UNIVERSITY OF CAPE TOWN

MASTER'S THESIS

Radio Frequency Interference: Simulations for Radio Interferometry Arrays

Author:
Chris FINLAY

Supervisor:
Prof. Bruce BASSETT

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Mathematics and Applied Mathematics

April 8, 2021

Declaration of Authorship

I, Chris FINLAY, declare that this thesis titled, "Radio Frequency Interference: Simulations for Radio Interferometry Arrays" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Signed by candidate

Date:

UNIVERSITY OF CAPE TOWN

Abstract

Science Faculty

Department of Mathematics and Applied Mathematics

Master of Science

Radio Frequency Interference:
Simulations for Radio Interferometry Arrays

by Chris FINLAY

Radio Frequency Interference (RFI) is a massive problem for radio observatories around the world. Due to the growth of telecommunications and air travel RFI is increasing exactly when the world's radio telescopes are increasing significantly in sensitivity, making RFI one of the most pressing problems for astronomy in the era of the Square Kilometre Array (SKA).

Traditionally RFI is dealt with through simple algorithms that remove unexpected rapid changes but the recent explosion of machine learning and artificial intelligence (AI) provides an exciting opportunity for pushing the state-of-the-art in RFI excision. Unfortunately, due to the lack of training data for which the true RFI contamination is known, it is impossible to reliably train and compare machine learning algorithms for RFI excision on radio telescope arrays currently.

To address this stumbling block we present RFIsim, a radio interferometry simulator that includes the telescope properties of the MeerKAT array, a sky model based on previous radio surveys coupled with an RFI model designed to reproduce actual RFI seen at the MeerKAT site. We perform an in-depth comparison of the simulator results with real observations using the MeerKAT telescope and show that RFIsim produces visibilities that mimic those produced by real observations very well. Finally, we describe how the data was key in the development of a new state-of-the-art deep learning RFI flagging algorithm in Vafaei et al. (2020.) [69] In particular, this work demonstrates that transfer learning from simulation to real data is an effective way to leverage the power of machine learning for RFI flagging in real-world observatories.

Acknowledgements

I would like to acknowledge my supervisor, Bruce Bassett, for all the guidance he has given me throughout this journey. I would also like to acknowledge Ben Hugo, Simon Perkins and others in the Radio Astronomy Research Group for helping me understand many radio interferometry concepts. Without this help the journey would have been much tougher. Further, I would also like to acknowledge my colleagues Nadeem Oozeer, Isaac Sihlangu and others. Many interesting discussions were had delving into the specifics of our work. Finally, I would like to acknowledge my wife, Cara Finlay, for her unwavering support through the toughest times including the write up of this thesis.

Contents

List of Figures

List of Tables

This thesis is dedicated to my grandfather, Alan Finlay, and grandmother, Jean Finlay. They both, unfortunately, passed away during the write up of this thesis. They lived their whole lives together and as such passed away together. They supported me through my studies and in anything that I did. Their love was truly unconditional, I wouldn't be here without you. I love you both dearly. Granny and Grandpa, you will live in our hearts forever.

Chapter 1

Motivation and Summary

1.1 Motivation

In the context of radio astronomy, Radio Frequency Interference (RFI) is any unwanted signal present in data collected from a radio telescope. An unwanted signal would be any signal that is not of astronomical origin with amplitude greater than the astronomical signal or the noise in the telescope. Some common examples that are external to a telescope array, at mid-frequency ($\sim$ 1 - 5 GHz), include Global System for Mobile Communications (GSM) signals, WiFi and Bluetooth, Global Navigation Satellite System (GNSS) signals, and Distance Measuring Equipment (DME) signals from airplanes. Another more inadvertent example is the spark plug from petrol-based vehicles and then there are telescope-based RFI sources such as leakage from improperly shielded electronics that run the telescope.

Radio observatories across the world such as the Giant Metrewave Radio Telescope (GMRT) [63], the Karl G. Jansky Very Large Array (JVLA) [46], the upcoming Square Kilometre Array (SKA) [23] and its precursor the MeerKAT, [36], have gone to great lengths to reduce the presence of RFI around their telescopes. The process of mitigating the contribution of RFI includes positioning the antennas far away from substantial human habitation, introducing government regulations [56] that prohibit the use of certain bands in the spectrum or areas of land, and using natural shields like mountains to protect the antennas from RFI.

The problem of RFI is multifaceted. Radio telescopes are set up and designed for receiving faint signals from distant celestial objects. In many cases, the received signal from RFI is orders of magnitude stronger than the astronomical signal astronomers are interested in. This may even lead to a saturation of the receivers which renders that data completely unusable. For nearly all cases the RFI sources do not lie in the far-field of the telescope. This creates a problem when using contaminated data for imaging, since imaging algorithms assume the signal to originate in the far-field. The RFI signal originating in the near-field will not resolve to the source location and will instead contribute to the background noise of the entire image. The extent of this

problem then depends on the magnitude of the RFI.

There are two different paths that can be taken to minimize the issue of RFI. The simplest approach is to flag the RFI. This is the process of determining if a given data sample is contaminated with RFI signal and flagging it for complete removal. This is usually the first step in the radio astronomy data reduction process. The alternative is to try and subtract [4], or block the contribution from RFI sources. This can be done pre-correlation by blocking contaminated samples from entering the correlator or post-correlation by calculating the expected contribution from RFI sources.

Traditionally astronomers would manually flag their data. This is a very labour intensive process that involves an astronomer physically looking at much of their data and through experience alone determining if a data sample is contaminated and then flagging it for removal. As radio datasets grew larger this process became a very laborious and so it led many to develop methods to automatically flag data. In [57] the flagging process is done in the UV-plane. The `SumThreshold` algorithm implemented in `AOFlagger` [43] is a time-frequency domain based method that is in common use today, including in the MeerKAT data reduction pipeline [59]. In recent years many deep learning (machine learning) based approaches have been investigated for flagging RFI automatically [2, 71, 69]. These approaches also work in the time-frequency domain. For a more extensive overview of the problem of RFI and its mitigation we refer the reader to [9] and the references therein.

RFI comes in many forms and characteristics so with such a broad variety it is hard to develop an algorithm that works in all of these cases. In the case of flagging, the removal of all RFI is very hard and often leads to an excess number of false positives. On the other hand, in the case of subtraction, some of the sources of RFI are just completely unknown so there is little hope in trying to predict/model their contribution. The overarching problem here is that it is very difficult to determine what data are contaminated with RFI, and exactly what the source of that RFI is in all cases. The development of algorithms for RFI flagging or subtraction is therefore hindered by a lack of ground truth. We cannot in good faith compare algorithms when we do not know the true answer.

The above-mentioned complexities of RFI detection and excision motivated the work presented here. If we can simulate visibility data that are contaminated by known sources of RFI, where the process is modeled from the signal emission source all the way through the telescope and other associated effects, we would then have a testing bed, with ground truth answers, for algorithm development. We cannot simulate every possible source of RFI, however, we can make strides towards simulating some of the more common and well-known sources. Once a good foundation has been developed this can be the starting point for more complicated RFI simulations

down the road. To serve the goal of creating a genuine testbed for RFI flagging/subtraction algorithm development we will need realistic RFI contaminated visibilities where we know the exact contribution from both RFI and astronomical sources. Such modeling has been done for single dish simulations in [1] however RFI sources were not modeled and such simulations have not been done for interferometric arrays. To this end, the goal of this thesis will be to create a radio interferometry simulator that includes a physical RFI model to generate realistic interferometric data.

1.2 Summary

Chapter 2 covers the introductory material needed to develop the simulator. We present an introduction to sources of radio emission leading to an explanation of the fundamentals of radio interferometry. The subject of electromagnetic wave polarization is then explained which provides the necessary background to introduce the Radio Interferometry Measurement Equation (RIME). The RIME is the main computational tool used to build the simulator. The chapter finishes with an introduction to Radio Frequency Interference (RFI) and how it can be incorporated into the simulator.

Chapter 3 describes the methods used to acquire the data and computational tools required to build the simulator. It starts with building a sky model from two radio sky surveys covering the entire celestial sphere. It then goes on to describe how a realistic RFI model is developed. This includes the positions and movements of RFI sources as well as descriptions of their emission. Next, the acquisition of the main components of the telescope, specifically for MeerKAT, is described. These components are the telescope array geometry, its primary beam model, and the time dependent bandpass. Finally, the computational tools, Montblanc and Dask, are described and the inclusion of the RFI model, telescope components and sky model are discussed.

Chapter 4 presents the basic results of this work including a real-world application thereof. Initially, the data produced by the simulator are compared with real MeerKAT observations. The data are compared from a few specific views. The first is a cut across frequency at an arbitrary time. This view allows one to compare the telescope response over frequency, the frequency distribution of RFI, and the relative power of both RFI and astronomical sources. In the next section, the data are compared using spectrograms for both the magnitude and phase of the visibilities. Using this view one can compare the time variations of the visibilities at all simulated frequencies. The comparison is finalized with a view of the visibility phases over time for both astronomical sources and RFI at two baselines of similar orientation but vastly different lengths. The various features present in this view are extensively explained. The chapter is finished off by describing how the final simulator was used in a paper to develop a state-of-the-art RFI mitigation

algorithm that uses deep learning to leverage the near-boundless amount of data the simulator can produce.

1.3 Contributions of This Thesis

The main contributions of this thesis are the derivation of a brightness matrix and phase delay term for near-field sources as well as a radio interferometry simulator that includes a realistic RFI model for the MeerKAT telescope. Another large contribution of this thesis is the creation of a ground truth dataset that was used to train and test a machine learning model for RFI mitigation in [69]. The derivation of RIME components applicable to near-field sources such as RFI is done in Section 2.5.4. We explain the specifics of our realistic RFI model in Section 3.2. In Sections 3.7 and 3.8 we explain how the first and second versions of RFIsim are implemented. The contribution of the RFI dataset to the development of a RFI mitigation algorithm is described further in Section 4.2. The repository for RFIsim is located at https://github.com/chrisfinlay/RFIsim. The `master` branch includes the code for the first version and the `curved_wavefront` branch includes the code for the second version.

Chapter 2

Introduction

2.1 Radio Emission

Radio emission is electromagnetic radiation with frequencies in the radio band. The radio band is defined differently in different disciplines but is generally in the 3 MHz to 300 GHz frequency range. The radio band is also further split up into sub-bands such as is defined by the Institute of Electrical and Electronics Engineers (IEEE) Standard Letter Designations, [13]. Table 2.1 shows the IEEE set of radio band designations.

IEEE Designation	Frequency Range (GHz)
HF	0.003 - 0.03
VHF	0.03 - 0.3
UHF	0.3 - 1
L	1 - 2
S	2 - 4
C	4 - 8
X	8 - 12
K_u	12 - 18
K	18 - 27
K_a	27 - 40
V	40 - 75
W	75 - 110

TABLE 2.1: IEEE radio band designations with their frequency ranges. The MeerKAT telescope currently operates in the L-band and UHF-band with the S-band in the testing phase.

Source: [21]

Karl G. Jansky was the first person to discover radio waves emanating from the Milky Way in 1931 and as such is considered one of the founding fathers of radio astronomy. Due to his great prominence in the field a physical unit was named after him. In radio astronomy, when we measure the strength of a source's radio emission, we are measuring the spectral flux density. This

has units of Watts per square metre per Hertz, however, in radio astronomy the unit Jansky is preferred. The relation is 1 Jy = 10^{-26} $Wm^{-2}Hz^{-1}$. The integral over frequency gives us the flux density of the source, also referred to as the intensity.

Radio astronomers split radio emission into two main categories, continuum emission and line emission. Certain physical phenomena, known as emission mechanisms, lead to distinct types of radio emission. The study of an astronomical source's radio emission allows astronomers to infer properties of that source. We will discuss some of these emission mechanisms in further detail in this section.

2.1.1 Continuum Emission

Continuum emission is electromagnetic radiation that covers a very broad frequency range [25]. Continuum emission is split into thermal and non-thermal emission. The separation of which is determined by the underlying physical process. Different types of continuum emission lead to different contributions to a source's spectral profile. A spectral profile characterizes the frequency dependence of a source's spectral flux density. Sections of a spectral profile can be described by the spectral index. This concept is described in Section 2.1.3. We will briefly touch on the most important sources of continuum emission in the radio band in this section.

2.1.1.1 Black-body Radiation

When matter is in local thermal equilibrium (LTE) with a temperature above absolute zero it will emit electromagnetic radiation. A black-body is an idealised object that absorbs all incident radiation, with non reflected or transmitted, from any angle of incidence. A black-body in LTE at temperature T will emit electromagnetic radiation according to Planck's law [47]. The profound finding by Planck is that the emission spectrum is only dependent on the temperature of the black-body. In radio astronomy, an astronomical object in local thermal equilibrium can often be modelled as a black-body with an emissivity less than 1, a grey-body. The Cosmic Microwave Background [44] (CMB) is the most famous example of black-body radiation in astronomy. Planck's law is mathematically given by 2.1.

$$B_\nu(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT} - 1} \tag{2.1}$$

In Equation 2.1 B_ν is the spectral brightness of the black-body [67], h is Planck's constant, c is the speed of light in a vacuum, k is Boltzmann's constant, T is

the temperature of the black-body and ν is the emission frequency. B_ν has units of Watts per square metre per steradian per Hertz.

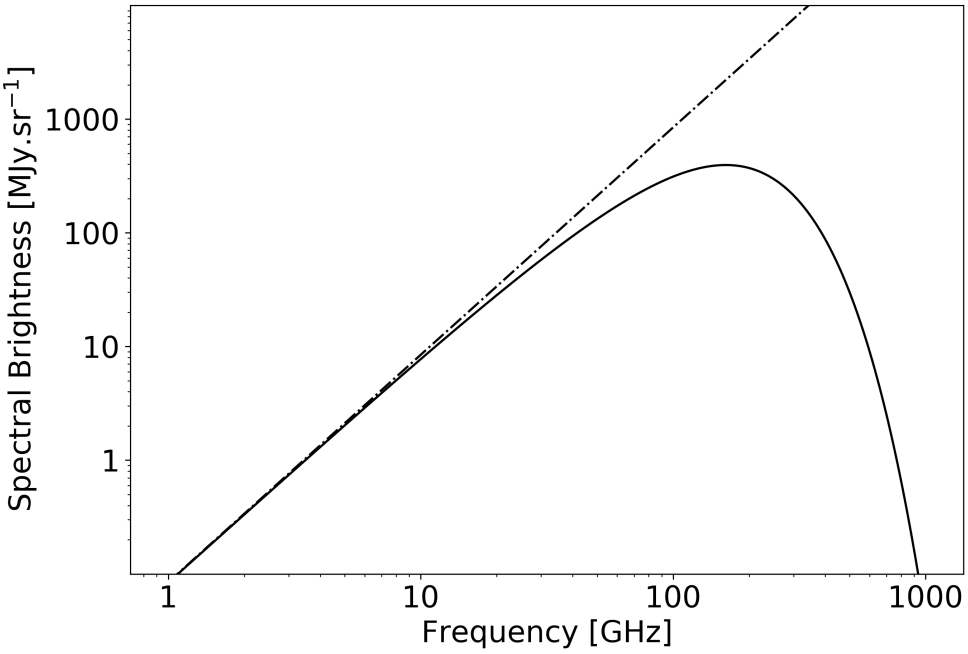

FIGURE 2.1: This plot shows the spectral distribution of black-body radiation for a body at 2.725 K, the temperature of the CMB. The black curve marks the Planck distribution and the dot dash line marks the Rayleigh-Jeans approximation.

Source: [14]

At frequencies where quantum effects can be ignored [14], $\nu \ll kT/h \approx 20T_{\text{source}}$ GHz, we can use the Rayleigh-Jeans approximation to express spectral brightness as a temperature. Here T_{source} is the temperature of the source that is in LTE. The Rayleigh approximation is shown in equation 2.2. This is referred to as the brightness temperature.

$$T = \frac{B_\nu c^2}{2k\nu^2} \tag{2.2}$$

2.1.1.2 Bremsstrahlung

Bremsstrahlung, the direct translation of which is braking radiation, is the result of a charged particle accelerating. When a collection of bremsstrahlung emitting particles are in local thermal equilibrium it is referred to as thermal bremsstrahlung [18]. In the context of radio astronomy this type of radiation

appears in regions of ionised hydrogen known as HII regions. HII regions contain hot plasma that is made up of ionised hydrogen. In these regions the ions are constantly deflecting off one another in free-free interactions, causing bremsstrahlung.

2.1.1.3 Synchrotron Emission

Synchrotron radiation is the result of charged particles, moving at relativistic speeds, being accelerated by magnetic fields, hence the alternative appellation magnetobremsstrahlung. When the charged particles are moving at non-relativistic speeds it is called cyclotron emission. Since the velocity distribution of relativistic electrons, in most astronomical sources of magnetobremsstrahlung, follow a power law, not a Maxwellian distribution indicating a LTE, magnetobremsstrahlung is considered a source of non-thermal emission [18].

Synchrotron radiation emits most of its radiation at a specific frequency. The way continuum emission results from this is that there is a broad distribution of velocities among the particles. The superposition of radiation from the collective of charged particles leads to a continuum of radiation due to the continuous distribution of particle velocities. Synchrotron emission can also be partially polarized [38].

One of the most well known examples of synchrotron emission is in Active Galactic Nuclei (AGN). Figure 2.2 shows a radio image of Cygnus A with an AGN at its core. The AGN is the source of relativistic outflows of plasma and magnetic fields that form what are called the jets [15].

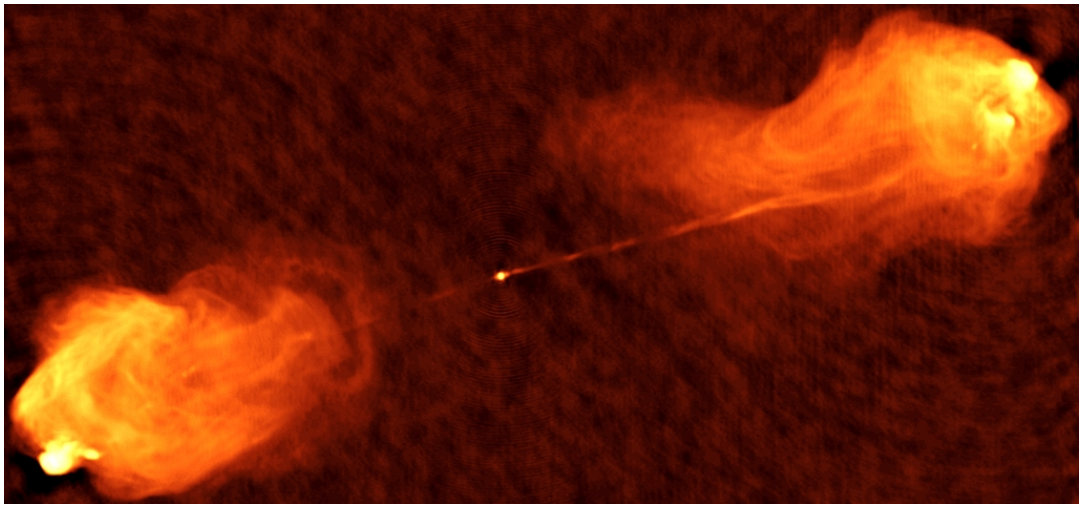

FIGURE 2.2: This is a radio image of Cygnus A at 5 GHz and a resolution of 0.4 arcseconds. The data was captured with the Very Large Array (VLA) and processed into this image by Rick Perley.

Source: [15]

2.1.2 Line Emission

Spectral lines are narrow, $\Delta\nu \ll \nu$, features in the emission spectrum of a source [19]. There are absorption lines and emission lines. These are caused by energy transitions that are quantum in nature. The emission/absorption frequency is proportional to the energy difference from the transition, due to $E = h\nu$.

2.1.2.1 Neutral Hydrogen (HI)

Hydrogen is the most abundant element in our universe and is the fundamental building block for all other elements that are created within stars. HI refers to hydrogen in its neutral, atomic state. HI emission is caused by the hyperfine-structure transition of the hydrogen atom. In this transition the spin of the electron and proton change from being parallel to antiparallel as shown in Figure 2.3. This is actually an extremely improbable event to the extent that it is considered a forbidden transition having a transition rate of 2.9×10^{-15} s^{-1}. Due to the immense quantity of HI in the universe we do in fact see HI emission from nearly all galaxies. Its emission frequency is precisely known to be 1,420,405,751.7667 $\pm$ 0.0009 Hz [29].

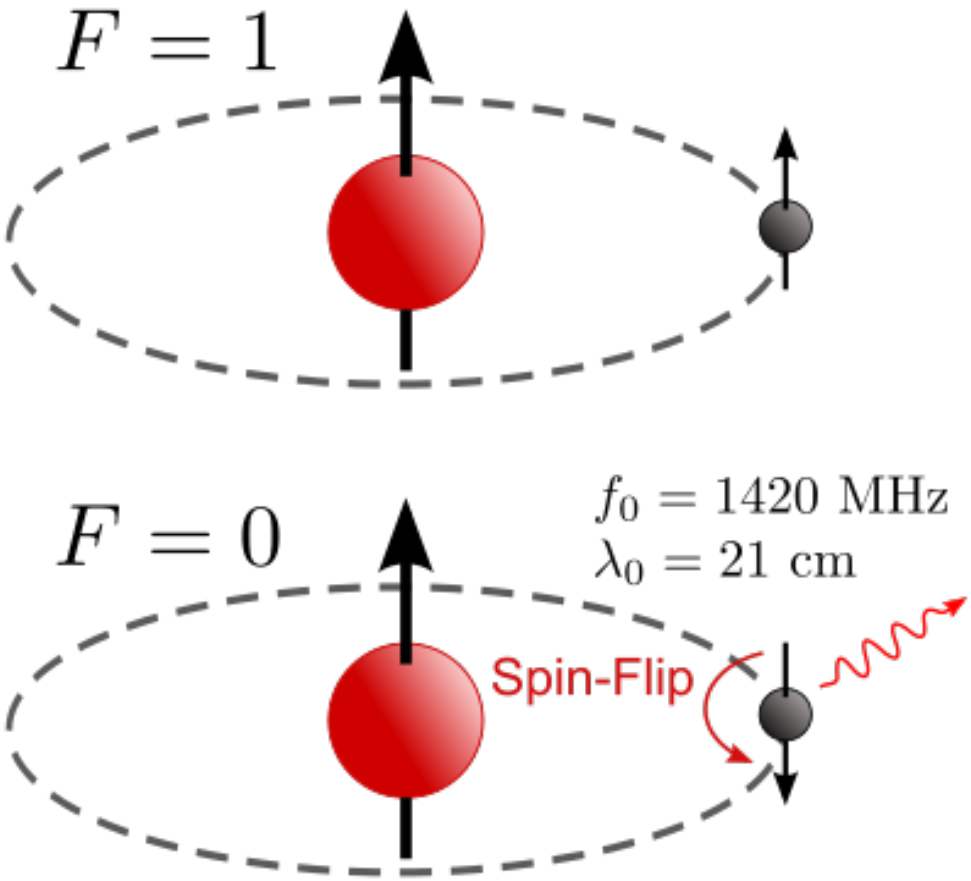

FIGURE 2.3: Schematic diagram representing the spin-flip transition of a neutral hydrogen atom.

Source: [68]

2.1.3 Spectral Index

The spectral index of a source is a parameter that describes the source's spectral flux density, I_ν, dependence on frequency. It is defined in terms of a derivative in Equation 2.3 [19].

$$\alpha(\nu) = \frac{d \log I_\nu}{d \log \nu} \tag{2.3}$$

In general the spectral index varies over frequency, however over the frequency range covered by a typical narrowband observation a constant value is a good approximation. The value of α gives a clue as to the type of emission. Figure 2.4 below shows an example spectrum of the galaxy M82 over a large frequency range which shows different type of emission from the

galaxy.

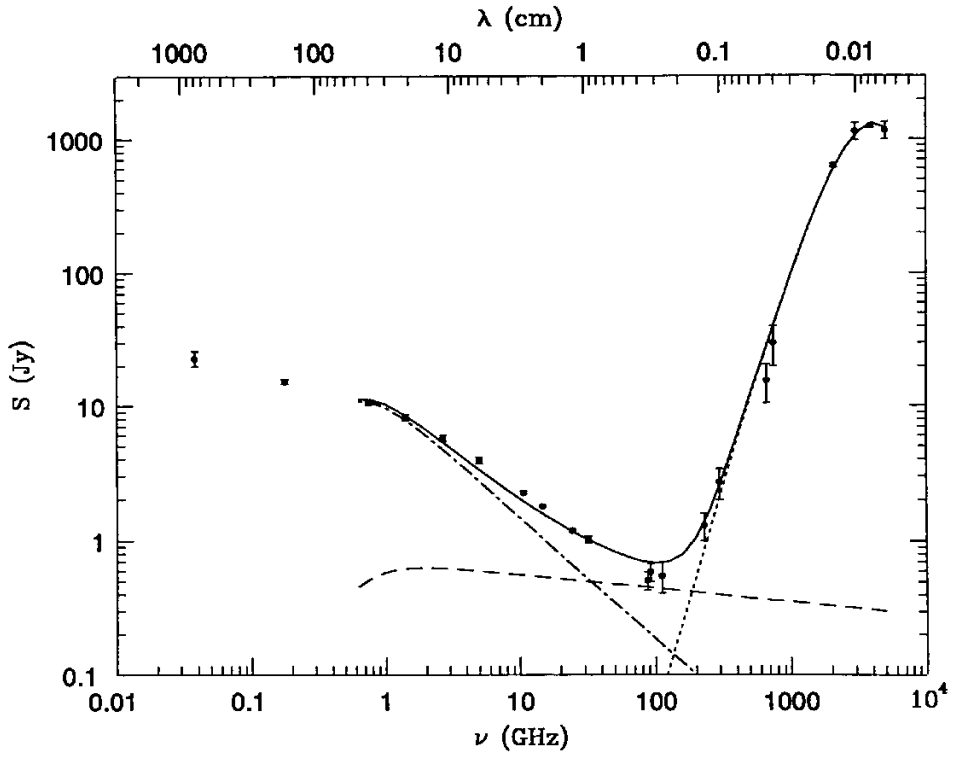

FIGURE 2.4: The spectral flux density spectrum of galaxy M82. The black dots (with error bars) are measured values. The solid black line shows a fitted curve to the spectrum. The dot-dash line shows the contribution from synchrotron emission to the overall spectrum. The dashed line shows the contribution from thermal brehmsstralung induced by hot, ionized HII regions. The dotted line shows the contribution from grey-body radiation. The galaxy has dust that absorbs higher energy photons that heat them up causing them to radiate thermal black-body emission.

Source: [18]

The spectral index as defined in Equation 2.3 is the slope of the spectrum when expressed on a log-log plot as is done in Figure 2.4. For smaller sections a straight line approximation can be used. This leads to the power law approximation that is commonly used to model the spectral flux density. The power law is defined in the equation below where the spectral index is assumed constant.

$$I_\nu \propto \nu^\alpha \tag{2.4}$$

We will see in Section 3.1 how this is used to model the spectral profile of astronomical sources in our simulations.

2.2 Radio Interferometry

Radio Interferometry as a technique for Radio Astronomy was first developed in the 1940s. It was born out of the need for greater angular resolution. The angular resolution of a telescope is limited due to the diffraction limit by the size of its aperture. For a single dish the angular resolution is inversely proportional to the diameter of the dish. A radio dish can only be so large before it becomes immovable and impractical. The solution was to introduce more smaller, more manageable dishes and link them together to simulate a much larger dish. This process of simulating a larger dish is called aperture synthesis and is achieved through the use of a radio interferometer. In this case the resolution of the telescope is inversely proportional to the largest distance between antennas in the interferometric array. In this section we will describe how an interferometer works and explain some of the formalism that will be used throughout this thesis.

2.2.1 Radio Interferometers

As explained previously, an interferometer is an array of individual antennas that work in unison to achieve the resolution of a very large dish. To do this an interferometer tracks a specific reference point in the sky known as the *phase centre*. This is done so that plane waves arriving from distant sources in that particular direction arrive in phase across the array of antennas. In order to do this, time delays need to be inserted in the signal chain of each antenna, so as to make the array appear to be in a plane perpendicular to the direction of the phase centre. The time delayed signals from all the antennae are then correlated with one another to produce visibilities. Let us start with a basic diagram shown in Figure 2.5.

We would like to work out something called the *geometric delay*, τ_g, which is the the time delay needed for signals, from the phase centre, to arrive in phase at the correlator. We begin by defining the baseline vector, $\vec{b}$ which points from antenna 1 to antenna 2. The direction to the phase centre is denoted with the unit vector $\hat{s}_0$. The path length difference will be $c\tau_g$. Using the diagram in Figure 2.5 we can determine the path length difference to be the projection of the baseline vector onto the unit vector $\hat{s}_0$. Expressed mathematically it is the following:

$$c\tau_g = \vec{b} \cdot \hat{s}_0 \tag{2.5}$$

Rearranging this to find the time delay τ_g we get:

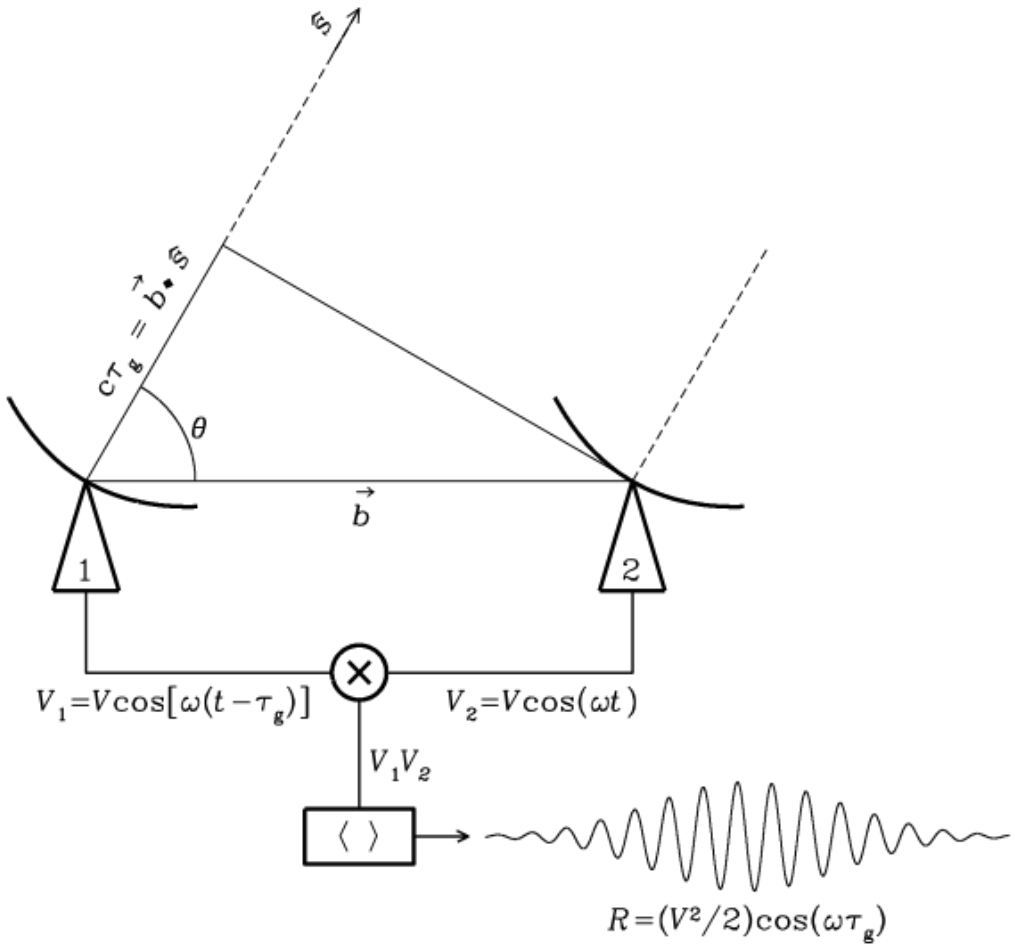

FIGURE 2.5: Diagram of a single baseline interferometer. The baseline vector $\vec{b}$ points from antenna 1 to antenna 2. $\hat{s_0}$ points toward the phase centre.

