkowski metric and we have [141]

$$\mathrm{d}s^2 = \eta_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = -c^2\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2. \tag{1.2}$$

This can also be expressed in spherical coordinates and is given as

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + \mathrm{d}r^2 + r^2\mathrm{d}\Omega^2 \tag{1.3}$$

where r is the radial distance coordinate and we also have the angular coordinate term

$$\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2 \tag{1.4}$$

where θ and ϕ are the angular separations on the 2-dimensional sky e.g. the *right-ascension* and *declination* coordinates which are conventionally used coordinates for Earth-based observations, although isotropy suggests that angular coordinates for any reference frame's origin will yield the same results.

We ideally want to generalise the 3-dimensional spatial manifold to allow for the possibility of *spatial expansion* and *spatial curvature*. Both of these need to be accounted for in the metric.

- **Spatial Expansion**

To account for the fact that the Universe is expanding, the space-time metric includes a scaling factor a such that comoving observers have constant spatial coordinates r, θ and ϕ but the proper distances between given coordinates increases in proportion to $a(t)$. The scale factor is a function of time only and increases as the Universe evolves and expands. For a flat universe with expansion on sufficiently large scales where the cosmological principle applies, the Minkowski metric can be simply corrected by multiplying through the spatial dimensions by $a(t)$.

- **Spatial Curvature**

Devising a metric for space-time which contains *spatial* curvature is potentially complex. Thankfully, things are greatly simplified under the assumption of the cosmological principle. This means that at a given constant time, curvature will be the same everywhere and we can assume that the spatial elements of space-time will be a 3-dimensional, maximally symmetric manifold. The most general metric for such a manifold is given by [63]

$$\mathrm{d}l^2 = \left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\Omega^2\right). \tag{1.5}$$

Here the curvature parameter k is what defines the form of this metric and there exist three possibilities, zero curvature (flat infinite space), positive curvature (3-sphere finite space), or negative curvature ('saddle'-like infinite space), given by $k = 0, +1, -1$ respectively.

A metric that describes a 4-dimensional homogeneous and isotropic space-time which is expanding with time is therefore needed for our Universe. The most generic form of this metric is agreed upon. What is disputed is the name given to it! Various permutation of Friedmann, Lemaître, Robertson and Walker are used with some names omitted in various versions. To avoid offending descendants of any of these cosmology greats, I will play safe and refer to it as the Friedmann-Lemaître-Robertson-Walker (FLRW) metric [84][122][181] which is given by

$$\boxed{\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a^2(t)\left(\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\Omega^2\right).} \tag{1.6}$$

The above (unperturbed) FLRW metric is crucial in cosmology for correctly describing the background Universe and therefore for interpreting distance measurements.

1.2 Cosmological Observables & Parameters

1.2.1 Redshift

Wavelength shifts of light waves from distant receding galaxies tend to show that the light is shifted to the red end of the electromagnetic spectrum, hence the term *redshift* which is defined by the fractional difference between the observed wavelength of light and the emitted [66]

$$z + 1 \equiv \frac{\lambda_{\mathrm{obs}}}{\lambda_{\mathrm{emit}}} \tag{1.7}$$

where z is the unitless quantity for redshift and λ is the wavelength of either the observed or emitted photons. Hence by knowing what the emitted rest wavelength is for a photon, which we can measure in the laboratory for various elements, we can obtain the redshift z by measuring the observed wavelength of distant signals. From special relativity, redshift is related to the velocity v along the line-of-sight (LoS) by

$$z+1=\sqrt{\frac{1+\beta}{1-\beta}} \tag{1.8}$$

where $\beta \equiv v/c$. In the $v \ll c$ limit, we can Taylor expand (1.8) and ignore higher terms to arrive at the well-known approximation $z \approx v/c$.

By looking at the redshift of momenta of photons, one can demonstrate that redshift is in general a direct consequence of an expanding universe [63] i.e.

$$\lambda(t) = a(t)\lambda_{\rm obs} \tag{1.9}$$

and since the overwhelming majority of measured spectra are redshifted (i.e. $\lambda_{\rm obs} > \lambda(t)$), this alone is evidence that the Universe is expanding. However, redshift alone does not tell us the precise distances to objects. We need to assume some cosmological expansion history $a(t)$ in order to make redshift-based distance measurements. Doing this relies on making some direct distance measurements and determining a distance-redshift relationship.

1.2.2 Distances

Comoving Distance

The cosmological principle only holds in the comoving system and therefore the most common distance measurement used in cosmology is the comoving distance $d_{\rm c}$ or often referred to as χ. This is defined such that if an object in free-fall i.e. not gravitationally bound, is at rest at time t and a comoving distance $d_{\rm c}$ away from an observer, then they remain at this comoving distance. Put mathematically, $d_{\rm c}(t) = d_{\rm c}(t_0)$. For a particle of light with $\mathrm{d}s^2 = 0$ along the radial line-of-sight ($\mathrm{d}\Omega^2 = 0$), equation (1.6) gives the comoving distance as

$$d_{\rm c}(t_0) = \int_{t_{\rm e}}^{t_0} \frac{c\,\mathrm{d}t}{a(t)} = \int_0^r \frac{\mathrm{d}r'}{\sqrt{1-kr'^2}}. \tag{1.10}$$

Proper Distance

The comoving distance differs for bound objects with fixed lengths e.g. a ruler with a fixed 'proper' size will have a decreasing size defined by its comoving distance in an expanding universe. Proper distance can therefore be linked to comoving distance by

$$d_{\rm pr}(t) = a(t)d_{\rm c}(t_0) \tag{1.11}$$

The proper distance we can think of as a chain of infinite neighbouring observers placed radially out to an object, instantaneously exchanging a light signal at the same time t. Again from

equation (1.6), but now with $a(t)$ outside the integral since it is measured for a constant time t, we find

$$d_{\rm pr}(t) = a(t)\int_0^r \frac{{\rm d}r'}{\sqrt{1-kr'^2}} \tag{1.12}$$

which is in agreement with equations (1.10) and (1.11). Equation (1.12) also shows that the proper distance of a luminous source at present time ($a(t) = 1$) is just

$$d_{\rm pr}(t) = a(t)d_{\rm pr}(t_0) \tag{1.13}$$

which shows that this is a physical distance measurement that scales with expansion.

In terms of practically measuring distances, astronomers typically use either an objects known *luminosity* or a known *angular size*.

Luminosity Distance

For an astrophysical object with known absolute luminosity L we can relate this to the observed flux $F = L/4\pi r^2$ when the object is at a radial distance r away from us (assuming Euclidean geometry). Therefore we can define the luminosity distance as

$$d_{\rm L} = \sqrt{\frac{L}{4\pi F}}. \tag{1.14}$$

However, in an expanding universe the energy emitted from each photon $E_{\rm em}$ is redshifted such that the energy that reaches us is

$$E_0 = E_{\rm em}a_{\rm em} = \frac{E_{\rm em}}{(1+z)}. \tag{1.15}$$

An additional effect of an expanding universe comes from the stretching of the time interval $\delta t_{\rm em}$. The actual time interval we observe is δt_0 given by

$$\delta t_0 = \frac{\delta t_{\rm em}}{a_{\rm em}} = \delta t_{\rm em}(1+z). \tag{1.16}$$

This redefines the flux we observe since the luminosity is reduced by these two effects

$$L_{\rm em} = \frac{E_{\rm em}}{\delta t_{\rm em}} \Rightarrow L_0 = \frac{E_0}{\delta t_0} = \frac{1}{(1+z)^2}\frac{E_{\rm em}}{\delta t_{\rm em}}. \tag{1.17}$$

So the effect on the observed flux from an object with absolute luminosity L at a proper distance $d_{\rm pr}$ away is a $1/(1+z)^2$ reduction i.e. $F = L/4\pi r^2(1+z)^2$. Taking the radial distance to be the proper distance $d_{\rm pr}$ and using (1.14) we therefore find the relation between luminosity distance and proper distance is given as

$$d_{\rm L} = (1+z)d_{\rm pr} \tag{1.18}$$

Angular Diameter Distance

The final definition of measurement relevant for this work uses known sizes of objects or features (often known as standard yardsticks) and their observed angular size to predict a distance. This is known as angular distance and for an object with known length l with an observed angular size of $\delta\theta$ (where we assume $\delta\theta \ll 1$) we have the angular distance defined by

$$d_{\mathrm{A}} = \frac{l}{\delta\theta}. \tag{1.19}$$

Measuring the distance between two objects at time t with the same radial coordinate r and angular coordinate ϕ but separated by an angle $\delta\theta$ the metric from (1.6) is $l = a(t)r\delta\theta$. Rewriting the scale factor in terms of redshift and using equation (1.19) this gives

$$d_{\mathrm{A}}(z) = \frac{r(z)}{1+z} = \frac{d_{\mathrm{L}}(z)}{(1+z)^2}. \tag{1.20}$$

For further detailed discussion on cosmological distances, I refer the reader to [103].

1.2.3 Hubble Parameter

A useful quantity in cosmology, and one frequently used in the context of distance measurement and expansion, is the Hubble parameter which is defined as

$$H(t) = \frac{\dot{a}}{a} \tag{1.21}$$

where the dot above the scale factor $\dot{a}$ represents differentiation with respect to time. By taking the proper distance (1.13) and differentiating we get $\dot{d}_{\mathrm{pr}}(t) = \dot{a}(t)d_{\mathrm{pr}}(t_0)$. Then dividing through by (1.13) gives

$$v_{\mathrm{pr}}(t) = H(t)d_{\mathrm{pr}}(t) \tag{1.22}$$

where $v_{\mathrm{pr}} \equiv \dot{d}_{\mathrm{pr}}$. This equation is known as Hubble's law and is a theoretical description of what Edwin Hubble discovered in 1929 [106], that more distant galaxies have a greater recession velocity. By specifying this equation to present time we can define the Hubble constant $H_0 \equiv H(t = t_0)$. This is often measured in units of $\mathrm{km\,s^{-1}Mpc^{-1}}$ i.e. for each Mpc of distance, the velocity of a distant object increases by some velocity measured in $\mathrm{km\,s^{-1}}$. The precise value of the Hubble constant is an active area of research in the cosmology community since differing measurement techniques disagree and have introduced a 3.4σ tension between their measured H_0 values [80][144]. It is customary to use $H_0 = 100h\,\mathrm{km\,s^{-1}Mpc^{-1}}$ where h is a dimensionless number to parameterise our ignorance.

Since $a = 1/(1+z)$ we have $\mathrm{d}a = -\mathrm{d}z/(1+z)^2$. This helps derive a relationship between redshift and time where

$$\frac{\dot{a}}{a} = H(z) = \frac{\mathrm{d}a}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}t}\frac{1}{a} = -\frac{1}{(1+z)}\frac{\mathrm{d}z}{\mathrm{d}t}. \tag{1.23}$$

This then allows the comoving distance to be written in terms of redshift and the Hubble parameter. Using the above and equation (1.10) we get

$$d_{\mathrm{c}} = \int_{t_{\mathrm{e}}}^{t_0} \frac{c\,\mathrm{d}t}{a(t)} = \int_0^z \frac{c\,\mathrm{d}z'}{H(z')} \tag{1.24}$$

The more general Hubble parameter $H(z)$ can be thought of as the Hubble constant measured by an observer at redshift z and we have $H(z) = H_0 E(z)$ where

$$E(z) \equiv \sqrt{\Omega_{\mathrm{R}}(1+z)^4 + \Omega_{\mathrm{M}}(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda} \tag{1.25}$$

and Ω_i is the energy density parameter (discussed in next section) for radiation (R), matter (M), curvature (k) and dark energy (Λ). The Hubble distance (or horizon distance) is defined as $d_{\mathrm{H}} = c/H_0$. These further parameters allow the comoving distance to be written in some further commonly found forms

$$d_{\mathrm{c}} = \frac{1}{H_0}\int_0^z \frac{c\,\mathrm{d}z'}{E(z')} = d_{\mathrm{H}}\int_0^z \frac{\mathrm{d}z'}{E(z')}. \tag{1.26}$$

1.3 Friedmann Cosmological Models

The FLRW metric in equation (1.6) is described by just two parameters, the scale factor a and the curvature parameter k. Assuming homogeneity, the curvature parameter is taken to be a constant for the Universe. This means that the scale factor is the only time-dependent parameter and thus encodes all the dynamics of the Universe.

1.3.1 Friedmann Equations

Under the assumptions of the cosmological principle, Einstein's field equations (1.1) can be simplified and a prediction for the evolution of the Universe i.e. a description of how a evolves, can be made;

$$\boxed{H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3c^2}\varepsilon - \frac{kc^2}{a^2} + \left[\frac{\Lambda c^2}{3}\right].} \tag{1.27}$$

This is known as the Friedmann equation [83] where we have the energy density ε which due to homogeneity, is only a function of time. Another Friedmann equation, referred to as the conservation (or fluid equation) is given as

$$\dot{\varepsilon} + 3\frac{\dot{a}}{a}\left(\varepsilon + p\right) = 0. \tag{1.28}$$

This can be considered as a consequence of thermodynamics from which the first law states $\mathrm{d}E + p\mathrm{d}V = T\mathrm{d}S$ [40]. For a reversible expansion i.e. $\mathrm{d}S = 0$, using $E = \varepsilon V$ in a volume with radius a and considering time differentials one can derive this conservation equation. We assume the content of the Universe, given by the energy momentum tensor $T_{\mu\nu}$, is made up of non-interacting components. Therefore this energy conservation holds for each individual component and their solutions to the equation will be independent.

By differentiating (1.27) and rearranging with use of (1.28), we get the final Friedmann equation referred to as the acceleration equation

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3c^2}(\varepsilon + 3p) + \left[\frac{\Lambda c^2}{3}\right]. \tag{1.29}$$

This immediately suggests that for a universe with no accelerated expansion, we require $\varepsilon+3p=0$ (assuming $\Lambda=0$ also). It is often useful to cast these Friedmann equations in terms of energy density parameters $\Omega=\varepsilon(t)/\varepsilon_{\rm c}(t)$ where $\varepsilon_{\rm c}$ is the critical energy density given by

$$\varepsilon_{\rm c}(t)\equiv\frac{3c^2H^2(t)}{8\pi G}. \tag{1.30}$$

This allows the Friedmann equation in (1.27) to be expressed in present day terms as

$$H_0^2\,(1-\Omega_0)=-kc^2 \tag{1.31}$$

thus requiring $\Omega_0=1$ for a flat universe.

1.3.2 Equation of State

By assuming a simple relationship between pressure and density i.e. an equation of state w where $p\equiv w\varepsilon$, we can solve the conservation equation (1.28) obtaining [184]

$$\varepsilon(t)=\varepsilon_0\left(\frac{a(t)}{a_0}\right)^{-3(1+w)}\Rightarrow\varepsilon\propto a(t)^{-3(1+w)}. \tag{1.32}$$

We can then explore these solutions by assuming different content in the Universe with different equations of state. For example cold (non-relativistic) matter is defined as anything that exerts negligible pressure, therefore will have $p=0$ [17] leading to $w_{\rm M}=0$ meaning the energy density of matter scales like $\varepsilon_{\rm M}\propto a^{-3}$. This makes intuitive sense, suggesting that the energy density of matter scales as the inverse of volume. The equation of state for radiation, or anything moving with relativistic velocities such as neutrinos, is given as $w_{\rm R}=1/3$ [40]. This gives an energy density evolution as $\varepsilon_{\rm R}\propto a^{-4}$ which can be understood as a scaling with inverse volume and an additional a^{-1} scaling from redshift as previously shown by equation (1.18).

These results are already telling us something about the composition of the Universe at different times. As the Universe evolves and expands, the scale factor a increases. Therefore matter will begin to dominate over radiation due to it falling off more slowly.

1.3.3 Late-Time Acceleration & Dark Energy

Since Hubble published his findings [106] on distance against redshift which appeared to suggest that the Universe is expanding, one of the follow-up questions has been whether the rate of this expansion is slowing down. One would expect the expansion to be decelerating since the Universe is filled with content which feels the effect of gravity which acts only to attract objects towards each other. So it was believed that evidence would exist which shows that the Universe was expanding faster in the past [191]. As the Universe evolved, it was predicted that gravity, caused by the energy density term in the Friedmann equation (1.27), would slow the expansion rate.

In 1998, from observations of type 1-a supernovae (SNeIa) two independent teams [179][165] reached the unexpected conclusion that the expansion of the Universe is not slowing down, but accelerating. Since then it has been assumed that there must be more to the Universe than

meets the eye. There must either exist some exotic constituent that opposes gravity and drives the late-time accelerated expansion, or we are faced with the unsettling conclusion that the laws of gravity are incomplete. Cosmologists often use the phrase *dark energy* to describe the plethora of possibilities for solving this problem.

Perhaps the simplest form of dark energy which causes the standard model to predict an accelerated expansion is the cosmological constant [161]. This is included in the bracketed terms of Einstein's equations (1.1) as the Λ term. It was originally proposed by Einstein as a way of obtaining a static universe since the acceleration equation (1.29) requires $w = -1/3$ for static solutions which cannot be obtained with a content of matter and radiation alone. Including this term to explain accelerated expansion however, has the required effect of repulsive gravity.

The cosmological constant can equivalently be thought of as an additional form of fluid instead of a modification to general relativity e.g. $T^{\text{Tot}}_{\mu\nu} = T^{\text{M}}_{\mu\nu} + T^{\text{R}}_{\mu\nu} + T^{\Lambda}_{\mu\nu}$. This translates into an extra energy density ε_Λ which acts as a constant energy density despite the expansion of the Universe. This recasts the Friedmann equations essentially setting $\Lambda = 0$ in the 'optional' brackets and having a cosmological constant 'fluid' appear as an extra constituent alongside matter and radiation. For example, the conservation equation (1.28) for the Λ component becomes

$$\dot{\varepsilon}_\Lambda + 3\frac{\dot{a}}{a}\left(\varepsilon_\Lambda + p_\Lambda\right) = 0 \tag{1.33}$$

where since $\dot{\varepsilon}_\Lambda = 0$ by definition, we find the cosmological constant has a negative effective pressure $\varepsilon_\Lambda = -p_\Lambda$ with equation of state $w_\Lambda = -1$. This additional and exotic fluid is most generally explained as being the energy density of the vacuum. Interestingly, a vacuum energy is something that is independently predicted by QFT [77] but there exist some serious tensions between what QFT predicts as the value for the energy density and what value is needed to explain the accelerated expansion we see [222]. Despite this we can still hypothesise the existence of such a fluid and as discussed in Section 1.4 this leads to some excellent agreement with observational data.

It is common to use the energy density parameter Ω for describing the contents of a universe. Since we can have mixes of different independent components e.g. matter, radiation and cosmological constant $\varepsilon = \varepsilon_{\text{M}} + \varepsilon_{\text{R}} + \varepsilon_\Lambda$, the energy density parameter is also given by the sum of the individual contributions. So for a present day universe we have

$$\Omega_0 = \Omega_{\text{M},0} + \Omega_{\text{R},0} + \Omega_{\Lambda,0}. \tag{1.34}$$

As shown in Section 1.3.2, these energy densities evolve with expansion and as will be discussed, their relative abundances have important consequences for the dynamics of the Universe.

1.4 ΛCDM

The previous sections have outlined the basics of the standard model of concordance cosmology, the ΛCDM model. The cosmological constant Λ represents the fact that this model has an exotic component which drives the accelerated expansion of the late Universe and the CDM stands for

Cold Dark Matter which makes up the majority of the matter content (introduced in Section 1.4.2).

Including the inflationary paradigm to describe the rapid expansion in the early Universe [89], ΛCDM is extremely successful at explaining astrophysical observations with few parameters [176]. I discuss some of the main examples of supporting evidence in this section. The obvious criticism of ΛCDM is that these two main components represent 'exotic' forms of matter-energy that lack a description from the standard model of particle physics built from QFT which describes 'ordinary' matter and radiation with great accuracy [221].

1.4.1 The Cosmological Constant (Λ)

While a great deal of uncertainty exists over the origin of dark energy, the evidence for accelerated expansion is robust. There exist a number of independent probes that converge on the same conclusion, which is the existence of dark energy in the form of a cosmological constant Λ (introduced in Section 1.3.3). I outline three of the major pieces of supporting evidence below.

- **Evidence (i) - SNeIa**

Supernovae are a fantastic tool for cosmologists [180]. When type Ia reach a peak in their light curves, their absolute luminosity is approximately a known constant. This means they are standardisable candles and as long as a light-curve for a supernova can be obtained, then an excellent approximation can be made on the absolute luminosity and from this a luminosity distance can be obtained [184]. By obtaining redshifts for these supernovae a distance-redshift measurement can be made. This was the method conducted in [179] which first suggested an accelerated expansion $\ddot{a} > 0$. More recent measurements have been done with SNeIa [203] which have reinforced these earlier results. Figure 1.1 shows supernovae data from the recent Dark Energy Survey[3] (DES) and its agreement with a ΛCDM-like cosmology which favour values of $\Omega_M \sim 0.3, \Omega_\Lambda \sim 0.7$.

- **Evidence (ii) - CMB**

Observations of the Cosmic Microwave Background (CMB), which is radiation from approximately 10^{13} seconds after the big-bang ($z \sim 1100$), have been conducted since the 1960's [163]. More recently, precise measurements of the acoustic peaks in the angular power spectrum have been made by the WMAP and *Planck* satellite telescopes [100][9] (see Figure 1.2 for these latest results provided by [9]). The precise positions of these peaks provide a wealth of cosmological information (see [105] for a more complete review of this probe). Perhaps most importantly, for the purposes of this section, is the position of the first peak which put simply gives the angular scale for the strongest fluctuations in CMB temperature which we can accurately predict based on what we know about the conditions of the Universe at this time. Measurements strongly agree with this prediction which would not be the case if there existed spatial curvature. Setting $k = 0$

[3] www.darkenergysurvey.org

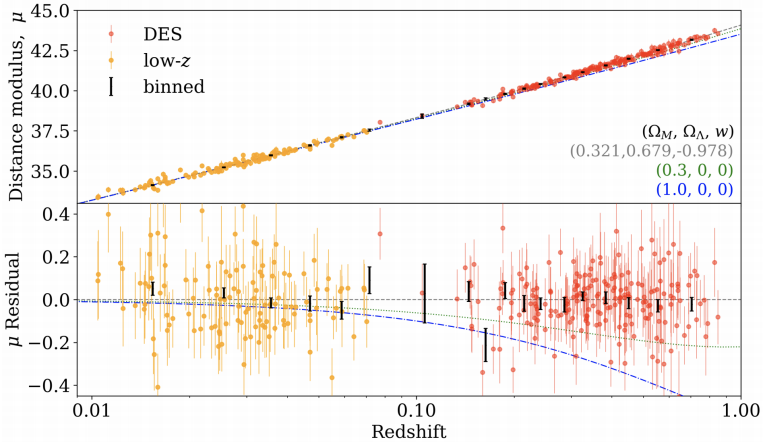

Figure 1.1: Hubble diagram for the DES-SN3YR sample [3]. The dashed grey line shows the best fit model, while the green and blue dotted lines show models with no dark energy and matter densities $\Omega_M = 0.3$ and 1.0 respectively. Bottom panel is residuals to the best fit model.

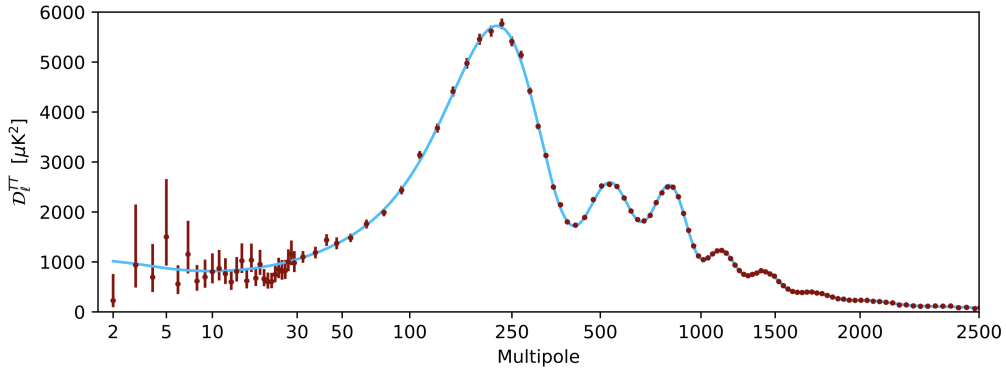

Figure 1.2: CMB power spectrum from Planck [9]. The relation between ℓ and θ is $\theta = \pi/\ell$. The positions and amplitudes of these peaks provide constraints on cosmological parameters. For example the first peak position provides a measurement of the horizon scale at recombination. Using the angular diameter distance to this measured scale of $\ell = 220.6 \pm 0.6$ we get a good agreement with a ΛCDM model with near perfect flatness ($\Omega_k = 0.0007 \pm 0.0019$).

in the Friedmann equation (1.27) requires a balance between the Hubble parameter and energy density. Including the energy density of matter and radiation alone is insufficient to achieve this balance thus a cosmological constant term Λ or additional energy density ε_Λ must be included to allow for flatness.

- **Evidence (iii) - BAO**

The same physics which leaves a characteristic scale on the CMB angular power spectrum, also leaves an imprint in the late-time matter density which results in a preferred separation r_s between density peaks. Through observing the positions of galaxies, which trace the underlying matter density and should therefore exhibit this preferred separation (see Figure 1.3 for an example of this measured BAO-'bump'), r_s has been used as a standard ruler to create a

distance-redshift relationship. This distance-redshift relation from BAO also hints at an accelerated expansion strongly consistent with a $w_\Lambda = -1$ cosmological constant. For an example, see recent results from SDSS-III BOSS DR12 [10].

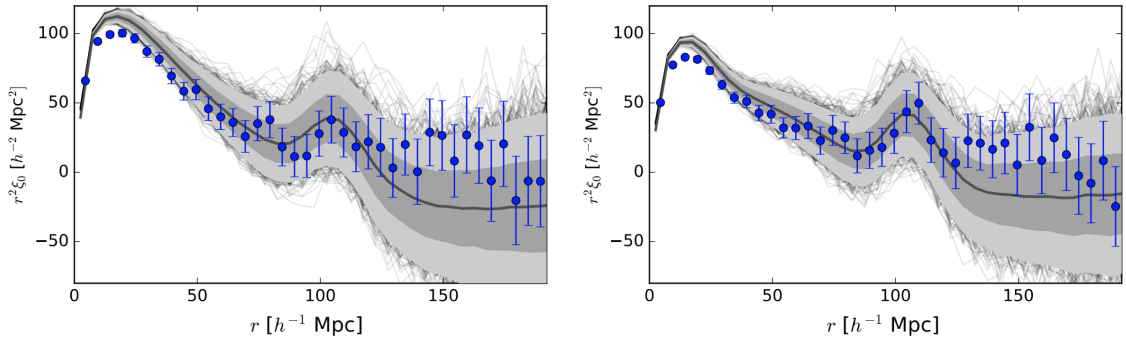

Figure 1.3: BAO measurement in configuration space of the monopole ξ_0 from [30]. Data shown as blue points and mock catalogues as grey lines. Panels on the left (right) show the pre(post)-reconstruction catalogs. The shaded regions represent the 68% and 95% boundaries of the distributions of correlation functions around the mean.

These are three key probes used to infer cosmological parameters but often the best way to get tight constraints on such values is through cross-correlations of them, a process demonstrated in [20] and shown in Figure 1.4 produced by the Supernova Cosmology Project [203]. In addition to the above list of probes, there are additional tools which cosmologists can use, e.g. weak lensing ([130] for review) and probing galaxy clustering, something I discuss in the Section 1.5. Both of these are achievable with photometric imaging surveys such as the Dark Energy Survey (DES) [1] as demonstrated in [2].

1.4.2 Cold Dark Matter (CDM)

Our best description of particle physics is the standard model which elegantly explains baryonic matter. However, there is strong evidence from cosmological and astrophysical probes [242][183][121] of matter beyond this standard model which must interact via gravity with baryonic matter. But since this matter shows no signs of interaction though the other forces, it is something that should be difficult to directly detect, hence its name, *dark* matter. The most successful model is from the hypothesis that dark matter is some form of weakly interacting particle with non-relativistic velocities [202]. This is why it is referred to as 'cold' to distinguish it from 'hot' dark matter models where the particle can have relativistic velocities e.g. in the form of a massive neutrino [67].

The understanding of this content is fundamental to cosmology since dark matter forms the 'skeleton' on which galaxies grow. Hence in large part, probing large-scale structure (discussed in Section 1.5), involves attempting to map the underlying dark matter density. The latest cosmological probes [9] are consistent with a ΛCDM model of cosmology with present day contents of $\Omega_\Lambda = 0.689$ and $\Omega_M = 0.311$. For a flat universe, which the same data is also consistent

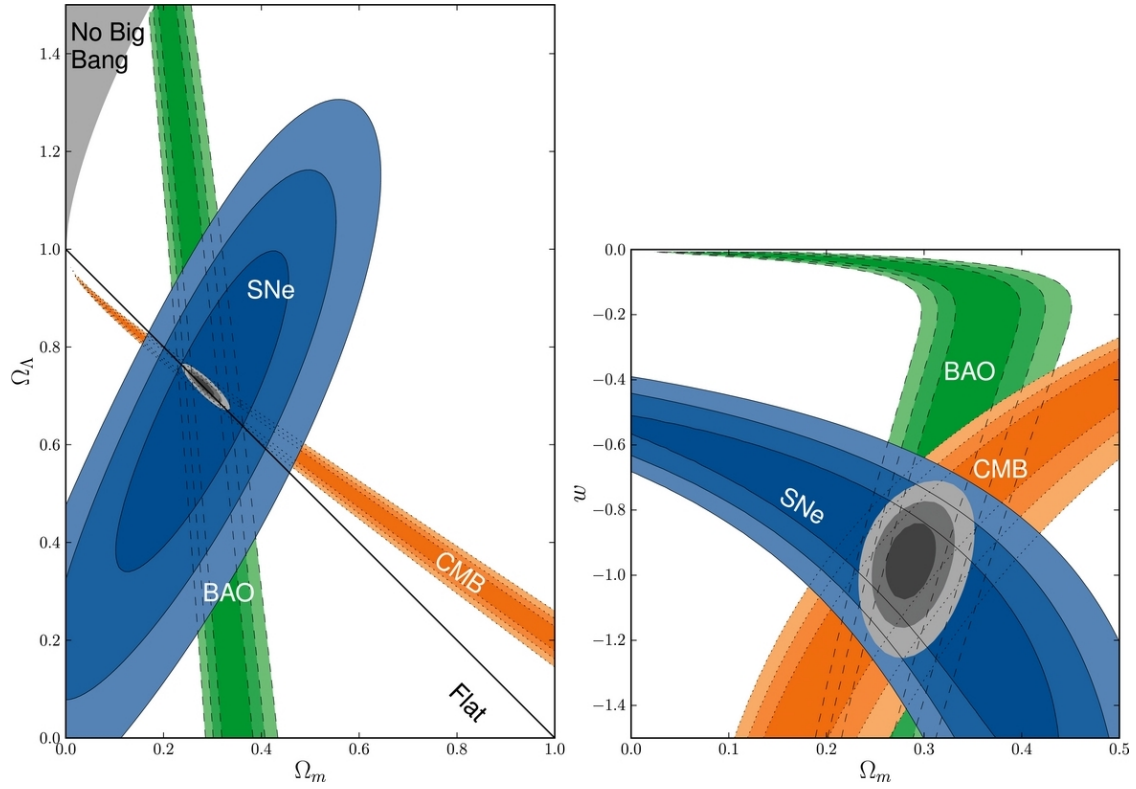

Figure 1.4: Constraints on cosmological parameters from three combined probes; Supernovae (SNe), Cosmic Microwave Background (CMB) and Baryon Acoustic Oscillations (BAO). Produced by the Supernova Cosmology Project [203]. 68.3%, 95.4%, and 99.7% confidence regions shown.

with, this amounts to the Universe's energy-matter content being 68.9% dark energy and of the 31.1% remaining matter, 26.0% is in the form of dark matter.

This dominance of dark matter in the Universe's matter density presents a practical problem. Cosmological theories will be largely concerned with this main substance and its behaviour yet observations with telescopes are only capturing the visible light resulting from baryonic interactions. This is something cosmologists are required to be mindful of when analysing observational data and involves considering that an intrinsic *bias* may exist in measurements. Something that is formalised in Section 1.5.1.

1.5 Large-Scale Cosmic Structure

A further piece of evidence in support of the ΛCDM model comes from probing how the matter in Universe is distributed, referred to as Large-Scale Structure (LSS). Cosmologists are not necessarily interested in the exact structure of the Universe on these largest scales because the aim is not to develop a theory which exactly reproduces our Universe as we see it down to the precise position of every galaxy. Instead, the interest is in the statistical distribution of matter as it can be shown that dark energy, which drives the background expansion of the Universe, can affect the growth rate of structure.

1.5.1 Two-Point Correlation Function & Power Spectrum

Since we wish to analyse the Universe as one representation in an ensemble of possible universes, we require a statistical formalism to quantitatively describe the LSS observations we conduct which then allows us to compare observational data with theoretical models. At any position $\vec{x}$ with matter density given by $\rho(\vec{x})$ we can describe perturbations from a background mean density $\overline{\rho}$ by [160]

$$\delta(\vec{x}) = \frac{\rho(\vec{x}) - \overline{\rho}}{\overline{\rho}}. \tag{1.35}$$

The two-point correlation function is then defined by

$$\xi(r) \equiv \langle \delta(\vec{x}) \delta(\vec{x} + \vec{r}) \rangle \tag{1.36}$$

where $\vec{r}$ defines the separation between two points in the density field and due to homogeneity and averaging over all directions, ξ only depends on the modulus $r = |r|$. The correlation function is a measurement of the excess probability of objects being separated by $\vec{r}$ compared with a randomly distributed density field. It is often convenient to work with the Fourier space equivalent of the correlation function $P(k)$, referred to as the power spectrum which is related to the Fourier transformed over density field $\delta(\vec{k})$ by

$$\left\langle \delta(\vec{k}) \, \delta(\vec{k}') \right\rangle = (2\pi)^3 P(k) \, \delta_{\mathrm{D}}^3(\vec{k} - \vec{k}') \tag{1.37}$$

where δ_{D} is the Dirac delta and the Fourier transform of the over-density is given by [218]

$$\delta(\vec{k}) = \int \delta(\vec{x}) e^{i\vec{k}\cdot\vec{x}} \mathrm{d}^3\vec{x} \tag{1.38}$$

and the inverse is given by

$$\delta(\vec{x}) = \frac{1}{(2\pi)^3} \int \delta(\vec{k}) e^{-i\vec{k}\cdot\vec{x}} \mathrm{d}^3\vec{k}. \tag{1.39}$$

As with the correlation function, the power spectrum is only dependent on the modulus of the wavenumber $k = |k|$.

As outlined in the previous section 1.4.2, the majority of the matter content in the Universe is in the form of dark matter which due to its weakly interacting nature is invisible to telescopes. We therefore rely on tracers of the underlying matter which most commonly is emission from galaxies. In the linear regime we can describe the relation between these two tracer fields by a single linear factor referred to as the bias b_{g} [114] where

$$\delta_{\mathrm{g}} = b_{\mathrm{g}} \delta \quad \Rightarrow \quad P_{\mathrm{g}}(k) = b_{\mathrm{g}}^2 P(k). \tag{1.40}$$

and the 'g' subscript is indicative of the galaxy tracer method.

1.5.2 Structure Growth

The Cosmological Principle is a reasonable assumption on the largest scales but if we begin to examine the finer structure of the Universe, we find that its matter content is not homogeneous or isotropic and our very existence is owed to these subtle inhomogeneities. The theory of

inflation provides an initial source of perturbations in the primordial density field, the imprint of which we see in the CMB. After inflation dark matter perturbations begin to grow under the influence of gravity but at this early stage, baryonic matter remains tightly coupled to photons and it is not until the decoupling epoch that baryonic matter begins to also mimic dark matter and undergo gravitational collapse.

The framework to describe going from a primordial power spectrum encapsulating these initial density perturbations, to a late-Universe matter power spectrum begins with either inflationary physics or use of the CMB temperature anisotropies. Then using linear perturbation theory and solving the coupled Einstein-Boltzmann equations, a description of the evolution of matter perturbations is produced [17]. The Boltzmann equation, describing the collisions between the various constituents of the Universe and the Einstein equations, which track the evolution of cosmological perturbations, are too complex to solve analytically. This therefore, is generally done numerically with various code packages such as CAMB [125] or CLASS [124]. These solutions can provide a transfer function $T(k)$ for making predictions for a linear late-time matter power spectrum $P_{\rm lin}(k)$ based on a primordial power spectrum $P_{\rm p}(k)$ which is the power spectrum of primordial matter density fluctuations

$$P_{\rm lin}(k) \propto T^2(k) P_{\rm p}(k). \tag{1.41}$$

The transfer function can thus be thought of as a description for how modes in the matter density evolve i.e.

$$T(k) = \frac{\delta(k)}{\delta_{\rm p}(k)} \tag{1.42}$$

where $\delta_{\rm p}$ are the primordial density fluctuations and δ the late-time fluctuations we see in LSS. This is conventionally normalised such that $T(k \to 0) = 1$.

Matter-Radiation Equality

As introduced in Section 1.3.2, the Universe passes through different phases which can be defined by which constituent is dominating its content. The early Universe was dominated by radiation but since this scales as $\varepsilon_{\rm R} \propto a^{-4}$ and falls away faster than matter ($\varepsilon_{\rm M} \propto a^{-3}$), there will be a point where matter overtakes and becomes the dominant constituent. The point of matter-radiation equality i.e. $\varepsilon_{\rm M}(a_{\rm eq}) = \varepsilon_{\rm R}(a_{\rm eq})$ is an important point in the Universe's history and has a large influence on how structure grows. By assuming flatness and utilising the critical energy density (1.30), the first Friedmann equation (1.27) can be written

$$\left(\frac{\dot{a}}{a}\right)^2 = H_0^2 \frac{\varepsilon(t)}{\varepsilon_{\rm c,0}} \tag{1.43}$$

Then using equation (1.32) which for a flat universe ($\varepsilon_{\rm c,0} = \varepsilon_0$) is just $\varepsilon = \varepsilon_0 a^{-(1+3w)}$, we get an expression for the growth rate of the scale factor purely as a function of the equation of state parameter w

$$\dot{a} = \frac{H_0}{\sqrt{a^{1+3w}}}. \tag{1.44}$$

Figure 1.5 shows a plot for equation (1.44) for matter and radiation and it shows how in the

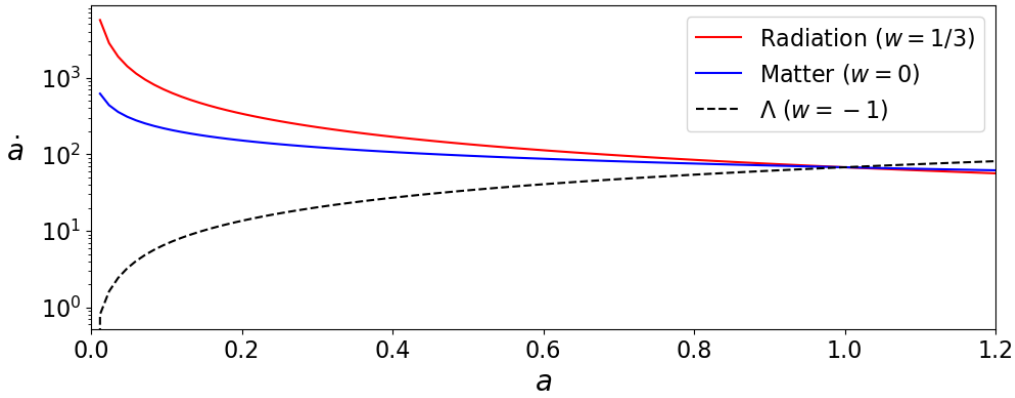

Figure 1.5: Evolution of the scale factor a for different constituents of the Universe. Calculated from using equation (1.44).

early Universe at low-a where radiation dominates, expansion occurs at a faster rate. This fast radiation-driven expansion prevents dark-matter perturbations from collapsing and suppresses growth. This means that modes which cross the horizon during the radiation dominated era ($k < k_{\rm eq}$) exhibit different behaviour to those that cross after this equality ($k > k_{\rm eq}$). See left-hand panel of Figure 1.6 for a plot of the matter power spectrum (blue line). Here $k_{\rm eq}$ is the scale of the horizon at matter-radiation equality occurring at approximately $k_{\rm eq} = 10^{-2} h{\rm Mpc}^{-1}$ and depends on the matter content of the Universe, $k_{\rm eq} \propto \Omega_{\rm M} h^2$. As an approximation, we find [66]

$$T(k) \sim \begin{cases} 1, & k \ll k_{\rm eq} \\ k^{-2}, & k \gg k_{\rm eq} \end{cases} \tag{1.45}$$

and therefore from (1.41), on large scales (small-k) the matter power spectrum increases as $P(k) \propto P_{\rm p}(k)$. Generally, for a simple single-field model of inflation, the primordial power spectrum is taken to be proportional to a power law $P_{\rm p}(k) \propto k^{n_{\rm s}}$, where $n_{\rm s}$ is the scalar spectral index. Data from the Planck satellite is consistent with a near scale-invariant $n_{\rm s} \sim 1$ power spectrum. The slight deviation from unity ($n_{\rm s} = 0.9649 \pm 0.0042$) is supportive of the inflationary paradigm [9]. However, on small scales (large-k), the power spectrum turns over since these modes were able to cross the horizon early and exist in the radiation dominated era for a long time, becoming damped due to the faster expansion driven by radiation domination.

Evidence of Dark Energy from LSS

Figure 1.6 shows the matter power spectrum on the left, run using `Nbodykit`[4] [92] and the Boltzmann solver package CLASS [124]. On the right is the transfer function produced using the Bardeen-Bond-Kaiser-Szalay (BBKS) fit [26][17] given by

$$T(x) = \frac{\ln(1+0.171x)}{0.171x}\left(1 + 0.284x + (1.18x)^2 + (0.399x)^3 + (0.490x)^4\right)^{-1/4} \tag{1.46}$$

[4] https://nbodykit.readthedocs.io

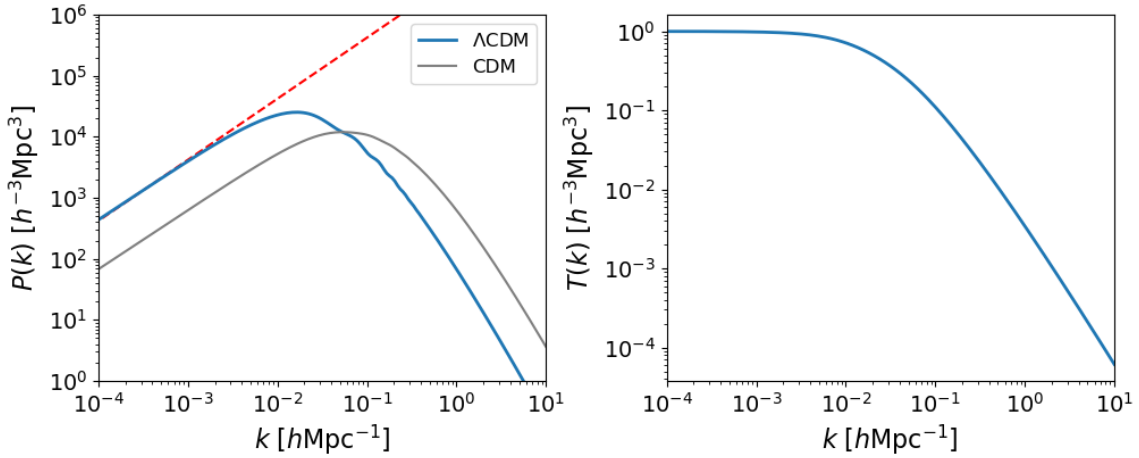

Figure 1.6: The linear matter power spectrum (left) for a ΛCDM cosmology (blue thick line) with Planck15 data [7] and a universe with no cosmological constant i.e. $\Omega_\Lambda = 0$ (thin grey line). Red dashed line shows the $P_k \propto k$ asymptotic relation predicted for a scale-invariant primordial power spectrum. The transfer function is shown on the right as predicted by fitting function (1.46) outlined in [26]. Both at redshift $z = 0$.

where $x \equiv k/k_{\rm eq}$. Figure 1.6 should immediately demonstrate how probing LSS can provide constraints on cosmological models. The thin grey line shows the power spectrum for a universe with $\Omega_\Lambda = 0$ and $\Omega_{\rm M} = 1$ which changes the position of the peak and therefore the position for the matter-radiation equality scale. Observational data in this context (see Figure 1.7), is consistent with a matter-radiation equality scale of around $0.01h{\rm Mpc}^{-1} < k_{\rm eq} < 0.02h{\rm Mpc}^{-1}$ and thus consistent with a ΛCDM universe.

1.5.3 Redshift Space Distortions

Typically in LSS surveys, the aim is to record coordinates of emission (e.g. from galaxies) that trace the underlying dark matter distribution. While obtaining angular coordinates is technically challenging, theoretically the process is fairly straightforward. However, obtaining a reliable radial distance is more involved. One option is to measure a luminosity distance (equation (1.14)) but this can only be done for objects for standardisable luminosity. Most commonly therefore, surveys rely on redshift and a well constrained distance-redshift relation to obtain this third coordinate for the tracer data.

However, if relying on redshifts, consideration must given to the inherent *peculiar velocity* of the galaxy caused by local density perturbations. The true radial velocity $\vec{v}$ of an object at true distance $\vec{r}$ can be split into a *Hubble flow* contribution and its peculiar velocity contribution

$$\vec{v} = H_0\vec{r} + \vec{v}_{\rm p}\,. \tag{1.47}$$

Here the peculiar velocity term $\vec{v}_{\rm p}$ is more dominant for galaxies closer to us due to the Hubble flow contribution $H_0\vec{r}$ being small for nearby galaxies. Since these peculiar velocities are correlated to density perturbations, any attempted measurement of a density field using redshift will therefore be distorted, known as the *Kaiser* effect [115].

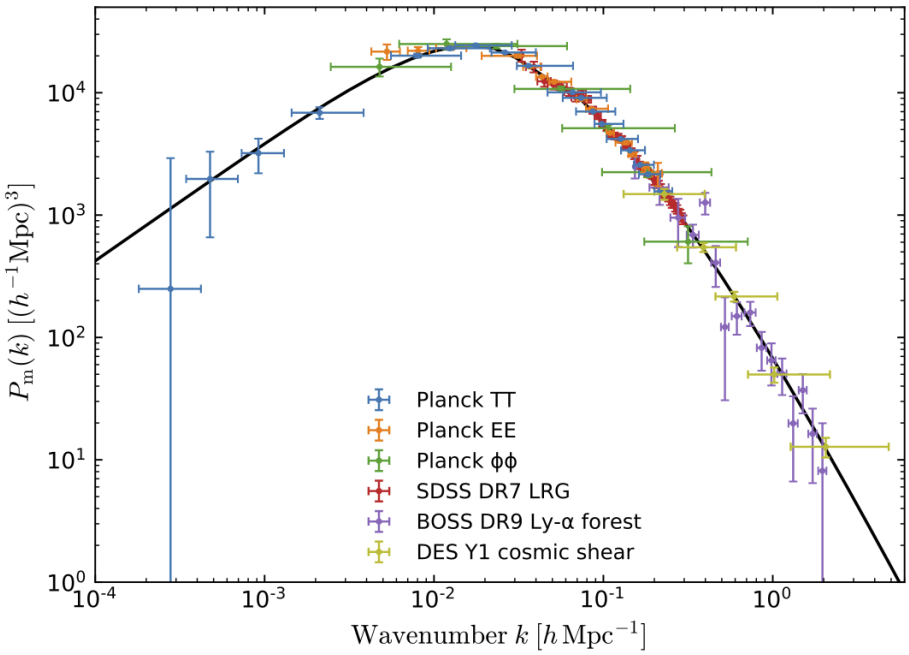

Figure 1.7: Summary of observational data constraining the matter power spectrum at $z = 0$ and its agreement with a theoretical linear power spectrum (black line) predicted by ΛCDM. Plot produced by [9].

The impact this has on the density field measured in redshift space is that on large scales, objects tend to fall in to high density regions which squashes the density field and the clustering amplitude becomes stronger along the LoS as shown in the left diagram of Figure 1.8.

The conclusion from this is that when making observations in redshift space, corrections need to be made to account for this effect. For example, the galaxy power spectrum introduced in equation (1.40) is amended to [115][17]

$$P_{\rm g}(k,\mu) = b^2\left(1+\beta\mu^2\right)^2 P(k) \tag{1.48}$$

where the term μ is introduced to account for the anisotropic effect of Redshift Space Distortions (RSD) and is defined as the cosine of the angle between the LoS and the wave vector $\vec{k}$. As one would expect, for modes perpendicular to the LoS we have $\mu = 0$ and we recover the isotropic power spectrum. Equation (1.48) also depends on $\beta = f/b$, where f is the growth rate and defined as $f \equiv {\rm d}\ln\delta/{\rm d}\ln a$ and approximated by $f \sim \Omega_{\rm M}(z)^{\gamma}$. Here γ is the growth rate index with $\gamma \sim 0.545$ for ΛCDM [160] and $\Omega_{\rm M}(z) = H_0^2\Omega_{\rm M,0}(1+z)^3/H(z)^2$ [186]. For a full derivation of the above I refer the reader to the review in [91].

RSD effects are also apparent in the non-linear regime on small scales. As shown by the diagram on the right in Figure 1.8, at the centre of a density peak the peculiar velocities can potentially be greater than the velocity caused by the Hubble flow and this has the effect of turning structures inside out and stretching them along the LoS. These stretched structures that we observe are called the fingers of god [109].

To investigate the effect of RSD on the observed matter density field it is useful to expand the

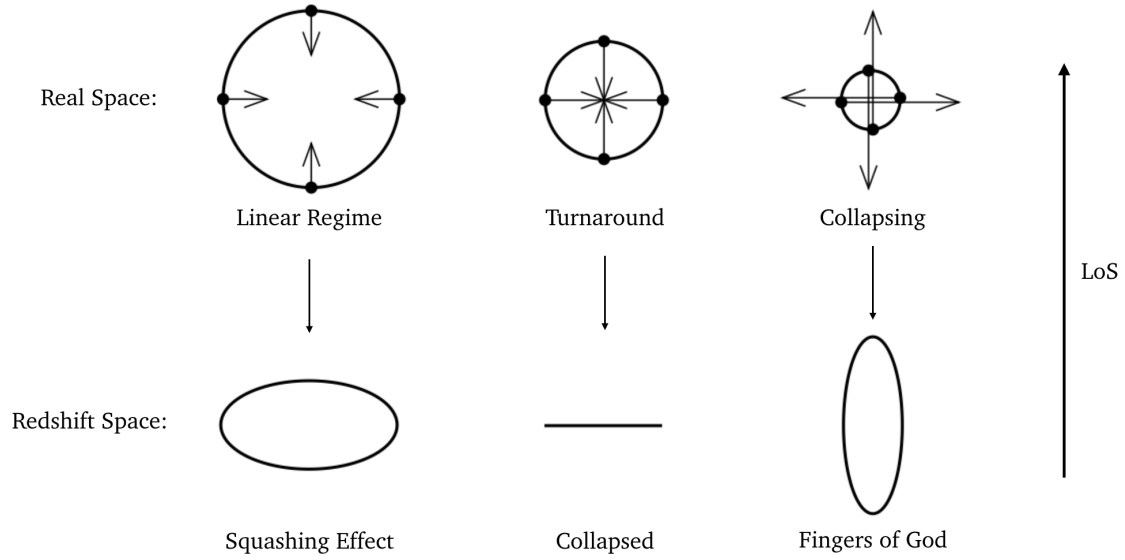

Figure 1.8: The effects of redshift space distortions on density fields observed in redshift space. Image adapted from [91].

anisotropic power spectrum into its multipoles $P_\ell(k)$ using the Legendre polynomials $\mathcal{L}_\ell$ [186]

$$P_{\rm g}(k,\mu) = \sum_\ell P_{g,\ell}(k)\mathcal{L}_\ell(\mu) = P_{g,\ell=0}(k)\mathcal{L}_0(\mu) + P_{g,\ell=2}(k)\mathcal{L}_2(\mu) + P_{g,\ell=4}(k)\mathcal{L}_4(\mu). \tag{1.49}$$

We only need to include the monopole, quadrupole and hexadecapole ($\ell = 0,2,4$ respectively) since equation (1.48) has terms no higher than μ^4 and is an even function of μ which kills the odd multipoles. The Legendre polynomials we need are $\mathcal{L}_0 = 1$, $\mathcal{L}_2 = (3\mu^2-1)/2$ and $\mathcal{L}_4 = (35\mu^4 - 30\mu^2 + 3)/8$ and each multipole is given by

$$P_\ell(k) = \frac{2\ell+1}{2}\int_{-1}^{1} {\rm d}\mu P_{\rm g}(k,\mu)\mathcal{L}_\ell(\mu). \tag{1.50}$$

Plugging in these and the Kaiser power spectrum from (1.48) we derive equations for the three multipoles

$$P_{{\rm g},\ell=0}(k) = \left(1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2\right) b^2 P(k), \tag{1.51}$$

$$P_{{\rm g},\ell=2}(k) = \left(\frac{4}{3}\beta + \frac{4}{7}\beta^2\right) b^2 P(k), \tag{1.52}$$

$$P_{{\rm g},\ell=4}(k) = \frac{8}{35}\beta^2 b^2 P(k). \tag{1.53}$$

See Appendix A.1 for a more detailed derivation of these multipole moments of the power spectrum.

By choosing realistic values for the monopole pre-factor in equation (1.51), we can examine the impact of RSD on the power spectrum. Using $b = 2$ and $f = 0.774$ we get $P_{{\rm g},\ell=0}/P_{\rm g} \sim 1.28$ which is clearly a significant factor. The quadrupole and hexadecapole are therefore seen as 'smoking guns' for RSD and would be measured as zero if no RSD existed.

Using RSD

Since RSD are a measurement of the cosmological velocity field which is determined by gravitational potential, they are a source of information on density. They can also be used to measure the growth rate f which is of great interest to cosmologists too. As mentioned, it is generally considered a good approximation that [120]

$$f(a) \sim \Omega_M^{\gamma}(a). \tag{1.54}$$

This tells us that the growth rate is largely determined by the matter content Ω_{M} of the Universe which makes intuitive sense since Λ is uniformly distributed and should not act to directly effect perturbations in the matter density. It is worth pointing out that if assuming flatness, then we have $\Omega_{\mathrm{M}} = 1 - \Omega_{\Lambda}$ and therefore the matter density must fall for an increasing Ω_{Λ}. The effect then from an increasing cosmological constant is a therefore a slowing growth rate.

