21cm foregrounds and polarization leakage: a user’s guide on cleaning and mitigation strategies

Steven Cunnington,1,* Melis O. Irfan,2,3 Isabella P. Carucci,3 Alkistis Pourtsidou,1,2 Jérôme Bobin3

1School of Physics and Astronomy, Queen Mary University of London, Mile End Road, London E1 4NS, UK
2Department of Physics & Astronomy, University of the Western Cape, Cape Town 7535, South Africa
3AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France

Accepted XXX. Received YYY; in original form ZZZ.

ABSTRACT

The success of HⅠ intensity mapping is largely dependent on how well 21cm foreground contamination can be controlled. In order to progress our understanding further, we present a range of simulated foreground data from four different ~ 3000 deg2 sky regions, with and without effects from polarization leakage. Combining these with underlying cosmological HⅠ simulations creates a range of single-dish intensity mapping test cases that require different foreground treatments. This allows us to conduct the most generalized study to date into 21cm foregrounds and their cleaning techniques for the post-reionization era. We first provide a pedagogical review of the most commonly used blind foreground removal techniques (PCA/SVD, FastICA, GMCA). We also trial a non-blind parametric fitting technique and discuss potential hybridization of methods. We highlight the similarities and differences in these techniques finding that the blind methods produce near equivalent results, and we explain the fundamental reasons for this. The simulations allow an exact decomposition of the resulting cleaned data and we analyse the contribution from foreground residuals. Our results demonstrate that polarized foreground residuals should be generally subdominant to HⅠ on small scales (k > 0.1 h Mpc⁻¹). However, on larger scales, results are more region dependent. In some cases, aggressive cleans severely damp HⅠ power but still leave dominant foreground residuals. We also demonstrate the gain from cross-correlations with optical galaxy surveys, where extreme levels of residual foregrounds can be circumvented. However, these residuals still contribute to errors and we discuss the optimal balance between over- and under-cleaning.

Key words: cosmology: large scale structure of Universe – cosmology: observations – radio lines: general – methods: data analysis – methods: statistical

1 INTRODUCTION

Mapping the cosmic neutral hydrogen (HⅠ) from the post-reionization era is as an excellent way to probe the large-scale structure of the Universe. By mapping the redshifted 21cm signal from HⅠ residing within galaxies, the underlying 3-dimensional matter density can be inferred and cosmological information can be extracted, in a similar fashion to optical galaxy surveys. A novel technique allowing to do this is intensity mapping (Bharadwaj et al. 2001; Battye et al. 2004; Chang et al. 2008).

In this work we focus on so-called single-dish intensity mapping (Battye et al. 2013), which uses the auto-correlation data of a telescope array (e.g. the Square Kilometer Array (SKA) – SKA Cosmology SWG et al. (2020)), as opposed to the more traditional interferometric mode of operation. Unlike a conventional spectroscopic galaxy survey that has to resolve galaxies and conduct spectroscopy to infer a redshift with sufficient precision, intensity mapping does not resolve galaxies but records the diffuse, unresolved HⅠ. This has the advantages of being able to rapidly observe very large volumes of the Universe, and is not as susceptible to high levels of shot-noise. The resulting maps have a relatively low-angular resolution due to the radio telescope beam, which is related to the dish diameter for single-dish observations. This damps the HⅠ power spectrum for modes perpendicular to the line-of-sight but despite this, many large cosmological scales can still be resolved. Furthermore, the spectroscopic resolution in these radio observations is excellent and thus modes can in principle be resolved to very small scales along the line-of-sight (Villaescusa-Navarro et al. 2017).

There are unique challenges to HⅠ intensity mapping which conventional galaxy surveys largely avoid. Whilst intensity mapping is unlikely to be limited by shot noise, there is instrumental (thermal) noise. Assuming enough observation time, and a well controlled system temperature, this noise should be sub-dominant relative to the HⅠ signal and well approximated as Gaussian white-noise. As noted in Harper et al. (2018), complications from other systematics such as 1/f noise pose a more complex challenge. However, analysis of recent HⅠ intensity mapping data from MeerKAT suggest this should be a controllable systematic (Li et al. 2020a). A further issue is contamination from human-made Radio Frequency Interference (RFI) such as global navigation satellites (Harper & Dickinson 2018).

Another major challenge, and the focus of this paper, is foreground...
contamination from astrophysical sources and how they interact with the telescope. The main source of 21cm foreground signals comes from Galactic synchrotron (sourced by cosmic-ray electrons accelerated by the Galactic magnetic field), free-free emission (sourced by free electrons scattering off ions largely within our Galaxy but weaker emission can also come from extragalactic sources), and point-sources (extragalactic objects emitting strong radio signals e.g. AGNs).

Some of these foregrounds can be orders of magnitude more dominant than the H\textsubscript{i} signal, but their spectrum evolves slowly through frequency. This is in contrast to cosmic H\textsubscript{i}, which varies with redshift and thus oscillates to a near-Gaussian approximation with frequency. The fact that the raw-foregrounds are smooth continuums through frequency means they can in principle be removed with modelling or source separation (Liu & Tegmark 2011; Wolz et al. 2014; Shaw et al. 2015; Alonso et al. 2015). However, large-scale foreground signals typically have some degeneracy with the cosmological H\textsubscript{i}, and require some form of treatment in order not to bias power spectra measurements and cosmological parameter estimation (Wolz et al. 2014; Cunnington et al. 2020a; Soares et al. 2020).

In reality the challenge of separating the H\textsubscript{i} signal from the foregrounds becomes even more complicated by the foreground’s response to the instrument. Unless instrumental effects from spectral response and chromaticity from the beam are controlled, the spectral smoothness of the foregrounds can be degraded. The most potentially concerning instrumental effect is from polarization leakage (Jelic et al. 2008, 2010; Moore et al. 2013). Cosmological H\textsubscript{i} is unpolarized and thus attempts are made to sufficiently calibrate telescopes to avoid polarized signals (Liao et al. 2016). However, a sufficient level of calibration is not guaranteed and even a small amount of polarized synchrotron leaking into the observational data can dominate the H\textsubscript{i} signal. Furthermore, the Faraday rotation that interferes with the polarization state is expected to be frequency-dependent, which means these leaked signals will not have such a smooth spectrum and will be harder to single out (Carucci et al. 2020a).

Previous investigations into foreground cleaning generally involve introducing a single set of foreground simulations which cleaning techniques can then be tuned to. However, these rarely include instrumental response effects such as polarization leakage (although there are some exceptions e.g. Shaw et al. (2015); Carucci et al. (2020a)). These idealized simulated foregrounds require much less aggressive cleans than what is usually needed in real data analyses from pathfinder intensity mapping experiments (Masui et al. 2013; Switzer et al. 2013; Wolz et al. 2017; Anderson et al. 2018).

In this work, we add an extra layer of complication. Whilst we cannot yet provide full end-to-end simulations that directly mimic a realistic experiment, we are able to present results where it is necessary to use aggressive foreground cleans akin to those employed in real data. By using a range of different sky regions, both with and without polarization leakage effects, we create a variety of cases in which different levels of foreground cleaning are required and different problems arise. We introduce and apply a range of different foreground cleaning methods and compare the results. Since we are dealing with simulations, we have full control over the data and provide analysis into problems concerning damping of H\textsubscript{i} power as well as the biases and errors introduced from foreground residuals.

This paper is structured as follows. In Section 2 we present our foreground simulations and the sky regions we consider. In Section 3 we present our method for producing the underlying cosmological H\textsubscript{i} intensity maps (and the accompanying optical galaxy data for cross-correlations) that we aim to recover. We provide a generalized review of foreground cleaning methods in Section 4 and identify the exact methods we apply to our simulated data. We then present our results in Section 5 and conclude in Section 6.

2 FOREGROUND SIMULATIONS

We begin by identifying four different regions on the sky which should exhibit different foreground behaviour or have certain interest to the cosmological community. We desire each of our regions to have the same size that is largely dictated by the size of our $1\ h^{-3}\mathrm{Gpc}^3$ H\textsubscript{i} simulation box (to be described in detail in Section 3). At the central redshift of this simulation ($z = 0.39$) these dimensions are equivalent to a sky size of $54.1 \times 54.1 \sim 2972\ \text{deg}^2$. This is similar to the sky area proposed in MeerKLASS (Santos et al. 2017),
a wide area survey using the MeerKAT telescope, which is the pathfinder for the Square Kilometre Array (SKA). We choose a frequency range of 899 – 1184 MHz (0.2 < z < 0.58), again consistent with a MeerKAT-like observation performed in the L-band. The four sky regions we investigate are:

[1] **Stripe 82**: A small, 300 deg² field imaged numerous times by galaxy surveys. To ensure consistency in sky sizes, this region is a 2927 deg² patch centred on the Stripe 82 field.

[2] **Galactic Plane**: Centre of our galaxy where combined emission from all foregrounds is expected to be highest. Typically avoided by surveys but we include this to create a test-case with a strong source of contamination.

[3] **South Celestial Pole (SCP)**: Low declination region away from Galactic Plane where combined emission from all foregrounds is expected to be low.

[4] **Barnard’s Loop**: An area expected to be dominated by free-free emission to allow investigation into the impact from this foreground.

These regions are outlined in Figure 1 on the full-sky (top-map) and individually shown in the flattened maps (bottom-row). We assume the maps are sufficiently small in size that they can be projected onto a Cartesian grid with minimal distortion. For each of these regions we simulate effects from polarization leakage, which we discuss further in Section 2.4.

The total observed temperature data are a combination of the cosmological H signal, the foregrounds and instrumental noise, all binned into pixels whose position is defined by \( \theta \) at each frequency channel \( v \):

\[
\delta T_{\text{obs}}(v, \theta) = \delta T_{\text{Hi}}(v, \theta) + \delta T_{\text{FG}}(v, \theta) + \delta T_{\text{noise}}(v, \theta). \tag{1}
\]

In this work, we will vary \( \delta T_{\text{FG}} \) between each region whilst keeping the other components fixed. The foreground signal can be further decomposed into the contributions from the different sources of foregrounds i.e. Galactic synchrotron emission, Galactic free-free emission, extragalactic point sources and polarization leakage:

\[
\delta T_{\text{FG}} = \delta T_{\text{syn}} + \delta T_{\text{free}} + \delta T_{\text{point}} + \delta T_{\text{pol}}. \tag{2}
\]

We introduce our simulation approach for each of these components in this section (except the \( \text{Hi} \) contribution which is discussed in Section 3). A full-sky realisation of \( \delta T_{\text{FG}} \) is openly available in Carucci et al. (2020b).

We chose the frequency range 899 – 1184 MHz and separate this range into 285 measurement bands. The effective resolution is determined by the beam equation given by

\[
\theta_{\text{FWHM}} = \frac{1.22c}{vD_{\text{dish}}}, \tag{3}
\]

where \( c \) is the speed of light and we assume a dish size of \( D_{\text{dish}} = 15 \text{m} \), which will be the size of the SKA-MID dishes and is approximately equivalent to MeerKAT’s dishes too. Since the beam size is dependent on the frequency \( v \), each channel is smoothed by a different amount. However, foreground removal algorithms (discussed in Section 4) perform better on data with a common resolution. We therefore smooth the intensity maps to a constant beam size given by the minimum frequency \( v_{\text{min}} = 899 \text{MHz} \), which from Equation 3 gives an effective resolution for our maps of \( \theta_{\text{FWHM}} = 1.55 \text{deg} \). We chose to create our simulations at \( N_{\text{side}} = 2048 \) and then to interpolate our 54.1 x 54.1 patches onto 256 x 256 pixel arrays.

We make use of the Planck Legacy Archive\(^2\) FFP10 simulations within our simulations and, as they are given in \( T_{\text{CMB}} \), the following conversion to the Rayleigh-Jeans regime is used:

\[
T_{\text{RJ}} = \frac{x^2 e^x}{(e^x - 1)^2} T_{\text{CMB}}, \tag{4}
\]

where \( x = h \nu/k_B T_{\text{CMB}} \), with \( h \) the Planck constant and \( k_B \) the Boltzmann constant.

### 2.1 Simulated Synchrotron Emission

We use the FFP10 simulations of synchrotron emission at 217 and 353 GHz for our purposes as these maps are provided at \( N_{\text{side}} = 2048 \). These maps are formed from the source-subtracted and destriped 0.408 GHz map (Remazeilles et al. 2015). Despite the 0.408 GHz survey data having a resolution of 56 arcmin, Remazeilles et al. (2015) provide a \( N_{\text{side}} = 2048 \) version of the data by filling in the higher resolution detail with a Gaussian random field.