Source: [34]

$$\tau_g = \frac{\vec{b} \cdot \hat{s}_0}{c} \tag{2.6}$$

We have now determined how to calculate the necessary time delay for a specific baseline in order to point in a specific direction, the phase centre. The phase centre follows a source as it rises and sets. Therefore τ_g varies with time. The phase delay needs to be constantly calculated and inserted in the signal chain in order to keep the target at the phase centre. An interferometer that does this is called a phase tracking interferometer or a fringe stopping interferometer. This is because the insertion of this time delay stops fringes from moving over the target direction.

2.2.2 Coordinate Systems

Radio interferometry makes use of a number of different coordinate systems. In this thesis we will make use of two coordinate systems for astronomical source positions and three coordinate systems for antenna positions and baselines.

For astronomical sources we will use the standard celestial coordinates right ascension, α, and declination, δ, with the J2000[1] epoch. The other astronomical coordinate system is composed of cosine coordinates and is described in the next subsection. For antennas and baselines we have three coordinate systems that we will use. There are ENU, XYZ and uvw. These are all measured in the same distance units and are simply rotations of one another. These are described in section 2.2.2.2. The objects described by these coordinate systems remain static in their respective systems, except for antennas in the uvw frame which change over time.

2.2.2.1 Celestial Coordinates

Let us start with a standard Cartesian coordinate system, except instead of using $\{x, y, z\}$ we will use $\{u, v, w\}$. Our coordinate system will have its origin at the centre of the interferometric array. This is somewhat arbitrary but helps with imagining things. We will align the u-axis with East, the v-axis with North and the w-axis will point toward the phase centre. This is shown in Figure 2.6.

Using this coordinate system we can describe positions of antennas in the standard way. A baseline is described by the difference in position vectors of two different antennas.

Now we need to be able to describe source positions. Since astronomical sources are so far away that we cannot measure radial depth we simply assume them to be projected onto the surface of a sphere known as the celestial sphere. Let the radius of this sphere be $R = \sqrt{u^2 + v^2 + w^2}$. We would like to factor this radial depth out as it is arbitrary. From this alone we can define the sky coordinate system using direction cosines. To be clear we are defining a new coordinate system using $\{l, m, n\}$ that is perfectly aligned with $\{u, v, w\}$ except we will remove the radial factor, R, and therefore it will have no units. This is shown in equation (2.7) below.

[1] J2000 epoch refers to a particular standard where source positions are defined relative to the equinox position in the year 2000.

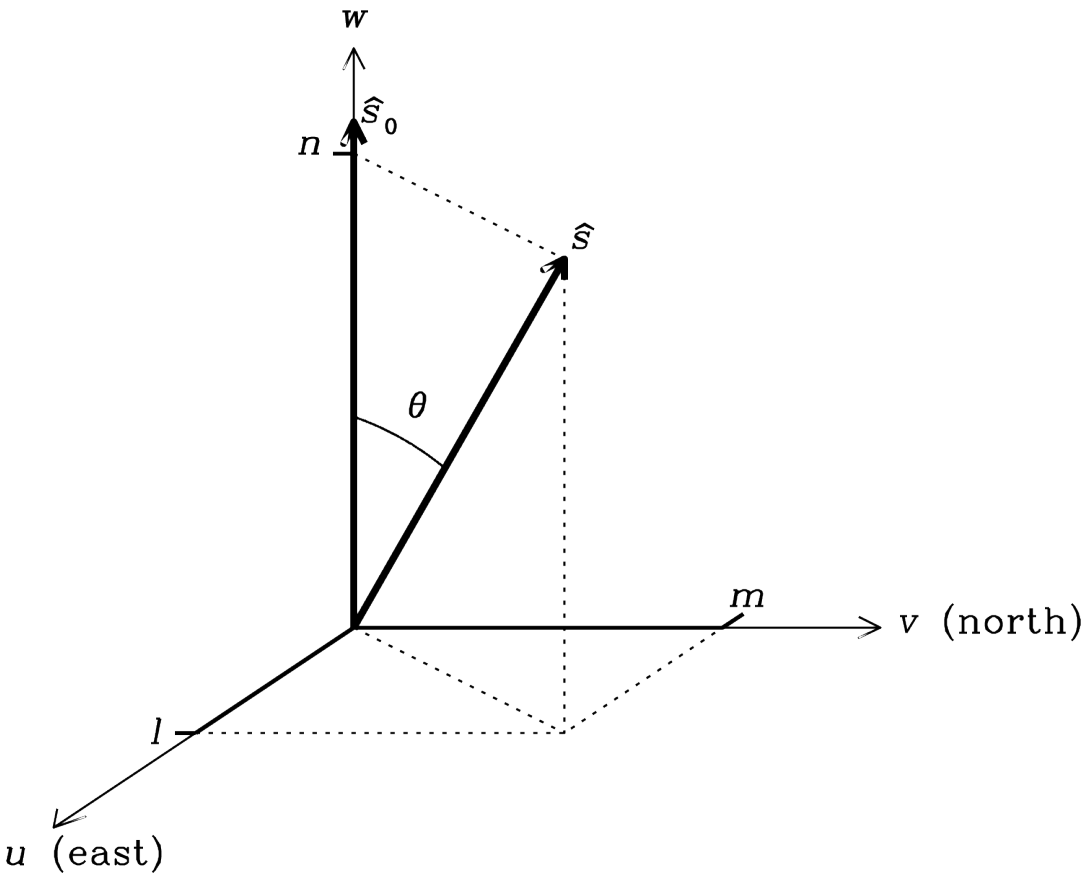

FIGURE 2.6: This diagram shows the two fundamental coordinate systems in radio interferometry. The uvw coordinate system is used to describe baselines and the lmn coordinate system is used to describe astronomical source positions. We see from this diagram that the w-axis is aligned with the n-axis and the u and v axes are also aligned with the l and m axes. In this diagram the angle θ denotes the angle between the direction $\hat{s}$ and n and w.

Source: [35]

$$\begin{aligned} l &= \frac{u}{R} = \cos(\theta_u) \\ m &= \frac{v}{R} = \cos(\theta_v) \\ n &= \frac{w}{R} = \cos(\theta_w) \end{aligned} \tag{2.7}$$

In equation 2.7 the angles *theta* with a subscript are the angles between the axes uvw denoted by the subscript and the direction of the source $\hat{s}$. From equation (2.7) above we can easily show that $l^2 + m^2 + n^2 = 1$ which is simply the equation of a sphere with unit radius as expected. It is to be noted that the origin of this new coordinate system is $(0, 0, 1)$.

Throughout the rest of this thesis we will use $\boldsymbol{l} = (l, m, n)^T$ to denote the column vector of direction cosines and $\boldsymbol{u} = (u, v, w)^T$ to denote the positional column vector of an antenna or a baseline depending on the context.

We will now show the transformation from celestial coordinates in terms of *right ascension*, α, and *declination*, δ, to direction cosines. It is here for reference purposes only and will not be derived. First let the phase centre position, which will be mapped to $(0, 0, 1)$, be (α_0, δ_0) and (α, δ) be some arbitrary source position. We then have

$$\begin{aligned} l &= \cos(\delta)\sin(\alpha - \alpha_0) \\ m &= \sin(\delta)\cos(\delta_0) - \cos(\delta)\sin(\delta_0)\cos(\alpha - \alpha_0) \end{aligned} \tag{2.8}$$

2.2.2.2 Baseline Coordinates Over Time

The coordinates of a baseline are dependent on the position of the phase centre of the array. Since the phase centre will be moving, relative to a stationary antenna, due to the diurnal cycle then the baseline coordinates for a single pair of antennas will change over time and be dependent on the position of the antennas on the Earth. Let us start from a basic antenna coordinate system that can be defined anywhere on the Earth. The ENU coordinate system stands for East, North, Up and is a Cartesian system measured in metres. The origin of the coordinate system is preferentially chosen to be the centre of the array or some reference antenna, however, it should not be at one of the poles. From this local coordinate system we can transform through a single rotation to a different one we will name XYZ. This new coordinate system XYZ will have the XY-plane sitting parallel to the Earth's equator, the equatorial plane, and the Z-axis will point towards the North Pole. The X-axis will point toward the point on the surface of the Earth with latitude of $0°$ and longitude of $0°$. This type of reference frame is known as an Earth Centred Earth Fixed (ECEF) frame. Equation 2.9 shows the rotation to perform this transformation. Let the latitude be given by λ and the longitude by ϕ.

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} -\sin\phi & -\cos\phi\sin\lambda & \cos\phi\cos\lambda \\ \cos\phi & -\sin\phi\sin\lambda & \sin\phi\cos\lambda \\ 0 & \cos\lambda & \sin\lambda \end{pmatrix} \begin{pmatrix} E \\ N \\ U \end{pmatrix} \tag{2.9}$$

Now that we have our antennas in an ECEF frame we can again transform coordinates to our final uvw frame. This is done using Equation 2.10. First we need to define $H_0 = GMST - \phi - \alpha$ where $GMST$ is the Greenwich Mean Sidereal Time, ϕ again is the longitude of the antenna array and α is the right ascension of the phase centre. δ is the declination of the phase centre.

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} \sin H_0 & \cos H_0 & 0 \\ -\cos H_0 \sin\delta & \sin H_0 \sin\delta & \cos\delta \\ \cos H_0 \cos\delta & -\sin H_0 \cos\delta & \sin\delta \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{2.10}$$

We now have all of our antennas in the uvw frame from which we can take differences between antenna positions to achieve baselines. The units have remained the same throughout as we have only performed rotations. The uvw frame described here is not the same as is used in Taylor [66] and Swenson [67]. They use a frame, $u'v'w'$, to describe baseline vectors divided by the observing wavelength. To obtain a $u'v'w'$ baseline coordinate from a pair of uvw antenna coordinates, with the antennae labelled as j and k, equation 2.11 should be used. Throughout this thesis we will use the uvw frame described in this section.

$$\begin{pmatrix} u' \\ v' \\ w' \end{pmatrix} = \frac{1}{\lambda} \left[\begin{pmatrix} u_j \\ v_j \\ w_j \end{pmatrix} - \begin{pmatrix} u_k \\ v_k \\ w_k \end{pmatrix} \right] \tag{2.11}$$

2.2.3 Visibilities

Let us consider a quasi-monochromatic plane wave, of frequency ν, hitting a pair of antennas separated by some distance. This scenario is depicted in Figure 2.5. Let the voltage induced at each antenna be directly proportional to the electric field of the incoming electromagnetic (EM) wave. The measured voltages, V_j and V_k, as a function of time at antennas j and k will be

$$V_j = V\cos\big(2\pi\nu(t-\tau)\big), \quad V_k = V\cos\big(2\pi\nu t\big) \tag{2.12}$$

Let us now take the cross-correlation of these two signals. $\langle\cdot\rangle$ denotes a time average over an as yet unspecified time.

$$\langle V_j V_k \rangle = V^2 \Big\langle \cos\big(2\pi\nu(t-\tau)\big)\cos\big(2\pi\nu t\big) \Big\rangle \tag{2.13}$$

$$= \frac{V^2}{2} \Big\langle \cos\big(4\pi\nu t - 2\pi\nu\tau\big) + \cos\big(2\pi\nu\tau\big) \Big\rangle \tag{2.14}$$

$$= \frac{V^2}{2}\cos\big(2\pi\nu\tau\big) \tag{2.15}$$

Going from equations (2.13) to (2.14) the cosine addition rule has been used then the double angle formulas for cosine and sine have been used and finally the cosine addition rule is used again. This then leaves the multiplication of

the signals in a nice form that we can determine the necessary averaging interval from. Allowing the averaging interval to be sufficiently greater than $(2\nu)^{-1}$ we can see that the $\cos\left(4\pi\nu t - 2\pi\nu\tau\right)$ term will average out to 0 leaving only the final term in equation (2.15). This is the output of something called a cosine correlator. If one were to place a 90° phase delay in the signal chain of one of the antennas and follow this same calculation one would find the output to be $(V^2/2)\sin(2\pi\nu\tau)$. This is then called a sine correlator. Together these form something called a complex correlator as the output can then be represented by a complex number as

$$\frac{V^2}{2}e^{i2\pi\nu\tau} \tag{2.16}$$

We now consider the full response of the telescope as it receives signal from the entire sky. In practice it is only a small region around the phase centre that we are interested in. This is because the directional response of a parabolic dish is generally very small, around 1° on the sky. Remembering that the output from our correlator is proportional to V^2 which is proportional to the power received i.e. the *brightness distribution*, $I_\nu(\boldsymbol{l})$, at frequency ν. We then have

$$r(\tau) = \int I_\nu(\boldsymbol{l})e^{-i2\pi\nu\tau}d\Omega \tag{2.17}$$

Let τ be the geometric delay induced by a point source in direction $\boldsymbol{l}$. τ_g is the geometric delay that is inserted in the signal chain due to tracking the phase centre in direction $\boldsymbol{l}_0$, this is identical to $\hat{s}_0$ in section 2.2.1. $\boldsymbol{u}$ will be the baseline vector. Recall equation (2.6). We will rewrite this in our new notation.

$$\tau = \frac{\boldsymbol{u}^T\boldsymbol{l}}{c} \tag{2.18}$$

What we call the visibility is the response shown in equation (2.17) except delayed by τ_g.

$$\begin{aligned} r(\tau - \tau_g) &= \int I_\nu(\boldsymbol{l}) \exp\left[-2\pi i\nu(\tau - \tau_g)\right] d\Omega \\ &= \int I_\nu(\boldsymbol{l}) \exp\left[-\frac{2\pi i\nu}{c}\boldsymbol{u}^T(\boldsymbol{l} - \boldsymbol{l}_0)\right] d\Omega \\ V_\nu(u,v,w) &= \iint I_\nu(l,m,n) \exp\left[-\frac{2\pi i}{\lambda}\left(ul + vm + w(n-1)\right)\right] \frac{dldm}{n} \end{aligned} \tag{2.19}$$

Going from the second step to the third in equation (2.19) is a matter of noting $c = \lambda\nu$, $\boldsymbol{l}_0 = (0,0,1)^T$ and expanding the vectors into their respective components. $d\Omega = dldm/n$, however this is not immediately obvious. $1/n$ is the Jacobian correction going from integrating over a curved surface to a flat surface of the same dimension.

Equation (2.19) holds for a specific frequency however because of the principle of superposition we can sum over frequency components without having to consider additional terms. Since $n(l,m)$ we find that $I(l,m,n) = I(l,m)$. This leaves us with the following

$$V(u,v,w) = \iint I(l,m)\exp\Big[-\frac{2\pi i}{\lambda}\big(ul+vm+w(n-1)\big)\Big]\frac{dldm}{n}, \quad n = \sqrt{1-l^2-m^2} \tag{2.20}$$

This looks very much like a Fourier transform but is is not. In the case of an East-West coplanar array configuration there would never be a w-term. In this specific case we find that $V(\boldsymbol{u})$ is the Fourier transform of the modified sky brightness distribution $I(\boldsymbol{l})/n$. Equation (2.20) is what is known as the van Cittert-Zernike (vCZ) theorem.

2.2.3.1 Fourier Transform Scenario

If one assumes the field of view (FoV) of a telescope, defined by the primary beam of an antenna, to be small enough such that $l^2 + m^2 \ll 1 \implies n \approx 1$ then we find that the w-term in the exponential and the Jacobian correction, $1/n$, of equation (2.20) vanishes. This leaves us with the following equation (2.21) which is an exact Fourier transform.

$$V(u,v) = \iint I(l,m)\exp\Big[-\frac{2\pi i}{\lambda}\big(ul+vm\big)\Big]dldm \tag{2.21}$$

2.3 Polarization

Polarization is a property of transverse waves like an electromagnetic (EM) wave. In the case of an EM wave it describes the direction of the electric field vector relative to its Poynting[2] vector as the wave propagates.

From Figure 2.7 above we can see the EM wave shown by the red curve can be composed by a superposition of the two waves described by the blue and green curves. If one were to allow a phase shift between these two waves

[2] A Poynting vector points in the direction of wave propagation.

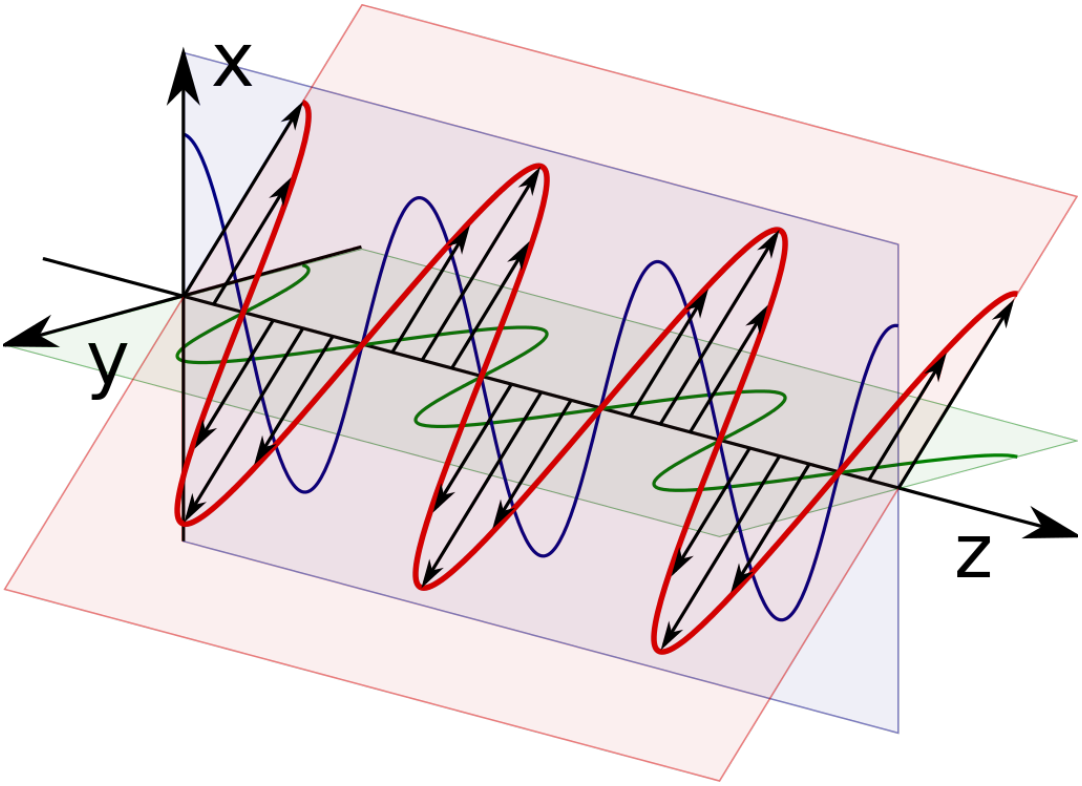

FIGURE 2.7: This is a diagram depicting a linearly polarized EM wave propagating along the Z-axis. The red shaded plane is the polarization plane with the red curve showing the amplitude of the electric field vector. The blue and green shaded planes with their accompanying curves show the X and Y components of the electric field vector respectively.

Source: [39]

then the red curve would no longer lie in a plane but would trace out an ellipse in the XY-plane. When this phase shift is 90° the resulting wave will be circularly polarized. The diagram in Figure 2.8 shows an example of a right hand circularly polarized EM wave propagating in the Z-direction.

In general the polarization of an EM wave can be characterised by a polarization ellipse. It therefore needs 3 numbers to describe [12]. From the progression of this section we can see that one could parameterise the polarization ellipse with two amplitudes and a phase difference. There are many ways of parameterising the polarization ellipse. In 1852 George Gabriel Stokes came up with what are now termed the Stokes parameters. We will describe these in the following subsection.

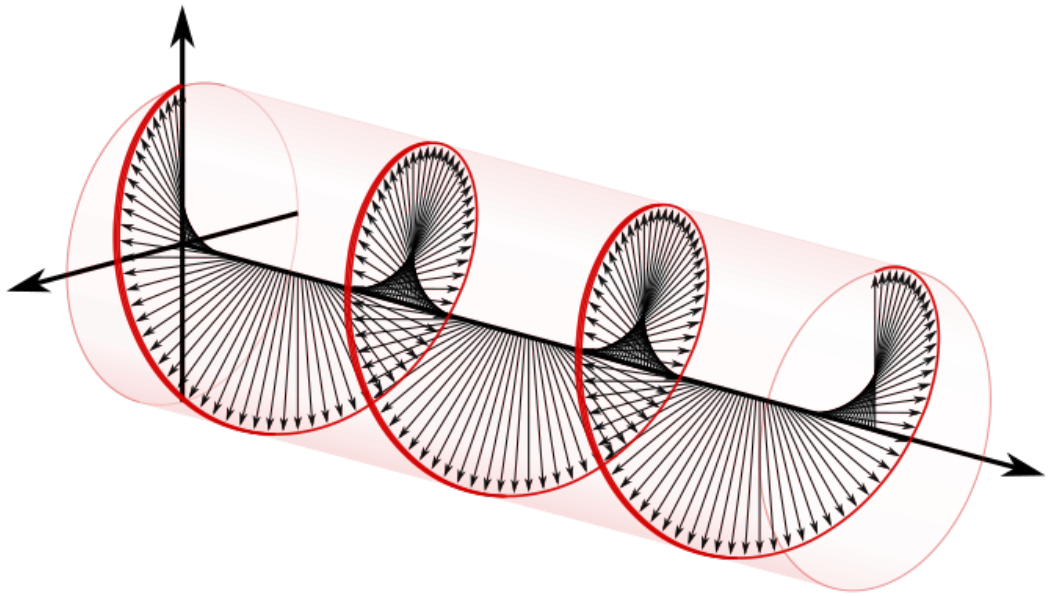

FIGURE 2.8: This diagram shows an example of a right hand circularly polarized EM wave propagating in the z-direction. The arrows show the direction of the electric field vector at different moments in time.

Source: [22]

2.3.1 Stokes Parameters

The Stokes parameters fully define the polarization state of some EM wave [62]. They are four real-valued numbers namely I, Q, U and V. I is the total intensity of the wave and is therefore always greater than 0. Q, U and V are the different polarization intensities. They are more succinctly described in Figure 2.9 below.

If we define the direction of propagation to be along the z-axis in a standard Cartesian basis then the four Stokes parameters are defined as follows. Here $\langle \cdot \rangle$ denotes an expectation value or time averaging of the signal and e_x and e_y are the electric field strength in the x and y directions respectively.

$$\begin{aligned} I &= \langle e_x e_x^* \rangle + \langle e_y e_y^* \rangle \\ Q &= \langle e_x e_x^* \rangle - \langle e_y e_y^* \rangle \\ U &= \langle e_y e_x^* \rangle + \langle e_x e_y^* \rangle \\ V &= i\langle e_y e_x^* \rangle - i\langle e_x e_y^* \rangle \end{aligned} \tag{2.22}$$

We can invert these relationships to recover the correlation products in terms of the Stokes parameters. These are given below in equation (2.23).

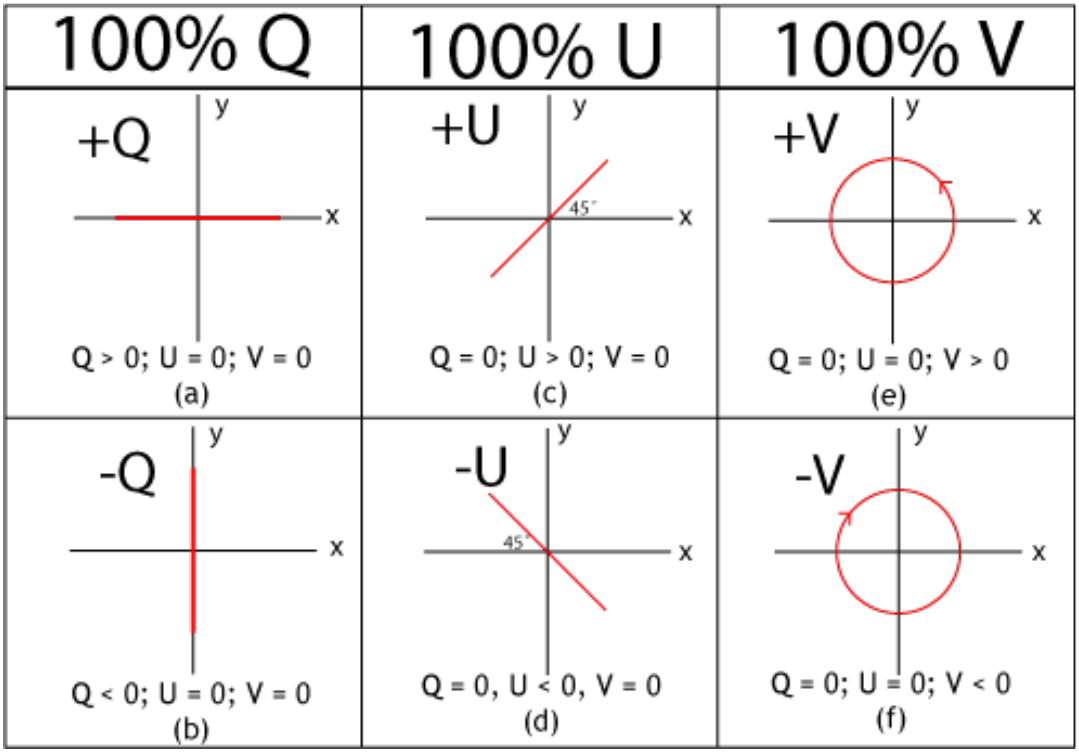

FIGURE 2.9: The polarized Stokes parameters. Both Q and U make up the linear polarization portion. They can be combined to create $L = Q + iU$. This is considered a generalised complex intensity that encapsulates the entire linear polarization portion. V is the circularly polarized portion and when V is negative we have right hand circular polarization. A total polarization intensity, I_p, is defined as $I_p = \sqrt{Q^2 + U^2 + V^2}$.

Source: [41]

$$\begin{aligned}
\langle e_x e_x^* \rangle &= \frac{1}{2}\Big(I + Q\Big) \\
\langle e_y e_y^* \rangle &= \frac{1}{2}\Big(I - Q\Big) \\
\langle e_x e_y^* \rangle &= \frac{1}{2}\Big(U + iV\Big) \\
\langle e_y e_x^* \rangle &= \frac{1}{2}\Big(U - iV\Big)
\end{aligned} \tag{2.23}$$

These definitions have been given with respect to a Cartesian basis. They can also be defined with respect to a circular basis. When measuring the polarization of an EM wave the basis is defined by the antenna/feed design. MeerKAT and subsequently the SKA use linearly polarized feeds which lead to a Cartesian basis. This is achieved with two linear feeds orientated at 90° to one another (orthogonal). The JVLA uses circularly polarized feeds which lead to the circular basis. The orthogonal feed pair in this case are

a right hand circularly polarized feed and a left hand circularly polarized feed. For completeness we will show the Stokes parameters defined in terms of a circular basis where (l, r) denote the orthogonal components left hand circular and right hand circular respectively. Again these are with respect to an increasing phase convention.

$$\begin{aligned} I &= \langle e_l e_l^* \rangle + \langle e_r e_r^* \rangle \\ Q &= \langle e_r e_l^* \rangle + \langle e_l e_r^* \rangle \\ U &= i\langle e_r e_l^* \rangle - i\langle e_l e_r^* \rangle \\ V &= \langle e_r e_r^* \rangle - \langle e_l e_l^* \rangle \end{aligned} \tag{2.24}$$

We will see in the following section on the Radio Interferometry Measurement Equation that these four correlation products are what make up the *brightness matrix*.

2.4 Radio Interferometry Measurement Equation

The Radio Interferometry Measurement Equation (RIME), first developed by [28], is a way to describe signal propagation effects in interferometric arrays. It allows one to concisely describe these effects in the signal chain by matrix multiplications. This makes it extremely useful for calibration purposes where the idea is to invert these effects to arrive at the original signal before its measurement.

The original RIME formulation uses a 4-vector of Stokes parameters which in turn needs 4 x 4 matrices to describe each propagation effect. This is known as the Mueller formalism. It was, however, revised in Hamaker [27] to a much simpler formalism that made use of 2 x 2 matrix composed of the 4 Stokes parameters. This new formalism then made use of 2 x 2 matrices, named Jones matrices, to describe propagation effects in a far more intuitive way. The Mueller formalism is more general than the Jones formalism however the Jones formalism is far more intuitive and easy to work with. In this section we will derive the 2 x 2 RIME formalism as we will use it to simulate visibilities. This derivation will follow closely to [60].

Before starting with this derivation we would like to explicitly define the notation that will be used. It is to be noted that all scalars and components of higher order objects are complex quantities unless otherwise stated.

- Scalar quantities will be lower case in normal weight font such as the phase delay κ. The only exception to this will be the Stokes parameters.
- Vectors will be lower case in a bold font like the electric field vector $\boldsymbol{e}$.
- Matrices will be upper case in a bold font like the gains $\boldsymbol{G}$.