We can also use the growth rate as an effective test of modified gravity theories. We find $\gamma \sim 0.55$ for general relativity and ΛCDM cosmological parameters, however the growth index changes if one originates the accelerated expansion to modifications of the general relativity equations [90].

Perhaps most importantly for cosmological surveys, if one assumes a well constrained growth rate parameter f, RSD can be used by taking the ratio of the multipoles as a way of measuring β i.e. [206]

$$\frac{P_2(k)}{P_0(k)} = \frac{\frac{4}{3}\beta + \frac{4}{7}\beta^2}{1 + \frac{2}{3}\beta + \frac{1}{5}\beta^2}. \tag{1.55}$$

Thus from this, a good approximation can be made for biases since $\beta = f/b$.

1.6 Summary

In this chapter I have outlined the theoretical framework for the most successful model we have to explain the Universe's beginning stages, evolution and fate along with the best supporting evidence. The concordance (ΛCDM) model is one which begins with Einstein's theory of general relativity and with a few well-reasoned assumptions, a simple model is crafted whose parameters can be constrained to match observations.

We are far from a complete theory though. While evidence from SNeIa, CMB, BAO and LSS all agree nicely with a ΛCDM model and indicates that we have a good idea for how the Universe works, it leaves us with the uncomfortable conclusion that we can not explain what ~95% of it is made of. The so-called 'dark sector' means the success of this model rests on being able to find missing matter (i.e. a dark matter particle) which is consistent with the best theories of particle physics and QFT. Furthermore, and arguably more of a challenge, is attempting to construct a mechanism for the origins of dark energy. If rigidly sticking to the ΛCDM model, then this is best explained as a cosmological constant with some exotic fluid which increases in content as the Universe expands or put differently, has a non-changing density i.e. $\dot{\varepsilon}_{\Lambda}(t) = 0$. However,

attempting to explain this in-line with QFT as the energy density of the vacuum results in some of the biggest tensions[5] seen in science!

The future of cosmology is therefore tasked with further testing this model. Precision observational cosmology will allow for this and will simultaneously test alternative theories by restricting the parameter space available. This thesis will mostly examine how we can use probes of LSS to contribute to these objectives. As precision cosmology progresses our ability to probe the non-linear regime of LSS becomes more achievable. While these non-linear scales are extremely useful and contain a wealth of information, this thesis will be largely focussed on the linear regime. This is largely down to the fact that a survey of neutral hydrogen using an intensity mapping technique with a single-dish, mostly probes linear scales (as outlined in the following chapter).

Under a linear approximation we can use techniques such as BAO probes to constrain cosmological parameters or RSD to understand growth histories. In the linear regime, modelling is simplified too allowing for an easier comparison between data and theory. The conventional approach to obtaining this observational data relies on resolving galaxies and measuring their redshift to obtain 3-dimensional coordinates for many point-like tracers of the underlying matter density. However, accurately obtaining redshifts for the millions of galaxies required to beat down shot-noise is a time-expensive process. As I will discuss in the following chapter, alternative survey methods exist which allow a more complete measurement of the underlying cosmological structure thus improving statistical errors and providing potential to further constrain cosmological parameters.

[5] Depending on the chosen method of calculation this can be as high as $\varepsilon_{\text{vacuum}}/\varepsilon_{\Lambda} \sim 10^{120}$. [222]

CHAPTER 2

HI (21CM) INTENSITY MAPPING

Hydrogen is the most abundant element in the Universe comprising around 75% of its baryonic mass. It exists in various chemical forms but the one of most interest for this thesis is the neutral hydrogen atom which is hydrogen's most simple atomic form. It consists of one positively charged proton and one negatively charged electron whose charges exactly cancel to give it electric neutrality [223]. Isolated neutral hydrogen is often referred to as atomic hydrogen or more concisely HI (pronounced H-one).

HI has a quantum structure whereby its single electron can exist at two hyperfine levels in its 1*s* ground state. This hyperfine structure relates to the spin alignment between the electron and proton. When the spin of the electron is parallel with the spin of the proton, the hydrogen atom is in a slightly higher energy state ($\sim 5.87\,\mu$eV difference) than when the spins are anti-parallel. The spontaneous un-alignment of spin will therefore produce a quantized photon with this energy which carries a frequency of ~1420MHz and a wavelength ~21cm. Hence, the radiation from this process is referred to as 21cm emission. This change in energy state is extremely rare and the mean lifetime of the excited state is around 10^7 years [76][225]. Fortunately, the abundance of HI in the Universe, at both late and earlier epochs, is sufficient for this redshifted 21cm radiation to be a significant signal.

The story of hydrogen's history is one which closely reflects the story of our Universe's history as a whole. In this chapter I will discuss this history which will lay the foundation for a discussion on why 21cm emission from HI is such a useful tool for cosmologists. I will then introduce the specific techniques of interest in my research, which look to map 21cm signals from unresolved galaxies in the low-redshift, late-Universe.

2.1 Cosmic History of Hydrogen

The richness of information hydrogen signals contain can only truly be appreciated by understanding the timeline of events which explain its abundance and location at various epochs

of the Universe's history. Outlined in the following sub-sections is a basic overview of the key points in the Universe's timeline which affect the evolution of HI .

2.1.1 The Early Universe

For a full review on the early Universe, I refer the reader to [89][128].

- **Big-Bang/Inflation** ($t = 0$): Shortly after the poorly understood big-bang beginning, the Universe underwent a rapid phase of spatial expansion, known as *inflation*. Within the first second, quarks begin to join and form protons and neutrons, thus technically forming the first Hydrogen-1 (protium) isotope consisting of a single proton [126].

- **Nucleosynthesis** ($10\text{s} < t < 10^3\text{s}$): The process where protons and neutrons could fuse to forge the first light elements. However, the Universe is still too hot and dense for electrons to be captured and form neutral atoms thus the Universe remains an ionised plasma.

- **Matter-Radiation Equality** ($t \sim 10^{12}\text{s}$): As discussed in the previous chapter, the epoch at which the abundance of matter and radiation is equal i.e. $\varepsilon_{\text{M}} = \varepsilon_{\text{R}}$. Matter begins to dominate but the Universe remains too hot for atomic nuclei to form neutral atoms.

- **Decoupling/Recombination** ($t \sim 10^{13}\text{s}$): At redshift $z \sim 1100$ the Universe expanded and cooled to $\sim 4000\text{K}$ allowing for free electrons to bind with atomic nuclei, thus forming the first neutral hydrogen (HI) atoms. This is also the epoch where photons decouple from matter, putting an end to the constant Thompson scattering, allowing photons to stream away. These early photons are what are visible today in the CMB.

2.1.2 Dark Ages & the Epoch of Reionization

After these early Universe processes, HI exists relatively unchanged throughout a period which is known as the *dark ages*. During this time, the formation of stars and galaxies has not yet occurred and HI is one of the rare sources of new signals through its spontaneous 21cm emission. The challenges of detecting these faint 21cm signals from the Dark Ages are large but remain one of the few windows of discovery into this poorly understood epoch.

The ending of the dark ages coincides with what is referred to as the *epoch of reionization* (EoR). The precise point in the Universe's past when the 'first light' from stars first began to shine is unknown [172][42]. At around 400Myr after the big-bang, UV-radiation from these cosmic dawn stars began reionizing the Universe. This initially formed ionized bubbles around these first objects which grew in size as the ionizing radiation extended its reach. As this process continued, supplemented by newly forming galaxies and UV emitting stars, the gas between galaxies, referred to as the *intergalactic medium* (IGM), came to be dominated by ionized hydrogen (HII). See [137][237] for a full review of the epoch of reionization which is a very active area of research with the potential to answer questions regarding the formation and composition of early stars and galaxies, as well as helping understand early structure formation.

2.1.3 HI in the Post-Reionization Universe

The only neutral hydrogen to remain at the end of the EoR was that locked away in the *interstellar medium* (ISM) of galaxies, self-shielded against the ionizing UV radiation which is unable to penetrate the denser environments of massive galaxies [25]. This leaves a situation in the post-reionization era where the only 21cm signals from HI will come from within galaxies which are biased tracers of the underlying large scale cosmic structure. If this hypothesis is correct, then HI is an excellent potential tool for cosmologists aiming to investigate the late Universe.

Due to the practical challenges of detecting 21cm emission (discussed in section 2.2), our current HI observational data is limited. Therefore based on observational evidence alone, it is difficult to claim with certainty that HI is not found in random large clumps outside galaxies in the late Universe which would systematically bias LSS measurements which used HI as a tracer. Fortunately the sophistication of computer simulated models is growing rapidly [200][197][81] and a large amount can be learned from their output. Much of our understanding on the exact distribution and abundance of HI within the post-reionization Universe therefore comes from theoretical modelling with computer simulations [168][212]. It is understood from such simulations that in the post-reionization Universe, the majority of HI resides within dark matter halos [213]. These halos are defined as gravitationally bound regions of dark matter into which baryonic matter collapses and galaxies form [219]. More will be discussed on the relationship between HI and dark matter halos in section 2.3.3.

Given these findings from simulations, HI should be a reliable tracer of underlying structure in the late Universe. Therefore the aim is to begin systematically detecting and mapping it on the largest scales.

2.2 Mapping Unresolved 21cm Emission

Theoretically HI is observable from Earth out to a redshift of $z \sim 50$. Above this, the relevant observed wavelengths from 21cm emission, which are redshifted to around 10m, are unable to penetrate our atmosphere due to its ionosphere. This thesis is interested in using HI to explore the late-Universe and I will therefore be focussing on detecting these signals in the post-reionization epoch. Here the majority of HI resides inside dark matter halos in dense clouds referred to as damped Lyman-α systems which are embedded in galaxies. The 21cm emission from HI is unfortunately relatively weak and therefore using it to detect and resolve enough galaxies for precision cosmology is an enormous challenge beyond the capabilities of current telescopes. To date only a few resolved galaxies at low redshifts ($z < 0.2$) have been detected using HI [82] and it is likely we will have to wait for the Square Kilometre Array (SKA)[1] [189][23] until a HI galaxy survey is competitive with a conventional optical galaxy survey [234][187].

Fortunately, a different approach can be adopted whereby the combined, unresolved HI emission is measured on large angular scales; this is known as *intensity mapping* [29][47][167][214]. While information on small scale density fluctuations is lost under intensity mapping, scales

[1] skatelescope.org

of interest for probing large scale structure and phenomena such as BAO are measurable. It is argued that this approach is similar to that of a conventional optical galaxy redshift survey which observes photons from billions of stars but reduce it to a single galaxy point. Similarly, albeit on a grander scale, intensity mapping takes HI emission from multiple sources and reduces it to a single broad pixel. Much like data from the CMB, the result from HI intensity mapping is therefore a large scale map of fluctuations whose statistical distribution carries a wealth of cosmological information. Unlike the CMB however, intensity maps can be 3-dimensional providing the added advantage of a signal which is a function of redshift.

Conventional optical redshift surveys have led the way for the last two decades in our understanding of LSS [24][164][104], however intensity mapping potentially has some distinct advantages. Firstly, galaxies in an optical survey are only entered into a catalogue if they are detected with high confidence thus throwing away some of the emission. Conversely, intensity mapping integrates radiation from all galaxies down to the faintest emitters and it is expected that several galaxies will contribute to each pixel thus providing a statistically strong signal. Secondly, an optical survey needs to rely on high signal-to-noise detection to conduct spectroscopy and obtain high precision redshift measurements. By their very nature, intensity mapping surveys are spectroscopic experiments but the HI signal, which falls at frequencies of 1420MHz and below due to redshift, is an isolated transition and hence robust against line confusion. This fortunate advantage, coupled with the high frequency resolution of modern radio telescopes, allows intensity mapping surveys to observe large swathes of cosmic structure faster than their optical counterparts with excellent redshift resolution.

It is worth noting that there are further examples of emission which can be utilized with line-intensity mapping. In this thesis I focus solely on 21cm emission from HI, which is the most useful for large scale cosmology. However, as an example, emission from rotational carbon monoxide (CO) transitions, the [CII] fine-structure line, or the Lyman-α line, are signals that can be 'intensity-mapped'. For further reading, I suggest the review in [118] which discusses these additional line-intensity mapping strategies and the prospects they hold for the wider astrophysics community.

2.2.1 Intensity Response for Radio Telescopes

The signal captured by the telescope antenna is received as an intensity pattern, which even for a set *pointing* (i.e. fixed sky coordinates), has some angular dependence. Figure 2.1 shows a sketch of this intensity pattern where the raw signal (shown by the back solid line) has been modelled by $P(\theta) = (D/\lambda)^2 \mathrm{sinc}^2(\theta D/\lambda)$ which is an approximation of the radiation pattern received by a uniformly illuminated one-dimensional aperture [56]. For this plot I used $D = 15$m for the receiver dish diameter and $\lambda = 21$cm for radiation wavelength which approximately emulates the SKA [23] at very low redshift. The majority of the power is concentrated in the main central lobe (referred to as the primary *beam*) and the angular resolution of a radio telescope is defined as the full-width-half-maximum (FWHM) of this beam, whose angular size is shown by the shaded regions in Figure 2.1.

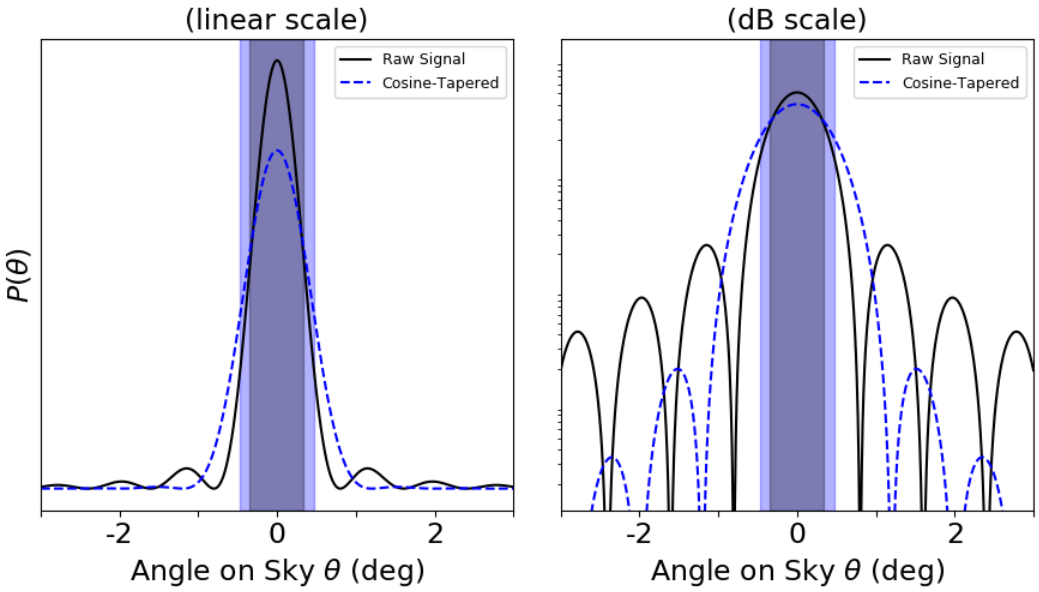

Figure 2.1: Approximate radiation patterns for a uniformly illuminated one-dimensional aperture. Main central lobe i.e. the primary beam, is centred at 0° and the beam width is defined as the FWHM of this central lobe (show by the central darker shaded region). Side-lobes are also visible with first side lobe peaking around ±1.2°. The side-lobes can be mitigated by tapering but this broadens the beam width as shown by the lighter shaded region which is the FWHM for the central lobe of the cosine tapered pattern. Side-lobes after tapering are only still visible on the decibel scale which is essentially equivalent to a $\log_{10}$ scale in this example.

In reality the beam pattern can usually be calibrated by measuring the telescope's response to a bright, isolated point signal. However, there is also the additional complication of *side-lobes* as shown in Figure 2.1. These refer to the signals that are detected outside the primary beam usually caused by diffraction. A large fraction of received power coming from the side-lobes results in a degradation in the telescopes pointing capability i.e. how well it can determine the location of a radiation source [225]. The side-lobe responses are often mitigated by tapering at the edge of the aperture which broadens the beam [72]. This is also shown in Figure 2.1 as the blue dashed line. While this does an excellent job at minimising side-lobe contributions, it broadens the primary beam and thus deteriorates the telescope's resolution.

Generally, for a signal with observed wavelength λ (i.e. $\lambda = (1+z)21\text{cm}$) incident on a circularly symmetric aperture of diameter D_{dish} with uniform illumination, its beam width in radians can be approximated by

$$\theta_{\text{FWHM}} = 1.22\frac{\lambda}{D_{\text{dish}}}. \tag{2.1}$$

Often the factor 1.22 is replaced with a generalised scalar a where $a \sim 1.0-1.3$ depends on the illumination of the aperture and the amount of tapering carried out to mitigate contamination from the side-lobes [22]. For a Gaussian shaped main beam, the solid angle Ω_{pix} covered per pointing is related to the beam width and is given by [225]

$$\Omega_{\text{pix}} = 1.133\theta^2_{\text{FWHM}}. \tag{2.2}$$

For further literature on the intensity response from a radio telescope, I refer the reader to [107][56] which provide full derivations of the models used in Figure 2.1 and for equations (2.1) and (2.2) which are beyond the scope of this thesis.

2.2.2 Interferometer or Single-Dish

From equation 2.1, it is clear that to maximise resolution, the dish size (or more generally the *baseline*) needs to be as large as possible. Arrays of radio receivers in the form of interferometers are now common in radio astronomy as a way to maximise baselines without the impracticality of building large single-dishes. For an interferometer, the baseline D is given by the maximum distance between two antennas in the array. In extreme cases this can allow radio telescopes spanning continents to link together and obtain impressive resolutions. This was shown in the recent work by the Event Horizon Telescope which allowed 8 radio receivers to image the environment around a blackhole 55 million light-years away at the centre of Messier 87 [8]. In high resolution radio astronomy interferometers are therefore the preferred observational strategy and operational single-dish radio telescopes are more rare due to the dish size needed for a competitive survey [112]. As an example, for observations of 21cm radiation with sub-arcminute resolution, a single-dish requires a diameter greater than 880m. Currently the largest *steerable* radio telescope is the Green Bank Telescope (GBT)[2] which has a 100m diameter dish.

However, in an intensity mapping approach, resolutions around the degree scale are acceptable since they still resolve the cosmological scales of interest e.g. the BAO scale at $110h^{-1}$Mpc [30] ($\sim 2.8^\circ$ at $z = 1$). Thus single-dish telescopes still have a place in cosmology and new single-dish intensity mapping experiments are being commissioned e.g. the BAO in Neutral Gas Observations (BINGO)[3] telescope [28] and the Five hundred meter Aperture Spherical Telescope (FAST)[4] [37]. Furthermore, future radio telescopes which are primarily interferometers are making plans to operate in *single-dish mode* where each receiver in the array conducts intensity mapping observations. Both the SKA and its pathfinder MeerKAT [190][169] have plans to operate in single-dish mode and offer a rapid way to survey large volumes of sky. Pre-existing single-dish telescopes have already contributed enormously to our understanding of intensity mapping. While not being purpose built for this approach, some have still managed to make the first detections of the cosmological signal using a 21cm intensity mapping approach. The GBT and the Parkes Observatory[5] have both have made intensity mapping detections [46][162][135][205][18], relying on cross-correlations with an overlapping optical galaxy redshift survey (discussed in more detail in section 2.2.4). Outlined in Table 2.1 are the main radio telescopes with connections to HI intensity mapping with some basic specifications.

[2] greenbankobservatory.org
[3] bingotelescope.org
[4] fast.bao.ac.cn
[5] parkes.atnf.csiro.au

Telescope	Location	Dish Diameter	Redshift Range	First Light
Parkes[201]	New South Wales, Australia	64m	$0.06 < z < 0.10$	1961
GBT [46]	West Virginia, USA	100m	$0.6 < z < 1.0$	2000
FAST [37]	Guizhou, China	500m	$0.0 < z < 0.49$	2016
MeerKAT [169]	Karoo, South Africa	13.5m	$0.00 < z < 1.45$	2016
BINGO [28]	*Serra do Urubu, Brazil*	40m	$0.13 < z < 0.48$	*~2020*
SKA-MID [189]	*Karoo, South Africa*	15m	$0 < z < 3$	*~2025*
CHIME [145]	British Columbia, Canada	1024×N/A*	$0.8 < z < 2.5$	2017
HIRAX [146]	*Karoo, South Africa*	1024×6m	$0.8 < z < 2.5$	*~2022*

*CHIME is not a conventional dish design and instead has four semi-cylinders (100m long, 20m wide) populated with 1024 receivers

Table 2.1: Some important examples of radio telescopes with links to HI intensity mapping arranged chronologically by first-light. The top section are all single-dish receivers (or will be used in single-dish mode for intensity mapping). The bottom section are interferometers where the number of receivers in the array is indicated in the Dish Diameter column. Italicised means they are not yet operational at time of writing.

2.2.3 Telescope Systematics

In order for the field of intensity mapping to mature into a competitive probe of cosmology, a detailed understanding of the source of measurement uncertainties is required. Unfortunately HI line emission, even when integrated over many galaxies is relatively faint meaning current intensity mapping observations are prone to systematic limitations; the main of which are introduced below.

- **Thermal Noise**:

  The dominant contribution to the overall noise comes from *thermal* noise (also often referred to as *instrument* noise). This is caused by the thermal motion of electrons in the resistors which produce a current [94]. Since this current has a mean value of zero and is uncorrelated with received signals it can be accurately modelled as Gaussian white noise [225]. The amplitude of this thermal noise contribution is defined by $\sigma_{\rm noise}$, the standard deviation of the Gaussian distribution, which is dependent on characteristics of the telescope and survey parameters. For a single-dish intensity mapping survey, $\sigma_{\rm noise}$ can be modelled by [11]

  $$\sigma_{\rm noise} = T_{\rm sys}\sqrt{\frac{4\pi f_{\rm sky}}{\Omega_{\rm pix} N_{\rm dish}\, t_{\rm obs}\delta\nu}}\,. \tag{2.3}$$

  Here $T_{\rm sys}$ is the total system temperature for the particular telescope (discussed in [23] and [189]), $f_{\rm sky}$ is the fraction of sky covered by the survey, $\Omega_{\rm pix}$ is the pixel solid angle (see equation (2.2)), $N_{\rm dish}$ the number of dishes in the survey, $t_{\rm obs}$ is total observation time and $\delta\nu$ is the frequency bandwidth. See [171][43] for a more complete discussion on thermal noise.

- **Red** $(1/f)$ **Noise**:

  Radio telescopes can also be further contaminated by gain fluctuations from amplifiers which create noise which is correlated across all frequency channels [36]. The impact of these

fluctuations is usually simulated in the frequency domain using a $1/f$ power spectrum, hence the name $1/f$ noise. While this is a challenge for cosmology with single-dish intensity maps, it is predicted that since this noise will be strongly correlated along the frequency direction, it should be possible to remove the noise in a similar way to 21cm foreground cleaning methods (discussed at length in later Chapters). See [94][50] for a complete discussion of $1/f$ noise in the context of HI intensity mapping.

- **Radio Frequency Interference (RFI) Noise**:

  Unfortunately for the purposes of radio astronomy, much of our planet is awash with radio signals which are used in our communication technologies. These human-made signals interfere with the cosmological ones we are aiming to detect. This is a particular problem for single-dish intensity mapping where discriminating between emission picked up within the main beam or the side-lobes is difficult. Forecasts for an SKA-like single-dish intensity mapping experiment have shown that RFI emission (in particular from Global Navigation Satellite Systems) will exceed the expected HI signal at all frequencies within SKA Band 2 ($0 < z < 0.5$) [23], therefore careful consideration and removal/mitigation is required. For a more detailed discussion, I refer the reader to [93].

- **Beam Smoothing**:

  As discussed, the primary beam from a radio telescope targeting 21cm signals can often be broad unless huge baselines are employed to compromise for the relatively large wavelengths. The effects of this wide beam can be well-modelled in cosmological simulations by convolving the simulated data with a symmetrical 2-dimensional Gaussian kernel whose FWHM matches that of the telescope beam (see equation (2.1)). This smoothing of data is considered a systematic because it leads to a loss of information contained in small perpendicular modes and as we probe higher redshifts, the number of resolvable scales decrease.

- **Foregrounds**:

  Another large systematic involved in HI intensity mapping comes from natural signals also present in our Universe. Most astrophysical sources emit some form of radio radiation [56] and due to the inherently faint HI signal it is easy for non-cosmological signals in the $1420\text{MHz}/(1+z)$ ranges to dominate. We call such dominant signals *foregrounds*. Foregrounds will be discussed in Chapter 3 and are a main topic of research in Chapter 4, so I will leave a more detailed discussion of them to these chapters.

2.2.4 Cross-Correlation with Optical Surveys

One could understandably ask the question, why invest time and funding in HI intensity mapping experiments when we could just invest more into improving optical redshift survey efficiency? After all, we already have an in-depth knowledge of their technology and we know they can provide large contributions to precision cosmology. In a way this has already been answered earlier in this section when I outlined some advantages of intensity mapping in terms of their

excellent redshift resolution, rapid survey time and a more inclusive use of the full signal even from faintest emitters. However, even ignoring these advantages, one could still argue that it is beneficial to pursue this new approach.

Having a different method for surveying the large scale cosmic structure means benefits can be gained from cross-correlations. An excellent advantage of cross-correlations is that major systematics that affect one probe may not necessarily affect the other. Take a basic example where the density fluctuations in the observed data are comprised of a true cosmological signal $\delta^{\rm cos}$ and an additive systematic component i.e. noise $\delta^{\rm sys}$. For both a HI intensity map survey and an optical galaxy survey this can be formalised by [117]

$$\delta_{\rm g} = \delta_{\rm g}^{\rm cos} + \delta_{\rm g}^{\rm sys}, \tag{2.4}$$

$$\delta_{\rm HI} = \delta_{\rm HI}^{\rm cos} + \delta_{\rm HI}^{\rm sys}. \tag{2.5}$$

When we cross-correlate the HI and galaxy data we get a product of these terms and we expect the cosmology and systematic cross-terms to be uncorrelated and drop out. However, unlike an auto-correlation where one would expect the systematic terms to correlate and contaminate the measurement, in a cross-correlation it is likely the systematics $\langle \delta_{\rm g}^{\rm sys} \delta_{\rm HI}^{\rm sys} \rangle$ will be uncorrelated and also drop out, as shown below;

$$\begin{aligned} \langle \delta_{\rm g} \delta_{\rm HI} \rangle &= \left\langle \delta_{\rm g}^{\rm cos} \delta_{\rm HI}^{\rm cos} \right\rangle + \left\langle \delta_{\rm g}^{\rm cos} \delta_{\rm HI}^{\rm sys} \right\rangle + \langle \delta_{\rm g}^{\rm sys} \delta_{\rm HI}^{\rm cos} \rangle + \langle \delta_{\rm g}^{\rm sys} \delta_{\rm HI}^{\rm sys} \rangle \\ &= \left\langle \delta_{\rm g}^{\rm cos} \delta_{\rm HI}^{\rm cos} \right\rangle. \end{aligned} \tag{2.6}$$

The more differing the telescopes are that are in cross-correlation, the less likely it will be that their systematics will be correlated.

Given that future surveys are on course to gather orders of magnitude more data than previously obtained, it is inevitable that precision cosmology will see a reduction in statistical uncertainty. This makes the attention on systematic uncertainty all the more paramount as it is likely that it could begin to dominate the error budget [119]. This highlights one of the appeals of HI intensity mapping for future precision cosmology, since its radio telescopes are fundamentally different in design to a conventional optical telescopes and therefore provide an assured way of improving constraints on cosmological measurements.

Multiple tracers of underlying LSS also produce opportunities to limit *cosmic variance* which is the variance caused by the survey's finite volume. This so-called *multi-tracer* approach works because two different tracers with different biases will have a ratio independent of the underlying dark matter field they trace [193][6]. Thus cosmic-variance caused by stochasticity in the particular realization of the dark matter field we observe can be limited because we are only concerned with effects on the bias of the dark matter tracers, not on dark matter itself. The effectiveness of this technique depends on a number of factors such as the ratio of the different biases and their non-linearity [86]. Intensity mapping is an ideal candidate to be used with optical LSS surveys in this multi-tracer approach and forecasts into the benefits have been produced [12][226].

First Detections Using Cross-Correlations

While cross-correlations offer tantalising possibilities for limiting systematics and constraining biases in future surveys, they also offer an excellent way of conducting some of the first successful HI intensity mapping detections of cosmological structure, thus vindicating the intensity mapping method. We would expect all the systematics outlined in section 2.2.3 to be uncorrelated with optical survey systematics (e.g. stellar contamination, galactic extinction, photometric calibration etc. [182]) and therefore be mitigated in a cross-correlation measurement. As the field of HI intensity mapping matures, some of its earliest detections have relied on cross-correlations with optical LSS data.

The first statistically significant detection came in [162] which used data from the Parkes Telescope survey HIPASS [27], which was cross-correlated with the Six Degree Field Galaxy Redshift Survey (6dFGS) [113] observed with the Anglo-Australian Telescope (AAT). While most of the signal came from correlations along the LoS, it was still an early indication that HI is a biased tracer of LSS since it is correlating with galaxies which are a known biased tracer of LSS. The HIPASS survey was originally a HI galaxy survey but in this work they used the 21cm spectral intensity data to claim a detection using the 21cm intensity field [162]. However, some could reasonably argue that this approach is not the same as the 'true' intensity mapping technique which involves mapping the whole patch of sky and integrating all 21cm emission, rather than just that from targeted galaxies.

The GBT began its contributions to early detections around a decade ago and arguably provided the first 'proper' intensity mapping detection. Since the GBT is able to probe lower frequencies than Parkes, their observations were made at higher redshifts ($z \sim 0.8$). Work from [46] presented results from the GBT intensity map cross-correlations with DEEP2 [62] optical redshift data obtained with the twin 10m Keck telescopes in Hawaii. This work began contributing measurements of the neutral gas density $\Omega_{\rm HI}$ which is particularly hard to measure at around redshifts of $z = 1$ (as later shown in Figure 2.3 where very few data exist at $z \sim 1$). A further detection was made using the GBT data in [135] which was cross-correlated with galaxies in the WiggleZ Dark Energy Survey [69] observed using the AAT at a redshift range of $0.6 < z < 1.0$ over 41deg.sq. These same GBT observations were also auto-correlated in [205] to provide an upper bound on the 21cm signal. More recently in [18], intensity maps from Parkes made a lower redshift detection at $0.057 < z < 0.098$ cross-correlated with earlier data from the AAT's 2dF galaxy survey [55].

Plans are in place to continue adding to this list of successful intensity mapping detections of cosmological structure and with new purpose built telescopes in operation (Table 2.1) the rate of observations is only likely to increase.

2.3 HI Cosmological Formalism

Here I introduce a framework which will be used throughout the thesis for relating the 21cm signals received from intensity mapping surveys to the cosmological structure which they trace.

This formalism is drawn from literature on this topic e.g. [4][25][28][43][14][36][170][229].

2.3.1 Characterising the 21cm Signal

Radio telescopes observe sources which have an associated power (or *luminosity*) measured as the amount of energy radiated per second. The flux defines how much luminosity is incident on a receiver for its given area and it is an extremely important relation in observational cosmology. The flux can be shown to be [158]

$$\mathrm{d}F(\nu) = \frac{\mathrm{d}L(\nu_{21})}{4\pi d_{\mathrm{L}}^2} = \frac{\mathrm{d}L(\nu_{21})}{4\pi d_{\mathrm{c}}^2(1+z)^2} \tag{2.7}$$

where dL is the luminosity from a HI source emitted with a frequency $\nu_{21} \sim 1.4$GHz and the relation is shown for both luminosity distance d_{L} and comoving distance d_{c} (see section 1.2.2) regarding cosmological distances.

Radio telescopes are designed such that they are sensitive to a certain range of frequencies defined by the *bandwidth*. When observing the 21cm emission from a redshifted galaxy source it will typically have an extended line profile caused by the galaxy's intrinsic rotation, creating an observed signal such as the one shown in Figure 2.2. Since the bandwidth can usually be much narrower than the 21cm line profile an intensity distribution is produced which shows separate measurements at many points across the receiver passband and shows the 21cm spectral feature [207]. This bandwidth thus needs to be considered when characterising the full observed signal received by the telescope and it is common to use flux density which is the flux per bandwidth. In radio astronomy, it is generally very common to use the unit of *Janskys* (Jy) to quantize flux density where 1Jy is defined as 10^{-26} Wm^{-2} Hz^{-1} (with W being watts, a measure of power).

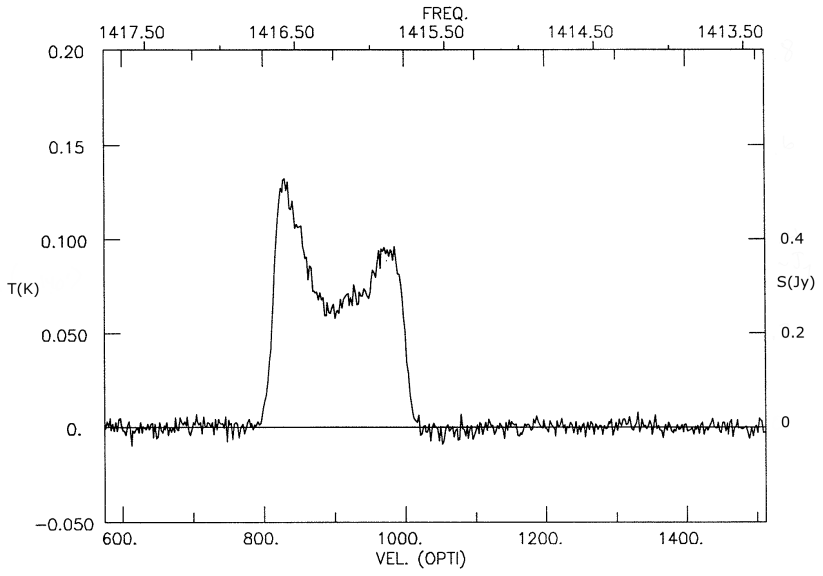

Figure 2.2: Example of an integrated HI spectrum from UGC 11707 demonstrating the typical two-horned profile of a rotating spiral galaxy [99]. Velocity measured in km s^{-1} and frequency in MHz.

For cases where the HI emitter is an unresolved point source, telescopes will agree on flux density measurements. However, where the source of radiation is widespread enough such that it fills the telescope beam, as is often the case in intensity mapping, telescopes of differing dish size provide different flux density measurements. Therefore radio astronomers often use *brightness* or specific intensity which is the flux density per solid angle.

With this understanding of the observables from radio telescope data, it is possible to measure the total HI mass $M_{\rm HI}$ of a galaxy [68]. Firstly, consider a system of neutral hydrogen which has a number density of HI atoms given by $n_{\rm HI} = n_0 + n_1$, where the subscripts 0 and 1 represent the lower and upper energy levels of the hyperfine structure respectively which have corresponding energies E_0 and E_1. The atoms in excited state E_1 will spontaneously return to the lower level E_0 with a certain probability A_{10} known as an *Einstein coefficient*. This coefficient is defined such that $n_1 A_{10}$ is the number of such spontaneous transitions per second in a unit volume. For the 21cm spontaneous emission we have $A_{10} \sim 2.869 \times 10^{-15} s^{-1}$ which is what makes any single transition such a rare event [225]. A useful starting point in deriving the signals received from such emission is the *emissivity* which is defined as the energy per unit time, per unit volume, per unit frequency, per unit solid angle. From atomic physics we know the emissivity of the 21cm transition to be given by [68][225]

$$\mathrm{d}\varepsilon = \frac{\mathrm{d}E_{\rm e}}{\mathrm{d}t\mathrm{d}V_{\rm e}\mathrm{d}\nu_{\rm e}\mathrm{d}\Omega_{\rm e}} = \frac{\mathrm{d}L_{\rm e}}{\mathrm{d}V_{\rm e}\mathrm{d}\nu_{\rm e}\mathrm{d}\Omega_{\rm e}} = \frac{h_{\rm P}\nu_{21}A_{10}}{4\pi}\frac{n_1}{n_{\rm HI}}n_{\rm HI}\varphi(\nu_{\rm e}) \tag{2.8}$$

where the e subscripts denote the quantities at time of emission. $\varphi(\nu_{\rm e})$ is the line profile which is assumed to be very narrow with width $\mathrm{d}\nu_{\rm e}$ such that it can be approximated that $\varphi(\nu_{\rm e}) = 1/\mathrm{d}\nu_{\rm e}$. $h_{\rm P}$ is Planck's constant. The relative population of the levels of upper and lower states is governed by an excitation temperature referred to as *spin temperature* and the relationship is defined by the equation [56]

$$\frac{n_1}{n_0} = \frac{g_1}{g_0}\exp\left(-\frac{h_{\rm P}\nu_0}{k_{\rm B}T_{\rm s}}\right) \tag{2.9}$$

where g_i is the statistical weights for the spin states with upper and lower spin states given as $g_1 = 3$ and $g_0 = 1$. The spin temperature is $T_{\rm s}$ and observations show that for HI in low redshift galaxies this can be as large as 300K [52] which means $h_{\rm P}\nu_{21}/k_{\rm B}T_{\rm s} \ll 1$ therefore

$$\frac{n_1}{n_0} \approx \frac{g_1}{g_0} = 3 \quad \text{and} \quad \frac{n_{\rm HI}}{n_1} = \frac{n_{\rm HI}}{n_0}\frac{n_0}{n_1} = \frac{n_0+n_1}{n_0}\frac{1}{3} = \frac{4}{3}. \tag{2.10}$$

Hence the emissivity from a system of HI becomes

$$\frac{\mathrm{d}L_{\rm e}}{\mathrm{d}V_{\rm e}\mathrm{d}\nu_{\rm e}\mathrm{d}\Omega_{\rm e}} = \frac{3h_{\rm P}\nu_{21}A_{10}}{16\pi}n_{\rm HI}\varphi(\nu_{\rm e}). \tag{2.11}$$

This is all assuming that the HI clump is optically thin[6] and also that the spin temperature of the gas is much larger than the background temperature (usually the CMB) which evidence from 21cm absorption by damped Lyman-α systems support [116]. This means we can neglect self-absorption of the HI. The above clump is in the comoving frame and therefore $\mathrm{d}V_{\rm e}$ is the

[6] This is only likely to be a poor assumption when the largest disc galaxies are seen close to edge on [4]

comoving volume element of the clump and also $n_{\rm HI}$ is a comoving number density. From the discussion above we can define the HI specific intensity (or brightness) $I_{\rm HI}$ quantitatively as

$$I_{\rm HI} = \frac{{\rm d}F_{\rm o}}{{\rm d}\nu_{\rm o}{\rm d}\Omega_{\rm o}} = \frac{{\rm d}L_{\rm e}}{4\pi(1+z)^2 d_{\rm c}^2(z){\rm d}\nu_{\rm o}{\rm d}\Omega_{\rm o}}. \tag{2.12}$$

In contrast to the emissivity, these parameters involve observed quantities and therefore have o subscripts to denote this. We can then use equation (2.11) to relate the emissivity of a HI clump to its brightness. We can also realise that the solid angle subtended by a spherical surface for any of its interior points is 4π meaning we can integrate over the solid angle at emission such that $\int {\rm d}\Omega_{\rm e} = 4\pi$. This gives the following for the brightness

$$I_{\rm HI} = \frac{3h_{\rm P}\nu_{21}A_{10}}{16\pi}\frac{n_{\rm HI}}{(1+z)^2 d_{\rm c}^2(z)}\frac{{\rm d}V_{\rm e}}{{\rm d}\nu_{\rm o}{\rm d}\Omega_{\rm o}} \tag{2.13}$$

where I have also utilised the assumption that $\int \varphi(\nu_{\rm e}){\rm d}\nu_{\rm e} = 1$ for the thin line profile. We can finally integrate over the volume of the whole clump to obtain an expression in terms of $M_{\rm HI}$, the total HI mass of the clump, using $M_{\rm HI}/m_{\rm H} = N_{\rm HI} = \int n_{\rm HI}{\rm d}V_{\rm e}$, where $m_{\rm H}$ is the mass of the hydrogen atom giving

$$\boxed{I_{\rm HI} = \frac{3h_{\rm P}\nu_{21}A_{10}}{16\pi m_{\rm H}}\frac{1}{(1+z)^2 d_{\rm c}^2(z)}\frac{M_{\rm HI}(z)}{{\rm d}\nu_{\rm o}{\rm d}\Omega_{\rm o}}.} \tag{2.14}$$

This important relation will be used extensively in the simulations of HI intensity maps throughout this thesis and is an excellent way of modelling a radio telescope's HI signal from output masses often found in simulated galaxy catalogues.

Rayleigh-Jeans Law

Radiation emitted from a source in *local* thermodynamic equilibrium will have a brightness $I(\nu)$ equal to that of a blackbody at temperature $T(\nu)$ and is therefore given by Planck's law as

$$I(\nu) \equiv \frac{2h_{\rm P}\nu^3}{c^2}\frac{1}{\exp\left(\frac{h_{\rm P}\nu}{k_{\rm B}T}\right)-1} \tag{2.15}$$

where $k_{\rm B}$ is Boltzmann's constant and c is the speed of light [68]. This however, is a spectral *distribution* of the radiation from a blackbody and to map the *line* intensity of HI, we are simply interested in the $\nu_{21}/(1+z)$ frequency. It is customary in radio astronomy to quantify the brightness of an extended source by the *brightness temperature* defined as the temperature one gets from the Rayleigh-Jeans Law for a brightness of $I(\nu)$ [225]. The Rayleigh-Jeans Law can be derived by first making the approximation that $k_{\rm B}T \gg h_{\rm P}\nu$, which for low-frequency radio astronomy including HI signals is a very reasonable approximation. Then, utilising the fact that $\exp(x) \approx 1+x$ in the limit $x \to 0$, it can be shown that the intensity $I(\nu)$ is directly proportional to the thermodynamic temperature of the blackbody. In the context of HI, assuming it also behaves like a blackbody, we get the relation

$$T_{\rm HI}(\nu) = \frac{I_{\rm HI}c^2}{2k_{\rm B}\nu^2} = \frac{I_{\rm HI}\lambda^2}{2k_{\rm B}}. \tag{2.16}$$

The HI brightness from equation (2.13) can be further simplified into a form which is widely used in the literature [28][43][231]. Firstly consider that from $\nu = \nu_{21}/(1+z) = c/\lambda_{21}(1+z)$ we get the following

$$\frac{\mathrm{d}\nu}{\mathrm{d}z} = -\frac{c}{\lambda_{21}(1+z)^2}. \tag{2.17}$$

From the comoving distance equation derived in (1.25) we have $\mathrm{d}z = \mathrm{d}r\,H(z)/c$ and plugging this into the above gives

$$\mathrm{d}r = \frac{\lambda_{21}(1+z)^2}{H(z)}\mathrm{d}\nu_\mathrm{o} \tag{2.18}$$

Using this result and by rewriting equation (2.13) with the volume factor to $\mathrm{d}V_\mathrm{e} = \mathrm{d}A_\mathrm{e}\mathrm{d}r$, the brightness can be simplified to

$$I_\mathrm{HI} = \frac{3h_\mathrm{P}cA_{10}}{16\pi}\frac{1}{H(z)}n_\mathrm{HI} \tag{2.19}$$

where I have also recognised that the comoving area element can be written $\mathrm{d}A_\mathrm{e} = \mathrm{d}\Omega_\mathrm{o}d_\mathrm{c}^2$. Therefore in the Rayleigh-Jeans limit we can relate the brightness temperature to the HI number density by

$$\boxed{T_\mathrm{HI} = \frac{3h_\mathrm{P}c^3A_{10}}{32\pi k_\mathrm{B}\nu_{21}^2}\frac{(1+z)^2}{H(z)}n_\mathrm{HI}} \tag{2.20}$$

where I have once again used the relation $\nu = \nu_{21}/(1+z)$.

2.3.2 Power Spectrum

The brightness temperature defined by (2.20) can be spilt into a background homogeneous part and a fluctuating part with $T_\mathrm{HI} = \overline{T}_\mathrm{HI}(1+\delta_\mathrm{HI})$. It is the fluctuating part δ_HI that is of interest for the purpose of LSS investigation and therefore the cosmological quantity of interest is the *over-temperature* defined as

$$\delta T_\mathrm{HI}(\vec{\theta},z) = \overline{T}_\mathrm{HI}(z)\,\delta_\mathrm{HI}(\vec{\theta},z) = T_\mathrm{HI}(\vec{\theta},z) - \overline{T}_\mathrm{HI}(z) \tag{2.21}$$

where $\vec{\theta}$ is the angular coordinate for a given voxel (3D pixel) and z is the redshift to it. $\overline{T}_\mathrm{HI}(z)$ is the mean HI brightness temperature at redshift z. Since HI is expected to be a biased tracer of the late Universe's matter density field δ_M, we can say $\delta T_\mathrm{HI} = \overline{T}_\mathrm{HI}b_\mathrm{HI}\delta_\mathrm{M}$. Therefore (ignoring RSD for now and assuming a linear deterministic bias) we can expect the power spectrum of the HI fluctuations to take the form

$$P_\mathrm{HI}(k,z) = \overline{T}_\mathrm{HI}^2(z)\,b_\mathrm{HI}^2(z)\,P_\mathrm{M}(k,z). \tag{2.22}$$

The mean HI brightness temperature $\overline{T}_\mathrm{HI}$ is often represented as the below which can be derived from (2.20) using $n_\mathrm{HI} = \Omega_\mathrm{HI}(z)\rho_{c,0}/m_\mathrm{H}$ where $\rho_{c,0} = 3c^2H_0^2/8\pi G$;

$$\overline{T}_\mathrm{HI}(z) = 180\Omega_\mathrm{HI}(z)h\frac{(1+z)^2}{E(z)}\mathrm{mK}. \tag{2.23}$$

Constraining the HI density $\Omega_\mathrm{HI} \propto \overline{T}_\mathrm{HI}$ along with the bias b_HI is therefore particularly important for the success of HI intensity mapping and seen as one of the early challenges for the field

as it matures [171]. Since higher abundance of HI provides a stronger cosmological signal, constraining $\Omega_{\rm HI}$ is also important for forecasting the signal-to-noise of surveys [43].

At certain redshifts the values of $\Omega_{\rm HI}$ alone are relatively well constrained as shown in Figure 2.3 which is data gained from targeted HI galaxy surveys and from observations of damped Ly-α systems [170]. Measuring $\Omega_{\rm HI}$ at 'mid'-redshift ranges of $z \sim 1$ is difficult since HI galaxies are very faint at these distances thus hard to detect and Lyman-α observations are challenging for $z < 2$ [224].

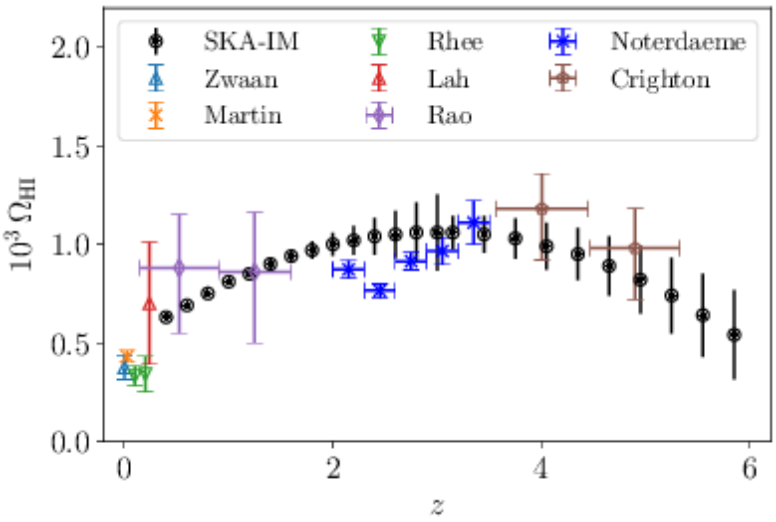

Figure 2.3: Measurements of the HI density $\Omega_{\rm HI}$. Plot from [23] and data obtained from [241][175] [134][148][178][57]. Black IM points show predictions for constraints from intensity mapping survey with the SKA as forecasted by [23] and following methodology outlined in [170]. This shows how $\Omega_{\rm HI}$ is more constrained at low redshifts where we can do targeted HI galaxy observations and at higher redshifts ($z > 2$) where we can rely on Lyman-α surveys. Around $z \sim 1$ it is likely we will rely on intensity mapping for $\Omega_{\rm HI}$ constraints.

Intensity mapping can be useful for measuring HI abundance at these mid-redshift ranges and forecasts are shown in Figure 2.3 (black points) for a SKA intensity mapping survey. Measurements have already been conducted using the GBT intensity maps in cross-correlation as discussed in section 2.2.4. The cross-correlation power spectrum can be written as [169]

$$P_{\rm HI,g}(k,z) = \overline{T}_{\rm HI} b_{\rm HI} b_{\rm g} r P_{\rm M}(k,z) \tag{2.24}$$

where I introduce the cross-correlation coefficient r which is defined as [229]

$$r(k) = \frac{P_{\rm HI,g}(k)}{\sqrt{P_{\rm HI}(k)\,P_{\rm g}(k)}} \tag{2.25}$$

to reflect the fact that both tracers may exhibit some stochasticity, however on large scales we expect $r \to 1$ [135]. As shown in [135], cross-correlations provide a constraint on the measurement of the $\Omega_{\rm HI} b_{\rm HI}$ prefactor of $\Omega_{\rm HI} b_{\rm HI} r = [4.3 \pm 1.1] \times 10^{-4}$ at redshift $z = 0.8$.

As can be seen from equation (2.22) and the quoted measurements of HI abundance using intensity mapping there exists a degeneracy between $b_{\rm HI}$ and $\Omega_{\rm HI}$. While $\Omega_{\rm HI}$ is individually

relatively well constrained at *certain* redshifts by several observations, the cosmological bias is less so and is a larger source of uncertainty in the $P_{\rm HI}$ signal. Previous work [135][170] has discussed using RSD in the HI auto-correlation to break this degeneracy. Similarly to (1.48), we can include RSD in HI auto-power spectrum so it is written as

$$P_{\rm HI}(k,z) = \overline{T}_{\rm HI}(z)^2 b_{\rm HI}(z)^2 \left[1+\beta_{\rm HI}(z)\mu^2\right]^2 P_{\rm M}(k,z). \tag{2.26}$$

Assuming some fixed fiducial cosmology the $\overline{T}_{\rm HI} b_{\rm HI}$ degeneracy can theoretically be broken and allows for a separate measurement of $b_{\rm HI}(z)$ and $\Omega_{\rm HI}(z)$, as shown in [170].

2.3.3 HI Halo Model

Along with measurements for $\Omega_{\rm HI}$ and $b_{\rm HI}$, theoretical models can also be used to predict the abundance and spatial distribution of HI. These are often halo-based models e.g. [213][153][157] and they principally rely on being able to connect the HI mass within the dark matter halo to the halo's mass. This is formalised by the HI-halo mass (HIHM) function $M_{\rm HI}(M,z)$ which assumes HI mass is only a function of the halo mass M and redshift. While there may also be some local environmental dependence affecting this function and also some stochasticity, in the context of intensity mapping where low resolutions are being considered, these will have minimal impact on this relation [43]. In this context the HI abundance can be modelled using [212]

$$\overline{\rho}_{\rm HI}(z) = \Omega_{\rm HI}(z)\rho_{\rm c,0} = \int_0^\infty n(M,z) M_{\rm HI}(M,z) {\rm d}M \tag{2.27}$$

where $\overline{\rho}_{\rm HI}$ is the mean HI density and $\rho_{\rm c,0}$ is the critical density of the Universe today. $n(M,z)$ is referred to as the *halo mass function* which defines how many halos of mass M exist for a redshift z [208] and is generally a fitted function based on results from N-body simulations. This halo-based model assumes that the majority of HI within the late Universe resides inside dark matter halos which, as I briefly introduced in section 2.1.3, is a reasonable assumption.

It is predicted that a large impact on the bias is the size of dark matter halos with only the larger halos being able to gain sufficient density for self-shielding from ionizing radiation [43]. The bias can also be modelled in this halo-based approach and is given as [153]

$$b_{\rm HI}(z) = \frac{1}{\overline{\rho}_{\rm HI}(z)} \int_0^\infty b(M,z) n(M,z) M_{\rm HI}(M,z) {\rm d}M = \frac{\int_0^\infty b(M,z) n(M,z) M_{\rm HI}(M,z) {\rm d}M}{\int_0^\infty n(M,z) M_{\rm HI}(M,z) dM} \tag{2.28}$$

where $b(M,z)$ is the halo bias discussed in detail in [209], but put simply governs the fact that higher mass halos are more clustered than lower mass ones and clustering is in general higher at higher redshift [139]. While this simple prescription is sufficient, especially for the purposes of an intensity mapping study where small scales are of little interest, it does break down in greater detail where further factors influence the bias such as halo formation history, spin, angular momentum, shape etc. [154]. These additional influencing factors on the bias are referred to as *assembly* (or secondary) bias and I refer the reader to [73][132] for more detail on this concept.

2.4 Simulating 21cm Cosmology

The development of reliable simulations is important for a novel observational approach such as intensity mapping. Forecasting the benefits with simulations is encouraging, explorative and useful for the purposes of lobbying further investment or collaboration. Producing a large number of mocks is also crucial in the analysis of LSS data and used in the computation of covariance matrices. Furthermore, reliable simulations is an excellent route to understanding an observational technique's systematics as the technique matures. This is no different for intensity mapping where early observational data and simulated signals can be compared to understand the range of systematics the technique is affected by. All of this suggests that the simulation of 21cm observational data is crucial for the development of the field and the eventual reliability of the resulting conclusions.

2.4.1 Building Structure

Simulating large scale cosmological structure is most accurately done using an N-body simulation where dark matter particles are evolved, allowing the formation of halos through accretion and repeated mergers, building the so-called merger tree history. The positions and velocities of the particles at several discrete time-steps are saved and these outputs form the cosmic density field. Dark matter only N-body simulations have existed as a reliable tool for understanding large scale structure for some time now [199], but added precision can be gained from inclusion of baryonic gas particles and relevant dissipative processes or at least inclusion of analytical prescriptions in *semi-analytical* models. The resulting density fields can be processed with a halo-finding algorithm to obtain the locations and basic properties of the dark matter halos [31]. Simulating galaxy formation inside these identified halos on the peaks of the density field can be done using a number of different ways with varying ranges of precision and hence processing power. I briefly summarise the different simulation approaches below in approximate descending order of simplicity. See [34][81][197] for a full review.

