

# Needlet Karhunen–Loève (NKL): a method for cleaning foregrounds from 21 cm intensity maps

John Podczerwinski<sup>★</sup> and Peter T. Timbie

*Department of Physics, University of Wisconsin-Madison, 1150 University Avenue, Madison, WI 53706, USA*

Accepted 2023 October 31. Received 2023 October 25; in original form 2023 August 6

## ABSTRACT

This paper introduces a technique called needlet Karhunen–Loève (NKL), which cleans both polarized and unpolarized foregrounds from H I intensity maps by applying a Karhunen–Loève transform on the needlet coefficients. In NKL, one takes advantage of correlations not only along the line of sight, but also between different angular regions, referred to as ‘chunks’. This provides a distinct advantage over many of the standard techniques applied to map space that one finds in the literature, which do not consider such spatial correlations. Moreover, the NKL technique does not require any priors on the nature of the foregrounds, which is important when considering polarized foregrounds. We also introduce a modified version of Generalized Needlet Internal Linear Combination (GNILC), referred to as MGNILC, which incorporates an approximation of the foregrounds to improve performance. The NKL and MGNILC techniques are tested on simulated maps which include polarized foregrounds. Their performance is compared to the GNILC, generalized morphological component analysis, independent component analysis, and principal component analysis techniques. Two separate tests were performed. One at  $1.84 < z < 2.55$  and the other at  $0.31 < z < 0.45$ . NKL was found to provide the best performance in both tests, providing a factor of 10–50 improvement over GNILC at  $k < 0.1 \text{ hMpc}^{-1}$  in the higher redshift case and  $k < 0.03 \text{ hMpc}^{-1}$  in the lower redshift case. However, none of the methods were found to recover the power spectrum satisfactorily at all baryon acoustic oscillations scales.

**Key words:** techniques: image processing – techniques: interferometric.

## 1 INTRODUCTION

One of the main endeavours of observational cosmology is to measure statistical properties of the spatial distribution of matter in the Universe. Such measurements are of interest as they could provide us with information about dark energy, inflationary physics, the growth of structure, early star formation, and more. The distribution of matter in the Universe is traced by neutral hydrogen (H I). Using 21 cm emission or absorption to measure these density fluctuations as a function of redshift is a technique referred to as 21 cm (or H I) intensity mapping (IM) (Bharadwaj & Sethi 2001; Battye, Davies & Weller 2004; Madau, Meiksin & Rees 1997). Most IM instruments are radio interferometers; some notable examples are the Canadian Hydrogen Intensity Mapping Experiment (CHIME, Bandura et al. 2014; Tianlai Chen 2012), the Hydrogen Epoch of Reionization Array (HERA, DeBoer & HERA 2015), and the Murchison Widefield Array (MWA, Morales 2005). The antenna elements that make up these interferometers can take several forms, including parabolic dishes (Tianlai, CHIME, and HERA), cylindrical reflectors (CHIME and Tianlai) or phased arrays (MWA). Single-dish instruments are also used, with some examples being the Five hundred meter Aperture Spherical Telescope (FAST, Bigot-Sazy et al. 2015), the Greenbank Telescope (GBT, Switzer et al. 2013),

and the More Karoo Array Telescope (MeerKAT, Santos et al. 2016). FAST and GBT both consist of a single, very large dish. On the other hand, MeerKAT would average signals from 64 smaller (13.5 m diameter) dishes. MeerKAT can also operate as an interferometer.

Measurement of the 21 cm line would provide valuable cosmological information. At low redshifts ( $z \leq 6$ ), IM would serve as a complement to galaxy redshift surveys. At such redshifts, cosmologists would be particularly interested in measuring baryon acoustic oscillations (BAO). The BAO would serve as a standard ruler, providing information about the expansion of the Universe and dark energy. Measurements during the epoch of reionization  $6 < z < 20$  would provide information about the formation of the earliest stars. Lastly, measurements during the cosmic dark ages ( $20 < z < 1100$ ), before the formation of the first luminous objects, would provide insights into the physics of inflation. It should be noted that no other probe aside from IM is capable of mapping the cosmic dark ages.

Although promising, IM is still a young technique, with its systematic effects and calibration requirements still being understood. In particular, foregrounds from Galactic synchrotron emission are an unsolved challenge for IM experiments. In the case of Galactic synchrotron emission, these foregrounds are up to a factor of  $10^5$  brighter than the H I signal.

Luckily, the unpolarized component of these foregrounds is expected to be spectrally smooth, while the H I signal is expected to have a high level of chromaticity. In other words, the foregrounds

<sup>★</sup> E-mail: [jpodczerwins@wisc.edu](mailto:jpodczerwins@wisc.edu)

are confined to a relatively small subsection of delay space (Datta, Bowman & Carilli 2010). In this context, delay refers to the Fourier dual of spectral frequency. This delay-space quarantining aids in the avoidance and removal of the foregrounds. Unluckily, this region of delay space will correspond to large spatial scales, which are important for measuring the BAO.

One must also consider the polarized component of Galactic synchrotron radiation. This component is affected by Faraday rotation in the Galaxy, introducing higher chromaticity than the unpolarized component. Moreover, the chromaticity of this component is expected to vary with line-of-sight direction. In particular, lines of sight closer to the Galactic plane will suffer from more severe chromaticity due to stronger Faraday rotation. Moreover, the power of the Galactic synchrotron radiation (both polarized and unpolarized components) relative to the signal is expected to vary with angular scale (Alonso, Ferreira & Santos 2014), with the largest scales having the worst contamination. In principle, these polarized foregrounds could be avoided altogether as long as the beams of the telescope have rotational symmetry and low cross-polar levels. However, such refined beams are not achievable in practice, and one must contend with some amount of polarization leakage. This leakage is expected to be on the per cent level for IM experiments (Moore et al. 2013, 2017).

Over the years, many foreground removal methods have been proposed. A helpful review of many of these may be found in Liu & Shaw (2020). Methods that require no priors on the H I, noise or foregrounds are described as ‘blind’, and those requiring priors are described as ‘non-blind’. One will also find methods being tested both on raw visibilities or on maps synthesized from data. However, unlike blind and non-blind, there is not a clear distinction here. Some methods may be applicable both in map and visibility spaces.