Given an EM wave propagating along the z-axis we can describe its electric field vector by a 2-vector, named a Jones vector, as follows:

$$\boldsymbol{e} = \begin{pmatrix} e_x \\ e_y \end{pmatrix} \tag{2.25}$$

Now let the measured complex voltage on each orthogonal feed be

$$\boldsymbol{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix} \tag{2.26}$$

Assuming that all effects in the signal chain from the original signal all the way to the measured voltage are linear effects we can describe the transformation by a matrix multiplication. Linear mappings applied to a Jones vector are called a Jones matrix. All the possible actions of applying Jones matrices to Jones vectors is called the Jones calculus. When the electric fields are weak nonlinearities in the effects can be ignored as the linear portion is dominant and therefore leads to very accurate approximations of the total effect. Let $\boldsymbol{J}$ be the linear mapping from the electric field vector, $\boldsymbol{e}$, to the measured voltage vector $\boldsymbol{v}$. We then have the following

$$\boldsymbol{v} = \boldsymbol{J}\boldsymbol{e} \tag{2.27}$$

Let us take antennas j and k and do a cross-correlation of their voltage vectors by taking the outer product of the two vectors and averaging over some time. $\boldsymbol{V}_{jk}$ will be the result and we call this the visibility on the baseline defined by the distance vector between antenna j and k. It is the visibility contribution from a single point source. Here $\langle \cdot \rangle$ denotes averaging. Taking $\boldsymbol{v}$ to be a column vector we can do this cross-correlation as follows

$$\begin{aligned} \boldsymbol{V}_{jk} &= \langle \boldsymbol{v}_j \boldsymbol{v}_k^\dagger \rangle \\ &= \langle \boldsymbol{J}_j \boldsymbol{e} (\boldsymbol{J}_k \boldsymbol{e})^\dagger \rangle \\ &= \langle \boldsymbol{J}_j \boldsymbol{e}\boldsymbol{e}^\dagger \boldsymbol{J}_k^\dagger \rangle \\ &= \boldsymbol{J}_j \langle \boldsymbol{e}\boldsymbol{e}^\dagger \rangle \boldsymbol{J}_k^\dagger \end{aligned} \tag{2.28}$$

In order for the last step in equation (2.28) to hold we need the effects encapsulated by $\boldsymbol{J}_j$ for all antennas to be constant over the averaging interval. We will assume this to be true.

Equation (2.28) defines the visibility for a single point source but our antennas receive signals from across the entire sky. To describe the complete visibility from all contributing sources we need to integrate over the entire sky. The all-sky RIME is shown in equation (2.29) below.

$$\boldsymbol{V}_{jk} = \iint \boldsymbol{J}_j \langle \boldsymbol{e}\boldsymbol{e}^\dagger \rangle \boldsymbol{J}_k^\dagger d\Omega \tag{2.29}$$

We can in fact break up $\boldsymbol{J}$ into a series like $\boldsymbol{J}_n \boldsymbol{J}_{n-1} ... \boldsymbol{J}_1$ since the composition of linear mappings is another linear mapping. The reason one may want to do this is so that different effects can be attributed to their own $\boldsymbol{J}_n$. One would have to apply each $\boldsymbol{J}_n$ to $\boldsymbol{e}$ in the order in which they affect the signal. This is because matrix multiplications are in general not commutative. We will delve further into some examples of the most basic individual $\boldsymbol{J}_n$ in subsections 2.4.2, 2.4.3 and 2.4.4.

2.4.1 Brightness Matrix

Let us take a closer look at the $\langle \boldsymbol{e}\boldsymbol{e}^\dagger \rangle$ term from equation (2.28).

$$\begin{aligned} \langle \boldsymbol{e}\boldsymbol{e}^\dagger \rangle &= \left\langle \begin{pmatrix} e_x e_x^* & e_x e_y^* \\ e_y e_x^* & e_y e_y^* \end{pmatrix} \right\rangle \\ &= \begin{pmatrix} \langle e_x e_x^* \rangle & \langle e_x e_y^* \rangle \\ \langle e_y e_x^* \rangle & \langle e_y e_y^* \rangle \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} I + Q & U + iV \\ U - iV & I - Q \end{pmatrix} \\ &:= \boldsymbol{B} \end{aligned} \tag{2.30}$$

We have just defined the *brightness matrix*, $\boldsymbol{B}$, for a single point source in terms of the Stokes parameters of the source. It is to be noted that we have used what is called in the literature the *Convention-1/2* as opposed to the *Convention-1*. These conventions define whether the factor $1/2$ should be used as in the definition above. For a more in depth review on the conventional controversy of the brightness matrix please see section 7.2 from [60]. In short it does not matter all that much as absolute flux calibrators are used to calibrate the instrument and therefore the factor will simply be absorbed into the antenna gains.

2.4.2 Geometric Delay Term

The geometric delay or phase delay term denoted by $\boldsymbol{K}$ is a scalar matrix. Let us define some arbitrary origin for the coordinate system in which we will measure positions of antennas. Let the position vector of antenna j be denoted by $\boldsymbol{u}_j$ where:

$$\boldsymbol{u}_j = \begin{pmatrix} u_j \\ v_j \\ w_j \end{pmatrix} \tag{2.31}$$

and be measured in metres. As shown in section 2.2.2 we have positions of sources in the sky measured as direction cosines. We will denote the vector of direction cosines for a single source by $\boldsymbol{l}_s$ such that:

$$\boldsymbol{l}_s = \begin{pmatrix} l_s \\ m_s \\ n_s \end{pmatrix} \tag{2.32}$$

This vector is unitless. We are now in a position to construct a geometric delay term. We will define $\boldsymbol{K}_{js}$ in terms of a complex scalar called the phase delay, κ.

$$\kappa_{js} = \frac{2\pi}{\lambda}\left(u_j l_s + v_j m_s + w_j(n_s - 1)\right) \tag{2.33}$$

$$\boldsymbol{K}_{js} = \boldsymbol{I}e^{-i\kappa_{js}} \tag{2.34}$$

Here $\boldsymbol{I}$ is the identity matrix. In the case of an ideal interferometer and only free space between the source and the receivers the only Jones term would be the geometric delay. Let us consider this case below.

$$\begin{aligned} \boldsymbol{V}_{jks} &= \boldsymbol{K}_{js}\boldsymbol{B}_s\boldsymbol{K}_{ks}^{\dagger} \\ &= (\boldsymbol{I}e^{-i\kappa_{js}})\boldsymbol{B}_s(\boldsymbol{I}e^{i\kappa_{ks}}) \\ &= \boldsymbol{B}_s e^{-i(\kappa_{js}-\kappa_{ks})} \\ &= \boldsymbol{B}_s e^{-\frac{2\pi i}{\lambda}\left((u_j-u_k)l_s+(v_j-v_k)m_s+(w_j-w_k)(n_s-1)\right)} \\ \boldsymbol{V}_{jks} &= \boldsymbol{B}_s e^{\frac{2\pi i}{\lambda}\left(u_{jk}l_s+v_{jk}m_s+w_{jk}(n_s-1)\right)} \end{aligned} \tag{2.35}$$

Where $\boldsymbol{u}_{jk} = \boldsymbol{u}_j - \boldsymbol{u}_k$. We find the result to be exactly what we want except for the $1/n$ term. It is simply the visibility contribution from a single point source as shown in the original derivation in section 2.2.3. Since the $1/n$ term is only source dependent one could define an apparent brightness matrix, $\boldsymbol{B}_{app} = \boldsymbol{B}_s/n$.

2.4.3 Direction-Dependent Effects

Direction-dependent effects (DDEs) are corruptions in the signal chain that are dependent on source positions. In other words these are effects that vary over the sky. The main DDE is the primary beam of each antenna. We will discuss this effect first and then describe a few other examples of DDEs. For a more in-depth review of DDEs, we refer the reader to [61]. This is an excellent text on which much of this section is based.

2.4.3.1 Primary Beam

The primary beam of an antenna is characterised by its directional sensitivity. A number of factors affect how this term is composed. The main contributing factor is the shape of the aperture/dish which affects both polarization feeds. The individual feeds (the linear or circular receivers that allow us to measure the EM field vector) also have an effect as one may be more sensitive than the other or have a slightly different geometry. If the feeds were identical then this would be a scalar matrix. In reality the feeds are not identical and are also susceptible to polarization leakage (when one feed is measuring some signal that should be measured only by the orthogonal/other feed) caused by crosstalk[3] or feed orientation. The Jones term for the primary beam, $\boldsymbol{E}$, is given as follows.

$$\boldsymbol{E}(\boldsymbol{l}) = \begin{pmatrix} E_{xx}(\boldsymbol{l}) & E_{xy}(\boldsymbol{l}) \\ E_{yx}(\boldsymbol{l}) & E_{yy}(\boldsymbol{l}) \end{pmatrix} \tag{2.36}$$

In the case of no polarization leakage the above matrix would be diagonal. There is a more general point to be noted here that applies for all Jones terms. Corruptions that affect both feeds identically manifest as scalar matrices (a scalar multiple of the identity matrix). Ones that affect each feed independently but not identically manifest themselves as diagonal matrices. Whenever there is some sort of polarization leakage the associated Jones term will have off diagonal elements.

2.4.3.2 Pointing Errors

Pointing errors occur when an antenna is not pointing directly in the desired direction, usually the phase centre. Again $\boldsymbol{E}(\boldsymbol{l})$ will be the primary beam attenuation in direction $\boldsymbol{l}$. Pointing errors are included in the primary beam

[3] Crosstalk is an effect characterised by electric signals in one circuit creating an undesired effect in another circuit.

term through additional parameters namely δl and δm. These are the offsets of the pointing in the l and m directions respectively.

$$\boldsymbol{E}(l, m, \delta l, \delta m) = \boldsymbol{E}(l + \delta l, m + \delta m) \tag{2.37}$$

2.4.3.3 Parallactic Angle

The parallactic angle effect is caused by using Altitude-Azimuth (AltAz) [4] antenna mounts that effectively lead to a sky rotation from the perspective of the telescope. From the perspective of the sky the primary beam and receivers/feeds are rotating. As a result it is sometimes called the *feed rotation* term and takes the form of a rotation matrix for linear feeds. The angle γ is the parallactic angle which is the angle between the great circle through a celestial object and the zenith, and the hour circle of the object [64].

$$\boldsymbol{L}(\gamma) = \boldsymbol{R}(\gamma) = \begin{pmatrix} \cos\gamma & \sin\gamma \\ -\sin\gamma & \cos\gamma \end{pmatrix} \tag{2.38}$$

2.4.3.4 Faraday Rotation

Faraday rotation occurs when a magnetic field is present in a medium of free electrons and an EM wave passes through it leading to a rotation of its field vector. It manifests itself as a rotation matrix, for linear feeds, where the angle of rotation, β, is proportional to ν^{-2} as well at the electron density and magnetic field strength of the medium along the line of sight to the source.

$$\boldsymbol{F}(\beta) = \boldsymbol{R}(\beta) \tag{2.39}$$

2.4.3.5 Ionospheric Phase Delay

Ionospheric phase delays are particularly important at low frequencies (<$\sim$1 GHz) since they are proportional to ν^{-1}. Being a phase delay these effects are also captured by a scalar matrix much like the geometric delay term. The

[4] An Altitude-Azimuth mount is a type of antenna mounting where the telescope is able to move about two axes, namely altitude and azimuth. This stands in contrast to an equatorial mount which has its axes of rotation aligned with the equatorial plane. In an equatorial mount one axis moves through declination and the other through right ascension.

term is dependent on both frequency and the total electron content, T, along the line of sight to the source.

$$\boldsymbol{Z}(T,\nu) = \boldsymbol{I}e^{-i\alpha T/\nu} \tag{2.40}$$

2.4.4 Direction-Independent Effects

Direction-independent effects (DIEs) are corruptions in the signal chain that do not depend on source position. DIEs affect the total signal that arrives at an antenna. The effects are typically due to the amplification of the analogue signal, bandpass filter used to channelize the signal and leakage of signal from one orthogonal feed to its complement. Much of this section is based on [61] so please refer to this text for further details.

2.4.4.1 Gains

Gains is a general term used for DIEs, however, we will use it to refer to the overall amplification and bandpass here. The amplifier tends to affect the total signal (all frequencies) received and varies with time. The bandpass of a telescope is the frequency response of the instrument. When the bandpass is stable in time we can usually, and in practice do, break the gains into a time varying component and a frequency varying component. Since the signal from each feed is independently processed and the signal chain is different for each we end up with a diagonal Jones matrix. The bandpass, frequency varying component, is usually designated with $\boldsymbol{B}(\nu)$. It is an unfortunate clash with the brightness matrix. The time varying component, overall amplification, is designated with $\boldsymbol{G}(t)$. The Jones matrices look as follows:

$$\boldsymbol{B}(\nu) = \begin{pmatrix} b_x(\nu) & 0 \\ 0 & b_y(\nu) \end{pmatrix}, \quad \boldsymbol{G}(t) = \begin{pmatrix} g_x(t) & 0 \\ 0 & g_y(t) \end{pmatrix} \tag{2.41}$$

2.4.4.2 Polarization Leakage

Polarization leakage can be caused by a number of things. It is mainly caused by misalignment of the feeds so they are not perfectly orthogonal and cross talk between complimentary feeds. These effects can exhibit themselves as both DIEs and DDEs. The direction-dependent polarization leakage is usually absorbed into the primary beam term as stated in section 2.4.3. For DIEs

an explicit term for leakage like this can be used to reduce the degrees of freedom in the model. The parameter d is the percentage of polarization leakage present.

$$\boldsymbol{D}(d) = \begin{pmatrix} 1 & d \\ -d & 1 \end{pmatrix}, \quad d \ll 1 \tag{2.42}$$

2.4.5 Full Sky RIME

In the previous subsections we have gone into detail explaining a number of sources of signal corruption. They are very useful if one wants to gain information about those specific corruptions from the radio data. One could also use other instruments to probe a number of these effects to simulate data for these specific scenarios.

To compose a RIME one must choose the set of signal corruptions, expressed as Jones terms, one wants to include. These Jones terms can be both DDEs and DIEs. We will label these $(\boldsymbol{J}_1)_j$ to $(\boldsymbol{J}_N)_j$ for corruptions on the path to antenna j. Let the subscript inside the brackets denote the order in which the corruptions affect the incoming signal. The DDEs should go first and then the DIEs. Let $\boldsymbol{B}_s$ be the brightness matrix for a source s. We can now apply the Jones terms, in the correct order, to each source's brightness matrix and take a sum over all point sources in the sky. This is shown in equation 2.43.

$$\boldsymbol{V}_{jk} = \sum_s (\boldsymbol{J}_1)_j ... (\boldsymbol{J}_N)_j \boldsymbol{B}_s (\boldsymbol{J}_N)_k^\dagger ... (\boldsymbol{J}_1)_k^\dagger \tag{2.43}$$

The summation in equation 2.43 can be moved inside the DIE Jones terms for computational efficiency.

The most basic yet still realistic form of the RIME includes only the most fundamental corruptions. These include the geometric delay, (since this is present even in an idealised instrument) the primary beam and the gains (both including polarization leakage). In practice the sky is modelled as a collection of point sources turning the integral into a sum. In order to be processed on a computer the problem needs to be discretized in some way and modelling the sky as point sources is a very convenient way.

$$\boldsymbol{V}_{jk} = \boldsymbol{G}_j \Big(\sum_s \boldsymbol{E}_{js} \boldsymbol{K}_{js} \boldsymbol{B}_s \boldsymbol{E}_{ks}^\dagger \boldsymbol{K}_{ks}^\dagger \Big) \boldsymbol{G}_k^\dagger \tag{2.44}$$

2.5 Radio Frequency Interference

As explained in the motivation, Radio Frequency Interference (RFI) is any unwanted signal present in the data collected form a radio telescope. In the L-band, the main sources of RFI at the MeerKAT site are Distance Measuring Equipment (DME) from aeroplanes, Global System for Mobile Communications (GSM) signals from cellular phones and their towers, Global Navigation Satellite Systems (GNSS) and some telecommunications satellites. Each one of these sources occupy specific frequency bands. Figure 2.10 shows the channels occupied by known sources of RFI at the MeerKAT site. These are overlaid on a histogram of RFI flagging across frequency.

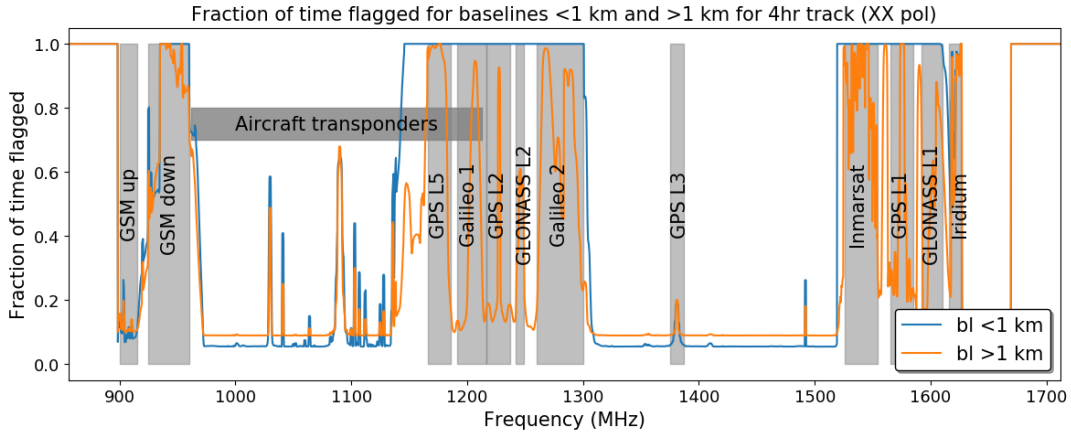

FIGURE 2.10: This Figure shows the RFI occupancy of the frequency channels in the MeerKAT L-band. The blue curve is the average across baselines shorter than 1 km and the orange curve is the average of baselines longer than 1 km. The grey boxes are bands designated to different RFI sources.

Source: [51]

Table 2.2 summarizes the RFI sources applicable to MeerKAT L-band and their frequency ranges.

2.5.1 RFI Signal

Figure 2.11 shows the typical frequency ranges and duty cycles of different sources of RFI. Satellites in the L-band are the sources of RFI that we will be most interested in for this thesis.

L-band RFI sources	
RFI Source	Frequency (MHz)
DME	962 - 1213
GSM	900 - 915 925 - 960
GNSS	1164 - 1300 1559 - 1610
Telecommunications	1525 - 1559 (Inmarsat) 1626.5-1660.5 (Inmarsat) 1616 - 1626.5 (Iridium)

TABLE 2.2: Frequency ranges of the most prominent RFI sources present on the MeerKAT site in the L-band.

Source: [58], [51], [55], [65]

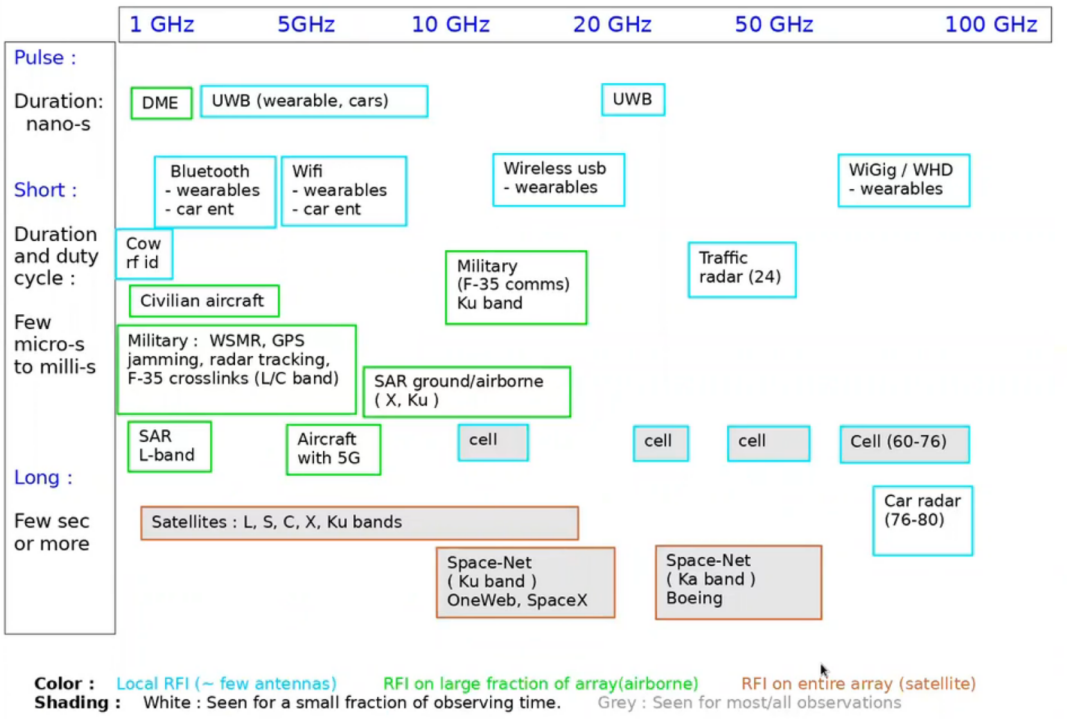

FIGURE 2.11: This Figure shows a variety of RFI sources and classifies them according to their emission frequencies and time scale (duty cycle) of their signals. For this thesis we are only interested in the low frequency section of this Figure around 1 GHz. In this figure there are a number of acronyms used. They are expanded as follows: DME - Distance Measuring Equipment, UWB - Ultra-Wide Band, USB - Universal Serial Bus, WHD - Wireless High Definition, RFID - Radio Frequency IDentification, WSMR - White Sands Missile Range, GPS - Global Positioning System, SAR - Synthetic Aperture Radar. Finally cell is short for cellular.

Source: [49]

2.5.2 Satellites

Satellites fall into one of four altitude classes. Low Earth orbit (LEO) satellites orbit at an altitude of less than 2000 km. Medium Earth orbit (MEO) satellites orbit at altitudes in the range of 2000 - 35,786 km. Geosynchronous orbit (GEO) satellites orbit at 35,786 km.

2.5.2.1 GNSS Satellites

There are four GNSS systems with global coverage. These are the BeiDou Navigation Satellite System (BDS), Galileo, the Global Navigation Satellite System (GLONASS) and the Global Positioning System (GPS).Details about these systems are summarized in the following Table 2.3.

GNSS Systems			
System	Owner	Altitude (km)	Frequency (GHz)
BDS	China	21,150	1.561098 (B1) 1.589742 (B1-2) 1.20714 (B2) 1.26852 (B3)
Galileo	EU	23,222	1.559-1.592 (E1) 1.164-1.215 (E5a/b) 1.260-1.300 (E6)
GLONASS	Russia	19,130	1.593-1.610 (G1) 1.237-1.254 (G2) 1.189-1.214 (G3)
GPS	USA	20,180	1.563-1.587 (L1) 1.215-1.2396 (L2) 1.164-1.189 (L5)

TABLE 2.3: Frequencies and altitudes of GNSS systems from around the world.

Source: [55]

All four of these GNSS satellite constellations use Right-Hand Circularly Polarized (RHCP) signals, [26], [5]. This is done because of the effects of Faraday Rotation. In section 2.4.3.4 we saw that in the Cartesian basis that Faraday rotation rotates the electric field vector in the xy-plane. For a circularly polarized wave such an effect results in only a phase shift.

2.5.2.2 Telecommunications Satellites

Iridium and International Maritime Satellites (Inmarsat) are telecommunications satellite constellations. Their emission frequencies are indicated in

Table 2.2. The Iridium satellite constellation orbits the Earth at an altitude of approximately 780 km, [31]. Iridium satellite signals are also RHCP, [32]. Inmarsat satellites are in geostationary orbit [72] and emit RHCP waves in the L-band. In other bands Inmarsat uses a combination of RHCP and LHCP for their uplink and downlink channels [33].

2.5.3 Near-field Sources

EM radiation coming from a source being received by an antenna will behave differently based on the distance between the source and antenna. This distance is often broken into regions where different effects are dominant. On the highest level it is broken into the near-field and the far-field. The near-field is further broken down into the reactive near-field and radiative near-field [10]. These regions are depicted in Figure 2.12. Let D be the size of the receiver and λ the wavelength of emission, we then have the boundary between the reactive and radiative near-field regions as $0.62\sqrt{\frac{D^3}{\lambda}}$, [10]. In some literature another region called the transition region is included between the radiative near-field and the far-field, [54]. This would appear to be for electromagnetically short antennas ($D < \lambda/4$). In this case the boundary between the reactive and radiative regions is $\frac{\lambda}{2\pi}$.

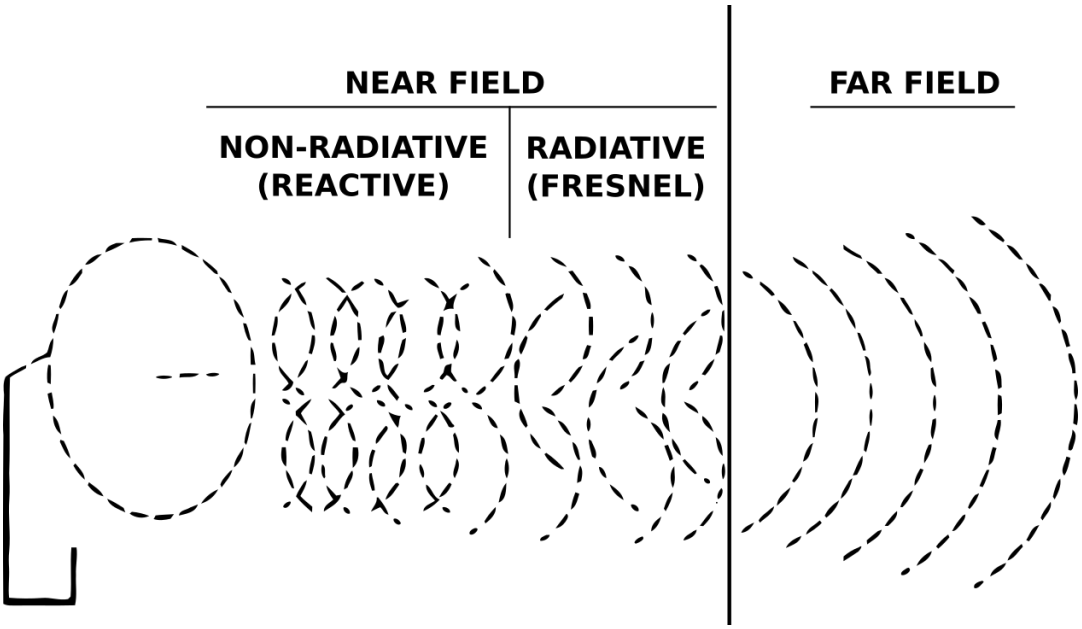

FIGURE 2.12: This diagram depicts the different field regions. A transition region may sometimes be included between the radiative near-field and the far-field.

Source: [24]

In our derivation of the visibilities in section 2.2.3 we assumed the incoming EM waves to be plane waves. This is a perfectly acceptable assumption for sources that are in the far-field. For sources in the near-field this is not an acceptable assumption and therefore the curvature of the wavefront needs to be considered. The Fraunhofer distance, d_F, defines the boundary between the near-field and far-field. The Fraunhofer distance is defined as follows:

$$d_F = \frac{2D^2}{\lambda} \tag{2.45}$$

where D is the largest dimension of the radiating or receiving antenna[5] and λ is the wavelength of the radiation. For the case of MeerKAT the longest baseline is $B_{max} \approx 8$ km and observing wavelengths in the L-band around 20 cm. When using $D = B_{max}$ and plugging in these typical values we arrive at:

$$\begin{aligned} d_f &= \frac{2(8 \times 10^3 \text{ m})^2}{2 \times 10^{-1} \text{ m}} \\ &= 640,000 \text{ km} \end{aligned} \tag{2.46}$$

which is much farther away than satellites in geostationary orbit, the farthest commonly used orbit (about the Earth) for satellites is at 35,786 km [48] above the Earth's surface. Using the values above we can also check which near-field region a RFI source resides. We will use the boundary for electromagnetically long antennas as $D > \lambda$. In our case it is the following:

$$\begin{aligned} R &= 0.62\sqrt{\frac{D^3}{\lambda}} \\ &= 992 \text{ km} \end{aligned} \tag{2.47}$$

From these two calculations we can see that all satellites in geosynchronous orbit or lower fall in the near-field. The GNSS satellites all fall into the radiative near-field and Iridium, the telecommunications satellites, fall just within the reactive near-field. We will only be taking into consideration the curvature of a wavefront due to being in the near-field. There exist other further complications such as the primary beam shape in the near-field but we will not take this into account.

2.5.4 Curved Wavefronts

The visibility contribution due to a satellite therefore does not follow the vCZ theorem and must be reformulated. This is because the vCZ theorem

[5] In the case of astronomy we only have a receiving antenna.

is founded on the assumption that sources lie in the far-field. We start with the most basic case of a point source radiating monochromatic EM waves isotropically with a power output of P. Let the positions of the RFI source and the antenna be $\boldsymbol{x}$ and $\boldsymbol{u}$ measured in the same reference frame. The measured intensity/flux density at a distance $R = |\boldsymbol{u} - \boldsymbol{x}|$ away from this source will then be $I = P/4\pi R^2$. We know that $I \propto |E|^2$ so the electric field strength of the EM wave $|E| \propto \sqrt{I}$. The EM wave can be described by a phasor as

$$E(\boldsymbol{x}, \boldsymbol{u}, t) = |E(\boldsymbol{x}, \boldsymbol{u})| \exp\left[-2\pi i \left(\frac{|\boldsymbol{u} - \boldsymbol{x}|}{\lambda} - \nu t \right) \right] \tag{2.48}$$

The positions of antennas j and k will be given by $\boldsymbol{u}_j$ and $\boldsymbol{u}_k$ respectively. We can now cross correlate, with a time delay τ_0, the electric field strength at each antenna to obtain our visibility.

$$V_{jk} \propto \left\langle E_j(t - \tau_0) E_k^*(t) \right\rangle \tag{2.49}$$

$$V_{jk} = \sqrt{I_j}\sqrt{I_k} \exp\left[-\frac{2\pi i}{\lambda} \left(|\boldsymbol{u}_j - \boldsymbol{x}| - |\boldsymbol{u}_k - \boldsymbol{x}| - w_{jk} \right) \right] \tag{2.50}$$

Where I_j and I_k are the spectral flux density at antennas j and k respectively. w_{jk} is the w-component of the baseline vector $\boldsymbol{u}_{jk}$ which comes from $\nu\tau_0 = \boldsymbol{u}_{jk}^T \boldsymbol{l}_0/\lambda = w_{jk}/\lambda$ since $\boldsymbol{l}_0 = (0, 0, 1)$. We see that the magnitude of the visibility is the geometric mean of the intensity values at each antenna.

Another way to motivate this result would be to consider that the power received at the telescope is proportional to r^{-2}, where r is the distance between the telescope and the RFI source. The amplitude of our result is nothing but the geometric mean of the received power at each antenna. This provides us with the expected amplitude of the visibility up to a real-valued proportionality constant. To obtain the phase we can refer back to how we arrive at the value of τ in section 2.2.1. We simply look at the path length difference and divide by the speed of light. The path length difference is nothing more than the difference of the distances between the RFI source and the antennas, $|\boldsymbol{x} - \boldsymbol{u}_j| - |\boldsymbol{x} - \boldsymbol{u}_k|$.

$$\tau = \frac{|\boldsymbol{u}_j - \boldsymbol{x}| - |\boldsymbol{u}_k - \boldsymbol{x}|}{c} \tag{2.51}$$

Now taking the phase delay given by $\tau - \tau_0$ as is done in Section 2.2.3 one can define a new geometric phase delay term for near-field sources as:

$$\boldsymbol{K}_{jks} = \boldsymbol{I} \exp\left[-\frac{2\pi i}{\lambda} \left(|\boldsymbol{x}_s - \boldsymbol{u}_j| - |\boldsymbol{x}_s - \boldsymbol{u}_k| - w_{jk} \right) \right] \tag{2.52}$$

We can now replace the phase delay term of any near-field source in an individual RIME with the one given in Equation 2.52. To implement this, one would need to have the position of the near-field source and the individual antennas in the same reference frame and in 3D so that the distance between the source s and antenna j, given by $R_{js} = |\boldsymbol{x}_s - \boldsymbol{u}_j|$, can be calculated.