- Halo Ocupation Distribution (HOD) [159][35]: The number of galaxies N assigned to a halo of mass M is determined by a probability distribution $P(N|M)$.
- Halo Abundance Matching (HAM) [88]: All halos above some mass cut-off host galaxies. The mass of the halo determines the likely mass of the galaxy based on the stellar mass-halo mass (SMHM) relation [142].
- Semi-Analytic Models [54]: Often referred to as hierarchical models, these depart from the previous empirical approaches and look to include physical processes caused by baryonic physics. For example how much gas accretes into halos, how much hot gas cools and turns into stars, how feedback processes from supernovae and AGN eject gas from the halo etc. are approximated with analytic prescriptions that are traced through the merging history of dark matter halos [197][219].

- Hydrodynamical Simulations [215]: The most physical approach involves full numerical calculations solving the equations of gravity and hydrodynamics for each particle and include simulation of baryonic processes, potentially including star formation, radiative transfer, feedback from supernovae, AGN and star formation etc. [197].

Hydrodynamical simulations therefore offer the most precise results. However, the problem with them is that they require enormous amounts of processing power to resolve the small-scale events which affect the large scale distribution of particles. Therefore compromises are made when large simulations (on Gpc scales) are required and in these cases, simpler halo-based, dark matter only N-body simulations are often preferred. Advances in computational power and numerical efficiencies in recent years are however allowing hydrodynamical simulations to compete more with semi-analytical and empirical galaxy formation models [200].

2.4.2 Assigning HI

As with most simulations of cosmological observables, there are a number of ways of modelling the HI signal with a trade-off between accuracy and computational expense. This can be done perhaps most simply in map-space taking a given over-density dark matter field $\delta_{\rm M}$ and transforming this into an over-temperature map $\delta T_{\rm HI} = \overline{T}_{\rm HI} b_{\rm HI} \delta_{\rm M}$ with some assumed ficucial models of $\overline{T}_{\rm HI}$ and $b_{\rm HI}$. This is of course highly empirical and reliant on accurate models for the bias and mean brightness temperature. However, it presents a rapid option for producing multiple mocks of intensity maps especially if the simulation of the matter density field is similarly efficient e.g. a *lognormal* approach [14].

A further step-up in complexity is to use a set of physical prescriptions to assign HI to individual galaxies produced in an N-body simulation, an approach adopted in e.g. [150][239]. Generally these approaches work by deriving a gas content and splitting the gas masses into fractions of HI, HII, Helium etc. within an analytical framework. For the purposes of intensity mapping where small scales are un-probed due to the beam, precision over realistic shape profiles and ISM detail is not overly important and this approach represents a good model of HI on large scales.

The most physical way of simulating the distribution and abundance of HI is with hydrodynamical simulations where coupled gas and dark matter particles are simultaneously evolved. With some assumption on neutral hydrogen gas fractions, which can be constrained from observations [151], HI distribution can be studied at various redshifts. Work done in [213], using state-of-the-art magneto-hydrodynamic simulations as part of the IllustrisTNG Project, is an example of how simulations are contributing to our understanding of the distribution and abundance of HI in our late Universe.

Results from simulations such as these can be compared to the small amount of data we have as a cross-check. As an example, observation and model can be compared for $N_{\rm HI}$ the number of atoms along the LoS and HI column density distribution function $f_{\rm HI}(N_{\rm HI})$ which is the the number of absorbers per unit column density, per unit absorption length (see [174][212] for further details). Since the HI column density distribution function can be constrained through

observations of the Lyman-α forest, this represents a good test for simulations. Figure 2.4 shows a comparison between data from these observations and the results using the IllustrisTNG [213] and there is excellent agreement between the two.

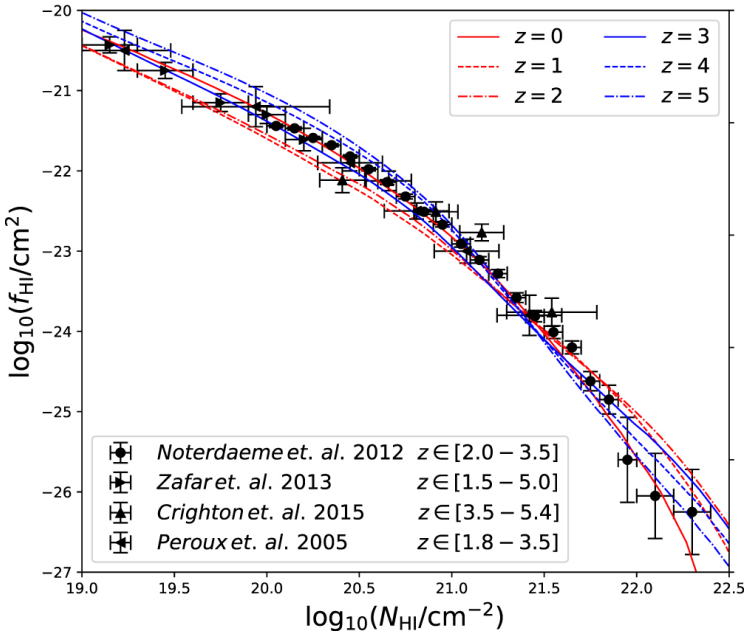

Figure 2.4: Comparison between observational and simulated data for the HI column density distribution function. Observational data from [166][148][57][236] are shown by the black data points and the coloured lines are for the simulated data produced using IllustrisTNG [213].

Furthermore, observational data such as the HI abundance presented in Figure 2.3 has also been compared with the results from [213] and they largely agree with these measurements. These agreements with observational data allow confidence to be placed in conclusions from simulations.

2.4.3 Mocks for HI-Galaxy Cross-Correlations

Results from [213] suggest that around 99% of HI resides in halos at $z < 2$, falling to 88% at $z = 5$. This makes intuitive sense since at higher redshift the gas in the IGM, outside of the halos, is denser and has had less time for UV radiation to ionise its HI content. This implies that simulating the HI signal in the late Universe can be done sufficiently by just focusing on HI content of each dark matter halo. This has important consequences for the efficiency of developing mock simulations which aim to study cross-correlations between HI intensity maps and galaxy redshift surveys.

Simulating HI intensity maps can be done rapidly and relatively accurately as I discussed above and as presented in [14]. However, producing a galaxy catalogue alongside these intensity maps that shares the same underlying clustering properties extends the complexity somewhat. However, the cosmology community have been simulating optical galaxy catalogues for some time now with increasing proficiency and computational resource [53][131][235][58][21]. It is

Parameter	Value	Parameter	Value
M_{10}	$4.58^{11} \pm 0.19$	M_{11}	$1.56^{+0.53}_{-2.70}$
N_{10}	$9.89^{-3} \pm 4.89$	N_{11}	$0.009^{+0.06}_{-0.001}$
b_{10}	0.90±0.39	b_{11}	$-1.08^{+1.52}_{-0.08}$
y_{10}	0.74±0.03	y_{11}	$4.07^{+0.39}_{-2.49}$

Table 2.2: Best-fit free parameter values for the HI-halo mass function in equation 4.6. Values obtained from [156].

typical for each of these catalogues to output the hosting halo's mass for each galaxy within the catalogue. It is possible to use these halo masses to infer some HI content but this relies on a realistic model of the relation between HI and its hosting halo properties. The evolution of HI content within dark matter halos is an active area of research and the HIHM relation can take a variety of forms. Generally it is a fitting function with free parameters which have redshift dependence constrained by observation. Below is an example of a HIHM relation I use in this thesis and I refer the reader to [156] for further details.

$$M_{\rm HI} = 2N_1 M\Big[\Big(\frac{M}{M_1}\Big)^{-b_1} + \Big(\frac{M}{M_1}\Big)^{y_1}\Big]^{-1} \tag{2.29}$$

Here M_1, N_1, b_1 and y_1 are all the free parameters tuned to provide a best fit. These parameters are redshift dependent and given by

$$\begin{aligned} \log_{10} M_1 &= \log_{10} M_{10} + \frac{z}{z+1} M_{11}\,, \\ N_1 &= N_{10} + \frac{z}{z+1} N_{11}\,, \\ b_1 &= b_{10} + \frac{z}{z+1} b_{11}\,, \\ y_1 &= y_{10} + \frac{z}{z+1} y_{11}\,, \end{aligned} \tag{2.30}$$

where each of the values for the equations are provided in Table 2.2. By using a HIHM function catalogues of galaxies can be gridded into a map of HI brightness using equation (2.14) and with further modelling of telescope effects such as the beam, instrumental noise and further systematics, realistic simulations of intensity maps can be produced. These intensity maps will of course then share the same underlying clustering signal as the catalogue of galaxies from which they were produced thus allowing for studies into galaxy-HI intensity map synergies. This is a method used throughout this thesis for simulating data and will therefore be discussed in more detail in subsequent chapters.

The approach of simulating the HI signal was also something investigated in [213] where they derive HI masses $M_{\rm HI}(M,z)$ from halo masses M given by some HIHM function. The halos are produced using a 'computationally-cheap' N-body simulation and the derived HI masses are placed onto a grid using the coordinates for the centre of the hosting halo. These grids are then processed to produce 2-dimensional maps of HI signal at a chosen frequency. Figure 2.5 shows results for the power spectra of these maps given by the blue line. Also shown by the orange line are results using the same process except the HI signal is created by spatial distribution of HI in the hydrodynamical IllustrisTNG simualtion [200]. These results show little difference

in terms of power spectrum shape between both methods. However, the amplitudes can show significant difference between 10% and 40% offset. This is explained in [213] by inaccuracies of their $M_{\rm HI}(M,z)$ HIHM function. Unless this is reproducing perfectly the correct HI abundance, then $\overline{T}_{\rm HI}$ will be off by some factor which is what is these results are showing.

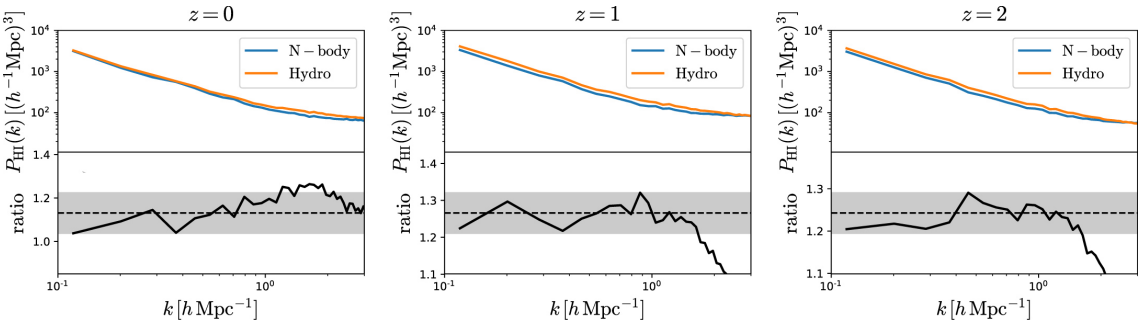

Figure 2.5: Comparison between power spectra at different redshifts for full hydrodynamical HI simulations (orange line) and a computationally cheaper method where HI mass and position is derived from halo properties simulated in an N-body simulation (blue line). The lower panel in each plot shows the ratio between the two power spectra across all scales. Plot adapted from original in [213].

This is evidence in support of a method which produces HI intensity maps on the back of pre-existing optical galaxy catalogues and their hosting halo masses. It suggests that this approach should reproduce a similar shape of power spectrum as a full hydrodynamical simulation. An aspect which needs careful consideration is the offset in amplitudes and this is especially the case when using halo properties from optical galaxy catalogues. This is because optical surveys only resolve the brightest galaxies above some detection threshold. Therefore simulations of their catalogues only need to produce bright galaxies typically high in mass and for this reason optical galaxy simulations tend to have quite a large halo mass resolution e.g. MICE[7] [78] whose dark matter halos are only resolved down to a few $10^{11}\,M_\odot\,h^{-1}$. This lack of low mass halos means the full HI abundance is unlikely to be reproduced by HIHM-based simulation. Results from Figure 2.5 would suggest that this is likely to affect the amplitude of the power spectrum which to correct would require a well constrained model of $\Omega_{\rm HI}$.

2.5 Summary

There is strong evidence from simulations [168][212][213] and now observations [162][46][135][18] that HI is a biased tracer of the underlying dark matter density. In this chapter I have introduced the novel technique of using maps of unresolved 21cm emission from HI to map LSS. This provides a different approach to probing LSS from the conventional optical galaxy redshift survey which is the current dominant source of observational data. As Figure 2.6 shows, intensity mapping is a technique which is slowly gathering momentum and with telescopes designed

[7] maia.ice.cat/mice/

specifically for intensity mapping now beginning to gather data (as shown in Table 2.1), this trend is unlikely to change.

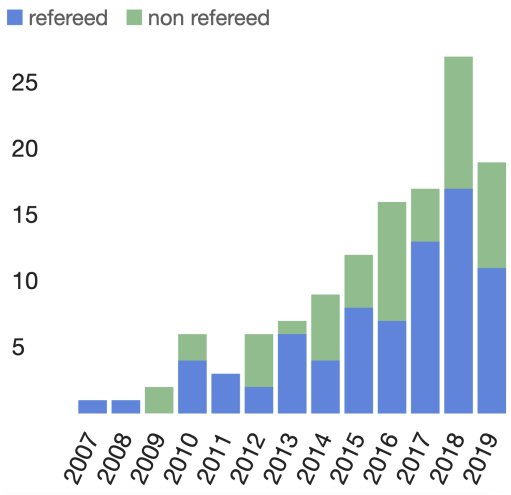

Figure 2.6: The growth in HI intensity mapping publications with publication year against number of publications. Results obtained from SAO/NASA Astrophysics Data System (ADS), with a search for papers with 'intensity map' in the title and either '21', 'HI' or 'neutral hydrogen' in the abstract. Search done 25th July 2019.

As shown by equation (2.14), the HI brightness $I_{\rm HI}$ can be shown to be a function of HI mass $M_{\rm HI}(z)$ along with some telescope parameter dependence. Using this under the Rayleigh-Jeans approximation (equation (2.16)) provides an expression for the HI brightness temperature $T_{\rm HI}$, the conventional observable in radio astronomy. This provides a formalism with which to begin modelling HI signals assuming reliable HI masses can be simulated which, as I outlined in sections 2.4.2 and 2.4.3, should be possible. I have also discussed some of the key systematics associated with radio telescopes undertaking intensity mapping observations (section 2.2.3) and including these in simulations is perhaps just as important as modelling the cosmological signal correctly (as I will show in Chapter 4).

We are soon likely to enter a phase in precision cosmology where uncertainties in results are systematic dominated rather than statistically dominated. An excellent way to limit systematics is in cross-correlations where two separate probes e.g. a galaxy survey and an intensity mapping survey, will see their systematical errors reduced due to the inherently different telescope designs both operate. This means each telescope's systematics can be considered independent from one another leading to the reduced error under cross-correlation (as demonstrated by equation (2.6)). I have laid the foundational framework for simulating and analysing the 21cm signal in the context of intensity mapping and also under cross-correlation with an optical redshift survey. The rest of the thesis will call upon this central formalism for developing a pipeline to design and forecast various methods associated with this novel technique.

CHAPTER 3

CLUSTERING-BASED REDSHIFT ESTIMATION WITH HI INTENSITY MAPS

Cunnington S., Harrison I., Pourtsidou A., Bacon D., (2019), Mon. Not. Roy. Astron. Soc., 482, 3341

This chapter presents work from the above published article [59] with amendments made to the original published script for the purposes of this thesis. For all parts of this chapter, including text and figures, I am the principal author with contributing edits from my co-authors (except Figures 3.7, 3.11 and Section 3.2.2 which were principally authored by Ian Harrison.)

Precision cosmology requires accurate galaxy redshifts, but next generation optical surveys will observe unprecedented numbers of resolved galaxies, placing strain on the amount of spectroscopic follow-up required. In this chapter I show how useful information can be gained on the redshift distributions of optical galaxy samples from spatial cross-correlations with intensity maps of unresolved HI (21cm) spectral line emission. I construct a redshift distribution estimator, which is tested using simulations. I utilise the S^3-SAX catalogue which includes HI emission information for each galaxy, which I use to construct HI intensity maps. I also make use of simulated LSST and *Euclid*-like photometry enabling the application of HI clustering calibration to realistic simulated photometric redshifts. While taking into account important limitations to HI intensity mapping such as lost k-modes from foreground cleaning and poor angular resolution due to large receiver beams, I show that excellent constraints on redshift distributions can be provided for an optical photometric sample.

3.1 Introduction

As I discussed in Chapter 1, according to the standard cosmological model (see Section 1.4), dark energy is responsible for the current acceleration of the Universe's expansion [179][165]. The next step towards constraining our cosmological model relies on precise measurements of the 3-

dimensional large-scale structure. The majority of this structure is in the form of underlying dark matter which does not interact with light and is therefore invisible to our telescopes. However, making the well reasoned assumption that light emitting galaxies act as a biased tracer of this underlying dark matter distribution, we can use large optical surveys to construct catalogues of galaxies. We then process and analyse these catalogues to construct a 3-dimensional map of the Universe. This relies heavily on having a good method for measuring the radial distance out to all these galaxies, i.e. having a good estimate of the galaxy redshifts.

There exist two approaches to measuring redshifts in optical catalogues, *spectroscopy* and *photometry*. Spectroscopy is the more accurate of the two but is time consuming since it relies on gathering a large number of photons for any one galaxy. An estimation of redshift is then obtained by observation of known emission or absorption lines in the spectral energy distribution (SED). With the rapidly increasing orders of magnitude of galaxy numbers detected by forthcoming surveys such as the Large Synoptic Survey Telescope[1] (LSST) and *Euclid*[2]-like surveys, a time-expensive method such as spectroscopy is unlikely to be a viable method for measuring the redshift for these large populations.

Often surveys need to settle for the photometry approach [41], which is faster but not as accurate as spectroscopy. This relies on obtaining the SED from broad-band photometry i.e. measuring the amount of flux collected in each of the telescope's broad colour filters, and relies on strong galaxy spectral features such as the 4000Å break being detectable. Obtaining photometric redshifts can therefore be thought of as spectroscopy with extremely low resolution; for example the LSST plans to operate with six colour filters, *ugrizy* [133]. Photometric redshift methods can generally be categorised into either template fitting methods, where various spectrum templates are fitted to find a close match, or opting for machine learning methods where a training set is used to derive a relation between redshifts and colour magnitudes [185]. Opting for a photometric approach means a far greater number of galaxies can have estimated redshifts, but more detailed consideration must be taken of the redshift error associated with this technique.

A method to calibrate photometric redshifts, without the need for verification from time-expensive spectroscopic follow-up, is to use clustering-based redshift estimation. The general idea is to use a pre-existing 'reference' sample for which some precise redshift information has already been gained, and which spatially overlaps with the photometric sample which can be treated as having unknown redshift. Then by utilising the spatial clustering of galaxies within the overlapping samples through cross-correlations, we can constrain the 'unknown' (photometric) redshift distribution. In other words, where there is strong angular clustering between the unknown sample and a slice of the known sample at a particular redshift, one can infer that the unknown sample is well represented in that particular redshift bin. From this principle we can build an estimated redshift distribution for the unknown sample, giving much more constrained redshift information for the particular population of galaxies.

There is now a significant amount of literature on clustering-based redshift estimation, with

[1] www.lsst.org
[2] www.euclid-ec.org

[147] being one of the first to demonstrate the method on simulations. The method has since been refined with simulations by [136][33][192][210], and others have more recently applied the approach to real data [143][173]. Most recently the Dark Energy Survey have applied the clustering redshifts method to their Year 1 Data [85][61].

The appeal of this idea is that when LSST and *Euclid*-like surveys deliver unprecedented galaxy catalogue sizes but lack well-constrained redshift information, we do not need to rely on time-consuming spectroscopic follow-up on every galaxy, or representative sub-samples which are not biased with respect to the full survey. Instead we can utilise a pre-existing, spatially overlapping, catalogue for which there is precise redshift information and use this as the reference sample in a clustering-based redshift estimation.

However, there is no reason why the reference sample needs to be a sample of resolved galaxies. The idea should work just as well if one cross-correlates with any tracer of large scale structure. The idea that we will investigate in this paper is the use of HI intensity maps (see Chapter 2). Intensity maps and photometric galaxy surveys are highly complementary to one another, with photometric surveys having high spatial but low spectral resolution, and intensity maps high spectral but low spatial resolution. Even though in the epoch of reionization (approximately $6 \leq z \leq 15$) it is thought that the power spectrum measured from HI intensity mapping will be largely shaped by the pattern of ionized regions, in the post-reionization era i.e. once reionization is complete ($z < 6$), some HI will still remain in collapsed objects and the HI power spectrum will therefore be a measure of the underlying matter power spectrum [232].

It is apparent therefore that cross-correlations can be beneficial for both optical surveys and radio HI intensity mapping experiments. Radio can help calibrate photometric redshifts, and optical galaxy surveys can help radio HI intensity mapping surveys by mitigating systematic effects and residual foreground contamination [171][170].

This chapter therefore aims to extend previous work [13] and investigate the use of HI intensity maps for clustering-based redshift estimation. I took a simulation-based approach and attempted to recover the redshift distribution for an optical galaxy catalogue that was treated as the 'unknown' redshift sample. This was done through cross-correlations with HI intensity maps (the 'reference' sample) which was simulated from the same catalogue so that they share a clustering signal. I can then compare the estimated redshift distribution with the true distribution of that catalogue.

3.2 Simulations

For this work I principally make use of the S^3-SAX simulation [150] for investigating the limitations of using HI intensity maps for clustering-based redshift estimation. However, I also make use of other simulations depending on the specific requirements of our tests. When seeking to demonstrate the calibration capability on photometric redshifts we require a catalogue which has robustly simulated photometry (discussed in Section 3.2.2.1). When seeking to test the HI intensity maps at low resolutions we require a simulation covering a much larger sky area (discussed in Section 3.4.3.1).

I begin by discussing the S^3-SAX catalogue which is used for the majority of this work since it contains simulated HI information for all its galaxies. This is a semi-analytic simulation of a sky field with apparent HI emission properties for approximately 2.8×10^8 galaxies in a virtual observing cone whose properties have been derived from the Millennium dark matter simulations [199]. The catalogue I extract from S^3-SAX contains galaxies spanning 36deg^2 and extends up to a redshift of $z = 3$, which approximately covers the redshift range of forthcoming stage-IV photometric telescopes which could benefit from our type of clustering-based redshift estimation. From this catalogue I use the columns for right ascension, declination, apparent redshift (which includes peculiar velocities), and HI-mass.

By using a galaxy catalogue from a simulation like this, we can construct realistic HI intensity maps from the integrated effect of apparent properties of each contributing galaxy. Furthermore, since the S^3-SAX catalogue already considers cosmological effects such as redshift space distortions, these will propagate into our adapted catalogues making them a robust reflection of a realistic clustering-based redshift experiment.

From the S^3-SAX catalogue we can construct two samples (explained in Sections 3.2.2 and 3.2.1 respectively) which I will refer to as

- Optical galaxy catalogue (subscripted with g)
- HI intensity maps (subscripted with HI).

The optical galaxy catalogue is the catalogue that I will be treating as our sample of 'unknown' redshifts, and for which I will try to recover the true redshift distribution. I will only need the galaxy positions from this catalogue, and from these I can construct a number density field $n_{\rm g}$ by binning each galaxy into a pixel.

The intensity maps will be thin slices in chosen intervals of redshift space and as is commonly the case with intensity maps, each slice will be a field of brightness temperature $T_{\rm HI}$ where regions of higher temperature indicate a higher matter density. Figure 3.1 shows the distribution of galaxy HI brightness ($I_{\rm HI}$) contained within our full S^3-SAX catalogue.

For maps produced using the S^3-SAX simulation we use a resolution of 2 pixels per arcminute which corresponds to 720 × 720 pixels maps for our 36deg^2 patch of sky. I also restrict the catalogue to redshifts of $0 < z < 3$ and use 30 redshift bins giving bin widths of $\Delta z = 0.1$. For the number of S^3-SAX galaxies contained within these ranges this gives an average number density of 4.6 galaxies per voxel.

3.2.1 Simulating HI Intensity Maps

While traditional optical galaxy surveys aim to resolve their targets and build a catalogue of discrete objects above some lower flux detection limit, intensity mapping instead collects flux from all sources of emission, even the very faint ones, to build a continuous map of intensity. I therefore choose not to place any limits on which HI emitting galaxies to include in our simulation to make this as realistic as possible. In other words, every galaxy within the S^3-SAX catalogue that has a non-zero amount of HI emission, regardless of how faint, is included as a

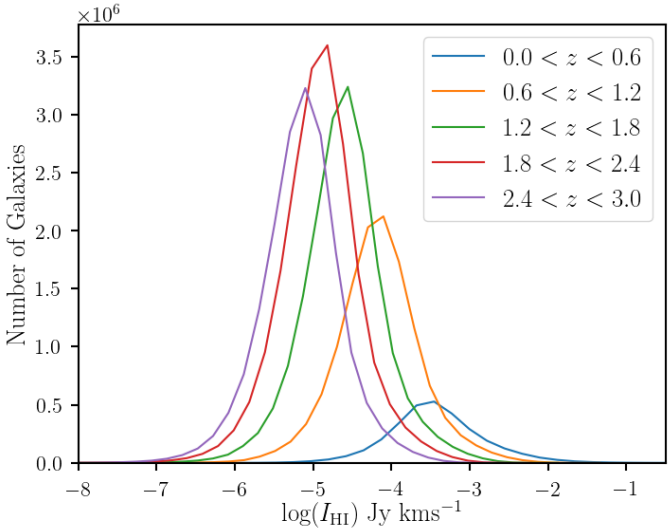

Figure 3.1: HI brightness histograms for galaxies in the S^3-SAX catalogue for different redshift bins. This shows the range of fluxes which will contribute to our HI intensity maps.

contributor to our HI intensity map. Note however that we can only include galaxies which are above the simulation completeness limit. In the case of the S^3-SAX catalogue, the simulation is complete for galaxies with cold hydrogen masses (HI + H II) above $10^8 M_\odot$.

We express our HI intensity map data $T_{\rm HI}^{\rm obs}$ in the form of a brightness temperature with two angular dimensions ($\theta_{\rm ra}$ and $\theta_{\rm dec}$, jointly represented by $\vec{\theta}$) and a radial dimension which is the redshift (z). The intensity map can be decomposed into three different map contributions

$$T_{\rm HI}^{\rm obs}(\vec{\theta},z) = s(\vec{\theta},z) + f(\vec{\theta},z) + n(\vec{\theta},z). \tag{3.1}$$

Here s represents the true HI signal we are aiming to detect, f are the radio foregrounds and n is noise associated with instrument systematics. The overall aim for successful intensity mapping is to therefore isolate $s(\vec{\theta},z)$ by subtracting or minimising the other unwanted components. I will discuss each of these components further in the following sections together with our method for simulating the "cleaned" data, i.e. the HI maps after some foreground cleaning technique has been applied.

3.2.1.1 Signal

Much of this is discussed in more detail in Chapter 2 but I reiterate here for convenience. To construct the intensity mapping signal I start with the HI mass $M_{\rm HI}$ of each galaxy, which is estimated in the S^3-SAX catalogue. Note that in any case one can use the formula outlined in [28] to infer $M_{\rm HI}$ from the raw signal $S_{\rm obs}{\rm d}v$, which is the flux integrated over a velocity width to capture the full HI signal that is stretched in frequency due to the galaxy's rotational velocity;

$$M_{\rm HI} = \frac{2.35\times10^5 M_\odot}{1+z}\frac{S_{\rm obs}{\rm d}v}{\rm Jy\ km\ s^{-1}}\left(\frac{{\rm d_L}(z)}{\rm Mpc}^2\right). \tag{3.2}$$

This can be inferred from the brightness equation outlined in (2.14). I then place the galaxies into a data cube with coordinates $(\theta_{\rm ra}, \theta_{\rm dec}, z)$ by binning each galaxy's HI mass into its relevant pixel so I end up with a gridded HI mass map $M_{\rm HI}(\vec{\theta}, z_c)$.

I can then convert this into a brightness field for a frequency width of $\delta\nu$ subtending a solid angle $\delta\Omega$ (which is effectively the pixel size)

$$I_{\rm HI}(\vec{\theta}, z) = \frac{3h_{\rm P}\nu_{21}A_{10}}{16\pi m_{\rm H}} \frac{1}{(1+z)^2 d_{\rm c}^2(z)} \frac{M_{\rm HI}(\vec{\theta}, z)}{\delta\nu\delta\Omega}. \tag{3.3}$$

which is a repeat of equation (2.14) I derived in Chapter 2 where $h_{\rm P}$ is the Planck constant, A_{10} the Einstein coefficient which quantifies the rate of spontaneous photon emission by the hydrogen atom, $m_{\rm H}$ is the mass of the hydrogen atom, ν_{21} the rest frequency of the 21cm emission and $d_{\rm c}(z)$ is the comoving distance out to redshift z (we will assume a flat universe).

As already mentioned, it is conventional in radio astronomy, in particular intensity mapping, to use brightness temperature which can be defined as the flux density per unit solid angle of a source measured in units of equivalent blackbody temperature. Hence, the intensity $I_{\rm HI}(\vec{\theta}, z)$ can be written in terms of a black-body temperature in the Rayleigh-Jeans approximation $T = Ic^2/(2k_{\rm B}\nu^2)$ where $k_{\rm B}$ is the Boltzmann constant. Using this we can estimate the brightness temperature at redshift z

$$T_{\rm HI}(\vec{\theta}, z) = \frac{3h_{\rm P}c^2A_{10}}{32\pi m_{\rm H}k_{\rm B}\nu_{21}} \frac{1}{[(1+z)d_{\rm c}(z)]^2} \frac{M_{\rm HI}(\vec{\theta}, z)}{\delta\nu\delta\Omega}. \tag{3.4}$$

Note I have used the notation $T_{\rm HI}$ to distinguish this raw signal from the true observed data $T_{\rm HI}^{\rm obs}$ outlined in (3.1), which includes the foreground and noise components. Lastly, to model the low angular resolution of an intensity map, I convolve $T_{\rm HI}$ with a telescope beam in Fourier space making use of the convolution theorem. Our telescope beam is modelled as a symmetric, two-dimensional Gaussian function with a full width half maximum of $\theta_{\rm FWHM}$ acting only in the directions perpendicular to the line of sight (as the frequency/redshift resolution is excellent).

Our clustering-based redshift method will cross-correlate optical galaxies with 2D angular intensity maps at various redshifts. I therefore choose to slice the intensity maps into thin tomographic redshift bins and collapse these to a 2D slice. The width of each tomographic redshift bin needs to be thin enough that I can make certain thin bin assumptions, yet wide enough that I allow for sufficient structure to obtain a strong cross-correlation signal. By thin bin assumptions I am referring to cosmological quantities such as the bias, which I assume to be constant within the width of the bin. This is discussed in more detail in Section 3.3. An example of a completed intensity map tomographically sliced and collapsed into a 2D angular map is shown in far-left map of Figure 3.2.

3.2.1.2 Foregrounds

Foregrounds are the main focus of investigation in Chapter 4, and I therefore refer the reader there for a more complete discussion. In this chapter I use a more concise introduction and treatment of foregrounds.

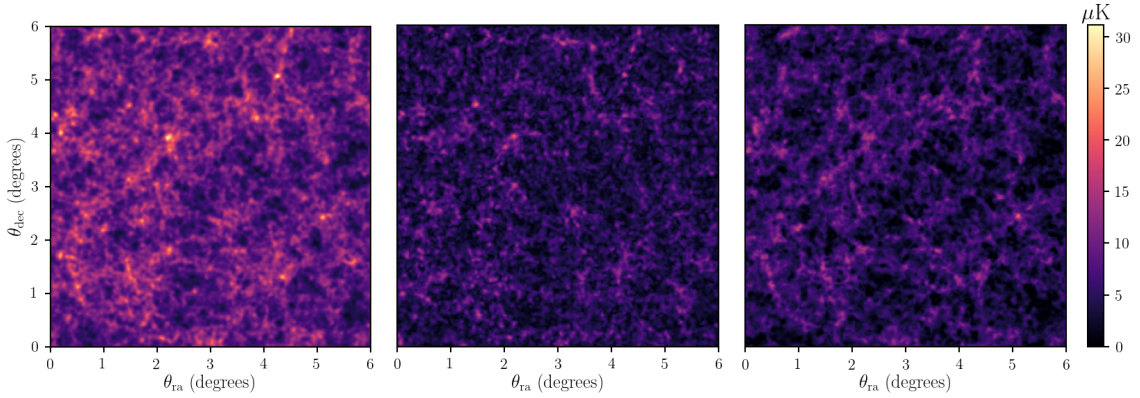

Figure 3.2: Example of a HI intensity map from our simulation using S^3-SAX catalogue galaxies. This particular example is a slice taken at $1.3 < z < 1.4$ with $\theta_{\rm FWHM} = 4'$. Far-left map shows the raw signal with no foreground contamination, centre map shows the same signal but with some large radial modes removed from the data to simulate some of the effects of a foreground clean as explained in Section 3.2.1.2. Differences can be seen by eye between these two but I also include the residuals in the far-right map to clarify the impact.

Arguably the biggest obstacle facing intensity mapping is the presence of foregrounds which emit signals below 1400 MHz and which can be several orders of magnitude brighter than the HI signal we are aiming to detect [228][188]. The term 'foregrounds' is perhaps misleading as some of these contaminants do not necessarily lie in front of the HI emitters. However, it is a term that is widely used in other literature so I will also adopt it here. The different types of foregrounds include

- Galactic synchrotron: Caused by high-energy cosmic ray electrons accelerated by the Galactic magnetic field. This is the most dominant of the foregrounds and has the added complication of being polarized.
- Point sources: Emission from extragalactic point radio sources e.g. AGNs. These can potentially cluster in the same way as the HI signal.
- Galactic & extragalactic free-free emission: caused by free electrons accelerated by ions, which trace the warm ionised medium both within the Milky Way and in the broader cosmic field.

Modelling and addressing the foreground removal problem with dedicated simulations is a very active area of research (see, for example, [228][196][11][230]). The conclusion of such work is that component separation methods can in principle be used to remove these foregrounds. The general idea is that the HI signal spectra fluctuate in frequency whereas the foreground spectra are expected to be smooth with long frequency coherence thus making them distinguishable. However, foreground cleaning based on this approach is typically more efficient on small scales i.e. small radial modes. On larger scales the HI signal is more similar to the foregrounds, so the result of these types of foreground cleaning can render larger radial modes useless. This has particular importance in the context of using HI intensity maps for clustering redshift estimation since information in these modes which could be utilised by the estimator are lost.

In this chapter, rather than simulating full foreground maps, adding these onto our signal to contaminate it and then applying some removal technique, I will instead bypass this step and aim directly to simulate the aforementioned effects of foreground cleaning by removing large radial modes from the data. I follow [13] in imposing that any comoving radial wavenumber $k_\parallel$ below a certain scale $k_\parallel^{\rm FG}$, where

$$k_\parallel^{\rm FG} \approx \frac{\pi H(z)}{c(1+z)\xi} \tag{3.5}$$

is rendered inaccessible by foreground cleaning; I therefore remove these modes. Here, ξ parametrises the characteristic frequency scale over which foregrounds are separable from the signal. In other words a smaller value of ξ means more large scale signal is lost, hence we need to remove a higher number of modes. To allow finer control of this foreground removal in fourier space I chose to increase the resolution in the line of sight direction by splitting the redshift bins into 5 pixels each giving our S^3-SAX intensity map cube 150 pixels along the line of sight. I will investigate the effect of different values of ξ in Section 3.4.2.

Hence, the recipe for simulating the effect of foreground cleaning can be summarised as

(i) Fourier transform the $T_{\rm HI}$ data cube;
(ii) Eliminate (set to zero) all pixels where $k_\parallel < k_\parallel^{\rm FG}$;
(iii) Inverse Fourier transform back.

The result of this process is an intensity map cube with some large radial modes lost. We can visualise the effects this has on the tomographic slices of intensity maps in Figure 3.2.

This method of simulating foreground removal is a crude approximation of the problem and I appreciate that this approach assumes all other modes above $k_\parallel^{\rm FG}$ are cleaned with 100% efficiency, which is of course an optimistic expectation. This has particular relevance for the auto-correlations since we expect foreground systematics to be a much bigger problem compared to the cross-correlation. However, the main issue when using intensity mapping for clustering-based redshift estimation is a loss of signal-to-noise on foreground dominated modes, rendering them ineffective when using them in correlation functions. With this in mind, we can test the main limitations of foreground cleaning by subtracting large radial modes from our reference sample.

3.2.1.3 Noise

Systematic effects and noise typically associated with radio surveys will impact intensity maps. Again, simulating those in detail would be a paper in its own right; for example [94] investigate the effects of $1/f$ noise (an instrumental effect that results in multiplicative gain fluctuations) in single-dish observations.

We can partially justify omitting survey specific additive systematic effects since we would expect these to drop out in any cross-correlations between intensity maps and optical surveys. I discussed this in section 2.2.4 and demonstrated this with an example in equation (2.6).

Here I used the fact that the signal-noise cross terms are uncorrelated and that the noise maps from each survey will be uncorrelated too. Strictly speaking, while this argument is valid for the expected cross-correlation, it is not valid for the uncertainties on that cross-correlation. For example if we have some survey-specific large-scale noise, it will cancel out in the cross-correlation, but will still dominate the error budget on that cross-correlation on large scales. It is also worth noting that we make use of auto-correlations in our work too, and in these situations the above argument does not apply. However, for the purposes of this chapter, I assume all instrumental systematics are either negligible or drop out and do not cause any contamination in the results. I leave a full simulation involving noise maps, which will look into whether realistic telescope noise levels are sub-dominant, for future work.

3.2.2 Optical Galaxy Sample

It is important that the optical galaxy samples have realistic redshift distributions which tail off at higher redshifts where resolved detection becomes more difficult. We therefore choose not to use all galaxies in the simulated catalogue, but instead randomly exclude galaxies in each redshift bin until a model distribution is achieved. This also means that the optical galaxy redshift distribution will differ from the distribution of the galaxies which contribute to the HI intensity maps, where we use all galaxies available. This makes for a more realistic test of this method too. For our optical model redshift distribution we use

$$\frac{\mathrm{d}N_\mathrm{g}}{\mathrm{d}z} = z^\beta \exp\left(-\left(\frac{z\alpha}{z_\mathrm{m}}\right)^\gamma\right) \tag{3.6}$$

where we use $\alpha = \sqrt{2}$, $\beta = 2$ and $\gamma = 1.5$ to make the distribution representative of a typical stage-IV optical large scale structure survey such as LSST or *Euclid*. For the mid-redshift parameter z_m we use the mid-redshift for the particular simulated catalogue we are applying this to. For our S^3-SAX catalogue, this will be $z_\mathrm{m} = 1.5$.

3.2.2.1 Simulated Optical Photometry

Large community efforts are put into simulating photometric redshift uncertainty to allow investigation of non-linear effects on the power spectra (and biases) of various tracers, and the testing and validation of photometric redshift estimation codes (which typically produce highly non-Gaussian estimates with significant tails and catastrophic outliers) [138][129]. However, for the HI clustering redshifts considered here it is necessary to simulate HI emission which is correctly correlated with the optical emission measured by photometric surveys. This is particularly difficult as the galaxies making up much of the HI signal in intensity maps are expected to be within $\approx 10^9\, h^{-1} M_\odot$ halos [213], orders of magnitude below the halo masses relevant (and hence simulated) for optical surveys, particularly in simulation boxes large enough to supply the wide and deep light-cones relevant to intensity mapping experiments. Given the potential utility of HI clustering redshifts, and other interest in cross-correlation of stage-IV radio and optical surveys e.g. [12][96][170], such a simulation is clearly needed, and we expect it to be pursued in further work.

For now, we take two approaches. For our principal results making use of the S^3-SAX simulation, we defer this problem, estimating the full redshift distribution for the sample, rather than binning according to an estimated photometric redshift. As described in section 3.4.4 we also investigate the ability to calibrate realistic simulated redshifts for the LSST telescope, but using HI intensity maps which only contain HI emission from the optically selected galaxies which we generate ourselves by using a HI-mass halo-mass relation.

3.3 Estimator Formalism

In this section I discuss the formalism associated with our method and provide a step-by-step construction of the estimator we use to make redshift predictions for the 'unknown' optical photometric sample.

Firstly, from the optical galaxy catalogue, I take the true galaxy redshifts and build a normalised redshift distribution given by

$$\frac{\mathrm{d}N_{\mathrm{true}}}{\mathrm{d}z}(z) = \frac{N_{\mathrm{g}}(z)}{\sum_i N_{\mathrm{g}}(z_i)} \frac{1}{\Delta z} \tag{3.7}$$

where $N_{\mathrm{g}}(z)$ is the galaxy count in a given redshift bin. I normalise by dividing through by all galaxies in each i-bin and by the redshift bin width Δz. The aim of this work is to be able to recover this true redshift distribution. In this work the chosen approach for doing this is to utilise angular correlation functions. I start by binning the HI intensity map into thin tomographic redshift slices and take the observable HI brightness temperature fluctuations δT_{HI} for each slice defined as

$$\delta T_{\mathrm{HI}}(\vec{\theta}, z) = T_{\mathrm{HI}}(\vec{\theta}, z) - \overline{T}_{\mathrm{HI}}(z)\,, \tag{3.8}$$

where a barred quantity denotes the average value for the particular field. I also take the optical galaxy count overdensity δ_{g} for the full redshift range defined as

$$\delta_{\mathrm{g}}(\vec{\theta}) = \frac{n_{\mathrm{g}}(\vec{\theta}) - \overline{n}_{\mathrm{g}}}{\overline{n}_{\mathrm{g}}}\,. \tag{3.9}$$

I then calculate the angular cross-correlation between each HI slice $\delta T_{\mathrm{HI}}(\vec{\theta}, z)$ and the unknown-redshift optical galaxy overdensity $\delta_{\mathrm{g}}(\vec{\theta})$;

$$w_{\mathrm{g,HI}}(z) = \langle \delta_{\mathrm{g}}(\vec{\theta}) \delta T_{\mathrm{HI}}(\vec{\theta}, z) \rangle \tag{3.10}$$

where the angled brackets signify an averaging over all positions in the field. This approach is therefore only focusing on the zero-lag of the angular correlation function, as I am only averaging over pixels in each map which share the same position $\vec{\theta}$. Previous clustering redshift works using resolved galaxy positions for both samples tend to extend beyond the zero-lag and attempt to gain more signal from the full-correlation function at extending separations. They then weight their correlation function such that it delivers the best signal-to-noise. For example, [85] and [61] average their correlation function w over a separation range such that

$$\overline{w}(z) = \int_{R_{\mathrm{min}}}^{R_{\mathrm{max}}} W(R)\, w(R, z)\, dR \tag{3.11}$$

where R is the separation distance between galaxies being correlated and $W(R) \propto R^{-1}$ is a weighting function, whose integral is normalised to unity and constructed to give higher weight to smaller scales; this maximises the signal-to-noise of the correlation function. They choose to use integration limits of 500 kpc and 1500 kpc and discuss how including larger scales tends to give a poorer signal-to-noise while smaller scales are more likely to suffer from non-linear bias.

Since I am using low resolution maps and correlating pixels rather than resolved galaxies, the choice is somewhat simplified. The low resolutions I use, which are constrained by the intensity mapping instrument's beam size, mean that often one or two pixels are representative of the preferred separations probed by the resolved optical galaxy clustering redshift methods. Also, given that the weight function prioritises smaller scales, the full-correlation function method will be very similar to using the zero-lag at the low resolutions I work with. I experimented with this using a maximum separation of $R_{\max} = 1500$ kpc which, for the resolutions used on the S^3-SAX catalogue, corresponds to 1 pixel of separation for $z \geq 1.95$, 2 pixels for $0.95 < z < 1.95$ and only reaching 16 and 9 pixels of separation for the lowest two redshift bins. Only very small deviations from the zero-lag approach would therefore be expected given this and I do in fact find that the results from the two approaches converge in the regime where the full correlation function is tuned to maximise the signal-to-noise ratio.

Where we have strong correlation I infer that the particular redshift bin is well represented in the overall redshift distribution i.e. I suppose

$$\frac{\mathrm{d}N_{\mathrm{g}}}{\mathrm{d}z}(z) \propto w_{\mathrm{g,HI}}(z)\,. \tag{3.12}$$

To understand the full version of this equation and build an estimator for $\mathrm{d}N_g/\mathrm{d}z$ we must consider the clustering amplitudes (bias terms), the underlying dark matter density, and the relationship between them. We can begin by looking at the δ_{g} and δT_{HI} fields separately. Firstly, under the assumption of linear and deterministic biasing (expected to be accurate on large scales), we have

$$\delta_{\mathrm{g}} = \int_0^{z_{\max}} b_{\mathrm{g}}(z)\delta(\vec{\theta}, z)\frac{\mathrm{d}N_{\mathrm{g}}}{\mathrm{d}z}(z)\mathrm{d}z \tag{3.13}$$

where b_{g} is the bias for the optical galaxies, δ is the dark matter over-density field and $\mathrm{d}N_g/\mathrm{d}z$ represents the normalised redshift distribution. Similarly, for the HI brightness temperature fluctuations we have

$$\delta T_{\mathrm{HI}} = \int_0^{z_{\max}} \overline{T}_{\mathrm{HI}}(z) b_{\mathrm{HI}}(z)\delta(\vec{\theta}, z)\frac{\mathrm{d}N_{\mathrm{HI}}}{\mathrm{d}z}(z)\mathrm{d}z\,. \tag{3.14}$$

We can slice the reference intensity maps into appropriately thin redshift bins,

$$\frac{\mathrm{d}N_{\mathrm{HI}}}{\mathrm{d}z}(z) = \Theta(z_1, z_2)\,, \tag{3.15}$$

$$\Theta(z_1, z_2) = \begin{cases} 0 & z < z_1 \\ 1 & z_1 \leq z \leq z_2 \\ 0 & z > z_2\,, \end{cases} \tag{3.16}$$

where I have used the top-hat function Θ to take a slice of the HI intensity map. I now cross-correlate $\delta_{\rm g}$ and $\delta T_{\rm HI}$ for the redshift range chosen by Θ,

$$\langle \delta_{\rm g} \delta T_{\rm HI} \rangle = \iint \overline{T}_{\rm HI}(z') b_{\rm g}(z) b_{\rm HI}(z') \langle \delta(\vec{\theta}, z) \delta(\vec{\theta}, z') \rangle \frac{{\rm d}N_{\rm g}}{{\rm d}z}(z_{\rm c}) \Theta(z_1, z_2) {\rm d}z {\rm d}z'. \tag{3.17}$$

The top-hat function Θ restricts the integral to a thin redshift range and at this point I assume that I have picked a sufficiently thin bin width such that all terms become constant over this redshift range with central redshift $z_{\rm c}$, leading to

$$\langle \delta_{\rm g} \delta T_{\rm HI} \rangle = \overline{T}_{\rm HI}(z_{\rm c}) b_{\rm g}(z_{\rm c}) b_{\rm HI}(z_{\rm c}) \langle \delta(\vec{\theta}, z_{\rm c}) \delta(\vec{\theta}, z_{\rm c}) \rangle \frac{{\rm d}N_{\rm g}}{{\rm d}z}(z_{\rm c}) \Delta z. \tag{3.18}$$

Here, Δz appears from the Limber approximation [127]. This assumes the coherence length of the correlation function is relatively small and inside a thick enough bin will not significantly evolve. Therefore, I assume zero correlation outside the redshift range, so I just integrate over the small dz segments where non-zero signal exists. Δz therefore represents the bin width. $\langle \delta_{\rm g} \delta T_{\rm HI} \rangle$ is the zero-lag angular cross-correlation statistic where I average over all positions in the field as expressed in equation (3.10) i.e. $w_{\rm g,HI} \equiv \langle \delta_{\rm g} \delta T_{\rm HI} \rangle$, so writing in this form gives

$$w_{\rm g,HI}(z_c) = \overline{T}_{\rm HI}(z_{\rm c}) b_{\rm g}(z_{\rm c}) b_{\rm HI}(z_{\rm c}) w_{\rm DM}(z_{\rm c}) \frac{{\rm d}N_{\rm g}}{{\rm d}z}(z_{\rm c}) \Delta z, \tag{3.19}$$

where $w_{\rm DM} = \langle \delta \delta \rangle$ is the dark matter auto-correlation function. We can make use of the auto-correlation of the intensity maps to eliminate the dark matter density auto-correlation $w_{\rm DM}$ from equation (3.19). This auto-correlation is derived using similar steps to those above and is given by

$$w_{\rm HI,HI}(z_{\rm c}) = \overline{T}^2_{\rm HI}(z_{\rm c}) b^2_{\rm HI}(z_{\rm c}) w_{\rm DM}(z_c). \tag{3.20}$$

Effects from foreground contamination and noise, which should largely drop-out in the cross-correlation, could potentially affect this auto-correlation. For deriving this estimator we assume these effects are minimal. Dividing equation (3.19) through by $w_{\rm HI,HI}(z_{\rm c})$ we therefore get

$$\frac{w_{\rm g,HI}(z_{\rm c})}{w_{\rm HI,HI}(z_{\rm c})} = \frac{1}{\overline{T}_{\rm HI}(z_{\rm c})} \frac{b_{\rm g}(z_{\rm c})}{b_{\rm HI}(z_{\rm c})} \frac{{\rm d}N_{\rm g}}{{\rm d}z}(z_{\rm c}) \Delta z. \tag{3.21}$$

Rearranging we get our final estimator for the redshift distribution

$$\boxed{\frac{{\rm d}N_{\rm g}}{{\rm d}z}(z_{\rm c}) = \frac{w_{\rm g,HI}(z_{\rm c})}{w_{\rm HI,HI}(z_{\rm c})} \overline{T}_{\rm HI}(z_{\rm c}) \frac{b_{\rm HI}(z_{\rm c})}{b_{\rm g}(z_{\rm c})} \frac{1}{\Delta z}.} \tag{3.22}$$

Since Δz is defined and $w_{\rm g,HI}$, $w_{\rm HI,HI}$ can be measured, we just need to know the factor $\overline{T}_{\rm HI} b_{\rm HI} / b_{\rm g}$ to recover our redshift distribution.

In our simulations $\overline{T}_{\rm HI}$ can easily be obtained, since we know the brightness temperature $T_{\rm HI}$ from each galaxy and therefore the average brightness temperature for the map. However, in reality, the actual observable is the brightness temperature fluctuation defined in equation (3.8). $\overline{T}_{\rm HI}$ is really an unknown quantity that needs to be inferred from our measurements. I discussed this in section 2.3.2 and in equation (2.23) which is repeated below

$$\overline{T}_{\rm HI} = 180 \Omega_{\rm HI}(z) h \frac{(1+z)^2}{H(z)/H_0} {\rm mK}, \tag{3.23}$$

with $\Omega_{\rm HI}$ the HI density (abundance). In principle $\Omega_{\rm HI}$ can be measured using the auto-correlation HI power spectrum with redshift space distortions, assuming a fixed fiducial cosmology [135][170]. This then gives a measurement of $\overline{T}_{\rm HI}$. In this work for simplicity I will assume $\overline{T}_{\rm HI}$ is known (or can be modelled accurately) and just use the mean of our catalogue brightness temperatures. Note that $\overline{T}_{\rm HI}$ is a global quantity which is defined, and can be measured, independently of a clustering redshift experiment, unlike similar normalisations for optical clustering redshift sub-samples, which will be unique to the tracer selection of each experiment. The only remaining factor to address is therefore the bias ratio, which I discuss in the following section.

3.3.1 Bias Treatment

Since using HI as a tracer of large scale structure as a way to explore cosmology is a relatively new concept, it is still unclear how biased this tracer is. In order to obtain the relevant factor $b_{\rm HI}/b_{\rm g}$, I take the simple approach of measuring the angular auto power spectra C_ℓ for both the optical number density field and the intensity maps. If we restrict to the large linear scales and neglect redshift space distortions, we can obtain the bias factor through

$$\frac{b_{\rm HI}(z)}{b_{\rm g}(z)} = \frac{1}{\overline{T}_{\rm HI}}\sqrt{\frac{C_{\rm HIHI}(\ell,z)}{C_{\rm gg}(\ell,z)}}\,. \tag{3.24}$$

It is worth pointing out that this method uses the power spectra in each redshift bin for both intensity maps and opticals. This therefore relies on the optical galaxies being binned by redshift, which is information I am assuming is poorly constrained, so the question of circularity arises. An approach that is viable is to bin the optical galaxies using the photometric redshifts, undergo our whole clustering redshift approach with this approximate bias ratio, and then refine and repeat so that self-consistency is reached.

The exact form of the neutral hydrogen bias is an area of active research [155][211][45][213] and recent detections in [18] relied on measurements from the ALFALFA survey [98] to obtain $b_{\rm HI}$ (see also the work by [152]). Furthermore, modelling bias amplitude differences between the reference and unknown samples is a problem that appears universal to clustering redshift methods. For example spectroscopic surveys cross-correlated with photometric surveys have not fully constrained these biases and offer a range of proposed solutions for addressing this in practice. In the context of this work, a further solution could be to build a model for the HI bias through its cross-correlation with a spectroscopic or weak-lensing survey. Again, it is worth pointing out that the HI bias may be determined independently of the clustering redshift survey, rather than in analyses where samples of optical galaxies are used, where the bias must be determined for the galaxy types making up that particular sample, which will be a function of the experiment.

For now we rely on the approach as outlined in equation (3.24) where we assume we can successfully obtain thin redshift slices in the optical sample and obtain perfect foreground removal (of the relevant modes) for the HI sample.