Principal component analysis (PCA), independent component analysis (ICA, Chapman et al. 2012a), and Generalized Morphological Component Analysis (GMCA, Chapman et al. 2012b) are the prominent blind methods that one finds being applied to maps. In the literature, one can also find non-blind methods being applied in map space. Examples include Generalized Needlet Internal Linear Combination (GNILC, Olivari, Remazeilles & Dickinson 2015), Gaussian Process Regression (GPR, Mertens, Ghosh & Koopmans 2018), and the ‘semibind’ Singular Value Decomposition (SVD) method proposed in Zuo, Chen & Mao (2023). On the other hand, tests of cleaning methods on visibilities are less commonly found in the literature. One blind method that can be used in visibility space is ‘foreground avoidance’. In this method, one exploits the fact that the foregrounds tend to be confined to a region of certain region  $k$ -space, referred to as ‘the wedge’ (Ewall-Wice et al. 2016). Power spectrum modes within this region of  $k$ -space are then excluded from the analysis. Non-blind methods have also been tested in visibility space. The beam projection plus Karhunen–Loève (KL) transform proposed in Shaw et al. (2015) and the GPR method used in Soares et al. (2021) and Mertens et al. (2018) are examples we have found in the literature.

However, for the case of post-epoch of reionization (EOR) surveys, such avoidance methods come at the cost of losing valuable information about BAO. The BAO are located roughly in the wavenumber range  $0.03 \leq k \leq 0.4 \text{ hMpc}^{-1}$  (Bull et al. 2015). Much information about large length-scales such as these would be lost when taking a foreground avoidance approach.

So far, when real data are considered, the foregrounds have usually been handled in a conservative way. For instance, in Paul et al. (2023), the first detection of H I using IM without cross-correlating with galaxies, the analysts chose to use foreground avoidance rather

than removal. Foreground avoidance was also used in Amiri et al. (2023). In this work, the authors cross-correlated data from the CHIME telescope with galaxies and quasars from the Extended Baryon Acoustic Oscillation Survey (eBOSS). In this work, it was found that the foreground avoidance method excluded length-scales important for characterizing the BAO. Wolz et al. (2021) present results from cross-correlating GBT data with the eBOSS survey. In this case, the authors opted to use FastICA paired with a transfer function to compensate for signal loss. In Cunnington et al. (2022), the authors cross-correlated maps from MeerKAT with WiggleZ galaxies. The authors of this paper cleaned their data with PCA and used a transfer function to compensate for signal loss. Anderson et al. (2018) cross-correlated maps from the Parkes telescope with galaxies from the Two-degree-Field Galaxy Redshift Survey (2dF) galaxy survey. These authors correlated maps from different seasons to reduce noise bias, and applied a transfer function to compensate for signal loss.

One can also find numerous papers in the literature testing these various methods on simulated data. Such tests are usually conducted at lower redshifts ( $z \lesssim 0.6$ ) and take into account beam effects by convolving with a Gaussian profile of appropriate width. In Carucci, Irfan & Bobin (2020), tests were conducted using simulations that assumed a telescope similar to MeerKAT operating in single-dish mode at low redshifts ( $0.09 \leq z \leq 0.58$ ) and surveying the full sky. This test included contributions from polarized foregrounds simulated using the Cosmological Realizations for Intensity Mapping Experiments (CRIME) software package (Alonso et al. 2014). In this instance, GMCA recovered the angular power spectrum of the H I plus noise with errors of 5 per cent down to  $\ell \approx 25$ . Moreover, they found that GMCA provided lower errors than ICA (Carucci et al. 2020). In Olivari et al. (2015), GNILC was tested at low redshifts ( $0.13 \leq z \leq 0.48$ ) in a survey covering half of the sky. These simulated maps included no polarized foregrounds. In this case, GNILC was found to recover the angular power spectrum of the H I to within 6 per cent error down to  $\ell = 30$ . So, like GMCA, GNILC also seemed to struggle at large angular scales. Recently, GNILC, GMCA, and ICA were all tested on simulations of the Baryon Acoustic Oscillations from Integrated Neutral Gas Observations (BINGO) experiment (Marins et al. 2022). In this test, all of the techniques were found to provide statistically equivalent results. In Soares et al. (2021), GPR was tested on simulated maps at low redshifts ( $0.18 < z < 0.58$ ). The simulated maps included polarized foregrounds generated using CRIME. These tests were conducted on Stripe82 and a  $3000 \text{ deg}^2$  region at the South Celestial Pole. In these tests, GPR and PCA provided similar performance, recovering the power spectrum with errors around 10 per cent for all spatial scales considered.

A notable cleaning method applied in visibility space is the combined beam projection and KL transform proposed and tested in Shaw et al. (2015). This test assumed a simplified version of the CHIME instrument and was conducted at higher redshifts ( $1.84 < z < 2.55$ ) where polarized foregrounds are more severe. The test used simulated maps produced by the Cosmology in the Radio Band (CORA) software package. This package makes different assumptions than CRIME, resulting in less severe chromaticity in the polarized foregrounds. In this paper, instrumental effects were accounted for in a more realistic way, by generating visibilities from maps using simulated beams from cylinder telescopes. This method was found to effectively recover the H I power spectrum down to  $k \approx 0.02 \text{ hMpc}^{-1}$ . This result appears quite promising, since the BAO would be recovered at all length-scales. However, the weakness of this approach is that it requires a highly accurate beam model along

with priors on the unpolarized foregrounds, HI, and noise. It was found that this approach breaks down when main beam uncertainties exceed 0.1 per cent. In addition, amplifier gains must be known to better than 1 per cent within each minute. Such accurate beam and gain calibration pose a significant challenge and may not be possible in practice.

In summary, published map-space tests of cleaning methods on simulated polarized foregrounds have only been conducted at low redshifts. Moreover, these papers usually take beam effects into account in a simple way. The only exception we found is in the work of Hothi et al. (2020), who performed tests on maps generated from visibilities. Such tests are informative, but not the whole story since these polarized foregrounds will become more severe as redshift increases. On the other hand, Shaw’s method was found to work extremely well when faced with polarized foregrounds at higher redshifts, but required precise knowledge of the beam. It is reasonable to imagine that the previously described methods used in map space might be more robust to beam mis-calibration than Shaw’s method. There are two reasons for this. One is that the beam projection part of beam projection/KL approach removes the polarized foregrounds by projecting onto the null space of the polarized beam matrix. This step may not work when the beams are not well understood. Another reason is that the KL part of the SVD/KL cleaning requires one to have priors on how the unpolarized foregrounds will contribute to the total visibility. On the other hand, the map-space applicable methods described earlier in this paper (aside from the ‘semiblind’ SVD method from Zuo et al. 2023) either do not require a foreground model, or, in the case of GPR, estimate one from the data.