We can also define a brightness matrix for an isotropically emitting RFI source where $I = P/4\pi R^2$ and has some particular polarization $\pm Q$, $\pm U$, or $\pm V$. So for example when using linear feeds the brightness matrix of a RHCP source, where $I = -V$, would be:

$$\boldsymbol{B}_{jks} = \frac{1}{2} \begin{pmatrix} I + Q & U + iV \\ U - iV & I - Q \end{pmatrix} \tag{2.53}$$

$$= \frac{1}{2} \begin{pmatrix} P_s/(4\pi R_{js} R_{ks}) & -iP_s/(4\pi R_{js} R_{ks}) \\ iP_s/(4\pi R_{js} R_{ks}) & P_s/(4\pi R_{js} R_{ks}) \end{pmatrix} \tag{2.54}$$

$$\tag{2.55}$$

This is a baseline dependent brightness matrix since the intensity at each antenna is dependent on the distance of that antenna to the RFI source. In Equation 2.53 R_{js} and R_{ks} are the distances from antennas j and k respectively to the RFI source labelled s, and P_s is the power emitted by the RFI source s.

In conclusion we have found that to include a near-field source in a given RIME calculation that one needs to augment the phase delay term to the one given in Equation 2.52 and the brightness matrix to that given in Equation 2.53.

Chapter 3

Methods

Our goal is to simulate realistic visibilities as measured by the MeerKAT radio telescope including RFI. Currently MeerKAT is operating with UHF and L band receivers. The UHF-band receiver works in the 580–1015 MHz range and the L-band receiver works in the 856-1712 MHz range. MeerKAT is composed of 64 dishes with two orthogonal (horizontal and vertical polarizations) feeds each. We are looking to include an astronomical sky model, a RFI model, the telescope array geometry and location, the primary beam, the bandpass, and time variation in the gains. The final simulator should accept a right ascension, a declination, an observation time, and observation length at the minimum. It can potentially accept further parameters to control the problem size such as reducing the frequency channels or number of sources. In this chapter we will put together the required elements to simulate realistic visibilities contaminated by RFI.

3.1 Astronomical Sky Model

An astronomical sky model consists of a few fundamental things. It must define an intensity distribution over the entire celestial sphere. Depending on the complexity of the intended simulations, more information may be needed. When using the RIME one is building the sky's intensity distribution from point sources. One can discretize the sky and include every point in the sky or only include the points that will meaningfully contribute to the received signal. We will use the latter approach as it is much more efficient thanks to astronomical source catalogues. The RIME uses polarization information so obtaining all four Stokes parameters is desired. A minimum requirement here is the total intensity Stokes parameter.

3.1.1 Catalogue Sources

The NRAO[1] VLA[2] Sky Survey (NVSS), [20], and the Sydney University Molonglo Sky Survey (SUMSS), [40], were both used to provide source intensities, positions and source shapes. The brightness of each source was measured at 1.4 GHz and 843 MHz for these surveys respectively.

The NVSS catalogue was obtained from the NRAO's anonymous FTP server[3] and the SUMSS catalogue from the University of Sydney's webpage[4]. Both these catalogue text files were processed using Python into separate Pandas[5] `DataFrames`. A `DataFrame` is a table-like structure that can be saved as a comma separated value (CSV) file for use elsewhere. NVSS covers the sky north of -40° declination and SUMSS covers the sky south of -30°. The NVSS catalogue was reduced such that it only had sources north of -30° and perfectly complemented the SUMSS catalogue with no overlap. All of the necessary data columns were available in both catalogues, however, for many sources in the NVSS catalogue the position angle was absent. The missing position angles were replaced with a random angle drawn from a uniform distribution on the interval 0 to 180 as they are in degrees. The catalogues were finally joined together into a single `DataFrame` and saved to CSV[6]. A statistical summary of the final sky model data is shown in Table 3.1 below.

	Right Ascension (RA) [deg]	Declination (DEC) [deg]	Intensity [mJy]	Major Axis [arcsec]	Minor Axis [arcsec]	Position Angle [deg]	Spectral Index
mean	179.341	5.883	26.056	56.199	39.340	89.939	-0.721
std	105.004	35.568	218.543	32.173	17.089	52.468	0.310
min	0.000	-88.990	2.000	13.900	13.900	0.000	-2.081
25%	87.040	-21.064	3.400	31.300	24.400	44.270	-0.930
50%	179.021	4.751	6.700	51.200	39.700	89.886	-0.720
75%	272.280	32.534	1.740	71.100	51.100	135.678	-0.511
max	359.998	89.819	342000.0	1156.600	839.000	180.00	0.591

TABLE 3.1: A summary of the Sky model catalogue that was put together from the SUMSS and NVSS catalogues available online. The total catalogue length is 1 838 394 sources.

[1] National Radio Astronomy Observatory
[2] Karl G. Jansky Very Large Array
[3] https://www.cv.nrao.edu/nvss/anonftp.shtml
[4] http://www.astrop.physics.usyd.edu.au/sumsscat/
[5] Pandas is a Python package that allows for the manipulation of table-like data programmatically.
[6] https://github.com/chrisfinlay/RFIsim/blob/master/utils/astronomical/catalogues/SUMSS_NVSS_Cle

3.1.2 Spectral Profiles

Currently the MeerKAT telescope is observing in the L and UHF bands. We are interested in simulating visibilities for the L-band receiver (856-1712 MHz). Over this frequency range, for extragalactic sources with non-inverted or peaked spectra, the spectral profile is well approximated by a power law. Equation 2.4 shows the form of a power law with the use of a spectral index, α. NVSS and SUMSS cover complimentary portions of the sky at a single frequency each. The coverage overlap of these surveys could allow one to calculate spectral indices of some sources, however, this would not necessarily be a good approximation to the entire distribution. The aim of these simulations is to produce realistic visibilities that are representative of true visibilities. It is therefore not necessary to use the true spectral index for each source as a representative sample will suffice. In [70] spectral indices were measured for 107,488 sources. The surveys used in [70] had reference frequencies between 30 MHz and 15 GHz, easily covering the L-band. Using the database we found that the mean spectral index is $\mu_\alpha = -0.72$ with a standard deviation of $\sigma_\alpha = 0.31$. Assuming that the distribution of spectral indices is approximately normal, we can draw samples for each source from a normal distribution with the mean and standard deviation just mentioned. This is what was done for each source. Equation 3.1 below shows how the entire spectrum is generated from a spectral index and a reference measurement such as is provided by NVSS and SUMSS. A reference measurement in this case is a measured spectral flux density at a specific frequency. ν_0 is called the reference frequency here. I_{ν_0} is the spectral flux density at the reference frequency.

$$I_\nu(\nu) = I_{\nu_0}\left(\frac{\nu}{\nu_0}\right)^\alpha \tag{3.1}$$

3.2 Radio Frequency Interference

3.2.1 Satellites

Two-Line Element sets (TLEs) are a data format used for encoding orbital elements of any Earth-orbiting object. TLEs are used in conjunction with simplified perturbations models such as SGP, SGP4, SDP4, SGP8, and SDP8 to predict the position and velocity of an Earth-orbiting object, [30]. The United States Air Force (USAF) tracks all Earth-orbiting objects (that it can detect) and makes corresponding TLE files available to the public through the website Space Track[7]. Of course it does not make every single object's TLE available as some of these are of military or a classified nature.

[7] https://www.space-track.org/

There are a number of Python packages available for using these TLEs to predict satellite paths. At the time of developing the code for the first version of RFIsim the *de facto* Python package was named `PyEphem` [52]. This package uses older prediction algorithms from the 1980s. Its successor, `Skyfield` [53], is more modern, having higher precision and was used in the second version of RFIsim.

3.2.2 Ground Based RFI

A minimal effort was made to include ground based RFI. Looking at the location of the MeerKAT telescope with Google Maps, we identified the four closest towns as Carnarvon, Vanwyksvlei, Williston and Brandvlei. The GPS locations for each of these towns were used with only a single source of RFI in each. These are included to represent various sources of RFI that may be present in populated regions. An area map is included in Figure 3.1 for reference.

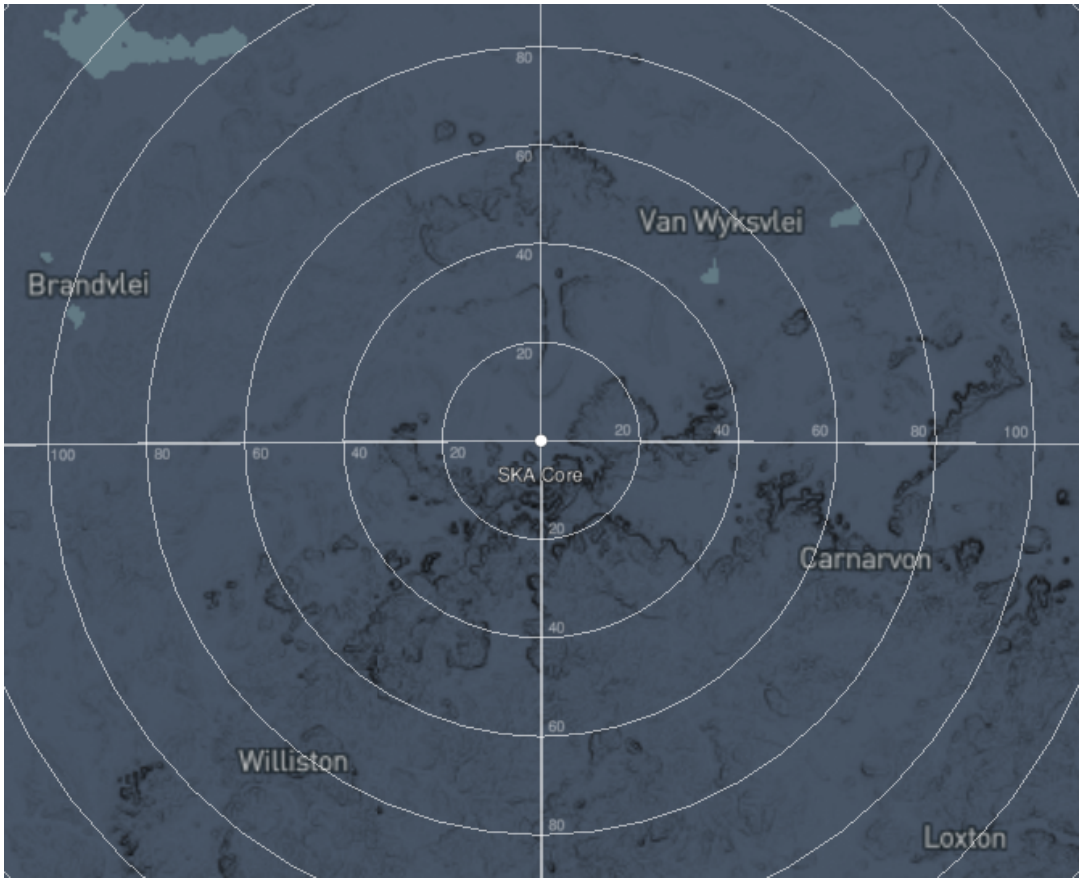

FIGURE 3.1: This is a map of the local area around the MeerKAT telescope site. The four closest towns are shown in the map. The white circles show radial distance bands, separated by 20km spacings, from the MeerKAT site.

Source: ComRADGIS [17]

3.2.3 Signal

When considering the signal emitted by RFI sources one must look at its spectral profile, polarization and its time variation.

Firstly we will need to chose a power output from the RFI sources in terms of Janskys. In [16] the conversion from dBW to Jy is described. Using this same logic we can calculate some estimates for the expected flux density from a range of RFI sources. The minimum received power from a GNSS satellite on the surface of the Earth is in the range -164.5 to -155 dBW [8, 5, 6, 7]. Using this we find the expected Jansky value to be in the range of 7.5×10^4 Jy to 1.5×10^6 Jy. This provided an initial range which was fine tuned by running simulations with varying RFI intensity to compare with real data. The following probabilistic model, given in equation 3.2, was used to provide a wide range of possible RFI intensities that include the example range given.

$$I_s = 10^{\text{Uniform}[4,7]} \tag{3.2}$$

Uniform$[a, b]$ is a uniform distribution over real values in the interval a to b including the boundaries. The model in equation 3.2 produces a uniform distribution over the exponents.

A project by the name of Historical Probabilities of RFI (HPRFI) [59] led by Isaac Sihlangu of the Science Data Processing (SDP) team at SARAO looked at the occurrence frequency of RFI flags over a number of dimensions including the frequency axis. This provided a probability distribution over frequency for where RFI may occur. Samples from this distribution were stored in a separate file. For each RFI source a random sample was drawn to specify its central emission frequency. Figure 3.2 shows this probability distribution.

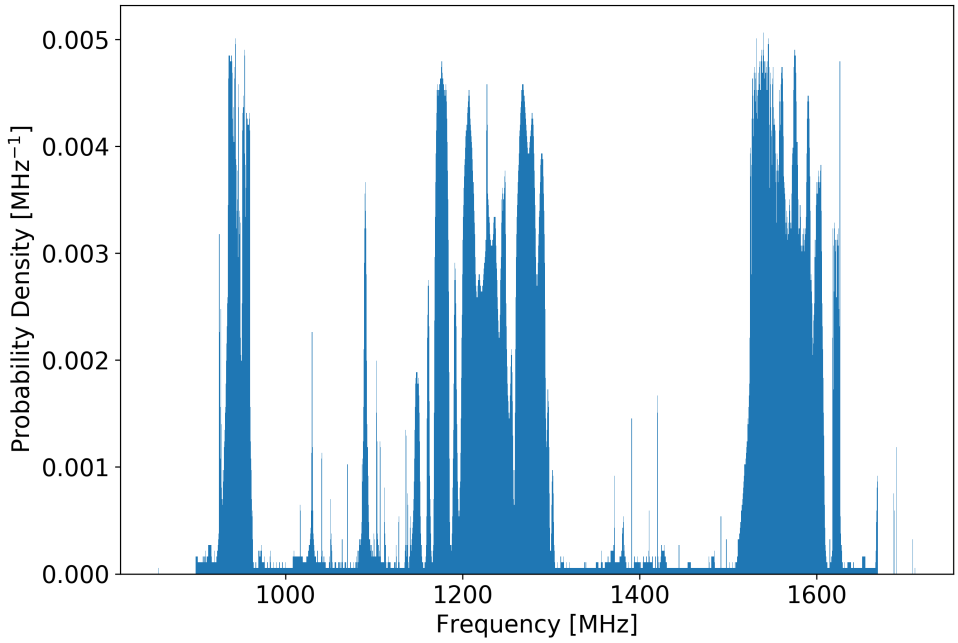

FIGURE 3.2: This is a normalized histogram of RFI flagging occurrences for the MeerKAT telescope in the L-band. This histogram is calculated by counting the number of times a frequency channel is flagged as RFI contaminated by the different flaggers present in the SDP data reduction pipeline at SARAO.

Source: [59]

In Section 2.5, we saw the different frequency bands that are occupied by known RFI sources. We can see that these are some of the most strongly populated in Figure 3.2.

Once a central emission frequency was chosen for each RFI source, a spectral profile was generated for each. The following probabilistic model was used to produce a wide variety of spectral profiles. This was done to avoid simple filtering algorithms being able to take advantage of a limited spectral profile dictionary. These profiles were considered normalized and then multiplied randomly drawn intensity as specified in equation 3.2.

$$I_\nu(\nu) = \sum_{i=1}^{N} A_i \exp\left[-\frac{1}{2}\left|\frac{\nu - \nu_i}{\sigma_i}\right|^{n_i}\right] \tag{3.3}$$

We have five random variable objects in 3.3, these are the number of communication channels N, the emission power of each channel A_i, channel position ν_i, pseudo-channel width σ_i, and the channel profile shape number n_i. Each random variable was assumed independent[8] and the probability distribution for each is listed below.

$$N \sim \text{UniformInteger}[2, 5] \tag{3.4}$$
$$A_i \sim \text{Uniform}[0.7, 1.3] \tag{3.5}$$
$$\nu_i \sim \text{Uniform}[\nu_0 - 4\text{MHz}, \nu_0 + 4\text{MHz}] \tag{3.6}$$
$$\sigma_i \sim \text{Uniform}[0, 700\text{KHz}] \tag{3.7}$$
$$n_i \sim \text{Uniform}[1, 5] \tag{3.8}$$

UniformInteger$[a, b]$ is a uniform distribution over the integers between and including the bounds a and b. Sampling from the above probabilistic model generates a mutli-channel spectrum for a single RFI source. This is done for each RFI source present in the simulation. Each channel has a profile that can take on a range of shapes. Both exponential and gaussian profiles are included in the range of shapes possible. This makes it somewhat realistic however not so simple as to be able to be filtered out with ease. Figure 3.3 shows 3 example spectral profiles generated using this model.

[8] In reality this would certainly not hold true for the channel positions ν_i as channel overlap would be avoided for a single RFI source.

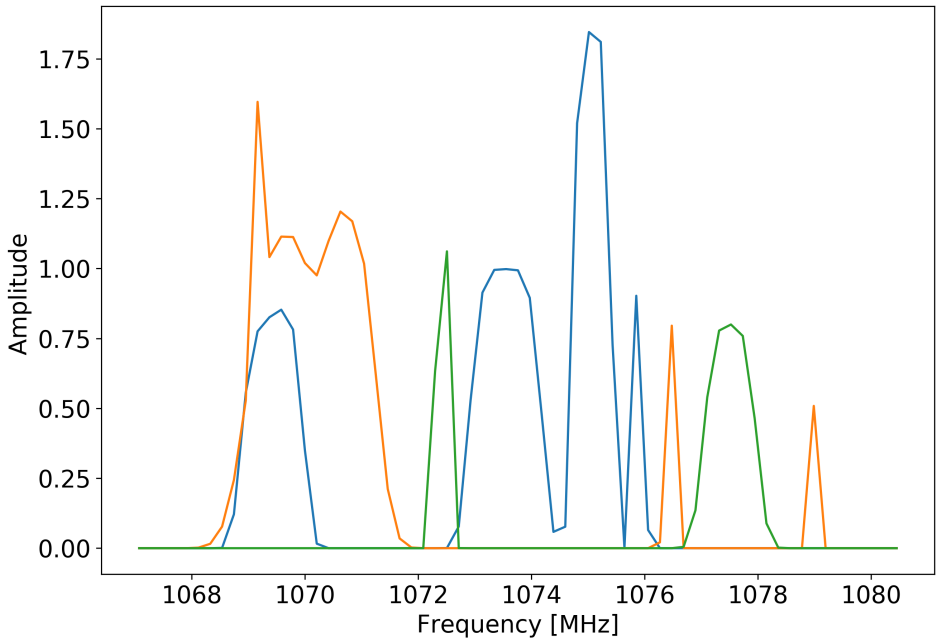

FIGURE 3.3: Here we have three example spectra that are drawn from the probabilistic model described in this section. The amplitudes here are unitless and will be multiplied by an intensity chosen according to equation 3.2. These examples show a range of channel numbers and the variability in channel shape.

In Section 2.5.2 we saw that GNSS satellites and some telecommunications satellites emit right hand circularly polarized (RHCP) radiation. In the first version of RFIsim all RFI sources were chosen to be fully polarized, however, the type of polarization was chosen at random between $\pm$Q, $\pm$U and $\pm$V. In the second version of RFIsim all satellite based RFI sources are chosen to be fully polarized with RHCP.

With regard to the time variability of each RFI source, a model very similar to that described in Section 3.5.2 below was used. It is a model that comprises of sampling sinusoids with random amplitudes, periods, and phases from specific ranges. Equation 3.9 shows the probabilistic model used. Each variable was assumed independent.

$$I_s(t) = \sum_i^N A_i \cos\left[\frac{2\pi t}{T_i} + \phi_i\right] \tag{3.9}$$

where $A_i \sim \mathrm{U}[0,1]$, $T_i \sim \mathrm{U}[1,10]$ (T_j is in seconds here) and $\phi_i \sim \mathrm{U}[0,2\pi]$. Here $\mathrm{U}[a,b]$ represents an uniform distribution on the interval a to b. The number of sinusoids N is selected by the user providing a very flexible model

for amplitude variation. In the simulator a default of 20 was used for N. 3The resulting time series is then scaled and shifted such that it lies between zero and one. Figure 3.4 shows three example time series.

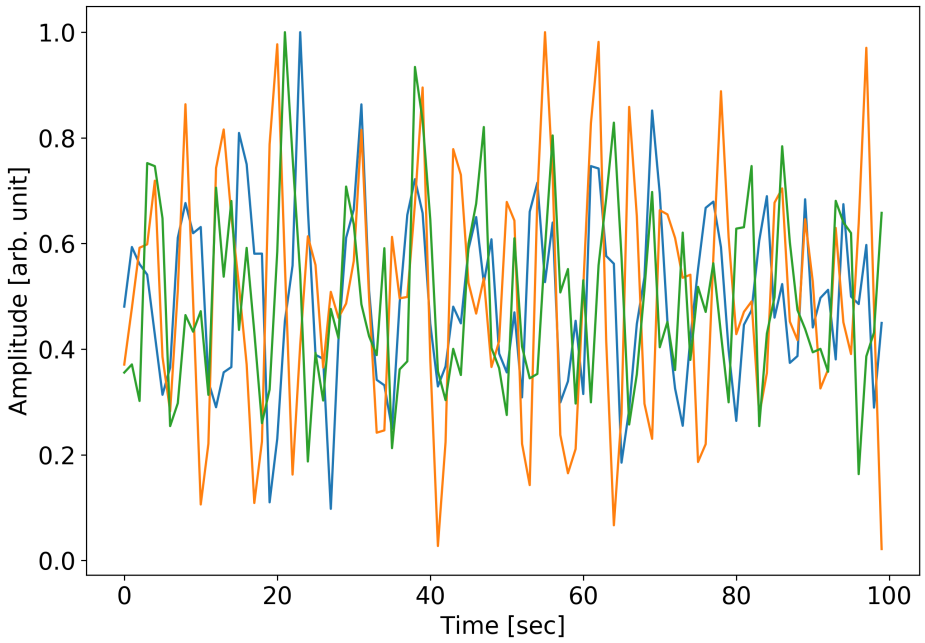

FIGURE 3.4: Here we have three example time series that are drawn from the probabilistic model described in Equation 3.9. Each time series is then scaled and shifted to lie in the range of zero to one.

The final signal for an RFI source is modelled by taking the product of the overall amplitude from Equation 3.2, the frequency dependent component from Equation 3.3 and the time dependent component from Equation 3.9.

3.3 Array Configuration

Since we are interested in simulating visibility data for the MeerKAT telescope we used the antenna positions for the entire 64-dish array. Figure 3.5 shows the antenna positions, relative to a reference antenna, for all 64 dishes.

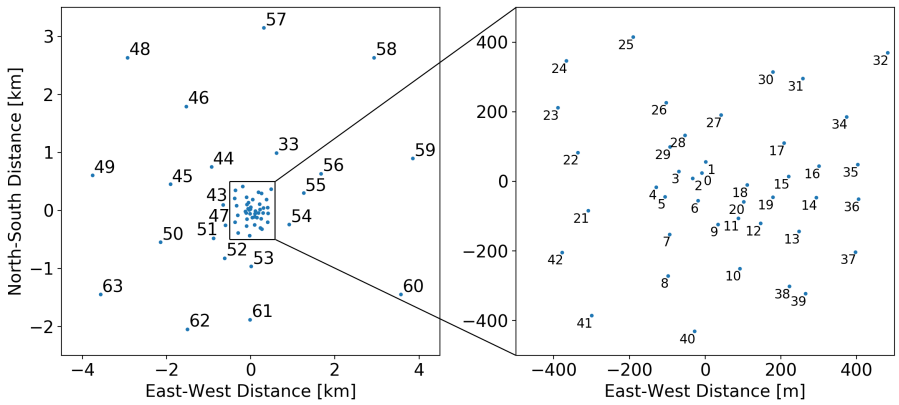

FIGURE 3.5: Antenna positions for all 64 MeerKAT dishes. The left panel shows all 64 antennas. The right panel is a zoomed in plot of the section shown in the left panel. The antennas in the right panel are part of what is called the core of the array.

A GitHub package exists to produce *uvw*-coordinates for the MeerKAT array. The package we used is named `uvgen`[9] and was created by Sphesihle Makhathini working under the Radio Astronomy Research Group (RARG) at the South African Radio Astronomy Observatory (SARAO). This package includes the GPS location of the MeerKAT array, 30°43'15.6"S - 21°24'39.6"E, and each antenna position relative to a reference antenna. Given a date, time, observation length, time step and pointing direction on the celestial sphere it is able to produce a set of *uvw*-coordinates for each point in time, over the intended observational period. An example set of *uv* tracks is shown in Figure 3.6.

[9] https://github.com/SpheMakh/uvgen

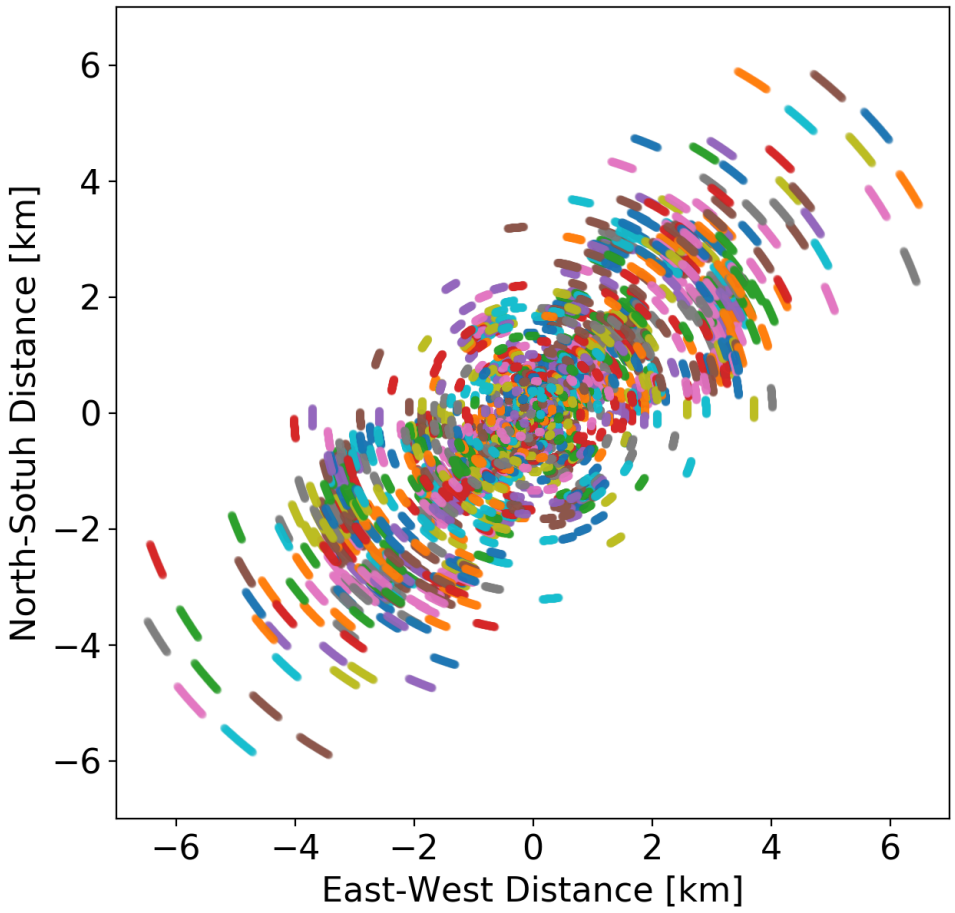

FIGURE 3.6: *uv* tracks for a 20 minute observation with 8 second integration times on the 64-dish MeerKAT telescope. Each colour track is a different baseline. The pointing direction (phase centre) is at the south celestial pole, this leads to circular tracks centred around the origin as seen here.

3.4 Direction-Dependent Effects

In section 2.4.3 a number of DDEs were discussed. In our simulations the only DDE included is that of the primary beam. In order to do this a primary beam model was needed.

3.4.1 Primary Beam

In [3] the primary beam of a MeerKAT antenna is measured and Zernike polynomials are fitted out to 5 degrees from the centre. All four polarizations of the primary beam were measured and subsequently fitted. The resulting model is accessed through a software package named `eidos`[10]. This model was used to create a beamcube[11] that could be used as data to fit a model to.

RFI signals are usually orders of magnitude more intense than astronomical signals. Because of this RFI that enters the primary beam in its far sidelobes still contributes significantly to the overall signal received. Since we are interested in simulating the effect of RFI we need to have a beam model that reaches out much further than 5 degrees. As a result a basic model was fitted to the beamcube, allowing us to evaluate the beam as far out as necessary, instead of using the Zernike polynomial representation directly. The basic fitted model was evaluated out to 30°[12] when being used as a beamcube for simulations. The goal of RFIsim is to be realistic enough to produce all of the correct characteristics of RFI in visibility data. It is therefore enough to use a basic theoretical model that has sidelobes where the RFI can pass over.

Theoretically, for a circular aperture with uniform illumination across the entire aperture, the diffraction pattern is an Airy disk. The analytical form of the normalised Airy disk with its angular and frequency dependence is as follows in Equation 3.10.

$$I(\theta, \nu; d) = \left[\frac{2J_1\left(2\pi\nu d \sin\theta / c\right)}{2\pi\nu d \sin\theta / c} \right]^2 \tag{3.10}$$

where J_1 is the Bessel function of the first kind, d is the diameter of the dish/aperture, ν is the observing frequency and θ is the angle from the pointing centre. It is therefore circularly symmetric and in terms of direction cosines, (l, m), we have $\sin\theta = \sqrt{l^2 + m^2}$.

[10] https://github.com/ratt-ru/eidos

[11] A beamcube is a 3-d array of primary beam attenuation values for the two direction coordinates, l and m, and along the frequency axis.