From our simulations I find that the bias factor is scale independent only at large scales, as expected. As the left panel of Figure 3.3 shows for an example redshift bin, we appear to have a

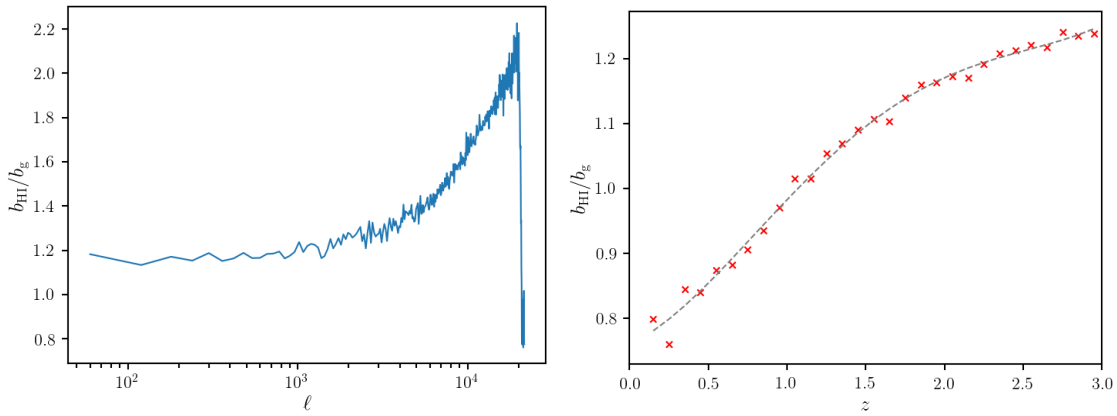

Figure 3.3: [Left] The bias ratio $b_{\rm HI}/b_{\rm g}$ as a function of angular scale at redshift $z = 2$. This ratio is only constant on the largest scales so we therefore choose to measure this bias at scales with $\ell < 10^3$. [Right] The bias ratio $b_{\rm HI}/b_{\rm g}$ as per equation (3.24) in each redshift bin with the grey dashed line showing a polynomial fit to the data points. As expected, the bias ratio that I use in our estimator evolves with redshift.

constant bias ratio on scales $\ell < 10^3$. I find a similar relation holds in all redshift bins. I therefore chose to take the mean value for this bias ratio at the angular scales of $\ell < 10^3$. The right panel of Figure 3.3 shows how the mean value of the bias ratio used in our estimator evolves with redshift.

One final point is that in the estimator, I choose to focus on the zero-lag of the correlation function, which includes small scales. However, we are only estimating the large-scale linear biases. This should not cause a problem since the small scale non-linear bias contributions are integrated out due to the low resolutions I am using. These are approximately 2 pixels per arcminute (or 2 pixels per 1 Mpc at $z \sim 1$). As we will see in the next section, our results show that the use of the zero-lag statistic in conjunction with large-scale linear biases appears not to cause any issues; but consideration should be given to this point when choosing bin sizes in real survey analyses.

3.4 Results & Discussion

Here I present the analysis and findings on the viability of using HI intensity mapping for clustering-based redshift estimation. Throughout I use the estimator as laid out in Section 3.3 and in particular equation (3.22) and proceed to investigate some of the properties that affect this method.

- I begin in Section 3.4.1 by examining the effect of HI-bright sources on the method using basic mock-catalogues which I simulate.
- In Section 3.4.2 I carry out the first test of the method using the adapted S^3-SAX catalogue (introduced in Section 3.2) which I construct realistic HI intensity maps from (albeit over a small sky area) and put particular emphasis on some of the effects from foreground cleaning.

- Section 3.4.3 looks at the Gaussian beam size θ_{FWHM} and whether increasing this to realistic amounts (comparable to some single-dish experiments such as the SKA) is too damaging to the redshift predictions. This relies on extending the simulation to a larger sky area so I make use of the MICE catalogue [78][58][79][44][102] which has a wider light-cone than S^3-SAX.
- I then finish in Section 3.4.4 by looking at how this method can provide excellent information on the error associated with stage-IV photometric redshifts. For this I use simulated LSST-like photometric redshifts from [19].

3.4.1 Bright HI-Rich Sources

Correlation functions in conventional optical surveys consider separation between different resolved point-like positions of galaxies. For intensity mapping, where we have different intensity objects binned into pixels, care needs to be taken when computing correlation functions for fields where there is not much signal or where extremely bright sources are dominating over the rest of the signal. This will cause the shot noise in the sample to increase since this is effected by the HI mass present i.e. the strength of the signal for intensity maps [45][198].

Having a HI-rich galaxy fall in a particular bin, whose signal vastly dominates over everything else in the field, could result in the rest of the field having essentially zero relative contribution to the signal. This can lead to the correlation function being shot-noise dominated. We want to try to avoid our fields having such extreme non-Gaussian properties, which constitute a poor representation of the underlying density field.

I investigated the effects of this behaviour by producing mock intensity maps and then contaminated them with dominant bright sources to see how this would affect the correlation functions and impact our clustering redshift method. I did this with a simplified model where we generate galaxies with a given distribution in redshift, simulate HI intensity maps with these galaxies, and then attempt to recover the redshift distribution with our clustering-based method. To initially ensure that no galaxy's flux was too dominant over the rest of the field I assigned all 10^7 galaxies in our mock a uniformly random HI flux emission between 0 and 1 (units are irrelevant for this mock example). For this simple model the input redshift distribution could be recovered since the intensity maps being produced were very uniform with Gaussian-like properties (see the blue triangle lines in Figure 3.4). However, it is possible that some galaxies will be several orders of magnitude brighter than the rest of the field as supported by the simulated fluxes from the S^3-SAX catalogue (Figure 3.1). So to investigate the effects of bright dominant sources I reassigned 1% of the galaxies in the mock catalogue a much higher HI flux emission, with uniformly random values between 1 and 10,000. At this point scaling problems were encountered in our mock situation along with large noise when recovering the redshift distribution, as shown in Figure 3.4. This shows that if bright sources dominate, they can contaminate the field and affect the results of the distribution recovery. The large scaling problem, shown by the red solid line in Figure 3.4, can be overcome since we are free to run a post-normalisation on the results to correct these scaling issues (shown by the dashed lines). However, the shape of the distribution still carries a large amount of noise for the contaminated case.

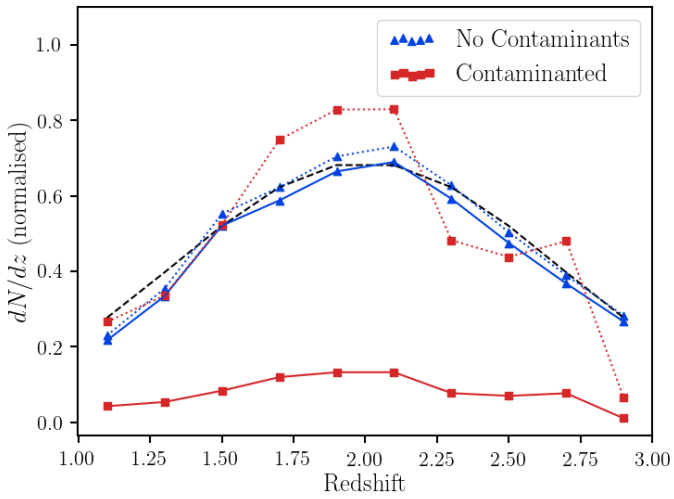

Figure 3.4: Mock simulation with an input redshift distribution (black dashed line) which I aim to recover. In the case where I have bright contaminating sources in our intensity maps (red square lines) our estimator struggles to recover this distribution presenting noise and scaling problems. However, results are improved when I remove these bright contaminants (blue triangle lines). The dotted lines in both cases show the results but normalised to unity to match the amplitude of the true redshift distribution which is also normalised to unity.

Therefore where possible, one should aim to avoid working with intensity maps where bright sources dominate the field and induce this extra noise in the correlation functions. An example of where this should be considered is when choosing which areas of redshift space to probe. Even when HI is mapped in a continuous way, as done with intensity mapping, the signal is still coming from discrete sources. At very low redshifts the survey volume is small, so the number of galaxies contributing to the intensity map is low making them more prone to bright source contamination and shot noise. This in turn makes them more likely to have non-Gaussian like fields leading to a poor distribution estimation for that redshift bin. It is therefore imperative to choose a redshift space region, and redshift bin width, which include sufficient numbers of contributing galaxies so that one does not produce shot-noise dominated intensity maps. For this reason I exclude low redshifts ($z < 0.1$) from all the catalogues I use and select a sufficient redshift bin width of either $\Delta z = 0.05$ or 0.1 depending on redshift range of the particular catalogue.

In reality, for intensity mapping experiments that are also performing HI galaxy surveys like SKA, it would be possible to remove the HI flux from a very bright source since it would likely be resolved in the HI galaxy survey. This flux-cutting approach represents an alternative way to alleviate the problem.

3.4.2 Foreground Removal

As described in Section 3.2.1.2, a key challenge when considering using HI intensity mapping methods for precision cosmology is foreground contamination. In this work I simulate some of the effect that foreground removal is expected to have on the recovery of the HI signal, which is

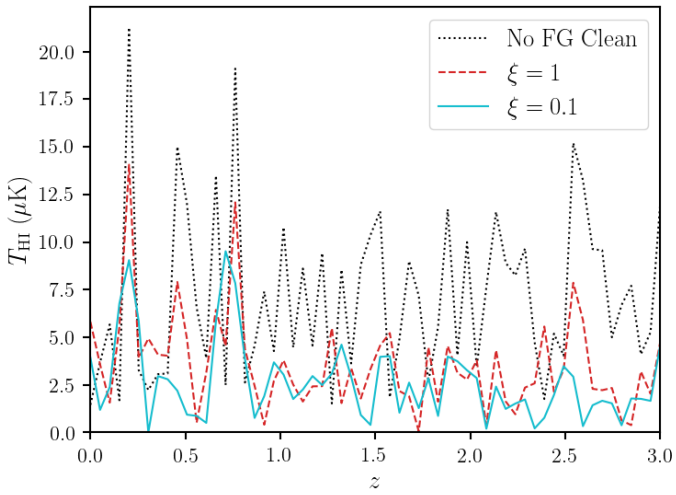

Figure 3.5: Demonstrates the effects of large radial mode removal (one of the effects expected from a foreground clean) and how lowering the parameter ξ, which translates to assuming a harsher foreground clean, gives data less representative of the original signal (the dotted black line). Done for random line of sight on our S^3-SAX catalogue.

to render a certain proportion of large radial modes useless. In reality, all modes would suffer some degree of foreground contamination as foreground cleaning can never separate signal and foregrounds with 100% efficiency. But it is largely considered that the large scale modes in the line-of-sight direction are the least separable from foregrounds [196] and therefore these will be rendered useless.

I follow the recipe laid out in Section 3.2.1.2 and eliminate any radial wavenumber that has $k_\parallel < k_\parallel^{\rm FG}$, where $k_\parallel^{\rm FG}$ is defined in equation (3.5), to emulate the main impact of a foreground clean on our data. The ξ parameter in equation (3.5) parametrises the foreground removal whereby a lower ξ equates to more radial modes being lost, signifying a harsher foreground clean.

Figure 3.5 shows an example of the effect that this simulated foreground removal has on a random line of sight through redshift and shows, as we expect, a suppression of the large radial modes which gets more severe for a higher ξ. The impact this has on the actual maps was displayed in Figure 3.2. The expectation is that much of the angular clustering information still remains in the smaller scale modes that are left behind, which can still be exploited for a clustering-based redshift estimation.

Figure 3.6(a) presents the first result from a redshift estimation attempt using our method on the S^3-SAX catalogue. For the case with no foreground contamination I find that it is still beneficial (i.e. it improves the goodness-of-fit) to nullify just one slice of pixels in k-space that contains the largest radial modes. This represents information at $0 < k_\parallel < 0.7 \times 10^{-3} h\text{Mpc}^{-1}$ scales and since these scales are so large, no useful information exists there to be used in the estimator's matching process. In other words these scales just contribute noise and therefore it is not surprising that their removal improves results. However, as I start to subtract more slices of pixels and eliminating information at larger values of $k_\parallel$ we get a reduction in estimator

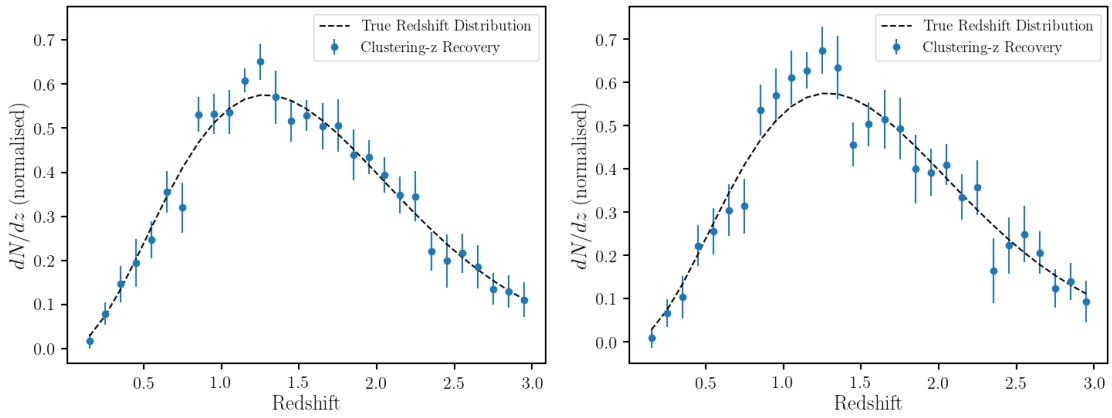

Figure 3.6: Results of using the HI intensity maps to recover the redshift distribution for the 'unknown' optical galaxies. The dashed lines show the true distribution which we seek to recover and the points are the estimator's prediction using a tomographic sliced intensity map at the particular redshift. Here I have used the S^3-SAX catalogue with $\theta_{\rm FWHM} = 4'$. (a) is the case with no foreground contamination and (b) is an example where I have applied our low-$k_\parallel$ cut with $\xi = 0.1$ to simulate a foreground clean. Error bars are obtained through jackknifing over 25 samples as explained in equation (3.25).

performance as desired to emulate a foreground clean. Figure 3.6(b) shows an example with our simulated foreground clean where I have used $\xi = 0.1$.

For these plots I have used a jackknifing technique to obtain the error bars. This was done by gridding the maps into an array of n smaller sub-samples, with $n = 25$. I then measure our estimator, which I here denote as $\hat{x}_i$, on the map but omit the i-th sub-sample. I repeat the procedure, averaging over the estimators obtained from omitting sub-samples, and obtain a standard deviation via

$$\sigma_{\rm error} = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{x}_i - \overline{\hat{x}})^2}\,. \tag{3.25}$$

Figure 3.6 suggests that even with quite a harsh foreground clean, a reasonable estimation of the redshift distribution of the optical galaxies can be made. A value of $\xi \approx 0.1$ corresponds to a cut that would target more complicated foreground residuals arising from leaked polarised synchrotron. Due to Faraday rotation these would exhibit a frequency structure which is not as spectrally smooth as other foreground contaminants hence making them more likely to remain after a mode cut [13].

The exact scales that are rendered inaccessible after a successful foreground clean is a subject still open for debate i.e. the most realistic value of ξ is unclear. Work by [196] proposes a foreground cleaning method which claims to render scales with $k_\parallel < 0.02\,h{\rm Mpc}^{-1}$ ($\xi \approx 0.05$ at $z = 1.5$) inaccessible, whereas there is more encouraging recent work by [240] which suggests that foreground cleaning is possible where information from these small $k_\parallel$ modes may not necessarily be lost at all. They propose using an extended method, Robust Principal Component Analysis (RPCA), which utilises the sparsity of the frequency covariance for the HI signal.

In Figure 3.7 we examine how various values of ξ affect the precision of our redshift distri-

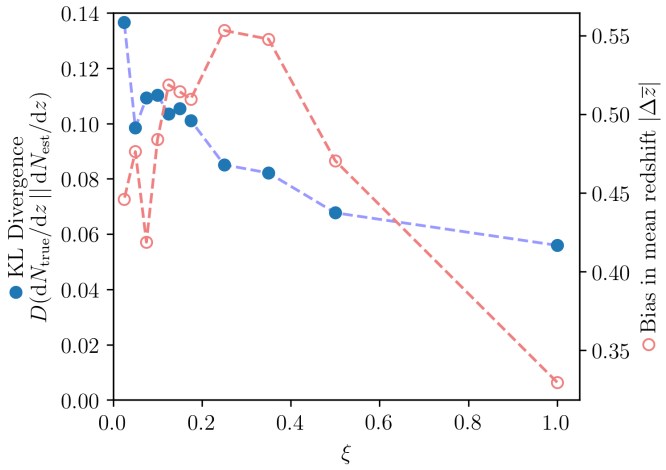

Figure 3.7: Test of estimator performance for differing levels of foreground cleaning parametrised by ξ. Shown is the Kullback-Liebler divergence D giving the information loss when describing the true redshift distribution with the estimated one (filled blue dots, left axis), and the bias in mean recovered for the redshift distribution (empty red dots, right axis). We see that the ability of the clustering estimator to recover the true distribution deteriorates as we increase the amount of foreground cleaning assumed (i.e. as we decrease ξ).

bution estimation, by analysing the Kullback-Liebler (KL) Divergence for different values of ξ as a figure of merit. The KL divergence $D(P\,||\,Q) = \sum_i P_i \log(P_i/Q_i)$ measures the information lost when an approximating discrete distribution Q is used to describe a true distribution P, providing a well-motivated way of estimating the goodness-of-fit across a whole distribution. Also shown is the mean recovered redshift for the distribution as a function of the same ξ. The plot is encouraging in showing that even when approaching conservative levels of foreground cleaning ($\xi \approx 0.1$), the degradation in performance is not significant when compared to the $\xi = 1$ case.

3.4.3 Varying Beam Size

Interferometric intensity mapping experiments such as CHIME (0.26° - 0.52°) [145] or HIRAX (0.08° - 0.17°) [146] have relatively good angular resolution. However, the proposed HI intensity mapping surveys using MeerKAT or SKA-MID in single-dish mode [190][189] are expected to have quite large beams and therefore a low angular resolution (greater than 1.4°). It is worth reiterating here that SKA will also operate as an interferometer, but I choose to focus on its use as a single-dish intensity mapping experiment to test the limitations of large receiver beams. In general, a single-dish intensity mapping experiment will typically have a beam size given by

$$\theta_{\mathrm{FWHM}} \approx \lambda / D_{\mathrm{dish}}\,, \tag{3.26}$$

where λ is the observing wavelength and D_{dish} is the dish diameter. So for an SKA-like intensity mapping experiment in single-dish mode, with dish diameters of $D_{\mathrm{dish}} = 15$m, targeting the redshifted $\lambda = 21$ cm signal we would expect to have a $\theta_{\mathrm{FWHM}} \approx 2$ deg at a median redshift of

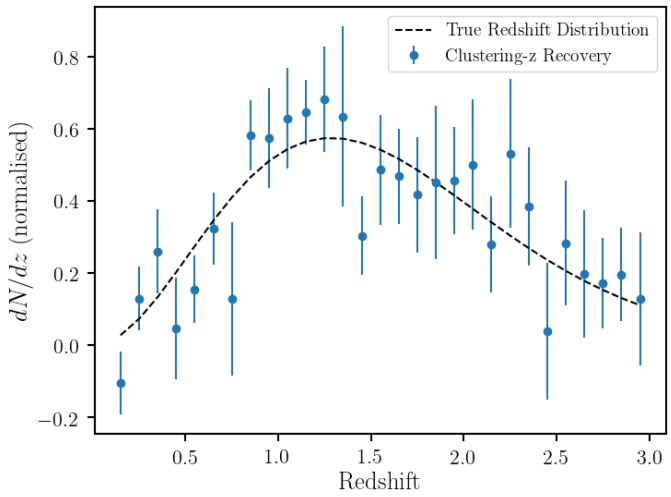

Figure 3.8: Increasing the beam size to $\theta_{\mathrm{FWHM}} = 16'$ for our S^3-SAX sample, which is equivalent to reducing the resolution of our experiment, causes errors to increase as predicted by equation (3.27).

$z = 1.5$. Unfortunately, for our simulations using the S^3-SAX catalogue we are limited to a small sky coverage of 6×6 square degrees, and this limits the extent to which we can increase our beam size. Since the error on our redshift estimation $\sigma_{N(z)}$ will be inversely proportional to the square root of the number of effective pixels in our field, and the number of effective pixels will just be the area of the whole field A divided by the area of our beam $\approx \theta^2_{\mathrm{FWHM}}$, we can estimate

$$\sigma_{N(z)} \propto \frac{\theta_{\mathrm{FWHM}}}{\sqrt{A}}. \tag{3.27}$$

We therefore find an increase in error as we explore lower resolutions. Even with no simulated foreground clean and only increasing the beam size to $\theta_{\mathrm{FWHM}} = 16'$, we are quadrupling our error and we find a large deterioration in the precision of our prediction as shown in Figure 3.8.

3.4.3.1 Testing on Larger Sky Area

Because of the rapid increase in error shown in Figure 3.8 from increasing the beam size to $\theta_{\mathrm{FWHM}} = 16'$, I proceeded to perform a scaled up test of clustering-based redshift estimation on larger sky areas to check if we can successfully go to higher values of θ_{FWHM}. To do this we require access to a catalogue with much larger sky coverage, so I chose to use the MICE simulation [78][58][79][44][102], which is a cosmological N-body dark matter only simulation resulting in a ≈ 200 million galaxy catalogue over a $5{,}000$ deg^2 area up to a redshift $z = 1.4$.

For these larger sky maps we use the HEALPix package [87] where the pixelation ensures that each pixel covers the same surface area as every other pixel. I handle the maps in HEALlPix RING ordering scheme with resolution `nside` $= 512$, which corresponds to $12 \times 512^2 = 3{,}145{,}728$ pixels across the full sky. Since the MICE catalogue covers angular coordinates in range $0 <$ ra, dec < 90 deg, these only fill $1/8^{\mathrm{th}}$ of the sky so I use 393,216 pixels for each map. 28 redshift bins are used between the redshift range of $0 < z < 1.4$ giving bin sizes of $\Delta z = 0.05$. For the number of MICE

galaxies contained within these ranges this gives an average number density of 18.6 galaxies per voxel.

Like I did when creating our optical galaxy sample from the S^3-SAX catalogue, I use equation (3.6) as our model for an optical redshift distribution with a mid redshift of $z_{\rm m} = 0.7$. This creates a realistic distribution in redshift for our opticals which tails off at higher redshift and that differs from the redshift distribution of the galaxies which contribute to the HI intensity maps.

Since this catalogue does not have apparent HI emission-line properties for each galaxy, we must derive our own HI masses for each galaxy. I therefore take each galaxy's halo mass as simulated by the MICE catalogue and convert this into a predicted HI mass by following the redshift dependent prescription laid out in [156]

$$M_{\rm HI} = 2N_1 M\Big[\Big(\frac{M}{M_1}\Big)^{-b_1} + \Big(\frac{M}{M_1}\Big)^{y_1}\Big]^{-1}, \tag{3.28}$$

where M is the galaxy's halo mass; M_1, N_1, b_1 and y_1 are all free parameters with redshift dependence tuned to provide a best fit; I refer the reader to [156] for details. From this I then follow the steps laid out in Section 3.2.1.1 and produce mock intensity maps.

It is important to highlight that in MICE, which is primarily a simulation for optical telescopes, the halos are only resolved down to a few $10^{11}h^{-1}M_\odot$ [58], and to build realistic intensity maps one would ideally want to go lower than this to ensure that HI emission from fainter galaxies is included in the intensity maps. However, for now it is sufficient to use this catalogue to demonstrate the potential of our method; improving the mass halo resolution will primarily change the bias on our over-density field representation, which is already well sampled.

An example of a HI intensity map produced from MICE is shown in Figure 3.9. Using these simulated intensity maps binned into suitable tomographic redshift slices of width $\Delta z = 0.05$, I attempt to recover the redshift distribution of an unknown optical galaxy population produced from this large sky catalogue. Figure 3.10 shows the results when using an angular resolution which varies with redshift as described by (3.26) to make the test representative of an SKA-like single-dish intensity mapping experiment beam. I also note that the increased shot noise from the higher mass cut applied to the MICE catalogue is highly sub-dominant to the beam size effect.

These results demonstrate that even with a large beam corresponding to an SKA-like single-dish HI intensity mapping experiment, an accurate redshift estimation can be made for the optical population. For cosmological HI intensity mapping surveys, telescopes may cover a sky area over 10,000 deg^2 (larger than the sky coverage from the MICE catalogue galaxies), which suggests that our results represent conservative forecasts since increased sky size should lower the errors as suggested by (3.27). Furthermore, it is worth reiterating that intensity mapping experiments such as CHIME and HIRAX will have better angular resolution (probing angular scales as low as 0.26° and 0.08°, respectively).

I note that these large sky maps do not include simulated foreground cleaning due to the added complexity of not being able to use the flat-sky approximation. However, the results obtained from Section 3.4.2 suggest that foreground contamination should not be a critical

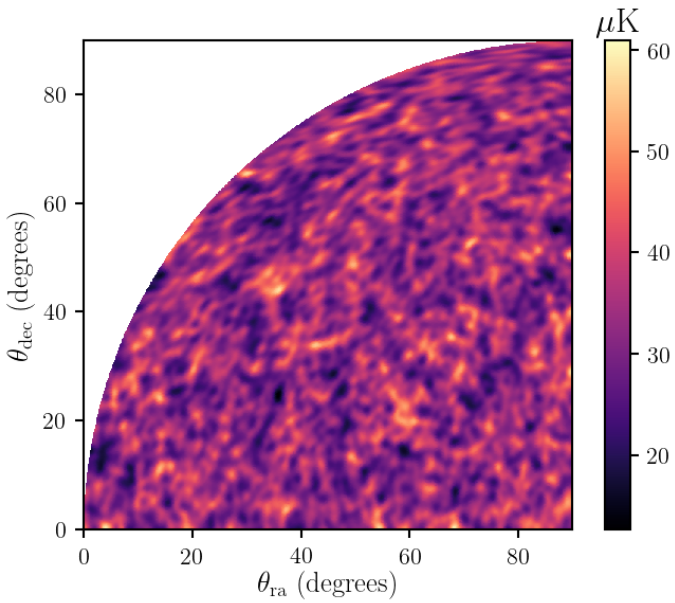

Figure 3.9: Large sky HI intensity map using MICE catalogue galaxies with halo masses converted into predicted HI masses. Since this is now a much larger patch of sky, we can no longer make the flat-sky approximation, and therefore I use a HEALPix projection for the map. This particular example is a slice taken at $0.60 < z < 0.65$ with $\theta_{\rm FWHM} \approx 1.3°$.

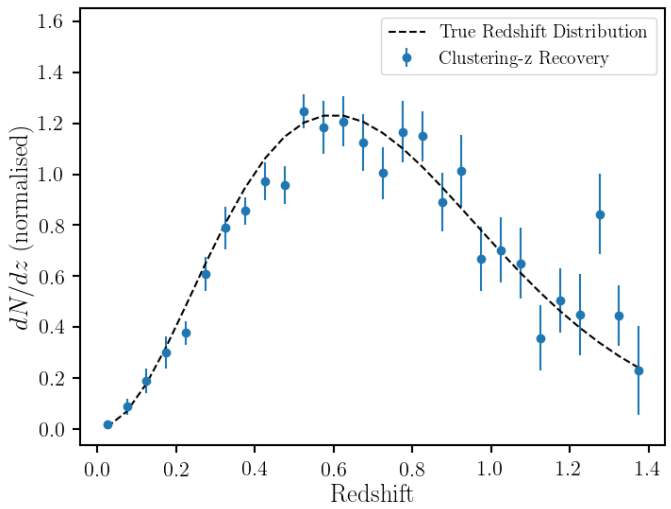

Figure 3.10: Results from using the large sky HI intensity maps to recover the optical redshift distribution. Here I have used the MICE catalogue with a frequency dependent beam size given by equation (3.26) for an SKA-like single dish experiment with a dish size of $D_{\rm dish} = 15$m.

problem for a clustering-based redshift estimation with intensity maps. This is investigated in more detail in the following Chapter 4, where more robust simulations are conducted.

3.4.4 Improvement on Photometric Redshift Measurements

The main aim of this work is to offer a new way of improving upon photometric redshifts, which is a major challenge for upcoming stage-IV optical telescopes like *Euclid* and LSST. In order to investigate this and build a pipeline we need a catalogue of galaxies for which there are robustly simulated photometric redshifts. As discussed in Section 3.2.2.1, in lieu of a simulation containing all the ingredients we would like, we choose to use a simulation that has robustly simulated photometry and then add HI emission to all galaxies using an analytical formula. Since we wish to emphasise the value of using our clustering redshift technique on future stage-IV surveys, we choose to make use of the simulated photometric redshifts from [19] (A^2A). This includes simulated LSST and *Euclid*-like photometry from the mock catalogues generated by [138] using the GALFORM semi-analytic code on light-cones extracted from the Millennium Simulation [199]. The A^2A catalogues also include photometric redshift estimates obtained using the BPZ estimation code [32], which is what we use when comparing a redshift distribution obtained using photometric redshifts against our clustering-based method.

Firstly, I show the performance that we can expect from an LSST-like experiment when trying to estimate the redshift distribution using photometric redshifts. This is displayed in Figure 3.11, and it shows significant deviation from the true redshift distribution. Here I bin the photometric galaxies using most-likely photometric redshift estimates obtained using the BPZ estimation code. Of course in reality LSST redshift catalogues will involve calibrations of and improvements over raw BPZ redshifts from the LSST bands, but here we simply seek to show how HI intensity mapping calibration can be one of these methods.

The A^2A catalogue I use extends to redshift $z = 3$ and covers a sky area of just over 25 deg^2. To simulate the HI mass for each galaxy I use equation (4.6) again as I did for the MICE catalogue in Section 3.4.3.1. From this I can again follow the steps laid out in Section 3.2.1.1 and produce mock intensity maps. The A^2A simulation has a mass resolution of $1.72 \times 10^{10}\, h^{-1} M_{\odot}$, which as discussed earlier means the simulated HI emission will not include faint HI emitters. This lack of completeness in our simulated intensity maps is not ideal but is likely to cause results to be worse than if we had more complete intensity maps; these would be a better representation of the underlying mass density and hence improve the precision of the correlation functions.

With only 2,950,025 galaxies in our A^2A catalogue, an angular resolution which is identical to our SAX simulation of 2 pixels per arcminute and 30 redshift bins over a $0 < z < 3$ range with $\Delta z = 0.1$, this gives a low number density of galaxies of 0.27 galaxies per voxel. Despite this a clustering redshift recovery is still possible.

One way of demonstrating the improvements we can make in constraining this distribution is to select a sub-population of galaxies between chosen photometric redshift limits. We can then examine the accuracy of the redshift distribution inferred from our clustering redshift method for this sub-population.

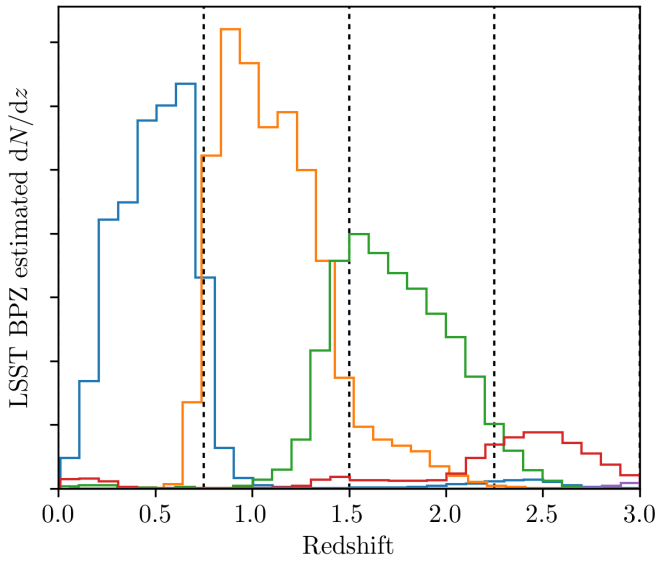

Figure 3.11: The performance of a simple redshift estimation with LSST bands from the A^{2}A catalogue. Here galaxies are binned (into the four bins indicated by vertical dashed lines) according to their most likely estimated redshift from running BPZ, with the histograms being of their true redshifts. This is equivalent to stacking $P(z)$ for individual galaxies in the case of Gaussian $P(z)$ with widths given by the BPZ widths.

The pink shaded regions in Figure 3.12 show various redshift intervals from which we are aiming to select galaxies. I select the galaxies using their known photometric redshifts and display their distribution with the orange line. From the photometric information alone one would conclude that a suitable population of galaxies has been selected with the desired redshift range. However, the black dashed line shows the true distribution, which extends significantly outside the claimed redshift interval. With our HI clustering redshift method we can estimate this true distribution thus allowing the experiment to calibrate the error on the photometric selection accurately. Figure 3.12 highlights both the need for methods that calibrate the photometric redshifts, and the potential success which our approach can have in providing this.

A further speculative approach, which is unlikely to go beyond a thought experiment level due to computational cost, would be to explore selecting galaxies from the unknown sample that maximise the correlation function signal. One could then claim that these galaxies fall within a certain redshift range based on the fact that they improve the correlation function with their inclusion. This can be put most simply by considering Figure 3.12. One could take the galaxies that make up the photometric redshift population as the 'first-guess' for exactly which galaxies lie within the target redshift range. Then, using a sophisticated trial-and-error approach, one could remove or add galaxies that bring the true distribution (predicted by the HI-clustering redshift estimation) into agreement with the targeted redshift range. As mentioned this would be a computationally expensive process but the final result, assuming one could avoid noise contaminating the final distribution, would be a population of resolved galaxies all of which have been predicted to fall within a redshift range, which one could arguably make thin.

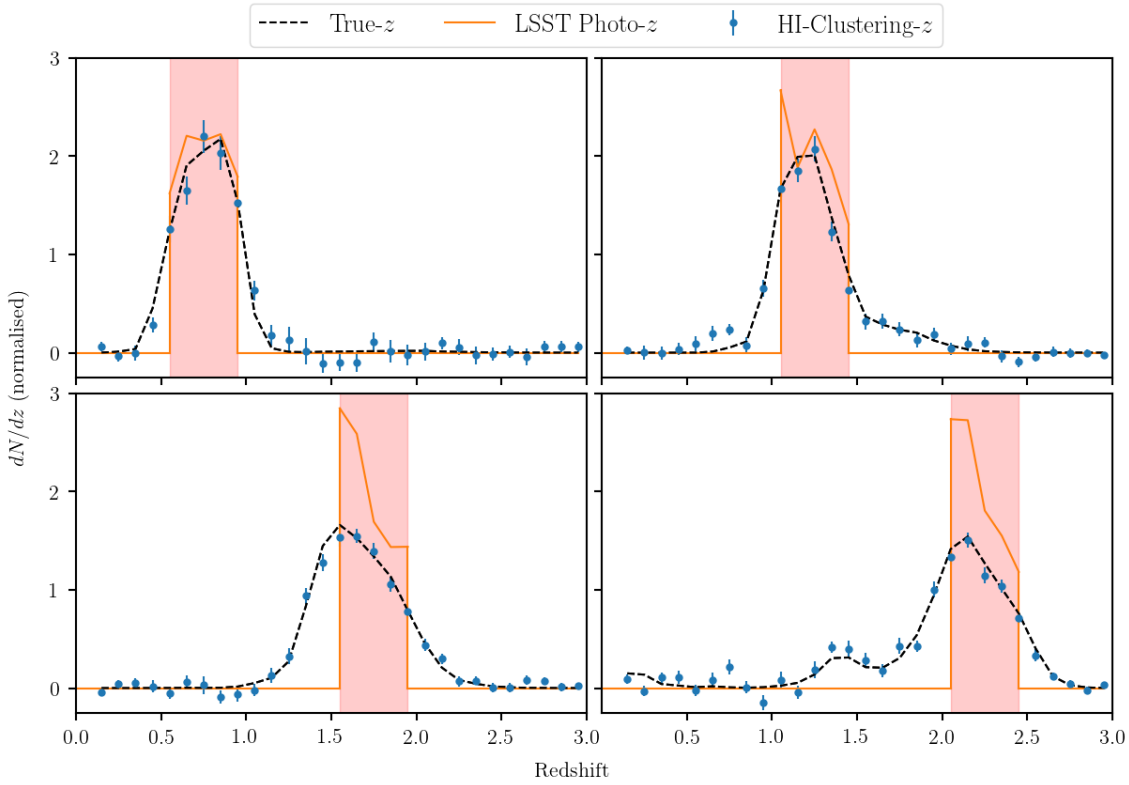

Figure 3.12: Complement to Figure 3.11 where I am now selecting galaxies based on their photometric redshift estimates. The pink shaded regions show the range in photometric redshift which galaxies are selected from. The orange line shows the distribution of these chosen galaxies according to their BPZ photometric redshift from LSST bands. The black-dashed line shows the true distribution, and the blue points show our HI clustering redshift estimate. This was done using the A^{2}A catalogue adapted to include HI emission information using equation (4.6). Given the small sky area, intensity map resolution was set to $\theta_{\rm FWHM} = 2'$.

Were this idea found to be feasible it would extend the clustering redshift method to be able to not just calibrate photometric redshift errors, but also actually improve the redshift estimates constraining them on the same scales as the bin width size ($\Delta z = 0.1$ and below).

3.5 Summary

By utilising realistic simulations of HI emission from galaxies, I have constructed HI intensity maps and provided evidence that they can be used to estimate the redshift distribution of a sample of optically resolved galaxies via the clustering cross-correlation method (Figure 3.6(a)). Our estimator uses the zero-lag element of the cross-correlation function between the intensity map and optical galaxy count field, rendering it computationally inexpensive. This computational efficiency, coupled with the fact that intensity mapping will be a much faster probe compared to a spectroscopic survey, means that the method I have presented is a rapid option for constraining the redshift distribution for a large population of galaxies. Next

generation surveys are promising to provide larger galaxy catalogues than ever, meaning that fast options for redshift constraints are likely to be in demand.

Given that experiments such as HIRAX, MeerKAT and the SKA have plans to operate as intensity mapping experiments in the near future, and CHIME is already taking data, a HI clustering redshift method has particular relevance for stage-IV optical surveys such as *Euclid* and LSST, which will all run at similar times. While surveys such as *Euclid* are planning to run their own spectroscopic experiments, these are time-consuming, and LSST will be purely photometric, so in each case HI intensity mapping clustering redshifts are likely to be useful. *Euclid* and LSST redshift ranges are accessible to planned intensity mapping surveys such as CHIME, HIRAX, MeerKAT and the SKA and the sky overlap between many of these optical and radio surveys is excellent too. Our results from Section 3.4.3.1 suggest that intensity mapping, even with the poor angular resolution that single-dish experiments are anticipated to have, can provide helpful redshift constraints on optical populations. It is also likely that these particular results are pessimistic since the intensity mapping experiments will most likely cover a larger sky area than that in the MICE simulation I used. Furthermore, in future the limit on halo mass resolution in simulations will decrease, emulating realistic HI intensity maps which include more faint galaxies, thus boosting the precision of the cross-correlations.

I have discussed the issue of modelling the linear bias, which is a problem that is inherent in all clustering redshift methods. This is arguably a more serious problem for the case of HI intensity mapping however, since the auto-correlations could potentially be further biased by contaminating foregrounds. I have made it clear that our idealistic approach of measuring the bias in our simulations would be difficult in reality; however, utilising cross-correlations with lensing data is one possible way to tackle this issue.

I also discussed some of the effects of foreground cleaning necessary for HI intensity maps to undergo. In the context of a clustering redshift method, the largest problem this poses is that a foreground clean on intensity mapping data affects large radial modes where the foregrounds are less distinguishable from the HI signal. I investigated this problem by removing large radial modes from our intensity maps to emulate this loss of information.

Our results, depicted in Figures 3.6(b) and 3.7, show that even with the loss of large portions of radial modes (low ξ), reasonable predictions of the redshift distribution can still be made. The fact that many large radial modes can be subtracted without too much damage to our method demonstrates that a lot of the useful matching information is in the small radial modes still exploited in the cross-correlations.

Further encouragement comes from recent work by [240], which proposes a foreground removal method that will not result in such losses of long-wavelength modes. This, together with our results, suggests that foreground contamination should not be an insurmountable problem for clustering-based redshift estimation involving HI intensity maps. However, this is investigated in more detail in the following Chapter 4 where more robust simulations are used.

I make this claim with a few caveats however, as there are still aspects of the foreground problem that require further exploration (some of these caveats are addressed in the following Chapter 4). Firstly, I have only investigated the impact foreground cleaning has on large radial

modes. Our method of simulating a foreground clean represents a basic approach to what is a very complex problem. It is true that foregrounds can also affect smaller scales similar to the beam size especially if considering impact from polarisation leakage. Furthermore, in the case of interferometers, additional complications need to be considered that are caused by the 'foreground wedge'. This is an effect that renders an area of k-space, known as the 'horizon-wedge', liable to foreground contamination that can be picked up from antennae with far side-lobe responses [194]. In Chapter 4 I will incorporate simulated foreground maps into our intensity maps and then proceed with a foreground cleaning algorithm; this is the only way to provide a fully realistic test of the effects of foreground removal.

Using simulated photometric redshifts from the A^2A catalogue I highlighted the potential improvements that could be made using clustering redshift estimation, as shown in Figure 3.12. This plot summarises the main point of the paper since it identifies that photometric redshifts have limitations in accuracy (especially at higher redshift) signalling the need for some accurate method of calibration, which clustering-based redshift estimation with HI intensity mapping offers.

Producing this work has also highlighted the need for catalogue simulations capable of being used to build realistic intensity maps, which also include simulated optical photometry, and cover a large sky area. This has been discussed throughout but I reiterate that a simulation which included

- simulated photometry for optically resolved galaxies so estimates using photometric redshifts can be done;
- simulated HI information for each galaxy for simulating realistic intensity maps;
- low halo-mass resolution ($\approx 10^9 h^{-1} M_\odot$) so intensity maps include integrated HI emission from faint galaxies;
- large sky-coverage ($\approx 10,000\,\mathrm{deg}^2$) to allow for testing low resolutions associated with a typical intensity mapping experiment's beam size

would be hugely beneficial not just for extending upon this work, but also for further exploration of potential synergies between optical and radio surveys.

The absence of such a simulation was significant when I extended our method to larger sky areas and quantified photometric redshift improvements. I used MICE and A^2A respectively and settled for generating our own HI emission for each galaxy using an analytical formula (equation (4.6)). Both of these catalogues however do not have sufficient halo mass resolution for realistic HI intensity maps. I have argued that this is only a limitation on current simulated tests and there is no reason to suppose that this will have over-inflated the effectiveness of this method. On the contrary, it is likely that obtaining lower mass-resolution, more complete HI intensity maps would improve our results since the more realistic intensity maps would be a closer representation of the underlying mass density providing the potential for more precise correlation functions.

Given that we are expecting huge increases in galaxy number densities from upcoming galaxy surveys, the strain placed on spectroscopic follow-up is also going to increase, therefore

motivating clustering-based redshift estimation methods. I believe that using HI intensity maps within such clustering redshift methods provides an exciting possibility that warrants further investigation.

In the following chapter, I will continue some of this investigation by including simulated 21cm foregrounds and a pipeline for removing their contamination. This will be provide an understanding of the impact of foregrounds on methods such as HI clustering-based redshift estimation and other cross-correlation techniques.

CHAPTER 4

FOREGROUND CONTAMINATION IN HI INTENSITY MAPS

Cunnington S., Wolz L., Pourtsidou A., Bacon D., (2019), Mon. Not. Roy. Astron. Soc., 488, 5452

This chapter presents work from the above published article [60] with amendments made to the original published script for the purposes of this thesis. For all parts of this chapter, including text and figures, I am the principal author with contributing edits from my co-authors.

The future of precision cosmology could benefit from cross-correlations between intensity maps of unresolved neutral hydrogen (HI) and more conventional optical galaxy surveys. A major challenge that needs to be overcome is removing the 21cm foreground emission that contaminates the cosmological HI signal. In this chapter, I again use N-body simulations to simulate HI intensity maps and optical catalogues which share the same underlying cosmology. In an extension from the previous chapter (Chapter 3), I also add simulated foreground contamination and use state-of-the-art reconstruction techniques to investigate the impacts that 21cm foregrounds and other systematics have on these cross-correlations. I find that the impact a FASTICA 21cm foreground clean has on the cross-correlations with spectroscopic optical surveys with well-constrained redshifts is minimal. However, problems arise when photometric surveys are considered: I find that a redshift uncertainty $\sigma_z \geq 0.04$ causes significant degradation in the cross power spectrum signal. I diagnose the main root of these problems, which relates to arbitrary amplitude changes along the line-of-sight in the intensity maps caused by the foreground clean and suggest solutions which should be applicable to real data. These solutions involve a reconstruction of the line-of-sight temperature means using the available overlapping optical data along with an artificial extension to the HI data through redshift to address edge effects. I then put these solutions through a further test in a mock experiment that uses a clustering-based redshift estimation technique to constrain the photometric redshifts of the optical sample as carried out in Chapter 3. I find that with my suggested reconstruction, cross-correlations can be utilized to make an accurate prediction of the optical redshift distribution.

4.1 Introduction

Methods involving detection of galaxies to trace large-scale structure are reliable providing that the galaxy samples obtained by a survey have a sufficient number density. If not, the measurements will suffer from significant statistical errors due to Poisson shot noise. Obtaining a large number of resolved galaxies with precise redshifts is expensive; spectroscopic redshifts with a redshift uncertainty $\sigma_z \sim 0.001$ rely on long integration times making this a slow process. Photometric redshifts offer a less precise alternative but can be obtained much more quickly allowing dense catalogues of galaxies to be built [41][74]. It is for this reason that future stage-IV surveys such as *Euclid*[1] [16] will heavily rely on photometric redshifts, and the Large Synoptic Survey Telescope (LSST)[2] [133] will be entirely reliant on them.

As an alternative, radio intensity mapping techniques, which do not rely on resolving individual sources, offer the prospect of more complete tracer maps with the redshift precision of a spectroscopic survey. In complete contrast to optical surveys, HI intensity mapping provides excellent constraints along the radial line-of-sight but poor angular resolution. This complementarity, together with the fact that cross-correlations are expected to alleviate survey-specific systematic effects, makes synergies between intensity mapping and optical galaxy surveys mutually beneficial.

The observed frequency (<1420MHz) of the photons emitted from HI places the signal in the radio part of the electromagnetic spectrum. Therefore radio dishes are the conventional choice of receiver for detecting these photons at low redshifts of $z < 3$. First detections using the HI intensity mapping technique have already been achieved in cross-correlation with optical galaxies in [162], [135] and [18].

The most prominent example of a next generation radio observatory is the Square Kilometre Array (SKA)[3] [23]. The mid-frequency instrument, SKA1-MID (where 1 stands for Phase 1), will be an array of 197 dish receivers that can operate in interferometer and single-dish mode. The low frequency instrument, SKA1-LOW, will probe the high redshift Universe, targeting the Epoch of Reionisation. As with any interferometer, it is the largest separation (or baseline) which determines the resolution of the instrument; hence baselines of up to 150 km are proposed to maximize resolution. Conversely, it is the smallest baselines between receivers which determines the largest scales that can be probed. The SKA1-MID instrument aims to perform a wide ($\sim 20,000$ deg^2) HI intensity mapping survey in single-dish mode. This compromises angular resolution but probes the large scales needed for cosmology.

The redshifted 21cm line signal from HI benefits from being particularly isolated in frequency, and there are few examples of spectral lines that could lead to potential line confusion, making HI intensity mapping particularly robust for redshift experiments. However, a major challenge for HI intensity mapping comes from foreground emission (e.g. synchrotron radiation), which can be orders of magnitude larger than the cosmological signal. Foregrounds are spectrally smooth signals which emit in the same range as the redshifted HI . Blind foreground removal

[1] www.euclid-ec.org/
[2] www.lsst.org/
[3] www.skatelescope.org/

21cm IM Survey		Photo-z Survey	$f_{\rm sky}$	$z_{\rm min}$	$z_{\rm max}$
MeerKAT	×	DES	0.1	0	1.45
TIANLAI	×	DECaLS	0.15	0	1.5
SKA1-MID	×	*Euclid*	0.2	0.35	2
SKA1-MID	×	LSST	0.4	0.35	3
HIRAX	×	*Euclid*	0.2	0.8	2
HIRAX	×	LSST	0.5	0.8	2
CHIME	×	*Euclid*	0.35	0.8	2
CHIME	×	LSST	0.5	0.8	2.5

Table 4.1: Examples of cross-correlation opportunities between 21cm intensity mapping surveys and optical photometric redshift surveys, with (approximate) estimates for their sky and redshift overlap. $f_{\rm sky}$ refers to the fraction of full sky for which these surveys can overlap. $z_{\rm min}$ and $z_{\rm max}$ represent the common redshift overlap range.

techniques, which require no prior knowledge or templates of the foregrounds, can be used to exploit the smooth form of most foreground signals along the line-of-sight to isolate and remove them.

In this chapter I investigate how foreground removal can impact important cosmological measurements. Several studies have investigated how foreground removal can be carried out without detrimental impact on the HI auto-correlation power spectrum recovery [228][195][11]. In this chapter I aim to place particular emphasis on the foreground removal's impact on cross-correlation measurements with optical galaxy surveys. Examples of some future optical-21cm cross-correlation possibilities are outlined in Table 4.1. In order to investigate the impact of foregrounds on cross-correlations, I utilize mock galaxy catalogues built from N-body simulations of dark matter particles. This approach allows for both optical and HI intensity map data to share the same underlying simulated cosmology, with realistic parameters (such as number density of galaxies) corresponding to the specifications of current and forthcoming surveys.

The plan of the chapter is as follows: In Section 4.2 I describe how I simulate the cosmological signals, both the resolved optical galaxy number density maps and the overlapping HI intensity maps. Section 4.3 explains how I simulate the 21cm foregrounds, which are then added into the HI cosmological signal to contaminate the intensity maps. Section 4.4 then explains the processes used for removing these foregrounds and details the FASTICA approach that I opt to use on the simulations. In Section 4.5 I analyze my results and demonstrate what impact foreground cleaning can have on a cross-correlation power spectrum. In Section 4.6 I extend these findings and apply them to a practical experiment which utilizes these cross-correlations to constrain photometric redshifts using HI intensity maps. I conclude and discuss in Section 4.7.

4.2 Cosmological Signals & Their Simulation

In order to probe the large-scale cosmic structure and map the matter over-density δ, we rely on luminous sources which trace the underlying matter density. In optical galaxy redshift surveys we use number density fields $n_{\rm g}(\vec{\theta}, z)$ where resolved galaxies can be counted in voxels

(3-dimensional pixels) at angular position $\vec{\theta}$ with a redshift z which is used for defining the line-of-sight (LoS) distance. We can then calculate the over-density of galaxies $\delta_{\rm g}$, which we assume is a linearly biased tracer of the matter over-density δ. As already introduced in previous Chapters 1 and 3, which I re-define here for completeness, the galaxy over-density can be defined by

$$\delta_{\rm g}(\vec{\theta},z) \equiv \frac{n_{\rm g}(\vec{\theta},z)-\overline{n}_{\rm g}(z)}{\overline{n}_{\rm g}(z)} = b_{\rm g}(z)\delta(\vec{\theta},z)\,, \tag{4.1}$$

where a barred quantity represents a mean average and $b_{\rm g}$ is the (linear) galaxy bias.

For HI intensity maps there are no resolved luminous sources, only combined brightness temperatures in a given voxel. Assuming that HI is also a biased tracer of the underlying matter density we can write

$$\delta_{\rm HI}(\vec{\theta},z) \equiv \frac{T_{\rm HI}(\vec{\theta},z)-\overline{T}_{\rm HI}(z)}{\overline{T}_{\rm HI}(z)} = b_{\rm HI}(z)\delta(\vec{\theta},z)\,, \tag{4.2}$$

where the linear bias factor is now $b_{\rm HI}$. Note that the mean brightness temperature $\overline{T}_{\rm HI}$ is also an unknown quantity, degenerate with $b_{\rm HI}$. Since the observable is a temperature fluctuation, it is customary to work with temperature fluctuation maps where

$$\delta T_{\rm HI}(\vec{\theta},z) \equiv T_{\rm HI}(\vec{\theta},z)-\overline{T}_{\rm HI}(z) = \overline{T}_{\rm HI}(z)\,b_{\rm HI}(z)\delta(\vec{\theta},z). \tag{4.3}$$

It is these quantities, $\delta_{\rm g}$ and $\delta T_{\rm HI}$, which can be used to make cosmological measurements e.g. auto-power spectra $P_{\rm gg} \sim \langle|\tilde{\delta}_{\rm g}|^2\rangle$ or cross-power spectra $P_{\rm g,HI} \sim \langle{\rm Re}\{|\tilde{\delta}_{\rm g}\tilde{\delta}T^*_{\rm HI}|\}\rangle$. Here the tilde notation $\tilde{\delta}$ represents the Fourier transform of the matter over-density.

An important measurement in cosmology, and one I heavily focus on in this chapter, is the angular clustering of a matter density tracer. In order to apply this with HI intensity maps, we measure the angular power spectrum by decomposing the temperature fluctuations into spherical harmonics $Y^m_\ell(\hat{\mathbf{n}})$:

$$\delta T_{\rm HI}(\hat{\mathbf{n}},\nu) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{m=\ell} a_{\ell m}(\nu)\,Y^m_\ell(\hat{\mathbf{n}})\,. \tag{4.4}$$

The harmonic coefficients $a_{\ell m}(\nu)$ describe the amplitudes of the fluctuations in spherical harmonic space; we can then define the angular power spectrum between tracers X and Y as

$$C^{XY}_\ell(\nu_1,\nu_2) = \langle a^X_{\ell m}(\nu_1)\,a^{Y*}_{\ell m}(\nu_2)\rangle\,. \tag{4.5}$$

Consideration must also be given to data that does not cover the full sky and instead comes from only the footprint covered by the survey. The simulations I use will have partial sky coverage and therefore emulate this problem. This has consequences for the power spectrum and results in correlated multipoles which bias the measurement. In this chapter I am not particularly interested in making precise comparisons of a measured power spectrum to say one predicted by a ΛCDM model. Instead I am interested in the comparison of a power spectrum free of 21cm foregrounds to one contaminated by them, which should both be biased by cut skies in the same way. However, to ensure an accurate treatment of the cut skies I will use the pseudo-C_ℓ method of angular power spectrum measurement [216][217] and use the unified pseudo-C_ℓ framework `NaMaster`[4] [15] and its `python` wrapper `pymaster`.

[4] https://github.com/LSSTDESC/NaMaster

If the tracer fields are Gaussian, the power spectrum (4.5) is a complete statistical representation of the fields. The power spectrum can either represent the HI intensity map auto-correlation where $X = Y =$ HI, or the cross-correlation with the optical galaxies where $X =$ g and $Y =$ HI. Hence, in order to use simulations to study the impact 21cm foregrounds can have on cross-correlation cosmological measurements such as $C_\ell^{\mathrm{g,HI}}$, we require a simulation which includes HI emission and resolved optical galaxies.