These 217 and 353 GHz synchrotron maps can be used to determine the synchrotron spectral index map at \( N_{\text{side}} = 2048 \). The spectral index map used by FFP10 is the ‘Model 4’ synchrotron spectral index map of Miville-Deschênes et al. (2008), which has a resolution of ~ 5 degrees. This map was formed from 0.408 GHz intensity data and 23 GHz polarization data. However, as we are simply trying to determine the accuracy of our foreground mitigation strategies, the accuracy of the synchrotron spectral index map does not come into consideration.

We will however, need a higher resolution view of the synchrotron spectral index than 5 degrees and so we also choose to fill in the higher resolution detail with a Gaussian random field. Taking the synchrotron multipole scaling relation from Santos et al. (2005), our \( N_{\text{side}} = 2048 \) synchrotron spectral index map is constructed as

\[
\beta_{\text{sy}} = \beta_{\text{model 4}} + \beta_{\text{SS}}, \tag{5}
\]

where

\[
C_\ell^{\beta_{\text{SS}}} = 7 \times 10^{-6} \left( \frac{1000}{\ell} \right)^{2.4} \left( \frac{v_1^2}{v_1 v_2} \right)^{2.8} \exp \left( - \log \left( \frac{v_1}{v_2} \right)^2 \right), \tag{6}
\]

where \( v_1 \) is 130 MHz, \( v_2 \) is 580 MHz and \( v_2 \) is 1000 MHz. Our Gaussian random field is identical to that found in Santos et al. (2005) with the exception of the amplitude, which we alter to suit the magnitude of the synchrotron spectral index as opposed to the emission amplitude.

### 2.2 Simulated Free-Free Emission

We take our simulated free-free amplitude \( a_{\text{ff}} \) from the FFP10 217 GHz free-free simulation at \( N_{\text{side}} = 2048 \). This map is a composite of the Dickinson et al. (2003) free-free template and the WMAP MEM free-free templates; the details of which can be found in (Miville-Deschênes et al. 2008). Our free-free emission is modelled by a power law

\[
T_{\text{ff}}(v, \theta) = a_{\text{ff}}(\theta) \left( \frac{v}{v_0} \right)^{\beta_{\text{ff}}}, \tag{7}
\]

where the free-free spectral index is \( \beta_{\text{ff}} = -2.13 \) and constant across all map pixels.

\[^1\text{skatelescope.org}\]

\[^2\text{pla.esac.esa.int/pla}\]
2.3 Simulated Point Sources

We use the empirical model of Battye et al. (2013), which fits a polynomial to a selection of radio source counts at 1.4 GHz. The specific details of assembling this model of the Poisson and clustering contributions at 1.4 GHz can be found in Olivarri (2018). Following the method of Olivarri et al. (2018), we then scale the 1.4 GHz point source map to our frequencies using a power law where the spectral index varies following a Gaussian distribution centred at -2.7, with a standard deviation of 0.2.

Battye et al. (2013) expect point sources over 10 mJy to be bright enough to be identified within the National Radio Astronomy Observatory Very Large Array Sky Survey (Lacy et al. 2019) and so, removed. In this work we consider a 100 mJy upper bound on source extraction.

2.4 Simulated Polarization Leakage

The magnetic fields within our Galaxy’s interstellar medium can cause Faraday rotation effects which change the polarization angles of light. If this were a consistent effect, it would not be hugely problematic for foreground consideration. However, Faraday rotation is a frequency-dependent effect as demonstrated by Jelic et al. (2010); Moore et al. (2013). If any spectrally fluctuating polarization intensity is leaked into the total intensity it would be difficult to subtract without large loss to the unpolarized H\textsc{i} signal. Further observational data analysis would be needed to compute this instrumental effect with the use of the CRIME\textsuperscript{3} software, which provides maps of Stokes Q emission at each frequency and we fix the polarization leakage to 0.5\% of the Stokes Q signal. Further observational data analysis would be needed to constrain a reasonable choice for this fraction, but given previous estimations our choice is sensible (Liao et al. 2016). Details for the rotation calculation of the Stokes Q/U signal such that this leakage will not remain constant across all the observational channels.

We simulate this instrumental effect with the use of the CRIME\textsuperscript{3} software, which provides maps of Stokes Q emission at each frequency and we fix the polarization leakage to 0.5\% of the Stokes Q signal. Further observational data analysis would be needed to constrain a reasonable choice for this fraction, but given previous estimations our choice is sensible (Liao et al. 2016). Details for the rotation calculation of the Stokes Q/U signal such that this leakage will not remain constant across all the observational channels.

2.5 Simulated Noise

In this work, only Gaussian noise is considered, with a zero mean and standard deviation of

\[
\sigma(v) = T_{\text{sys}}(v) \left( \delta_v \frac{\Omega_p}{3} \frac{N_{\text{dish}}}{n_{\text{pix}}} \right)^{-1/2},
\]

where \(\delta_v\) is the width of each frequency band (Hz), \(t_{\text{tot}}\) is the total survey time (s), \(N_{\text{dish}}\) is the number of dishes and \(\Omega_p\) are the pixel and survey solid angle, respectively (Alonso et al. 2014). For the pixel solid angle only the beam FWHM expressed in radians is required

\[
\Omega_p = 1.13 \theta_{\text{FWHM}}^2.
\]

while for the survey solid angle the fraction of the sky covered is needed. If the angular area of the observed sky \(A_{\text{sky}}\) is given in square degrees, we have

\[
\Omega_a = 4\pi \frac{A_{\text{sky}}}{41253}.
\]

The system temperature in each band \(T_{\text{sys}}(\nu)\) is a combination of the receiver noise temperature \(T_{\text{rec}}\) and the sky temperature \(T_{\text{rec}}\) and the sky temperature (Santos et al. 2015):

\[
T_{\text{sys}}(\nu) = 1.1 \times 60 (\frac{300}{\nu [\text{MHz}]})^{2.55} + T_{\text{rec}}.
\]

The specific receiver and survey properties used here are based on a MeerK\textsuperscript{L}ASS-like survey and are summarized in Table 1.

| Quantity | Value | 1 MHz | 1000 hrs | 64 | 25 K | 2927 deg\textsuperscript{2} |
|----------|-------|-------|----------|----|-----|----------------|
| \(\delta_v\) | MHz | | | | | |
| \(t_{\text{tot}}\) | hr | | | | | |
| \(N_{\text{dish}}\) | | | | | | |
| \(T_{\text{rec}}\) | K | | | | | |
| \(A_{\text{sky}}\) | deg\textsuperscript{2} | | | | | |

3 COSMOLOGICAL SIMULATIONS

We use the same simulated cosmological H\textsc{i} signal data for each patch. Specifically, we make use of the MultiDARK-PLANCK cosmological N-body simulation (Klypin et al. 2016), which evolved 3840\textsuperscript{3} dark-matter particles in a 1000\textsuperscript{3} h\textsuperscript{-3} Mpc\textsuperscript{3} volume with the adopted cosmology complying with Planck Collaboration et al. (2016). The cosmological parameters used are therefore \(\Omega_M = 0.307\), \(\Omega_b = 0.048\), \(\Omega_{\Lambda} = 0.693\), \(\sigma_8 = 0.823\), \(n_s = 0.96\) and Hubble parameter \(h = 0.678\). This data has been processed into the MultiDark-Galaxies data (Knebe et al. 2018), which are galaxy catalogues publicly available from the Skies & Universes web page\textsuperscript{4}. It is from these catalogues that we build the simulated H\textsc{i} intensity maps and an overlapping map of resolved optical galaxies.

Each snapshot from the MultiDark-Galaxies simulation represents a different redshift and evolved state of the cosmological density field and the galaxies therein. We opt to use the catalogues at \(z = 0.39\) and take this as the effective re\textsuperscript{2}dshift (\(z_{\text{eff}}\)) for our data. This is analogous to real surveys assuming a central effective redshift provided that the width of the bin is small enough so that cosmological quantities can be assumed constant within it.

We still need to assume some redshift range however, since we require a frequency range from which to produce the foregrounds. We therefore assume our data has redshift range of \(0.2 < z < 0.58\), which for the H\textsc{i} intensities with \(\nu = 1420\) MHz/(1 + z), will convert to a frequency range of \(899 < \nu < 1184\) MHz. This frequency range is probed with the L-band from the MeerKAT telescope and is thus representative of a near-term intensity mapping survey (Santos et al. 2017).

The MultiDark data we use are for a Cartesian box with galaxy coordinates in physical distances. We thus work in this Cartesian regime throughout this investigation. This is common practice in large-scale structure surveys, where either a small enough sky is surveyed that a flat-sky approximation is valid, or where curved sky effects are accounted for (Castorina & White 2018; Blake et al. 2018). At the effective redshift \(z_{\text{eff}} = 0.39\), the redshift range of \(0.2 < z < 0.58\) we assume for our data converts to a physical distance of 925 h\textsuperscript{-1} Mpc. We therefore trim the MultiDark data cube to this distance along one dimension, keeping the others the same. This results in a data cube with physical size \(L_x, L_y, L_z = 1000, 1000, 925\) h\textsuperscript{-1} Mpc where we use the convention that \(x\) and \(y\) are the angular dimensions perpendicular to the

\[\text{intensitymapping.physics.ox.ac.uk/CRIME.html}\]

\[\text{www.skiesanduniverses.org}\]
line-of-sight and $z$ is parallel to the line-of-sight. We use the plane-parallel approximation throughout. The data cube is gridded into volume-pixels (voxels), with $n_x = n_y = 256$ along the angular dimensions and $n_z = 285$ along the radial dimension. The choice of radial binning allows the 899 < $v$ < 1184 MHz frequency range we assume to have a frequency resolution of $\Delta v = 1$ MHz. As we have already mentioned, the approximate sky coverage of our data is just under 3000 deg$^2$, which is fairly representative of proposed intensity mapping surveys like MeerKLaSS (Santos et al. 2017).

3.1 H$\text{\textsc{i}}$ Intensity Maps

To produce the intensity maps from the MultiDark data we utilize the catalogue produced from applying the SAGE (Croton et al. 2016) semi-analytical model to the data. We summarise our method below for how we produce the intensity maps from the MultiDark-SAGE catalogue. For a more complete description of this process, we refer the reader to Cunnington et al. (2020b), where an identical methodology was employed. The MultiDark-SAGE catalogue is fully outlined in (Knebe et al. 2018).

Firstly the cold gass mass for each galaxy is converted into a H$\text{\textsc{i}}$ mass, which is then binned into the relevant voxel according to the galaxy’s coordinates. This gridded H$\text{\textsc{i}}$ mass is then converted to a H$\text{\textsc{i}}$ brightness temperature $T_{\text{H}}(x)$. Since intensity mapping surveys will detect signal down to the faintest emitters, it is common in simulations to rescale the $T_{\text{H}}$ temperature of the field up to a realistic (expected) value. This is required because simulations have finite capabilities and often do not resolve halos down to masses of $\sim 10^9 h^{-1} M_\odot$ where H$\text{\textsc{i}}$ is still predicted to reside (Villaescusa-Navarro et al. 2018; Spinelli et al. 2020). To determine this value we utilize the results of the GBT-WiggleZ cross-correlation analysis (Masui et al. 2013), where it was found that the H$\text{\textsc{i}}$ abundance is $\Omega_{\text{H}} b_{\text{H}} r = [4.3 \pm 1.1] \times 10^{-4}$, and assume it is constant with redshift. We also take the cross-correlation coefficient to be $r = 1$ and use a H$\text{\textsc{i}}$ bias fit from Villaescusa-Navarro et al. (2018). For our effective redshift this equates to $b_{\text{H}}(z_{\text{eff}}) = 1.105$.

Lastly, in order to emulate the effects from the radio telescope beam we smooth each channel using $\theta_{\text{FWHM}} = 1.55$ deg (as discussed in Section 2). The observable field for intensity mapping is the over-temperature field defined as

$$\delta T_{\text{H}}(z) = T_{\text{H}}(z) - \langle T_{\text{H}} \rangle = \langle T_{\text{H}} \rangle b_{\text{H}}(z) \delta_M(z),$$

where $\delta_M(z)$ is the underlying matter density.