Although they may end up being more robust to calibration issues than the visibility-space beam projection/KL method, the available methods usable in map space are missing certain strengths of the beam projection/KL approach. One particular strength of beam projection/KL is that it takes full advantage of the available priors. In particular, during the KL step, one considers correlations between all baselines in the telescope. In this way, one is exploiting correlations in both frequency and spherical harmonics (different baselines are sensitive to different  $l$  modes). This is much different from commonly used map-space methods, such as PCA, GNILC, ICA, and GPR, which, as employed so far in the literature, only consider correlations along the line of sight. Although there are some exceptions, such as the commonly used GMCA method does incorporate angular correlations by enforcing sparsity of components in the wavelet domain. Another exception to this is the ‘semiblind’ SVD method. This method provides excellent recovery of the HI signal, but, like Shaw’s method, suffers from requiring priors on the foregrounds present in the maps.

However, one must also keep in mind that working in map-space presents certain advantages not available in visibility space. For one, map space allows for specific pixels to be selected. As such, one can mask out pixels with particularly strong foregrounds or artefacts. Moreover, one can perform a spherical-harmonic or spherical wavelet transform of the data, allowing for precise separation by angular scale. Such operations are not possible in visibility space. The closest thing one could do is separate the visibilities into  $m$  modes, where the  $m$  refers to the azimuthal  $m$  found in spherical harmonics (Shaw et al. 2015). Such freedom could be useful when cleaning foregrounds from maps, due to their dependence on line-of-sight direction and angular scale.

Based on this review of the literature, there several tasks that ought to be performed.

(i) Create a foreground cleaning method that is usable in map space and uses both frequency and angular correlations. This method should not require a prior on the foregrounds.

(ii) Test the available map-based methods on polarized foregrounds at redshifts  $z \gtrsim 0.6$ .

(iii) Test the robustness of various map-based methods against beam mis-calibration.

We make progress on task (i) by introducing the needlet Karhunen–Loève (NKL) method of foreground removal. In NKL, different sections of pixel/spherical harmonic space are separated via a needlet transform. The needlet coefficients are then cleaned using a KL transform that exploits both angular and frequency correlations. This is different from most other map-based methods, which only consider frequency correlations. Moreover, this is all done without needing any priors on the foregrounds present in the map. Not requiring a foreground prior is quite useful, as polarized foregrounds are not well understood. Moreover, NKL cleans the foregrounds differently depending on location and angular scale. This is also a desirable feature, as the foreground brightness and chromaticity are expected to vary with line of sight and angular scale.

In this paper, we also make some progress on task (ii) of the list by testing GNILC, GMCA, ICA, and PCA at redshifts ( $1.84 < z < 2.55$ ). PCA, ICA, and GMCA were chosen as these are all prominent in the literature. GNILC was chosen since it has similarities to the NKL method proposed in this paper.

We leave task (iii) for future work.

In Section 2, we describe the process of foreground cleaning in an abstract way and provide additional motivation for the NKL technique. In Section 3, we provide a detailed description of the NKL technique. In this section, we also introduce modified GNILC (MGNILC), a slightly modified version of GNILC which uses a foreground approximation acquired from the data. In Section 4, we describe the maps we used to perform our tests. In Section 5, we present results acquired by testing GNILC, ICA, PCA, GMCA, and NKL on simulated maps. We then summarize our results and conclude in Section 6.

## 2 FOREGROUND REMOVAL TECHNIQUES

In this paper, we will compare the performance of the proposed NKL technique with that of other techniques commonly found in the literature. In particular, we will be considering PCA (Cunnington et al. 2021), ICA (Wolz et al. 2014), GMCA (Carucci et al. 2020), and GNILC (Olivari et al. 2015). Although not quite obvious at first, these techniques are in fact very similar. In particular, all techniques tested in this paper will model the foregrounds (or signal plus noise in the case of GNILC) as a mixture of template maps.

### 2.1 Foreground removal using templates

All techniques considered in this paper begin by assuming that the maps produced by an IM experiment are given by

$$\mathbf{X} = \mathbf{f} + \mathbf{h} + \mathbf{n}. \quad (1)$$

In this formula,  $\mathbf{X}$  is a  $n_{\text{ch}} \times n_{\text{p}}$  matrix, where  $n_{\text{ch}}$  is the number of frequency channels and  $n_{\text{p}}$  is the number of pixels in each map. Moreover,  $\mathbf{f}$  represents the foregrounds,  $\mathbf{h}$  represents the HI signal and  $\mathbf{n}$  represents the noise. In this paper, we will use bold roman sans serif font for matrices. For vectors, we will use bold italic font, or bold roman when upper case greek letters are used. We will

also define

$$\mathbf{s} = \mathbf{h} + \mathbf{n}. \quad (2)$$

It is then assumed that the foregrounds can be expressed as a mixture of templates:

$$\mathbf{f} = \mathbf{A}\mathbf{S}. \quad (3)$$

In this formula,  $\mathbf{S}$  is a  $n_t \times n_p$  matrix of templates, where  $n_t$  is the number of templates and  $\mathbf{A}$  is a  $n_{ch} \times n_t$  ‘mixing matrix’ which encodes how the templates are combined at each frequency channel. The foreground removal process then becomes a matter of fitting  $\mathbf{A}\mathbf{S}$  to  $\mathbf{X}$ , subject to some sort of regularization.

In PCA, one seeks to find templates that capture as much variation in the data as possible. In particular, these templates are really just the dot product of the data with the eigenvectors corresponding to the largest  $n_t$  eigenvalues of the covariance matrix. In this context, the covariance matrix is usually generated assuming that the brightness of the sky along each line of sight were independently drawn from some distribution. This assumption is not true, but is close enough to reality that PCA can still provide reasonable results. In the context of 21 cm IM, one usually estimates the covariance from the data as

$$\hat{\mathbf{C}} = \frac{1}{n_p}(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T, \quad (4)$$

where row  $i$  of  $\bar{\mathbf{X}}$  is the average of row  $i$  of  $\mathbf{X}$ . In this paper, we will use hats to denote covariance estimates. On the other hand, we will use matrices without hats to denote true covariances. In this case,  $\hat{\mathbf{C}}$  is a  $n_{ch} \times n_{ch}$  matrix. In ICA, one seeks templates that are statistically independent. In GMCA, one assumes that the templates  $\mathbf{S}$  ought to be sparse in some wavelet domain. ILC methods such as GNILC seek to find a filter that has a unit response to  $\mathbf{h} + \mathbf{n}$ , while minimizing the variance of the residuals (Olivari et al. 2015). Readers can refer to Marins et al. (2022) for a more rigorous description of how these standard techniques are regularized.