[12] This value was chosen due to computational constraints that are described in section 3.7.2

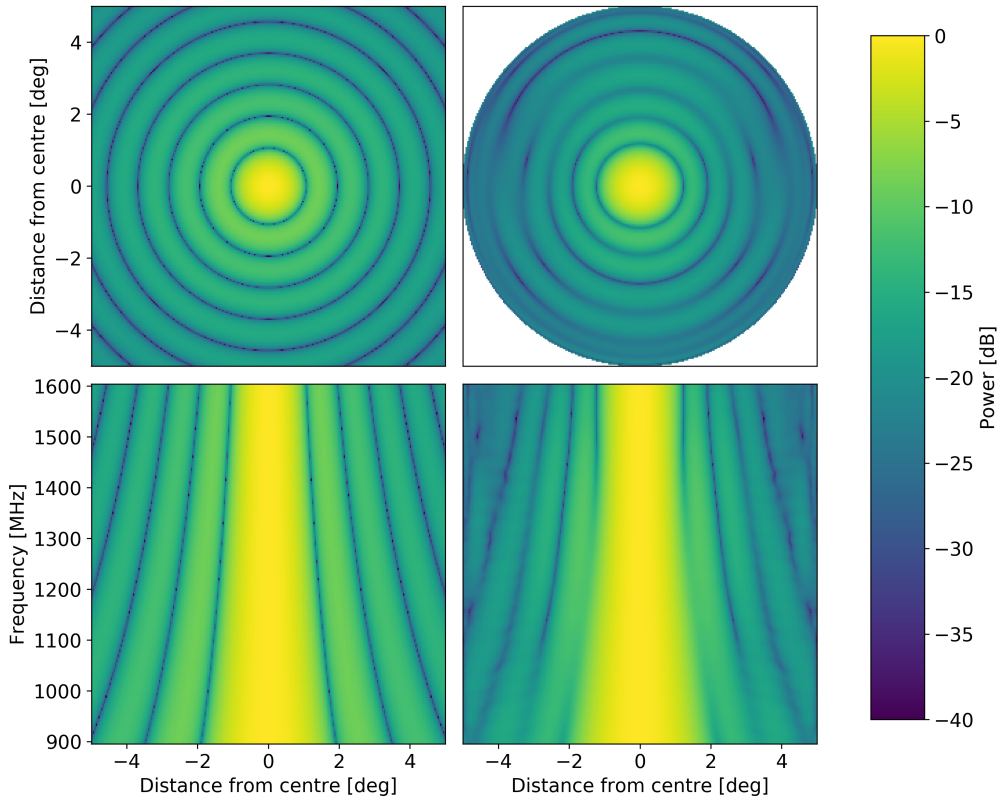

FIGURE 3.7: Zernike polynomial representation of the magnitude of the MeerKAT primary beam in the HH-polarization is shown on the right. Airy disk fit to the Zernike polynomial representation of the MeerKAT primary beam in the HH-polarization is shown on the left. The upper panels show the spatial dependence of the beam. The lower panels show a cut through the centre of the beam over the frequency range 900 MHz to 1600 MHz. We found that for both the Zernike polynomial representation and the Airy disk fit that the main lobe width increases with decreasing frequency as expected.

An Airy disk was fitted to the HH component of the Zernike polynomial beamcube where the dish diameter d was allowed to vary. The fitted function was then used as the HH and VV components of the primary beam for RFIsim. A comparisons of the fitted HH beam and the Zernike polynomial beam data are shown in Figures 3.7 and 3.8. Figure 3.7 shows 2D comparisons across the lm-plane at $\nu = 1604$ MHz and $l\nu$-plane at $m = 0$. Figure 3.8 shows a 1D slice at $\nu = 1305$ MHz with the fitted model overlaid on the Zernike polynomial beam data.

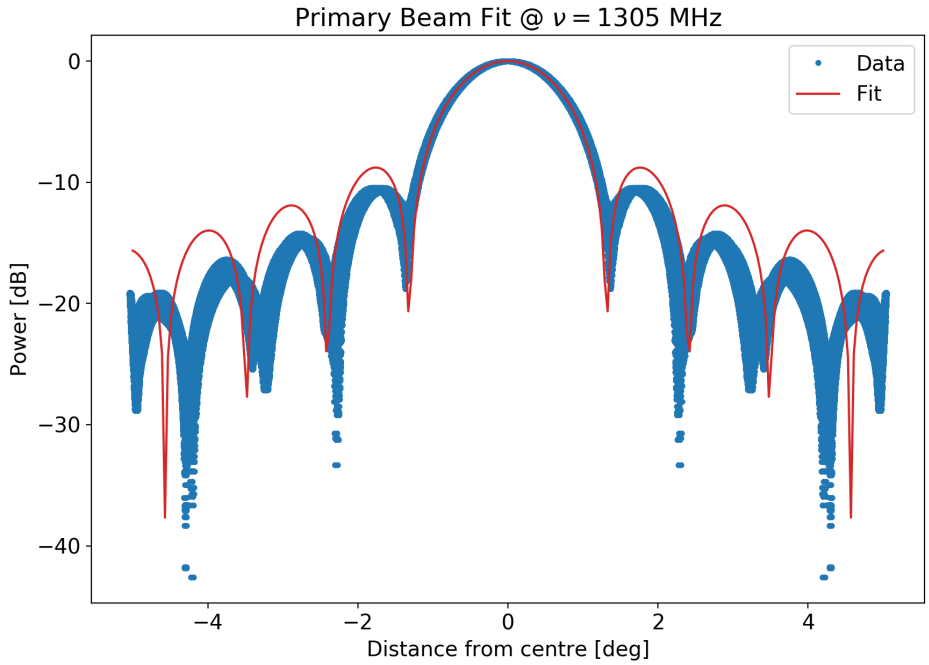

FIGURE 3.8: Radial profile of the Zernike polynomial representation of the MeerKAT primary beam in the HH-polarization at a frequency of 1305 MHz is shown as blue dots. The Airy disk fit is shown as the red line. The Airy disk model manages to fit the main lobe of the beam very well however the sidelobe levels are too large in the model. The spacing of the sidelobes in the model do not match that of the data well either. For our purposes these are not serious issues as we are interested in the general behaviour of RFI emission entering through the sidelobes.

Equation 3.10 applies specifically to the HH and VV polarizations. Theoretically the HV and VH components should be zero everywhere as the H, x, and V, y, components of the incoming EM-wave are independent. In practice the signal measured at each feed is not independent due to polarization leakage described in Section 2.4.3.1 leading to non-zero cross-polarization components, HV and VH. The H and V receivers themselves are not identical which leads to potentially different beam shapes and gains for the HH and VV auto-polarization components.

An attempt was made to create an analytical model that would fit well to the cross-polarization components. The best of these models is shown in Equation 3.11.

$$I(l,m,\nu;I_0,\beta,\sigma) = I_0 lm \frac{J_1\left(\beta\nu\sqrt{l^2+m^2}\right)}{\beta\nu\sqrt{l^2+m^2}} \exp\left[-(l^2+m^2)/2\sigma^2\right] \tag{3.11}$$

Equation 3.11 shows a model for the real or imaginary part of the cross-polarization beam components and is parameterized by I_0, β and σ. I_0 controls the overall amplitude and polarity of the quadrupole (clover leaf shape). β controls the beam of the lobes and σ controls the width of the overall gaussian envelope. An example of the cross-polarization model is shown in Figure 3.9 along with some examples of the data. Unfortunately this model could not be reliably fit to the cross-polarization components using standard optimization techniques. A more flexible model is potentially needed.

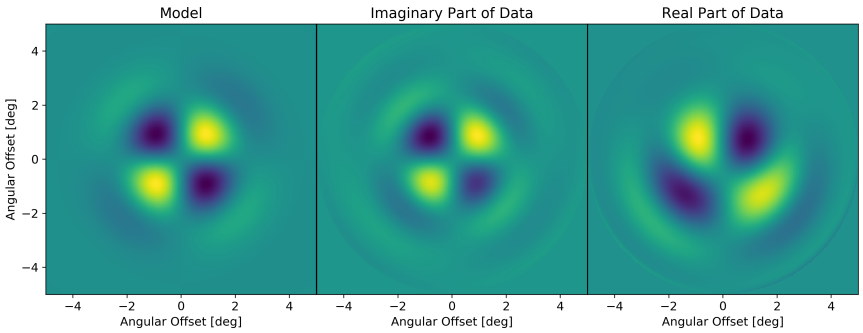

FIGURE 3.9: The left hand panel shows an example of the model in equation 3.11 with the parameter values (chosen by eye) of $I_0 = -1$, $\beta = 0.0716$ and $\sigma = 0.0262$. The middle and right-hand panels show the imaginary and real components of the Zernike polynomial data, from [3], respectively. All three panels are shown for a frequency of 1221 MHz. From this Figure we can see that the cross-polarization model works well in some cases (the imaginary component in this case) and badly in others (the real component in this case).

3.5 Direction-Independent Effects

DIEs can vary over any number of things that are not direction dependent as the name suggests. We will look to create a model for gains that vary over polarization, frequency, time and antenna. We will assume that the matrix valued gains function, $\boldsymbol{G}_i(\nu, t)$ for antenna i is separable as $\boldsymbol{G}_i(\nu, t) = \boldsymbol{G}_i(\nu)g_i(t)$. The frequency dependent part is referred to as the bandpass. This section is broken down into obtaining the bandpass for all polarizations and each antenna and then a time dependent part for each antenna.

3.5.1 Bandpass

To obtain realistic DIEs (for MeerKAT simulations) a measurement of the MeerKAT L-band bandpass was needed. To do this, a real observation of

a calibrator source (PKS J1939-6342) from the SARAO archive was used. A spectral model for this source is obtained from [50]. The spectral model is given by equation 3.12.

$$\log_{10} S = -30.767 + 26.491 \log_{10} \nu - 7.0977(\log_{10} \nu)^2 + 0.60533(\log_{10} \nu)^3 \tag{3.12}$$

Where S is the spectral flux density in units of Jy and ν is the frequency in units of MHz.

To obtain the gains from a calibrator observation we can look to the RIME to form a basic model. The calibrator source is ideally chosen to be a bright point source with no other bright sources nearby. This is needed for a high signal-to-noise ratio (SNR) and to have an accurate spatial model. The singular point source nature allows us to ignore any primary beam effects by pointing all the antennas directly at the source (such that the primary beam's maximum is directly on the source).

We have an unpolarized source with spectral flux density I_ν sitting at the phase centre so its position in the lm-plane is $(0, 0)$. At the phase centre $\boldsymbol{K}_i\boldsymbol{K}_j^\dagger = \boldsymbol{I}$ for all baselines ij. Additionally with the antennas pointed directly at the source (the maximum of the primary beam) we have $\boldsymbol{E} = \boldsymbol{I}$, because the beam is normalised. Let $\boldsymbol{G}_i$ be the entire 2x2 DIE term for antenna i. Our brightness matrix will be the identity multiplied by $I_\nu/2$. We are assuming an unpolarized source and therefore the remaining Stokes parameters will be 0. Equation 3.15 shows the calculation using the RIME for the auto-correlations (correlation of an antenna with itself).

$$\boldsymbol{V}_{ii} = \boldsymbol{G}_i\boldsymbol{E}_{is}\boldsymbol{K}_{is}\boldsymbol{B}_s\boldsymbol{K}_{is}^\dagger\boldsymbol{E}_{is}^\dagger\boldsymbol{G}_i^\dagger \tag{3.13}$$

$$\boldsymbol{V}_{ii} = \frac{I_\nu}{2}\boldsymbol{G}_i\boldsymbol{G}_i^\dagger \tag{3.14}$$

$$\implies \boldsymbol{G}_i = \sqrt{2\boldsymbol{V}_{ii}/I_\nu} \tag{3.15}$$

We can see from equation 3.15 that we need to perform a matrix square root. To do this we can make use of the eigendecomposition [11] of the matrix. Since the auto-correlations $\boldsymbol{V}_{ii}$ are Hermitian matrices (being the result of $\boldsymbol{X}\boldsymbol{X}^\dagger$) we know that its eigendecomposition will be of the following form.

$$\boldsymbol{V}_{ii} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\dagger, \qquad \text{where} \quad \boldsymbol{Q}^\dagger = \boldsymbol{Q}^{-1} \tag{3.16}$$

$$\implies \sqrt{\boldsymbol{V}_{ii}} = \boldsymbol{Q}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{Q}^\dagger \tag{3.17}$$

Because $\mathbf{\Lambda}$ is a diagonal matrix its square root is nothing more than the square root of its elements. Doing this for every frequency channel and antenna separately we can generate the bandpass of every antenna and polarization component. The result is shown in Figure 3.10.

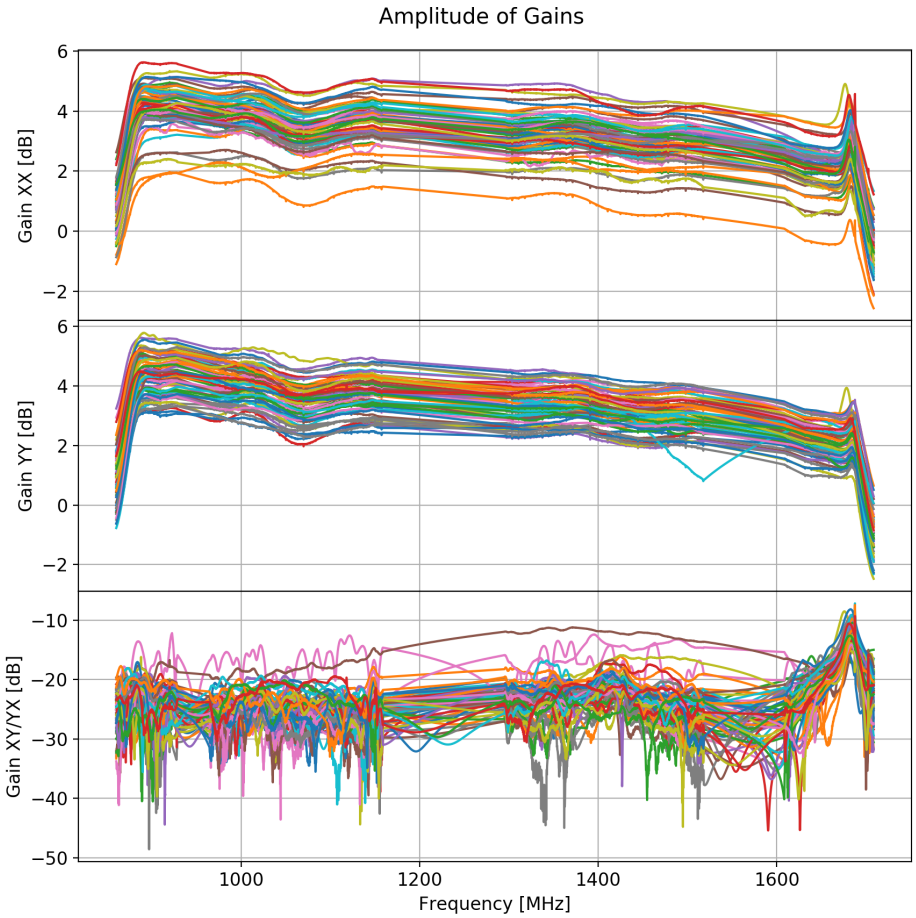

FIGURE 3.10: These are the bandpasses for all 64 MeerKAT antennas. The top panel shows the bandpass magnitude for the HH (xx) component with every antenna as a separate curve. The middle panel shows the magnitude of the VV (yy) component of the bandpass. The bottom panel shows the magnitude of the cross-polarization components of the bandpass. To calculate the bandpass a calibrator observation of PKS J1934-6342 from 23:18 19th April 2018 was used. It was the calibration portion of a target observation that had a 0.25011 Hz sampling rate across 4096 frequency channels for a period of 10 minutes. To achieve these plots the RFI was manually flagged from the visibilities, the missing portions were linearly interpolated and then each bandpass was averaged over the entire 10 minute period. After this the bandpass was calculated using some simple matrix maths outlined in this Section.

3.5.2 Time Dependence

In this section we are interested in creating a realistic model of the gain variation in time. There is no theoretical model to work from here so we will perform an analysis on a calibration dataset and make use of institutional knowledge to develop a basic model that captures the main time dependent features of the gain variations.

We will work from a 30 minute dataset broken into three 10 minute sections over a period of just over 4 hours. These are the calibration portions of a target observation where the calibrator was PKS J0408-6545. Figure 3.11 shows the gain variations, averaged over the frequency range 1.387 GHz to 1.518 GHz, as a percentage away from the time average for each antenna.To achieve these plots the RFI was manually flagged from the visibilities, the missing portions were linearly interpolated across frequency and then the gains for each frequency and time step were calculated using the same matrix math outlined in Section 3.5.1. The result is the matrix valued (polarized) gains across frequency and time. With these two objects, variations over time (from the mean) can be calculated using equation 3.18.

Let $G_i(\nu, t)$ be the gains for antenna i across both frequency and time and $\bar{G}_i(\nu) = \frac{1}{T}\int_0^T G_i(\nu, t)dt$ the time averaged gains i.e the bandpass. Combining these as described in Equation 3.18 we obtain the curves shown in Figure 3.11.

$$\delta G_i(\nu, t) = \frac{G_i - \bar{G}_i}{\bar{G}_i} \tag{3.18}$$

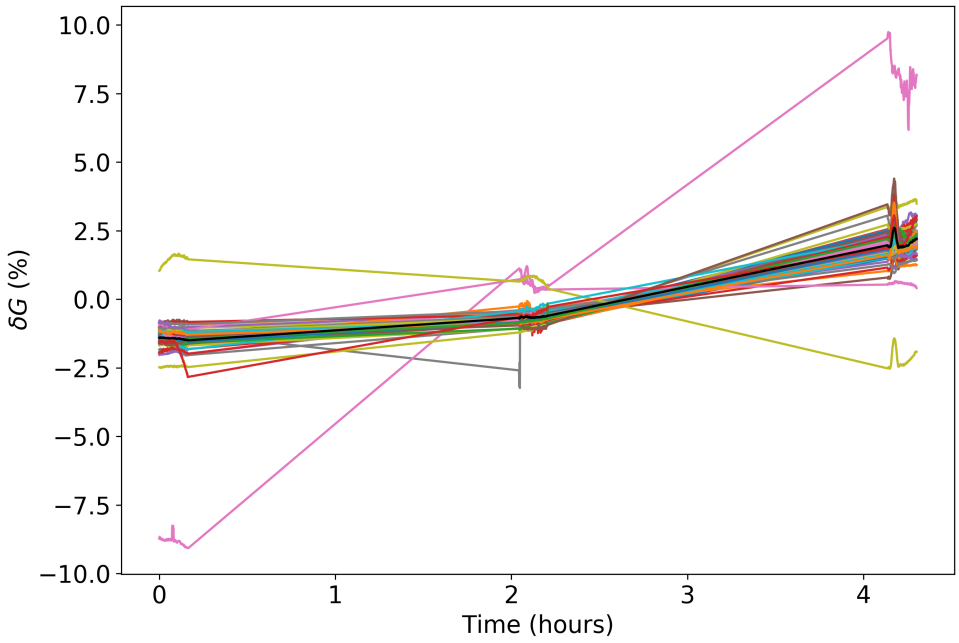

FIGURE 3.11: These are the gain variations over time for all 64 MeerKAT antennas in the HH polarization. Each curve is the result of averaging the bandpass variations from the mean across 630 frequency channels in the range 1.387 GHz and 1.518 GHz. The black curve is the average across all antennas. To calculate the gain variations a calibrator observation of PKS J0408-6545 from 16:15 19th April 2018 was used.

Looking at Figure 3.11 above one could come up with a number of basic models for the time variation such as a Wiener process or Brownian motion with drift. A simple heuristic that is important to consider is that gains should not grow or decay in an unbounded manner as this is not something that happens in reality (Due to gains being adjusted on the fly). With such information one can determine that the models previously mentioned are not suitable when simulating long observations. In light of this we have made use of a combination of many sinusoids. With such a method, random variations (that are bounded) can be generated with different amplitudes on different time scales. From Figure 3.11 we also found that the typical gain variation is less than 5% over a four hour period with the worst case scenario giving variations of 20% on the same time scale. The following Equation 3.19 is used to produce the time dependence of the gains [42] in RFIsim.

$$g_i(t) = 1 + \frac{1}{N}\sum_{j}^{N} A_j \cos\left[\frac{2\pi t}{T_j} + \phi_j\right] \tag{3.19}$$

where $A_j \sim \mathrm{U}[0, 0.1]$, $T_j \sim \mathrm{U}[1, 10]$ (T_j is in hours here) and $\phi_j \sim \mathrm{U}[0, 2\pi]$. More complex versions can be implemented through potentially different probability distributions or even joint distributions between these variables. A new set of variables is drawn for each antenna such that they are independent of one another. A few example curves generated using Equation 3.19 are shown in Figure 3.12.

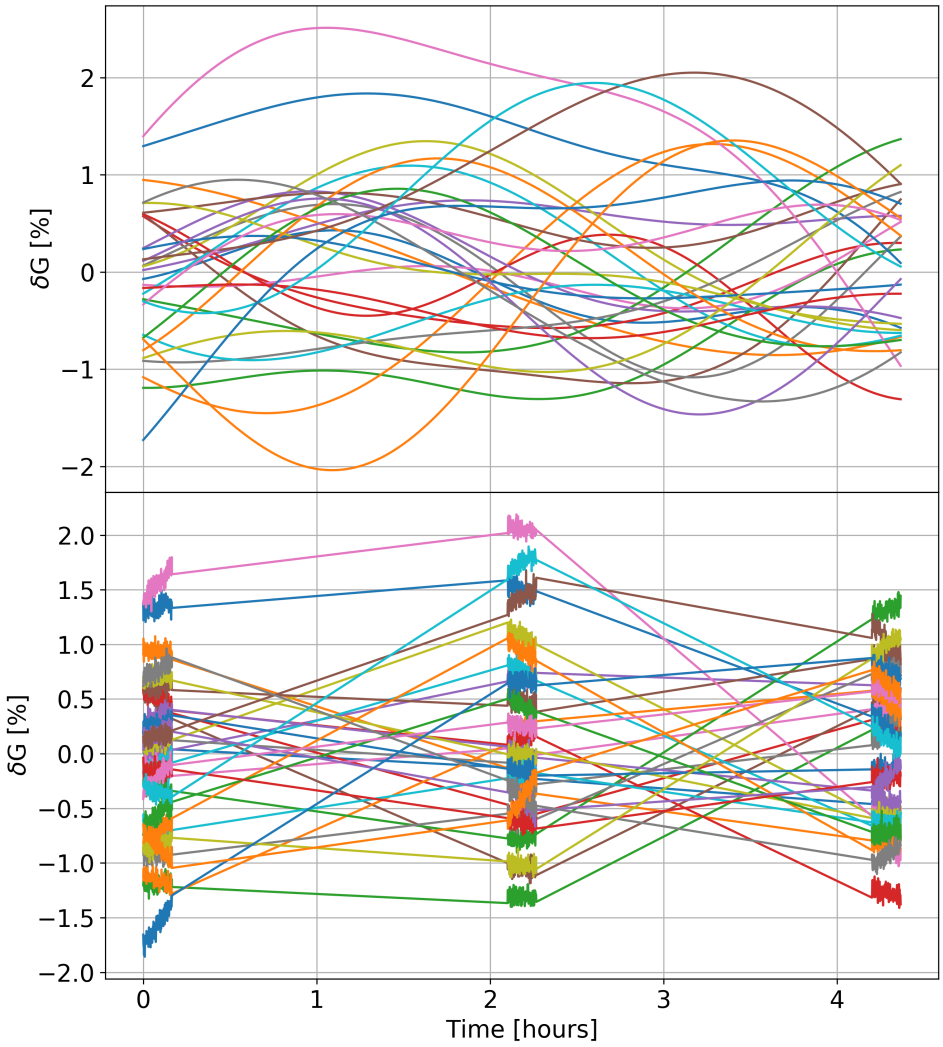

FIGURE 3.12: Here are 32 independent samples drawn from the gain time variation model shown in Equation 3.19. The top panel shows the value of the gain deviations for every time step over the 4.5 hour period. The bottom panel shows the same 32 samples as the top panel however only three 10 minute sections are shown with noise added to show the resemblance with Figure 3.11. For these samples $N = 20$ sinusoids were used. We can see the desired behaviour is present. The gain deviations vary smoothly over time with deviations in this set of samples not exceeding 3%.

3.6 Measurement Set Conversion

In radio astronomy, one of the most commonly used software tools is Common Astronomy Software Applications (CASA), the successor to Astronomical Image Processing System ++ (AIPS++). CASA makes use of a visibility data file format called a Measurement Set (MS). Due to CASA's prevalence in the community it was seen to be of great benefit to write a program that could convert the output of RFIsim to this file type. This allows our simulation data to be subjected to standard data reduction and analysis performed by CASA. The main components of a MS file will be outlined below.

A measurement set is essentially a relational database. It is made up of a set of directories and tables that are stored in a single directory we call a MS file. The `MAIN` table contains the visibility data and is kept in the top directory of the MS file. All metadata for the telescope and observation are stored in separate directories under the top directory. These are called Subtables. When necessary the MAIN table can make reference to these Subtables so as to avoid data redundancy. We will not describe all the possible metadata that can be stored in a MS file as it is a vast list to accommodate nearly any type of radio observation. Please refer to [37] for the full MS version 2.0 description.

The full Python script for converting RFIsim data to MS is given in `https://github.com/chrisfinlay/RFIsim/blob/master/utils/helper/H5_to_MS.py`.

3.7 Montblanc

Montblanc, [45] is a software package written by Simon Perkins, a member of RARG at SARAO. It is an implementation of the RIME that runs on GPUs by splitting up the problem into chunks that fit into the GPU's memory. It has a Python API (Application Programming Interface) to specify a problem. Problem specification is done through a number of different base classes that are used to define data sources and data sinks. Table 3.2 below shows the functions used in this work that form part of the `SourceProvider` class with further explanation in Section 3.7.1. For inclusion of a primary beam model Montblanc offers the `FitsBeamSourceProvider` class which is discussed in Section 3.7.2 below. For getting the output data from Montblanc there is a `SinkProvider` class which is discussed in Section 3.7.3.

Throughout this section and the next the following abbreviations will be used when referring to the size of array dimensions

- N_time : Number of time steps.
- N_ant : Number of antennas.

- N_bl : Number of baselines.
- N_freq : Number of frequency channels.
- N_pol : Number of polarizations. (Always 4, sometimes split into 2x2)
- N_src : Number of sources. This will sometimes be augmented to specify point sources or gaussian sources in Montblanc.

3.7.1 Data Sources

When referring to data sources in the context of Montblanc we are talking about any data that is needed to simulate a specific observation. It is important to note the distinction between data sources, astronomical sources and RFI sources. Data sources are any data needed to run the simulation of which astronomical sources and RFI sources are a subset of. Montblanc is able to calculate visibilities for a selection of source distributions. It is able to handle points sources, gaussian sources, and Sérsic[13] profile sources. Fortunately the sky model used, described in section 3.1, includes gaussian shape parameters for each source. We therefore used these directly in Montblanc. All RFI sources were modelled using point sources. Table 3.2 shows all of the functions used in the data source provider class.

3.7.2 Primary Beam

Montblanc has a class named `FitsBeamSourceProvider`. This class allows for the inclusion of a primary beam model which has been evaluated on a 3D grid across two spatial dimensions and a frequency axis. The Flexible Image Transport System (FITS) is a file type that allows multidimensional arrays to be stored along with metadata in its header. It is a very common file type within astronomy. This is a very convenient data input form as no analytical definition for the beam is needed and therefore a directly measured beam can be used in the simulations.

In order to use a beamcube in RFIsim, Montblanc has to perform 3D interpolation to estimate the beam values at a specific location in space and frequency. It is important to note that because beams tend to have oscillating sidelobes a minimum sampling density is needed for the beam such that the interpolation process does not smooth over the sidelobes. When one wants to include a beam model that reaches out very wide this becomes a memory constraint.

[13] A Sérsic source profile is one where the intensity distribution of the source is described by the equation $\ln I(R) = \ln I_0 - kR^{1/n}$. Here I_0 is the intensity at the centre ($R = 0$), k is a constant and n is called the Sérsic index. The higher the value of n the more centrally concentrated the profile is.

Montblanc Function Summary		
Function Name	Output Dimensions	Description
`frequency`	(N_freq)	The frequency per channel in Hz.
`point_lm`	(N_psrc, 2)	The (l, m) coordinates for each point source.
`point_stokes`	(N_psrc, N_time, N_freq, 4)	The Stokes parameters, (I, Q, U, V), for each point source, time step, and frequency channel.
`gaussian_lm`	(N_gsrc, 2)	The (l, m) coordinates for each gaussian source.
`gaussian_shape`	(N_gsrc, 3)	The shape parameters for each gaussian source. This requires the l-projection and m-projection of the minor axis and the ratio of the major axis over the major.
`gaussian_stokes`	(N_gsrc, N_time, N_freq, 4)	The Stokes parameters, (I, Q, U, V), for each gaussian source, time step, and frequency channel.
`direction_ ... _independent_effects`	(N_time, N_ants, N_freq, 4)	The DIE term for each time step, frequency channel, antenna and polarization. The last dimension is ordered as $(G_{00}, G_{01}, G_{10}, G_{11})$.
`uvw`	(N_time, N_ants, 3)	The (u, v, w) coordinates for each antenna at each time step.

TABLE 3.2: Functions for passing data to Montblanc

In our particular case we wish to include the contribution from RFI sources that are in the far sidelobes. Typically, in purely astronomical simulations, beam models are only included out to the first or second sidelobe. This is usually enough as the attenuation at these beam positions is so large such that astronomical sources provide a negligible contribution. This is not the case for RFI sources as the received power is much greater due to RFI sources being so much closer to us.

In order to strike a balance between accuracy, performance and FoV the fitted beam from Section 2.4.3.1 was evaluated with a total of 513 pixels across a 60° diameter giving a roughly 7 arcminute resolution. Given that the sidelobe

spacing is around 1° this provided more than enough resolution for accurate interpolation. Variations in the beam across frequency are much smoother so only 100 channels across a 1GHz range (800MHz-1800MHz) were used giving a frequency resolution of 10MHz.

Montblanc requires the FITS based beam models to to be split up into 8 different files. Each polarization component requires a separate file for the real and imaginary components so considering 4 polarization components this gives us 8 separate files.

3.7.3 Data Sinks

The data sink base class is named `SinkProvider` and provides a class method named `model_vis`. This function simply provides access to the output of the RIME evaluation. The output has dimensions of (N_time, N_bl, N_freq, N_pol). In this function the output data is saved to disk. All data was saved in binary to Hierarchical Data Format (HDF5)[14] datasets.

3.7.4 Limitations of Montblanc

Montblanc is a very powerful and versatile RIME implementation that allowed us to create MeerKAT specific simulations with ease, however there are some fundamental limitations with regard to simulating the contribution of RFI sources. Because Montblanc is designed to simulate astronomical sources it has baked-in assumptions that are not suitable for simulating RFI sources. Firstly it expects all sources to lie in the far-field and secondly that sources do not change position over time with respect to the celestial sphere. These limitations are further explained in the subsections below.