3.2 Overlapping Optical Galaxy Data

We also utilize the MultiDark-Galaxies for creating an overlapping optical spectroscopic catalogue, which we will use for investigating cross-correlating techniques between H$\text{\textsc{i}}$ intensity maps and optical galaxy data. For this purpose we use the SAG (Cora 2006) semi-analytic model; that is because this catalogue has magnitude outputs for each of the SDSS $ugriz$ broad bands, which can be utilized to construct a realistic optical galaxy data set.

Whilst the MultiDark-SAG catalogue does also possess the cold gas mass outputs, and therefore could have also been used to produce the intensity maps, it has fewer galaxies ($\sim 3.8 \times 10^5$) compared with the MultiDark-SAGE catalogue ($\sim 7 \times 10^7$) at the snapshot redshift of $z = 0.39$. We prefer using MultiDark-SAGE for the intensity maps because it has a higher number of galaxies. Since both SAGE and SAG catalogues are generated from the same underlying MultiDark simulated density field, they should still produce sufficient cross-correlation signals. The MultiDark-SAG catalogue is fully outlined in (Knebe et al. 2018).

Optical galaxy surveys generically operate by constructing a catalogue of resolved galaxies whose luminosity is above some threshold determined by the telescope’s sensitivity. As a rather crude emulation of this method, which is sufficient for this investigation, we use the sum of the magnitudes from the five SDSS $ugriz$ bands and select the highest total magnitudes from the simulation until a target $N(z)$ redshift distribution is achieved. Following Mandelbaum et al. (2011), we construct a realistic target distribution by assuming a double Gaussian where 77.6% of the galaxies are in the first Gaussian and the remaining 22.4% are in the second Gaussian. The Gaussian’s are centred at $(z) = 0.595$ and $(z) = 0.558$ with standard deviations of $\sigma_z = 0.236$ and $\sigma_z = 0.112$ respectively. Since we are simulating a spectroscopic redshift galaxy sample, we assume all redshifts have been measured correctly to the precision required for correct binning into our Cartesian grid. Imposing the redshift bin limits for our simulated survey of $z_{\text{min}} = 0.2$ and $z_{\text{max}} = 0.58$ provides the redshift distribution. We then finally stipulate that $2 \times 10^6$ galaxies will be detected in the optical survey. The over-density field for the optical galaxies is given as

$$\delta_g(z) = \frac{n_g(z) - \langle n_g \rangle}{\langle n_g \rangle} = b_g(z) \delta_M(z),$$

where $b_g$ is the linear bias for the optical galaxy field. For these simulations galaxy maps and the simulated H$\text{\textsc{i}}$ intensity maps (presented in Section 3.1) we checked that both measured power spectra, and their cross-correlation, are modelled well by commonly used anisotropic redshift space clustering models (see e.g. Soares et al. (2020)), thus validating their use as our underlying cosmological data.

4 METHODS FOR FOREGROUND CLEANING

Here we discuss some of the most popular and well studied approaches to 21cm foreground cleaning. Our focus in this work is on single-dish observations, in the context of cosmological analysis and we are therefore ultimately trying to optimize a power spectrum measurement. All foreground removal methods aim to utilize the fact that the foreground contributions are slowly varying with frequency (unlike cosmological H$\text{\textsc{i}}$) and are orders of magnitude more dominant than the H$\text{\textsc{i}}$. Thus, the general approach is identifying a set of smooth functions that represent the dominant foreground contributions and subtracting these from the data to leave the cosmic H$\text{\textsc{i}}$ signal. The method for estimating this set of smooth functions is largely where the techniques diverge into the wide library of foreground removal options available today (see e.g. Liu & Shaw (2020) for a more detailed summary).

Blind component separation methods dominate the literature concerning foreground removal techniques, and we also use them in our analysis. Blind separation means little input information is needed and the process exploits the fact that relatively few dominant uncorrelated (or statistically independent) (or sparse) sources should contain the majority of the foreground emission in the observed signal. The advantage of such an approach is that it does not require a detailed understanding of the foreground signals, e.g. their precise amplitude through frequency, and how they respond to instrumental systematics. Given that we are a long way from fully understanding sky emission at the ~21cm wavelengths and that the intensity mapping technique is still in its infancy (meaning instrumental response and systematics are poorly understood), it is sensible for blind meth-
ods to be the preferred choice (Masui et al. 2013; Wolz et al. 2017; Anderson et al. 2018).

The raw observed sky signal in intensity mapping\(^5\) can be decomposed into contributions from the cosmological \(H_\text{i}\), the foregrounds, and the thermal noise from the instrument (as in Equation 1). These observed data can be represented by a matrix \(X_{\text{obs}}\) with dimensions \(N_r \times N_\nu\) where \(N_r\) is the number of frequency channels along the line-of-sight and \(N_\nu\) the number of pixels. In this approach the 2D \((N_\nu^a \times N_\delta^a)\) angular pixel space is turned into a \(N_\theta = N_\nu^a \times N_\delta^a\) long 1D vector to make the foreground cleaning formalism more concise. This is also typically the way curved-sky data is handled in HEALPix maps (Zonca et al. 2019).

We make the assumption that the data matrix \(X_{\text{obs}}\) can be represented as a linear system

\[
X_{\text{obs}} = \hat{A} S + R, \tag{14}
\]

where \(\hat{A}\) represents the estimated set of \(N_{\text{FG}}\) smooth functions (often referred to as the mixing matrix) with shape \([N_\nu, N_{\text{FG}}]\) that evolve the \(N_{\text{FG}}\) separable source maps \(S\) through frequency. Generally the sources can be identified by projecting the mixing matrix along the observed data\(^6\)

\[
S = (A^T \hat{A})^{-1} A^T X_{\text{obs}}. \tag{15}
\]

\(N_{\text{FG}}\) is the pre-selected number of separable sources which we expect our foreground emission to be contained within. The remaining signal from subtracting the smooth functions and sources from the data is in the residual term \(R\) and this is used for the cleaned intensity map data:

\[
X_{\text{clean}} = R = X_{\text{obs}} - \hat{A} S. \tag{16}
\]

This will contain cosmological \(H_\text{i}\), noise and typically some residual foreground emission. The resulting cleaned intensity maps can be summarized as

\[
\delta T_{\text{clean}} = \delta T_{\text{obs}} - \hat{A} S = \delta T_{\text{obs}} - \sum_{n=1}^{N_{\text{FG}}} \hat{A}_n (\nu) S_n (\theta). \tag{17}
\]

As we will see in this investigation, the optimal choice of \(N_{\text{FG}}\) can vary considerably. Generally speaking, an \(N_{\text{FG}}\) that is too low will result in too much foreground signal remaining in the residual component, and an \(N_{\text{FG}}\) that is too high will result in too much cosmological \(H_\text{i}\) leakage into the subtracted component causing a loss of true signal. Finding an optimal balance is the aim of a successful foreground clean, and a key focus in our investigation.

There are many existing methods for estimating \(\hat{A}\) for a given choice of \(N_{\text{FG}}\), and we explore some in the remainder of this section. In this section we aim to introduce some of the most popular blind source separation techniques, and highlight their similarities. We will use an SVD-based technique (or equivalently PCA – an equivalence we will explain) as our default foreground cleaning method which we introduce next. We then explore some related techniques with extended sophistication and test them on our simulated data.

We emphasize that the methods we outline are in no way an exhaustive list, and many more methods exist for foreground removal that could be applicable to 21cm intensity mapping e.g. GNILC (Oliveri et al. 2016), SMICA (Delabrouille et al. 2003), RPCA (Zuo et al. 2016) etc. (see the list in the Appendix of Leach et al. 2008) for more information). One further notable approach is Gaussian Process Regression (GPR) (Mertens et al. 2018), which has been recently used on real data but for a higher redshift, epoch or reionization survey (Mertens et al. 2020). Investigating this method with low-redshift 21cm intensity mapping data is very interesting and will be the focus of future work.

### 4.1 PCA (\& SVD)

Principal Component Analysis (PCA) is a widely used technique in statistics, closely related to Singular Value Decomposition (SVD). It provides a hierarchical coordinate system to represent high-dimensional correlated data by transforming it to a dimensional basis that maximises the variance. These new basis vectors are the principal components. In the context of correlated foreground emission in 21cm data, due to their large amplitude and highly correlated frequency structure, it is likely that the foreground signals can be reconstructed from just a few of these principal components. Hence, the first few \(N_{\text{FG}}\) dominant basis vectors found in this process represent the estimate for the set of smooth functions in Equation 14, which can then be removed from the observational data.

The steps for performing PCA to construct an estimate of the foreground contamination \(\hat{X}_{\text{FG}}\), which is then removed from the data, can be concisely outlined as follows:

1. The data is mean-centred, i.e. the mean at each frequency is subtracted from the data for each frequency channel.

2. The covariance matrix of the mean-centred data is calculated:

\[
C = X_{\text{obs}}^T X_{\text{obs}} / (N_\theta - 1). \tag{18}
\]

3. The eigen-decomposition of the covariance matrix is computed:

\[
CV = AV, \text{ where } A \text{ is the diagonal matrix of eigenvalues ordered by descending magnitude.} \tag{19}
\]

4. The first \(N_{\text{FG}}\) columns from the eigenvector matrix \(V\) represent the set of smooth functions to construct the mixing matrix, i.e. \(\hat{A} = BV\) where \(B\) is a selection matrix with 1 along the diagonal for modes to be removed and 0 elsewhere.

5. The projection of the selected eigenvectors along the mean-centred data provides the eigen-sources, \(S = A^T X_{\text{obs}}\), which are combined with the mixing matrix to provide the reconstructed foreground estimation \(\hat{X}_{\text{FG}} = AS\).

#### 4.1.1 Singular Value Decomposition

The singular value decomposition (SVD) is a unique matrix decomposition of the data (note that PCA and SVD are inherently related). The SVD of the observed data \(X\) is given by\(^7\)

\[
X = U \Sigma V^T, \tag{18}
\]

where \(U\) and \(V\) are unitary matrices with orthonormal columns and \(\Sigma\) is a diagonal matrix whose entries represent the singular values. It can be demonstrated how closely related the SVD is to an eigenvalue decomposition. By considering Equation 18, and given that \(X^T = V \Sigma U^T\), the covariance can be written as

\[
C = \frac{X^T X}{N_\theta - 1} = \frac{V \Sigma U^T U \Sigma V^T}{N_\theta - 1}. \tag{19}
\]

\(^7\) The more general form for SVD is \(X = U \Sigma V^*\), however in the context of 21cm data we are always dealing with real-valued matrices where \(V^* \equiv V\).
A fundamental property of SVD stipulates that $U^TU$ is a unitary matrix ($U^TU = I$). With a little rearranging we get

$$\mathbf{CV} = \frac{\mathbf{V}\Sigma^2}{N_\theta - 1}, \quad (20)$$

which we can recognise as an eigenvalue decomposition of the correlation matrix $\mathbf{C}$ (as shown in step 3, in Section 4.1) where the columns of $\mathbf{V}$ are the eigenvectors and the singular values $\Sigma$ are proportional to the positive square roots of the eigenvalues.

In real intensity mapping data, in order to mitigate the high-levels of thermal noise present in Pathfinder experiments, it is necessary to cross-correlate data from different observation runs e.g. $X_A \times X_B$. In this situation, the covariance matrix $X_A^T X_B$ is no longer symmetric and an SVD is required where the left and right singular vectors in $\mathbf{U}$ and $\mathbf{V}$ are used to reconstruct foreground estimates in each run (Switzer et al. 2013). In this work we do not explore such a situation and therefore the SVD and PCA can be seen as equivalent treatments.

4.1.2 Polynomial Fitting

A related, and essentially equivalent, method to PCA is polynomial fitting (Ansari et al. 2012). Although similarities exist, this is not to be confused with parametric fitting (see Section 4.4) and conventionally refers to a blind approach to foreground cleaning. The approach works by identifying a set of smooth fitted functions $f_k$ where polynomials are used as basis functions e.g.

$$f_k(\log(\nu)) = [\log(\nu)]^{k-1}. \quad (21)$$

Then, by least-squares fitting these functions to each line of sight, the foreground contribution can approximated. Since previous work has already demonstrated the theoretical equivalence this has with PCA (e.g. Alonso et al. 2015 that also provides simulation tests) we do not include this in our investigation.