Both Shaw’s work and the NKL technique described in this paper make use of the KL transform (Tegmark, Taylor & Heavens 1997) for foreground cleaning. The KL transform uses the covariance matrices of both the foregrounds and signal to clean the data. These covariances either come from priors or are estimated from the data itself. For NKL, we estimate the foreground covariance from the data and get the HI and noise covariances from priors. Let  $\mathbf{C}_{FG}$  be the foreground covariance and  $\mathbf{C}_S$  be the signal covariance. For now, let us consider only frequency correlations, resulting in matrices of size  $n_{ch} \times n_{ch}$ . In practice, the covariances used in NKL will have larger dimension since they will include angular correlations as well. However, this discussion will be clearer if we consider only the  $n_{ch} \times n_{ch}$  case, and a mixing matrix plus template model can still be used to describe the case involving angular correlations.

The signal model  $\mathbf{C}_S$  includes the statistics of whatever components the analyst would like to recover from beneath the foregrounds. In this paper, we will take  $\mathbf{C}_S = \mathbf{C}_{HI}$  unless otherwise noted. It should be noted that the freedom to choose  $\mathbf{C}_S = \mathbf{C}_{HI}$  will provide NKL with an advantage over the other techniques mentioned in this paper. These techniques make no distinction between the HI and noise. The KL technique works by solving the generalized eigenvalue problem

$$\mathbf{C}_{FG}\Phi = \mathbf{C}_S\Phi\Lambda. \quad (5)$$

In this formula,  $\Phi$  is a matrix of eigenvectors and  $\Lambda$  is a diagonal matrix of eigenvalues. It turns out that the eigenvectors  $\Phi$  obtained are a solution to the optimization problem (Ghojogh, Karray &

Crowley 2023):

$$\max_{\Phi} \text{tr}(\Phi^T \mathbf{C}_{FG} \Phi) \quad (6)$$

subject to  $\Phi^T \mathbf{C}_S \Phi = \mathbf{I}$ .

So, the KL transform finds modes that have as high a ratio of foreground to signal as possible. In particular, the eigenvalues  $\lambda_i$  indicate the expected ratio of foreground to signal power at that particular mode. One then cleans the data by removing modes that are foreground dominated. In the language of templates and mixing matrices, we find that the template matrix is

$$\mathbf{S} = \mathbf{P}_s(\mathbf{X} - \bar{\mathbf{X}}). \quad (7)$$

In this formula,  $\mathbf{P}_s$  is a  $n_t \times n_{ch}$  matrix whose rows are the foreground-dominated eigenvectors. Moreover, let the symbol  $\mathbf{P}$  to represent a  $n_{ch} \times n_{ch}$  matrix whose rows are the eigenvectors generated by the generalized eigenvalue problem. The mixing matrix is then

$$\mathbf{A} = \mathbf{P}_s^{-1}, \quad (8)$$

where  $\mathbf{P}_s^{-1}$  is a  $n_{ch} \times n_t$  matrix containing only columns of  $\mathbf{P}^{-1}$  that correspond to foreground-dominated modes.

## 2.2 Discussion of techniques

Before actually testing any techniques, it will be beneficial to discuss differences between the techniques considered in this paper. In particular, we will discuss these differences and try to provide some motivation for why certain techniques may provide better performance than others.

First, one should note that GNILC requires a model for the HI and noise, making it a ‘non-blind’ technique. This is different from GMCA, PCA, and ICA, where the only free parameter provided by the user is the number of templates to use.

Another aspect of these techniques to consider is locality. In the standard version of GMCA, the matrices  $\mathbf{A}$  and  $\mathbf{S}$  are meant to capture the foregrounds at all pixels and angular scales (Carucci et al. 2020). PCA and ICA are typically conducted in a similar way, where one estimates the covariance  $\mathbf{C}$  using the entire data set. On the other hand, GNILC cleans the maps in a more fine-grained way. In particular, it divides the data into needlet coefficients, and then cleans the data one coefficient at a time. This allows for the foregrounds to be cleaned differently depending on the location and angular scale in question.

One can imagine reasons why such a fine-grained treatment might provide advantages. For instance, the ratio of foreground to HI power is expected to vary with angular scale. In particular, large angular scales will suffer worse contamination than smaller ones. When working with unpolarized foregrounds, one would expect similar chromaticity at all lines of sight. However, when dealing with polarized foregrounds, we expect for the chromaticity to vary with line of sight. As such, one would expect for pixels close to the Galactic plane to require more templates to clean than ones far from the plane. Thus, it seems likely that a more fine-grained approach would work better when dealing with polarized foregrounds.

It is certainly possible to imagine changing PCA, ICA or GMCA to make them more local. For instance, one could implement a scale- and location-dependent version of PCA where the covariance matrix is estimated for some neighbourhood around each needlet coefficient. In addition, a technique called L-GMCA has been proposed Bobin (2017) in which one uses different mixing matrices for different regions of the map. For now though, we will consider only the global versions of these techniques.

So, we see reasons why a more localized approach would likely provide better results when cleaning foregrounds from maps. So far, GNILC is the most local approach that has been found in the literature. In Olivari et al. (2015), this approach was tested on low- $z$  simulated maps without any polarization leakage included (Olivari et al. 2015). This study found GNILC outperforming PCA at angular scales ( $\ell > 30$ ). However, GNILC provided worse performance than PCA at scales larger than that. Interestingly, Carucci et al. (2020) found that GMCA also struggled at large angular scales  $\ell < 50$ , at least in the case when polarized foregrounds were included.

In the next section, we introduce NKL. Similarly to GNILC, NKL is a non-blind technique which acts on needlet coefficients. However, while GNILC only considers statistics along the line of sight, NKL also considers angular correlations in needlet space.