In many 21cm studies it is sufficient to simulate wide continuous intensity maps through Gaussian realizations of a HI power spectrum. However, for this chapter we need an optical galaxy catalogue which shares the same underlying cosmology as the HI intensity maps, since we are looking to exploit a shared clustering signal between resolved optical galaxies and HI emission for cross-correlated measurements. It is also preferable to have the optical galaxy simulation as a resolved catalogue of sources so that $N(z)$ distributions can be built precisely from individual galaxy redshifts. We can then choose to degrade the redshift accuracy in order to emulate a photometric imaging survey.

In order to achieve this I use a similar method to that used in Chapter 3 which I discuss again here for completion. I use existing N-body galaxy simulations and exploit certain components of them, e.g. HI mass or halo mass to simulate HI brightness temperatures which I can build intensity maps from. Utilizing N-body simulations also allows for a more robust representation of a survey catalogue than Gaussian realized signals. With this in mind we ideally require a catalogue which has the following features:

- low halo-mass resolution ($\approx 10^9 h^{-1} M_\odot$) so that intensity maps include integrated HI emission from faint galaxies;
- HI information for each galaxy for simulating realistic intensity maps;
- deep redshift and wide sky coverage ($0 < z < 3$, $\sim 20{,}000\, deg^2$) to allow for testing low resolutions associated with the typical beam size of a SKA-like intensity mapping experiment;
- simulated photometry for optically resolved galaxies so that realistic cross-correlation forecasts can be made between intensity maps and photometric galaxy surveys.

A simulation including all of the above is not currently available, and is unlikely to be available in the near future. This is largely due to the fact that low halo mass resolution with sufficient galaxy number densities over large sky volumes would require N-body simulations that would be exceptionally computationally expensive.

In this chapter I therefore utilize two simulated catalogues with differing characteristics. One catalogue contains HI signal with a low halo mass resolution and simulated HI masses for every galaxy. The other is a more conventional optical survey catalogue with simulated photometry but which is not as resolved in halo mass. I will now describe the two catalogues in detail.

Catalogue	Box Volume $[(\mathrm{Mpc}/h)^3]$	m_p $[M_\odot/h]$	$\mathrm{N_{gal}}$	f_sky	z_max
GAEA	500^3	8.6×10^8	201×10^6	0.5	0.5
MICE	3072^3	2.9×10^{10}	497×10^6	0.125	1.4

Table 4.2: Summary of the two different mock galaxy catalogues I will be using. Both are built from N-body simulations for which I provide the box size and particle mass m_p.

- **GAEA Simulation**[5]

I make use of the GAEA semi-analytic model [239][233][101]. The catalogue was built using the Millennium Simulation [199], which is a cosmological N-body simulation that used $N = 2160^3$ particles of mass $m_\mathrm{p} = 8.6\times10^8 h^{-1}M_\odot$ within a comoving box of size 500^3 $(\mathrm{Mpc}/h)^3$ with a cosmology consistent with WMAP1 data. In particular, the values of the adopted cosmological parameters are: $\Omega_\mathrm{B} = 0.045$, $\Omega_\mathrm{m} = 0.25$, $\Omega_\Lambda = 0.75$, $H_0 = 100h\,\mathrm{Mpc}^{-1}\mathrm{km\ s}^{-1}$, $h = 0.73$, $\sigma_8 = 0.9$ and $n_s = 1$. The GAEA catalogue is built replicating this same box, but selecting galaxies from the nearest snapshot corresponding to its co-moving distance from the observer.

GAEA used an algorithm to identify halos which allowed for a halo mass resolution of $1.7\times10^{10}M_\odot h^{-1}$ which resulted in just over 2×10^8 galaxies with a continuous sky coverage $f_\mathrm{sky} = 0.5$. Redshifts are limited to $0 < z < 0.5$ which means we will only be able to study cross-correlations within this small range, but this should still allow for multiple redshift/frequency bins given the completeness within this range. GAEA also includes simulated HI masses for all its galaxies, which can be used to generate realistic HI brightness temperatures. I discuss this further in Section 4.2.1.

- **MICE Simulation**[6]

I also make use of the MICECATv2.0 simulation(Chapter 3 used an earlier version for the large-sky maps) released as part of the MICE-Grand Challenge Galaxy and Halo Light-cone Catalogue [78][58][79][44][102], which is a cosmological N-body dark matter only simulation containing 4096^3 dark-matter particles of mass $m_\mathrm{p} = 2.93\times10^{10}h^{-1}M_\odot$ in a box-size of 3072^3 $(\mathrm{Mpc}/h)^3$. They resolved halos down to a few $10^{11}M_\odot h^{-1}$ and used a hybrid Halo Occupation Distribution (HOD) and Halo Abundance Matching (HAM) technique for galaxy modelling resulting in just under 5×10^8 galaxies. The simulation's sky footprint is 90×90 deg^2 filling an octant of sky ($f_\mathrm{sky} = 0.125$) up to a redshift $z = 1.4$. The assumed cosmology is a flat concordance ΛCDM model with $\Omega_\mathrm{m} = 0.25$, $\Omega_\Lambda = 0.75$, $\Omega_\mathrm{B} = 0.044$, $n_\mathrm{s} = 0.95$, $\sigma_8 = 0.8$ and $h = 0.7$ consistent with WMAP 5-year data.

Since the MICE catalogue does not have simulated HI masses for each galaxy, we must derive our own. I therefore take each central galaxy's halo mass as simulated by MICE and convert this into a predicted HI mass by following the redshift dependent prescription laid out in [156]

$$M_\mathrm{HI} = 2N_1 M\Big[\Big(\frac{M}{M_1}\Big)^{-b_1} + \Big(\frac{M}{M_1}\Big)^{y_1}\Big]^{-1}, \tag{4.6}$$

[5] http://astrosims.flatironinstitute.org/gaea
[6] http://maia.ice.cat/mice/

where M is the galaxy's halo mass; M_1, N_1, b_1 and y_1 are all free parameters with redshift dependence tuned to provide a best fit; I refer the reader to [156] for details. Each central galaxy then has a HI mass from which I can generate a HI brightness temperature signal. While this prescription would not be ideal for small scale studies of HI distribution, since I am assuming that all HI lies within central galaxies, it suits our purposes because I will be smoothing out any small scale imprecisions when I simulate the effect of an intensity mapping beam.

From these catalogues, which I summarize in Table 4.2, I will produce both HI intensity maps (Section 4.2.1) and a detected optical galaxy catalogue (Section 4.2.2), which will share the same underlying dark-matter distribution. It is this shared clustering signal which I will look to utilize in the cross-correlation tests. I emphasize once more that I use these two separate N-body simulations since each one has unique advantages. For example the semi-analytical GAEA has replication of the particle box sample which delivers larger sky sizes and also has HI masses for each galaxy at lower mass resolution. Both of these features contribute to delivering more robust simulations of large-beam HI intensity maps. In contrast MICE uses a HOD/HAM approach over a larger box size, so is arguably more realistic in its cosmological signal in that no replication is required, but perhaps less realistic in that it distributes synthetic galaxies into simulated halos using a statistical approach rather than simulating baryonic process to drive galaxy evolution, as performed in semi-analytic models. MICE also includes some simulated photometric redshifts which I utilize for forecasting the impacts of HI foregrounds in cross-correlations with a photometric survey.

4.2.1 HI Intensity Map Simulation

I aim to express the HI intensity map data $T_{\rm HI}$ in the form of a brightness temperature with two angular dimensions ($\theta_{\rm ra}$ and $\theta_{\rm dec}$, jointly represented by $\vec{\theta}$ for notation purposes) and a radial dimension, the redshift z. To do this I follow the same recipe laid out in Chapter 3 and [59] which I repeat here for completeness.

To construct $T_{\rm HI}$ I start with the HI mass $M_{\rm HI}$ of each galaxy, which is given in the GAEA catalogue and generated using halo masses and equation (4.6) for MICE. I then place the galaxies into a data cube with coordinates $(\theta_{\rm ra}, \theta_{\rm dec}, z)$ by binning each galaxy's HI mass into its relevant voxel so I end up with a gridded HI mass map $M_{\rm HI}(\vec{\theta}, z_c)$.

I can then convert this into an intensity field for a frequency width of $\delta\nu$ subtending a solid angle $\delta\Omega$ (which is effectively the pixel size)

$$I_{\rm HI}(\vec{\theta}, z) = \frac{3h_{\rm P}A_{12}}{16\pi m_{\rm H}} \frac{1}{[(1+z)d_{\rm c}(z)]^2} \frac{M_{\rm HI}(\vec{\theta}, z)}{\delta\nu\,\delta\Omega} \nu_{21}\,, \tag{4.7}$$

where $h_{\rm P}$ is the Planck constant, A_{12} the Einstein coefficient which quantifies the rate of spontaneous photon emission by the hydrogen atom, $m_{\rm H}$ is the mass of the hydrogen atom, ν_{21} the rest frequency of the 21cm emission and $d_{\rm c}(z)$ is the comoving distance out to redshift z (I will assume a flat universe).

It is conventional in radio astronomy, in particular intensity mapping, to use brightness temperature which can be defined as the flux density per unit solid angle of a source measured in units of equivalent black body temperature. Hence, the intensity $I_{\rm HI}(\vec{\theta},z)$ can be written in terms of a black-body temperature in the Rayleigh-Jeans approximation $T = Ic^2/(2k_{\rm B}\nu^2)$ where $k_{\rm B}$ is the Boltzmann constant. Using this we can estimate the brightness temperature at redshift z

$$T_{\rm HI}(\vec{\theta},z) = \frac{3h_{\rm P}c^2A_{12}}{32\pi m_{\rm H}k_{\rm B}\nu_{21}}\frac{1}{[(1+z)d_{\rm c}(z)]^2}\frac{M_{\rm HI}(\vec{\theta},z)}{\delta\nu\,\delta\Omega}\,. \tag{4.8}$$

For cosmology studies one aims to make measurements at different redshifts. I therefore choose to slice the intensity maps into thin tomographic redshift bins and collapse these to a 2D slice which can be auto-correlated or cross-correlated with another survey map. I will often discuss binning by frequency (ν) or redshift (z). To clarify, these are interchangeable expressions since $z+1 = \nu_{21}/\nu_{\rm obs}$. For consistency however, bin widths will always be constant in redshift. The width of each tomographic redshift bin needs to be thin enough that we can make certain thin bin assumptions, yet wide enough that we allow for sufficient structure to obtain a strong cross-correlation signal. By thin bin assumptions I am referring to cosmological quantities such as the bias, which I assume to be constant within the width of the bin ($\Delta z = 0.02, 0.05$ for GAEA and MICE respectively).

In order to ensure the HI intensity map amplitudes are in agreement with what is theoretically predicted, I choose to rescale each redshift bin so that it agrees with a model average temperature $\overline{T}_{\rm HI}$. For example [28] gives this average temperature as

$$\overline{T}_{\rm HI}(z) = 180\Omega_{\rm HI}(z)h\frac{(1+z)^2}{H(z)/H_0}\,{\rm mK} \tag{4.9}$$

where $\Omega_{\rm HI}$ is the HI density (abundance). In principle $\Omega_{\rm HI}$ can be measured using the auto-correlation HI power spectrum with redshift space distortions, assuming a fixed fiducial cosmology [135][170]. For this chapter I use a fit for the HI density [23]

$$\Omega_{\rm HI}(z) = 0.00048 + 0.00039z - 0.000065z^2\,. \tag{4.10}$$

In radio HI intensity mapping the observable signals detected by a telescope are brightness temperature fluctuations,

$$\delta T_{\rm HI}(\vec{\theta},z) = T_{\rm HI}(\vec{\theta},z) - \overline{T}_{\rm HI}(z)\,. \tag{4.11}$$

I will therefore convert all the intensity maps to these quantities.

4.2.1.1 Receiver Noise

As we are aiming to simulate realistic observations, we need to include the effects of instrumental (thermal) noise. For the case of a single-dish intensity mapping experiment instrumental noise can be modelled as uncorrelated Gaussian fluctuations. Following [11] and [189] I add onto the observable maps a Gaussian random field with rms

$$\sigma_{\rm noise} = T_{\rm sys}\sqrt{\frac{4\pi f_{\rm sky}}{\Omega_{\rm pix}N_{\rm dish}\,t_{\rm obs}\delta\nu}}\,. \tag{4.12}$$

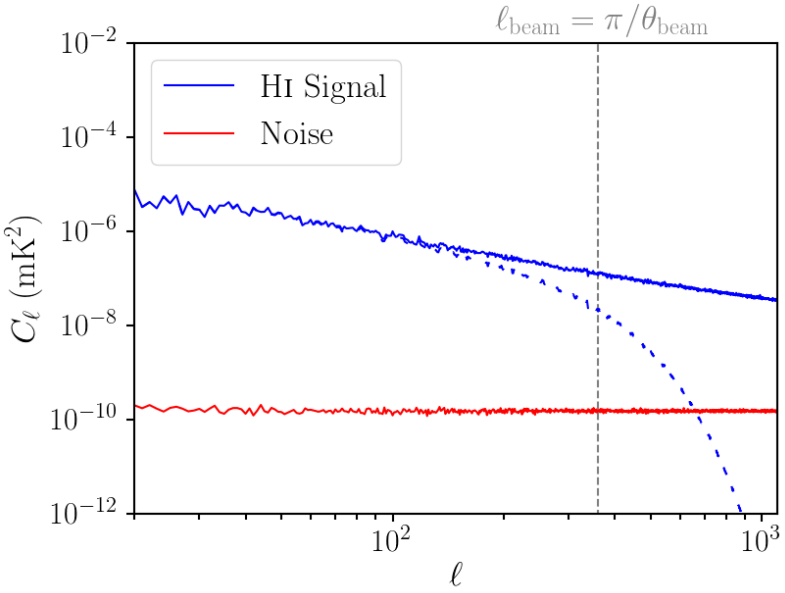

Figure 4.1: Angular power spectrum at a redshift of $z = 0.25$ ($\nu = 1136$ MHz) for both the cosmological signal (blue solid line) for a HI intensity map produced using the GAEA catalogue, and instrumental noise (red solid line). Also included is the effect of a $\theta_{\rm FWHM} = 0.5^{\rm o}$ Gaussian convolution (blue dashed line) which shows a degradation in the cosmological HI signal on smaller scales (high ℓ). The grey vertical dashed line shows the angular scale of this beam. We see that instrumental noise begins to dominate at around $\ell > 700$.

Here $T_{\rm sys}$ is the total system temperature which is the sum of the sky and receiver noise, $T_{\rm sys} = T_{\rm rcvr} + T_{\rm sky}$ with $T_{\rm rcvr} = 0.1\,T_{\rm sky} + T_{\rm inst}$ and $T_{\rm sky}(\nu) \approx 60(300{\rm MHz}/\nu)^{2.55}$ K. I set $T_{\rm inst} = 20$ K which is representative of SKA1-MID for the redshift range $0 < z < 0.58$. $\Omega_{\rm pix} = 1.133\theta^2_{\rm FWHM}$ is the solid angle for the intensity mapping beam. I also assume SKA1-MID-like values for the remaining variables in the noise model with the fraction of sky $f_{\rm sky} = 0.41$ (representative of an SKA-LSST overlap), the number of dishes $N_{\rm dish} = 197$ and the total observation time $t_{\rm obs} = 10,000$ hours. Lastly, $\delta\nu$ is the frequency bandwidth for a particular redshift bin. Figure 4.1 shows the level of this noise in relation to the cosmological HI signal. We can see that the noise only begins to dominate when the signal has the telescope beam effects (discussed in next section) included, and this is only at small scales (high ℓ).

A complete noise simulation would require the inclusion of red noise (or $1/f$ noise) which originates from time correlated gain fluctuations unique to radio receivers [94]. Here I assume that using component separation techniques, this noise can be removed [94]. There is also an argument to include the effects of cross-shot noise caused by HI emitting galaxies, which provide signal in the intensity maps, also being present in the optical galaxy sample [231]. I assume these additional noise effects are sub-dominant at the scales of interest and do not include them in my simulations.

4.2.1.2 Beam Resolution

To model the low angular resolution of an intensity map, I convolve $\delta T_{\rm HI}$ with a telescope beam in Fourier space making use of the convolution theorem. The telescope beam is modelled as a symmetric, two-dimensional Gaussian function with a full width half maximum of $\theta_{\rm FWHM}$ acting only in the directions perpendicular to the LoS (as the frequency/redshift resolution is excellent). The beam size can be determined by the dimensions of the radio receiver and the frequency which is being probed [13]:

$$\theta_{\rm FWHM} = \frac{1.22c}{\nu D_{\rm max}}, \tag{4.13}$$

where $D_{\rm max}$ is the maximum baseline of the radio telescope; for a single dish receiver, $D_{\rm max}$ is given by the dish diameter. The GAEA redshift range of $0 < z < 0.5$ would mean we are looking at beam sizes of $0.99^{\rm o} < \theta_{\rm FWHM} < 1.45^{\rm o}$ for the intensity maps, where I have assumed a maximum baseline of $D_{\rm max} = 15$ m which is representative of the SKA1-MID dishes [23]. The MICE catalogue, which extends to redshifts of $z = 1.4$ will reach even larger beam sizes of $\theta_{\rm FWHM} = 2.36^{\rm o}$. Figure 4.1 shows how the beam effect can present challenges in that it causes instrumental noise to dominate at small scales and potentially destroys information there. I will include the beam scale in terms of multipole $\ell_{\rm beam}$ on some future power spectra plots (as done in Figure 4.1) as this is one of the most dominant effects on the results and on HI intensity mapping power spectra in general.

An example of a completed intensity map tomographically sliced and collapsed into a 2D angular map is shown in Figure 4.2. For all the full-sky maps I use HEALPix maps [87] where the pixelization ensures that each pixel covers the same surface area as every other pixel. I handle the maps in HEALPix RING ordering scheme with resolution `nside` = 512, which corresponds to $12 \times 512^2 = 3,145,728$ pixels across the sky.

4.2.2 Optical Galaxy Catalogue Simulation

For probing large-scale cosmic structure with resolved optical galaxies I use number density fields. While we ideally require a simulated catalogue with high number density and completeness for the HI intensity maps, it would be unrealistic to expect every one of the low mass galaxies to be resolved and detected by a conventional wide area optical survey. Therefore to make this a realistic test we need to introduce some detection threshold which results in only the brightest galaxies being included in the optical sample. We also desire to have realistic $N(z)$ redshift distributions which tail off at higher redshifts where resolved detection becomes more difficult. The way this is all achieved is by invoking a model redshift distribution, given by

$$\frac{dN_{\rm g}}{dz} = z^{\beta} \exp(-(z\alpha/z_{\rm m})^{\gamma}) \tag{4.14}$$

where I use $\alpha = \sqrt{2}$, $\beta = 2$ and $\gamma = 1.5$ [96] which are values typical of stage-IV optical large-scale structure survey such as LSST or *Euclid*. $z_{\rm m}$ is the mid-redshift for the particular simulated catalogue I am applying this to e.g. for MICE this would be $z_{\rm m} = 0.7$. I make the optical samples

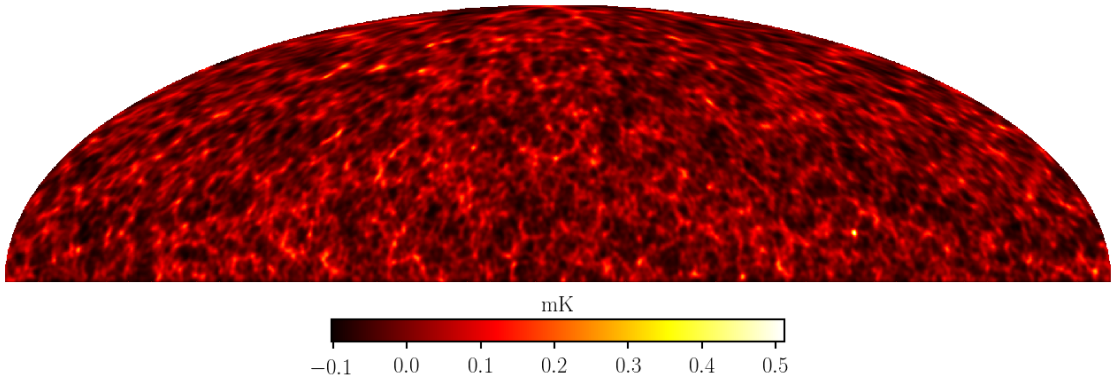

Figure 4.2: $\delta T_{\rm HI}$ intensity map at redshift $z = 0.25$ ($\nu = 1136$MHz) binned using constant redshift intervals of $\Delta z = 0.02$. This includes the effects of SKA-like noise and beam, outlined in Sections 4.2.1.1 and 4.2.1.2 respectively. At this frequency the beam size is approximately $\theta_{\rm FWHM} = 1.23^{\rm o}$. This example is done with the GAEA catalogue covering half of the sky ($f_{\rm sky} = 0.5$). This example does not include any foreground contamination.

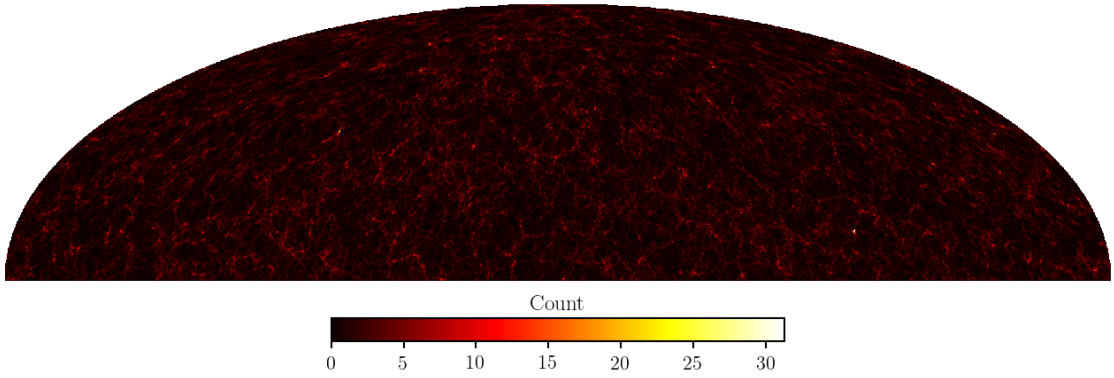

Figure 4.3: $n_{\rm g}$ optical galaxy number density field with galaxies binned by true redshift at $z = 0.25$ with $\Delta z = 0.02$. Unlike the intensity map in Figure 4.2, this map has no beam smoothing since it represents observations by an optical telescope. However, for this demonstration map only, I have downgraded the HEALPix resolution to `nside` = 128. This is to make the shared structure between this and the intensity map at the same redshift more apparent.

conform to this distribution by ordering galaxies by stellar mass in each redshift bin. Here I am using stellar mass as a crude approximation of optical brightness which for our purposes will be sufficient. I then pick the 'brightest' galaxies in each redshift bin until the model redshift distribution is achieved. This process gives final galaxy catalogues with 2.67×10^7 galaxies for GAEA, which is an average density of around 54 galaxies per square degree for each of the 24 redshift bins I use. For MICE I achieve a much denser catalogue with 3.97×10^8 galaxies over a smaller sky area giving 3.2×10^3 galaxies per square degree.

This optical sample makes no consideration of any classifications of galaxies. All are treated as point-like and either 'observed' or not. More investigation could be taken into certain classifications e.g. by colour; red and blue galaxies are expected to cluster differently and have different densities at different redshifts. This could plausibly have an effect on these studies and bias

the correlation function; this has been touched upon in a recent cross-correlation study using Parkes HI intensity maps and 2dF galaxies [18].

Figure 4.3 shows an example of a final over-density field for the optical data. This has been made using the GAEA catalogue at $z = 0.25$; similarities between this and Figure 4.2 should be apparent since these are both for the same dataset at the same redshift. I have shown this map with `nside` = 128 to make the clustering pattern more obvious.

4.3 21cm Foregrounds & Their Simulation

I test the effects on HI intensity maps of four main foregrounds:

(i) Galactic synchrotron
(ii) Extragalactic point sources
(iii) Galactic free-free emission
(iv) Extragalactic free-free emission

Each of these processes emit signals in the frequency region of the redshifted HI signal i.e. $\sim 1420/(1+z)$ MHz. Each of them are dominant over the HI signal which is inherently weak. In some cases, such as galactic synchrotron, the foregrounds can be several orders of magnitude higher in observed brightness temperature. It is therefore immediately apparent that this a major challenge for the success of the HI intensity mapping technique.

Extragalactic point source foregrounds (ii) are caused by objects beyond our own Galaxy emitting signals with wavelengths similar to the redshifted 21cm signal, a typical example being AGNs. (iii) & (iv) represent free-free emission which is caused by free electrons scattering off ions without being captured and remaining free after the interaction. In this weak-scattering interaction low-energy photons are produced which can enter the $21(1+z)$ cm wavelength window we are interested in. These free-free interaction signals can be detected both within (galactic free-free) and beyond (extragalactic free-free) our own Galaxy.

Lastly the synchrotron emission (i) occurs when high-energy electrons are subject to an acceleration perpendicular to their velocity by the application of a magnetic field. This foreground is typically caused by relativistic cosmic ray electrons accelerated by the galactic magnetic field. It is the galactic synchrotron which is by far the most dominant of the foreground types and is therefore the one we would like to concentrate most on removing.

4.3.1 Galactic Synchrotron

While it would be fairly straightforward to simulate Gaussian realizations of galactic synchrotron from a model power spectrum, it is far more robust to make use of existing data and use this to emulate the shape of the emission on the sky. This also allows us to study the impact of subtracting a foreground which has wide structures, potentially eliminating information at large angular scales.

Foreground	A	β	α	ξ
Galactic synchrotron	700	2.4	2.80	4.0
Point sources	57	1.1	2.07	1.0
Galactic free-free	0.088	3.0	2.15	35
Extra-galactic free-free	0.014	1.0	2.10	35

Table 4.3: Parameter values for foreground C_ℓ's (see equation (4.17)) with amplitude A given in mK^2. Pivot values used are $\ell_{\rm ref} = 1000$ and $\nu_{\rm ref} = 130\,\mathrm{MHz}$ as per [188].

Unfortunately, foregrounds within the frequency range of the redshifted 21cm signal ($400\,\mathrm{MHz} < \nu < 1420$ MHz) are less well studied than other foregrounds, for example those which impact the microwave background emission at higher frequencies ($\nu > 10\mathrm{GHz}$). Therefore, obtaining actual data maps of galactic emission at regular frequency intervals in the range we are interested is challenging.

Following a method which has been used in similar HI foreground studies [195][228][14] I use the Global Sky Model (GSM) [238] to generate maps $T_{1420}(\vec{\theta})$ and $T_{400}(\vec{\theta})$ which are emission maps at 1420 MHz and 400 MHz, then use these to construct a full-sky spectral index given by

$$\alpha(\vec{\theta}) = \frac{\ln T_{1420}(\vec{\theta}) - \ln T_{400}(\vec{\theta})}{\ln 1420 - \ln 400}. \tag{4.15}$$

This is then used to extrapolate the Haslam map [97], which is one of few all-sky maps for galaxy emission around these frequencies,

$$T_0(\vec{\theta}, \nu) = T_{\rm Haslam}(\vec{\theta}) \Big(\frac{\nu}{408\,\mathrm{MHz}}\Big)^{\alpha(\vec{\theta})}. \tag{4.16}$$

This can now be used to simulate a map of the sky at any desired frequency. However, since the Haslam map does not provide information beyond its own resolution ($\sim 1^{\rm o}$), we need a further process to improve the resolution of these maps for any meaningful investigation of small scales.

I add in this additional small scale information through Gaussian realizations of an angular power spectrum which models galactic synchrotron emission. Following [188] I make this construction using the angular power spectrum

$$C_\ell(\nu_1, \nu_2) = A\Big(\frac{\ell_{\rm ref}}{\ell}\Big)^\beta \Big(\frac{\nu_{\rm ref}^2}{\nu_1 \nu_2}\Big)^\alpha \exp\Big(-\frac{\log^2(\nu_1/\nu_2)}{2\xi^2}\Big), \tag{4.17}$$

where ξ is a parameter which regulates the spectral smoothness of the foreground such that smaller ξ cases are less smooth in frequency and are therefore more of a challenge to disentangle from the cosmological signal. The rest of the parameters are defined in Table 4.3. Figure 4.4(i) shows a full-sky map of the simulated galactic synchrotron emission for a frequency slice.

Galactic synchrotron has the added complication of being partially linearly polarized. This polarized portion can undergo Faraday rotation which changes the polarization angle of the radiation. The consequences for the HI signal have been studied in [111][110][140]. Generally speaking this polarization response tends to erode the spectral smoothness of the signal, since it is a frequency dependent effect, and the induced spectral structure is problematic for separating the foreground from the cosmological HI signal. This requires excellent instrumental

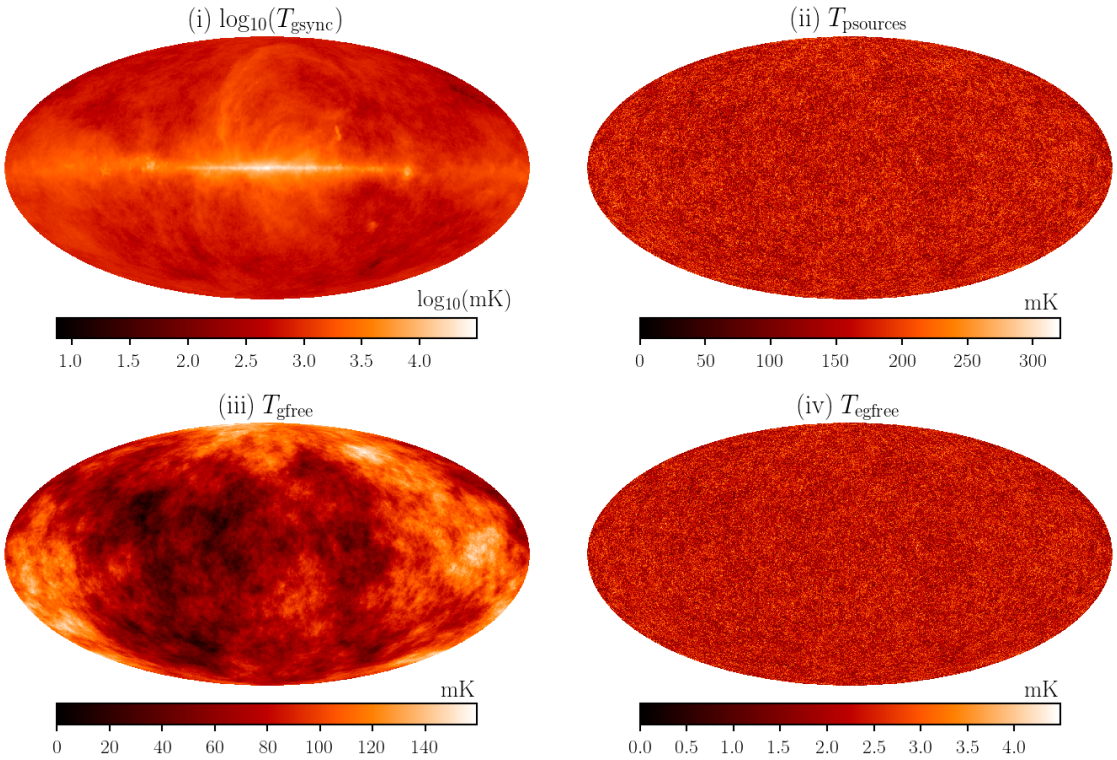

Figure 4.4: Full sky maps of each simulated foreground at a frequency of 1136 MHz ($z = 0.25$). These examples do not include noise or beam smoothing. All temperatures are in mK but the galactic synchrotron map (i) shows the logarithm of the temperatures.

calibration to avoid leakages of the polarization effects. The simulation of such polarization leakage is complex and instrument specific. For this chapter I do not simulate any polarization of the synchrotron emission, but I do opt to convolve all the maps at differing frequencies to a common resolution based on the maximum size of the instrument beam. This is thought to be an active step in mitigating the effects of polarization leakage. This is likely due to decreasing the consistency with which the magnetic field can be directed, thus making the Faraday rotation more stochastic and unable to create the structured frequency dependence which causes the contamination problem. Mitigating the effects of polarization leakage by further smoothing the maps is something that is carried out in the Green Bank Telescope HI intensity mapping data analysis [205].

4.3.2 Point Sources & Free-Free Emission

While galactic synchrotron dominates over all other HI foregrounds, it is still important to consider these additional contaminants since they still dominate over the HI signal. Extragalactic point sources and extragalactic free-free emission are isotropic in nature, since they are sources beyond our own Galaxy. Therefore it is realistic to simulate them with full-sky Gaussian realizations of the angular power spectrum I laid out in equation (4.17) using parameters from Table 4.3. This makes the assumption that the source of these foregrounds is Gaussian and also that

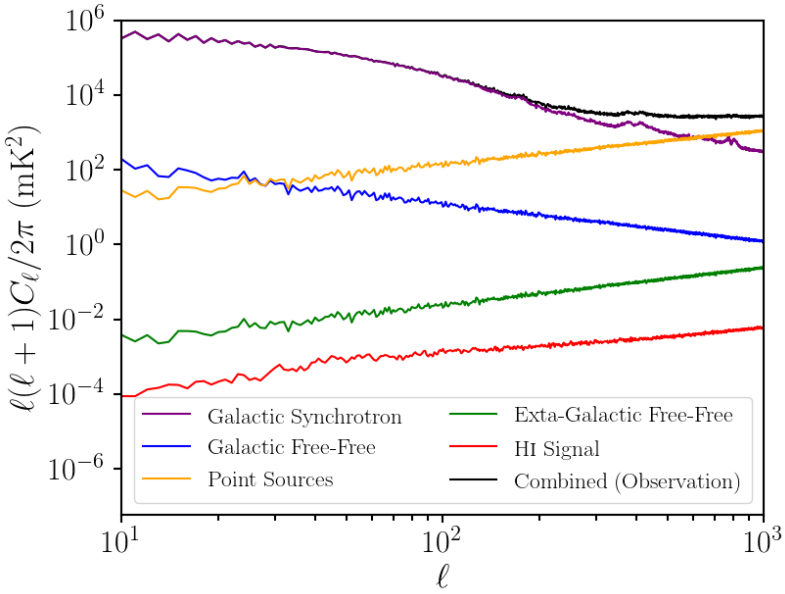

Figure 4.5: Angular power spectra for all the different simulated foregrounds, and the HI cosmological signal produced using the GAEA catalogue. The black solid line represents the combined signal from all foregrounds and the HI cosmological signal. All are at a frequency of 1136MHz ($z = 0.25$) and noise free with no beam effects added.

there is no angular correlation between point sources and HI emitting galaxies. While point sources will cluster with the underlying matter density, the continuum signals they emit mean that in any one redshift bin, angular correlation between point source signal and HI is likely to be small.

Galactic free-free emission is not expected to be perfectly isotropic and will have some galactic latitude dependence. However, because it has a low amplitude and very smooth frequency dependence, this will not be a difficult foreground to subtract and I therefore do not believe a more robust modelling is needed here.

For these three foregrounds, point sources, galactic free-free and extra-galactic free-free, I therefore use equation (4.17) and the parameters from Table 4.3 for the simulations. Figures 4.4(ii),(iii) and (iv) shows maps of these three different foregrounds using the isotropic Gaussian realization approach I have outlined. The lack of galactic latitude dependence is immediately apparent in contrast to the galactic synchrotron map in Figure 4.4(i).

To complete this discussion on HI foregrounds I include the angular power spectra measured for each of the produced foregrounds in Figure 4.5 along with the cosmological signal. This immediately highlights the challenge faced when attempting foreground subtraction as it demonstrates how dominant all the foregrounds are over the cosmological signal I am trying to extract.

4.3.3 Simulated Observable Signal

To summarize, the simulated sky signal is a composition of maps at certain frequencies (equivalently, redshifts) which can be described by

$$\delta T_{\rm sky}(\nu) = \delta T_{\rm HI}(\nu) + \sum_i \delta T_i^{\rm FG}(\nu) \tag{4.18}$$

where the first term comes from the signal described in Section 4.2.1 and the second term is the contribution from all the different foregrounds outlined previously. Once these maps are combined I smooth the total temperature map $\delta T_{\rm sky}$ using the Gaussian beam given by equation (4.13). I then add the simulated random noise from equation (4.12) to emulate basic instrumental systematics, resulting in the final simulated observation

$$\delta T_{\rm obs}(\nu) = \mathbf{S}_{\rm beam}\left(\delta T_{\rm HI}(\nu) + \sum_i \delta T_i^{\rm FG}(\nu)\right) + \delta T_{\rm noise}(\nu) \tag{4.19}$$

where $\mathbf{S}_{\rm beam}$ is the smoothing (or convolution) function.

4.4 Foreground Removal

While foregrounds pose a huge problem for the prospects of exploring cosmology with HI intensity mapping data, there are some features that help distinguish them from the cosmological 21cm signal. We can utilize the spectral smoothness of the foregrounds to separate them from the HI , which fluctuates with frequency. Figure 4.6 shows that along a LoS, the foregrounds are very smooth, whereas the expected signal from HI has a strong frequency dependence. It is this property that is utilized in a class of methods referred to as blind foreground subtraction. Less general 'non-blind' approaches would involve precise modelling of the foreground contamination. Given the lack of data for these foreground signals at the relevant frequencies, this approach is not currently viable.

It is apparent however, that a foreground clean based on this distinguishing spectral smoothness would be more successful for small wavelength radial modes, whereas for larger wavelength radial modes the HI signal is more similar to the foregrounds. Hence these types of foreground cleans can render large Fourier radial modes (or small $k_\parallel$) useless. Removing large-scale modes from HI intensity maps is therefore an expected effect of a foreground clean and was used as a toy model to emulate the effects of foreground cleaning in Chapter 3 and [59]. In this chapter I extend the foreground investigation by directly contaminating the maps with the foregrounds I outlined in Section 4.3, and then use state-of-the-art foreground removal techniques to recover the HI input data and study the impact this will have on fundamental cosmological measurements.

There are several blind foreground removal techniques, for example principle component analysis (PCA) and independent component analysis (ICA) whose distinctions are outlined in [11]. Further blind component separation methods include Generalalized Morphological Component Analysis (GMCA) [49] and Generalized Internal Linear Combination (GnILC) [177].

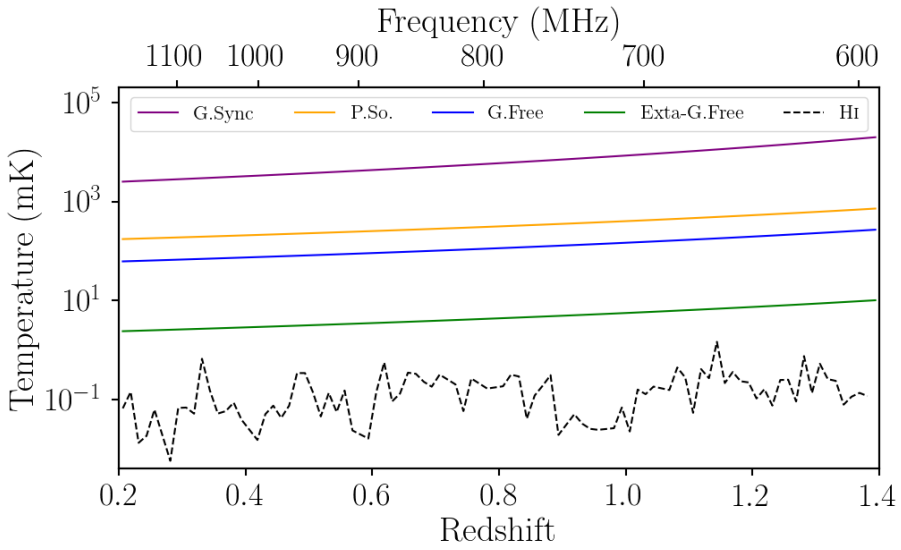

Figure 4.6: Observed brightness temperatures along a chosen LoS through frequency (or redshift). This is presented for the MICE catalogue with 100 redshift bins to show a large frequency range. The plot demonstrates the foreground smoothness in frequency (coloured solid lines), in contrast to the highly oscillatory fluctuations of the HI signal (black dashed line).

For this chapter I examine the FASTICA method [108][48][228][230], which I describe in the following section. [11] found there to be very little distinction between a PCA and ICA approach to foreground cleaning, so this choice of FASTICA as a foreground removal process should not affect the generality of my conclusions.

4.4.1 FASTICA Formalism

Here I introduce the basic principles of the Fast Independent Component Analysis (FASTICA) technique, which I will utilize for foreground removal. For a more complete derivation and discussion I refer the reader to [108]. In a blind foreground removal problem we assume that a raw observed signal, such as that outlined in equation (4.19), can be generalized into a linear equation where the elements making up the signal are statistically independent. That is

$$\mathbf{x} = \mathbf{A}\mathbf{s}. \tag{4.20}$$

The dimensions and basic description of each term in this equation are given as:

x $[N_z,\ 1]$: combined observed signal
A $[N_z,\ m]$: mixing matrix - determines the amplitudes of **s**
s $[m,\ 1]$: independent components (containing foregrounds)

Practically this system will have some trace residuals which have some frequency dependence which will include instrumental noise, any residual foregrounds which cannot be classified into an independent component (IC), and the cosmological HI signal. FASTICA aims to solve

equation (4.20) and identify each IC so that from the remaining residual, the HI signal can be reconstructed. For each LoS, sorted into N_z redshift bins and assuming m ICs are present, FASTICA assumes

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \varepsilon = \sum_{i=1}^{N_{\rm IC}=m} \mathbf{a}_i s_i + \varepsilon\,, \tag{4.21}$$

with $\varepsilon[N_z,\ 1]$ the residual (containing HI signal and noise).

Under the assumption that each independent component s_i is statistically independent, FASTICA attempts to solve equation (4.21) by utilizing the central limit theorem. This states that the greater the number of independent variables in a distribution, the more Gaussian that distribution will be i.e. the probability density function (PDF) of several independent variables is always more Gaussian than that of a single variable. Hence, if we can maximize any statistical quantity that measures non-Gaussianity, then we can identify statistical independence and form a prediction for $\mathbf{a}_i$ and s_i.

The parameter m must be pre-specified before calculations. This is the number of ICs that can be described by unique non-Gaussian descriptions and is not necessarily the number of different foregrounds one is aiming to find. It is typically assumed that $m \approx 4$ [48][228][11] and FASTICA then works by obtaining 4 data vectors which are as statistically independent as possible. With FASTICA going to a higher number of ICs than is required converges to the same result. However, the computational cost is increased so for efficiency, the lowest value for m which gives the best possible result is sought.

The FASTICA process considers all LoS simultaneously. Therefore for its calculations on maps with a number of pixels given by $N_{\rm pix}$, the ICs $\mathbf{s}$ in equation (4.21) are actually maps, and hence an array with size $[m,\ N_{\rm pix}]$, while $\mathbf{x}$ and ε are arrays of size $[N_z,\ N_{\rm pix}]$. Furthermore, as I will further explain below, FASTICA involves some expectation value calculations which rely on a number of samples and for this it uses the $N_{\rm pix}$ different LoS.

To obtain $\mathbf{s}$ we start by inverting equation (4.21), ignoring the residual term ε which will just be left over from signal not contained within the ICs. We can therefore write

$$\mathbf{s} = \mathbf{W}\mathbf{x}, \tag{4.22}$$

here $\mathbf{W}$ is the weighting matrix, defined as the inverse of $\mathbf{A}$ in equation (4.20). Under the assumption that the elements $\mathbf{s}$ are as statistically independent as possible, FASTICA then begins maximizing the non-Gaussianity. For a measure of Gaussianity it uses negentropy $J(y)$, which for a variable y, is based on typical entropy $H(y)$ defined as

$$H(y) = -\sum_i P(y = a_i) \log P(y = a_i)\,, \tag{4.23}$$

where $P(y = a_i)$ is the probability that y equals a possible value a_i. The modification made to obtain the negentropy $J(y)$ is

$$J(y) = H(y_{\rm G}) - H(y)\,, \tag{4.24}$$

where $y_{\rm G}$ is a unit-variance Gaussian random variable. In practice, negentropy is computationally hard to calculate and requires numerous realizations to obtain information on probability

distributions. However, using the maximum entropy principle, we can write

$$J(y) \approx -\sum_{i}^{n} k_i [\langle G_i(y)\rangle_\theta - \langle G_i(y_{\rm G})\rangle_\theta] \,, \tag{4.25}$$

where k_i are positive constants, G_i is referred to as the contrast function, and all pixels are utilized by averaging over them (this is denoted by $\langle\rangle_\theta$). For the contrast function, whilst practically any non-quadratic function will work, FASTICA mainly uses

$$G_1(y) = \frac{1}{a_1}\log\cosh(a_1 y),\; G_2(y) = -\frac{1}{a_2}\exp(-a_2 y^2/2)\,, \tag{4.26}$$

where $1 \leq a_1 \leq 2$ and $a_2 \approx 1$.

I reiterate that there is very little distinction between FASTICA and a PCA approach. In fact, FASTICA begins with whitening the data which involves performing a full PCA analysis. It is then that FASTICA imposes statistical independence to isolate the foreground contamination, whereas PCA presumes that the sources should be uncorrelated. This subtle distinction is actually mathematically equivalent in the case where all sources are Gaussian [11]. Hence any conclusions gleaned from using a FASTICA approach are likely valid for PCA and vice-versa.

In a nutshell, FASTICA delivers a method of reconstructing the foreground signals as m ICs and then the residual ε between this reconstruction and the raw observed input map is the recovered cosmological HI signal plus any receiver noise and residual foreground contaminants. A final point to include is that the mean temperature of the HI cosmological signal is a smooth function of frequency and is therefore incorporated into the ICs of the analysis. This information is therefore lost and the residual maps are required to be renormalised to some model mean temperature or treated as δT observables as in equation (4.11).

4.4.2 FASTICA Results

Here I seek to validate the FASTICA reconstruction process introduced in the previous Section 4.4.1 by presenting results from the simulations outlined in Section 4.3. Since neither of the cosmological simulations cover the full sky, I only add and remove foregrounds to the footprint covered by GAEA and MICE. Restricting the foreground analysis to these patches represents a more realistic emulation of a cosmological survey. However, I found no noticeable difference when I conduct the foreground removal over the full sky compared with conducting it over the cosmological simulation footprint.

Figure 4.7 shows the IC maps found after FASTICA has been applied. This is the only occasion where the foreground analysis is done for the full sky and I have chosen to do this purely for demonstrative purposes of the FASTICA process. It is interesting to note that the third and fourth ICs clearly seem to pick up the galactic synchrotron angular shape whereas the second IC shows structure across the sky. The first IC is largely contained in the top half of the map, where the HI cosmological signal lies for the GAEA catalogue. This suggests that it is this component which is collecting large radial modes which belong to the cosmological signal along with the $\overline{T}_{\rm HI}$ average which smoothly fluctuates and therefore is removed. Despite trying a number of different values

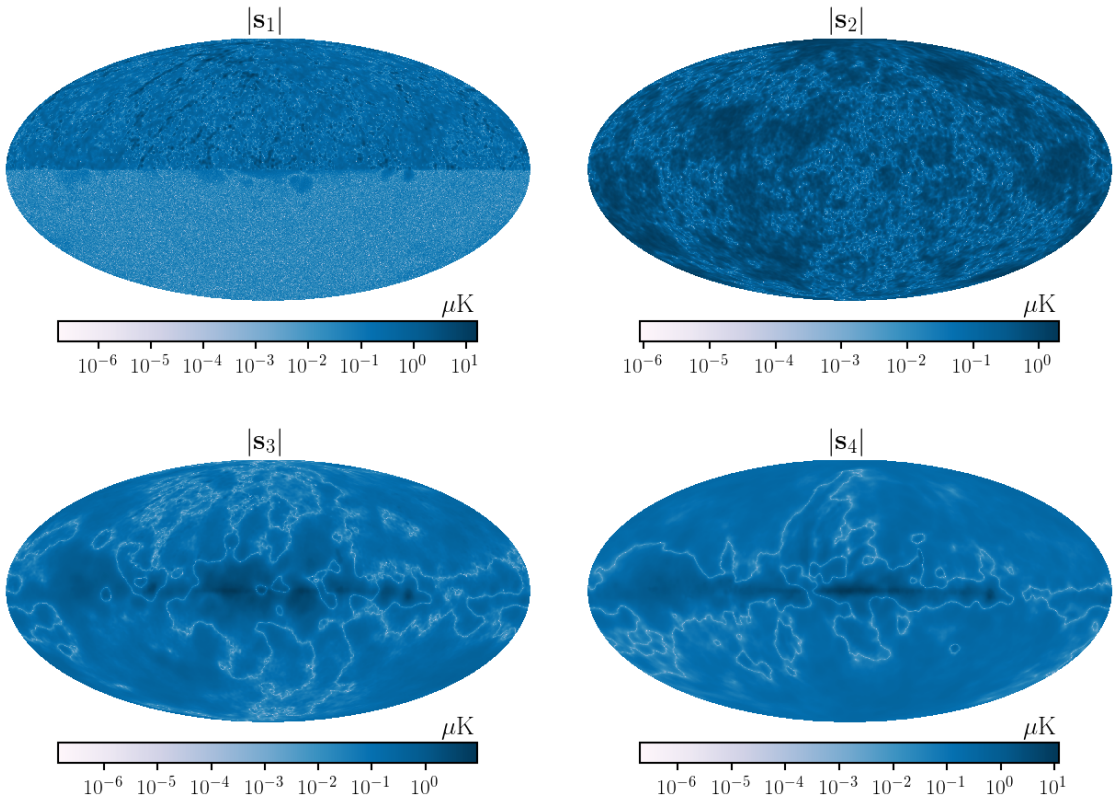

Figure 4.7: Independent component maps found using FASTICA with $m = 4$ on the GAEA simulation contaminated with foregrounds. This is for a constant beam of $\theta_{\text{FWHM}} = 0.5^{\text{o}}$ at all frequencies. Temperature fluctuations are given in μK but the true amplitudes for the estimated foregrounds are determined by their combination with the mixing matrix.

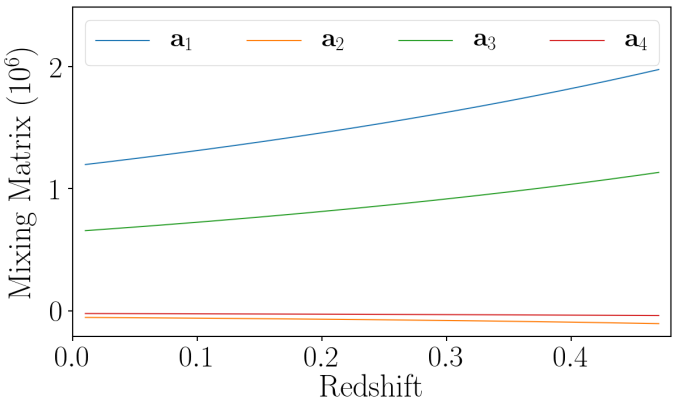

Figure 4.8: Mixing matrix elements as outlined by equation (4.21). Combination of these with the independent components in Figure 4.7 determines the subtraction to be made from the combined observed signal at each frequency.

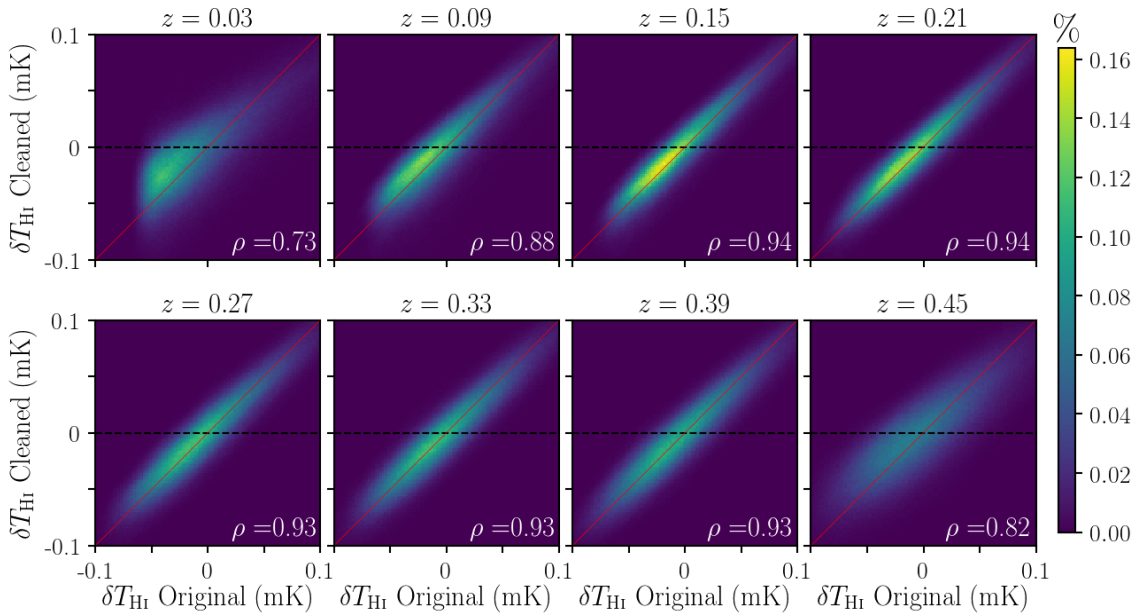

Figure 4.9: Histogram showing the original HI temperature against the FASTICA reconstructed value for each pixel in a range of redshift bins for the GAEA model. Each histogram has been normalized such that the histogram values sum to 100%. I also include the Pearson correlation coefficient ρ for each redshift to quantify the agreement. For a perfectly working foreground clean we would expect an entirely one-to-one ($\rho = 1$) agreement along the thin diagonal red line. We can see how FASTICA is less effective at extreme ends of redshift range with a wider dispersion of values.

of m (the number of ICs) it appears that it is always the case that some cosmological signal will be removed. These ICs from Figure 4.7 are then combined with the mixing matrix (displayed in Figure 4.8) as described in equation (4.21).

Figure 4.9 shows a pixel-by-pixel comparison between original values in the $\delta T_{\rm HI}$ intensity maps and the cleaned values for some selected redshift bins in the GAEA simulation. For a perfectly performing reconstruction we would obtain all values along the red diagonal line, i.e. all values would match their originals. We can see that this is not the case but largely FASTICA is performing reasonably well with a Pearson correlation coefficient of $\rho \geq 0.93$ for most redshifts. We expect a value of $\rho = 1$ for a perfectly performing foreground clean indicating perfect correlation between original and cleaned maps. Figure 4.9 also shows that this method of foreground cleaning performs better at the mid-ranges of redshift. This is not a redshift specific effect since we also see similar results in the MICE model where the best agreement is at redshift $z \sim 0.8$ which is the mid-redshift for its range. This suggests that there are some edge effects in the foreground removal process causing it be less effective at the extreme radial ends of the input data, a result previously noted e.g. [228].