4.1.3 Truncation Choice

Deciding where to truncate to, i.e. the number of $N_{FG}$ principal components to include in the foreground estimate (and hence remove) is key to an optimized blind foreground clean. By analysing the eigenvalues in $\mathbf{A}$ (or, equivalently, the singular values from the SVD), that estimate the amount of variance in the data captured in the corresponding principal components, an informed choice can be made. As discussed above, due to the nature of the foreground emission, most of the information is contained in a small sub-set of principal components where often $N_{FG} \sim 3 \rightarrow 20$ (depending on the foreground emission and instrument response) can produce a reasonable reconstruction. We can quantitatively analyse this choice with

$$R = \frac{\sum_{i=1}^{N_{FG}} \lambda_i}{\sum_{i=1}^{N_{FG}} \lambda_i}, \quad (22)$$

where $\lambda_i$ are the eigenvalues in $\mathbf{A}$, descending in magnitude. Since a higher number for $N_{FG}$ will remove more H$^i$ information, the aim for an optimal choice is to maximise $R \rightarrow 1$ for a minimal value for $N_{FG}$.

We show some values for $R$ (Equation 22) in Figure 2 for the different sky regions in our simulated data, both with and without polarization leakage (solid and hollow markers); we plot $1 - R$ to demonstrate the convergence, this means the closer to zero the particular combination of eigenvalues is, the better a representation of the full data that reconstruction will be since it is capturing more of the full data’s variance, and its reconstruction will capture more of the foreground contamination which can be removed. Figure 2 immediately shows how highly correlated the observed data is given that just one eigenvalue in all cases has $R \sim 1$ meaning nearly 100% of the signal can be represented with just 1 principal component ($N_{FG} = 1$). However, just a small amount of residual foreground, even at the sub-percent level, is enough to entirely dominate the H$^i$ signal. Therefore in all cases $N_{FG} = 1$ is not sufficient for a foreground clean. This plot gives an indication of how far one needs to go in the reconstruction. Almost all cases eventually reach a plateau where including more eigenvalues barely contributes to the reconstructed signal and it is here where PCA has likely reached its efficiency limit and will not be able to remove much more foreground.

In contrast, the H$^i$ information cannot be compressed into a small number of principal components due to its Gaussian-like nature. This is the main principle behind the blind source separation approach. The highly correlated information containing the majority of foregrounds can be removed using a few $N_{FG}$ modes, leaving the bulk of the H$^i$ information that is mostly evenly distributed among the remaining components. However, it does mean a fine balance needs to be attained in a successful foreground clean. Being too aggressive and choosing too high values for $N_{FG}$ will begin to remove H$^i$ information, typically large-scale line-of-sight modes.

In a simulation-based procedure, we can effectively analyse this problem since we have access to the separated pure-H$^i$ and pure-foreground simulated data. We can therefore calculate the contributions from these components remaining in the residuals after a foreground clean. The separated residuals are calculated using the

---

8 We also include the contribution from thermal noise in this calculation.
estimated mixing matrix $\hat{A}$, and projecting the pure-Hi (or pure-foreground) simulated data along this:

$$x_{\text{resid Hi}} = x_{\text{Hi}} - \hat{A}(\hat{A}^T\hat{A})^{-1}\hat{A}^T x_{\text{Hi}},$$  \hfill (23)

$$x_{\text{resid FG}} = x_{\text{FG}} - \hat{A}(\hat{A}^T\hat{A})^{-1}\hat{A}^T x_{\text{FG}}.$$  \hfill (24)

Note the $(\hat{A}^T\hat{A})^{-1}$ factor is not needed for the PCA method since, by construction, the mixing matrix is orthogonal and this quantity will equal the identity matrix. However, the vectors in the mixing matrix for the FASTICAand GMCA approach we use are not orthogonal, thus this quantity is needed to obtain the correct projection.

We utilise these separated residual calculations extensively in our results to analyse the performance from various cleaning methods under different situations. We also use this concept for the bottom-panel of Figure 2 where we demonstrate an estimation for the amount of eigen-information lost along the line-of-sight for each choice of $N_{\text{FG}}$. This calculated by computing the eigendecomposition of the residual Hi (Equation 23) and summing the eigenvalues. Dividing this by the sum of all the eigenvalues in the original Hi data gives a proxy for the amount of eigen-information remaining after subtracting $N_{\text{FG}}$ principal components, i.e. the portion of variance that is removed. This should further illustrate the challenge of foreground cleaning, a balance between removing foreground while trying to leave the Hi signal intact – however, we always expect some signal loss. Since foreground dominated data is typically decomposed into dominant eigenmodes containing highly correlated information along the line-of-sight subtracting $N_{\text{FG}}$ principal components generally removes large-scale modes along the line-of-sight in the Hi power spectrum, i.e. small $k_\parallel$ modes. Modelling this is non-trivial and a major challenge for precision radio cosmology.

4.2 FASTICA

Fast Independent Component Analysis (FASTICA) is another widely used method for foreground cleaning and has been tested on simulated data (Chapman et al. 2012; Wolz et al. 2014; Cunnington et al. 2019) and also on real data (Wolz et al. 2017). When we discuss FASTICA we are referring to the method developed in Hyvärinen (1999) and we use the package in Scikit-learn\(^9\) (Pedregosa et al. 2011).

While PCA is generalized for reducing dimensionality in data, FASTICA (and more generally independent component analysis) is more specifically used to separate mixed signals, and is therefore naturally suited to a blind source separation problem. FASTICA forms an estimate for the mixing matrix $\hat{A}$ by assuming the sources are statistically independent of each other. The method therefore aims to maximize statistical independence that can be assessed using the central limit theorem, which states that the greater the number of independent variables in a distribution, the more Gaussian that distribution will be (that is, the probability density function of several independent variables is always more Gaussian than that of a single variable). Hence, by maximising any statistical quantity that measures non-Gaussianity, we can identify statistical independence.

Before assessing non-Gaussianity, FASTICA begins by mean-centering the data then carries out a whitening step that aims to achieve a covariance matrix equal to the identity matrix for this whitened data (i.e. the components will be uncorrelated and their variances normalized to unity). Since this whitening step can be achieved with a PCA analysis, FASTICA is essentially an extension of PCA, and hence in most cases in the context of foreground cleaning, will provide very similar results.

For maximising non-Gaussianity, an approximation of the negentropy can be used\(^10\). We refer the reader to Hyvärinen & Oja (2000) for further detail on this aspect of the algorithm. In the context of 21cm foreground cleaning, the approximation of negentropy uses a set of optimally chosen non-quadratic functions which are applied to the data and averaged over for all available pixels. The maximization of negentropy by averaging over angular pixels means that for purely Gaussian sources, FASTICA will be unable to improve upon the initial PCA step carried out in the whitening step. This is because the Gaussian sources will have an equivalent zero negentropy. This explains the similarity in results often found between PCA and FASTICA when most of the simulated components are Gaussian fields (Alonso et al. 2015). It is in situations over very large skies, where the negentropy approximation will be more optimal and sufficient non-Gaussian structure exists in the foreground maps, where FASTICA will perhaps make discernible differences to the PCA-only performance.

To summarise, the components found using PCA are uncorrelated linear combinations of the data, which are identified by maximising the variance. FASTICA extends on this by finding components that are also uncorrelated linear combinations of the data but identified by maximising statistical independence, through estimates of non-Gaussianity in angular pixels.

4.3 GMCA

GMCA stands for Generalised Morphological Component Analysis (Bobin et al. 2007), it is a blind source separation algorithm exploiting the idea that the different components contributing to the signal are morphologically different. To enhance the morphological differences, the signal is projected into an adapted domain where we expect the components to have a sparse representation, i.e. to be described by few non-zero coefficients. When we find such a domain, the contrast between components increases, easing the separation process. Here, we make use of wavelets, which has recently been shown to be optimal for this context (Carucci et al. 2020a). GMCA has already been optimised and used with astrophysical data sets (e.g. Cosmic Microwave Background data (Bobin et al. 2013, 2014), high-redshift 21cm interferometric data (Patil et al. 2017), X-ray images of Supernova remnants (Picqueton et al. 2019)).

In practice, once the data $x_{\text{obs}}$ has been wavelet-transformed to $X^{\text{wt}}$, GMCA promotes sparsity in the requested $N_{\text{FG}}$ sources $S^{\text{wt}}$ by solving iteratively the minimization problem given by

$$\{\hat{A}, \hat{S}\} = \min_{\hat{A}, S^{\text{wt}}} \sum_{i=1}^{N_{\text{FG}}} \lambda_i \|S_i^{\text{wt}}\|_1 + \|X^{\text{wt}} - AS^{\text{wt}}\|_F^2,$$  \hfill (25)

where the first term is the $\ell_1$ norm, i.e. $\sum_{j,k} |S_{j,k}^{\text{wt}}|$: this constitutes a constraint for sparsity, mediated by the regularization coefficients $\lambda_i$. The latter act as sparsity-thresholds that in our case should be tuned by the difference in intensity between the foregrounds and the cosmological signal; we first estimate them with the median absolute deviation (MAD) method and progressively decrease towards a final noise-related level. The second term in Equation 25 is the standard Frobenius norm, that assures data-fidelity step by step.

Once the mixing matrix $\hat{A}$ has been estimated, we project the initial data $x_{\text{obs}}$ in pixel-space (following Equation 15 and Equation 16) to

\(^9\)https://scikit-learn.org/

\(^{10}\)Kurtosis can also be used as a measure of non-Gaussianity.
4.4 Non-Blind Parametric Fitting

In non-blind methods an estimator $\hat{A}$ for the mixing matrix is constructed using astrophysical, as opposed to statistical, knowledge about the foreground sources and has been previously explored (Ansari et al. 2012; Bigot-Sazy et al. 2015). A single frequency channel of a 21 cm intensity mapping experiment will consist of diffuse synchrotron emission, diffuse free-free emission, extragalactic point sources, the H\textsc{i} signal, instrumental noise and any other instrumental contributions (e.g. polarization leakage). Synchrotron and free-free emission are believed to be spectrally smooth with well-understood spectral forms that can both be expressed as power laws. Whilst the synchrotron spectral index is known to change across pixels, diffuse synchrotron emission is the signal identified with the largest signal-to-noise ratio within 21 cm intensity mapping experiments giving it the largest probability of an accurate characterisation.

As of such we propose a parametric fit which aims to parameterize the free-free and synchrotron foreground contributions explicitly. Diffuse synchrotron, diffuse free-emission and extragalactic point sources are strongly degenerate; free-free and synchrotron emission maps contain identical spatial features and all three spectra can be represented as power laws with similar spectral indices. Whilst we do not aim to explicitly fit for the extragalactic point sources we expect their contributions to be subsumed within the synchrotron and free-free emissions fits. We are essentially making the opposite assumption to ICA by relying upon the parameter degeneracy, if this assumption is correct then the residuals between our parametric assumption and ICA by relying upon the parameter degeneracy, if and free-free emissions fits. We are essentially making the opposite expectation their contributions to be subsumed within the synchrotron do not aim to explicitly fit for the extragalactic point sources we

The GMCA algorithm is a fully blind component separation method, which cannot take into account physical knowledge about the spectral forms of the sought-after components. For that purpose, Bobin et al. (2019) recently introduced a novel semi-supervised sparse component separation method that can constrain the spectral shape of certain components, based on few, say $L$, examples $\{ai\}_{i=1,...,L}$. The problem of source separation is updated so as to account for this additional constraint, as detailed in the following optimisation problem:

$$\{\hat{A}, \hat{S}\} = \arg \min_{A, S} \sum_{i=1}^{n_p} B_i \left( A_i \right)^{\alpha^i} + \sum_{i=1}^{n_p} ||a_i||_{\ell_1} + ||X - AS||_2^2,$$

(26)

where $n_p$ stands for the number of sources, whose spectra have a constrained shape. Following (Bobin et al. 2019), the additional constraint $B_i \left( A_i \right)^{\alpha^i}$, enforces the spectrum $A_i$ to belong to the so-called barycentric span of the examples $\{A_i\}_{i=1,...,L}$ and according to some distance. More precisely, the estimated spectrum $A_i$ will be computed as a particular weighted mean. The values of the corresponding weights are optimized during the separation procedure. This leads to the so-called semi-blind GMCA algorithm.