### 3 IMPLEMENTING NKL

In this section, we introduce the NKL technique for removing foregrounds from 21 cm maps. We will begin this section by providing background knowledge required for understanding how NKL works. In particular, the first subsection introduces needlets. In the second, we describe the way in which needlet coefficients are partitioned before cleaning is performed. Then, we describe the ways in which we can generate an approximation of the foregrounds from the data. In the final subsection, we provide a list of steps for performing NKL.

#### 3.1 The needlet transform

The NKL process begins by performing a needlet transform on each frequency slice of the 3D maps. needlets are wavelet-like functions that have a finite width in both  $\ell$  and pixel spaces. These functions were first presented in Narcowich, Petrushev & Ward (2006). For our purposes, we computed these coefficients using the PYS2LET software package, details of which can be found in Leistedt et al. (2013). Needlets are defined via

$$\psi_{jk}(\hat{n}) = \sqrt{\lambda_{jk}} \sum_l b\left(\frac{l}{B^j}\right) \sum_{m=-l}^l \bar{Y}_{lm}(\hat{n}) Y_{lm}(\xi_{jk}). \quad (9)$$

In this formula,  $b\left(\frac{l}{B^j}\right)$  is a bandpass function that is non-zero for  $B^{j-1} \leq l \leq B^{j+1}$ . In this paper, we choose  $B = 2$  and use  $4 \leq j \leq 8$ . The very lowest  $\ell$  modes are described using a ‘scaling function’ (Leistedt et al. 2013). The variable  $\xi_{jk}$  refers to the line of sight at which the needlet is centred. In this paper,  $\xi_{jk}$  refers to the location of pixel  $k$  in a HEALPIX map with  $n_{\text{side}} = 2^{j+1}$ . The variable  $\lambda_{jk}$  then refers to the solid angle of the HEALPIX pixel in question. Needlet coefficients can then be obtained simply by computing the following integral

$$\chi_{jk} = \int T(\hat{n}) \psi_{jk}(\hat{n}) d\Omega. \quad (10)$$

In this formula,  $T(\hat{n})$  is the function for which the needlet coefficients are being computed. In this work,  $T(\hat{n})$  will be the sky temperature. One can then reconstruct their temperature map using these coefficients as follows

$$T(\hat{n}) = \sum_{jk} \chi_{jk} \psi_{jk}(\hat{n}). \quad (11)$$

Marinucci et al. (2007) provide more detail on needlets and their use in cosmology.

Chunking example,  $n_{\text{side}} = 64, N = 32$

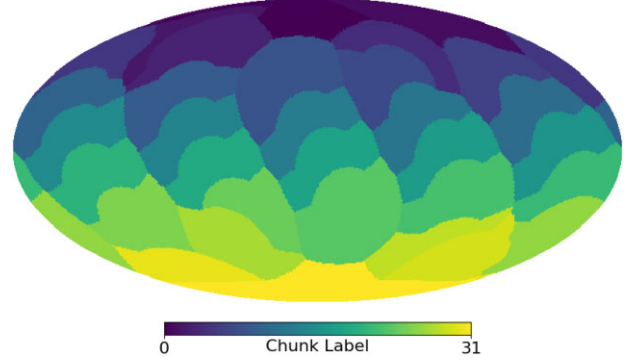


Figure 1. An example of dividing a needlet map into chunks.

#### 3.2 Partitioning of the needlet coefficients

An important part of NKL is the partitioning of the needlet coefficients. We refer to the groups of partitioned coefficients as ‘chunks’. In Fig. 1, we present an example of a chunked map. For now, let us suppose we are partitioning the coefficients for needlet scale  $j$ . Let us package the needlet coefficients at this scale of interest in a matrix  $\chi^{(j)}$ . This matrix will have dimensions  $n_{\text{ch}} \times n_{\text{co}}^{(j)}$ , where  $n_{\text{ch}}$  is the number of frequency channels and  $n_{\text{co}}^{(j)}$  is the number of needlet coefficients per frequency channel at scale  $j$ .

Next, let  $\chi_i^{(j)}$  be the row of  $\chi^{(j)}$  corresponding to frequency channel  $i$ . The process of partitioning the row into  $N$  chunks begins by selecting the coefficients located at columns  $q = a * n_{\text{co}}^{(j)} / N$  for  $a = 1, 2, \dots, N$ . Note here that  $q$  is indexing pixels in the HEALPIX map. Let us refer to these coefficients as ‘anchors’. Next, we assign all the other coefficients in the row to chunks according to which anchor point they have the smallest angular separation from. Adjacent neighbourhoods then swap coefficients until all contain the same number of coefficients.

We then use the same chunk assignments for all rows in  $\chi^{(j)}$ . This results in a new ‘chunked’ matrix  $\underline{\chi}^{(j)}$  which has dimensions  $n_{\text{ch}} N \times n_{\text{co}}^{(j)} / N$ . One row of  $\underline{\chi}^{(j)}$  will contain the coefficients of one chunk at one frequency channel. Column  $q$  will contain the  $q$ th element from each chunk at all frequency channels.

#### 3.3 Producing a foreground approximation

In order to perform a KL transform, we need estimates for  $\mathbf{C}_{\text{FG}}$  and  $\mathbf{C}_{\text{H1}}$ . The estimate of  $\mathbf{C}_{\text{H1}}$  is provided as a prior chosen by the analyst. For the purposes of this paper, we estimate  $\mathbf{C}_{\text{H1}}$  from maps simulated by CORA (Shaw et al. 2015). One may also want to estimate  $\mathbf{C}_{\text{N}}$  to include in the KL transformation or for later debiasing. In this paper, we estimate  $\mathbf{C}_{\text{N}}$  from the model used to simulate the noise. In a real experiment, this noise covariance could be determined from the data. On the other hand, we seek to estimate  $\mathbf{C}_{\text{FG}}$  from the data itself. This approach is motivated by two facts. First, as of now, polarized Galactic synchrotron radiation is not well enough measured or modelled to use an a priori covariance. In addition, a method that estimates the foregrounds directly from data will be more robust against errors in calibration and systematics that may cause the data to deviate from what would otherwise be an accurate a priori model. It should be noted that relatively small errors in the model of  $\mathbf{C}_{\text{FG}}$  can cause significant problems, due to the large dynamic range between the foregrounds and H1. In this paper, two possible approaches to foreground estimation are presented, PCA and Discrete Prolate Spheroidal Sequences (DPSS) Approximate

lazY filtEriNg of foregroUnds (DAYENU, Ewall-Wice et al. 2020). In the DAYENU approach, one assumes that the foregrounds are confined within some region of delay space. This assumption still holds in our CRIME generated foreground maps, at least for lines of sight outside the Galactic plane. This assumption can be useful, since it means that the HI for larger delays will not be affected by the filter. This is different from PCA, where some signal loss outside the foreground-dominated region of delay space may occur. Moreover, unlike PCA, the DAYENU method treats each pixel (or needlet coefficient) uniquely. In PCA, one is using the same templates to approximate the foregrounds for all pixels/coefficients. The main weakness of DAYENU, however, is that it does not work when all delays are contaminated by foregrounds.