Figure 4.10 indicates how well the HI auto power spectrum can be recovered with FASTICA and shows how varying the number of ICs affects the recovery. I show results from both simulations, and it is interesting to note the difference between the two. We see that with GAEA only 3 ICs are needed for a successful reconstruction, however for MICE even 4 ICs is not sufficient for

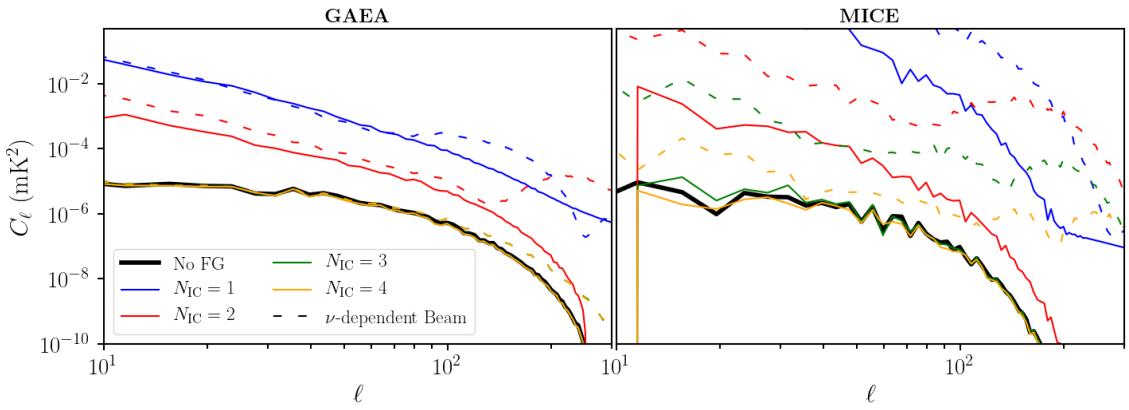

Figure 4.10: Impact of foregrounds on the HI auto-power spectrum for both the GAEA and MICE catalogues. Thick solid black line shows the original HI signal with no foregrounds. The coloured lines then show different values of m used i.e. the number of independent components assumed in the FASTICA process. Also included are the results from using a beam which varies with frequency (dashed lines) and how this damages performance. These results are for mid-range redshifts for each catalogue with $z = 0.25$ for GAEA and $z = 0.825$ for MICE.

good agreement at large scales (small-ℓ). I tested a larger number of ICs with little improvement. The difference in results is probably due to the fact that MICE has a smaller sky coverage (25% of GAEA) which means less samples to average over for negentropy estimation in equation (4.25). Furthermore, since MICE has a deeper redshift range (extending to $z = 1.4$ whereas GAEA is only up to $z = 0.5$) the constant beam size that I convolve with is much larger for MICE, $\theta_{\rm FWHM} = 2.36^{\rm o}$ against GAEA's $\theta_{\rm FWHM} = 1.46^{\rm o}$. This difference in beam size is also evident from the scales at which the power spectrum seems to degrade. Due to its larger beam, the MICE power spectrum begins to tail off at lower-ℓ than GAEA. Lastly this plot also includes results where each tomographic slice has been smoothed by a varying amount due to the frequency dependence of the beam. This is shown as the dashed lines, and it is evident that results are much worse when compared with the constant beam case. This is discussed further in the following section.

4.4.2.1 Frequency Dependent Beam Size

As previously outlined in Section 4.2.1.2, the intensity maps at different frequencies will have different beam sizes defined by equation (4.13), meaning intensity maps at lower redshift have less degradation of angular scales. However, since FASTICA finds m IC maps and then subtracts these from the total observation based on the mixing matrix A, trying to obtain e.g. 4 IC maps based on N_z intensity maps with different resolutions for each will cause problems because the IC map resolution will not match each of the intensity maps. This is exactly why we see poorer performance in Figure 4.10 in the case where there is a frequency dependent beam size (dashed lines) especially at smaller scales (large-ℓ) where the beam has a more dominant effect.

The way we resolve this issue is by carrying out a further convolution on the intensity maps such that each tomographic slice is smoothed to the same resolution. I therefore take the

maximum beam for the particular redshift range and smooth over all maps with this constant beam size. FASTICA then finds IC maps which, when subtracted from the observed signal, prove more effective for reconstructing the original HI signal as shown by the solid lines in Figure 4.10.

Artificially re-smoothing over all the intensity maps may appear to be a wasteful process in terms of loss of large-ℓ modes, but it is necessary for a successful FASTICA reconstruction. In fact, choosing a common resolution significantly larger than the max beam has additional benefits when dealing with real data, as an effective way of mitigating the effects of polarization leakage [205].

4.4.2.2 Increasing the Number of Frequency Bins

For both the GAEA and MICE simulations I am only using 24 redshift (frequency) bins with the bin width determined by a constant separation in redshift Δz. This may be seen as quite a low number of bins to be using in an intensity mapping simulation which uses an ICA process. This is largely out of necessity due to the choice of simulation approach: since I am using N-body simulations there are a finite number of galaxies to use from which to build intensity maps. By using bins which are too thin we risk under-sampling the intensity maps and making them an unrealistic emulation of a continuous field of emission.

The MICE catalogue contains $\sim 500 \times 10^8$ galaxies and I bin them into 393,216 angular pixels giving ~ 1272 galaxies per pixel. The GAEA catalogue has fewer galaxies ($\sim 200 \times 10^8$) and more pixels to bin into due to the larger sky and results in ~ 127 galaxies per pixel. These galaxies then need to be further binned into radial redshift bins and it is obvious that if I opted to use a large number of bins ($\gtrsim 100$) then certainly for the GAEA simulation we would be nearing the situation where there is on average 1 galaxy per voxel. This would be an inaccurate emulation of an intensity mapping experiment.

In practice when using real data, the typical approach would be to perform the FASTICA method on more maps (> 100 frequency channels), then re-stack these into fewer bins for cosmological analysis and cross-correlations with optical data. I trialed this with the MICE catalogue using 240 bins, and found that it made no improvement on the FASTICA foreground removal, hence justifying the choice of using 24 frequency bins in all my analysis.

4.5 HI × Optical Cosmology with Foregrounds

In this section I investigate the impact that HI foreground contamination and removal with FASTICA has on the cross-correlation power spectra $C_\ell^{\mathrm{g,HI}}$ with the simulated optical catalogues.

In recent work [38], a framework which models observational effects on 3D power spectra for HI-optical cross-correlations has been developed. This framework can be extended to include the effects of foreground removal and photometric redshift uncertainty. By doing this one could analytically model the foreground removal effects, as well as the photometric redshift effects, as a loss of small and large $k_\parallel$ modes respectively, and attempt quantitative corrections accordingly.

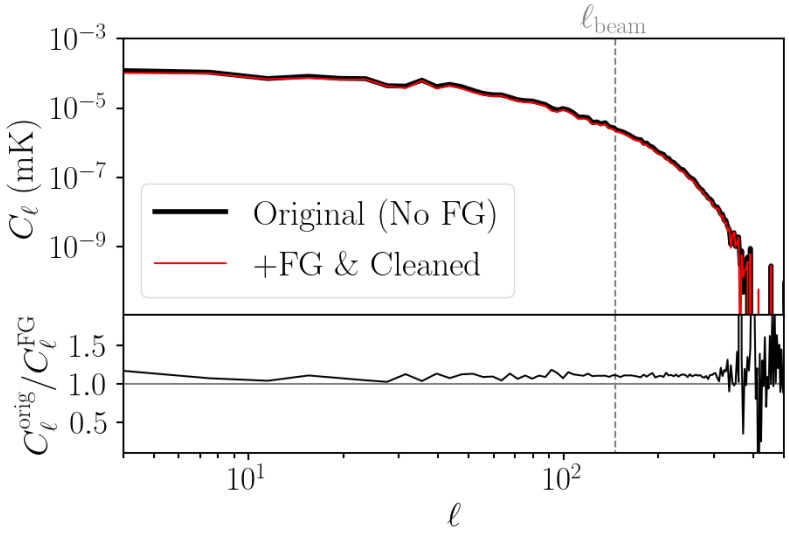

Figure 4.11: Cross-correlation angular power spectrum between the HI intensity map at redshift $z = 0.25$, with $\Delta z = 0.02$ bin width and the optical galaxies binned using their true redshifts. This is representative of a scenario in which spectroscopic redshifts are used in the optical survey. The original result with no foregrounds is shown as the black thick line and the case where foregrounds have been included then removed by FASTICA is shown as the red thin line. The bottom panel shows the ratio of the two spectra. This test was carried out on the GAEA simulation where the HI intensity maps have been re-smoothed with a constant maximum beam of $\theta_{\rm FWHM} = 1.46^{\rm o}$.

In this chapter I aim to use my simulations to investigate what corrections can be made to the data to extract the most information from these cosmological measurements.

To begin exploring how HI foregrounds can impact cross-correlations with optical surveys I first perform a best-case scenario test and cross-correlate with an optical survey which I assume has very well constrained redshifts; Figure 4.11 shows the result of this cross-correlation. Here I bin the optical galaxies from the GAEA simulation by their true redshift with constant bin width of $\Delta z = 0.02$. This is exactly matched to the frequency bins used for the 21cm intensity maps using $\nu = \nu_{21}/(1+z)$), so we have a sample of optical galaxies at $z = 0.25$ to cross-correlate with an HI intensity map at the same redshift. This shows that foregrounds should have little impact on optical spectroscopic cross-correlations. The bottom panel shows a small bias which in principle could be corrected for by constructing a foreground cleaning transfer function [204], but it is encouraging that these initial efforts have already reconstructed the cross-power to within 8.5% at scales below those unaffected by the beam ($\ell < \ell_{\rm beam}$). It is only at higher ℓ, way below the resolution of the beam ($\ell_{\rm beam}$), that we start to have large errors on C_ℓ. This is unsurprising since this is going beyond the scales of the radio instrument's resolution. I experimented with smoothing the optical field to replicate the HI intensity map resolution but find no mitigation of the noise we see at $\ell > 250$.

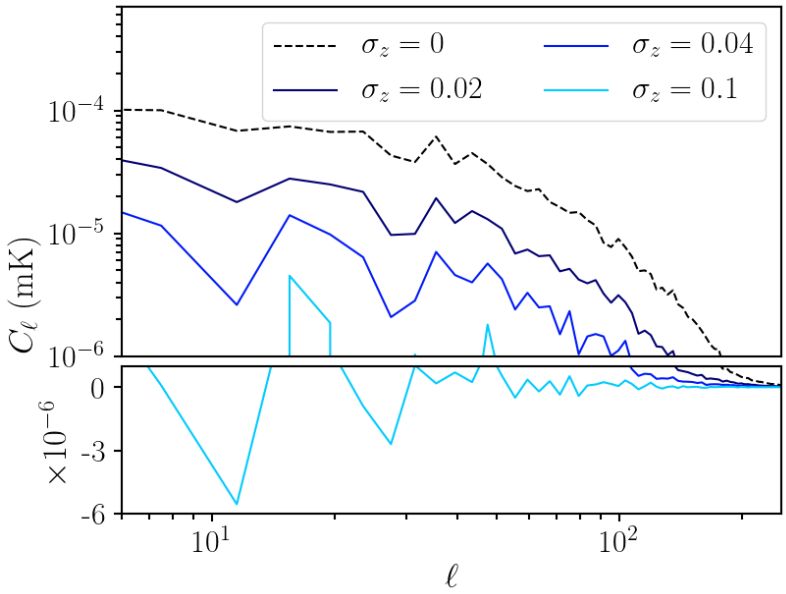

Figure 4.12: Cross-correlation between HI intensity maps with FASTICA reconstruction and an optical survey using GAEA at $z = 0.25$ with $\Delta z = 0.02$ bin width. I degrade the constraints on the optical galaxy redshifts by increasing the redshift error σ_z shown by going from dark to ligher blue. In other words I go from cross-correlating intensity maps with a spectroscopic-like ($\sigma_z \sim 0$) survey, to a photometric-like survey where there is significant uncertainty on the optical galaxy redshifts. This strongly affects the measured cross-correlation power spectrum. Plot includes a hybrid log-linear y-axis to fully demonstrate the degradation in power.

4.5.1 Optical Redshift Uncertainty

Future optical galaxy redshift surveys such as LSST and *Euclid* will rely on using photometric redshifts for estimating the radial position of each galaxy (note that *Euclid* will also perform a wide spectroscopic survey). It is therefore important to forecast the cross-correlation potential between HI intensity maps and photometric galaxy redshift surveys, taking into account foreground removal effects. The higher uncertainty on redshift measurement inherent in these photometric surveys, equates to a degradation in radial mode measurement on small scales. Since foreground removal also impacts radial modes but on larger scales, it is unclear whether combining these two effects will leave enough useful modes for a cross-correlation signal [226].

To investigate this I begin by simply introducing a Gaussian error on the optical redshifts for each galaxy and cross-correlate with foreground contaminated intensity maps. Figure 4.12 shows the effect on the cross-power spectrum when I introduce a Gaussian photo-z error σ_z into each of the optical galaxies. We can see how increasing the uncertainty in redshift (dark to light blue lines) rapidly degrades the agreement with the original (black-dashed line) where no redshift error is applied. [5] suggests a fiducial model of $\sigma_z = \sigma_{z_0}(1+z)$ is appropriate for an LSST-like instrument, where $\sigma_{z_0} = 0.05$. Therefore, the fact that Figure 4.12 suggests the cross-power spectra signal-to-noise will be damaged for $\sigma_z \sim 0.1$, which would correspond to LSST's photo-z error at $z = 1$, is cause for concern.

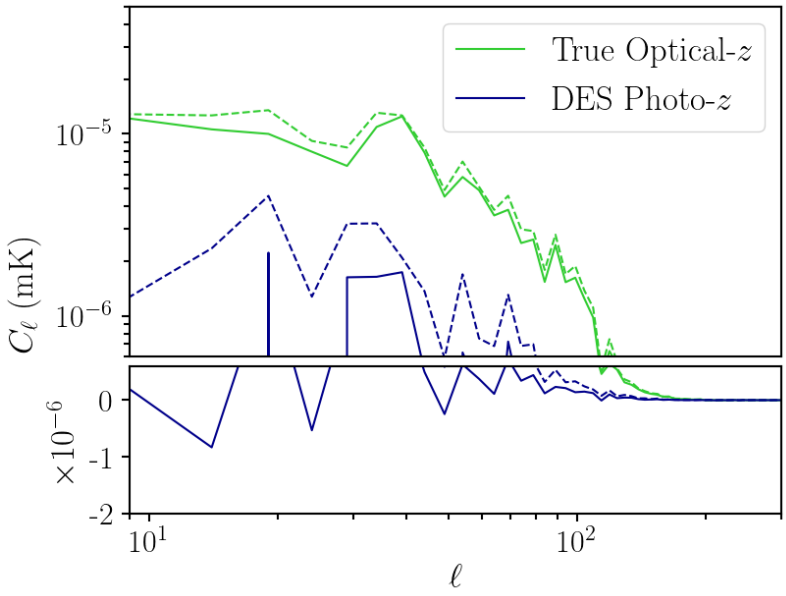

Figure 4.13: Cross-correlation between HI intensity maps with MICE optical galaxies. Dashed lines show the cases without HI foregrounds, solid lines show the impact of including them. I use the DES-like photometric redshifts available in MICE for the photo-z forecasts shown in blue and compare these with using ideal true redshifts (green). While a drop in signal is inevitable when using less constrained redshifts, including the effects of HI foregrounds (solid lines) degrades the signal further in the photo-z case. These tests have been performed at redshift $z = 0.725$ with $\Delta z = 0.05$ bin width.

We can further explore this with the use of some more robust photometric redshift simulations and compare to foreground free cross-correlations. Realistic photometry for a number of optical surveys is included within the MICEv2 simulation, for example the Dark Energy Survey (DES)[7]. I thus make use of the DES-like photometric redshifts available to make a more robust forecast of the cross-correlation between a photometric survey and HI intensity maps. I refer the reader to the MICE website[8] for more details on how these DES-like photometric redshifts were simulated.

Figure 4.13 shows the results when I include these simulated DES-like photometric redshifts in my simulations. The dashed lines show the cross-correlation power spectrum with the original HI intensity map with no foreground contamination. The solid lines then show the inclusion of foregrounds and a FASTICA reconstruction. What is clear from this plot is that while we still get a degradation in signal from using photometric redshifts (blue line) compared with true redshifts (green line), the signal deterioration accelerates in the case where HI foregrounds are included in the simulation.

The conclusion from the GAEA simulation using Gaussian photometric redshifts and MICE using DES-like photometric redshifts appears to be the same and both forecast damaging signal loss when FASTICA reconstructed intensity maps are cross-correlated with photometric redshift

[7] www.darkenergysurvey.org/
[8] http://maia.ice.cat/mice/

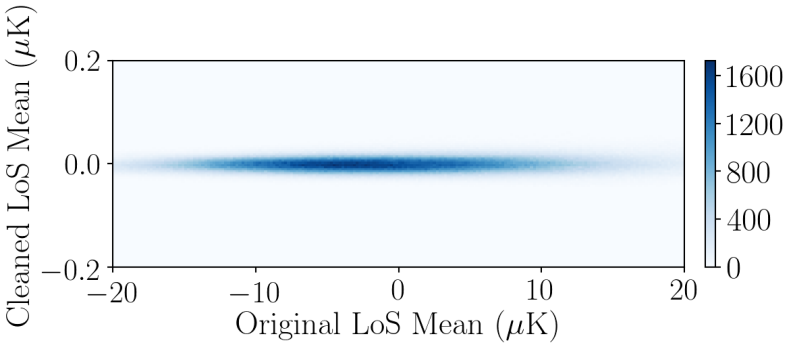

Figure 4.14: The mean δT temperatures along the line-of-sight (LoS) for the original HI intensity map against one with which has undergone a FASTICA foreground clean. This is shown for all available LoS in the GAEA simulation which for $f_{\rm sky} = 0.5$ and `nside` = 512 equates to over 1.5 million pixels (or LoS). The plot shows how FASTICA essentially removes any non-zero LoS mean present in the original HI signal and collapses it to zero.

surveys.

4.5.2 Mitigating the Effects of FASTICA

Here I begin investigating the precise reasons why combining the effects of HI foregrounds and the poor redshift constraints from photometric galaxy surveys is so detrimental to the cross-correlation signal. Generally, it can be considered unsurprising that combining an effect that removes information at large radial modes, with a survey which has poor constraints at small radial modes, can damp the amplitude of projected angular power spectra, as we see in Figures 4.12 and 4.13. The aim here is to quantify this explanation with the hope of being able to provide a solution.

It is interesting to look at the effects a foreground clean has along the LoS of the HI intensity mapping data. It is known that large radial modes are removed since this is where the contamination from foregrounds lies due to their smooth variation in frequency. Figure 4.14 shows the specific effect this has and illustrates how the foreground clean removes all information on the mean temperature along the LoS. My simulations are arranged such that the transverse mean of each map is zero but even with this setup it is of course still possible to have a large range of values for the LoS mean temperatures, which is what we see in Figure 4.14. However, we can see that the large range of LoS mean values present in the original HI signal (shown on the x-axis) are removed after the foreground clean to a much narrower range (shown on the y-axis). It is worth pointing out that the y-axis range is two orders of magnitude smaller than the x-axis. So essentially a blind foreground clean will destroy any non-zero mean along the LoS. The original line-of-sight means have a slight skewness away from zero and centre at around $-4\,\mu$K. This is caused by the presence of some dominant bright pixels which, when setting transverse means in each map to zero, can result in there being more negative temperatures than positive ones.

It is conceivable that an increase in the number of redshift bins could affect this LoS result, so I therefore conducted a test using the MICE catalogue and extended to 240 redshift bins

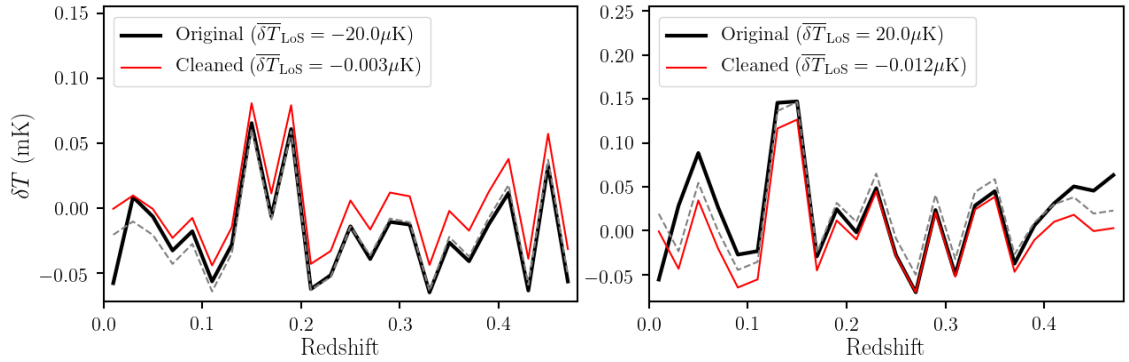

Figure 4.15: GAEA δT amplitudes along chosen lines-of-sight (LoS). Original mean values along the LoS are given in the legend along with the cleaned ones. The thick black line shows the original amplitude and the red solid line shows the impact of a foreground contamination and FASTICA foreground clean. The grey dashed line shows the amplitude with the LoS mean added back on as outlined in equation (4.29).

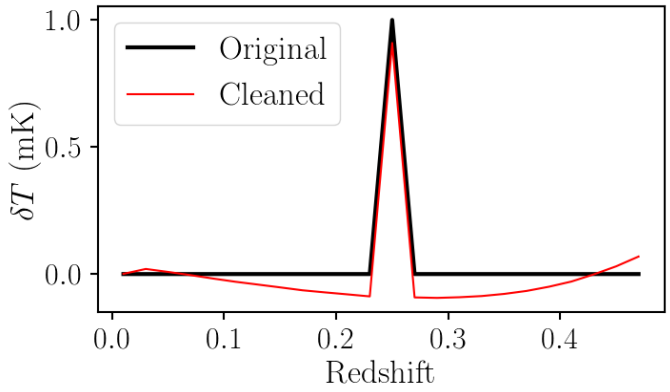

Figure 4.16: Effect of FASTICA on a test response function. For the GAEA model, all values along a chosen LoS have been set to 0 except one at $z = 0.25$ which is set to 1. This data is then subject to a FASTICA clean. An amplitude change from the LoS mean removal is apparent and there are also under-dense side-lobes either side of the temperature spike.

following the same procedure. Even with this more realistic number of redshift bins I still find a similar result to Figure 4.14 suggesting that this is not a feature of the relatively low number of redshift bins I am using.

In summary, the problem is that while FASTICA reconstructs the shape of the LoS signal, unfortunately it changes the amplitudes in an unpredictable manner based on the original LoS mean. The further from zero a particular LoS mean is, the greater the change in amplitude for pixels along this LoS. I attempt to model this by hypothesising that in a blind foreground clean the main resulting change is given by

$$\delta T_{\rm clean}(\vec{\theta}, \nu) \sim \delta T_{\rm orig}(\vec{\theta}, \nu) - \overline{\delta T}_{\rm LoS}(\vec{\theta}), \tag{4.27}$$

where $\overline{\delta T}_{\rm LoS}(\vec{\theta})$ is the mean fluctuation along a LoS for a pixel at position $\vec{\theta}$,

$$\overline{\delta T}_{\rm LoS}(\vec{\theta}) = \frac{\sum_i \delta T_{\rm orig}(\vec{\theta}, \nu_i)}{N_z}\,, \tag{4.28}$$

where the summation is over the N_z number of frequency (or redshift) bins.

Figure 4.15 shows the impact along the LoS resulting from the effect outlined in equation (4.27). I have chosen two pixels and show their δT values through redshift, taking two extreme examples for demonstrative purposes. The plot on the left is for a pixel where the original LoS mean $\overline{\delta T}_{\rm LoS}(\vec{\theta})$ is from the extreme low end from Figure 4.14. The plot on the right is for a pixel with a high $\overline{\delta T}_{\rm LoS}$. In both cases their LoS means are collapsed to zero for the reasons discussed above and the impact this has on the agreement between individual values through redshift is evident.

We can demonstrate that this is the main impact of a blind foreground clean by reversing the effect, i.e. adding back in the original LoS mean to each foreground-removed pixel:

$$\delta T_{\rm HI}(\vec{\theta}, \nu) = \delta T_{\rm clean}(\vec{\theta}, \nu) + \overline{\delta T}_{\rm LoS}(\vec{\theta})\,. \tag{4.29}$$

The corrected $\delta T_{\rm HI}$ should agree with the original signal $\delta T_{\rm orig}$. I have tested this and find this to be the case and show the results of this approach in Figure 4.15, where I have included the reconstructed LoS based on equation (4.29) shown by the gray dashed line.

Unfortunately, this LoS HI mean reconstruction is challenging in reality. The original $\overline{\delta T}_{\rm LoS}$ will be information buried in the foreground contaminated maps, and which is then lost after the foreground clean. So performing the process outlined in equation (4.29) would require some extra information to reconstruct these LoS means.

In a similar demonstration to Figure 4.15, I also analyse the FASTICA result on a test response function in the form of a Dirac-delta spike in temperature, shown in Figure 4.16. By manipulating the GAEA data such that all pixels along a chosen LoS are set to 0 except for one which is set to 1, we can gain a deeper insight into the effects of a foreground clean. The large side-lobes which form either side of the temperature spike can explain why the cross-correlation with photometric galaxy data is performing so badly. A galaxy at $z = 0.25$ with high measured redshift uncertainty, is likely to cross-correlate with the false under-temperature regions. This effect, compounded over many galaxies and temperature spikes, could cause signal loss.

As an additional problem, I also find that this kind of foreground removal is less successful at the extremes of the redshift range (something already concluded from Figure 4.9). Therefore reconstructing the LoS means will not be a sufficient correction on its own at the redshift edges of the data.

All this highlights the problems for the future success of HI intensity mapping cross-correlations with photometric galaxies. Nevertheless, photometric galaxy surveys are an important choice of probe to cross-correlate with given their complementary strengths, i.e. good angular resolution for optical and good radial resolution for HI intensity maps. I therefore suggest potential methods to mitigate the effects which a blind foreground clean has on HI intensity maps. These not only serve to drastically improve cross-correlations with photometric optical data, but also

provide additional improvements in cross-correlations with spectroscopic galaxy surveys, as well as HI intensity mapping auto-correlations. The two methods I propose are:

- **LoS Mean Reconstruction:** This is theoretically possible using optical galaxies which measure density along the LoS. By relating the optical over-density to the HI temperature we can make a prediction for the LoS mean HI temperature that has been removed and reverse the effect of this loss of information.
- **Artificial Extension of Redshift Range:** Introducing a buffer at either end of the data sets in the redshift (or frequency) direction will limit edge effects and as I will demonstrate, improves the general agreement with the original data.

I discuss both of these methods in more detail in the following sub-sections.

4.5.2.1 Line-of-Sight Mean Reconstruction

While recovering the exact LoS means from the intensity map data is not possible (they are inaccessible before the clean, and removed after it) we can make predictions of what they are from other data. Then by measuring the angular power spectrum of the LoS mean predictions, we can reverse the effects of the LoS mean loss. To understand this further, consider the hypothesis in equation (4.29) we can write

$$\langle \delta_{\rm g} \delta T_{\rm HI} \rangle = \langle \delta_{\rm g} \delta T_{\rm clean} \rangle + \langle \delta_{\rm g} \overline{\delta T}_{\rm LoS} \rangle \,, \tag{4.30}$$

and similarly for the auto-correlation we have

$$\langle \delta T_{\rm HI} \delta T_{\rm HI} \rangle = \langle \delta T_{\rm clean} \delta T_{\rm clean} \rangle + 2 \langle \delta T_{\rm clean} \overline{\delta T}_{\rm LoS} \rangle + \langle \overline{\delta T}_{\rm LoS} \overline{\delta T}_{\rm LoS} \rangle \,. \tag{4.31}$$

Therefore, for a cross-correlation we require the correction term $\langle \delta_{\rm g} \overline{\delta T}_{\rm LoS} \rangle$ and for an auto-correlation we require $2\langle \delta T_{\rm clean} \overline{\delta T}_{\rm LoS} \rangle + \langle \overline{\delta T}_{\rm LoS} \overline{\delta T}_{\rm LoS} \rangle$. We can utilise the optical number density fields to make estimates for these terms. This is because we can relate the optical over-density $\delta_{\rm g} = b_{\rm g} \delta_{\rm M}$ to temperature fluctuations $\delta T_{\rm HI} = \overline{T}_{\rm HI} b_{\rm HI} \delta_{\rm M}$ through

$$\delta T_{\rm orig}(z_i) = \frac{\overline{T}_{\rm HI}(z_i) b_{\rm HI}(z_i)}{b_{\rm g}(z_i)} \delta_{\rm g}(z_i) \,. \tag{4.32}$$

Then we relate this to each LoS mean by

$$\overline{\delta T}_{\rm LoS} = \frac{1}{N_z} \sum_i \frac{\overline{T}_{\rm HI}(z_i) b_{\rm HI}(z_i)}{b_{\rm g}(z_i)} \delta_{\rm g}(z_i). \tag{4.33}$$

This is all that is required to construct the correction terms for the cross- and auto-correlations outlined by equations (4.30) and (4.31). This approach does not require precise optical redshift information for the $\delta_{\rm g}(z)$. It is sufficient to use the poorly constrained photometric redshifts since the error on these should not heavily impact on the slowly varying summation kernel $\overline{T}_{\rm HI}(z) b_{\rm HI}(z) / b_{\rm g}(z)$.

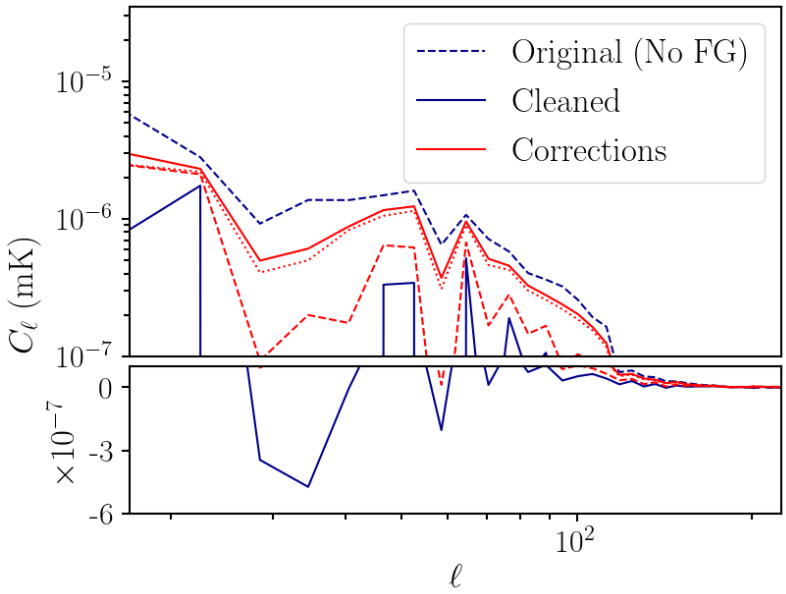

Figure 4.17: Cross-correlation angular power spectrum for the MICE simulation at redshift of $z = 1.075$ with $\Delta z = 0.05$ bin width and like Figure 4.13 I have used the DES-like photometric redshifts available in MICE. Again the impact from foregrounds is visible in the difference between the blue dashed line and blue solid line. However, the effectiveness of the corrective techniques that I outlined in Section 4.5.2, shown by the red lines, is encouraging. The dashed red line is for the LoS mean correction, the dotted red line represents the extended-z correction and the red solid line represents both corrections applied. Produced using bandpowers with 6 multipoles per bin for clarity.

The prefactor $\overline{T}_{\rm HI}(z)b_{\rm HI}(z)/b_{\rm g}(z)$ is not directly observable and therefore requires independent modelling or indirect measurement. $\overline{T}_{\rm HI}$ (equation (4.9) and discussed thereafter) is degenerate with $b_{\rm HI}$. Note that redshift space distortions can break this degeneracy and constrain $\Omega_{\rm HI}$ and consequently $\overline{T}_{\rm HI}$ [135]. For the purpose of testing this correction method I assume $\overline{T}_{\rm HI}$ has been accurately obtained, i.e. I simply use the model (4.9) which my simulated intensity maps have been designed to conform to. For the bias terms I determine them based on fiducial models where

$$b_{\rm g}(z) = 1 + 0.84z\,, \tag{4.34}$$

which was estimated from simulation results in [220] and used in the LSST Science book [5]. Following [23] I model the HI bias as

$$b_{\rm HI}(z) = 0.67 + 0.18z + 0.05z^2\,. \tag{4.35}$$

4.5.2.2 Artificial Extension to Redshift Range

While the reconstruction of the LoS means works reasonably well for the mid-range redshifts, improvements can still be made especially to the edge effects caused by a foreground clean. These edge effects have been previously noted and suggestions have been made to exclude these contaminated regions [228][227]. One simple solution to mitigate this effect and limit the data

excluded, is to extend the range of the data with the idea that the new artificial edges suffer the edge effect problems, but can then be removed from the rest of the data. I therefore take the full observed signal in the original N redshift bins given by

$$[z_1, z_2, ..., z_{N-1}, z_N]$$

and pad both ends with replicated reversed data to become

$$[z_N, z_{N-1}, ..., z_2, z_1, z_1, z_2, ..., z_{N-1}, z_N, z_N, z_{N-1}, ..., z_2, z_1].$$

So I have added reversed copies of the data to the beginning and the end of the original redshift range. This ensures the padded data includes continuous foregrounds since this is what a blind foreground clean needs to utilise in order to remove them.

Figure 4.17 shows the performance of these corrections on the MICE catalogue. I have shown this at a redshift of $z = 1.075$ which is closer to the extreme end of the redshift range for MICE and therefore has more need for correction. The solid blue line which shows the cross-correlation signal for FASTICA foreground cleaned map demonstrates how poor the signal is without any correction. The solid red line then shows that with the artificial extension to the redshift ranges and the LoS mean corrections to the power spectrum outlined by equation (4.30), the signal is significantly recovered and approaches the original signal with no foregrounds (blue dashed line).

I also demonstrate the more general improvement made across all redshift bins with Figure 4.18 which is for the GAEA simulation. Using the relative difference between original and clean power spectra as a gauge of performance (stated above the colour-bar), this shows how improved the signal is across all redshifts and scales with the corrections in place. We still see some poor disagreement in the very first redshift bin and slightly poorer performance for the last few bins, but the catastrophic discrepancies that we were seeing previously have been addressed.

These results are encouraging and suggest that with further refinement and understanding, cross-correlations between foreground cleaned intensity maps and photometric imaging surveys should be a useful probe of cosmology. I stress that the suggested corrections need further testing, preferably alongside real data to ensure they are reliable.

4.6 Clustering-Based Redshift Estimation

As a direct example of the potential impact that foreground removal can have on cross-correlations with photometric redshift surveys, I now aim to use these simulations to see if a photometric calibration method using such cross-correlations is still viable. This method utilises the shared clustering signal between photometric optical galaxies and overlapping HI intensity maps. This clustering-based redshift estimation process has previously been studied in [13] and [59] (see Chapter 3), but a full analysis including simulated foreground contamination has not yet been conducted.

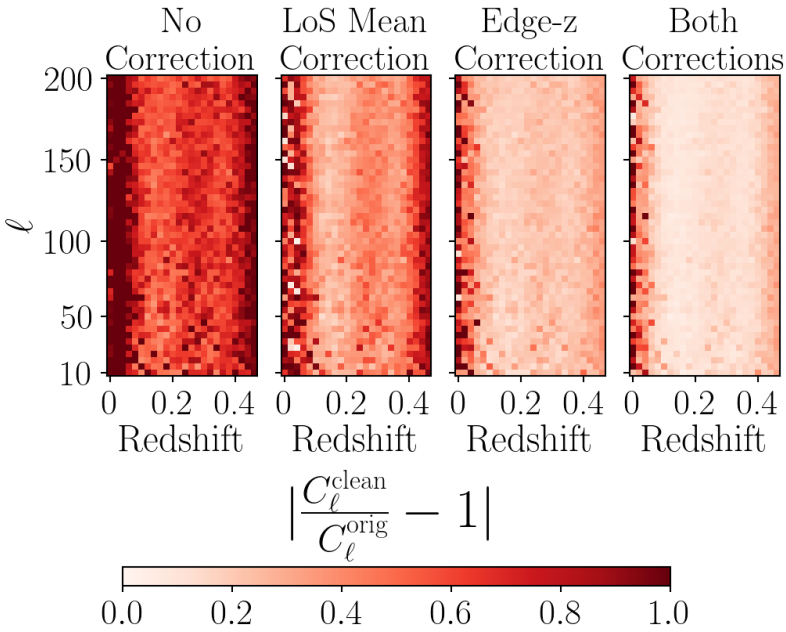

Figure 4.18: Demonstration of improvement on cross-correlation by including the corrections to the data outlined in Section 4.5.2. This is for the GAEA data-set and shows relative differences for cross-correlation of optical photometric-like data with HI intensity maps for the original (no foregrounds) and cleaned cases. For the optical sample I used a catalogue with redshift error of $\sigma_z = 0.06$.

Given the difficulties outlined in Section 4.5, such a method represents a stern test since the intensity maps are correlated with a population of optical galaxies where little redshift information is assumed. The only assumption made is that the optical galaxies are within the redshift range covered by the reference intensity maps. This is applicable to weak-lensing probes where wide redshift bins may be used, and where the aim is to obtain the source distribution which is required for precise measurements of cosmological shear. This wide redshift binning would mean huge degradation in small-scale radial modes, which is a major obstacle for this method given the increased noise due to the redshift uncertainty as outlined in Figure 4.12.

4.6.1 HI Clustering-z Method

In order to make a prediction for the redshift distribution of optical galaxies we require an estimator which utilises the shared clustering signal between the opticals and the HI intensity maps. I use the following estimator and refer the reader to [59] (and Chapter 3) where a full derivation is given:

$$\frac{\mathrm{d}N_{\rm g}}{\mathrm{d}z}(z) = \frac{w_{\rm g,HI}(z)}{w_{\rm HI,HI}(z)} \overline{T}_{\rm HI}(z) \frac{b_{\rm HI}(z)}{b_{\rm g}(z)} \frac{1}{\Delta z}. \tag{4.36}$$

Here I use angular correlations functions w where $w_{\rm g,HI}(z)$ is the cross-correlation between all the optical galaxies and an HI intensity map at a redshift z. Similarly, $w_{\rm HI,HI}(z)$ is the auto-correlation between two intensity maps at redshift z. An effective test of this estimator given the contamination of foregrounds is to use information from the C_ℓ power spectra since this

is a measurement of angular clustering which is what we want to utilise for estimating $\mathrm{d}N_\mathrm{g}/\mathrm{d}z$. An effective measurement for the angular correlation functions, which closely follows previous clustering redshift work [143] is given by

$$w_{XY}(z) = \int_{\ell_\mathrm{min}}^{\ell_\mathrm{max}} W(\ell) C_\ell^{XY}(z) \mathrm{d}\ell \,, \tag{4.37}$$

where $W(\ell)$ is a weight function which can be tuned to certain scales. For our purposes $W(\ell) = \ell$ is sufficient to give weight to smaller scales where more useful matching is expected to exist. Further investigation could be carried out into this to determine the function for $W(\ell)$ which delivers optimal weighting. As previously, the indexes X and Y can either be chosen to represent the HI intensity map auto-correlation where $X = Y = \mathrm{HI}$ or the cross-correlation with the optical where $X = \mathrm{g}$ and $Y = \mathrm{HI}$.

As before, $\overline{T}_\mathrm{HI}$ is the average brightness temperature which is known in the simulations. In reality however, the observable is a temperature fluctuation and $\overline{T}_\mathrm{HI}$ requires modelling as explained previously in equation (4.9). Again, for these purposes I assume an accurate modelling of $\overline{T}_\mathrm{HI}$ has been achieved, i.e. I simply measure the quantity in the simulations.

Finally, the estimator in equation (4.36) also requires the bias ratio $b_\mathrm{HI}/b_\mathrm{g}$. We can find this from the angular auto-correlation power spectra for the two samples:

$$\frac{b_\mathrm{HI}(z)}{b_\mathrm{g}(z)} = \frac{1}{\overline{T}_\mathrm{HI}(z)} \sqrt{\frac{C_\mathrm{HIHI}(\ell, z)}{C_\mathrm{gg}(\ell, z)}} \,. \tag{4.38}$$

However this relies on binning the galaxies by true redshift to measure the bias at that redshift. But I choose to assume that the optical sample has very poorly known (effectively unconstrained) redshifts, since it will be these surveys where redshift calibration is most in demand, so obtaining $C_\mathrm{gg}(z)$ accurately is not possible. For this study I therefore rely on fiducial models of the individual biases as laid out in equations (4.34) and (4.35).

4.6.2 HI Clustering-z Results

We are now ready to present a simple test of the HI clustering-based redshift estimation method and demonstrate its capability of recovering a redshift distribution using the HI intensity maps discussed in Section 4.2.1 for the simulated optical photometric sample with a detection threshold applied as discussed in Section 4.2.2. This is all in the presence of 21cm foreground contamination which has been cleaned using a FASTICA process. I also apply the corrections as outlined in Section 4.5.2 using only the photometric redshift information available.

I test this approach on both the GAEA and MICE based simulations and Figure 4.19 shows the results. In both cases I select optical galaxies based on their photometric redshifts in targeted redshift ranges shown as the pink shaded regions on the plots. Because these galaxies have been selected using their poorly constrained photometric redshifts, the true redshift distribution (black dashed line) extends way beyond these ranges. By cross-correlating with HI intensity maps and using the estimator outlined by equation (4.36) we can make a prediction of this true redshift distribution, shown by the blue data points. The grey shaded distributions show

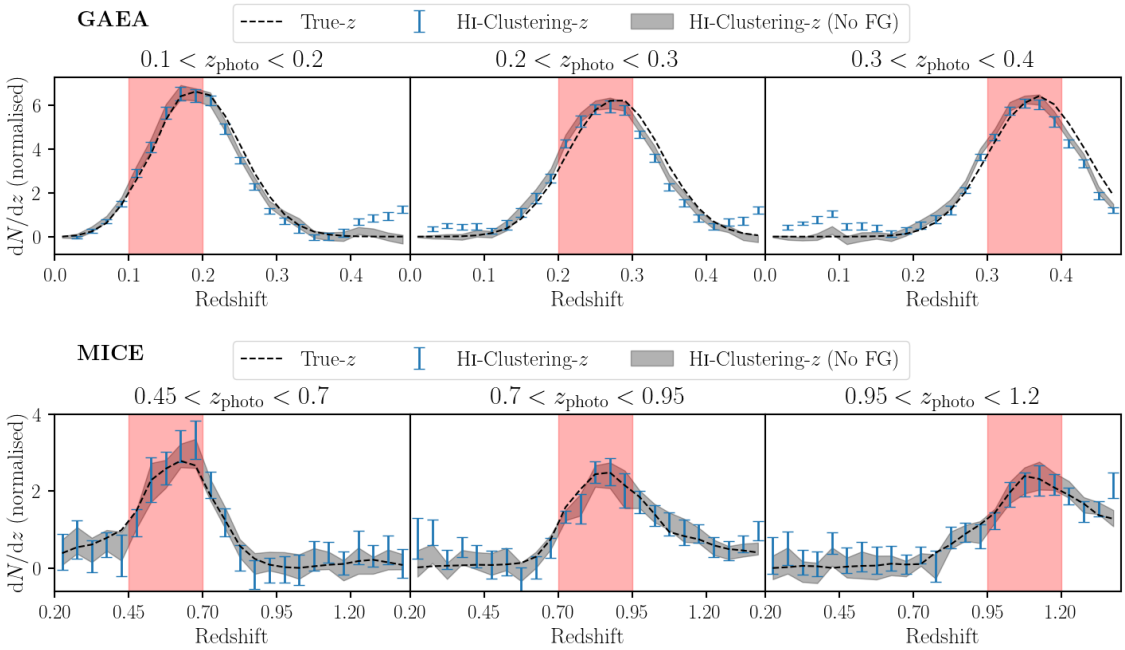

Figure 4.19: Clustering-based redshift estimation results using both the GAEA and MICE catalogues. The pink vertical shaded regions represent the optical sample chosen as galaxies whose photometric redshift lies within the targeted redshift ranges. The black dashed lines show the true redshift distributions of these galaxies. The blue data points give the estimated redshift distributions based on cross-correlations with HI intensity maps and using the estimator in equation (4.36). Intensity maps have foregrounds added and then removed with the FASTICA process with corrections made (Section 4.5.2). I also include the estimated distributions with errors from intensity maps absent from any foreground contamination, shown as the grey shaded distribution. The GAEA model uses a beam size of $\theta_{\rm FWHM} = 1.46^{\rm o}$, representative of an SKA-like beam for that redshift range. However, for MICE I have used a smaller beam size of $\theta_{\rm FWHM} = 1^{\rm o}$ because of its smaller sky coverage.

the result without any foreground contamination. I obtain the error bars for $\mathrm{d}N/\mathrm{d}z$ using a jackknifing technique, gridding the maps into an array of 25 smaller sub-samples. I then measure the estimator on the map but omit one of the 25 sub-samples. I repeat the procedure, averaging over the estimators obtained from omitting sub-samples, and obtain a standard deviation.

These results are very encouraging for the future of using shared clustering signals from HI intensity maps to calibrate photometric redshifts. A small bias is present which appears to skew the distribution, most evident in the GAEA results where the error is low. This will be caused by the fiducial bias models I use (equations (4.34) and (4.35)) in the estimator (4.36) not agreeing precisely with the simulated catalogues. More focused follow-up on this bias factor is required, as discussed in the previous section, and an improved approach which constrains the biases and mean HI temperature should mitigate this slight skewness.

Small discrepancies tend to exist at the extreme ends of the redshift distribution. When the true redshift distribution at these edges should be close to zero, often the estimator in the foreground contaminated case, predicts a non-zero quantity. These are due to residual edge

effects not fully mitigated by the correction outlined in Section 4.5.2.2. Because of this, it is difficult to place quantitative interpretation on the results in Figure 4.19 without these edge discrepancies skewing the measurement. I calculate the Median Absolute Deviation (MAD) which provides a robust measure of the variability of the recovered distributions. As the name suggests, this process takes the deviation of the estimated data from the true value, orders them in terms of their absolute deviation value, then takes the median quantity. The MAD values are therefore calculated for the differences between the true and estimated distributions i.e. $\mathrm{d}N/\mathrm{d}z_{\mathrm{true}} - \mathrm{d}N/\mathrm{d}z_{\mathrm{est}}$, since this measurement will not be too sensitive to the incorrect estimations near the edges. I find that for the three GAEA distributions shown in Figure 4.19 these MAD values are 0.199, 0.167 and 0.284 for $0.1 < z_{\mathrm{photo}} < 0.2$, $0.2 < z_{\mathrm{photo}} < 0.3$ and $0.3 < z_{\mathrm{photo}} < 0.4$ respectively. For MICE, the MAD values for the differences in true and estimated distributions are 0.129, 0.107 and 0.132. The similar values in each simulation demonstrate that the redshift prediction method is behaving consistently. The relatively low MAD values, under 5% of the normalised $\mathrm{d}N/\mathrm{d}z$ peak value, also suggest the discrepancies between true and estimated distributions are mostly small and is indicative of the estimator's precision. This represents an excellent test of cross-correlations between foreground affected HI intensity maps and photometric surveys. This is because this method relies on sufficient cross-signal existing for poorly constrained optical redshifts over wide redshift ranges. The relative success of this method suggests that the problems outlined in Section 4.5.1 will be surmountable.

I found that a key factor regarding the success of the clustering-based redshift estimation method using HI intensity maps is the combination of the sky area and the size of the instrumental beam. [59] found that the error on the estimation is directly proportional to the beam size and can be approximated by

$$\sigma_{N(z)} \propto \frac{\theta_{\mathrm{FWHM}}}{\sqrt{A}}, \tag{4.39}$$

where A is the area of the sky covered. Due to the smaller sky coverage in the MICE simulation I found that I was unable to use a constant beam size of $\theta_{\mathrm{FWHM}} = 2.36^{\mathrm{o}}$ which would be representative of an SKA-like beam probing redshifts up to $z = 1.4$. Instead I have only smoothed with a 1^{o} beam. However, having larger sky coverage in future simulations would mitigate this issue. It is interesting to note how the error does not increase too much in Figure 4.19 with the inclusion of foreground contamination in the analysis (comparison between blue data points and grey shaded distribution). This supports the claim that the error from this estimator is largely dominated by the sky area and beam size and explains the larger errors on the MICE plot compared with GAEA.

4.7 Summary

Forthcoming HI intensity mapping experiments will be able to contribute to cosmological studies through HI auto-correlations as well as cross-correlations with optical galaxy surveys. To ensure that HI intensity mapping is a competitive technique, it is important to understand 21cm foreground contamination, and the effects of foreground removal on the measurements.

In this chapter I have taken a simulations-based approach to investigate these issues, focusing on the foreground removal effects on HI intensity mapping cross-correlations with photometric galaxy surveys. By using existing N-body simulations and the galaxy catalogues produced from them, I constructed both optical galaxy catalogue data and HI intensity map data with the same underlying cosmological clustering signal. I then simulated the relevant 21cm foreground signals that are expected to contaminate the HI intensity maps, and used a state-of-the-art blind foreground removal process known as FASTICA. This approach allowed me to then examine what impact this type of foreground removal has on cosmological probes such as the clustering measured by the angular power spectrum C_ℓ.

The main conclusions are as follows:

- I have shown evidence that a FASTICA reconstruction will successfully allow accurate auto-correlation measurements as shown by previous work [228][195]. Figure 4.10 showcases the results for both the simulations, GAEA and MICE. The better result obtained for the GAEA model is likely due to its larger sky size allowing for more samples to average over in negentropy calculations.

- The auto-correlation tests I performed strongly suggest that a frequency dependent beam size will cause problems for independent component-like methods as demonstrated in Figure 4.10 and also shown by [11]. A solution to this is to re-smooth the intensity maps to match the beam size for the highest redshift when using these foreground removal techniques.

- FASTICA also delivers good results in cross-correlation with optical galaxy data where the redshifts for the opticals are very well constrained as they would be in a spectroscopic-like survey. In Figure 4.11 I used optical galaxies with true redshifts in cross-correlation with HI intensity maps. The figure shows the excellent agreement between using the original (no foregrounds included) intensity maps and the foreground cleaned ones.

- I find that further treatment is needed when cross-correlating foreground cleaned HI intensity maps with photometric-like optical galaxy surveys with poor redshift constraints. Figures 4.12 and 4.13 show the impact of combining foreground cleaned intensity maps with an imaging galaxy survey which has poorly constrained redshifts. This poor result is unsurprising and can be generally explained by the combination of eroded large-radial modes caused by the foreground cleaning, with eroded small radial-modes caused by the uncertainty in the photometric redshifts [226].

- More specifically, I find that a cause of the poor results when considering HI $\times$ Photo-z is the loss of LoS mean information when conducting the foreground clean. Figure 4.14 shows

how any prior off-zero LoS means are collapsed to zero which has the effect of unpredictably changing pixel values in the transverse maps, as demonstrated in Figure 4.15. As a possible treatment for this unwanted effect I proposed a LoS reconstruction that uses information from the optical galaxies as outlined by equation (4.29). This, coupled with artificially extending the redshift range to mitigate the edge effects caused by the foreground clean, improves results as shown by Figures 4.17 and 4.18.

- Finally, I conducted a comprehensive test of these methods by attempting to use foreground contaminated intensity maps for clustering-based redshift estimation of a photometric optical sample. By using FASTICA and my additional corrections I was able to accurately predict the redshift distributions for mock optical catalogues in both the models (Figure 4.19).

This chapter used two independent N-body simulations, where one (GAEA) used a semi-analytical approach to constructing a galaxy catalogue and the other (MICE) used a HOD/HAM hybrid method. The resulting catalogues formed the basis for constructing the optical and HI intensity map mock data. This means we can be confident that the conclusions I have made are unlikely to be specific to these simulations.

A limitation in using existing mock galaxy catalogues to generate HI intensity maps however comes from the finite number of galaxies available to sample in the map. The great advantage of HI intensity mapping is the frequency resolution which allows for numerous tomographic bins. While the catalogues I use are large ($>10^8$ galaxies), this finite number means care was needed when going to large numbers of tomographic bins. If the bin is too thin, it will contain a low number of galaxies (sparse galaxy density), and therefore a sparse signal in each pixel. This is not an accurate emulation of an intensity map which should provide a near continuous emission profile. Tests were carried out with a higher number of bins in some cases. For example I used 240 redshift bins for the MICE catalogue and tested if we still see the LoS mean destruction demonstrated by Figure 4.14. Even with this more realistic number of bins, I find similar results but cannot be certain that these are accurate simulations of combined emission maps since the number density of simulated galaxies becomes low (~ 5 per voxel) at this fine radial resolution. This is why I used relatively thick tomographic bins in this chapter ($\Delta z = 0.02$ for GAEA and $\Delta z = 0.05$ for MICE).

Throughout this chapter I have made assumptions that parameters such as the mean HI temperature ($\overline{T}_{\rm HI}$) can be precisely obtained. While I use a model for this parameter in the analysis, this same model was used in the construction of the HI intensity map signal, therefore its success is unsurprising. However, other parameters such as the clustering bias terms ($b_{\rm g}$ and $b_{\rm HI}$) are not directly fed into the simulated signals, so the success of modelling these as scale-independent biases in the clustering-based redshift estimation is encouraging.

Note that in this chapter I have not simulated any foreground polarization leakage effects. However, in many frequency channels I have smoothed the maps more than is required to simulate the instrument beam, which is a treatment previously used in real data to mitigate these effects [205] as discussed in Section 4.3.1. It is unclear whether the required level of instrument

calibration is achievable to avoid effects such as polarization leakage. Therefore one could argue that it will not necessarily be the foregrounds themselves that cause the biggest problems, but instead the leakage of them through imperfect instrument calibration [140][196]. Therefore, a follow-up study with simulations of realistic observations including polarization leakage and other instrument systematics such as $1/f$ gain fluctuations, beam side-lobes, radio-frequency interference etc. [95] will be an important step.

Furthermore, in this chapter I did not consider the clustering of point source foregrounds, which one could argue has potential to bias cosmological clustering measurements. Nor in the simulations did I simulate the anisotropy of galactic free-free emission which is expected to be stronger in the galactic plane. However, neither of these subtle features are likely to affect the frequency coherence of the signals which FASTICA uses to isolate them.