For this application, only the spectral shape of the synchrotron emission will be constrained. As described in Bobin et al. (2019), the distance used to compute the resulting weighted means should be chosen based on the shape of the spectra. For synchrotron emission, the Kullback-Leibler divergence is perfectly well-suited to power laws. The required examples will be composed of two spectral shapes that are obtained from the maximal and minimal expected values for the spectral index of synchrotron.

Accurately modelling synchrotron emission requires accommodating spectral forms which vary from pixel to pixel. Semi-blind GMCA works across groups of pixels, as opposed to at the individual pixel level. Therefore semi-blind GMCA is run on super-pixels of size 4 pixels by 4 pixels. This small area size is chosen to ensure monopoles must be removed from each map; for our particular simulations that means the unresolved extragalactic point source levels at each frequency. The spectral index for a particular emission is strongly tied to the monopole level of the maps and so the parametric fit we perform is tied to the zero-level of the observational data.
that the synchrotron spectral index remains constant within a super-pixel. Once in this patch example spectra need to be constructed to inform the algorithm how many emission mechanisms are present and what are their average, maximum and minimum spectral forms. The example synchrotron spectra used for this analysis are given in Table 2; note that the free-free spectral form has been fixed to its true value, and is not updated during the optimisation procedure. All estimates within this work are computed within the pixel domain; transforming the total emission into the wavelet domain and running semi-blind GMCA on the wavelet coefficients provided no further gain in accuracy for this particular method and data set.

Semi-blind GMCA is run on each super-pixel and provides one value for $\beta_{sy}$. Next, a least-squares fitting routine such as LmFit (python) can be used to determine the model parameters per full resolution pixel. Fitting each individual pixel is necessary as otherwise the synchrotron spectral index map produced will only have the resolution of the super-pixels. So we essentially perform two fits, one to super-pixels to get rough synchrotron spectral index estimates and then one to each full resolution pixel. The parameter space for $\beta_{sy}$ is limited to within $\pm1\%$ of the trial estimate calculated by semi-blind GMCA.

4.4.2 Least-squares fitting

Despite the fact that the total temperature across each pixel always features a combination of free-free and synchrotron emission, we find that fitting both these diffuse Galactic emissions is not optimal across the full sky. In low free-free signal-to-noise regions such as the SCP and Stripe 82 it is, in fact, better to assume free-free emission to be negligible and to only fit for synchrotron emission.

Within the synchrotron dominant regions of our simulation data, semi-blind GMCA is not required and a model of only synchrotron emission is fit to the total temperature per pixel. The parameter space for $\beta_{sy}$ is limited to within $\pm10\%$ of the total data spectral index. The parameter space for the synchrotron emission amplitude is limited to within $\pm50\%$ of the total temperature.

Having used either semi-blind GMCA or a least-squares fit to acquire the per-pixel synchrotron spectral index values we now have a complete mixing matrix which we use to calculate the diffuse Galactic emission amplitudes from the total temperature maps using Equation 15. We can subtract our free-free emission and synchrotron emission estimates from the total temperature data to leave maps of HI plus instrumental contributions.

4.5 Quantifying Foreground Removal Effects

Despite the range of different foreground cleaning methods available, none are perfect and will inevitably remove some cosmological HI signal or leave behind foreground residuals. We discuss some methods for investigating this both with simulations and real data.

4.5.1 Damping Cosmological HI

This usually occurs on large scales where the HI is most degenerate with the foregrounds. For idealized future surveys assuming excellent instrumental calibration, residual foregrounds should be well controlled and not exacerbated from effects such as polarization leakage. In these cases the effects from a low-$N_{HI}$ foreground clean are relatively straightforward and can be potentially modelled as some damping to the power spectrum (see e.g. Cunnington et al. (2020a,b); Soares et al. (2020)).

However, applying this to real data requires a high level of confidence in the modelling that builds upon a detailed understanding of the nature of foregrounds as well as systematic/instrumental effects, something we do not currently have. An alternative approach is to add the observed data itself to simulations (mocks), then apply a foreground clean and access the response the mock data had to this process. Signal loss can be quantified this way with a foreground transfer function, which is applied to the real data to compensate for these effects. This has been the approach of several of the HI intensity mapping detections so far ((Masui et al. 2013; Switzer et al. 2013; Anderson et al. 2018)).

Following Switzer et al. (2015) the transfer function can be constructed by adding mock data $M$ to the true observed data $X_{obs}$, which includes foregrounds. This can then be cleaned to provide $M_{clean}$, an estimate for the effects of removing the foregrounds on the mock map:

$$M_{clean} = [M + X_{obs}]_{PCA} - [X_{obs}]_{PCA}.\quad (27)$$

where the $[\cdot]_{PCA}$ notation represents performing a PCA clean, but in principle this could be done with any foreground cleaning method. Note that in Equation 27 the cleaned data $[X_{obs}]_{PCA}$ has been subtracted. This is necessary to reduce the variance in this estimation since the unwanted data-HI component will serve as additional, unwanted noise. The transfer function is then given by:

$$T(k) = \frac{\mathcal{P}(M_{clean} \cdot M)}{\mathcal{P}(M\cdot M)},\quad (28)$$

where $\mathcal{P}(\cdot)$ denotes an operator which measures the power spectrum in $(k_x, k_y)$ space. The angled brackets denote an averaging over a large number of mocks. The power spectrum is then corrected for by dividing through by this transfer function. This can also be utilized in a cross-correlation measurement with the only difference being that the power of 2 is dropped from Equation 28 because the effects of cleaning are only applied to the HI data. We employ the transfer function later in our analysis (Section 5.5) by constructing 100 lognormal mocks and using our MultiDark simulations as the “observed” data.

The foreground transfer function is thus used as a data-driven way of compensating for signal loss in the foreground removal. However since $T(k) \leq 1$, it will not be capable of addressing the issue caused by additive biases from foreground residuals, discussed in the following section.

4.5.2 Foreground Residuals

Whilst signal loss from over-cleaning can be modelled or compensated for with a foreground transfer function, foreground residuals produced from under-cleaning, which cause additive biases and boost errors, are more challenging to address. For near-future, pathfinder surveys (e.g. MeerKAT (Santos et al. 2017)) it is possible that the instrument response will not be sufficiently understood and polarization leakage effects could manifest, causing contamination from foreground residuals. Developing robust statistics which estimate the effects caused by these residuals will therefore be essential for future surveys (Switzer et al. 2015). There is not a large amount of research on this issue, since current data analysis usually has large thermal
noise and unknown systematic effects (Switzer et al. 2013). Alternatively, detections have been made using cross-correlations with optical galaxy data (e.g. Masui et al. (2013); Anderson et al. (2018)). In cross-correlation the residual foregrounds and survey-specific systems do not correlate with the optical galaxy data and instead, simply boost errors (we will study this in detail in Section 5.5). As the intensity mapping technique matures and calibration and signal-to-noise capabilities of surveys improve, we will aim to conduct precision cosmology using auto-correlation measurements. Therefore, we need to develop a pipeline for quantifying the foreground residual contamination.

As discussed in Section 4.1.3, the residuals can be exactly calculated (see Equation 23 and Equation 24) because we are using simulations where the original decomposed HI and foreground contributions are known. A direct comparison between $X_{\text{residHi}}$ and $X_{\text{residFG}}$ is then extremely useful (and a topic we investigate) where one ideally desires a situation where $X_{\text{residHi}}$ dominates $X_{\text{residFG}}$. A more dominant $X_{\text{residFG}}$ would increase additive biases due to the residuals correlating with each other. However, in real data, distinguishing the contribution between foreground residual and HI signal will be challenging. If one can develop a robust way of estimating the contribution from foreground residuals, then this can be effectively modelled in a similar way to the instrumental noise which causes additive power in auto-correlations along with boosting errors. So the HI auto-power spectrum could be expressed as

$$P_{\text{HI}}(k) = (T_{\text{HI}})^2 b_{\text{HI}}^2 P_m(k) + P_N(k) + P_{\text{residFG}}(k),$$

where $P_m$ is the matter power spectrum and $P_N$ and $P_{\text{residFG}}$ are the contributions from thermal noise and residual foregrounds. An estimation for the errors can then be analytically made with

$$\sigma_P(k) \sim \frac{P_{\text{HI}}(k) + P_N(k) + P_{\text{residFG}}(k)}{\sqrt{N_{\text{modes}}(k)}},$$

where $N_{\text{modes}}$ is the number of unique modes in each $k$-bin, included to account for cosmic variance.

### 5 RESULTS

Here we present our results from tests carried out on the simulated data sets and foreground removal methods outlined in the previous sections. For all our blind methods we use the values outlined in Table 3 for $N_{\text{FG}}$, unless otherwise stated.

To diagnose the performance of our foreground cleans we look at measurements of power spectra, both 1D $P(k)$ and 2D $P(k_\perp, k_\parallel)$, and compare these to equivalent foreground-free results where only the cosmological HI is being measured. The offset between the two then serves as a good indicator for how well the chosen method is performing. Following previous studies (e.g. Alonso et al. (2015); Carucci et al. (2020a)), we define below the weighted difference between subtracted foreground and no foreground cases as a metric to help assess the success of the foreground removal under all the different scenarios:

$$\varepsilon(k) = \frac{P_{\text{SubFG}}(k) - P_{\text{NoFG}}(k)}{P_{\text{NoFG}}(k)}.$$

Here $P_{\text{SubFG}}(k)$ is the measured power spectrum for the simulated intensity maps with foregrounds included and then cleaned, while $P_{\text{NoFG}}(k)$ is the measured power spectrum of the HI-only (foreground-free) intensity maps. We also analyse the 2D power spectrum and use an identical analysis in this basis where

$$\varepsilon_{2D}(k_\perp, k_\parallel) = \frac{P_{\text{SubFG}}(k_\perp, k_\parallel) - P_{\text{NoFG}}(k_\perp, k_\parallel)}{P_{\text{NoFG}}(k_\perp, k_\parallel)}.$$

We begin by plotting the auto-power spectra for each region in Figure 4 and perform a comparison between the foreground-free HI auto-power spectrum (black dashed line) and the foreground-cleaned results using the PCA method. This demonstrates some differences between the regions and also some clear differences for the cases with polarization leakage (red squares) and without (blue circles).

As expected most of the damping to the power comes from large scales (small-$k$). Because we use $N_{\text{FG}} = 3$ in all cases with no polarization leakage, the damping across all regions is approximately equal. However, the foreground residuals would still differ in each region, with the most likely case being that residuals will be highest in the galactic plane where foregrounds dominate the most. Figure 4 finally shows that results can be extensively worse when including polarization leakage effects, except perhaps the SCP, the region least affected by polarization. Results are much worse in the other three regions, for example the largest few modes in the Galactic Plane region are effectively damped to zero.

We show similar results in Figure 5 but here more information can be extracted on the nature of this contamination. This shows the weighted difference $\varepsilon_{2D}$ between the HI-only 2D power spectrum and the foreground cleaned one. Again we plot each region both without polarization leakage (top row) and with (bottom row). This gives an illustration into the effects of foreground under-cleaning and over-cleaning and the delicate balance between the two. From Equation 32 we can see that the blue regions are indicating modes that have higher power in the foreground-cleaned maps compared to the foreground-free ones. Thus blue areas indicate under-cleaned $k$-space. Conversely, red areas have lower power in the foreground-cleaned maps, indicating over-cleaning.

**Figure 5** therefore shows the effects of over-cleaning tend to manifest in low-$k$ modes, as expected. This is particularly evident when going to high $N_{\text{FG}}$ as in the polarized Galactic Plane case for example, where we see significant damping to low $k_\parallel$ modes, again as expected. This is because, in order to control contamination from polarization leakage, we are removing more principal components, each with different oscillating modes due to the instrumental response, but all will still have largely frequency correlated spectra. This inevitably removes the modes in HI which are also highly correlated in frequency i.e. low-$k_\parallel$ modes. This conclusion is quite general and not just specific to a PCA-based method. Generally, any method that utilizes the highly correlated nature of foreground signals will struggle to disentangle foregrounds and large HI modes parallel to the line-of-sight.