### 3.3.1 Estimating foregrounds using PCA

This process begins by estimating the frequency–frequency covariance of the needlet coefficient maps. For this paper, we performed PCA foreground estimates one needlet scale at a time and without any chunking. PCA was not performed on chunked maps, as our scheme for selecting the number of components to remove struggled in this case. It is possible that applying PCA to each chunk individually would provide a better approximation, but we leave this to future testing. Once again, we are making the assumption that each line of sight is drawn independently and identically from some distribution. Since the needlet coefficients at a particular frequency channel have 0 mean, we can estimate their covariance as

$$\hat{\mathbf{C}}^{(j)} = \frac{1}{n_{\text{co}}^{(j)}} \boldsymbol{\chi}^{(j)} (\boldsymbol{\chi}^{(j)})^T. \quad (12)$$

The only exception to this is for the ‘scaling function’ of the needlet transform, which contains the  $\ell = 0$  spherical harmonic mode. For the scaling function, one can just use equation (4). Due to their simpler frequency dependence, we will assume that the foregrounds are restricted to a subset of these modes. On the other hand, we will assume that the HI signal and noise ( $s^{(j)}$ ) will be more evenly spread out throughout all the modes. This naturally divides the eigenmodes into  $n_{\text{fg}}$  modes which are dominated by the foregrounds and  $n_{\text{ch}} - n_{\text{fg}}$  modes which are dominated by the signal plus noise. Let  $\mathbf{P}$  be a matrix whose rows are eigenvectors of  $\hat{\mathbf{C}}^{(j)}$ . Moreover, let  $\mathbf{P}_f$  be an  $n_{\text{fg}} \times n_{\text{ch}}$  matrix containing the rows of  $\mathbf{P}$  corresponding to the largest  $n_{\text{fg}}$  modes of  $\hat{\mathbf{C}}^{(j)}$ . Next, let  $(\mathbf{P}^{-1})_f$  be an  $n_{\text{ch}} \times n_{\text{fg}}$  matrix whose columns are the columns of  $\mathbf{P}^{-1}$  which act on the foreground-dominated rows of  $\mathbf{P}$ . A foreground estimate can then be obtained via

$$\hat{\mathbf{f}}^{(j)} = (\mathbf{P}^{-1})_f \mathbf{P}_f \boldsymbol{\chi}^{(j)}. \quad (13)$$

The value for  $n_{\text{fg}}$  can be estimated in different ways. In Zhang et al. (2016), a likelihood ratio test was used. In such a method, one essentially increases the number of parameters used until the resulting power spectrum begins to converge (Zhang et al. 2016). Alternatively, one can use their models of the HI and noise, along with the Akaike Information Criterion (AIC), to estimate  $n_{\text{fg}}$ . This was the approach taken in Olivari et al. (2015). The AIC is given by

$$\text{AIC} = 2k - 2\ln(L), \quad (14)$$

where  $k$  is the number of parameters in the model and  $L$  is the maximum-likelihood value of the model. The likelihood function used in this paper can be found in the appendix of Olivari et al. (2015). Thus, by minimizing the AIC, one rewards goodness of fit (via the second term) while also discouraging overfitting (via the first term). More information on the use of AIC in a context such as this

can be found in Olivari et al. (2015). In this paper, the number of modes chosen in any given situation will always be selected using AIC unless otherwise noted.

### 3.3.2 Estimating foregrounds using DAYENU

Another possible approach to estimating the foregrounds is to use the DAYENU model from Ewall-Wice et al. (2020). DAYENU models the foreground covariance as simply as possible, using the fact that the foregrounds should be highly concentrated at delays close to  $\tau = 0$ . In particular, DAYENU will model the foreground covariance in delay space as

$$\tilde{\mathbf{C}}_{\text{FG}}^{-1}(\tau, \tau') = \epsilon^{-1} \frac{1}{2\tau_w} \delta(\tau - \tau') \quad |\tau| < \tau_w. \quad (15)$$

In this equation,  $2\tau_w$  is the assumed full width of the foregrounds in delay space and  $\epsilon^{-1}$  is their assumed magnitude.

At a high level, our DAYENU-based scheme approximates the foregrounds at some line of sight in needlet space using the following steps:

(i) Estimate the width of the foregrounds in delay space for the needlet coefficient of interest by comparing the delay transform to priors. The full width will correspond to  $2\tau_w$ .

(ii) Use this width to generate the foreground covariance model  $\mathbf{C}_{\text{FG}}^{-1}(\nu, \nu')$ . We would like to emphasize here that this foreground covariance is estimated one line of sight at a time. It is also not the same as the foreground covariance that will be eventually used in the KL step of NKL.

(iii) Project the needlet coefficient onto the image of  $\mathbf{C}_{\text{FG}}^{-1}(\nu, \nu')$ . This acts as a bandpass filter in delay space, selecting the foreground-dominated region.

(iv) Perform steps 1–3 at all coefficients in the needlet map. This will provide a foreground approximation  $\hat{\mathbf{f}}_k^{(j)}$  from which one can estimate  $\hat{\mathbf{C}}_{\text{FG}}^{(j)}$ .

This list has given a high level explanation of the DAYENU-based scheme. Let us now consider the details.

First, recall the DAYENU covariance definition provided in equation (15). In this formula, will use the value of  $\tau_w$  acquired in the previous steps and will set  $\epsilon = 1$  for now. We will account for the magnitude of the foregrounds later on in this work.