3.7.4.1 Static Source Positions

Montblanc expects sources to maintain static positions with respect to the celestial sphere i.e. all sources move at the sidereal rate. This is not the case for RFI sources in general. To solve this, the contribution due to RFI was simulated a single time step at a time. This was a simple workaround to simulate moving sources as static sources with a different position at each time step. Unfortunately such a workaround resulted in greater computational costs as the problem needed to be load in data from scratch for each time step.

An alternative solution (within Montblanc) was to create a source for every time step such that N_src would become N_src×N_time. In this solution

[14] https://www.hdfgroup.org/solutions/hdf5/

the flux could be varied over time such that at any given time step only N_src sources would have non-zero flux. This solution was not chosen as it increases the problem size rapidly when one wants to simulate many time steps.

3.7.4.2 Far-field Sources

Montblanc was designed to simulate static astronomical sources. As such it expects direction cosine coordinates which can be converted to celestial coordinates using the inverse of Equation 2.8. Montblanc uses these when calculating the geometric phase delay term for each source over time. Montblanc does not provide the possibility to include a custom phase delay term so the incoming EM-waves are always treated as plane waves. Due to this limitation all RFI sources used in the Montblanc simulations are treated as lying in the far-field. Unfortunately no work around was established for this particular aspect. This limitation was deemed to be acceptable for the first iteration as it is the time variation of the phase that is the most distinctive aspect of the RFI contribution. The distinct time variation of phase for an RFI source is due to the fact that RFI sources do not move across the sky at the sidereal rate. This causes the visibility phase contribution to change must faster than for an astronomical source. This phenomenon is explained further in Section 4.1.3.

Additionally the primary beam model is one derived from the far-field diffraction pattern. This is a mostly overlooked detail that is deemed beyond the scope of the current thesis.

3.8 Dask

Dask is an open-source python package that provides a simple interface to perform calculations that cannot be fit into memory alone or on a single machine. It splits a calculation into smaller calculations that can be distributed across machines or performed successively on a single machine.

The basic flow of work is to define where on disk (stored in a non-volatile manner like a hard drive) specific elements such as the bandpass and source attributes are stored. These elements that are stored on disk need to be saved in a specific format that allows sections of the data to be loaded at a time. Functions are then defined that transform these inputs into a desired output. These functions can be chained together indefinitely until a function is defined that writes the output to disk again. Through such a definition Dask is able to build a computational graph where it automatically determines what data is needed to perform any specific subcalculation and how many computational resources are needed for it. Dask will then assign subcalculations to

specific pools of compute resources (CPU core with accompanying memory allocation). It will repeat this until all sections of the computational graph have been executed.

To emphasize the requirement for Dask over something more traditional like NumPy let us calculate the memory requirements for an example problem. A typical target observation will take place in 15 minute runs, this is 900 seconds. We would like to simulate snapshot (instantaneous) visibilities every second so that we can perform averaging after the fact to account for time-smearing[15] that will strongly affect moving RFI sources. MeerKAT currently has a 4096 frequency channel mode in the L-band, and 64 antennas that give 2016 independent baselines. Finally, let us assume we wish to simulate 1000 sources in total (RFI and astronomical). If we wish to perform calculations using the `complex128` format then each number in the calculation takes up 16 bytes of memory.

Let us first look at the brightness matrix. To remind the reader, the brightness matrix contains a source's intensity and polarization information. This particular object has 4 polarization components making up the matrix and this is repeated for each source, time step and frequency channel. This gives a total of N_pol$\times$N_src$\times$N_time$\times$N_freq$\times$N_bytes which is $4\times1000\times900\times4096\times16 \approx 220$ GiB. Now this is more memory than many servers have let alone consumer hardware. When we consider the phase delay term where this number is multiplied by the number of antennas we have reached terabyte territory. It should be noted that these two objects are the absolute minimum necessary to calculate visibilities in the ideal interferometer. In this calculation we have only considered the memory requirements to store them let alone perform the RIME calculation.

3.8.1 Configuration File

It was decided that all the details pertaining to a simulation are best defined in a configuration file. This makes it much easier for a typical user to define a simulation without having to change any of the python code. YAML Ain't Markup Language (YAML) was chosen for the configuration file due to its human friendly ease of reading with very minimal syntax. The configuration file is broken up into sections (which can be further broken up into subsections) within which specifics are defined pertaining to that section. Table C.1 in Appendix C shows the different sections with the options available in each.

[15]Smearing is a perceived reduction in visibility amplitude due to averaging a series of complex numbers with varying phases. Time-smearing refers to this effect when the averaging set is a time series.

3.8.2 Near-field Sources

The introduction of curved wavefronts for near-field sources was one of the primary motivators for putting together a second version of RFIsim. In the first version, which uses Montblanc, all sources were assumed to be in the far-field. This aspect is described in Section 3.7.4.2. To calculate phase delays between antennas caused by a single near-field source one must know the exact position of each antenna and the source in question. To do this we made use of `Skyfield`. `Skyfield` allowed us to input specific GPS coordinates and altitudes for antennas and ground-based RFI sources as well as calculate satellite positions with high precision. Using this information we could easily calculate the distances between near-field sources and antennas. These distances were needed to calculate the near-field phase delay term given in Equation 2.52 and near-field brightness matrix given in Equation 2.53. In the second version of RFIsim all RFI sources are treated as near-field sources where their phase delays are calculated using Equation 2.52.

3.8.3 Primary Beam Model

In the first version of RFIsim using Montblanc a primary beam model could only be included through a set of FITS files. This is a convenient way to include very complex beam models for which an analytical formula is not available, however, as described in Section 3.7.2 this also means that evaluations are approximations that depend on the resolution of the FITS files. The added memory requirements that are increased with the FITS file resolution is another downside. In [3] they use Zernike polynomials which produce very accurate beam representations with analytical formulae. Such methods allow one to include a beam model that is both very complex and without the large memory requirements. It is for this reason that the second version of RFIsim supports the inclusion of analytical beam models. Initially only a basic Airy disk model as was created in Section 3.4.1 was included however the user can create whichever beam model they desire such as a Zernike polynomial based model.

3.8.4 Codex Africanus

Codex Africanus[16] is another piece of software written by the RARG team at SARAO. It can be considered the successor to Montblanc. It was written to be more flexible and simpler than Montblanc. Much like Montblanc it is geared towards simulating astronomical sources, and hence does not include any of the features specific to simulating RFI sources that have been described in this thesis. It predates RFIsim, however, development on this

[16] https://codex-africanus.readthedocs.io/en/latest/

project has been much slower (in the areas applicable to RFIsim) as its scope is much greater and includes additional steps of the data reduction pipeline used in radio interferometry. A number of the RIME evaluations that occur under the hood in our own simulator could be replaced with those in Codex Africanus and in future this will probably be done. The main differences between RFIsim and Codex Africanus are as follows. Codex Africanus is intended for users that want to experiment with different sections of the radio interferometric data reduction process for algorithmic research purposes. As such it has been designed to be incredibly flexible and interacted with using an API. RFIsim on the other hand is designed to be very simple where the main interaction between the program and the user is through a configuration file. Additionally, RFIsim includes RFI specific components. Essentially Codex Africanus is an API that can be used in RFIsim to make it more robust in the future as well as expand the possible compute devices that can be used. Codex Africanus was not used from the beginning of RFIsim's second version as not all features were available at the time and for the purposes of learning. The library has now matured and the learning has been done so there are great benefits to making use of Codex Africanus in future versions of RFIsim.

3.9 Simulator Comparison

In Sections 3.7 and 3.8 we have described the workhouses that perform the RIME calculations and some of the specifics of how input data is included. We would like to use this section to clarify some of the differences present in the two versions of the simulator. Table 3.3 below gives the main differences between the first and second versions of the simulator.

Simulator Comparison		
	Version 1	Version 2
Computational back end	Montblanc	Dask
RFI sources	Far-field	Near-field
Brightness matrix equation	2.30	2.53
Phase delay term equation	2.34	2.52
Jones matrices	In-built	Self coded
Primary beam model	Discrete	Continuous
GPU accelerated	Yes	No
Code repository	https://github.com/chrisfinlay/RFIsim	https://github.com/chrisfinlay/RFIsim/tree/curved_wavefront
Development period	September 2018 - February 2019	June 2019 - August 2019

TABLE 3.3: A comparison of the main aspects of each version of the simulator.

An interesting difference to be noted between the two versions of RFIsim is that even though the first version was GPU accelerated and the second was not, we found the second version to run faster for a set problem. We believe this to be due to two different things. The first is that, in version 2, the computation of input data is predefined in a graph but not calculated therefore Dask is able to decide for itself when the most efficient time is to calculate the needed data for a specific subsection of the problem. The second reason why version 1 may be slower is that the primary beam model is discrete and therefore 3D interpolation needs to take place to calculate the associated Jones term. Since version 2 accepts an analytical model for the primary beam there is a much faster function evaluation that takes place. The downside to this is that for more complex beam models there may not be an analytical model available and a discrete can do a better job of approximating it.

Chapter 4

Results

In this chapter we will compare our simulations with real data, looking at the data from different views. This will be a mostly qualitative comparison. This is done in Section 4.1. We will also explain how this simulator was used for its intended purpose in supporting algorithm testing and development for RFI mitigation. This is done in Section 4.2.

4.1 Real and Simulated Data Comparisons

4.1.1 Bandpass

The bandpass is defined as the frequency response of the telescope. In this section, we will look at the magnitude of visibility data across frequency for a single time step. The comparison in Figure 4.1 are important for highlighting the following aspects. We want to see the general bandpass shape, the relative power of the RFI contaminated channels, and the frequency distribution of RFI contaminated channels.

When looking at the overall shape of the bandpasses we find them to be similar including the dip that is present in the 1050-1150 MHZ range and the small peak around 1700 MHz. The real data shown in Figure 4.1 was taken on a different day to the data used to calculate the bandpass in section 3.5.1, therefore, we can expect the bandpasses to differ to some degree.

When looking at the frequency distribution of the RFI contaminated channels we see that there is a good match with three main regions of contamination. We have a small region between 920 MHz and 960 MHz where GSM signal from cellphones and Distance Measuring Equipment (DME) signal from aircraft are present. The middle section between 1150 MHz and 1300 MHz and the later section between 1500 MHz and 1600 MHz are where GPS signal is present.

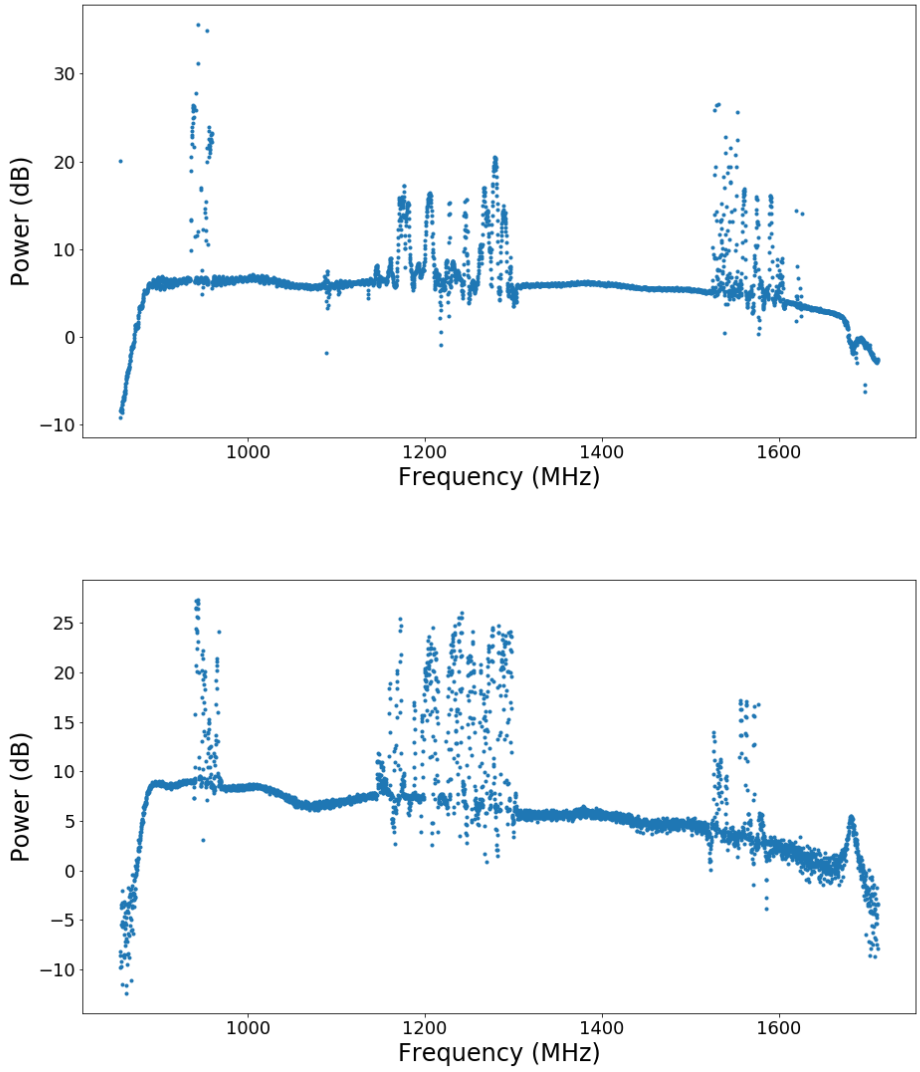

FIGURE 4.1: The above plots show the logarithmic visibility amplitudes over frequency for a single time step and single baseline. Each panel shows visibilities over the MeerKAT L-band from 856 MHz to 1712 MHz. The top panel shows an example of real data from a single calibration run observing PKS J1934-6342 on 22nd March 2018. The bottom panel is an example of a simulated observation of the same source on 5th November 2018. The differences in the narrowband RFI spectral profile and amplitude variation across the band are attributed to the random nature of the RFI modelling that is described in Section 3.2.

The joint distribution of RFI signal over both frequency and amplitude however can still be improved. We see that the RFI amplitudes in the middle section of the frequency band tend to be some 10 dBm weaker than the RFI

that is present in the smaller, lower frequency section. The frequency distribution of RFI was obtained from a Historical Probabilities of RFI project led by Isaac Sihlangu [58]. This project looked at the frequency of RFI flags over many dimensions, however, detected RFI amplitude was not one of them. We, therefore, had access only to the frequency distribution. The amplitude distribution was treated as independent of frequency, following a uniform distribution in log space, so as to produce a wide range of amplitudes. Since both these distributions were treated as independent, we did not achieve a joint distribution that allows for weaker RFI amplitudes in specific frequency bands. This is an area that can be improved in future work.

For further comparison plots of the bandpass, refer to the Figures in Appendix A.

4.1.2 Spectrograms

A spectrogram or waterfall plot is a two-dimensional plot of function values over both frequency and time. We will perform comparisons on the amplitude and phase spectrograms. The amplitude spectrogram is shown in Figure 4.2 with the real data shown on the top panel and the simulated data on the bottom panel. From these plots, one is able to identify the frequency distribution of RFI-contaminated channels much like when comparing the bandpass. After all, the plots in Figure 4.2 are nothing more than the bandpass plots from Section 4.1.1 extended over time. We will focus on the time-varying structures in this section.

The most important feature is the amplitude modulation of the RFI channels over time.This is faintly visible in the 1150-1300 MHz range for both plots in Figure 4.2. This is caused by RFI sources that move through the sidelobes of the primary beam and therefore the received power from these sources is modulated/attenuated accordingly.

For further comparison plots of the amplitude spectrogram refer to the Figures in Appendix B.

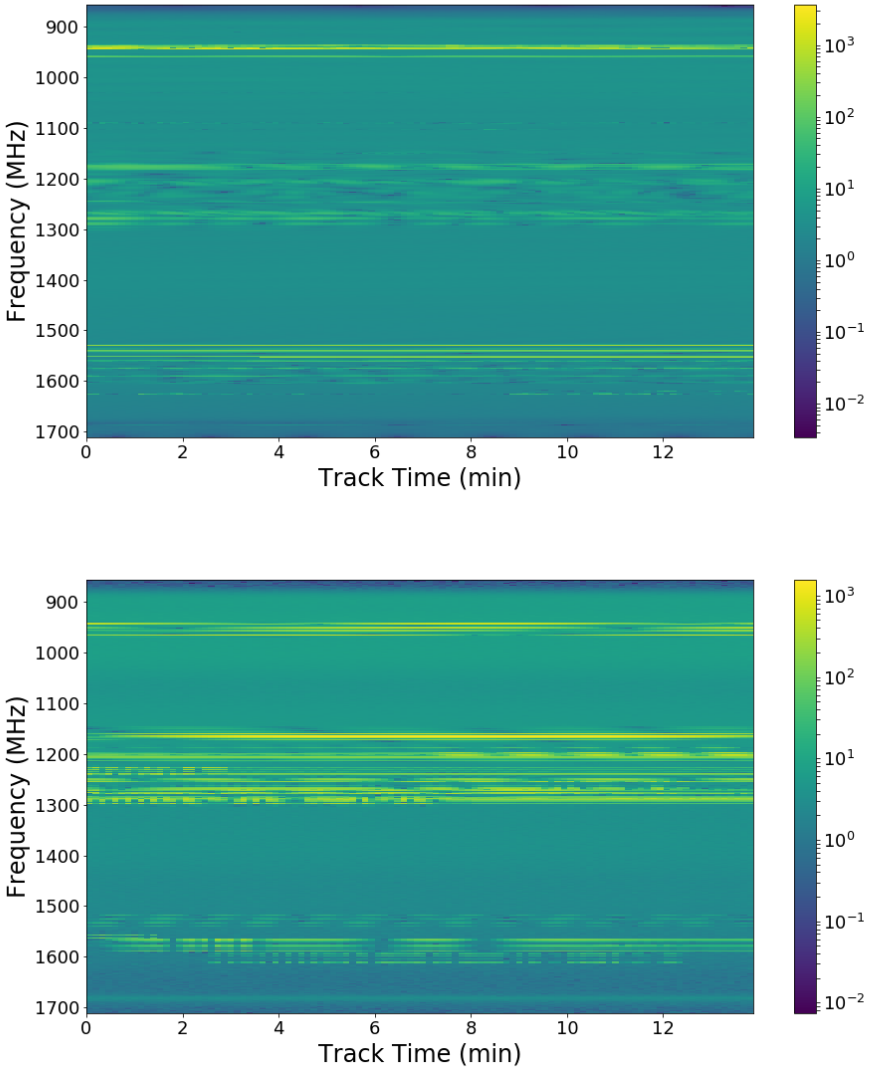

FIGURE 4.2: The above plots show the visibility amplitudes over both frequency and time for a single baseline. Each panel shows visibilities over the MeerKAT L-band for a period of $\sim$ 14 minutes. The top panel shows an example of real data from a single calibration run observing PKS J0408-6545 on 22nd March 2018. The bottom panel is an example of a simulated observation of the same source on 5th November 2018. Both of these plots have logarithmic color scaling. The differences in the narrowband RFI spectral profile and amplitude variation across the band are attributed to the random nature of the RFI modelling that is described in Section 3.2. The difference in RFI amplitude variation over time is due to the modelled RFI being different in location and movement bas compared to the real observation.

Figure 4.3 shows a comparison of uncalibrated visibility phase spectrograms.

The real data is plotted on the top panel and the simulated data on the bottom panel. From these plots, we found that the majority of the phase spectrogram is filled with values close to 0. This is because the uncontaminated channels are generally dominated by the signal from the source at the phase centre. When the field of view of a telescope is small, the dominating signal contribution to the visibilities will come from toward the phase centre, in a tracking interferometer. This is due to the primary beam strongly attenuating signals from off-centre sources.

In the uncontaminated channels, there are two main differences that can be seen. Firstly, in the real data, a phase offset is present, whereas, the simulated data has a background much closer to zero. Secondly, the real data has a slight phase gradient across frequency. This is mainly due to the lack of inclusion of certain effects in the simulations. This is discussed further in Section 4.1.3. The contaminated channels show phase wrapping[1] over time which is indicative of a strong source moving at a non-sidereal rate. The sidereal rate is the rate of the Earth's rotation relative to a static sky frame. The rate of phase wrapping is dependent on the angular velocity of the source that is parallel to the baseline.

[1] Phase wrapping is when the phase continuously increases/decreases such that it wraps back around to the original value due to modularity.

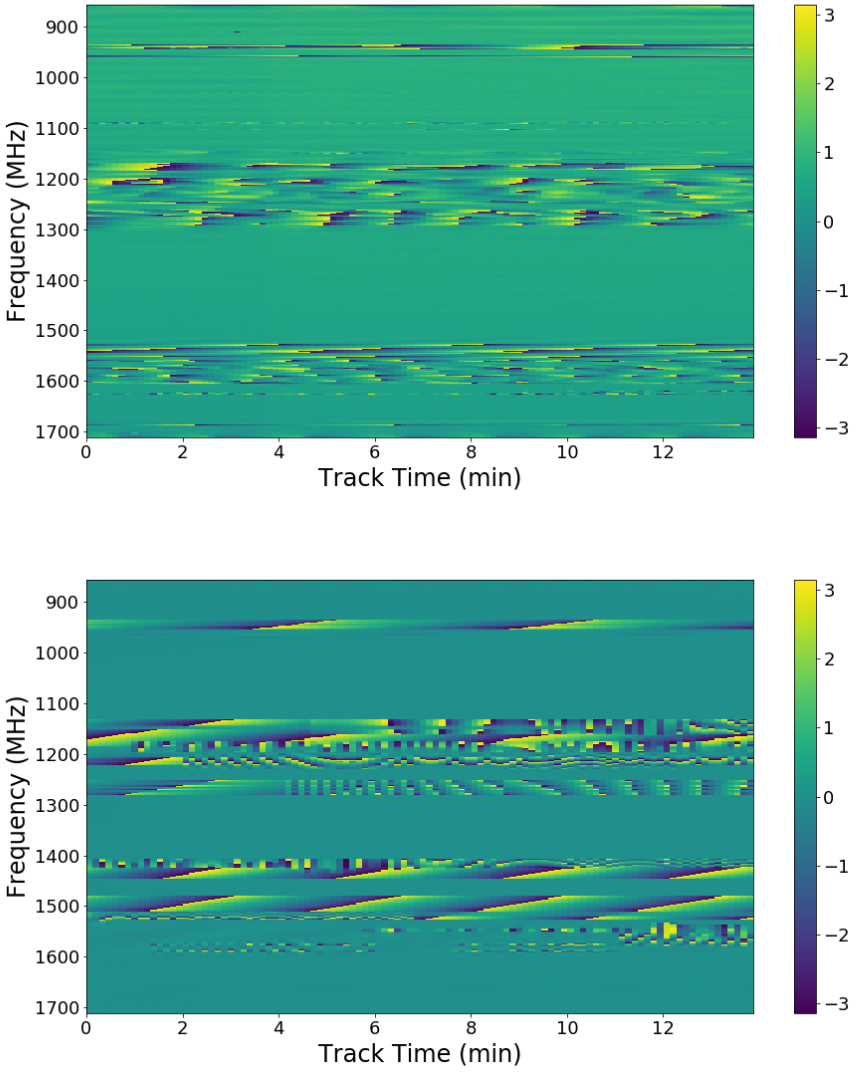

FIGURE 4.3: The above plots show the visibility phases over both frequency and time for a single baseline. Each panel shows visibilities' phase over the MeerKAT L-band for a period of 14 minutes. The top panel shows an example of real data from a single calibration run observing PKS J1934-6342 on 22nd March 2018. The bottom panel is an example of a simulated observation of the same source on 5th November 2018. Both of these plots have linear color scaling. The difference in RFI phase variation over time is due to the modelled RFI being different in location and movement as compared to the real observation. The difference in RFI channel width is due to the random nature of the RFI spectral profile model that allows the channel width to vary and potentially overlap at times.

4.1.3 Phase Time Variation

In this section, we will focus on Figure 4.4. The phase of the visibilities contains the positional information of sources in the sky. One of the fundamental aspects of RFI in comparison to astronomical sources is that RFI sources move across the sky at a different rate than astronomical sources. Astronomical sources move at the sidereal rate that the tracking of the telescope stays synchronized with. For uncontaminated data we do expect the phase to change over time at a particular rate[2], however, this rate of phase change is much slower than that of a source not moving at the sidereal rate.

For the plots in Figure 4.4, we chose two specific baselines from our simulations that had a minimal difference in orientation (<20°) as well as having at least an order of magnitude difference in baseline length. The same baselines and very similar frequencies were chosen for both the real and simulated datasets.

[2]This rate of phase change is governed by the changing baseline coordinates as described in section 2.2.2.2

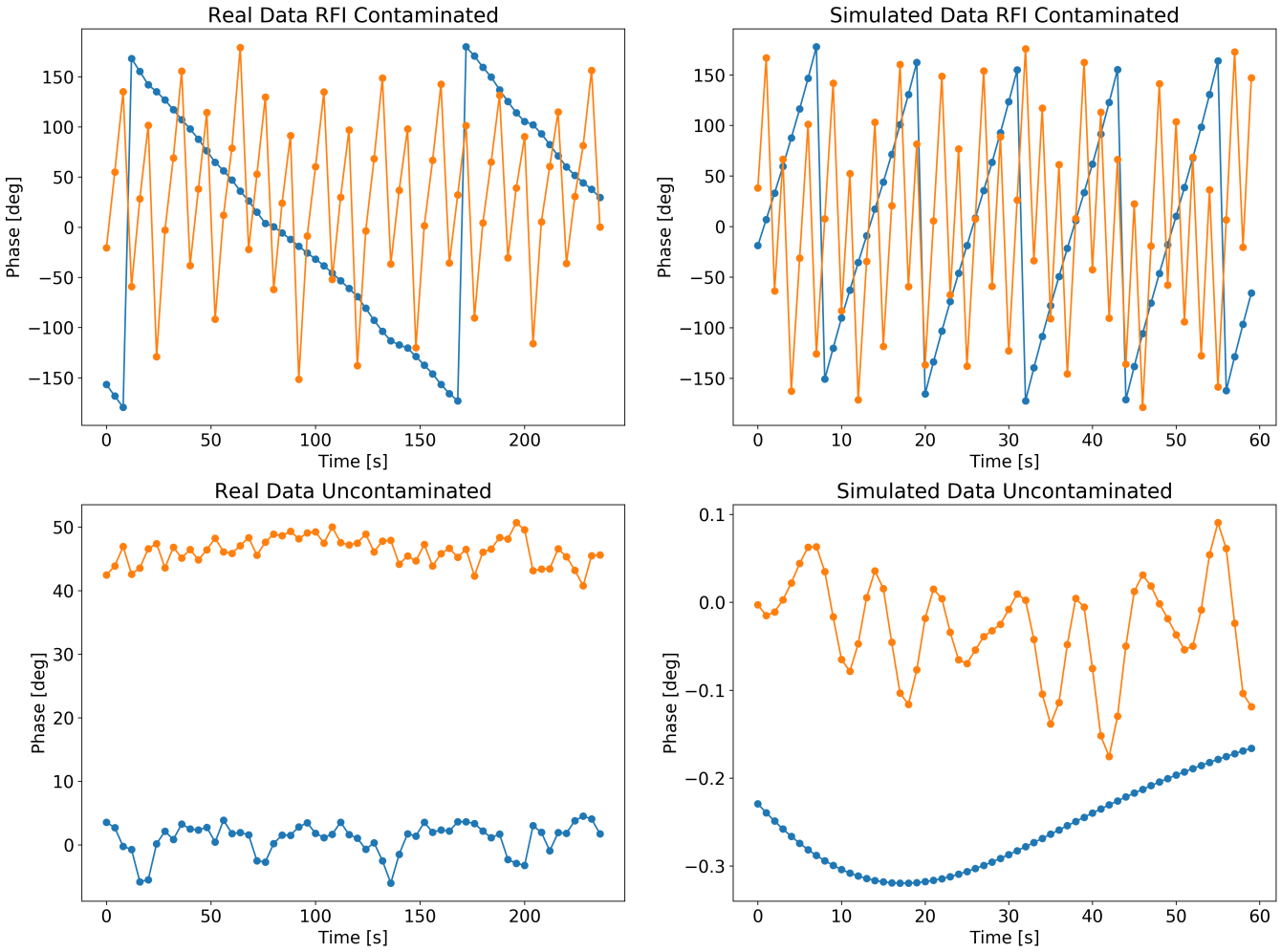

FIGURE 4.4: All the above panels show the phase variations over time of uncalibrated visibilities very close to the L1 GPS frequency band of 1.575 GHz. The right-hand side panels show real data from a calibration run on 20th April 2018 pointing at PKS J1934-6342. The left-hand side panels show simulated data for an observation on 30th March 2019 pointing at PKS J0157-1043. The blue curves are for a short baseline of 117m and the orange curves are for a long baseline of 3474m. The top panels show the phase of visibilities that are contaminated by RFI and the bottom panels show the same for an uncontaminated channel.

There are two distinct effects present in Figure 4.4. To understand these, one needs to keep in mind that each baseline views the received power from the

sky through a sinusoidal filter. We call these fringes. The telescope tracks a particular point in the sky so the fringes (on a time scale of minutes) remain in a constant position with respect to astronomical sources in the sky. Therefore anything not moving at the sidereal rate with the sky will be moving across the fringes at a rate that is dependent on the length of the baseline (longer baselines give shorter fringe spacing) as well as the velocity of the object relative to the baseline orientation. An RFI source like a satellite will not move at the sidereal rate and will therefore move across the fringes such that the phase is changing and wrapping around. That same source when viewed with a longer baseline will have the phase wrapping faster because of the tighter fringe spacing. We can see this effect in the two top panels when comparing the blue (short baseline) and orange (long baseline) curves. The difference between an RFI source that does not move at the sidereal rate and an astronomical source that does, is present when comparing the top panels with the bottom panels. We can see that the phase remains more or less constant for the uncontaminated visibilities present in the bottom panels. This is a very important fundamental difference between visibilities induced by astronomical sources and RFI sources. The description given above is simplified and explained as though only a single source is present in the data. This is not the case for any of this data, however, the main contribution will be from a single source for both cases. The received power contribution from RFI sources is generally orders of magnitude larger than the astronomical contribution due to it being much closer to the telescope. In the case of the uncontaminated data, the same argument given in section 4.1.2 can be used.

One last point to note in the two bottom panels is that the simulated data has phases closer to zero than the real data. This can be caused by ionospheric delay and phases present in the gains. These effects were not included in the simulations so such large phase offsets cannot be seen.

4.2 Application to RFI Mitigation Algorithms

The main goal of RFIsim was to create a test dataset for developing better RFI flagging and excision algorithms as compared to those that are in use today. There are a number of advantages to our simulated dataset over currently available datasets for testing RFI flagging and excision. First, let us consider what is currently available. One can either turn to a simulator or a human-curated dataset.