For the unpolarized cases, where a lower $N_{\text{FG}}$ is used, it is interesting to see that small-$k_\perp$ modes are damped to a similar level as the

### Table 3. Summary of the default values we use for $N_{\text{FG}}$ in our blind foreground removal methods for the different simulated data sets. These decisions are based on achieving a good balance between over-cleaning (which damps the HI signal power) and under-cleaning (which leaves foreground residuals in the map).

| Sky Region          | No Polarization | With Polarization |
|---------------------|-----------------|-------------------|
| Stripe 82           | 3               | 9                 |
| Galactic Plane      | 3               | 16                |
| SCP                 | 3               | 4                 |
| Barnard’s Loop      | 3               | 11                |
Figure 4. Measured power spectra for each simulated data set. For all four regions we show the foreground free case (black-dashed), cleaned foregrounds without polarization leakage (blue-circle markers) and cleaned foregrounds with polarization leakage (red-square markers). For the no polarization leakage cases, we fix the number of removed principal components to $N_{FG} = 3$ but for the polarization leakage, a more varied selection is needed (displayed in top-right of each panel). In all foreground cleaned cases, PCA is used.

Figure 5. The impact of a PCA foreground clean on the 2D power spectra. Plotted is the weighted difference $\epsilon_{2D}(k_\perp, k_\parallel)$ between the foreground free and cleaned 2D power spectra (outlined in Equation 32). Each panel represents a different sky-region (top-row without polarization leakage and bottom-row is with polarization leakage). This can be loosely interpreted as blue positive-pixels representing under-cleaned modes (modes with dominant foreground residual), and red negative-pixels representing over-cleaned modes. The values for $N_{FG}$ are the same as those used in Figure 4 and outlined in Table 3.

small-$k_\parallel$ modes. This will be because the foregrounds generally have quite large angular structures (see maps in Figure 1), therefore when removing the dominant principal components representing the foregrounds, any small-$k_\parallel$ modes will be degenerate with these and are therefore damped. This agrees with results from Soares et al. (2020) where a damping to $k_\perp$ modes was required to model the foreground contamination. For the polarization leakage cases in the Stripe 82 and Barnard’s Loop regions, we found that we could achieve decent results with lower $N_{FG}$. However, there was a small region in 2D $k$-space particularly affected by foreground residuals. This high-$k_\perp$ and low-$k_\parallel$ under-cleaned (blue) area is still evident in both regions of Figure 5 but for lower $N_{FG}$ this was much more dominant and we therefore chose a more aggressive clean to achieve a better balance. This is likely due to smaller-angular foreground structures in these maps which, with polarization leakage, create a very localized region in $k$-space that requires a high-$N_{FG}$ clean.

5.1 Comparing Blind Foreground Cleaning Methods

We now compare the cleaning methods we have introduced: PCA, FASTICA and GMCA. All methods rely on the assumption that we can decompose the signal linearly as in Equation 14 and estimate the mixing matrix $\hat{A}$, identifying the subspace of the data set where we expect foregrounds to live, which are then removed as per Equation 16.

In the top panels of Figure 6 we show the mixing matrices derived by the different methods applied to the same data cube. Each method provides a different estimation for $\hat{A}$, yet, the final cleaned
maps are remarkably similar in all cases. We found that there were no discernible differences in the maps and in the bottom panel we demonstrated this. We plot the PCA-cleaned Hi intensity map for one channel (bottom-left panel of Figure 6) and then show the difference between this PCA map and the corresponding FASTICA and GMCA counterparts (middle and right bottom panels). Their differences are orders of magnitude below the amplitudes of the cleaned map and this is true for all channels of all the sky regions explored, with and without the inclusion of polarization leakage. Given the similarity in the cleaned maps, it is unsurprising that at the 2-point statistics level, in all the scenarios considered, the three blind methods output essentially identical power spectra (as shown by the examples in Figure 7).

The difference in the top-panel between PCA and the other methods is an interesting demonstration of the subtle distinction in techniques, PCA is maximising the variance into as few modes as possible. The highest ranked mode represents the one that best fits the variance of the data. The second highest rank mode, will be the next best fit but is required to be orthogonal to the first, hence why PCA identifies two dominant smooth modes in Figure 6, which likely contain the synchrotron and free-free emission. The remaining modes are then more oscillatory and likely identify polarized residuals in a descending order of contribution to the total variance. Applying FASTICA and GMCA algorithms, we are instead identifying a pre-determined \( N_{\text{FG}} \) number of modes within which to maximise statistical independence or sparsity. They achieve this by identifying functions that can share out the contributions to the variance amongst these \( N_{\text{FG}} \) modes, with no requirement of orthogonality, providing they are maximizing independence and sparsity. This is allows the modes in FASTICA and GMCA to approximately follow the slope defined by the dominant spectral indices from synchrotron and free-free emission. These functions will still contain the polarized information demonstrated in the PCA functions, but be contained as sub-dominant oscillations within these modes.

Despite the differences in the identified mixing matrices, the similarity in final results from all three blind methods can be understood by considering the difference in the assumptions they make when linearly-decomposing the signal. PCA merely identify the eigenvectors corresponding to the largest eigenvalues in the frequency-frequency covariance of the data. FASTICA adds on this by promoting non-Gaussianity in the estimated sources, as a proxy for their statistical independence. GMCA promotes sparsity in the spatial domain, after having wavelet-transformed the patches, relying on highlighting the specific morphologies of the components to facilitate the source separation process. Since all three approaches result in essentially identically cleaned maps and power spectra, this leads to the conclusion that the FASTICA and GMCA assumptions do not hold for these data sets: we are dealing with fairly Gaussian (non-sparse) components in the spatial dimensions, thus FASTICA and GMCA are not optimized to improve upon their pre-processing PCA step. However, this statement cannot be generalized and is specific to our tested simulated data i.e the size and resolution of the patches we work with. For instance, Carucci et al. (2020a) show how FASTICA and GMCA behave differently in presence of non-continuous, RFI-flagged data on the full-sky. Also work is still needed to understand a realistic beam effect on the intensity maps, which could add complexity to the foreground removal process.

5.2 Non-Blind Parametric Fitting

The appeal of a parametric method is its ability to yield estimates for both the cosmological signal and each individual foreground. In
our parametric approach we have assumed knowledge of the free-free spectral index, enlisted two methods (a least-squares fitting and semi-blind GMCA) to determine the synchrotron spectral index, and finally solved a least-squares optimization to the data based on the synchrotron and free-free spectral forms. As our approach only considers synchrotron and free-free emission explicitly, all other data contributions get absorbed into either our cosmological, synchrotron or free-free estimates. Specifically, the instrumental noise is included in our H i estimate, point sources are contained in our synchrotron estimate due to their similar spectral forms and polarization leakage is seen to degrade the quality of the H i, free-free and synchrotron estimates.

With the free-free spectral form held at the true value, the key to our approach is accurate determination of the synchrotron spectral indices. Figure 8 shows the absolute percentage difference maps between the true synchrotron spectral indices and those estimated by our non-blind fit. The bottom row of maps shows an increase in percentage error when polarization leakage is present, most notably for Barnard’s Loop and the Galactic Plane. These regions possess the largest rotation measures in our simulations so this is to be expected. For high Galactic latitude regions, like the SCP, polarization leakage is not a prohibitive factor for accurate recovery of the synchrotron spectral index. Aside from polarization leakage, the presence of non-negligible free-free emission also degrades our estimate of the synchrotron spectral indices and so again Barnard’s Loop and the Galactic Plane show larger percent errors than the SCP and the Stripe 82. Nevertheless, the percentage errors for the majority of all four regions remain under 1% when polarization leakage is not present and for the SCP and Stripe 82 when polarization leakage is present.

As polarization leakage is not explicitly accounted for within our parametric fit, nor does it display any degenerate behaviour with the foregrounds we do account for, we simply have to rely on an observing strategy which avoids sky regions with strong rotation measures. In conclusion: our parametric fit cannot compete with the blind methods in the face of polarization leakage. Our leakage-free simulations reveal that for regions of the sky where both free-free and synchrotron emission are present at comparable magnitudes, such as across the Galactic Plane, the combination of their similar magnitudes, similar spectral forms and similar spatial patterns degrades the ability of semi-blind GMCA to identify the synchrotron spectral index.

Figure 7 presents the ‘best’ of the non-blind results, in that the total simulation data for these two regions are synchrotron emission dominated. For high Galactic latitude regions, when no polarization leakage is included, Figure 7 shows that our parametric fitting technique was worth consideration but is not capable of being used "as is" to get to the H i signal level. When polarization leakage is included the non-blind method can only produce comparable results to the blind methods within the SCP; for all other regions the non-blind residuals are too large to occupy the same axis ranges as the blind residuals.

### 5.3 Hybrid Foreground Cleaning

With several available methods for performing 21cm foreground cleaning, an obvious question to ask is whether any of them can be combined into a hybrid method to produce better results. Some meth-
ods already exist to this effect: our semi-blind GMCA approach is essentially a combination of GMCA plus parametric fitting based on astrophysical knowledge. Furthermore, hybridized techniques have been explored in Planck Collaboration et al. (2020) where SMICA was used in a semi-blind way. Semi-blind methods continue to be developed as well (Babin et al. 2019).

The approach we adopt here differs slightly since we investigate combinations of foreground cleaning methods by cross-correlating two maps cleaned using different approaches. The main potential benefits from this approach come from the possibility of the cross-correlation reducing the residual foregrounds.

Due to the inherently similar resulting maps in our three blind methods, it is not surprising that we found no benefit in combining these techniques. However, for cases where the semi-blind approach was providing a good fit to the foregrounds and thus producing a reliable foreground clean, we found that a cross-correlation between a map cleaned with this approach and one using PCA, does produce a reduction in residual foreground correlation, as we demonstrate in Figure 9. This shows the contribution to the power from the pure-foreground residuals (Equation 24) left after a clean using PCA (thick blue line) and semi-blind parametric fit (thin red line). The aim is for the foreground residuals to be as low as possible, ideally far below the original H\(i\) signal (black dashed line), which we include for reference. As shown, by using a hybrid approach (dashed purple line) that cross-correlates the two differently cleaned maps, foreground residuals are reduced. We found no discernible improvement in the final power spectrum measurement of all the components, i.e. this current approach would make no change to the measured \(\sigma\). However, this reduction in residuals is important, especially for future surveys where maximising precision and reducing error-bars is paramount (see discussion in Section 4.5.2). This technique could also yield benefits in more realistic situations containing more systematics. If the different methods respond differently to these systematics, then reductions in residual contamination could be significant. For now we highlight these results as an area for potential further investigation.

5.4 Balancing Foreground Residuals & \(H_i\) Damping

As intensity mapping surveys continue to produce data, the focus will be on how to optimize these surveys for constraining cosmological and astrophysical parameters. Until now we have used a consistent defined choice of \(N_{FG}\) for each region as outlined in Table 3. Here we begin to examine the consequences of varying this parameter and explore whether an optimal choice can be made. We seek a balance between over-cleaning foregrounds by using a high-\(N_{FG}\) that causes damping to \(H_i\) power on large scales, and under-cleaning using a low-\(N_{FG}\) leaving higher residual foregrounds that potentially bias results or boost errors.

In the case of cross-correlations (which we focus on in Section 5.5) this is more straightforward to analyse since residual foregrounds, provided they are not too large, will just boost the errors in the power spectrum measurement. Thus identifying an optimal number of \(N_{FG}\) modes to remove is primarily based on minimising errors. In auto-correlation, the process for optimizing the choice of \(N_{FG}\) is more difficult (see discussion in Section 4.5.2).

Figure 10 shows how the 2D power spectrum evolves with an increasing \(N_{FG}\) by analysing \(E_{2D}\), the weighted difference between foreground-cleaned and foreground-free power spectra (Equation 32). The results are only for the Galactic Plane region and include polarization leakage cleaned using a PCA method. This demonstrates how increasing the aggressiveness of the clean mitigates foreground residuals (blue regions) but at the expense of severely damping small-\(k\) modes (shown by red regions).

In Figure 11 the levels of foreground residuals for each region with polarization leakage are shown after PCA cleans with varying \(N_{FG}\). These are calculated using the methods outlined in Section 4.5.2, which reconstruct the exact maps of the foreground-only signal remaining in the cleaned data (shown by solid lines in Figure 11). For comparison, we also plot the \(H_i\)-only power spectra (dashed lines) calculated in a similar way by projecting the \(H_i\)-only simulated data along the \(N_{FG}\) subtracted eigenvectors to precisely reconstruct the residual-\(H_i\) in the maps after the PCA clean.