In frequency space, the DAYENU foreground covariance will be given by

$$\mathbf{C}_{\text{FG}}^{-1}(\nu, \nu') = \text{sinc}(2\pi\tau_w(\nu - \nu')). \quad (16)$$

When frequency samples are evenly spaced,  $\mathbf{C}_{\text{FG}}^{-1}(\nu, \nu')$  will be diagonalized by DPSS (Slepian 1978). Suppose one is dealing with sequences of length  $N_d$ . The DPSS sequences  $\{u^\alpha(N_d, W) | \alpha = 1, 2, 3, \dots, N\}$  will form an orthonormal basis that maximizes energy in a region  $[-W, W]$  of the Discrete Fourier Transform (DFT) domain. We call this ‘spectral concentration’ and denote it as  $s_c$ . More precisely, the spectral concentration for some sequence is

$$s_c = \frac{\int_{-W}^W |U(x)|^2 dx}{\int_{-1/2}^{1/2} |U(x)|^2 dx}. \quad (17)$$

In this formula,  $x$  is the variable of the DFT domain, which in our context is delay. The variable  $U$  represents the DFT of  $u$ , centred at  $(N_d - 1)/2$ .  $u^1$  is defined as a unit norm sequence that maximizes  $s_c$ . Then,  $u^2$  is created by finding a unit norm sequence that maximizes  $s_c$  while being orthogonal to  $u^1$ . In general,  $u^n$  is created by finding

a unit norm sequence that maximizes  $s_c$  while being orthogonal to  $u^i$  for  $i = 1, \dots, n - 1$ .

In the case of DAYENU, the DPSS sequences will maximize energy at delays  $-\tau_w < \tau < \tau_w$ .

The DAYENU model can then be used to estimate the foregrounds for each needlet coefficient in the following way. For the following steps, let  $\xi_k^{(j)}$  be a  $n_{\text{ch}} \times 1$  vector containing needlet coefficients for all channels at scale  $j$  and pixel  $k$ . The procedure to estimate the foregrounds at  $\chi_k^{(j)}$  is as follows.

(i) Use priors on the H I and noise to generate mock needlet coefficient maps. These mock maps will be denoted by  $\mathbf{h}^{m,(j)}$  and  $\mathbf{n}^{m,(j)}$  for the H I and noise, respectively. Moreover, let  $\mathbf{y}^{m,(j)} = \mathbf{h}^{m,(j)} + \mathbf{n}^{m,(j)}$ .

(ii) Randomly select  $N_p$  lines of sight from these mock coefficient maps. Then, apply two Blackman–Harris windows and a Fast Fourier Transform (FFT) to bring them into the delay domain. Moreover, the data should be shifted to place  $\tau = 0$  at the centre sample. We will denote such a transform with a tilde, that is,

$$\tilde{a} = \text{SHIFT} [\text{FFT} [aW^2]], \quad (18)$$

where  $W$  denotes a four term Blackman–Harris window of appropriate length. Let  $\mathcal{D}$  be a set containing the randomly chosen pixels.

(iii) For each  $i \in \mathcal{D}$ , perform a curve fit on  $\text{abs}(\tilde{y}_i^{m,(j)})$ . This curve fit is meant to give one estimate of the typical level of  $\text{abs}(\tilde{y}_i^{(j)})$  at each delay. It was found that using a third-order polynomial in log-space worked quite well. The curve produced by this fitting process at pixel  $i$  will be denoted  $r_i^{m,(j)}$ .

(iv) Average the fits obtained at each pixel to produce a typical H I plus noise curve in delay space:

$$\bar{r}^{(j)} = \frac{1}{N_p} \sum_{i \in \mathcal{D}} r_i^{m,(j)}. \quad (19)$$

In this formula,  $N_p$  is the number of pixels chosen and  $\mathcal{D}$  is the set of pixels used in the approximation.

(v) Perform a delay transform on the needlet coefficient of interest, giving  $\tilde{\chi}_k^{(j)}$ . Next, compare  $\text{abs}(\tilde{\chi}_k^{(j)})$  with  $\bar{r}^{(j)}$ . The delay at which the level of  $\text{abs}(\tilde{\chi}_k^{(j)})$  becomes comparable to that  $\bar{r}^{(j)}$  will correspond to the  $\tau_w$  used in DAYENU. For this paper, we select the delay at which  $\text{abs}(\tilde{\chi}_k^{(j)})$  is within a factor of 2 of  $\bar{r}^{(j)}$ .

(vi) Use the  $\tau_w$  obtained in the previous step to generate a DAYENU covariance as in equation (16). This DAYENU matrix will be denoted via  $\mathbf{C}^\nabla$ .

(vii) Compute the SVD of  $\mathbf{C}^\nabla$ , giving  $\mathbf{C}^\nabla = \mathbf{U}\mathbf{S}\mathbf{V}^\dagger$ . This SVD will be used to project onto the image of  $\mathbf{C}^\nabla$ . Let  $\bar{\mathbf{U}}$  be a matrix containing columns of  $\mathbf{U}$  for which the singular value is above some threshold. This threshold is determined based on the expected magnitude of the foregrounds. Since the foregrounds can be 5 orders of magnitude brighter than the signal, it is advisable to choose the cut-off to be at or below  $10^{-10} s_{\text{max}}$ , where  $s_{\text{max}}$  is the largest singular value of  $\mathbf{C}^\nabla$ . In this particular work, we use the same threshold for all angular scales and all needlet locations. The foregrounds are then approximated via

$$\hat{f}_k^{(j)} = \bar{\mathbf{U}}\bar{\mathbf{U}}^T \chi_k^{(j)}. \quad (20)$$

### 3.4 The NKL algorithm

In the literature, one finds standard foreground removal methods such as PCA, ICA, and GMCA being used in a way where one pair of  $\mathbf{A}$  and  $\mathbf{S}$  is generated for the entire data set. One exception to this in the literature is GNILC, which performs the cleaning in a more fine-grained way, sending the data to the needlet domain and

then cleaning each needlet coefficient individually. However, this method only takes advantage of frequency–frequency correlations in the data. Similarly to GNILC, NKL performs its cleaning in the needlet domain. The key difference, however, is that NKL also takes advantage of angular correlations in the needlet coefficients. Moreover, instead of using an ILC step, as in GNILC, NKL uses a KL transform. Below are step by step instructions for performing an NKL cleaning on some dataset.

(i) Generate an approximation of the foregrounds at the needlet scale of interest. In this paper, foreground approximations are generated using either DAYENU or PCA. This approximation will be denoted via  $\hat{\mathbf{f}}^{(j)}$ .