When looking for a simulator for radio observations that include RFI one will come across `HIDE`, [1], and `hera_sim`[3]. `HIDE` provides the RFI and astronomical contributions separately so both flagging and excision algorithms can be developed using this simulator. Unfortunately, `HIDE` is only a single dish simulator, and hence is of no use when one is interested in interferometric data. The RFI model included in `HIDE` is also very simple having only gaussian (or exponentially) shaped time-frequency pulses being included directly into the time-frequency data. `hera_sim` does simulate interferometric data, however, only for redundant arrays of non-steerable dishes. This makes it unsuitable for simulating data from a telescope such as MeerKAT which is not redundant and tracks a source in the sky. The RFI model included in `hera_sim` is better than that of `HIDE` however it does not include any moving sources of RFI like satellites.

On the side of human-curated data, we have real data that has been flagged by astronomers. There are three main problems with data that is flagged by astronomers. Firstly, an astronomer cannot determine the exact contribution of RFI in a given data point. They can only flag whether a given data point is contaminated. This is a major advantage for a simulator that can provide separate contributions. Secondly, astronomers are human so mistakes are possible and what one astronomer considers to be RFI may not be deemed the same by another astronomer. Finally, flagging RFI is a laborious process and hence getting large quantities of this data is very hard. As a result, human flagged data is not perfect and ends up lacking in both quality and quantity.

Having looked at the currently available sources of data we find a number of problems. They come down to a lack of an appropriate simulator for MeerKAT like telescopes, and human data not being of high enough quality and quantity, and not providing exact contributions from astronomical and RFI sources separately. These problems are solved by RFIsim. RFIsim produces visibilities for a steerable phase tracking interferometer, specifically MeerKAT in our case, which are contaminated with realistic RFI where the astronomical and RFI contributions are provided. With RFIsim, an arbitrarily sized dataset can be generated, and flags can be created by thresholding the exact RFI contributions. We are able to produce large ground truth datasets for testing both flagging and excision algorithms. With the current wave of

[3] https://github.com/HERA-Team/hera_sim

machine learning[4] (ML) and specifically deep learning[5] (DL), the ability to produce vast amounts of high fidelity data for training is advantageous for developing machine/deep learning algorithms.

To this end, I used RFIsim to produce a large training set[6] of approximately 1TB for the analysis in [69]. In this paper, lead by Alireza Vafaei, an improved RFI flagging algorithm is developed and compared with both an existing DL-based approach, [2], and a modified version of `AOFlagger` [43] that is used in the MeerKAT Science Data Processing (SDP) data reduction pipeline. The new algorithm presented in the paper is a DL-based approach. The contributions I made to this paper are summarised as having generated the datasets, using the simulator that is the subject of this thesis, that was used for training and testing the algorithm. I also provided any further information needed about the data. Additionally, I helped determine if any issues in different versions of the algorithm could be explained by simulation effects and/or how these effects could be exploited. Alireza lead the machine learning side of the paper including but not limited to model architecture decisions, implementation of the algorithm and the training and testing thereof.

In the paper, three different datasets are made use of for training and testing purposes. The first is 13 months of a `HIDE` dataset, where each data sample has 276 frequency channels and 400 time steps. The data are amplitude spectrograms from a single dish simulation. The second dataset has 100 samples generated using the simulator presented in this thesis. Each sample contained 4096 frequency channels, 800 time steps, 15 baselines, and 4 polarizations. The last dataset used was hand flagged (by a SARAO based radio astronomer) data from the Karoo Array Telescope 7 (KAT7). This data consisted of 12 hours of actual observations made in 2016 and each sample has 800 frequency channels, 42 baselines, and 4 polarizations. To produce a training/test set from the two simulator-based datasets mentioned above, a threshold was applied to the RFI contributions leading to what is called a mask. This is an array of ones and zeros where a one represents the presence of RFI and a zero, the lack thereof.

Five different algorithms were tested and compared. They were named: R-net5, R-net6, U-net16, U-net32, and SDP Flagger. The R-net models are the newly developed algorithms, U-net models are implementations of the algorithm proposed in [2] and finally, SDP Flagger is a SDP implemented version of `AOFlagger`.

[4] Machine learning is a branch of algorithms that improve/learn through experience such as from data examples.

[5] Deep learning is a branch of machine learning that makes use of many parameters to estimate an arbitrary function such as identifying the presence of RFI.

[6] A training set is a set of examples which include the input and expected output from a function. In our context, this would be a time-frequency spectrogram as input and an equal-sized array with zeroes where there is no RFI and ones where RFI is present.

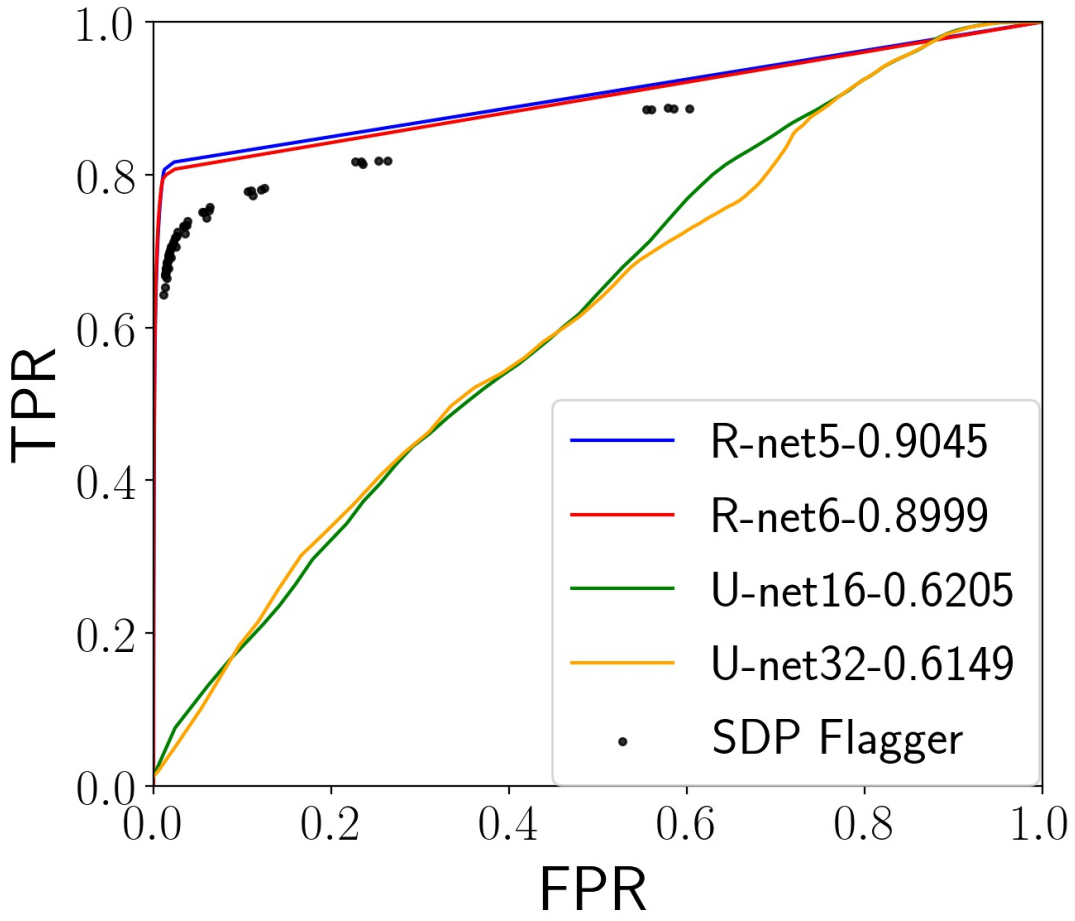

FIGURE 4.5: This graph shows the true positive rate of the classifiers against the false positive rate as the threshold of the classifier is varied. The black dots show the performance of the SDP Flagger as a pair of thresholding hyperparameters were varied.

Figures 4.5 & 4.6 are plots of the True Positive Rate (TPR) against the False Positive Rate (FPR) on the test set. In the case of RFI mitigation, the TPR is the percentage of RFI that is correctly flagged, and the FPR is the percentage of non-RFI that is flagged as RFI. These are commonly used to compare classification algorithms. In most ML classification models the output is thresholded to determine a positive or negative classification. By varying this threshold one is able to produce a curve on the TPR vs FPR graph. A random classifier will make a straight line from the bottom left corner to the top right. The Area Under the Curve (AUC) on the TPR vs FPR graph is used as a metric to compare classifiers. A perfect score is 1 which corresponds to a curve that moves vertically along the y axis from the point 0,0 to 0,1 and then straight across to the point 1,1. The AUC for U-net, R-net and transferred R-net models are given in the legend. One is not supplied for the SDP Flagger as there are multiple parameters to threshold so no unique curve is available.

To produce Figure 4.5 the algorithms were trained and tested on simulated MeerKAT data from RFIsim. We see that the R-net models narrowly beat the SDP Flagger on this data. We also see that the U-net models perform badly

in comparison to the R-net models and SDP Flagger. The two main benefits of using RFIsim are present here. Firstly, RFIsim produced a test set that was used to produce these plots. Thanks to this test set we are able to compare RFI algorithms with confidence. Secondly, RFIsim produced enough training data for a state-of-the-art RFI algorithm to be developed and trained.

In Figure 4.6, we see that the R-net models trained on MeerKAT simulations only (labelled as direct-R-net) and then tested on KAT7 data performed comparably with the SDP Flagger. This indicates that the MeerKAT simulations are presenting similar RFI features that are also found in the KAT7 dataset.

When training the R-net models labelled as transferred-R-net a training method known as transfer learning was employed. Transfer learning is a training method where a model is trained on one dataset that is similar to the desired task. Once this initial training is complete, training is completed on the data set of the desired task with the exception that only part of the model is allowed to change in this second training stage. On should note that in the second stage of transfer learning that a large portion of the model is frozen while training. This means that whatever features were learnt in the first stage are maintained and only how these features are used is changed in the transfer learned model. In this case, the MeerKAT dataset was used for the initial training, and then part of the KAT7 dataset was used for the second training stage. This means the model can leverage the large quantity of training data available from RFIsim as well as learn specifics of the KAT7 dataset from a small amount of KAT7 data. In Figure 4.6 we see that the transferred-R-net models outperformed the SDP Flagger and direct-R-net models. The important point to note in the results of this paper is that the R-net models successfully used transfer learning to learn RFI features from training on the MeerKAT simulations that could be reused on real KAT7 data.

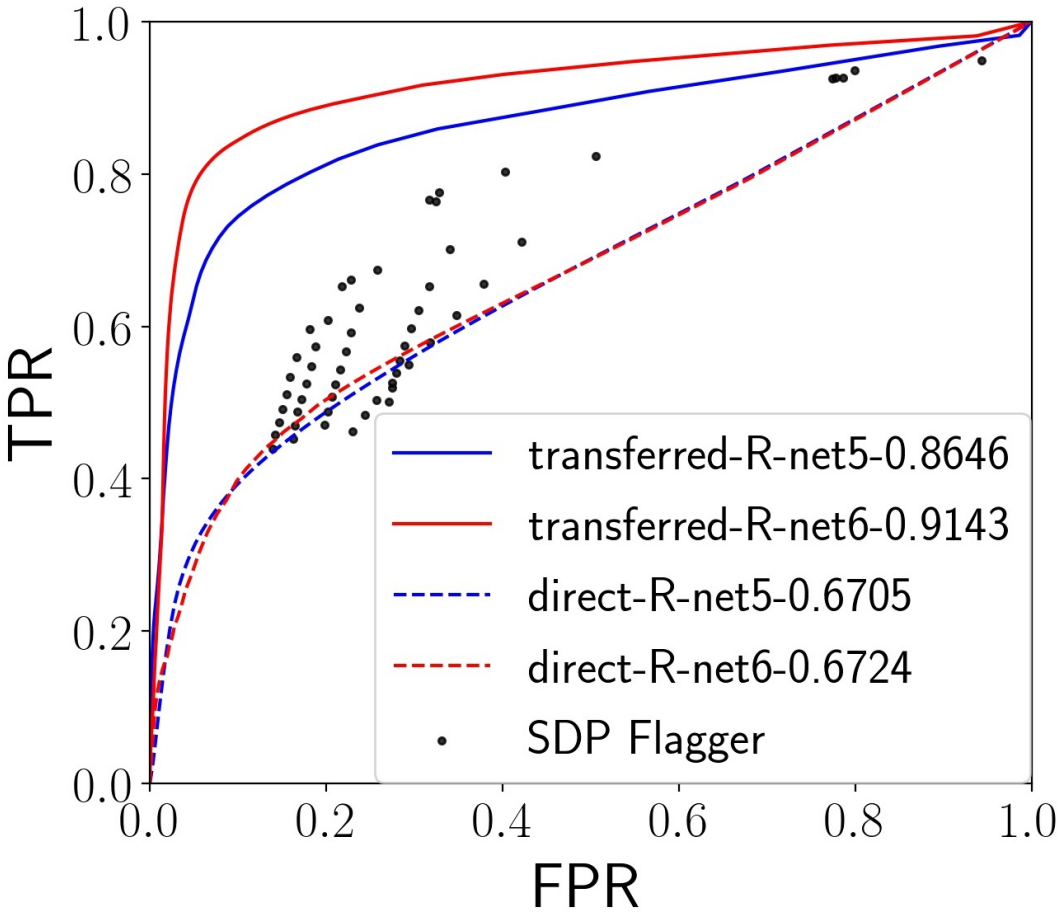

FIGURE 4.6: This graph shows the true positive rate of a classifier against the false positive rate as the threshold of the classifier is varied. The black dots show the performance of the SDP Flagger as a pair of thresholding hyperparameters were varied. The dashed curves show the performance of the R-netX models that were trained on MeerKAT data and then tested on KAT7 data. The solid curves show the performance of the R-netX models that were trained on MeerKAT data, then further trained with a restriction on some KAT7 data, and finally tested on withheld KAT7 data.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The aim of this thesis was to create a simulator for radio interferometer arrays in general and the MeerKAT telescope in particular, including a realistic Radio Frequency Interference (RFI) model of known GPS satellite orbits and towns. This can produce huge amounts of visibility data that can be used to train and test RFI mitigation and excision algorithms. This is needed because currently there exists no such simulator or dataset with known ground truth RFI contributions. Without this, machine learning algorithms cannot be reliably trained to excise RFI and, more importantly, algorithms cannot be legitimately compared with each other to determine which is best.

In this thesis we explained the basics of interferometry data simulation and how to include both moving and stationary types of RFI into these simulations. In our goal to create a realistic simulator we have included polarisation, a real astronomical sky model with data from NVSS and SUMSS, the measured MeerKAT bandpasses and primary beam model (matched and extended to arbitrary angular offset from the phase centre), as well as the actual Two Line Element (TLE) sets of known satellites. Comparing to real MeerKAT data, we find that RFIsim does produce realistic looking amplitude and phase plots with highly plausible RFI distributions.

Data produced by RFIsim has been used to successfully develop and test a state-of-the-art deep learning RFI mitigation algorithm, R-net. Because of the access to ground truth data, it was possible to unambiguously demonstrate that R-net outperforms existing RFI algorithms.

Finally, the best R-net model trained on data from RFIsim, was successfully applied to real KAT-7 data using a small amount of human hand-flagged data via transfer learning. This shows that simulated interferometric RFI data can play a key role in the design, training and optimisation of future RFI algorithms which will need a large amount of data to reach their full potential.

5.2 Possible Improvements and Future Work

The work in this thesis was sufficient to serve our original goals. The improvements that will be listed below may not represent significant advances but can provide further legitimacy to any derived results (results of algorithms development). One would also be able to determine if a particular RFI mitigation/excision algorithm is able to work in certain instances such as avoiding flagging/excising spectral lines that are present in the spectrum of a radio source.

Given that this simulator was put together from data products derived from a real telescope, there is potential to extend this work into high accuracy forecasting. One of the most valuable types of forecasting could be to determine the impact of RFI on an intended future real observation. This can lead to more efficient telescope scheduling. Such RFI mitigated scheduling could lead to significantly greater scientific output of a telescope.

There are some telescope based effects that were not included in RFIsim, as well as further improvements to currently implemented methods that we will now discuss further.

In Section 3.4.1 we created a basic primary beam model for the auto-polarization components (HH and VV). However, this model did not fit the data particularly well. Although it was sufficient for our purpose, this would need to be improved if one wishes to provide accurate forecasting. We also attempted to create a primary beam model for the cross-polarization components but found that the data was too inconsistent to support the basic model that was chosen. A possible solution to improve the entire primary beam model may be to have a Zernike polynomial model for the central region where measurements have been made and a basic model for the outer regions. One would of course need to make sure the transition between these regions is smooth. The primary beam model that we have talked about throughout this thesis refers to the far-field model for the beam[1]. Using Equation 2.45 we can determine that the near-field region is within 1215m[2] of an individual dish. If one wished to introduce RFI sources that are within/close to this distance one would have to include a near-field primary beam model to accurately simulate such a source.

When developing a ML algorithm, the resultant model will perform better when the training data is representative of the final task. For this reason, it is very useful to have a bandpass model in the simulations, that matches the telescope for which the model will be deployed. In section 3.5.1 a bandpass model is generated for each antenna, from a single observation. However, the bandpass for a particular antenna does change over time. The trouble

[1] One should note that the far-field for an individual antenna is different to the far-field for the entire array.

[2] To calculate this, a dish diameter of 13.5 m and wavelength of 30 cm was used.

with having the same bandpass for every simulated observation is that the distribution of simulated data is not necessarily an accurate representation of the real data distribution and thus the algorithm may end up performing poorly on real data. A solution to this would be to build a probabilistic model for the bandpass. This can be created by performing the same analysis as is done in section 3.5.1 for many observations. From there one can either sample directly from the extracted bandpasses or create something like a kernel density estimate[3] to sample from.

In RFIsim a couple of previously mentioned propagation effects were not included. These are the ionospheric phase delay and the Faraday rotation effect. The magnitude of these effects scale inversely with frequency, ν^{-1} and ν^{-2} respectively, so these effects, or lack thereof, will become more prominent when simulating at lower frequencies.

As mentioned above it would be useful to include astronomical sources with spectral lines so as to test an algorithms' propensity to flag/excise such a discontinuity in the frequency domain.

In RFIsim the RFI model is somewhat limited in positional variety since it only includes satellite based sources and that of a few nearby towns. It would be a very useful addition, especially for MeerKAT forecasting in the L-band, to include other sources such as aeroplanes with their Distance Measuring Equipment (DME) since the MeerKAT site is very close to the flight path between Cape Town and Johannesburg and is severely affected by this. The signal model included in our simulations has its complexities. However it can be fine-tuned to reflect reality even better. The emission strength and central emission frequency of a RFI source currently does not depend on the nature of that RFI source. It would be beneficial to include a RFI signal model that is more cognisant of the particular RFI source it is simulating.

In section 3.6 we describe a program that converts the output from RFIsim to the Measurement Set (MS) file type. This is very useful as our simulations are therefore able to be analysed, and even processed into images, using common astronomical software tools such as CASA. Such interfacing into existing software tools allows us to leverage the power of these tools and reach a broader audience. Including the ability to interface with other common astronomical tools such as FITS files and Local Sky Model (LSM) files to specify the astronomical sky model would be a useful and a welcome addition for many users of our simulation tool.

In section 2.2.1 we describe a fringe stopping, tracking interferometer. There are other types of interferometers such as drift scan interferometers, phased arrays and Very-Long-Baseline Interferometry (VLBI) which work on the same basis of correlating signals from different spatial locations, but differ in other

[3] A kernel density estimate is a non-parametric way to estimate the probability density function of a random variable. In our case the random variable is a multivariate being the bandpass.

key ways. Introducing the ability to simulate these types of telescopes would bring RFI contaminated radio interferometry simulations to a much broader audience.

Appendix A

Bandpass Comparison

In this appendix we show further examples of bandpass plots from real and simulated data. The figures on the left-hand side of the page are real observation data and those on the right-hand side are simulated data.

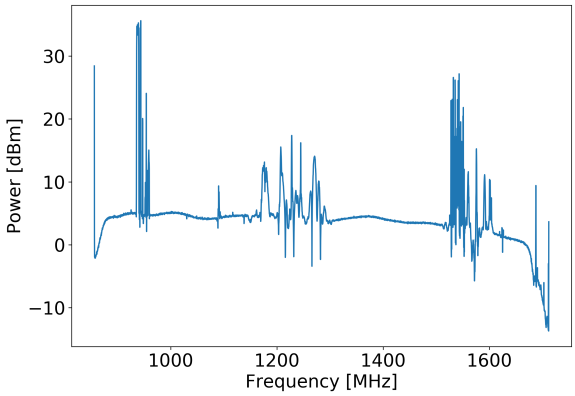

FIGURE A.1: This is the bandpass plot of a real observation of PKS 1934-6342.

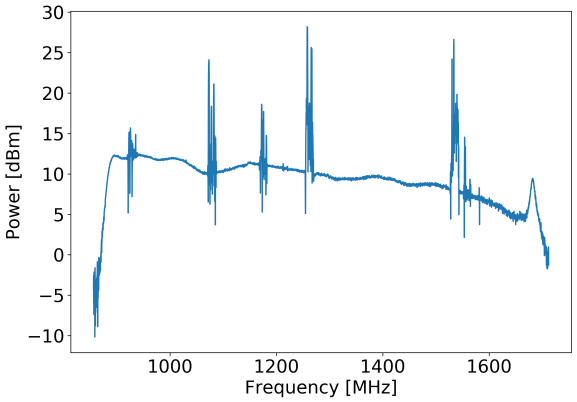

FIGURE A.2: This is the bandpass plot of a simulated observation of PKS J0006-0004.

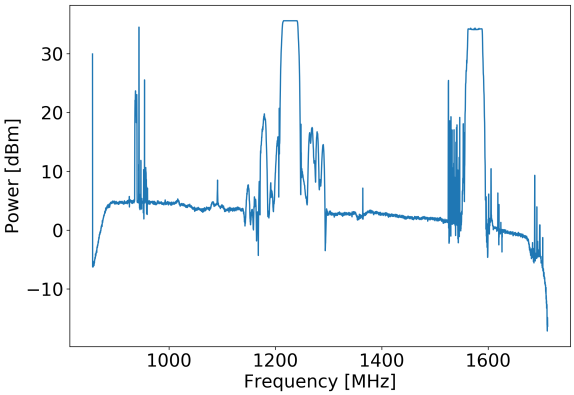

FIGURE A.3: This is the bandpass plot of a real observation of PKS J1424-4913.

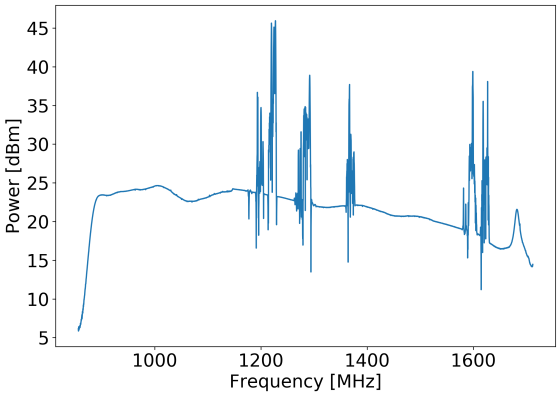

FIGURE A.4: This is the bandpass plot of a simulated observation of PKS J0050-0929.

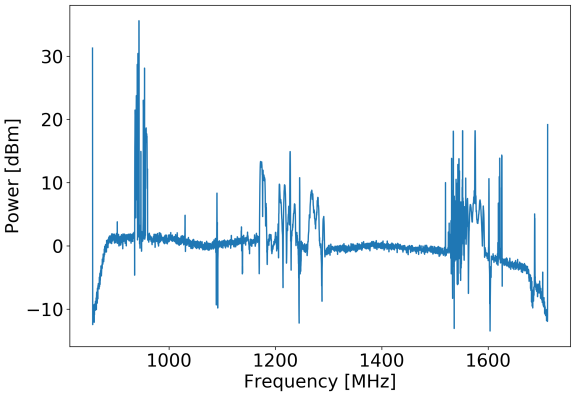

FIGURE A.5: This is the bandpass plot of a real observation of WMAP J1617-5847.

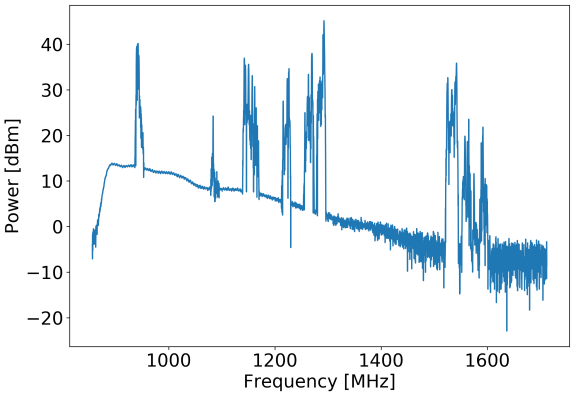

FIGURE A.6: This is the bandpass plot of a simulated observation of PKS J0157-1043.

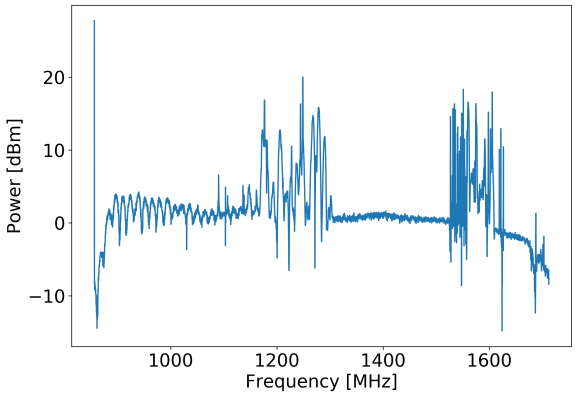

FIGURE A.7: This is the bandpass plot of a real observation of WMAP J1617-5847.

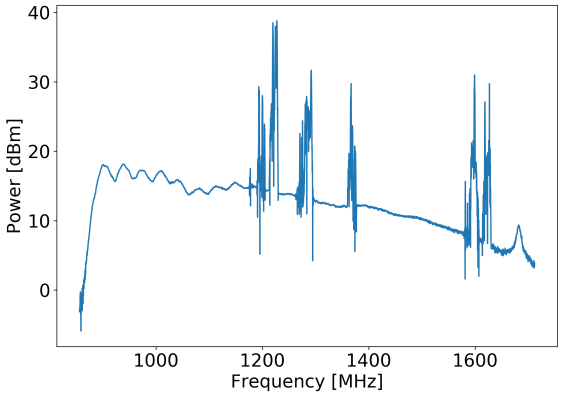

FIGURE A.8: This is the bandpass plot of a simulated observation of PKS J0050-0929.

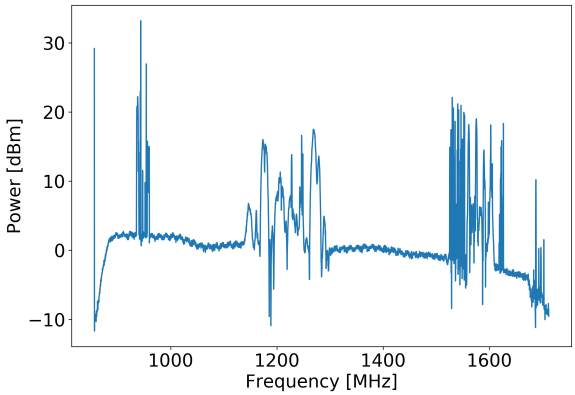

FIGURE A.9: This is the bandpass plot of a real observation of PKS J1218-4600.

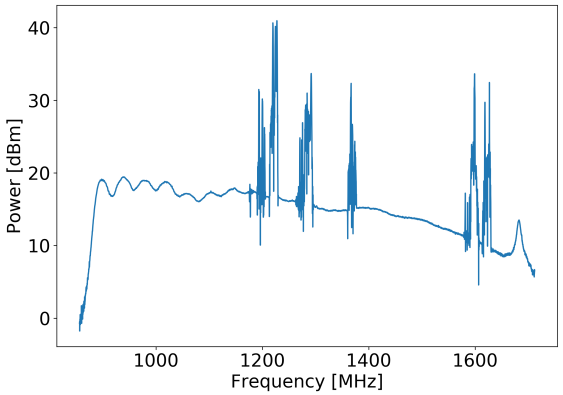

FIGURE A.10: This is the bandpass plot of a simulated observation of PKS J0050-0929.

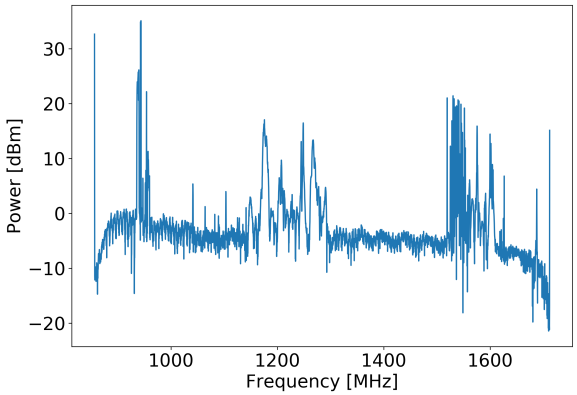

FIGURE A.11: This is the bandpass plot of a real observation of PKS J0506-6109.

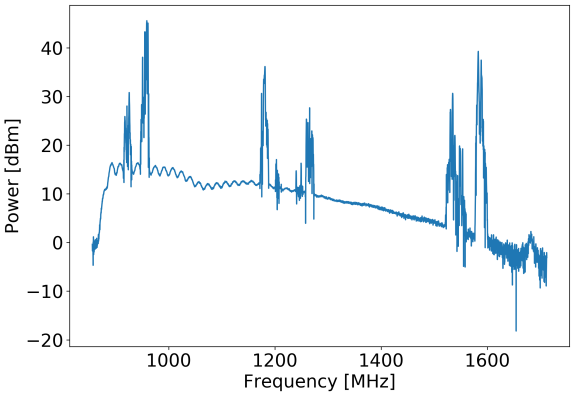

FIGURE A.12: This is the bandpass plot of a simulated observation of PKS J0050-0929.

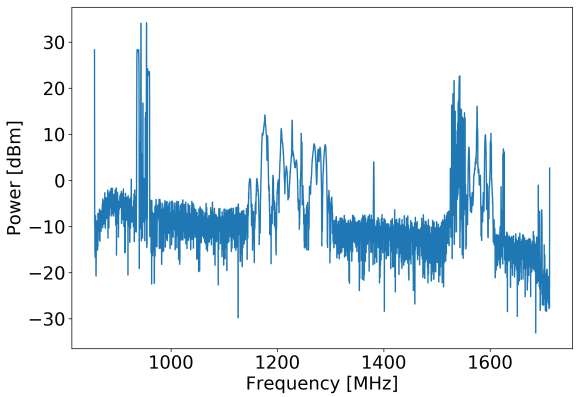

FIGURE A.13: This is the bandpass plot of a real observation of PKS J1830-3602.