Figure 11 shows that the residuals decrease with increasing \(N_{FG}\) as expected, but the \(H_i\) signal is also damped. The aim would be to at least reach a level where the \(H_i\) signal dominates over the residual foregrounds. In regions like the Galactic Plane a high \(N_{FG}\) is required to bring the foreground residuals below the \(H_i\)-only power and this is only achieved for the \(N_{FG} = 18\) and \(N_{FG} = 20\) cases on the largest scales. We can see how results are very different across the regions, for example in the SCP, where a very low-\(N_{FG}\) is required to achieve sub-dominant foreground residuals.

It is interesting to note how much the residual-\(H_i\) differs in each region for the same \(N_{FG}\) (e.g. comparing the Galactic Plane \(H_i\) residual with the SCP one for \(N_{FG} = 16\)) despite it being the same underlying simulated \(H_i\) data. The reason for this is down to how the eigenvectors are constructed in each case and we show the first six in Figure 12. For each region we see little distinction between the first two eigenmodes, which are likely picking out the synchrotron slope and the free-free emission. But in the SCP the modes start to oscillate after this, which suggests that smaller scale cosmological \(H_i\) or noise is leaking into these modes. Whereas in the Galactic Plane, the eigenmodes are relatively smooth (except for the long wavelength oscillations caused by polarization leakage) indicating that more large scale information will be removed. This is what we see in Figure 11 too, if we compare e.g. the \(N_{FG} = 16\) case for the two regions. In the Galactic Plane the small-\(k\) is severely damped relative to the SCP, but comparing the scales around \(k = 0.08\ h/\text{Mpc}\) we see that the Galactic Plane actually has slightly larger power in the residual \(H_i\), owing to the fact that the eigenmodes being removed are smoother and contain less small-scale \(H_i\) power.

Figure 11 demonstrates the potential problem from additive biases caused by residual foregrounds, which will correlate in the auto-power spectrum. Of course, in our simulated scenario we are able to decompose the contribution from residual foregrounds and \(H_i\) to the total power spectrum and make a well-informed choice on the optimal choice of \(N_{FG}\). However, if dealing with real data, residual foregrounds and \(H_i\) would not be easily separable at the required precision. This highlights a central challenge to using \(H_i\) intensity mapping in auto-correlation for data that include foregrounds that cannot be efficiently removed (see discussion in Section 4.5.2).

5.5 Cross-Correlation with Optical Galaxy Surveys

Previous \(H_i\) intensity mapping surveys with the GBT (Masui et al. 2013; Wolz et al. 2017) and Parkes telescopes (Anderson et al. 2018; Li et al. 2020b) have relied on cross-correlations with optical surveys for successful detections of cosmological \(H_i\). This is both due to noise and systematic effects in these pathfinder intensity mapping experiments, but also due to foreground residuals. As we have already stated, intensity mapping simulations are typically idealised and foregrounds can be removed relatively straightforwardly. However, the real data in these early experiments have shown that to achieve low enough foreground residuals for successful detections, more aggre-
Figure 10. Impact of a varying $N_{FG}$ on the efficacy of the foreground clean. We show the weighted difference $\epsilon_{2D}$ in foreground free and cleaned 2D power spectra, defined by Equation 32, where positive (negative) values represent under- (over-)cleaning. This is shown for a range of $N_{FG}$ values in a PCA clean on the Galactic Plane region with polarization leakage. We see too low $N_{FG}$ leaves large foreground residuals and increasing $N_{FG}$ removes these residuals at the cost of damping small $k_{||}$ modes.

Figure 11. Power spectra for the foreground residuals (solid lines) from a PCA clean with varying $N_{FG}$ modes removed from each region, all with polarization leakage effects included. For comparison, we also plot the foreground-free original $\text{H}/i.p.c$-only signal (black-dashed line) along with the residual $\text{H}/i.p.c$-only power spectra (dashed lines) after each PCA clean. Ideally one would require a scenario where the residual $\text{H}/i.p.c$ dominates over the residual foregrounds. Calculations outlined in Equation 24 and Equation 23.

S. Cunnington et al.

Figure 13 shows the improvements that can be made with cross-correlations. Here we cross-correlate the simulated optical data (outlined in Section 3.2) with the $\text{H}$ intensity map data contaminated attempt to mimic real-data experiments, which as discussed often need high levels of foreground cleaning (∼ 10 – 20 $N_{FG}$ modes removed from the data). As we have demonstrated in Figure 4 (third panel) and Figure 10, the $\text{H}$ auto-correlation is highly affected by the presence of such dominant foregrounds. Either dominant foreground residuals remain in the data from choosing $N_{FG}$ which is too low, or the largest scales are completely destroyed from choosing a higher $N_{FG}$.

Cross-correlations with optical galaxy surveys are important because they allow for systematics to be mitigated (they drop out in cross-correlation), and a detection can be more easily achieved. We demonstrate this process with our simulated data sets and in doing so we can investigate the optimal level of foreground cleaning required.

To do this we exclusively use our most dominant foreground region (the Galactic Plane), including polarization leakage. This is an sive foreground cleans are required compared to simulations. This is likely due to calibration issues and effects from polarization leakage or chromatic beams, which can cause some frequency decoherence in the otherwise continuous foreground signals.
with a polarized Galactic Plane foreground and PCA cleaned. The top-panel immediately shows that a foreground clean with much fewer modes removed is sufficient for a cross-correlation measurement that has reasonable agreement with the no-foreground case. This is in spite of the large levels of residual foregrounds that will inevitably be remaining from such a mild clean. This is also shown in the middle panel where the weighted difference $\varepsilon$ (Equation 31) is plotted for each variant of $N_{FG}$. This shows that $N_{FG} = 4$ and $N_{FG} = 6$ deliver a reasonably consistent agreement across all scales with the foreground-free power spectrum. This also shows how more aggressive cleans still damp the power spectrum even in this case of cross-correlation. This is especially noticeable at large scales but even at mid-range scales in the zoomed-in section at $0.17 < k < 0.3 \, h/\text{Mpc}$, we can see power being damped as $N_{FG}$ is increased.

It is perhaps tempting to conclude that an optimal choice of $N_{FG}$ in cross-correlations can be entirely based on what delivers the best agreement with the foreground-free data, i.e. the $\varepsilon$ which is closest to zero. However, as is already slightly discernible from the middle-panel, a low choice of $N_{FG}$ does provide a higher variance in the results, as shown from the $N_{FG} = 4$ case whose value for $\varepsilon$ fluctuates more than all other cases. This can be understood, by considering that in this situation a large amount of foreground residuals will be in the cleaned intensity map which cause more random and spurious correlations between the optical data and the foreground residuals. In other words, the higher levels of foreground residuals resulting from a lower $N_{FG}$ clean, will inevitably boost uncertainties in cross-correlations.

We investigated this further in the bottom panel of Figure 13 which show the estimated fractional errors $\sigma_P/|P(k)|$ for each data point. We calculated these using the same approaches used in real-data studies (e.g. Masui et al. (2013); Anderson et al. (2018)) by treating our MultiDark simulated data sets as real data. Alongside this we produced 100 lognormal mocks for both the H\textsc{i} intensity maps and optical galaxy maps. From this we can calculate a foreground transfer function (see Section 4.5.1). Applying this transfer function to each of the mocks and then measuring the variance in their results provides an estimate for the power spectrum errors $\sigma_P$.

The bottom panel of Figure 13 shows that errors are generally largest for the $N_{FG} = 4$ case, highlighting the important point that foreground residuals will boost errors. At low-$k$ we also see the fractional errors are extremely high for the $N_{FG} = 20$ case. This is because the power at these scales is damped so severely that the...
transfer function is essentially trying to recover a power spectrum from a position of \( P(k) \sim 0 \), which inevitably causes noisy results and boosts the variance. This demonstrates the balance required for an optimal choice of \( N_{\text{FG}} \). An \( N_{\text{FG}} \) which is too low causes too much foreground residual leading to large errors. An \( N_{\text{FG}} \) which is too high, damps the power spectrum drastically causing too much scatter in the power recovered through the transfer function. There is also a requirement for an \( N_{\text{FG}} \) balance at higher \( k \)-ranges. Looking at the zoomed-in section at \( 0.22 < k < 0.35 \) h/Mpc we clearly see errors are highest for \( N_{\text{FG}} = 4 \). They begin to decrease with an increasing \( N_{\text{FG}} \) but we eventually find this saturates and there is little or no improvement in error even with a big jump from \( N_{\text{FG}} = 10 \) to \( N_{\text{FG}} = 20 \). This means going arbitrarily high in \( N_{\text{FG}} \) will slowly stop improving errors but continue to bias results (increase \( |\varepsilon| \)).

The results from Figure 13 are therefore strong evidence that a fine balance must be reached for an optimal choice of \( N_{\text{FG}} \) in cross-correlations. This choice will also depend on the cosmological parameters being probed. For example, an investigation into primordial non-Gaussianity involves attempts to constrain the parameter \( f_{\text{NL}} \), which requires large scales (small-\( k \)) measurements. Going to \( N_{\text{FG}} \sim 10 \) in an attempt to minimise errors may not be plausible in this situation if the bias induced on these large scales is too strong. Conversely, probing something like the \( \text{H} \) abundance (\( \Omega_{\text{HI}} \)) which generally just affects the amplitude scaling of the power spectrum and can thus be probed at most scales, could potentially allow for a more aggressive clean that controls errors but still does not heavily bias the higher-\( k \) scales where this parameter can still be constrained. Therefore, it is unlikely that a universally optimum foreground treatment can ever be selected. This also supports conclusions from previous work that attempted to model the effects of foreground cleaning e.g. Cunnington et al. (2020a) and Soares et al. (2020). These works employed subtly different foreground modelling, which is likely due to the different range of scales they targeted given their science goals.

6 DISCUSSION

Evidence from pathfinder 21cm intensity mapping data (Masui et al. 2013; Wolz et al. 2017; Anderson et al. 2018) suggests that we will, at least initially, be requiring fairly aggressive foreground cleans in \( \text{H} \) intensity mapping surveys. This is potentially due to instrumental responses to foregrounds causing effects such as polarization leakage. Understanding the impact this has on probing of large-scale cosmic structure using \( \text{H} \) intensity maps is therefore paramount. In this paper we have provided a study into these issues by presenting a set of test data with a differing range of foreground contamination, both with and without effects from polarization leakage.

The contamination to the data from a foreground clean can manifest in two distinct ways: damping of cosmological \( \text{H} \) modes, and foreground residuals.

Damping cosmological \( \text{H} \):

Inevitably, some cosmological \( \text{H} \) modes will be degenerate with the 21cm foregrounds and their information will be contained in the same modes that are being removed. This has the effect of damping the \( \text{H} \) power spectrum mostly on large scales (small-\( k \)). For our mild cases (all regions without polarization leakage and even the SCP with polarization leakage), this effect is minimal (see Figure 4 and Figure 5). However, where higher \( N_{\text{FG}} \) cleans are needed this damping becomes severe and generally isolated to small-\( k \) modes (see Figure 10). In all cases though, a foreground transfer function (Section 4.5.1) should be able to compensate for this damping or alternatively, a foreground model with free nuisance parameters that can be marginalized over (Soares et al. 2020).

Foreground residuals:

Another challenging effect from foreground contamination comes from residuals left in the data after a clean. In mild cases where foregrounds are removed relatively easily, foreground residuals are small but we have shown evidence that there can still be an additive bias (see high-\( k \), values in Figure 7 with positive \( |\varepsilon| \) values). However, previous work (e.g. Cunnington et al. (2020b); Soares et al. (2020)) has shown that unbiased cosmological parameter estimates can be obtained even when contribution from residuals is not included, thus showing that for mild cases with low enough residuals, their effect is minimal. However, our results indicate that foreground residuals can be exacerbated by polarization leakage. For example, Figure 11 demonstrated how even with large \( N_{\text{FG}} \), the foreground residuals can still dominate over the remaining \( \text{H} \) signal in some cases. This would cause large additive biases and boost errors and needs to be modelled, as we outlined in Equation 29 and Equation 30. Quantifying the contribution from foreground residuals is not trivial with real data and this will be a key challenge for auto-correlation measurements. However, for cross-correlations with an overlapping optical galaxy survey the situation is somewhat simplified. Here the foreground residuals do not correlate with the foreground-free galaxy data, allowing less extreme cleans at the expense of a boost to errors (as demonstrated by Figure 13).