In future work I plan to include a further analysis into the effects of foreground removal on cosmological measurements including the 3D correlation function $\xi(s)$ and power spectrum $P(k)$ multipoles, extending the work of [38].

As HI intensity mapping data becomes available alongside the plethora of high precision optical datasets, we will be able to confirm conclusions derived from simulated mocks using real observations. Future measurements of HI $\times$ Photo-z data, for example from MeerKAT and DES [169][23] or TIANLAI [51] and DECaLS[9] [39], will be an excellent test for the claims in this chapter. I have demonstrated the potential of such experiments with the example of how cross-correlations can be used for photometric redshift calibration. This is a major challenge for forthcoming Stage-IV instruments utilising photometric optical samples, such as LSST and *Euclid*. I believe that photometric redshift calibration using HI intensity mapping data is an alternative method with great promise for tackling this challenge.

To summarise, I have shown evidence that a method such as FASTICA performs excellently at reconstructing the inherently weak HI signal in the presence of dominant 21cm foreground contamination. Even in cross-correlation with optical data with poorly constrained redshifts, with the suggested corrections it is possible to make good measurements of the cosmological signal. I have introduced a LoS mean reconstruction as a treatment for foreground cleaned intensity mapping signal loss, which improves the fidelity of cross-correlation measurements but which will benefit from further investigation and refinement. Foreground contamination is a challenge for HI intensity mapping, but this work alongside others demonstrates that it is a surmountable one. I look forward to providing even more realistic simulations, and testing my proposed methods with real data, in the near future. In the following chapter I look to continue the investigation into the impact of 21cm foreground contamination by considering the 3-dimensional power spectrum and its multipole expansion.

[9] http://legacysurvey.org/

CHAPTER 5

3D POWER SPECTRA & MULTIPOLES

Many of the results in this thesis, in particular Chapter 4, focussed on the angular 2D measurements of clustering such as the angular 2-point correlation function $w(\theta)$ or the angular power spectrum C_ℓ. However, a large amount of cosmological information can be gleaned from the analogous 3D measurements $\xi(r)$ and the power spectrum $P(k)$. I have therefore been contributing towards a framework that can measure 2-point statistics of 3D fluctuations mainly in the context of HI intensity maps and optical galaxy surveys with the aim of including observational effects.

This chapter discusses some ongoing work that aims to test and extend upon the work in [38] (hereafter Blake19) which looked at modelling the power spectrum for galaxy and HI intensity map data including many observational effects. By providing larger simulations, further testing of this pipeline has been possible on galaxy × HI cross-correlations and also their auto-correlations.

A further contribution to extending this has involved including 21cm foreground contamination and cleaning, something not included in the original Blake19 paper. This is discussed in section 5.2 where interesting impact on the power spectrum multipoles from foregrounds is apparent. The beam, in conjunction with foreground contamination, also produces some intriguing effects which I discuss.

The results produced in this section have come from a collaborative effort between myself, Chris Blake and my supervisors Alkistis Pourtsidou and David Bacon. This chapter uses code[1] originally developed by Chris which has been extended by myself and Chris to allow for the inclusion of my MICE-based pipeline of HI intensity maps and galaxy catalogue simulations. We both also extended it to incorporate my 21cm foreground simulations and FASTICA foreground removal techniques (as outlined in Chapter 4). The result in Figure 5.1 was first shown by Chris. The first results showing the effects of foregrounds were produced by me. The splitting of the modes into μ-wedges was first done by Chris (shown in Figure 5.6) which Alkistis developed into

[1] https://github.com/cblakeastro/intensitypower

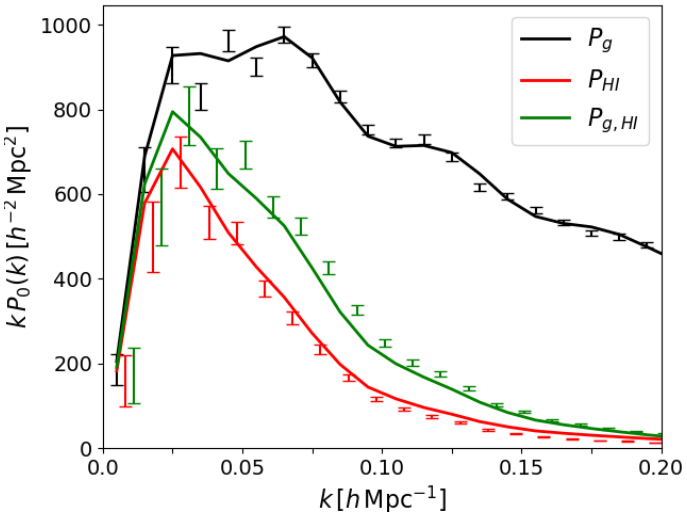

Figure 5.1: Replication of the Blake19 monopole $P_0(k)$ results at $z = 0.4$ with larger sky simulations built from MICE. Results are for both optical and HI auto-correlations (black and red lines) and cross-correlation (green line). Solid line shows the theoretical prediction with observational effects included.

a toy modelling of the foreground effects (shown by Figure 5.5) which I extended upon for the discussion in this chapter. All figures used in this chapter however, are versions which have been reproduced by me for the inclusion in this thesis.

5.1 Testing and Extending 3D Power Spectrum Measurement Pipelines

Work in Blake19 provided a formalism for how position-dependent selection functions, noise, weighting, smoothing, pixelization and discretization affect power spectra in HI × optical cross-correlations. As discussed in my introduction (Chapter 1), RSD cause the amplitude of the power spectrum to depend on the direction relative to the global LoS for a survey. Therefore the power spectrum can be parameterised by μ the cosine of the angle θ to the LoS, $\mu = \cos\theta$. This power spectrum $P(k,\mu)$ can then be quantified by its multipoles $P_\ell(k)$ where

$$P(k,\mu) = \sum_{\ell=0}^{\infty} P_\ell(k)\mathcal{L}_\ell(\mu) \tag{5.1}$$

and $\mathcal{L}_\ell(\mu)$ are the Legendre polynomials. From this we can separate the power spectrum into its multipole contributions. The monopole, quadrupole and hexadecapole (see Appendix A.1 for a derivation of these). By inverting equation (5.1) the power spectrum multipoles can be given as

$$P_\ell(k) = \frac{2\ell+1}{2}\int_{-1}^{1} \mathrm{d}\mu P(k,\mu)\mathcal{L}_\ell(\mu). \tag{5.2}$$

The Blake19 pipeline looks to include observational effects from;

- intensity mapping telescope beam

- frequency channel binning
- angular pixelization

which were modelled as damping effects on the power spectrum such that $P(\vec{k}) \rightarrow P(\vec{k})\, D^2(\vec{k})$, where the damping function is given by

$$D^2(\vec{k}) = \frac{1}{V} \int \mathrm{d}^3\vec{x}\, |\tilde{B}(\vec{k}, \vec{x})|^2. \tag{5.3}$$

Here $\tilde{B}$ is the Fourier transform of a dimensionless smoothing function which may vary with position $\vec{x}$ and is used to encapsulate of each the observational effects such that

$$\tilde{B}(\vec{k}, \vec{x}) = \tilde{B}_{\mathrm{beam}} \tilde{B}_{\mathrm{chan}} \tilde{B}_{\mathrm{ang}} = \exp\left(\frac{-k_\perp^2 |\vec{x}|^2 \sigma_\theta^2}{2}\right) \times \frac{\sin\left(k_\parallel s_\parallel(\vec{x})/2\right)}{k_\parallel \vec{s}_\parallel(\vec{x})/2} \times W_{\mathrm{ang}}\left(k_\perp |\vec{x}|\right) \tag{5.4}$$

is a combination from telescope beam[2], frequency channel, and angular pixelisation smoothing effects respectively. I refer the reader to Blake19 [38] for full derivations of these terms and a more detailed discussion of this topic.

Figure 5.1 shows results from using my MICE simulations (which I refer back to Section 4.2 for further details) on the extended multipole measurement pipeline developed in Blake19. We can see the theoretical predictions are agreeing nicely for these monopole measurements. There is arguably some discrepancy between simulation and theoretical model in the HI auto-correlation result (red line) which requires some additional treatment to improve the fit. This would most likely be improved with a more sophisticated non-linear model. We have tested these results at higher redshift and find we achieve a better fit thus suggesting non-linearities, which are less dominant at higher redshift, could be affecting results. This could also be impacting the cross-correlation result (green line) which has some mild discrepancy too. These results currently assume foregrounds have been perfectly cleaned, we therefore intend to include the effects of foregrounds, examine them, and then attempt to model them.

5.2 Impact of Foregrounds on Multipole Measurements

The effect of 21cm foregrounds upon measurements of the power spectrum multipoles is something that is yet to receive much detailed investigation. This was therefore one of our aims and we are in the process of extending the multipole pipeline from Blake19 to include foregrounds. To do this we are using the simulations in the work outlined in Chapter 4 and examining the difference between the multipoles of the power spectra using foreground contaminated maps, against completely foreground free ones.

The results for this are shown in Figure 5.2. As with Figure 5.1, these are produced using the simulations from the MICE maps. There are a few interesting conclusions to draw from these results, with foregrounds affecting all multipoles differently. Firstly for the monopole P_0, we see damping of power at lower values of k. It is these modes that are most likely to comprise large

[2] Clearly, smoothing effects from the radio telescope beam need only be included once for the cross-correlation i.e. $D^2(k) = 1/V \int \mathrm{d}^3x\, \tilde{B}_{\mathrm{beam}} \left|\tilde{B}_{\mathrm{chan}}\right|^2 \left|\tilde{B}_{\mathrm{ang}}\right|^2$ and not included at all for the optical auto-correlation.

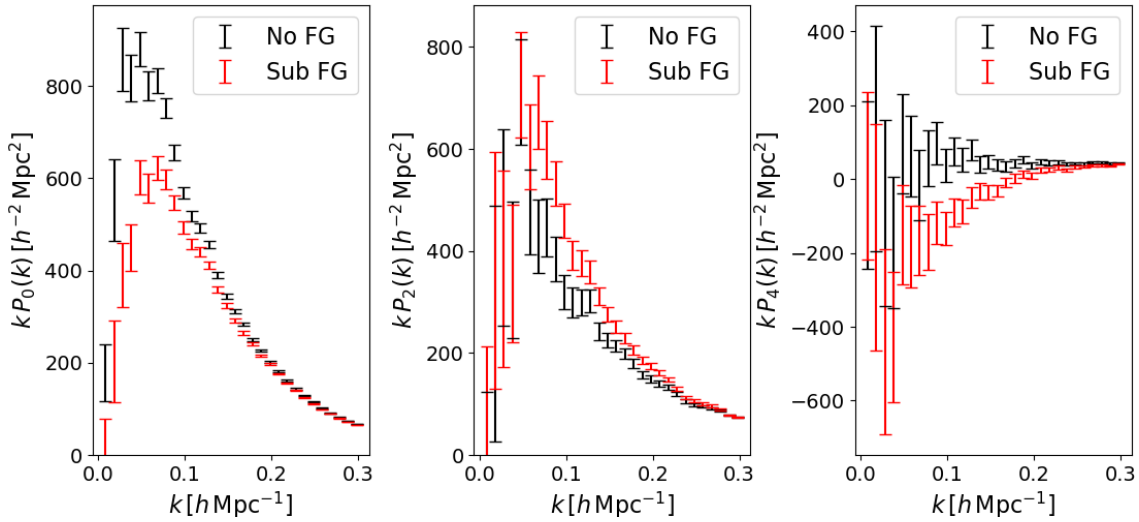

Figure 5.2: Comparison between HI auto-power spectrum multipoles at $z = 0.4$ with no foreground contamination (black points) and with simulated foregrounds added and then removed with FASTICA (red points). Monopole (P_0) shown in left panel, quadrupole (P_2) centre and hexadecapole (P_4) on the right.

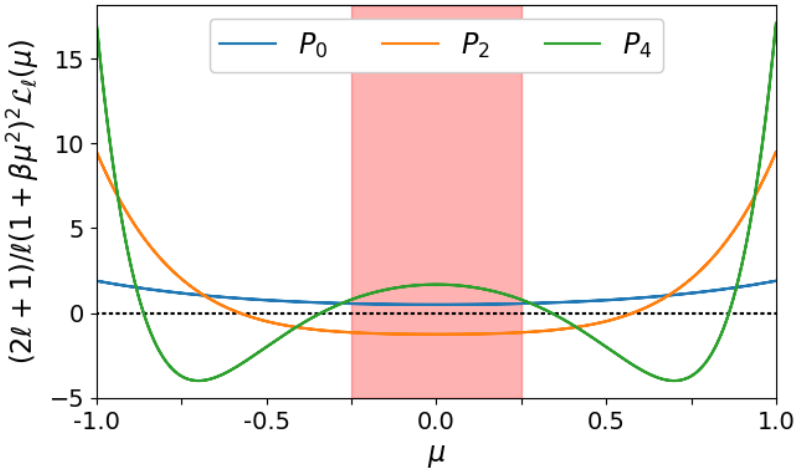

Figure 5.3: Integrands for the expanded multipole equations as a function of μ. Equations are outlined in Appendix equations (A.11), (A.12) and (A.13). The pink shaded region shows where $|\mu| < 0.25$ where large radial modes will dominate. These results use a value of $\beta = f/b_{\mathrm{HI}} \sim 0.95$, realistic for HI intensity maps at $z = 0.4$.

radial (small $k_\parallel$) modes, which are the ones that should be most affected by a foreground clean. The quadrupole P_2 seems to have an opposite effect and we actually get an enhancement of power from a foreground clean and the hexadecapole P_4 changes sign on large scales (small k).

To understand these observastions, it is useful to consider the parameter μ. Since we expect foregrounds to eliminate signal along the LoS, we can use μ to separate those parts of the signal with large contribution from radial modes since μ is the directional cosine of the modes i.e. $\mu = \cos(\theta)$ where θ is the angle to the LoS. These effects can then be understood by analysing the expanded multipole terms as a function of μ that are integrated over (shown in the Appendix

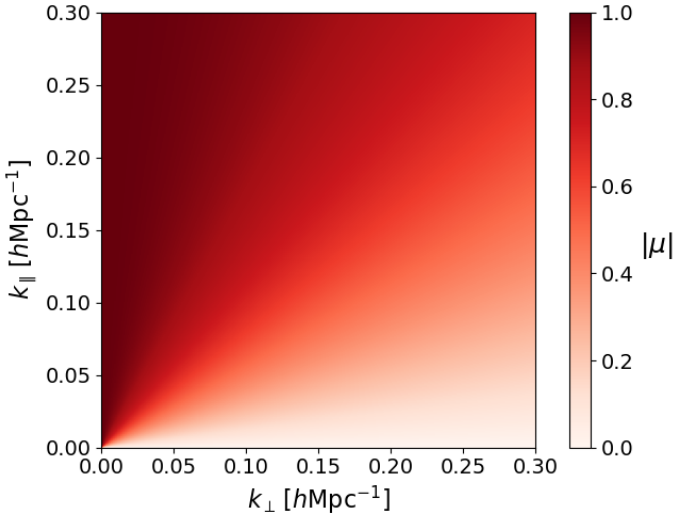

Figure 5.4: Demonstration of how μ, the directional cosine of modes, changes depending on the contributions from modes parallel and perpendicular to the LoS. This is calculated from $\mu = \cos\theta = k_\parallel / k = \mathrm{k}_\parallel / \sqrt{k_\parallel^2 + k_\perp^2}$.

equations (A.11), (A.12) and (A.13)). In Figure 5.3 I have plotted the integrands as a function of μ for these equations. Since $k_\parallel = k\mu$, we expect foregrounds to have the largest effect on modes with low-μ. This then explains some of the results we are seeing in Figure 5.2 as a foreground clean should have a similar effect to removing contributions to the multipoles from low-μ regions (e.g. $\mu < 0.25$ shown as the pink shaded region). Doing this removes a lot of the negative contribution in the quadrupole which is why we see an enhanced signal. Similarly, this also removes positive contributions to the monopole, hence why we see an overall damping here and the hexadecapole has enough positive contributions removed for its negative contributions to dominate.

Figure 5.4 shows how values of μ depend on the contributions from $k_\parallel$ and $k_\perp$ which are modes parallel and perpendicular to the LoS respectively. By considering this, we should be able to model the results we see in Figure 5.2 by calculating theoretical multipoles with low-μ modes removed. The theoretical multipoles are produced from an underlying matter power spectrum generated using using `Nbodykit` [92] which uses the Boltzmann solver package CLASS [124]. Figure 5.5 shows these theoretical models where for the subtracted foreground cases we have eliminated contributions with $|\mu| < \mu_{\mathrm{FG}}$ where we define μ_{FG} by

$$\mu_{\mathrm{FG}} = k_\parallel^{\mathrm{FG}} / k \tag{5.5}$$

where $k_\parallel^{\mathrm{FG}}$ is some parallel mode limit below which modes are rendered inaccessible by the foreground clean. The power spectra in Figure 5.5 are therefore given by

$$P_{\mathrm{HI},\ell}(k) = \frac{2\ell+1}{2\Delta\mu} \int_{|\mu|=\mu_{\mathrm{FG}}}^{|\mu|=1} \mathrm{d}\mu\, P_{\mathrm{HI}}(k,\mu)\mathcal{L}_\ell(\mu) \tag{5.6}$$

where $\Delta\mu \equiv 1 - \mu_{\mathrm{FG}}$ is the amount of μ-space being integrated over and if foregrounds are perfectly cleaned, then $\mu_{\mathrm{FG}} = 0$ and we recover the standard multipole expansion equation. For

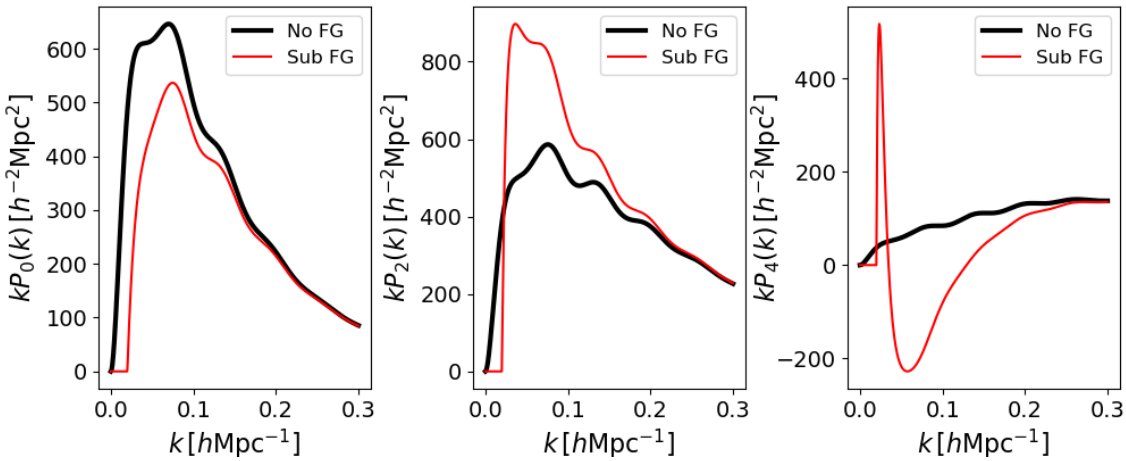

Figure 5.5: Theoretical multipole power spectra aiming to model Figure 5.2 by removing low-μ contributions to emulate a foreground clean. Here I have removed low-μ contributions for the foreground subtracted cases (red lines) as defined by equation 5.6. This uses a $\mu_{\rm FG}$ cut-off defined by $\mu_{\rm FG} = k_\parallel^{\rm FG}/k$ using $k_\parallel^{\rm FG} = 0.02h\mathrm{Mpc}^{-1}$. These results use a beam of $\theta_{\rm FWHM} = 0.44$deg which causes damping as outlined by equation (5.7).

this example we chose $k_\parallel^{\rm FG} = 0.02h\mathrm{Mpc}^{-1}$ consistent with previous work e.g. [196]. This chosen cut-off can be seen in Figure 5.5 where all power with $k < 0.02$ is lost which is because for these modes $|\mu_{\rm FG}| \geq 1$ according to equation (5.5) but since μ is the directional cosine and cannot be greater than 1, the integral in equation (5.6) collapses. This model also includes a damping effect from the telescope beam which affects small perpendicular modes such that $P_{\rm HI}(k,\mu)$ in equation (5.6) is given by

$$P_{\rm HI}(k,\mu) = \overline{T}_{\rm HI}^2 b_{\rm HI}^2 \left(1+\beta\mu^2\right)^2 e^{-(1-\mu^2)k^2\sigma^2} P(k) \tag{5.7}$$

and the exponential beam damping term comes from only perpendicular modes $k_\perp = k\sqrt{1-\mu^2}$ being smoothed with $\sigma = d_{\rm c}(z)\theta_{\rm FWHM}/2\sqrt{2\log 2}$. For the results in Figure 5.5 a beam of $\theta_{\rm FWHM} = 0.44$deg was used. While this is a rather crude toy model, it agrees nicely with the more robust results from simulations shown in Figure 5.2. We see the damping of the monopole, enhancement of the quadrupole, and sign reversal of the hexadecapole all as expected. This currently does not encapsulate other observational effects such as the telescope noise or pixelization but the aim is to extend this model and have a theoretical fit such as those used in Figure 5.1 (solid lines) but with foregrounds included in the fit.

Figure 5.6 shows the effect a foreground clean has on different 'wedges' defined by μ. This is also illustrative of how μ is an interesting parameter to examine in the context of foreground contamination. Since μ is the directional cosine of the modes and therefore $k_\parallel = k\mu$, a wedge with only low-μ included (as in the top-left plot of Figure 5.6) means only small $k_\parallel$ modes are included and these are the ones most affected by the foregrounds. These results in Figure 5.6 show that the impact of foregrounds becomes smaller as μ increases as expected.

Since the quadrupole and hexadecapole are seen as 'smoking-guns' for RSD, their detection using HI intensity mapping would be exciting progress. However, in order to make the first

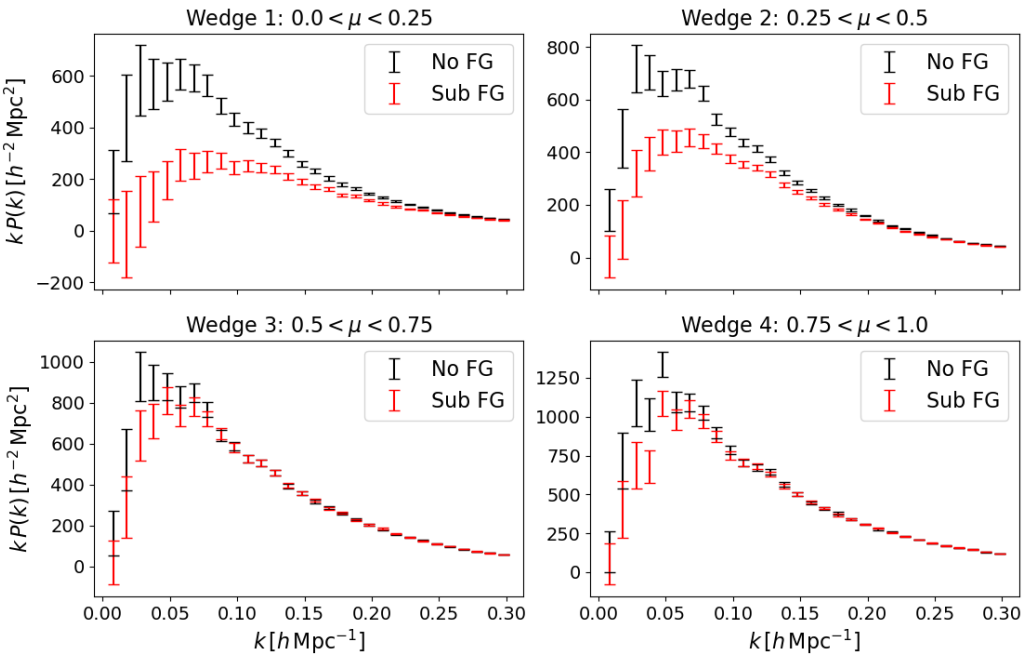

Figure 5.6: Multipoles separated into different 'wedges' defined by μ which is the directional cosine from the LoS i.e. $\mu = \cos(\theta)$ where θ is the angle from the LoS. Again I show the difference between no 21cm foregrounds (black line) and where foregrounds are added then removed with FASTICA (red line).

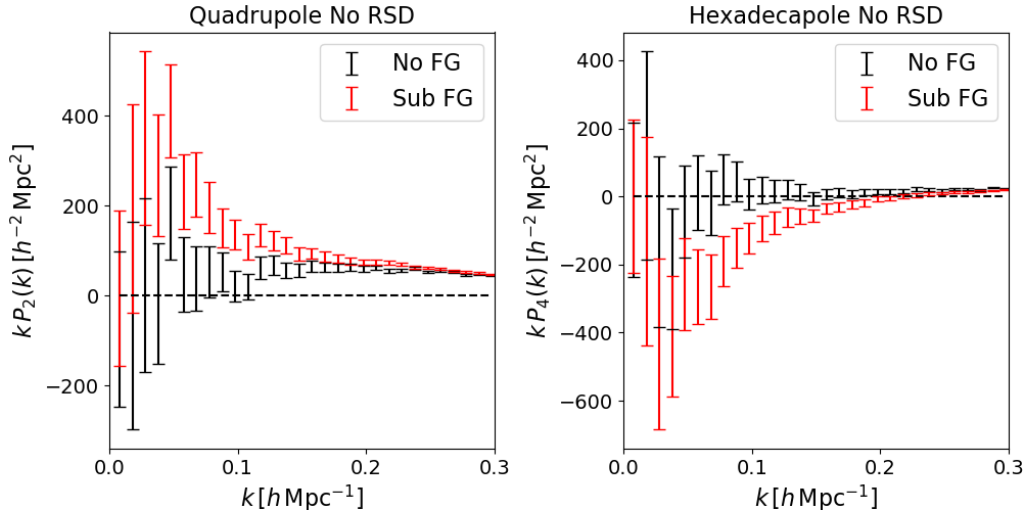

Figure 5.7: Quadrupole (P_2) and hexadecapole (P_4) for auto-correlations of HI intensity maps, produced using simulations with no RSD. Therefore, the results should be $P_2 = P_4 = 0$ which the foreground free results (black lines) are fairly consistent with but due to the presence of 21cm foregrounds (and some simulated systematic noise), a false signal appears for both.

detection of the quadrupole or hexadecapole, careful consideration would need to be given to the effects foreground removal has on intensity mapping data. What might appear as a detection, could just be systematics interacting with the Legendre polynomials to create an enhanced signal. In Figure 5.7 I have used the MICE simulation as done in the previous examples in this section, but instead set the galaxies peculiar velocities to zero thus resulting in no RSD. This should therefore result in a null quadrupole and hexadecapole. However, in the presence of foregrounds that have been removed using FASTICA, a false signal appears. Therefore any future claim of this kind of detection would need to be confident it is not just a foreground effect. Figure 5.7 does show some non-zero signal even in the no foreground case. This is being caused by the telescope beam which also introduces anisotropic effects since it only smooths modes in the perpendicular direction.

5.2.1 Increasing Beam

It is interesting to look at how the previous results in Section 5.2 change when the size of the telescope beam is increased. Previously I used a small beam of $\theta_{\rm FWHM} = 0.44$deg (corresponding to the size of a GBT-like beam [230]) but here I investigate the effects of increasing that beam to one which has $\sigma = 1$deg which is equivalent to a $\theta_{\rm FWHM} \sim 2.355$deg beam. It is straightforward to tweak the model outlined by equation (5.6) (and shown in Figure 5.5) to form a prediction for what effect this should have. Figure 5.8 shows these results and we can immediately see some differences appear. It is perhaps unsurprising that we see damping from the larger beam begin to affect more mid-range values of k since a larger beam will smooth larger perpendicular modes, thus affecting smaller $k_\perp$. However, less intuitive is the difference between the foreground free and foreground contaminated cases. It appears that increasing the beam renders foregrounds less of a problem for the quadrupole and hexadecapole when compared with Figure 5.5. For the monopole, the foregrounds still seem to have a damping effect for the lowest k values.

To understand this we can again analyse the integrands for the multipoles shown in Appendix equations (A.11), (A.12) and (A.13). These are shown in Figure 5.9 but this time I show how the beam term parameterised by σ (outlined in equation (5.7)) affects the integrands as a function of μ. Since the beam damping term is dependent on k, I have chosen a fixed mid-range value ($k = 0.15$) to demonstrate these results. [3]

Figure 5.9 shows that a larger beam damps contributions across all μ values, but it has more of an effect at low-$|\mu|$. It is the modes with low-$|\mu|$ which are most affected by foregrounds and this is why we see apparent mitigation of foreground effects for intensity maps with large beams. It is simply because the beam is damping foreground contaminated modes anyway rendering the foreground contamination a less dominant effect. For lower values of k, it is more likely that there will be smaller $k_\perp$ values which are less beam contaminated. Figure 5.4 shows that high values of μ can exist at these low-$k_\perp$ values where there is much less beam damping and this

[3] As one would expect, we find that larger values of k are affected more by the beam since the beam smooths small perpendicular scales, thus affecting large $k_\perp$ modes. Choosing a very small value for k for the results in Figure 5.9 would show little difference between each different σ case.

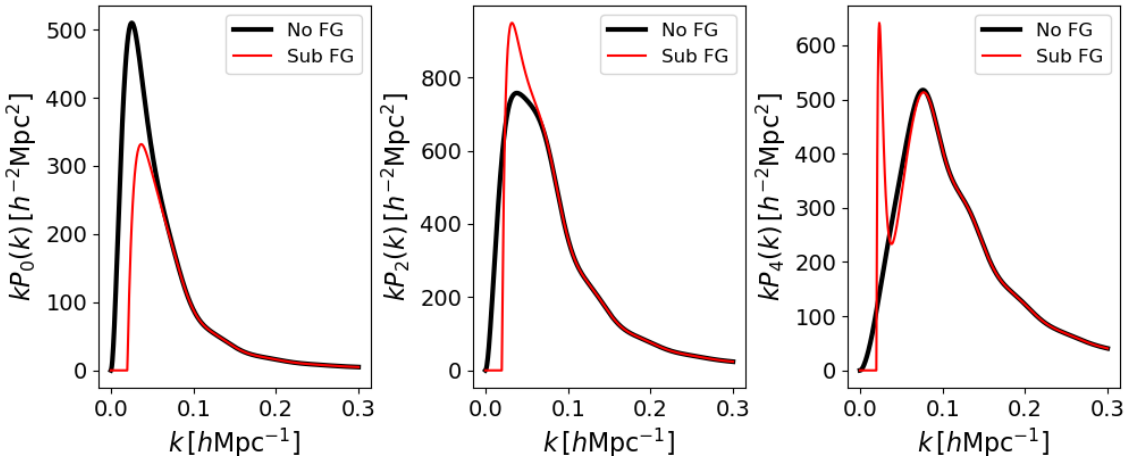

Figure 5.8: Same plot as Figure 5.5 but with an increased beam size of $\theta_{\mathrm{FWHM}} = 2.355$deg. We see more damping here at high-k in comparison with Figure 5.5 as expected from equation (5.7). The larger beam also causes less effects from foregrounds in the quadrupole and hexadecapole in comparison to the smaller beam case of Figure 5.5.

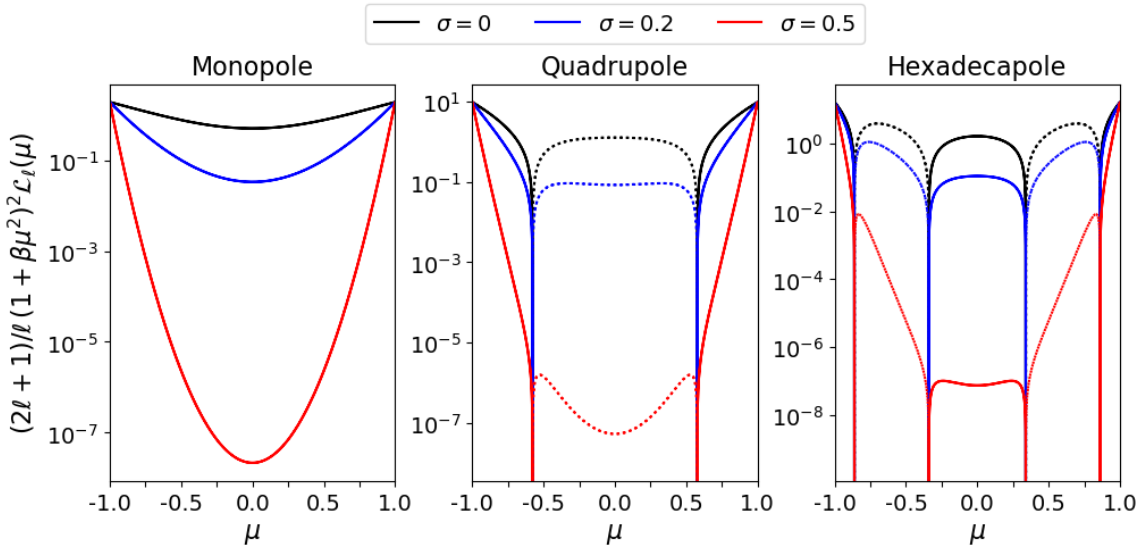

Figure 5.9: Effect of a changing beam size on the multipoles. Similarly to Figure 5.3, this shows the integrands for the expanded multipole equations as a function of μ but now showing the effect of increasing the beam, parameterised by σ as show in equation (5.7). Dotted lines represent negative values. Equations are outlined in Appendix equations (A.11), (A.12) and (A.13). These are results are for a set value of $k = 0.15$.

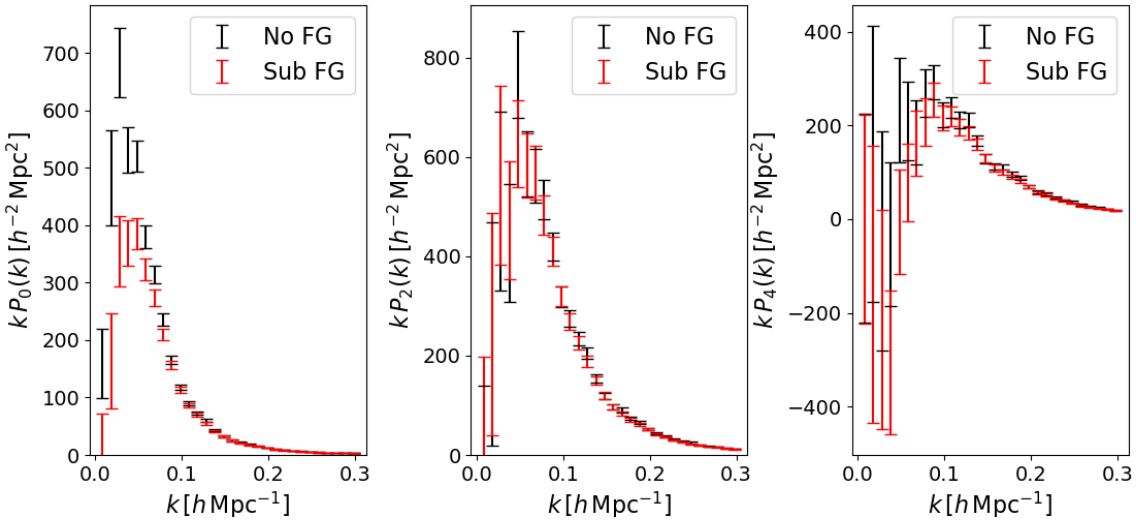

Figure 5.10: Comparison of simulated results of multipole measurements with and without foreground contamination. Same plot as Figure 5.2 but with a larger beam of $\theta_{\mathrm{FWHM}} = 2.355$deg. The extra damping of modes and less impact from foregrounds is something predicted by the model shown by Figure 5.8.

allows foreground effects to dominate. This is why we still see some foreground effects at low-k in Figure 5.8.

I have tested this understanding of the larger beam and its relationship with foreground contamination with our MICE HI intensity map simulations in the extended Blake19 multipole measurement pipeline. Figure 5.10 shows these results. These mostly agree with the model in Figure 5.8 where we only really see the foregrounds have an impact at the smallest k values. At larger k, the large beam damps more power causing it to dominate over the foreground effects and there is little distinction between the foreground free and foreground contaminated cases.

5.2.2 Smaller Sky Analysis

The results outlined in the previous section and in the Blake19 paper are in the context of larger skies ($\gtrsim 1000\mathrm{deg}^2$) where angular pixelization and curved sky geometries need consideration. However, the small-sky regime is likely to be more relevant in the near future for HI intensity mapping, with radio telescopes such as GBT or MeerKAT likely to provide maps on $\sim 100\mathrm{deg}^2$ scales initially; here flat sky approximations can be made. We therefore aim to investigate this and provide a more applicable pipeline to near-future HI intensity mapping/optical cross-correlation surveys.

Figure 5.11 shows multipole measurements with and without foregrounds using MICE maps which have been created to emulate the redshift range of GBT-like intensity maps ($0.6 < z < 1.0$) [230] but with a $100\mathrm{deg}^2$ sky area, which is likely to be achieved by near future observations. We see relatively similar foreground effects for the monopole and hexadecapole as was seen for the larger sky case with the same beam size in Figure 5.2. However, the quadrupole is less conclusive with more increased uncertainty, caused by the smaller sky size.

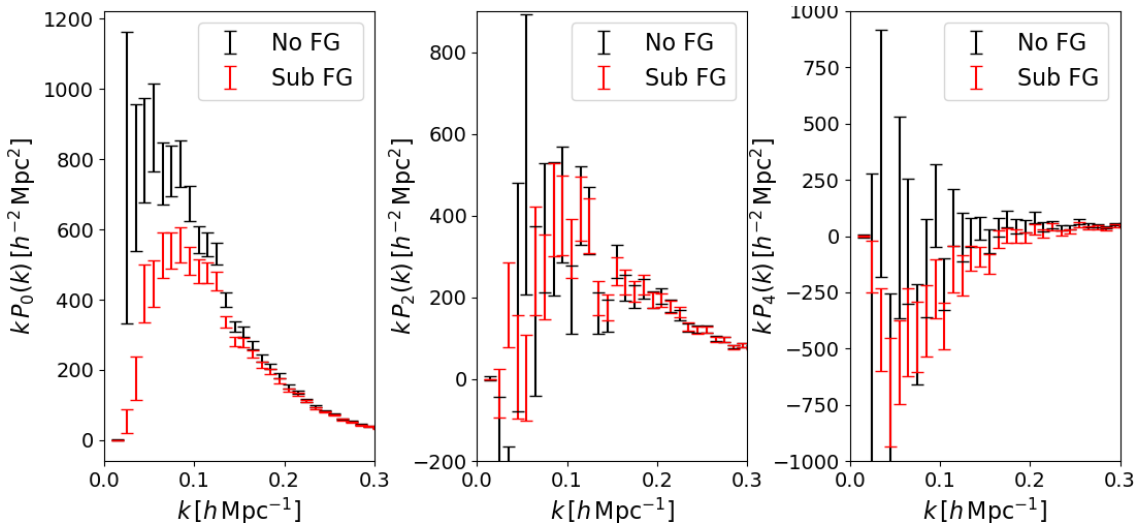

Figure 5.11: Multipole measurements for smaller sky ($100\mathrm{deg}^2$) intensity maps. Done with GBT-like beam ($\theta_{\mathrm{FWHM}} = 0.44\mathrm{deg}$) for redshifts of $0.6 < z < 1.0$. Black data points represent foreground free maps, red represent maps with foregrounds added and cleaned using FASTICA.

The future intention is to conduct more investigation into these smaller sky simulations and attempt to model foreground contamination in conjunction with other observational effects. It is likely that HI intensity maps similar to these I have simulated will be in cross-correlation with optical surveys in the near future and therefore developing a pipeline which theoretically predicts clustering measurements such as the power spectrum multipoles will be hugely beneficial.

CHAPTER 6

THESIS CONCLUSION

The future challenge for cosmology can be broadly summarised. The ΛCDM model accurately explains most data obtained from probes of CMB, BAO, SNeIa, LSS and more. However, the fact that this model relies on a dark sector, not currently included in the standard model of particle physics, means efforts should be dedicated to constraining the available dark sector candidates. The aim should also be to simultaneously confirm that we are on the right track with ΛCDM, or alternatively find tensions which force a deviation from this concordance model.

The focus of this thesis has been on the probe of LSS and there are current (e.g. DES and eBOSS) and future galaxy surveys (e.g. DESI, LSST and Euclid) with this objective. These galaxy surveys can generally be divided into two categories, photometric and spectroscopic redshifts. While spectroscopic surveys provide excellent redshift calibration, and thus good distance constraints, they are time-consuming experiments and not able to produce the galaxy number densities ideally required to maximise statistical precision. Conversely, photometric surveys provide larger data sets but their imaging technique struggles to constrain redshifts, often producing large systematic uncertainty.

A promising alternative is 21cm intensity mapping which has been a central theme of this thesis. Instead of maps of galaxy number densities, this method involves mapping the combined emission from 21cm radiation produced by HI within the galaxies with low angular resolution. The lack of angular precision is deemed acceptable for the purposes of probing LSS, since many of the scales of interest are still accessible. Intensity mapping is still in the development phases and is not yet a competitive cosmological probe. However, assuming the related systematics can be understood and controlled, we should see increasing use of radio data using this technique in cosmology.

Another central theme of this thesis has been cross-correlations between radio intensity maps and optical galaxy redshift surveys. The poor angular resolution yet excellent redshift resolution in intensity maps is inversely complemented by the poor redshift resolution yet

excellent angular resolution of optical surveys. Furthermore, the differing observation strategies between radio and optical mean each experience vastly different systematics. Under a cross-correlation, these differing systematics are uncorrelated and thus mitigated in the analysis. I have demonstrated that not only are these synergies crucial for conducting some of the first cosmological detections using HI intensity mapping but will continue to provide systematic reducing benefits with future datasets e.g. SKA × LSST/Euclid.

Broadly speaking, the main conclusions from this thesis are

- Using a clustering-based redshift estimation technique, HI intensity maps can be used to constrain the uncertainty on photometric redshifts as shown by Figures 3.12 and 4.19. For future imaging surveys such as weak lensing probes, which plan to extend observation depth where constraining redshift will be even more problematic with current methods, a HI clustering redshift technique is appealing.

- 21cm foregrounds contamination is problematic in cross-correlation with a photometric redshift galaxy survey. Because small $k_\parallel$ modes are damped by foreground cleaning and large $k_\parallel$ modes are unconstrained due to the redshift uncertainty from the photometric survey, a combination of these factors risks destroying any cross-correlation signal.

- Corrections can be applied to foreground cleaned intensity mapping data and I have demonstrated examples of these in Chapter 4. With sufficient treatment, foregrounds do not pose an insurmountable problem to the success of HI clustering redshift estimation and HI × photo-z cross-correlations in general.

- Foregrounds also affect the power spectrum multipoles in some interesting ways as I have begun to investigate in Chapter 5. The impact of foregrounds on the multipoles can be accurately modelled using the parameter $\mu = \cos\theta$ where θ is the angle to the LoS. Since foregrounds largely contaminate small $k_\parallel$ modes, the low-μ part of the signal is mostly affected. This provides an explanation of the effects on the multipoles as shown in Figure 5.2.

Future Work

All of the above conclusions warrant further investigation and my future work will aim to provide this. Intensity mapping observations are being conducted now with telescopes such as the GBT, CHIME and MeerKAT and hopefully future observations with HIRAX, BINGO and the SKA should not be far behind. I intend to contribute towards the first intensity mapping observations using MeerKAT both in auto-correlation and also in cross-correlation with optical surveys such as the WiggleZ data. This will require a detailed understanding of 21cm foregrounds and other systematics and potentially incorporating some of the techniques I have outlined into a pipeline which can extract the cosmological signal from this early data. I am also currently working with GBT intensity mapping data which is looking at cross-correlating with eBOSS Luminous Red Galaxies (LRG) and Emission Line Galaxies (ELG) galaxies. This will provide further understanding of the 21cm signal and is allowing work to be done on multipole expansion

of 21cm signal, binning with μ-wedges (both discussed in Chapter 5) and also new techniques such as HI × optical correlation functions in configuration space.

I also believe HI intensity maps will be able to benefit optical surveys in the future too, not just the other way around as has presently been the case with intensity mapping data relying on spectroscopic optical surveys to boost their signal. As I outlined in Chapter 3, arguably the largest challenge facing future photometric surveys such as LSST and Euclid is constraining their redshifts, especially at deeper distances where galaxies become fainter and the challenge is enhanced. I plan to further investigate the benefit intensity maps can have in HI clustering-based redshift estimation with particular attention to Euclid-like data and begin more robust forecasting of the benefits to be gained.

A crucial part of all these investigations lies in the robustness of the simulations used both for calculating uncertainty in current data and providing forecasts for the future. I therefore also plan to build upon my current suite of simulation techniques. However, as I emphasised in previous chapters, simulations that are fit for the purpose of HI × optical investigations ideally need low mass resolutions, over large volumes, with sophisticated halo/galaxy finding algorithms applied with output observables such as photometry and HI content. All of this makes them very computer resource-intensive and naturally some approximative methods are needed such as the ones I have used in this thesis. Detailed analysis of the robustness of these approximations is needed to ensure that the conclusion being drawn from them can be relied upon. I plan to investigate this in future work too.

This thesis has examined the future potential of HI intensity mapping as a probe of large scale cosmic structure, in particular, the benefits that can be gained from cross-correlating these data with conventional optical galaxy redshift surveys. I have shown how 21cm foregrounds present a major challenge to overcome and stimulated the need for further investigation into their effects and techniques for mitigation. But overall this work has shown that the excitement surrounding the potential of HI intensity mapping is warranted and offers an excellent opportunity to accelerate our understanding of the Universe.

APPENDIX A

APPENDICES

A.1 Power Spectrum Multipoles

This provides a complete derivation of the power spectrum multipoles. Firstly, the anisotropic $P(\vec{k})$ spectrum can be expanded in Legendre polynomials

$$P(\vec{k}) \equiv P(k,\mu) = \sum_{\ell} P_{\ell}(k)\mathcal{L}_{\ell}(\mu) \tag{A.1}$$

where $\mathcal{L}_{\ell}(\mu)$ is the ℓ^{th} Legendre polynomial and μ is the cosine of the angle between the wavevector $\vec{k}$ and the LoS. Given that the Legendre polynomials are orthogonal over $(-1,1)$, we have the identity

$$\int_{-1}^{1} \mathcal{L}_{\ell}(\mu)\mathcal{L}_{m}(\mu)\mathrm{d}\mu = \frac{2}{2\ell+1}\delta_{\ell m} \tag{A.2}$$

where $\delta_{\ell m}$ is the Kronecker delta. By multiplying both sides of (A.1), integrating and then rearranging we can derive a general expression for the multipoles given by

$$P_{\ell}(k) = \frac{2\ell+1}{2}\int_{-1}^{1} \mathrm{d}\mu P(k,\mu)\mathcal{L}_{\ell}(\mu). \tag{A.3}$$

For linear RSD we can expand the Kaiser equation in (1.48) to get

$$P(k,\mu) = b^2\left(1+2\beta\mu^2+\beta^2\mu^4\right)P_{\mathrm{M}}(k) \tag{A.4}$$

where $P_{\mathrm{M}}(k)$ is the underlying, isotropic matter power spectrum. The only non-zero multipoles come from $\ell = 0,2,4$ therefore the Legendre polynomials we need are given by

$$\mathcal{L}(\mu)_0 = 1 \tag{A.5}$$

$$\mathcal{L}(\mu)_2 = \frac{3\mu^2-1}{2} \tag{A.6}$$

$$\mathcal{L}(\mu)_4 = \frac{35\mu^4-30\mu^2+3}{8}. \tag{A.7}$$

Using these Legendre polynomials provides the three multipole equations we need as

$$P_{\ell=0}(k) = \frac{1}{2}\int_{-1}^{1} \mathrm{d}\mu P(k,\mu) \tag{A.8}$$

$$P_{\ell=2}(k) = \frac{5}{2}\int_{-1}^{1} \mathrm{d}\mu P(k,\mu)\frac{3\mu^2-1}{2} \tag{A.9}$$

$$P_{\ell=4}(k) = \frac{9}{2}\int_{-1}^{1} \mathrm{d}\mu P(k,\mu)\frac{35\mu^4-30\mu^2+3}{8}\,. \tag{A.10}$$

Plugging in the expanded Kaiser galaxy power spectrum from (A.4) and rearranging gives

$$P_{\ell=0}(k) = b^2 P_{\mathrm{M}}(k)\int_{-1}^{1} \mathrm{d}\mu\left[\frac{1}{2}+\beta\mu^2+\frac{\beta^2\mu^4}{2}\right] \tag{A.11}$$

$$P_{\ell=2}(k) = b^2 P_{\mathrm{M}}(k)\int_{-1}^{1} \mathrm{d}\mu\left[-\frac{5}{4}+\frac{15\mu^2}{4}-\beta\left(\frac{5\mu^2}{2}-\frac{15\mu^4}{2}\right)-\beta^2\left(\frac{5\mu^4}{4}-\frac{15\mu^6}{4}\right)\right] \tag{A.12}$$

$$\begin{aligned} P_{\ell=4}(k) = b^2 P_{\mathrm{M}}(k)\int_{-1}^{1} \mathrm{d}\mu\Bigg[&\frac{27}{16}-\frac{135\mu^2}{8}+\frac{315\mu^4}{16}+\beta\left(\frac{27\mu^2}{8}-\frac{135\mu^4}{4}+\frac{315\mu^6}{8}\right)\\ &+\beta^2\left(\frac{27\mu^4}{16}-\frac{135\mu^6}{8}+\frac{315\mu^8}{16}\right)\Bigg]. \end{aligned} \tag{A.13}$$

Using the integration identity

$$\int_{-1}^{1}\mu^n \mathrm{d}\mu = \frac{2}{n+1} \tag{A.14}$$

gives the results

$$P_{\mathrm{g},\ell=0}(k) = \left(1+\frac{2}{3}\beta+\frac{1}{5}\beta^2\right)b^2 P_{\mathrm{M}}(k) \tag{A.15}$$

$$P_{\mathrm{g},\ell=2}(k) = \left(\frac{4}{3}\beta+\frac{4}{7}\beta^2\right)b^2 P_{\mathrm{M}}(k) \tag{A.16}$$

$$P_{\mathrm{g},\ell=4}(k) = \frac{8}{35}\beta^2 b^2 P_{\mathrm{M}}(k)\,. \tag{A.17}$$

Certificate of Ethics Review

Project Title: Synergies Between Optical and Radio Redshift Surveys for Probing Large Scale Cosmic Structure

Name: Steve Cunnington **User ID:** 834240 **Application Date:** 15-Nov-2018 12:16 **ER Number:** ETHIC-2018-1124

You must download your certificate, print a copy and keep it as a record of this review.

It is your responsibility to adhere to the University Ethics Policy and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers and University Health and Safety Policy.

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

You are reminded that as a University of Portsmouth Researcher you are bound by the UKRIO Code of Practice for Research; any breach of this code could lead to action being taken following the University's Procedure for the Investigation of Allegations of Misconduct in Research.

Any changes in the answers to the questions reflecting the design, management or conduct of the research over the course of the project must be notified to the Faculty Ethics Committee. **Any changes that affect the answers given in the questionnaire, not reported to the Faculty Ethics Committee, will invalidate this certificate.**

This ethical review should not be used to infer any comment on the academic merits or methodology of the project. If you have not already done so, you are advised to develop a clear protocol/proposal and ensure that it is independently reviewed by peers or others of appropriate standing. A favourable ethical opinion should not be perceived as permission to proceed with the research; there might be other matters of governance which require further consideration including the agreement of any organisation hosting the research.

(A1) Please briefly describe your project:: **Using computer simulations of cosmological observables to make forecasts for future telescope surveys.**
(A2) What faculty do you belong to?: **Technology**
(A3) I am sure that my project requires ethical review by my Faculty Ethics Committee because it includes at least one material ethical issue.: **No**
(A5) Has your project already been externally reviewed?: **No**
(B1) Is the study likely to involve human participants?: **No**
(B2) Are you certain that your project will not involve human subjects or participants?: **Yes**
(C6) Is there any risk to the health & safety of the researcher or members of the research team beyond those that have already been risk assessed?: **No**
(D2) Are there risks of damage to physical and/or ecological environmental features?: **No**
(D4) Are there risks of damage to features of historical or cultural heritage (e.g. impacts of study techniques, taking of samples)?: **No**
(E1) Will the study involve the investigator and/or any participants in activities that could be considered contentious, unacceptable, or illegal, or in any other way harmful to the reputation of the University of Portsmouth?: **No**
(E2) Are there any potentially socially or culturally sensitive issues involved? (e.g. sexual, political, legal/criminal or financial): **No**
(F1) Does the project involve animals in any way?: **No**
(F2) Could the research outputs potentially be harmful to third parties?: **No**
(G1) Please confirm that you have read the University Ethics Policy and have considered the implications for your project.: **Confirmed**
(G2) Please confirm that you have read the UK RIO Code of Practice for Research and will conduct your project in accordance with it.: **Confirmed**
(G3) The University is committed to The Concordat to Support Research Integrity.: **Confirmed**
(G4) Submitting false or incorrect information is a breach of the University Ethics Policy and may be considered as misconduct and be subject to disciplinary action. Please confirm you understand this and agree that the information you have entered is correct.: **Confirmed**

FORM UPR16

Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the Research Degrees Operational Handbook for more information

UNIVERSITY OF PORTSMOUTH

Postgraduate Research Student (PGRS) Information				**Student ID:**	834240
PGRS Name:	Steven Cunnington				
Department:	ICG - Tech	**First Supervisor:**	David Bacon		
Start Date:		**01/10/2016**			
Study Mode and Route:	Part-time ☐	MPhil ☐	MD ☐		
	Full-time ☒	PhD ☒	Professional Doctorate ☐		

Title of Thesis:	Synergies Between 21cm & Optical Surveys for Probing Large Scale Cosmic Structure
Thesis Word Count: (excluding ancillary data)	47,527

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a)	Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO ☐
b)	Have all contributions to knowledge been acknowledged?	YES ☒ NO ☐
c)	Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO ☐
d)	Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO ☐
e)	Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO ☐

Candidate Statement:	
I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)	
Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	ETHIC-2018-1124
If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:	
Signed *(PGRS)*: 	**Date: 17/09/19**

UPR16 – April 2018

List of Tables

List of Figures

LIST OF ABBREVIATIONS, CONSTANTS & NOTATIONS

Notation	Description
AAT	Anglo-Australian Telescope
AGN	Active Galactic Nuclei
BAO	Baryon Acoustic Oscillations
BINGO	BAO In Neutral Gas Observations
c	Speed of light $\approx 3.00 \times 10^{8}\mathrm{m\ s^{-1}}$
CHIME	Canadian Hydrogen Intensity Mapping Experiment
CMB	Cosmic Microwave Background
D_{dish}	Radio telescope's dish diameter
DES	Dark Energy Survey
ELG	Emission Line Galaxies
FASTICA	Fast Independent Component Analysis
FLRW	Friedmann-Lemaître-Robertson-Walker metric
GBT	Green Bank Telescope
Gpc	Gigaparsec (unit of distance)
h_{P}	Planck's constant $\approx 6.63 \times 10^{-34}\mathrm{m^2 kg\ s^{-1}}$
HAM	Halo Abundance Matching
HI	Neutral Hydrogen (Pronounced 'H - one')
HIHM	HI-Halo Mass function
HOD	Halo Occupation Distribution
k_{B}	Boltzmann's constant $\approx 1.38 \times 10^{-23}\mathrm{m^2 kg\ s^{-2} K^{-1}}$
LRG	Luminous Red Galaxies
LSS	Large-Scale Structure
LSST	Large Synoptic Survey Telescope
LoS	Line-of-Sight
Mpc	Megaparsec (unit of distance)
QFT	Quantum Field Theory
RSD	Redshift Space Distortions
SKA	Square Kilometre Array
z	Redshift
θ_{FWHM}	Full-Width-Half-Maximum of radio telescope beam

BIBLIOGRAPHY

[1] T. ABBOTT ET AL., *The Dark Energy Survey: more than dark energy – an overview*, Mon. Not. Roy. Astron. Soc., 460 (2016), arXiv, 1601.00329.