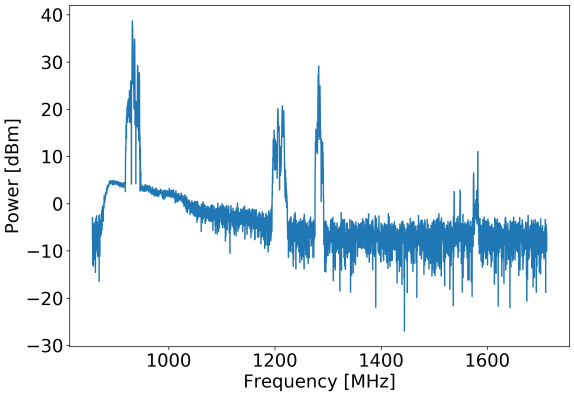

FIGURE A.14: This is the bandpass plot of a simulated observation of PKS J0006-8306.

Appendix B

Amplitude Spectrogram Comparison

In this appendix we show further examples of amplitude spectrograms from real and simulated data. The figures on the left-hand side of the page are real observation data and those on the right-hand side are simulated data.

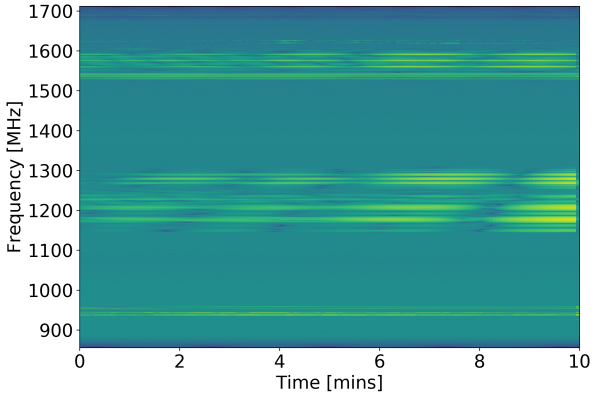

FIGURE B.1: This is the amplitude spectrogram of a real observation of NVSS J120624+641337.

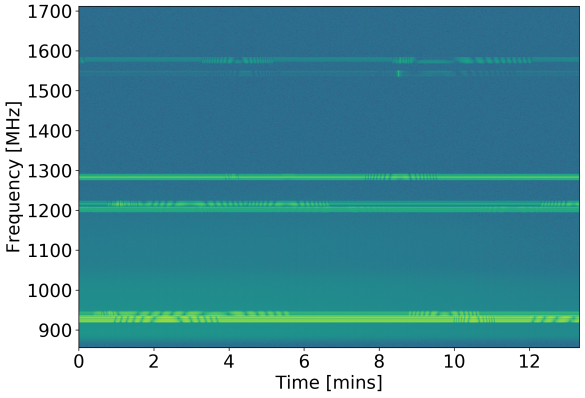

FIGURE B.2: This is the amplitude spectrogram of a simulated observation of PKS J0006-8306.

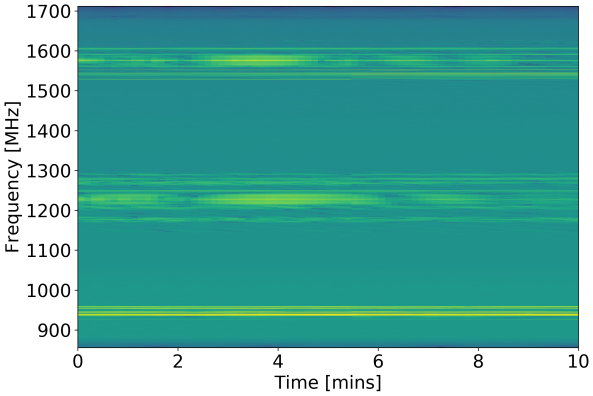

FIGURE B.3: This is the amplitude spectrogram of a real observation of PKS J1830-3602.

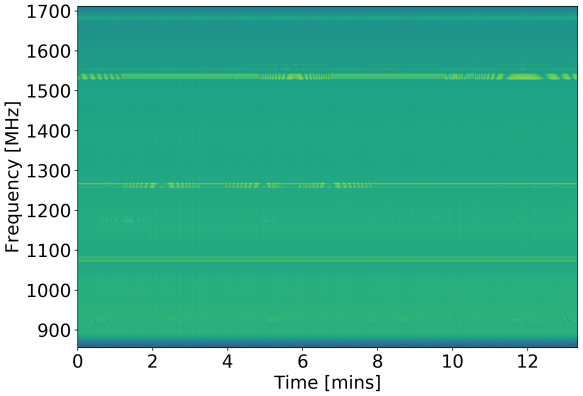

FIGURE B.4: This is the amplitude spectrogram of a simulated observation of PKS J0006-0004.

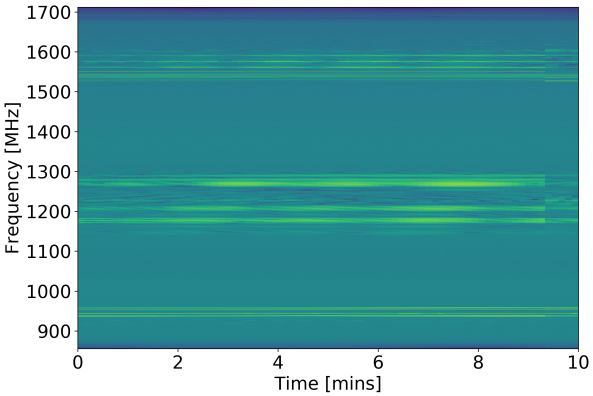

FIGURE B.5: This is the amplitude spectrogram of a real observation of PKS 1934-6342.

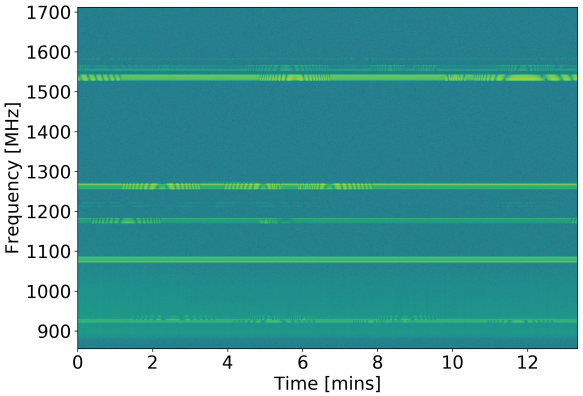

FIGURE B.6: This is the amplitude spectrogram of a simulated observation of PKS J0006-0004.

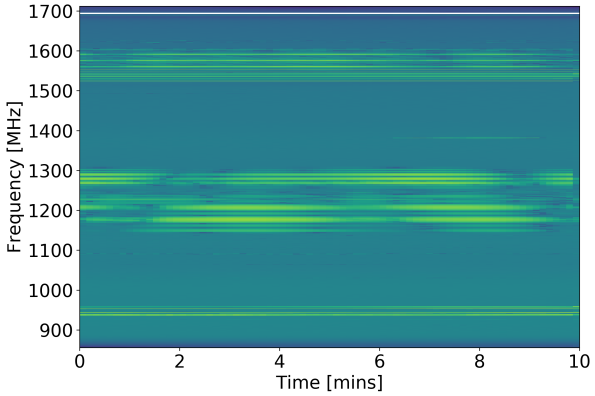

FIGURE B.7: This is the amplitude spectrogram of a real observation of NVSS J120624+641337.

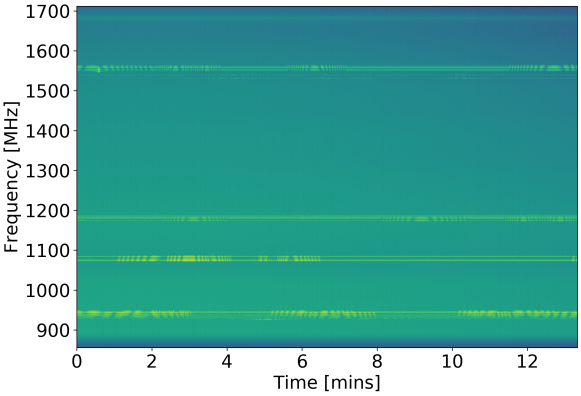

FIGURE B.8: This is the amplitude spectrogram of a simulated observation of PKS J0006-0004.

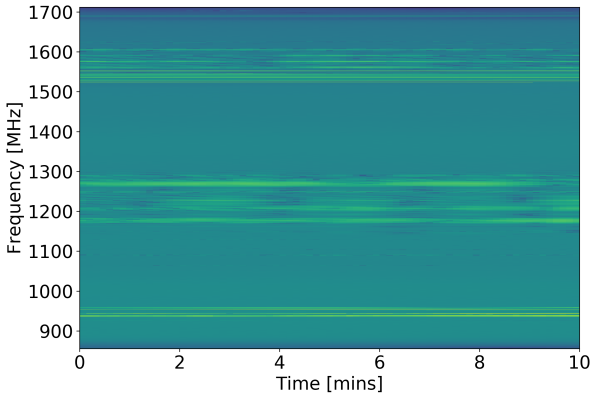

FIGURE B.9: This is the amplitude spectrogram of a real observation of NVSS J120624+641337.

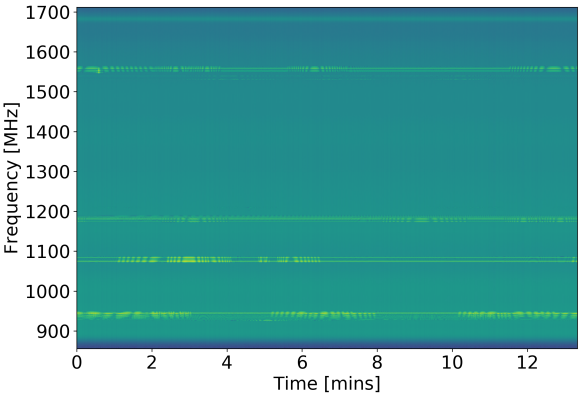

FIGURE B.10: This is the amplitude spectrogram of a simulated observation of PKS J0006-0004.

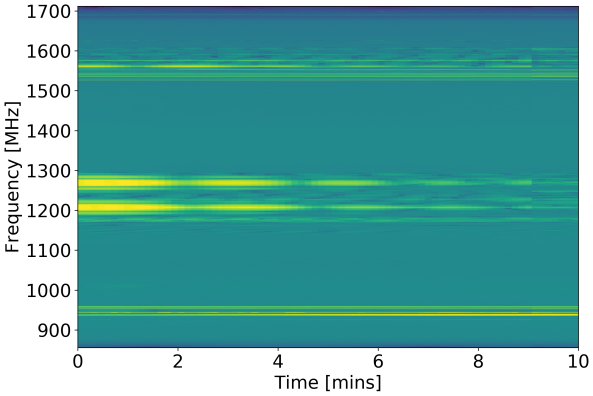

FIGURE B.11: This is the amplitude spectrogram of a real observation of PKS 1934-6342.

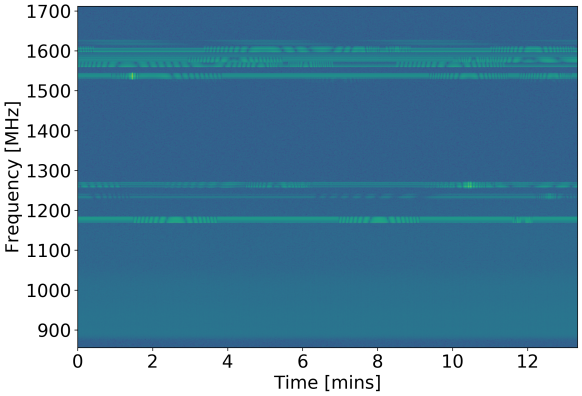

FIGURE B.12: This is the amplitude spectrogram of a simulated observation of PKS J0006-0004.

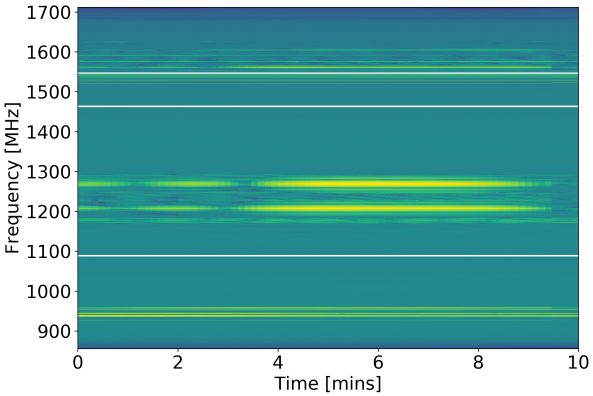

FIGURE B.13: This is the amplitude spectrogram of a real observation of PKS 1934-6342.

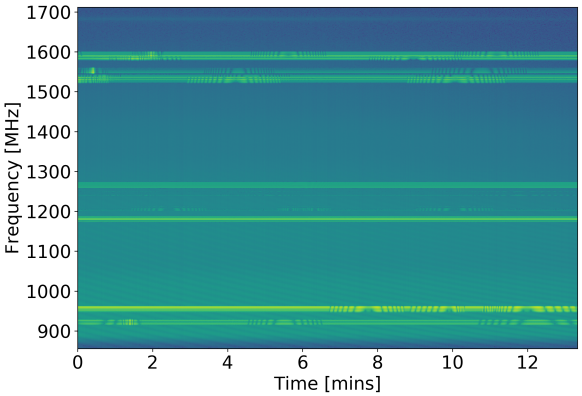

FIGURE B.14: This is the amplitude spectrogram of a simulated observation of PKS J0050-0929.

Appendix C

Dask Configuration File

This appendix contains a table describing the configuration file options for the second version of RFIsim.

TABLE C.1: Configuration file options and their descriptions.

Section-Subsection	Option	Description
`observation`	`target`	The right ascension and declination of the observation target.
`observation`	`start_date`	The date on which the observation should start. This should be given in dd/mm/yyyy format.
`observation`	`start_time`	The local time at which the observation should start. This should be given in HH:MM:SS format. `transit` can be given as an option. In this case the start time is chosen such that the transit of the target source occurs exactly half way through the observation.
`observation`	`duration`	The duration of the observation. This should be given in HH:MM:SS format. It can also be provided as an integer number of seconds.
`observation`	`time_steps`	The number of time steps to simulate. If this option is chosen then it supersedes the `duration` option.
`observation`	`int_time`	The number of seconds between each time step.

`observation`	`auto_corrs`	Whether to calculate the auto-correlations as well as the cross-correlations.
`telescope`	`GPS_coords`	The GPS coordinates as [latitude, longitude] for the reference antenna of the array.
`telescope`	`ENU_coords`	The East-North-Up (ENU) coordinates of each antenna relative to the reference antenna of the array. The reference antenna is defined to be at ENU=(0,0,0). This should be the path to a text file, csv file or npy binary file.
`telescope`	`frequencies`	The list of frequencies to use in the simulation. This should be the path to a text file, csv file or npy binary file.
`telescope`	`bandpass_xx`	The bandpass values for the xx-component, G_{00} of the gains matrix. This data should have the shape (N_ant, N_freq) where N_ant is the length of the data in `ENU_coords` and N_freq is the length of the data given in `frequencies`.
`telescope`	`bandpass_xy`	The bandpass values for the xy-component, G_{01} of the gains matrix. This data should have the shape (N_ant, N_freq) where N_ant is the length of the data in `ENU_coords` and N_freq is the length of the data given in `frequencies`.
`telescope`	`bandpass_yx`	The bandpass values for the yx-component, G_{10} of the gains matrix. This data should have the shape (N_ant, N_freq) where N_ant is the length of the data in `ENU_coords` and N_freq is the length of the data given in `frequencies`.
`telescope`	`bandpass_yy`	The bandpass values for the yy-component, G_{11} of the gains matrix. This data should have the shape (N_ant, N_freq) where N_ant is the length of the data in `ENU_coords` and N_freq is the length of the data given in `frequencies`.

`telescope`	`beam_xx`	The name of the function that should be used for the primary beam model for the xx-component.
`telescope`	`beam_xy`	The name of the function that should be used for the primary beam model for the xy-component.
`telescope`	`beam_yx`	The name of the function that should be used for the primary beam model for the yx-component.
`telescope`	`beam_yy`	The name of the function that should be used for the primary beam model for the yy-component.
`telescope`	`gain`	This is the overall gain value to be used when the bandpasses have been normalised.
`telescope`	`time_var`	The time scale on which the gains should maximally vary. This should be in HH:MM:SS format.
`astronomical`	`sky_model`	The path to the sky model data. This should be provided as a text file, csv file or a npy binary file.
`astronomical`	`radius`	The radius in decimal degrees around the target that sources from the sky model should be included.
`rfi`	`freq_dist`	The probability distribution over the frequency axis giving the likelihood that RFI will be present in a given frequency channel. This should be provided as a text file, csv file or a npy binary file.
`rfi-satellites`	`tle_dir`	The path to the directory in which the TLE files are stored for all the satellites.
`rfi-satellites`	`max_sats`	The maximum number of satellites to be included in the simulation.
`rfi-satellites`	`time_var`	The time scale of amplitude variations. This should be in HH:MM:SS format.
`rfi-satellites`	`freqs`	The frequencies at which satellites are emitting radio waves. (Not yet implemented.)

`rfi-cell_towers`	`GPS_coords`	The path to the directory in which the TLE files are stored for all the satellites. (Not yet implemented.)
`rfi-cell_towers`	`max_towers`	The maximum number of cellphone towers to be included in the simulation. (Not yet implemented.)
`rfi-cell_towers`	`time_var`	The time scale of amplitude variations. This should be in HH:MM:SS format. (Not yet implemented.)
`rfi-cell_towers`	`freqs`	The frequencies at which cell towers are emitting radio waves. (Not yet implemented.)
`rfi-planes`	`paths`	The path to the directory in which the TLE files are stored for all the satellites. (Not yet implemented.)
`rfi-planes`	`max_planes`	The maximum number of satellites to be included in the simulation. (Not yet implemented.)
`rfi-planes`	`time_var`	The time scale of amplitude variations. This should be in HH:MM:SS format. (Not yet implemented.)
`rfi-planes`	`freqs`	The frequencies at which planes' distance measuring equipment (DME) are emitting radio waves. (Not yet implemented.)
`process`	`time_chunk`	The size, on the time axis, of a chunk to be used in a subcalculation. This can be left to Dask but in some cases is best left to the user to decide. This is an integer value. -1 will assume the entire axis.
`process`	`freq_chunk`	The size, on the frequency axis, of a chunk to be used in a subcalculation. This can be left to Dask but in some cases is best left to the user to decide. This is an integer value. -1 will assume the entire axis.
`process`	`input_dir`	The path to the directory in which all input data is stored. All data being used as input defined above should be relative to this directory path.

`process`	`output_dir`	The path to the directory in which all output data will be stored. This includes any data produced in intermediary steps.

Bibliography

[1] J. Akeret et al. "Hide & Seek: End-to-End Packages to Simulate and Process Radio Survey Data". In: *Astronomy and Computing* 18 (2017), pp. 8–17.

[2] J. Akeret et al. "SEEK: Signal Extraction and Emission Kartographer". In: *Astrophysics Source Code Library* (2016).

[3] K. Asad et al. "Primary Beam Effects of Radio Astronomy Antennas - II. Modelling the Meerkat L-Band Beam". In: (Apr. 2019).

[4] R. Athreya. "A New Approach to Mitigation of Radio Frequency Interference in Interferometric Data". In: *The Astrophysical Journal* 696.1 (2009), p. 885.

[5] J.A. Ávila-Rodríguez. *BeiDou Signal Plan*. 2011. URL: https://gssc.esa.int/navipedia/index.php/BeiDou_Signal_Plan (visited on 05/30/2020).

[6] J.A. Ávila-Rodríguez. *Galileo Signal Plan*. 2011. URL: https://gssc.esa.int/navipedia/index.php/Galileo_Signal_Plan.

[7] J.A. Ávila-Rodríguez. *GLONASS Signal Plan*. 2011. URL: https://gssc.esa.int/navipedia/index.php/GLONASS_Signal_Plan.

[8] J.A. Ávila-Rodríguez. *GPS Signal*. 2011. URL: https://gssc.esa.int/navipedia/index.php/GPS_Signal_Plan.

[9] W.A. Baan. "RFI Mitigation in Radio Astronomy". In: *2011 XXXth URSI General Assembly and Scientific Symposium*. IEEE. 2011, pp. 1–2.

[10] P.J. Bevelacqua. *Antenna-Theory*. 2016. URL: http://www.antenna-theory.com/ (visited on 05/29/2020).

[11] A-J Boonstra and A-J Van der Veen. "Gain calibration methods for radio telescope arrays". In: *IEEE Transactions on Signal Processing* 51.1 (2003), pp. 25–38.

[12] M. Born and E. Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Elsevier, 2013.

[13] J.A. Bruder et al. "IEEE Standard for Letter Designations for Radar-Frequency Bands". In: *IEEE Aerospace & Electronic Systems Society* (2003), pp. 1–3.

[14] B.F. Burke, F. Graham-Smith, and P.N. Wilkinson. *An Introduction to Radio Astronomy*. Cambridge University Press, 2019.

[15] C.L. Carilli and P.D. Barthel. "Cygnus A". In: *The astronomy and astrophysics review* 7.1 (1996), pp. 1–54.

[16] A. Clegg. *RF Dialects: Understanding Each Other's Technical Lingo*. URL: http://www.iucaf.org/SSS2010/presentations/day2/Clegg(Units).ppt.

[17] *ComRADGIS*. URL: http://comradgis.kat.ac.za:8000/comradgis.

[18] J.J. Condon. "Radio Emission From Normal Galaxies". In: *Annual review of astronomy and astrophysics* 30.1 (1992), pp. 575–611.

[19] J.J. Condon and S.M. Ransom. *Essential radio astronomy*. Vol. 2. Princeton University Press, 2016.

[20] J.J. Condon et al. "The NRAO VLA Sky Survey". In: *The Astronomical Journal* 115.5 (May 1998), pp. 1693–1716. DOI: 10.1086/300337. URL: https://doi.org/10.1086/300337.

[21] National Research Council et al. *Handbook of Frequency Allocations and Spectrum Protection for Scientific Uses*. National Academies Press, 2007.

[22] Dave. *Circular Polarization*. 2010. URL: https://commons.wikimedia.org/wiki/File:Circular.Polarization.Circularly.Polarized.Light_Without.Components_Right.Handed.svg.

[23] P.E. Dewdney et al. "The Square Kilometre Array". In: *Proceedings of the IEEE* 97.8 (2009), pp. 1482–1496.

[24] G.M. Djuknic. *Method of Measuring a Pattern of Electromagnetic Radiation*. US Patent 6,657,596. Dec. 2003.

[25] The Editors of Encyclopaedia Britannica. "Radio source". In: (2019). URL: https://www.britannica.com/science/radio-source.

[26] D.O. Gerald J.K. Moernaut. "GNSS Antennas". In: *GPS World* (Feb. 2009).

[27] J.P. Hamaker. "Understanding Radio Polarimetry-IV. the Full-Coherency Analogue of Scalar Self-Calibration: Self-Alignment, Dynamic Range and Polarimetric Fidelity". In: *Astronomy and Astrophysics supplement series* 143.3 (2000), pp. 515–534.

[28] J.P. Hamaker, J.D. Bregman, and R.J. Sault. "Understanding Radio Polarimetry. I. Mathematical Foundations". In: *Astronomy and Astrophysics Supplement Series* 117.1 (1996), pp. 137–147. DOI: 10.1051/aas:1996146.

[29] H. Hellwig et al. "Measurement of the Unperturbed Hydrogen Hyperfine Transition Frequency". In: *IEEE Transactions on Instrumentation and Measurement* 19.4 (1970), pp. 200–209.

[30] F.R. Hoots and R.L. Roehrich. *Models for Propagation of NORAD Element Sets*. Tech. rep. Aerospace Defense Command Peterson AFB CO Office of Astrodynamics, 1980.

[31] Iridium Communications Inc. 2012. URL: https://iridium.it/en/iridium.htm (visited on 05/30/2020).

[32] Iridium Communications Inc. 2012. URL: https://www.iridium.com/products/antcom-iridium-gps-antenna/ (visited on 05/30/2020).

[33] Telecomm Strategies Inc. *Inmarsat 4F2 Attachment 1. Technical Description*. 2005. URL: https://licensing.fcc.gov/myibfs/download.do?attachment_key=-94644 (visited on 02/03/2020).

[34] *Interferometers I*. URL: https://www.cv.nrao.edu/course/astr534/Interferometers1.html.

[35] *Interferometers II*. URL: https://www.cv.nrao.edu/course/astr534/Interferometers2.html.

[36] J.L. Jonas. "MeerKAT - The South African Array With Composite Dishes and Wide-Band Single Pixel Feeds". In: *Proceedings of the IEEE* 97.8 (2009), pp. 1522–1530.

[37] A.J. Kemball and M.H. Wieringa. "MeasurementSet Definition Version 2.0". In: *URL:http://casa.nrao. edu/Memos/229.html* (2000).

[38] M.I. Large. "The Mechanisms of Radio Emission". In: *Radiotekhnika* (1963), pp. 126–137.

[39] *Linear Polarization*. URL: https://www.pngwing.com/en/free-png-xpapb.

[40] T. Mauch et al. "SUMSS : A Wide-Field Radio Imaging Survey of the Southern Sky – II . the Source Catalogue". In: 1130.2003 (2006), pp. 1117–1130.

[41] D. Moulton. *Stokes Paramters*. Jan. 2008. URL: https://commons.wikimedia.org/wiki/File:StokesParameters.png.

[42] Jan E Noordam and Oleg M Smirnov. "The MeqTrees software system and its use for third-generation calibration of radio interferometers". In: *Astronomy & Astrophysics* 524 (2010), A61.

[43] A.R. Offringa. "AOFlagger: RFI Software". In: *Astrophysics Source Code Library* (2010).

[44] A.A. Penzias and R.W. Wilson. "A Measurement of Excess Antenna Temperature at 4080 Mc/s". In: *The Astrophysical Journal* 142 (1965), pp. 419–421.

[45] S. Perkins et al. "Montblanc: GPU Accelerated Radio Interferometer Measurement Equations in Support of Bayesian Inference for Radio Observations". In: *CoRR* abs/1501.07719 (2015). arXiv: 1501.07719. URL: http://arxiv.org/abs/1501.07719.

[46] R.A. Perley et al. "The Expanded Very Large Array: A New Telescope for New Science". In: *The Astrophysical Journal Letters* 739.1 (2011), p. L1.

[47] M. Plank. *The Theory of Heat Radiation*. Dover Publication, New York, 1914.

[48] S. Pople. *Advanced Physics Through Diagrams*. Oxford University Press, 2001.

[49] U. Rau, R. Selina, and A. Erickson. "RFI Mitigation for the ngVLA: A Cost-Benefit Analysis ngVLA. Memo# 70". In: (2019).

[50] J.E. Reynolds. "A Revised Flux Scale for the AT Compact Array". In: *ATNF Internal* (1994).

[51] *RFI L-band*. 2020. URL: https://skaafrica.atlassian.net/servicedesk/customer/portal/1/topic/bc9d6ad2-8321-4e13-a97a-d19d6d019a1c/article/305332225.

[52] B.C. Rhodes. "PyEphem: Astronomical Ephemeris for Python". In: *ascl* (2011), ascl–1112.

[53] B.C. Rhodes. "Skyfield: High Precision Research-Grade Positions for Planets and Earth Satellites Generator". In: *ascl* (2019), ascl–1907.

[54] United States Occupational Safety and Health Administration. *Electromagnetic Radiation and How It Affects Your Instruments*. 1990. URL: https://www.osha.gov/SLTC/radiofrequencyradiation/electromagnetic_fieldmemo/electromagnetic.html (visited on 05/30/2020).

[55] J. Sanz Subirana, J.M. Juan-Zornoza, and M. Hernandez-Pajares. *GNSS Signal*. 2011. URL: https://gssc.esa.int/navipedia/index.php/GNSS_signal.

[56] South African Department of Science and Technology. "Astronomy Geographic Advantage Act". In: *Government Gazette* 516.31157 (2007).

[57] S. Sekhar and R. Athreya. "Two Procedures to Flag Radio Frequency Interference in the UV Plane". In: *The Astronomical Journal* 156.1 (2018), p. 9.

[58] I. Sihlangu. "The MeerKAT Radio Frequency Interference Environment". PhD thesis. Faculty of Science, 2019.

[59] I. Sihlangu, N. Oozeer, and B.A. Bassett. "Multidimensional RFI Framework for Characterising Radio Astronomy Observatories". In: *arXiv preprint arXiv:2008.08877* (2020).

[60] O.M. Smirnov. "Revisiting the Radio Interferometer Measurement Equation: I. a Full-Sky Jones Formalism". In: *Astronomy and Astrophysics* 527.14 (2011). ISSN: 00046361. DOI: 10.1051/0004-6361/201016082. arXiv: 1101.1764.

[61] O.M. Smirnov. "Revisiting the Radio Interferometer Measurement Equation: II. Calibration and Direction-Dependent Effects". In: *Astronomy and Astrophysics* 527.14 (2011), pp. 1–16. ISSN: 00046361. DOI: 10.1051/0004-6361/201116434. arXiv: 1101.1765.

[62] G.G. Stokes. "On the Composition and Resolution of Streams of Polarized Light From Different Sources". In: *TCaPS* 9 (1851), p. 399.

[63] G. Swarup et al. "The Giant Metre-Wave Radio Telescope". In: *Current science* 60.2 (1991), pp. 95–105.

[64] L.G. Taff. "Computational Spherical Astronomy". In: *A Wiley-Interscience Publication* (1981).

[65] Z. Tan et al. "Positioning Using IRIDIUM Satellite Signals of Opportunity in Weak Signal Environment". In: *Electronics* 9.1 (2020), p. 37.

[66] G.B. Taylor, C.L. Carilli, and R.A. Perley. "Synthesis Imaging in Radio Astronomy II". In: *ASPC* 180 (1999).

[67] R.A. Thompson, J.M. Moran, and G.W. Swenson Jr. *Interferometry and Synthesis in Radio Astronomy*. Springer Nature, 2017.

[68] Tiltec. *File:Hydrogen-SpinFlip.svg*. 2009. URL: https://commons.wikimedia.org/wiki/File:Hydrogen-SpinFlip.svg.

[69] A. Vafaei Sadr et al. "Deep Learning Improves Identification of Radio Frequency Interference". In: *Monthly Notices of the Royal Astronomical Society* 499.1 (2020), pp. 379–390.

[70] B. Vollmer et al. "The SPECFIND V2. 0 Catalogue of Radio Cross-Identifications and Spectra-Specfind Meets the Virtual Observatory". In: *Astronomy & Astrophysics* 511 (2010), A53.

[71] E.E. Vos et al. "A Generative Machine Learning Approach to RFI Mitigation for Radio Astronomy". In: *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2019, pp. 1–6.

[72] T. Wei et al. "Hybrid Satellite-Terrestrial Communication Networks for the Maritime Internet of Things: Key Technologies, Opportunities, and Challenges". In: *arXiv preprint arXiv:1903.11814* (2019).