Not all of the regions we have chosen are representative of a target region one would choose in a real survey. For example, a 3000 deg\(^2 \) survey would not be aimed at the Galactic Plane. However, this simulated data provides a test case where a high-\( N_{\text{FG}} \) clean is required, which in this sense, is similar to early pathfinder experiments. Therefore it provides a means to begin investigating some of these issues in a simulated setting where we have full control and can separate contributions to the final observed signal.

Our investigation suggests that drawing general conclusions or recommendations for an optimal foreground treatment is not possible. An optimal method depends on the region being targeted, the instrumental calibration (e.g. susceptibility to polarization leakage), and also on the scales being targeted depending on the survey’s key science goals. But if one can achieve near-perfect instrument calibration, the results do appear fairly general. The top tow of Figure 5 shows all regions can be cleaned by blindly removing 3 principal components and this delivers similar damping to \( \text{H} \) power and a similar level of residual bias. However, residuals may differ between region for smaller-\( k \) but without affecting the accuracy (measured by \( |\varepsilon| \) and instead just affect the precision (boosting errors).

We introduced and tested three commonly employed blind foreground cleaning techniques: FASTICA, GMCA and PCA, the latter being mathematically equivalent to SVD (Section 4.1.1), and a polynomial fit (Section 4.1.2). We found all three blind methods deliver essentially equivalent results in all cases. We discussed this in detail in Section 5.1 – in summary, this is due to FASTICA and GMCA performing an initial PCA which they then try to improve upon by imposing spatial statistical independence and sparsity respectively. However, perhaps due to the sky size we use and the resolution, the foregrounds not included in the initial PCA reconstruction are not sufficiently sparse or non-Gaussian to be identified by GMCA or FASTICA and no discernible improvement is made.

We also trialed, for the first time on low redshift intensity mapping data, a semi-blind approach to foreground removal (Section 4.4). Our tests revealed that this method would need further improvement to be
competitive with a blind approach (see Figure 7). Furthermore, the method in its current form, is not robust to polarization leakage effects and performs poorly when this is included. Thus, this approach would be reliant on a near-perfect calibration of the intensity mapping instrument. For these reasons, this approach would only likely be viable for future surveys where calibration strategies are highly optimized and astrophysical parameters can be tightly constrained, allowing more precise fits to the foregrounds. A potential benefit from this is the possibility of further combinations with a full-blind approach in a hybridization. Section 5.3 examined the cross-correlation between a semi-blind GMCA cleaned map and a PCA cleaned one. This revealed that whilst little improvement can be gained in the accuracy of the final power spectrum, the foreground residuals in the two are subtly different and result in a lower contribution in the cross-measurement (Figure 9).

We hope our findings can be useful for analysing HI intensity mapping data from the MeerKAT intensity mapping survey (Santos et al. 2017; Li et al. 2020a), and for preparing cross-correlation strategies for MeerKAT and the SKA (Pourtsidou et al, 2020).

ACKNOWLEDGEMENTS

We are grateful to Paula Soares for helpful comments on the draft. SC is supported by STFC grant ST/S000437/1. JB, IPC and MI are supported by the European Union through the grant LENA (ERC StG no. 678282) within the H2020 Framework Program. AP is a UK Research and Innovation Future Leaders Fellow, grant MR/S016066/1, and also acknowledges support by STFC grant ST/S000437/1. This research utilised Queen Mary’s Apocrita HPC facility, supported by EPSRC. We acknowledge the use of open source software (Jones et al. 2001; Hunter 2007; McKinney 2010; Van Der Walt et al. 2011; Harris et al. 2020). Some of the results in this paper have been derived using the healpy and HEALPix package. We thank New Mexico State University (USA) and Instituto de Astrofisica de Andalucia CSIC (Spain) for hosting the Skies & Universes site for cosmological simulation products.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

Alonso D., Ferreira P. G., Santos M. G., 2014, MNRAS, 444, 3183
Alonso D., Bull P., Ferreira P. G., Santos M. G., 2015, Mon. Not. Roy. Astron. Soc., 447, 400
Anderson C., et al., 2018, Mon. Not. Roy. Astron. Soc., 476, 3382
Ansari R., et al., 2012, Astron. Astrophys., 540, A129
Asoory J., et al., 2020, MNRAS, 495, 1788
Battye R. A., Davies R. D., Weller J., 2004, Mon. Not. Roy. Astron. Soc., 355, 1339
Battye R. A., Browne I. W. A., Dickinson C., Heron G., Maffei B., Pourtsidou A., 2013, MNRAS, 434, 1239
Bharadwaj S., Nath B., Nath B. B., Sethi S. K., 2001, J. Astrophys. Astron., 22, 21
Bigot-Sazy M. A., et al., 2015, Mon. Not. Roy. Astron. Soc., 454, 3240
Blake C., Carter P., Koda J., 2018, Mon. Not. Roy. Astron. Soc., 479, 5168
Bobin J., Starck J.-L., Fadili J., Moudden Y., 2007, IEEE Transactions on Image Processing, 16, 2662
Bobin J., Starck J. L., Sureau F., Basak S., 2013, A&A, 550, A73
Bobin J., Sureau F., Starck J. L., Rassat A., Paykari P., 2014, A&A, 563, A105
Bobin J., Acero F., Picqueton A., 2019, in CAMSAP.
Carucci I. P., Irfan M. O., Bobin J., 2020b, 21 cm intensity mapping: a 900 - 1300 MHz full-sky simulation, - , doi:10.5281/zenodo.3991818, https://doi.org/10.5281/zenodo.3991818
Carucci I. P., Irfan M. O., Bobin J., 2020a, MNRAS,
Castorina E., White M., 2018, Mon. Not. Roy. Astron. Soc., 476, 4403
Chang T.-C., Pen U.-L., Peterson J. B., McDonald P., 2008, Phys. Rev. Lett., 100, 091103
Chapman E., et al., 2012, Mon. Not. Roy. Astron. Soc., 423, 2518
Cora S. A., 2006, Mon. Not. Roy. Astron. Soc., 368, 1540
Croton D. J., et al., 2016, Astrophys. J. Suppl., 222, 22
Cunnington S., Wola L., Pourtsidou A., Bacon D., 2019, Mon. Not. Roy. Astron. Soc., 488, 5452
Cunnington S., Camara S., Pourtsidou A., 2020a, preprint, -(arXiv:2007.12126)
Cunnington S., Pourtsidou A., Soares P. S., Blake C., Bacon D., 2020b, Mon. Not. Roy. Astron. Soc., 496, 415
Delabrouille J., Cardoso J.-F., Patanchon G., 2003, Mon. Not. Roy. Astron. Soc., 346, 1089
Dickinson C., Davies R. D., Davis R. J., 2003, MNRAS, 341, 369
Harper S., Dickinson C., 2018, Mon. Not. Roy. Astron. Soc., 479, 2024
Harper S., Dickinson C., Battye R., Roychowdhury S., Browne I., Ma Y.-Z., Oliviari L., Chen T., 2018, Mon. Not. Roy. Astron. Soc., 478, 2416
Harris C. R., et al., 2020, Nature, 585, 357–362
Hunter J. D., 2007, Computing In Science & Engineering, 9, 90
Hypervirani A., 1999, IEEE transactions on neural networks, 10 3, 626
Jones E., Oliphant T., Peterson P., et al., 2001–, SciPy: Open source scientific tools for Python, http://www.scipy.org/
Klypin A., Yepes G., Gottlober S., Prada F., Hess S., 2016, Mon. Not. Roy. Astron. Soc., 457, 4340
Knebe A., et al., 2018, Mon. Not. Roy. Astron. Soc., 474, 5206
Lacy M., et al., 2019, arXiv e-prints, p. arXiv:1907.01981
Leach S., et al., 2008, Astron. Astrophys., 491, 597
Li Y., Santos M. G., Grainge K., Harper S., Wang J., 2020a, preprint, -(arXiv:2007.01767)
Li L., Staveley-Smith L., Rhee J., 2020b, preprint, -(arXiv:2008.04081)
Liao Y.-W., Chang T.-C., Kuo C.-Y., Masui K. W., Oppermann N., Pen U.-L., 2016, Mon. Not. Roy. Astron. Soc., 460, 1647
Liu A., Tegmark M., 2011, Phys. Rev. D, 83, 103006
Liu A., Shaw J. R., 2020, Publ. Astron. Soc. Pac., 132, 062001
Liu A., Tagman M., 2011, Phys. Rev. D, 83, 103006
Mandelbaum R., et al., 2011, Mon. Not. Roy. Astron. Soc., 410, 844
Masui K. W., Oppermann N., Pen U.-L., 2016, Astrophys. J., 833, 289
Miller J. B., 2016, Astrophys. J., 833, 289
Mertens F., Ghosh A., Koopmans L., 2018, Mon. Not. Roy. Astron. Soc., 478, 3640
Mertens F., et al., 2020, Mon. Not. Roy. Astron. Soc., 493, 1662
Millville-Deshénênes M.-A., Ysard N., Lavabre A., Ponthieux N., Macías-Pérez J. F., Aumont J., Bernard J. P., 2008, A&A, 490, 1093
Moore D. F., Aguirre J. E., Parsons A. R., Jacobs D. C., Peber J. C., 2013, Astrophys. J., 769, 154
Oliviari L., 2018, PhD thesis, University of Manchester
Oliviari L., Remazeilles M., Dickinson C., 2016, MNRAS, 456, 2749
Oliviari L., Dickinson C., Battye R. A., Ma Y. Z., Costa A. A., Remazeilles M., Harper S., 2018, MNRAS, 473, 4242
Oppermann N., et al., 2012, A&A, 542, A93
Patil A. H., et al., 2017, Apl, 838, 65
Pedregosa F., et al., 2011, J. Machine Learning Res., 12, 2825

MNRAS 000, 1–20 (2020)
Picquenot A., Acero F., Bobin J., Maggi P., Ballet J., Pratt G. W., 2019, A&A, 627, A139
Planck Collaboration et al., 2016, Astron. Astrophys., 594, A13
Planck Collaboration et al., 2020, Astron. Astrophys., 641, A4
Pourtsidou A., Bacon D., Crittenden R., 2017, Mon. Not. Roy. Astron. Soc., 470, 4251
Remazeilles M., Dickinson C., Banday A. J., Bigot-Sazy M.-A., Ghosh T., 2015, MNRAS, 451, 4311
SKA Cosmology SWG et al., 2020, Publ. Astron. Soc. Austral., 37, e007
Santos M. G., Cooray A., Knox L., 2005, ApJ, 625, 575
Santos M., et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14), p. 19 (arXiv:1501.03989)
Santos M. G., et al., 2017, in Proceedings, MeerKAT Science: On the Pathway to the SKA (MeerKAT2016): Stellenbosch, South Africa, May 25-27, 2016. (arXiv:1709.06099)
Shaw J., Sigurdson K., Sitwell M., Stebbins A., Pen U.-L., 2015, Phys. Rev. D, 91, 083514
Soares P. S., Cunnington S., Pourtsidou A., Blake C., 2020, preprint, -(arXiv:2008.12102)
Spinelli M., Zoldan A., De Lucia G., Xie L., Viel M., 2020, Mon. Not. Roy. Astron. Soc., 493, 5434
Switzer E., et al., 2013, Mon. Not. Roy. Astron. Soc., 434, L46
Switzer E. R., Chang T.-C., Masui K. W., Pen U.-L., Voytek T. C., 2015, Astrophys. J., 815, 51
Van Der Walt S., Colbert S. C., Varoquaux G., 2011, preprint, (arXiv:1102.1523)
Villaescusa-Navarro F., Alonso D., Viel M., 2017, Mon. Not. Roy. Astron. Soc., 466, 2736
Villaescusa-Navarro F., et al., 2018, Astrophys. J., 866, 135
Wolz L., Abdalla F., Blake C., Shaw J., Chapman E., Rawlings S., 2014, Mon. Not. Roy. Astron. Soc., 441, 3271
Wolz L., et al., 2017, Mon. Not. Roy. Astron. Soc., 464, 4938
Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, Journal of Open Source Software, 4, 1298
Zuo S., Chen X., Ansari R., Lu Y., 2018, Astron. J., 157, 4

This paper has been typeset from a \TeX/\LaTeX file prepared by the author.