[2] T. M. C. ABBOTT ET AL., *Dark Energy Survey year 1 results: Cosmological constraints from galaxy clustering and weak lensing*, Phys. Rev., D98 (2018), arXiv, 1708.01530.

[3] T. M. C. ABBOTT ET AL., *First Cosmology Results using Type Ia Supernovae from the Dark Energy Survey: Constraints on Cosmological Parameters*, Astrophys. J., 872 (2019), arXiv, 1811.02374.

[4] F. B. ABDALLA AND S. RAWLINGS, *Probing dark energy with baryonic oscillations and future radio surveys of neutral hydrogen*, Mon. Not. Roy. Astron. Soc., 360 (2005), arXiv, astro-ph/0411342.

[5] P. A. ABELL ET AL., *LSST Science Book, Version 2.0*, arXiv, 0912.0201.

[6] L. R. ABRAMO AND K. E. LEONARD, *Why multi-tracer surveys beat cosmic variance*, Mon. Not. Roy. Astron. Soc., 432 (2013), arXiv, 1302.5444.

[7] R. ADAM ET AL., *Planck 2015 results. I. Overview of products and scientific results*, Astron. Astrophys., 594 (2016), arXiv, 1502.01582.

[8] K. AKIYAMA ET AL., *First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole*, Astrophys. J., 875 (2019), arXiv, 1906.11238.

[9] Y. AKRAMI ET AL., *Planck 2018 results. I. Overview and the cosmological legacy of Planck*, arXiv, 1807.06205.

[10] S. ALAM ET AL., *The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample*, Mon. Not. Roy. Astron. Soc., 470 (2017), arXiv, 1607.03155.

[11] D. ALONSO, P. BULL, P. G. FERREIRA, AND M. G. SANTOS, *Blind foreground subtraction for intensity mapping experiments*, Mon. Not. Roy. Astron. Soc., 447 (2015), arXiv, 1409.8667.

[12] D. ALONSO AND P. G. FERREIRA, *Constraining ultralarge-scale cosmology with multiple tracers in optical and radio surveys*, Phys. Rev., D92 (2015), arXiv, 1507.03550.

[13] D. ALONSO, P. G. FERREIRA, M. J. JARVIS, AND K. MOODLEY, *Calibrating photometric redshifts with intensity mapping observations*, Phys. Rev., D96 (2017), arXiv, 1704.01941.

[14] D. ALONSO, P. G. FERREIRA, AND M. G. SANTOS, *Fast simulations for intensity mapping experiments*, Mon. Not. Roy. Astron. Soc., 444 (2014), arXiv, 1405.1751.

[15] D. ALONSO, J. SANCHEZ, AND A. SLOSAR, *A unified pseudo-C_ℓ framework*, Mon. Not. Roy. Astron. Soc., 484 (2019), arXiv, 1809.09603.

[16] L. AMENDOLA ET AL., *Cosmology and fundamental physics with the Euclid satellite*, Living Rev. Rel., 21 (2018), arXiv, 1606.00180.

[17] L. AMENDOLA AND S. TSUJIKAWA, *Dark Energy*, Cambridge University Press, 2015.

[18] C. J. ANDERSON ET AL., *Low-amplitude clustering in low-redshift 21-cm intensity maps cross-correlated with 2dF galaxy densities*, Mon. Not. Roy. Astron. Soc., 476 (2018), arXiv, 1710.00424.

[19] B. ASCASO, S. MEI, AND N. BENÍTEZ, *Apples to apples A^2 – I. Realistic galaxy simulated catalogues and photometric redshift predictions for next-generation surveys*, Mon. Not. Roy. Astron. Soc., 453 (2015), arXiv, 1503.01113.

[20] É. AUBOURG ET AL., *Cosmological implications of baryon acoustic oscillation measurements*, Phys. Rev., D92 (2015), arXiv, 1411.1074.

[21] S. AVILA ET AL., *Dark Energy Survey Year 1 Results: galaxy mock catalogues for BAO*, Mon. Not. Roy. Astron. Soc., 479 (2018), arXiv, 1712.06232.

[22] J. W. BAARS AND J. KÄRCHER, HANS, *Radio Telescope Reflectors: Historical Development of Design and Construction*, Springer International Publishing, 2018.

[23] D. J. BACON ET AL., *Cosmology with Phase 1 of the Square Kilometre Array: Red Book 2018: Technical specifications and performance forecasts*, Submitted to: Publ. Astron. Soc. Austral., arXiv, 1811.02743.

[24] D. J. BACON, A. R. REFREGIER, AND R. S. ELLIS, *Detection of weak gravitational lensing by large-scale structure*, Mon. Not. Roy. Astron. Soc., 318 (2000), arXiv, astro-ph/0003008.

[25] J. S. BAGLA, N. KHANDAI, AND K. K. DATTA, *HI as a Probe of the Large Scale Structure in the Post-Reionization Universe*, Mon. Not. Roy. Astron. Soc., 407 (2010), arXiv, 0908.3796.

[26] J. M. BARDEEN, J. R. BOND, N. KAISER, AND A. S. SZALAY, *The Statistics of Peaks of Gaussian Random Fields*, Astrophys. J., 304 (1986), pp. 15–61.

[27] D. G. BARNES ET AL., *The HI Parkes All Sky Survey: southern observations, calibration and robust imaging*, , 322 (2001), pp. 486–498.

[28] R. A. BATTYE, I. W. A. BROWNE, C. DICKINSON, G. HERON, B. MAFFEI, AND A. POURTSIDOU, *HI intensity mapping: a single dish approach*, Mon. Not. Roy. Astron. Soc., 434 (2013), arXiv, 1209.0343.

[29] R. A. BATTYE, R. D. DAVIES, AND J. WELLER, *Neutral hydrogen surveys for high redshift galaxy clusters and proto-clusters*, Mon. Not. Roy. Astron. Soc., 355 (2004), arXiv, astro-ph/0401340.

[30] J. E. BAUTISTA ET AL., *The SDSS-IV extended Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations at redshift of 0.72 with the DR14 Luminous Red Galaxy Sample*, Astrophys. J., 863 (2018), arXiv, 1712.08064.

[31] P. S. BEHROOZI, R. H. WECHSLER, AND H.-Y. WU, *The Rockstar Phase-Space Temporal Halo Finder and the Velocity Offsets of Cluster Cores*, Astrophys. J., 762 (2013), arXiv, 1110.4372.

[32] N. BENITEZ, *Bayesian photometric redshift estimation*, Astrophys. J., 536 (2000), arXiv, astro-ph/9811189.

[33] J. BENJAMIN, L. VAN WAERBEKE, B. MENARD, AND M. KILBINGER, *Photometric redshifts: estimating their contamination and distribution using clustering information*, Mon. Not. Roy. Astron. Soc., 408 (2010), arXiv, 1002.2266.

[34] A. J. BENSON, *Galaxy Formation Theory*, Phys. Rept., 495 (2010), arXiv, 1006.5394.

[35] A. A. BERLIND AND D. H. WEINBERG, *The Halo occupation distribution: Towards an empirical determination of the relation between galaxies and mass*, Astrophys. J., 575 (2002), arXiv, astro-ph/0109001.

[36] M. A. BIGOT-SAZY, C. DICKINSON, R. A. BATTYE, I. W. A. BROWNE, Y. Z. MA, B. MAFFEI, F. NOVIELLO, M. REMAZEILLES, AND P. N. WILKINSON, *Simulations for single-dish intensity mapping experiments*, Mon. Not. Roy. Astron. Soc., 454 (2015), arXiv, 1507.04561.

[37] M.-A. BIGOT-SAZY, Y.-Z. MA, R. A. BATTYE, I. W. A. BROWNE, T. CHEN, C. DICKINSON, S. HARPER, B. MAFFEI, L. C. OLIVARI, AND P. N. WILKINSON, *HI intensity mapping with FAST*, ASP Conf. Ser., 502 (2016), arXiv, 1511.03006.

[38] C. BLAKE, *Power spectrum modelling of galaxy and radio intensity maps including observational effects*, arXiv, 1902.07439.

[39] R. D. BLUM ET AL., *The DECam Legacy Survey*, in American Astronomical Society Meeting Abstracts #228, vol. 228 of American Astronomical Society Meeting Abstracts, June 2016, p. 317.01.

[40] S. BLUNDELL AND K. BLUNDELL, *Concepts in Thermal Physics*, Oxford University Press, 2010.

[41] M. BOLZONELLA, J.-M. MIRALLES, AND R. PELLO', *Photometric redshifts based on standard SED fitting procedures*, Astron. Astrophys., 363 (2000), arXiv, astro-ph/0003380.

[42] J. D. BOWMAN, A. E. E. ROGERS, R. A. MONSALVE, T. J. MOZDZEN, AND N. MAHESH, *An absorption profile centred at 78 megahertz in the sky-averaged spectrum*, Nature, 555 (2018), arXiv, 1810.05912.

[43] P. BULL, P. G. FERREIRA, P. PATEL, AND M. G. SANTOS, *Late-time cosmology with 21cm intensity mapping experiments*, Astrophys. J., 803 (2015), arXiv, 1405.1452.

[44] J. CARRETERO, F. J. CASTANDER, E. GAZTANAGA, M. CROCCE, AND P. FOSALBA, *An algorithm to build mock galaxy catalogues using mice simulations*, MNRAS, 447,646 (2015).

[45] E. CASTORINA AND F. VILLAESCUSA-NAVARRO, *On the spatial distribution of neutral hydrogen in the Universe: bias and shot-noise of the HI power spectrum*, Mon. Not. Roy. Astron. Soc., 471 (2017), arXiv, 1609.05157.

[46] T.-C. CHANG, U.-L. PEN, K. BANDURA, AND J. B. PETERSON, *Hydrogen 21-cm Intensity Mapping at redshift 0.8*, Nature, 466 (2010), arXiv, 1007.3709.

[47] T.-C. CHANG, U.-L. PEN, J. B. PETERSON, AND P. MCDONALD, *Baryon Acoustic Oscillation Intensity Mapping as a Test of Dark Energy*, Phys. Rev. Lett., 100 (2008), arXiv, 0709.3672.

[48] E. CHAPMAN, F. B. ABDALLA, G. HARKER, V. JELIC, P. LABROPOULOS, S. ZAROUBI, M. A. BRENTJENS, A. G. DE BRUYN, AND L. V. E. KOOPMANS, *Foreground Removal using FastICA: A Showcase of LOFAR-EoR*, Mon. Not. Roy. Astron. Soc., 423 (2012), arXiv, 1201.2190.

[49] E. CHAPMAN ET AL., *The Scale of the Problem : Recovering Images of Reionization with GMCA*, Mon. Not. Roy. Astron. Soc., 429 (2013), arXiv, 1209.4769.

[50] T. CHEN, R. A. BATTYE, A. A. COSTA, C. DICKINSON, AND S. E. HARPER, *Impact of* $1/f$ *noise on cosmological parameter constraints for SKA intensity mapping*, arXiv, 1907.12132.

[51] X. CHEN, *The tianlai project: A 21cm cosmology experiment*, International Journal of Modern Physics: Conference Series, 12 (2012), https://doi.org/10.1142/S2010194512006459.

[52] J. N. CHENGALUR AND N. KANEKAR, *Implications of 21cm observations for damped ly-alpha systems*, Mon. Not. Roy. Astron. Soc., 318 (2000), arXiv, astro-ph/0011540.

[53] S. COLE, S. HATTON, D. H. WEINBERG, AND C. S. FRENK, *Mock 2dF and SDSS galaxy redshift surveys*, Mon. Not. Roy. Astron. Soc., 300 (1998), arXiv, astro-ph/9801250.

[54] S. COLE, C. G. LACEY, C. M. BAUGH, AND C. S. FRENK, *Hierarchical galaxy formation*, Mon. Not. Roy. Astron. Soc., 319 (2000), arXiv, astro-ph/0007281.

[55] M. COLLESS, *First results from the 2dF galaxy redshift survey*, Phil. Trans. Roy. Soc. Lond., A357 (1999), arXiv, astro-ph/9804079.

[56] J. CONDON AND S. RANSOM, *Essential Radio Astronomy*, Princeton Series in Modern Observational Astronomy, Princeton University Press, 2016.

[57] N. H. M. CRIGHTON ET AL., *The neutral hydrogen cosmological mass density at z = 5*, Mon. Not. Roy. Astron. Soc., 452 (2015), arXiv, 1506.02037.

[58] M. CROCCE, F. J. CASTANDER, E. GAZTANAGA, P. FOSALBA, AND J. CARRETERO, *The MICE Grand Challenge lightcone simulation – II. Halo and galaxy catalogues*, Mon. Not. Roy. Astron. Soc., 453 (2015), arXiv, 1312.2013.

[59] S. CUNNINGTON, I. HARRISON, A. POURTSIDOU, AND D. BACON, *HI intensity mapping for clustering-based redshift estimation*, Mon. Not. Roy. Astron. Soc., 482 (2019), arXiv, 1805.04498.

[60] S. CUNNINGTON, L. WOLZ, A. POURTSIDOU, AND D. BACON, *Impact of foregrounds on HI intensity mapping cross-correlations with optical surveys*, Mon. Not. Roy. Astron. Soc., 488 (2019), arXiv, 1904.01479.

[61] C. DAVIS ET AL., *Dark Energy Survey Year 1 Results: Cross-Correlation Redshifts in the DES – Calibration of the Weak Lensing Source Redshift Distributions*, Submitted to: Mon. Not. Roy. Astron. Soc., arXiv, 1710.02517.

[62] M. DAVIS, J. A. NEWMAN, S. M. FABER, AND A. C. PHILLIPS, *The DEEP2 Redshift Survey*, in Proceedings, Workshop on Deep Fields: Garching, Germany, October 9-12, 2000, arXiv, astro-ph/0012189.

[63] P. DI BARI, *Cosmology and the Early Universe - PHYS6005 Lecture Notes*, University of Southampton, (2015).

[64] R. H. DICKE, P. J. E. PEEBLES, P. G. ROLL, AND D. T. WILKINSON, *Cosmic Black-Body Radiation.*, , 142 (1965), pp. 414–419.

[65] P. A. M. DIRAC, *The Quantum Theory of the Electron*, Proceedings of the Royal Society of London Series A, 117 (1928), pp. 610–624.

[66] S. DODELSON, *Modern cosmology*, Academic Press, 2003.

[67] S. DODELSON AND L. M. WIDROW, *Sterile-neutrinos as dark matter*, Phys. Rev. Lett., 72 (1994), arXiv, hep-ph/9303287.

[68] B. T. DRAINE, *Physics of the Interstellar and Intergalactic Medium*, Princeton University Press, 2011.

[69] M. J. DRINKWATER ET AL., *The WiggleZ Dark Energy Survey: Survey Design and First Data Release*, Mon. Not. Roy. Astron. Soc., 401 (2010), arXiv, 0911.4246.

[70] A. EINSTEIN, *Zur Elektrodynamik bewegter Körper*, Annalen der Physik, 322 (1905), pp. 891–921.

[71] A. EINSTEIN, *Die Grundlage der allgemeinen Relativitätstheorie*, Annalen der Physik, 354 (1916), pp. 769–822.

[72] S. W. ELLINGSON, *Antennas in Radio Telescope Systems*, Springer Singapore, Singapore, 2014, pp. 1–21.

[73] A. FALTENBACHER AND S. D. M. WHITE, *Assembly bias and the dynamical structure of dark matter halos*, Astrophys. J., 708 (2010), arXiv, 0909.4302.

[74] A. FERNANDEZ-SOTO, K. M. LANZETTA, H.-W. CHEN, S. M. PASCARELLE, AND N. YAHATA, *On the compared accuracy and reliability of spectroscopic and photometric redshift measurements*, Astrophys. J. Suppl., 135 (2001), arXiv, astro-ph/0007447.

[75] R. P. FEYNMAN AND J. VERNON, F. L., *The theory of a general quantum system interacting with a linear dissipative system*, Annals of Physics, 24 (1963), pp. 118–173.

[76] G. B. FIELD, *Excitation of the Hydrogen 21-CM Line*, Proceedings of the IRE, 46 (1958), pp. 240–250.

[77] L. H. FORD, *Quantum vacuum energy in general relativity*, , 11 (1975), pp. 3370–3377.

[78] P. FOSALBA, M. CROCCE, E. GAZTAÑAGA, AND F. J. CASTANDER, *The MICE grand challenge lightcone simulation – I. Dark matter clustering*, Mon. Not. Roy. Astron. Soc., 448 (2015), arXiv, 1312.1707.

[79] P. FOSALBA, E. GAZTAÑAGA, F. J. CASTANDER, AND M. CROCCE, *The MICE Grand Challenge light-cone simulation – III. Galaxy lensing mocks from all-sky lensing maps*, Mon. Not. Roy. Astron. Soc., 447 (2015), arXiv, 1312.2947.

[80] W. L. FREEDMAN, *Cosmology at a Crossroads*, Nat. Astron., 1 (2017), arXiv, 1706.02739.

[81] C. S. FRENK AND S. D. M. WHITE, *Dark matter and cosmic structure*, Annalen Phys., 524 (2012), arXiv, 1210.0544.

[82] W. FREUDLING, L. STAVELEY-SMITH, B. CATINELLA, R. MINCHIN, M. CALABRETTA, E. MOMJIAN, M. ZWAAN, M. MEYER, AND K. O'NEIL, *Deep 21-cm HI Observations at z 0.1: The Precursor to the Arecibo Ultra Deep Survey*, Astrophys. J., 727 (2011), arXiv, 1011.0877.

[83] A. FRIEDMANN, *Über die Krümmung des Raumes*, Zeitschrift fur Physik, 10 (1922), pp. 377–386.

[84] A. FRIEDMANN, *Über die Möglichkeit einer Welt mit konstanter negativer Krümmung des Raumes*, Zeitschrift fur Physik, 21 (1924), pp. 326–332.

[85] M. GATTI ET AL., *Dark Energy Survey Year 1 Results: Cross-Correlation Redshifts – Methods and Systematics Characterization*, Mon. Not. Roy. Astron. Soc., 477 (2018), arXiv, 1709.00992.

[86] H. GIL-MARIN, C. WAGNER, L. VERDE, R. JIMENEZ, AND A. F. HEAVENS, *Reducing sample variance: halo biasing, non-linearity and stochasticity*, Mon. Not. Roy. Astron. Soc., 407 (2010), arXiv, 1003.3238.

[87] K. M. GORSKI, E. HIVON, A. J. BANDAY, B. D. WANDELT, F. K. HANSEN, M. REINECKE, AND M. BARTELMAN, *HEALPix - A Framework for high resolution discretization, and fast analysis of data distributed on the sphere*, Astrophys. J., 622 (2005), arXiv, astro-ph/0409513.

[88] Q. GUO, S. WHITE, C. LI, AND M. BOYLAN-KOLCHIN, *How do galaxies populate Dark Matter halos?*, Mon. Not. Roy. Astron. Soc., 404 (2010), arXiv, 0909.4305.

[89] A. H. GUTH, *The Inflationary Universe: A Possible Solution to the Horizon and Flatness Problems*, Phys. Rev., D23 (1981), pp. 347–356.
[Adv. Ser. Astrophys. Cosmol.3,139(1987)].

[90] L. GUZZO ET AL., *A test of the nature of cosmic acceleration using galaxy redshift distortions*, Nature, 451 (2008), arXiv, 0802.1944.

[91] A. J. S. HAMILTON, *Linear redshift distortions: A Review*, in Ringberg Workshop on Large Scale Structure Ringberg, Germany, September 23-28, 1996, arXiv, astro-ph/9708102.

[92] N. HAND, Y. FENG, F. BEUTLER, Y. LI, C. MODI, U. SELJAK, AND Z. SLEPIAN, *nbodykit: an open-source, massively parallel toolkit for large-scale structure*, Astron. J., 156 (2018), arXiv, 1712.05834.

[93] S. HARPER AND C. DICKINSON, *Potential impact of global navigation satellite services on total power HI intensity mapping surveys*, Mon. Not. Roy. Astron. Soc., 479 (2018), arXiv, 1803.06314.

[94] S. HARPER, C. DICKINSON, R. BATTYE, S. ROYCHOWDHURY, I. BROWNE, Y.-Z. MA, L. OLIVARI, AND T. CHEN, *Impact of Simulated 1/f Noise for HI Intensity Mapping Experiments*, Mon. Not. Roy. Astron. Soc., 478 (2018), arXiv, 1711.07843.

[95] S. E. HARPER, *Simulation of Systematics in Future Single-Dish HI Intensity Mapping Experiments*, arXiv, 1805.06835.

[96] I. HARRISON, S. CAMERA, J. ZUNTZ, AND M. L. BROWN, *SKA weak lensing – I. Cosmological forecasts and the power of radio-optical cross-correlations*, Mon. Not. Roy. Astron. Soc., 463 (2016), arXiv, 1601.03947.

[97] C. G. T. HASLAM, C. J. SALTER, H. STOFFEL, AND W. E. WILSON, *A 408 MHz all-sky continuum survey. II - The atlas of contour maps*, , 47 (1982), p. 1.

[98] M. P. HAYNES ET AL., *The Arecibo Legacy Fast ALFA Survey: The alpha.40 HI Source Catalog, its Characteristics and their Impact on the Derivation of the HI Mass Function*, Astron. J., 142 (2011), arXiv, 1109.0027.

[99] M. P. HAYNES, D. E. HOGG, R. J. MADDALENA, M. S. ROBERTS, AND L. VAN ZEE, *Asymmetry in high-precision global H I profiles of isolated spiral galaxies*, , 115 (1998), p. 62.

[100] G. HINSHAW ET AL., *Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results*, Astrophys. J. Suppl., 208 (2013), arXiv, 1212.5226.

[101] M. HIRSCHMANN, G. DE LUCIA, AND F. FONTANOT, *Galaxy assembly, stellar feedback and metal enrichment: the view from the GAEA model*, Mon. Not. Roy. Astron. Soc., 461 (2016), arXiv, 1512.04531.

[102] K. HOFFMANN, J. BEL, E. GAZTANAGA, M. CROCCE, P. FOSALBA, AND F. J. CASTANDER, *Measuring the growth of matter fluctuations with third-order galaxy correlations*, MNRAS, 447,1724 (2015).

[103] D. W. HOGG, *Distance measures in cosmology*, arXiv, astro-ph/9905116.

[104] D. W. HOGG, D. J. EISENSTEIN, M. R. BLANTON, N. A. BAHCALL, J. BRINKMANN, J. E. GUNN, AND D. P. SCHNEIDER, *Cosmic homogeneity demonstrated with luminous red galaxies*, Astrophys. J., 624 (2005), arXiv, astro-ph/0411197.

[105] W. HU AND S. DODELSON, *Cosmic microwave background anisotropies*, Ann. Rev. Astron. Astrophys., 40 (2002), arXiv, astro-ph/0110414.

[106] E. HUBBLE, *A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae*, Proceedings of the National Academy of Science, 15 (1929), pp. 168–173.

[107] T. R. HUNTER AND P. J. NAPIER, *Antennas and Receivers in Radio Interferometry*, arXiv e-prints, (2016), arXiv, 1609.09376.

[108] A. HYVÄRINEN, *Fast and robust fixed-point algorithms for independent component analysis*, IEEE transactions on neural networks, 10 3 (1999), pp. 626–34.

[109] J. C. JACKSON, *Fingers of God: A critique of Rees' theory of primoridal gravitational radiation*, Mon. Not. Roy. Astron. Soc., 156 (1972), arXiv, 0810.3908.

[110] V. JELIC ET AL., *Foreground simulations for the LOFAR - Epoch of Reionization Experiment*, Mon. Not. Roy. Astron. Soc., 389 (2008), arXiv, 0804.1130.

[111] V. JELIC, S. ZAROUBI, P. LABROPOULOS, G. BERNARDI, A. G. DE BRUYN, AND L. V. E. KOOPMANS, *Realistic Simulations of the Galactic Polarized Foreground: Consequences for 21-cm Reionization Detection Experiments*, Mon. Not. Roy. Astron. Soc., 409 (2010), arXiv, 1007.4135.

[112] S. JOARDAR AND J. CLAYCOMB, *Radio Astronomy: An Introduction*, Mercury Learning and Information, 2016.

[113] D. H. JONES, W. SAUNDERS, M. READ, AND M. COLLESS, *Second data release of the 6dF Galaxy Survey*, Publ. Astron. Soc. Austral., 22 (2005), arXiv, astro-ph/0505068.

[114] N. KAISER, *On the spatial correlations of Abell clusters.*, , 284 (1984), pp. L9–L12.

[115] N. KAISER, *Clustering in real space and in redshift space*, Mon. Not. Roy. Astron. Soc., 227 (1987), pp. 1–27.

[116] N. KANEKAR, J. X. PROCHASKA, S. L. ELLISON, AND J. N. CHENGALUR, *A search for HI 21cm absorption in strong MgII absorbers in the redshift desert*, Mon. Not. Roy. Astron. Soc., 396 (2009), arXiv, 0903.4487.

[117] T. D. KITCHING, D. BACON, M. L. BROWN, P. BULL, J. D. MCEWEN, M. OGURI, R. SCARAMELLA, K. TAKAHASHI, K. WU, AND D. YAMAUCHI, *Euclid & SKA Synergies*, arXiv, 1501.03978.

[118] E. D. KOVETZ ET AL., *Line-Intensity Mapping: 2017 Status Report*, arXiv, 1709.09066.

[119] F. KÖHLINGER, H. HOEKSTRA, AND M. ERIKSEN, *Statistical uncertainties and systematic errors in weak lensing mass estimates of galaxy clusters*, Mon. Not. Roy. Astron. Soc., 453 (2015), arXiv, 1508.05308.

[120] O. LAHAV, P. B. LILJE, J. R. PRIMACK, AND M. J. REES, *Dynamical effects of the cosmological constant*, Mon. Not. Roy. Astron. Soc., 251 (1991), pp. 128–136.

[121] A. LEAUTHAUD ET AL., *New constraints on the evolution of the stellar-to-dark matter connection: a combined analysis of galaxy-galaxy lensing, clustering, and stellar mass functions from z=0.2 to z=1*, Astrophys. J., 744 (2012), arXiv, 1104.0928.

[122] G. LEMAÎTRE, *Expansion of the universe, A homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae*, , 91 (1931), pp. 483–490.

[123] G. LEMAÎTRE, *The Beginning of the World from the Point of View of Quantum Theory.*, , 127 (1931), p. 706.

[124] J. LESGOURGUES, *The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview*, arXiv, 1104.2932.

[125] A. LEWIS, A. CHALLINOR, AND A. LASENBY, *Efficient computation of CMB anisotropies in closed FRW models*, Astrophys. J., 538 (2000), arXiv, astro-ph/9911177.

[126] A. R. LIDDLE, *An introduction to modern cosmology*, John Wiley and Sons Ltd, 1998.

[127] D. N. LIMBER, *The Analysis of Counts of the Extragalactic Nebulae in Terms of a Fluctuating Density Field. II*, Astrophys. J., 119 (1954), p. 655.

[128] A. D. LINDE, *Inflation, quantum cosmology and the anthropic principle*, in Science and ultimate reality: Quantum theory, cosmology, and complexity, 2002, arXiv, hep-th/0211048.

[129] N. MACCRANN ET AL., *DES Y1 Results: Validating cosmological parameter estimation using simulated Dark Energy Surveys*, arXiv, 1803.09795.

[130] R. MANDELBAUM, *Weak lensing for precision cosmology*, Ann. Rev. Astron. Astrophys., 56 (2018), arXiv, 1710.03235.

[131] M. MANERA ET AL., *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: a large sample of mock galaxy catalogues*, Mon. Not. Roy. Astron. Soc., 428 (2012), arXiv, 1203.6609.

[132] Y.-Y. MAO, A. R. ZENTNER, AND R. H. WECHSLER, *Beyond Assembly Bias: Exploring Secondary Halo Biases for Cluster-size Haloes*, Mon. Not. Roy. Astron. Soc., 474 (2018), arXiv, 1705.03888.

[133] P. MARSHALL ET AL., *Science-Driven Optimization of the LSST Observing Strategy*, arXiv, 1708.04058.

[134] A. M. MARTIN, E. PAPASTERGIS, R. GIOVANELLI, M. P. HAYNES, C. M. SPRINGOB, AND S. STIERWALT, *The Arecibo Legacy Fast ALFA Survey: X. The HI Mass Function and Ω_{HI} From the 40% ALFALFA Survey*, Astrophys. J., 723 (2010), arXiv, 1008.5107.

[135] K. W. MASUI ET AL., *Measurement of 21 cm brightness fluctuations at z 0.8 in cross-correlation*, Astrophys. J., 763 (2013), arXiv, 1208.0331.

[136] D. J. MATTHEWS AND J. A. NEWMAN, *Reconstructing Redshift Distributions with Cross-Correlations: Tests and an Optimized Recipe*, Astrophys. J., 721 (2010), arXiv, 1003.0687.

[137] M. MCQUINN, O. ZAHN, M. ZALDARRIAGA, L. HERNQUIST, AND S. R. FURLANETTO, *Cosmological parameter estimation using 21 cm radiation from the epoch of reionization*, Astrophys. J., 653 (2006), arXiv, astro-ph/0512263.

[138] A. I. MERSON ET AL., *Lightcone mock catalogues from semi-analytic models of galaxy formation - I. Construction and application to the BzK colour selection*, Mon. Not. Roy. Astron. Soc., 429 (2013), arXiv, 1206.4049.

[139] H. J. MO AND S. D. M. WHITE, *An Analytic model for the spatial clustering of dark matter halos*, Mon. Not. Roy. Astron. Soc., 282 (1996), arXiv, astro-ph/9512127.

[140] D. F. MOORE, J. E. AGUIRRE, A. R. PARSONS, D. C. JACOBS, AND J. C. POBER, *The Effects of Polarized Foregrounds on 21cm Epoch of Reionization Power Spectrum Measurements*, Astrophys. J., 769 (2013), arXiv, 1302.0876.

[141] T. MOORE, *A General Relativity Workbook*, University Science Books, 2012.

[142] B. P. MOSTER, R. S. SOMERVILLE, C. MAULBETSCH, F. C. V. D. BOSCH, A. V. MACCIO', T. NAAB, AND L. OSER, *Constraints on the relationship between stellar mass and halo mass at low and high redshift*, Astrophys. J., 710 (2010), arXiv, 0903.4682.

[143] B. MÉNARD, R. SCRANTON, S. SCHMIDT, C. MORRISON, D. JEONG, T. BUDAVARI, AND M. RAHMAN, *Clustering-based redshift estimation: method and application to data*, arXiv, 1303.4722.

[144] E. MÖRTSELL AND S. DHAWAN, *Does the Hubble constant tension call for new physics?*, JCAP, 1809 (2018), arXiv, 1801.07260.

[145] L. B. NEWBURGH ET AL., *Calibrating CHIME, A New Radio Interferometer to Probe Dark Energy*, Proc. SPIE Int. Soc. Opt. Eng., 9145 (2014), arXiv, 1406.2267.

[146] L. B. NEWBURGH ET AL., *HIRAX: A Probe of Dark Energy and Radio Transients*, Proc. SPIE Int. Soc. Opt. Eng., 9906 (2016), arXiv, 1607.02059.

[147] J. A. NEWMAN, *Calibrating Redshift Distributions Beyond Spectroscopic Limits with Cross-Correlations*, Astrophys. J., 684 (2008), arXiv, 0805.1409.

[148] P. NOTERDAEME ET AL., *Column density distribution and cosmological mass density of neutral gas: Sloan Digital Sky Survey-III Data Release 9*, Astron. Astrophys., 547 (2012), arXiv, 1210.1213.

[149] P. NTELIS ET AL., *Exploring cosmic homogeneity with the BOSS DR12 galaxy sample*, JCAP, 1706 (2017), arXiv, 1702.02159.

[150] D. OBRESCHKOW, D. CROTON, G. DE LUCIA, S. KHOCHFAR, AND S. RAWLINGS, *Simulation of the Cosmic Evolution of Atomic and Molecular Hydrogen in Galaxies*, Astrophys. J., 698 (2009), arXiv, 0904.2221.

[151] D. OBRESCHKOW AND S. RAWLINGS, *Understanding the H2/HI Ratio in Galaxies*, Mon. Not. Roy. Astron. Soc., 394 (2009), arXiv, 0901.2526.

[152] A. OBULJEN, D. ALONSO, F. VILLAESCUSA-NAVARRO, I. YOON, AND M. JONES, *The HI content of dark matter halos at $z \approx 0$ from ALFALFA*, arXiv, 1805.00934.

[153] A. OBULJEN, E. CASTORINA, F. VILLAESCUSA-NAVARRO, AND M. VIEL, *High-redshift post-reionization cosmology with 21cm intensity mapping*, JCAP, 1805 (2018), arXiv, 1709.07893.

[154] A. OBULJEN, N. DALAL, AND W. J. PERCIVAL, *Anisotropic halo assembly bias and redshift-space distortions*, arXiv, 1906.11823.

[155] H. PADMANABHAN, T. R. CHOUDHURY, AND A. REFREGIER, *Theoretical and observational constraints on the HI intensity power spectrum*, Mon. Not. Roy. Astron. Soc., 447 (2015), arXiv, 1407.6366.

[156] H. PADMANABHAN AND G. KULKARNI, *Constraints on the evolution of the relationship between H i mass and halo mass in the last 12 Gyr*, Mon. Not. Roy. Astron. Soc., 470 (2017), arXiv, 1608.00007.

[157] H. PADMANABHAN, A. REFREGIER, AND A. AMARA, *A halo model for cosmological neutral hydrogen : abundances and clustering H i abundances and clustering*, Mon. Not. Roy. Astron. Soc., 469 (2017), arXiv, 1611.06235.

[158] J. A. PEACOCK, *Cosmological physics*, Cambridge University Press, 1999.

[159] J. A. PEACOCK AND R. E. SMITH, *Halo occupation numbers and galaxy bias*, Mon. Not. Roy. Astron. Soc., 318 (2000), arXiv, astro-ph/0005010.

[160] P. J. E. PEEBLES, *The Large-Scale Structure of the Universe / by P. J. E. Peebles*, Princeton University Press Princeton, N.J, 1980.

[161] P. J. E. PEEBLES AND B. RATRA, *The Cosmological constant and dark energy*, Rev. Mod. Phys., 75 (2003), arXiv, astro-ph/0207347.

[162] U.-L. PEN, L. STAVELEY-SMITH, J. PETERSON, AND T.-C. CHANG, *First Detection of Cosmic Structure in the 21-cm Intensity Field*, Mon. Not. Roy. Astron. Soc., 394 (2009), arXiv, 0802.3239.

[163] A. A. PENZIAS AND R. W. WILSON, *A Measurement of Excess Antenna Temperature at 4080 Mc/s.*, , 142 (1965), pp. 419–421.

[164] W. J. PERCIVAL ET AL., *The 2dF Galaxy Redshift Survey: The Power spectrum and the matter content of the Universe*, Mon. Not. Roy. Astron. Soc., 327 (2001), arXiv, astro-ph/0105252.

[165] S. PERLMUTTER ET AL., *Measurements of Omega and Lambda from 42 high redshift supernovae*, Astrophys. J., 517 (1999), arXiv, astro-ph/9812133.

[166] C. PEROUX, M. DESSAUGES-ZAVADSKY, S. D'ODORICO, T. S. KIM, AND R. G. MCMAHON, *A Homogeneous sample of sub-DLAs. 3. Total gas mass Omega(HI+HeII) at z>2*, Mon. Not. Roy. Astron. Soc., 363 (2005), arXiv, astro-ph/0507353.

[167] J. B. PETERSON ET AL., *21 cm Intensity Mapping*, arXiv, 0902.3091.

[168] A. POPPING, R. DAVE, R. BRAUN, AND B. D. OPPENHEIMER, *The Simulated HI Sky at low redshift*, Astron. Astrophys., 504 (2009), arXiv, 0906.3067.

[169] A. POURTSIDOU, *HI Intensity Mapping with MeerKAT*, PoS, MeerKAT2016 (2018), arXiv, 1709.07316.

[170] A. POURTSIDOU, D. BACON, AND R. CRITTENDEN, *HI and cosmological constraints from intensity mapping, optical and CMB surveys*, Mon. Not. Roy. Astron. Soc., 470 (2017), arXiv, 1610.04189.

[171] A. POURTSIDOU, D. BACON, R. CRITTENDEN, AND R. B. METCALF, *Prospects for clustering and lensing measurements with forthcoming intensity mapping and optical surveys*, Mon. Not. Roy. Astron. Soc., 459 (2016), arXiv, 1509.03286.

[172] J. R. PRITCHARD AND A. LOEB, *Constraining the unexplored period between the dark ages and reionization with observations of the global 21 cm signal*, Phys. Rev., D82 (2010), arXiv, 1005.4057.

[173] M. RAHMAN, A. J. MENDEZ, B. MÉNARD, R. SCRANTON, S. J. SCHMIDT, C. B. MORRISON, AND T. BUDAVÁRI, *Exploring the SDSS Photometric Galaxies with Clustering Redshifts*, Mon. Not. Roy. Astron. Soc., 460 (2016), arXiv, 1512.03057.

[174] A. RAHMATI, A. P. PAWLIK, M. RAICEVIC, AND J. SCHAYE, *On the evolution of the HI column density distribution in cosmological simulations*, Mon. Not. Roy. Astron. Soc., 430 (2013), arXiv, 1210.7808.

[175] S. M. RAO, D. A. TURNSHEK, AND D. NESTOR, *Damped lyman alpha systems at z<1.65: the expanded sdss hst sample*, Astrophys. J., 636 (2006), arXiv, astro-ph/0509469.

[176] M. REES, *Just Six Numbers*, SCIENCE MASTERS, Orion, 2014.

[177] M. REMAZEILLES, J. DELABROUILLE, AND J.-F. CARDOSO, *Foreground component separation with generalised ILC*, Mon. Not. Roy. Astron. Soc., 418 (2011), arXiv, 1103.1166.

[178] J. RHEE, M. A. ZWAAN, F. H. BRIGGS, J. N. CHENGALUR, P. LAH, T. OOSTERLOO, AND T. VAN DER HULST, *Neutral atomic hydrogen (HI) gas evolution in field galaxies at z 0.1 and 0.2*, Mon. Not. Roy. Astron. Soc., 435 (2013), arXiv, 1308.1462.

[179] A. G. RIESS ET AL., *Observational evidence from supernovae for an accelerating universe and a cosmological constant*, Astron. J., 116 (1998), arXiv, astro-ph/9805201.

[180] A. G. RIESS, W. H. PRESS, AND R. P. KIRSHNER, *A Precise Distance Indicator: Type IA Supernova Multicolor Light-Curve Shapes*, , 473 (1996), arXiv, astro-ph/9604143.

[181] H. P. ROBERTSON, *Relativistic Cosmology*, Reviews of Modern Physics, 5 (1933), pp. 62–90.

[182] A. J. ROSS ET AL., *Ameliorating Systematic Uncertainties in the Angular Clustering of Galaxies: A Study using SDSS-III*, Mon. Not. Roy. Astron. Soc., 417 (2011), arXiv, 1105.2320.

[183] V. C. RUBIN AND W. K. FORD, JR., *Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions*, Astrophys. J., 159 (1970), pp. 379–403.

[184] B. RYDEN, *Introduction to Cosmology*, Addison-Wesley, 2003.

[185] I. SADEH, F. B. ABDALLA, AND O. LAHAV, *ANNz2 - photometric redshift and probability distribution function estimation using machine learning*, Publ. Astron. Soc. Pac., 128 (2016), arXiv, 1507.00490.

[186] S. SAITO, *Galaxy Clustering in Redshift Space*, Lecture Series on Cosmology: MPA Garching, (2016).

[187] M. G. SANTOS, D. ALONSO, P. BULL, M. SILVA, AND S. YAHYA, *HI galaxy simulations for the SKA: number counts and bias*, arXiv, 1501.03990.

[188] M. G. SANTOS, A. COORAY, AND L. KNOX, *Multifrequency analysis of 21 cm fluctuations from the era of reionization*, Astrophys. J., 625 (2005), arXiv, astro-ph/0408515.

[189] M. G. SANTOS ET AL., *Cosmology with a SKA HI intensity mapping survey*, arXiv, 1501.03989.

[190] M. G. SANTOS ET AL., *MeerKLASS: MeerKAT Large Area Synoptic Survey*, in Proceedings, MeerKAT Science: On the Pathway to the SKA (MeerKAT2016): Stellenbosch, South Africa, May 25-27, 2016, arXiv, 1709.06099.

[191] B. P. SCHMIDT ET AL., *The High Z supernova search: Measuring cosmic deceleration and global curvature of the universe using type Ia supernovae*, Astrophys. J., 507 (1998), arXiv, astro-ph/9805200.

[192] S. SCHMIDT, B. MÉNARD, R. SCRANTON, C. MORRISON, AND C. MCBRIDE, *Recovering Redshift Distributions with Cross-Correlations: Pushing The Boundaries*, Mon. Not. Roy. Astron. Soc., 431 (2013), arXiv, 1303.0292.

[193] U. SELJAK, *Extracting primordial non-gaussianity without cosmic variance*, Phys. Rev. Lett., 102 (2009), arXiv, 0807.1770.

[194] H.-J. SEO AND C. M. HIRATA, *The foreground wedge and 21 cm BAO surveys*, Mon. Not. Roy. Astron. Soc., 456 (2016), arXiv, 1508.06503.

[195] J. R. SHAW, K. SIGURDSON, U.-L. PEN, A. STEBBINS, AND M. SITWELL, *All-Sky Interferometry with Spherical Harmonic Transit Telescopes*, Astrophys. J., 781 (2014), arXiv, 1302.0327.

[196] J. R. SHAW, K. SIGURDSON, M. SITWELL, A. STEBBINS, AND U.-L. PEN, *Coaxing cosmic 21 cm fluctuations from the polarized sky using m-mode analysis*, Phys. Rev., D91 (2015), arXiv, 1401.2095.

[197] R. S. SOMERVILLE AND R. DAVÉ, *Physical Models of Galaxy Formation in a Cosmological Framework*, Ann. Rev. Astron. Astrophys., 53 (2015), arXiv, 1412.2712.

[198] M. SPINELLI, A. ZOLDAN, G. DE LUCIA, L. XIE, AND M. VIEL, *The atomic Hydrogen content of the post-reionization Universe*, arXiv, 1909.02242.

[199] V. SPRINGEL ET AL., *Simulating the joint evolution of quasars, galaxies and their large-scale distribution*, Nature, 435 (2005), arXiv, astro-ph/0504097.

[200] V. SPRINGEL ET AL., *First results from the IllustrisTNG simulations: matter and galaxy clustering*, Mon. Not. Roy. Astron. Soc., 475 (2018), arXiv, 1707.03397.

[201] L. STAVELEY-SMITH, W. E. WILSON, T. S. BIRD, M. J. DISNEY, R. D. EKERS, K. C. FREEMAN, R. F. HAYNES, M. W. SINCLAIR, R. A. VAILE, R. L. WEBSTER, AND A. E. WRIGHT, *The Parkes 21cm Multibeam Receiver*, Publ. Astron. Soc. Aust., 13 (1996), pp. 243–248.

[202] G. STEIGMAN AND M. S. TURNER, *Cosmological constraints on the properties of weakly interacting massive particles*, Nuclear Physics B, 253 (1985), pp. 375–386.

[203] N. SUZUKI ET AL., *The Hubble Space Telescope Cluster Supernova Survey: V. Improving the Dark Energy Constraints Above z>1 and Building an Early-Type-Hosted Supernova Sample*, Astrophys. J., 746 (2012), arXiv, 1105.3470.

[204] E. R. SWITZER, T.-C. CHANG, K. W. MASUI, U.-L. PEN, AND T. C. VOYTEK, *Interpreting the unresolved intensity of cosmologically redshifted line radiation*, Astrophys. J., 815 (2015), arXiv, 1504.07527.

[205] E. R. SWITZER ET AL., *Determination of z∼0.8 neutral hydrogen fluctuations using the 21 cm intensity mapping auto-correlation*, Mon. Not. Roy. Astron. Soc., 434 (2013), arXiv, 1304.3712.

[206] A. N. TAYLOR AND A. J. S. HAMILTON, *Nonlinear cosmological power spectra in real and redshift space*, Mon. Not. Roy. Astron. Soc., 282 (1996), arXiv, astro-ph/9604020.

[207] A. THOMPSON, J. M. MORAN, AND G. SWENSON, JR, *Interferometry and Synthesis in Radio Astronomy*, vol. -1, 01 1991.

[208] J. L. TINKER, A. V. KRAVTSOV, A. KLYPIN, K. ABAZAJIAN, M. S. WARREN, G. YEPES, S. GOTTLOBER, AND D. E. HOLZ, *Toward a halo mass function for precision cosmology: The Limits of universality*, Astrophys. J., 688 (2008), arXiv, 0803.2706.

[209] J. L. TINKER, B. E. ROBERTSON, A. V. KRAVTSOV, A. KLYPIN, M. S. WARREN, G. YEPES, AND S. GOTTLOBER, *The Large Scale Bias of Dark Matter Halos: Numerical Calibration and Model Tests*, Astrophys. J., 724 (2010), arXiv, 1001.3162.

[210] M. P. VAN DAALEN AND M. WHITE, *A cross-correlation-based estimate of the galaxy luminosity function*, Mon. Not. Roy. Astron. Soc., 476 (2018), arXiv, 1703.05326.

[211] F. VILLAESCUSA-NAVARRO ET AL., *Neutral hydrogen in galaxy clusters: impact of AGN feedback and implications for intensity mapping*, Mon. Not. Roy. Astron. Soc., 456 (2016), arXiv, 1510.04277.

[212] F. VILLAESCUSA-NAVARRO, M. VIEL, K. K. DATTA, AND T. R. CHOUDHURY, *Modeling the neutral hydrogen distribution in the post-reionization Universe: intensity mapping*, JCAP, 1409 (2014), arXiv, 1405.6713.

[213] F. VILLAESCUSA-NAVARRO ET AL., *Ingredients for 21 cm Intensity Mapping*, Astrophys. J., 866 (2018), arXiv, 1804.09180.

[214] E. VISBAL, A. LOEB, AND J. S. B. WYITHE, *Cosmological Constraints from 21cm Surveys After Reionization*, JCAP, 0910 (2009), arXiv, 0812.0419.

[215] M. VOGELSBERGER, S. GENEL, V. SPRINGEL, P. TORREY, D. SIJACKI, D. XU, G. F. SNYDER, D. NELSON, AND L. HERNQUIST, *Introducing the Illustris Project: Simulating the co-evolution of dark and visible matter in the Universe*, Mon. Not. Roy. Astron. Soc., 444 (2014), arXiv, 1405.2921.

[216] B. D. WANDELT, E. HIVON, AND K. M. GORSKI, *Cosmic microwave background anisotropy power spectrum statistics for high precision cosmology*, arXiv, astro-ph/9808292.

[217] B. D. WANDELT, E. HIVON, AND K. M. GORSKI, *The pseudo-C_ℓ method: cosmic microwave background anisotropy power spectrum statistics for high precision cosmology*, Phys. Rev., D64 (2001), arXiv, astro-ph/0008111.

[218] Y. WANG, *Dark Energy*, Wiley Series in Cosmology, Wiley, 2009.

[219] R. H. WECHSLER AND J. L. TINKER, *The Connection between Galaxies and their Dark Matter Halos*, Ann. Rev. Astron. Astrophys., 56 (2018), arXiv, 1804.03097.

[220] D. H. WEINBERG, R. DAVE, N. KATZ, AND L. HERNQUIST, *Galaxy clustering and galaxy bias in a lambda-CDM universe*, Astrophys. J., 601 (2004), arXiv, astro-ph/0212356.

[221] S. WEINBERG, *A model of leptons*, Phys. Rev. Lett., 19 (1967), pp. 1264–1266.

[222] S. WEINBERG, *The cosmological constant problem*, Reviews of Modern Physics, 61 (1989), pp. 1–23.

[223] N. WIBERG, A. HOLLEMAN, AND E. WIBERG, *Holleman-Wiberg's Inorganic Chemistry*, Elsevier Science, 2001.

[224] G. M. WILLIGER, R. F. CARSWELL, R. J. WEYMANN, E. B. JENKINS, K. R. SEMBACH, T. M. TRIPP, R. DAVE, L. HABERZETTL, AND S. R. HEAP, *The Low-Redshift Lyman Alpha Forest toward 3C 273*, Mon. Not. Roy. Astron. Soc., 405 (2010), arXiv, 1002.3401.

[225] T. L. WILSON, K. ROHLFS, AND S. HÜTTEMEISTER, *Tools of Radio Astronomy*, Springer-Verlag, 2009.

[226] A. WITZEMANN, D. ALONSO, J. FONSECA, AND M. G. SANTOS, *Simulated multi-tracer analyses with HI intensity mapping*, arXiv, 1808.03093.

[227] L. WOLZ, F. B. ABDALLA, D. ALONSO, C. BLAKE, P. BULL, T.-C. CHANG, P. G. FERREIRA, C.-Y. KUO, M. G. SANTOS, AND R. SHAW, *Foreground Subtraction in Intensity Mapping with the SKA*, PoS, AASKA14 (2015), arXiv, 1501.03823.

[228] L. WOLZ, F. B. ABDALLA, C. BLAKE, J. R. SHAW, E. CHAPMAN, AND S. RAWLINGS, *The effect of foreground subtraction on cosmological measurements from Intensity Mapping*, Mon. Not. Roy. Astron. Soc., 441 (2014), arXiv, 1310.8144.

[229] L. WOLZ, C. BLAKE, AND J. S. B. WYITHE, *Determining the HI content of galaxies via intensity mapping cross-correlations*, Mon. Not. Roy. Astron. Soc., 470 (2017), arXiv, 1703.08268.

[230] L. WOLZ ET AL., *Erasing the Milky Way: new cleaning technique applied to GBT intensity mapping data*, Mon. Not. Roy. Astron. Soc., 464 (2017), arXiv, 1510.05453.

[231] L. WOLZ, S. G. MURRAY, C. BLAKE, AND J. S. WYITHE, *Intensity mapping cross-correlations II: HI halo models including shot noise*, arXiv, 1803.02477.

[232] S. WYITHE AND A. LOEB, *The 21cm Power Spectrum After Reionization*, Mon. Not. Roy. Astron. Soc., 397 (2009), arXiv, 0808.2323.

[233] L. XIE, G. DE LUCIA, M. HIRSCHMANN, F. FONTANOT, AND A. ZOLDAN, *H_2-based star formation laws in hierarchical models of galaxy formation*, , 469 (2017), arXiv, 1611.09372.

[234] S. YAHYA, P. BULL, M. G. SANTOS, M. SILVA, R. MAARTENS, P. OKOUMA, AND B. BASSETT, *Cosmological performance of SKA HI galaxy surveys*, Mon. Not. Roy. Astron. Soc., 450 (2015), arXiv, 1412.4700.

[235] X.-H. YANG, H. J. MO, Y. P. JING, F. C. VAN DEN BOSCH, AND Y.-Q. CHU, *Populating dark matter halos with galaxies: Comparing the 2dFGRS with Mock Galaxy Redshift Surveys*, Mon. Not. Roy. Astron. Soc., 350 (2004), arXiv, astro-ph/0303524.

[236] T. ZAFAR, C. PEROUX, A. POPPING, B. MILLIARD, J.-M. DEHARVENG, AND S. FRANK, *The ESO UVES Advanced Data Products Quasar Sample - II. Cosmological Evolution of the Neutral Gas Mass Density*, Astron. Astrophys., 556 (2013), arXiv, 1307.0602.

[237] S. ZAROUBI, *The Epoch of Reionization*, arXiv, 1206.0267.

[238] H. ZHENG, M. TEGMARK, J. S. DILLON, A. LIU, A. NEBEN, J. JONAS, P. REICH, W. REICH, D. A. KIM, AND A. R. NEBEN, *An improved model of diffuse galactic radio emission from 10 MHz to 5 THz*, Mon. Not. Roy. Astron. Soc., 464 (2017), arXiv, 1605.04920.

[239] A. ZOLDAN, G. DE LUCIA, L. XIE, F. FONTANOT, AND M. HIRSCHMANN, *H I-selected galaxies in hierarchical models of galaxy formation and evolution*, , 465 (2017), arXiv, 1610.02042.

[240] S. ZUO, X. CHEN, R. ANSARI, AND Y. LU, *21 cm Signal Recovery via Robust Principal Component Analysis*, Astron. J., 157 (2018), arXiv, 1801.04082.

[241] M. A. ZWAAN, M. J. MEYER, L. STAVELEY-SMITH, AND R. L. WEBSTER, *The HIPASS Catalogue: Omega(HI) and environmental effects on the HI mass function of galaxies*, Mon. Not. Roy. Astron. Soc., 359 (2005), arXiv, astro-ph/0502257.

[242] F. ZWICKY, *On the Masses of Nebulae and of Clusters of Nebulae*, Astrophys. J., 86 (1937), pp. 217–246.