(ii) Divide the needlet coefficients at the scale of interest into  $N$  ‘chunks’ of equal size. These chunks must be adequately large to provide accurate covariance estimates. The appropriate chunk size will depend both on the severity of the foregrounds and on the needlet scale in question. This gives  $\chi^{(j)}$ , which was defined in Section 3.2,

(iii) Divide the foreground approximation into chunks, providing  $\hat{\mathbf{f}}^{(j)}$ . Use  $\hat{\mathbf{f}}^{(j)}$  to generate an estimate of the foreground covariance at the scale of interest. This covariance can be estimated in the manner shown in equation (4). This covariance will be denoted  $\hat{\mathbf{C}}_{\text{FG}}^{(j)}$  and will be a square matrix of size  $n_{\text{ch}}N$ .

(iv) Use priors on the signal to create an estimate of  $\mathbf{C}_S^{(j)}$ . This matrix will have the same shape as the foreground covariance matrix described in the previous step.

(v) Use the covariance matrices acquired in the previous two steps to perform a KL transform on the needlet scale of interest. The steps for this process are described in Section 2.1.

### 3.5 Deviations from the ideal case

The NKL technique is based on the KL transform, which is described in equation (5). There are two deviations from this ideal case that we will consider here. The first is inaccuracies in our estimates of  $\mathbf{C}_S$  and  $\mathbf{C}_{\text{FG}}$ . The signal covariance,  $\mathbf{C}_S$ , is derived from priors and thus may be sensitive to the model chosen. We tested this in one case by using priors generated by CORA on maps generated using CRIME. We do not present the results in this paper, but we found that there was a negligible difference in performance.

On the other hand,  $\mathbf{C}_{\text{FG}}$  is estimated from the data. This foreground approximation will inevitably be contaminated by some signal and noise. More precisely, for scale  $j$ , we have

$$\hat{\mathbf{f}}^{(j)} = \underline{\mathbf{f}}^{(j)} + \underline{\mathbf{h}}^{(j)} + \underline{\mathbf{n}}^{(j)}. \quad (21)$$

In this formula,  $\underline{\mathbf{h}}^{(j)}$  and  $\underline{\mathbf{n}}^{(j)}$  represent the H I and noise present in the foreground approximation, respectively. We include underlines on these variables to emphasize that these maps have been partitioned as described in Section 3.2. In this section, we will consider the effect that  $\underline{\mathbf{h}}^{(j)}$  and  $\underline{\mathbf{n}}^{(j)}$  have on the statistics of the cleaned coefficients.

Let us begin by considering the generalized eigenvalue problem of equation (5). For now, let us assume perfect knowledge of the foregrounds, noise, and signal. Moreover, let us consider only one eigenmode,  $\phi_i$ , which has a corresponding eigenvalue,  $\lambda_i$ . Lastly, let us consider an analysis conducted only at one needlet scale,  $j$ . We will drop any  $j$  subscripts since the analysis is all done at the same scale. This becomes the generalized eigenvalue problem

$$\mathbf{C}_{\text{FG}}\phi_i = \lambda_i \mathbf{C}_S\phi_i. \quad (22)$$

However, when performing NKL, we will use  $\hat{\mathbf{C}}_{\text{FG}}$ . We use the hat symbol to denote covariance estimates derived from the data. This

gives a new eigenvalue problem

$$\hat{\mathbf{C}}_{\text{FG}}\phi'_i = \lambda'_i \mathbf{C}_S \phi'_i, \quad (23)$$

where apostrophes represent perturbed versions of the variables seen in equation (22). This estimate of the foreground covariance will vary from the true version as

$$\hat{\mathbf{C}}_{\text{FG}} = \mathbf{C}_{\text{FG}} + \Delta_{\text{FG}} + \mathbf{C}_{h'h'} + \mathbf{C}_{f_h'} + \mathbf{C}_{n'h'} + \mathbf{C}_{n'n'} + \mathbf{C}_{f_n'}. \quad (24)$$

In this formula,  $\Delta_{\text{FG}}$  represents foreground errors that would be present if  $\hat{\mathbf{C}}_{\text{FG}}$  were acquired from the true foregrounds. The  $\mathbf{C}_{f_h'}$  term is given by

$$\mathbf{C} = \frac{1}{N_p} \sum_q (h'_q f_q^T + f_q h_q'^T). \quad (25)$$

In this formula,  $f_q$  is the  $q$ th column of  $\underline{\mathbf{f}} - \bar{\underline{\mathbf{f}}}$  and  $h'_q$  is defined as  $\underline{\mathbf{h}}' - \bar{\underline{\mathbf{h}}}'$ . Note that the overline here has the same meaning as in equation (4). The variable  $N_p = n_{\text{co}}^{(j)}/N$  is the number of columns in  $\underline{\mathbf{f}}'$  and  $\underline{\mathbf{g}}'$ . This is the same as the number of needlet coefficients per chunk per frequency channel. Note that the other terms in equation (24), for example,  $\mathbf{C}_{n'h'}$ , are defined in a similar way.

Each of these terms ought to be a source of error in our resulting estimate of the HI signal. However, the most concerning of these is  $\mathbf{C}_{f_h'}$ , which will lead to a negative bias in our estimate of the HI signal. In principle,  $\mathbf{C}_{n'h'}$  should also lead to correlations between the noise and HI after cleaning has been done. However, we did not observe this leading to any noticeable bias in the resulting power spectra. Thus, we will focus on the contribution of  $\mathbf{C}_{f_h'}$ .

Let  $\mathcal{F}'_i = \phi_i'^T f_p$  and  $\mathcal{H}'_i = \phi_i'^T h_p$ . We will find that  $\mathbf{C}_{f_h'}$  will lead to a non-zero correlation between  $\mathcal{H}'_i$  and  $\mathcal{F}'_i$ . It turns out that this bias will be more severe for modes with larger eigenvalues  $\lambda_i$ . This comes into the picture through the variable

$$\alpha_{ik} = E[\bar{\mathcal{H}}'_i \mathcal{H}'_k] \approx \begin{cases} 0 & \lambda_i < 1 \\ \delta_{ik} & \lambda_i > 1, \end{cases} \quad (26)$$

where  $E[\ ]$  refers to an expectation value. It will be the case that

$$E[\mathcal{F}'_i \mathcal{H}'_k] \approx \delta_{ik} \frac{1}{N_p} \sum_{j \neq i} \frac{\lambda_j \alpha_{ii} + \alpha_{jj} \lambda_i}{\lambda_i - \lambda_j}. \quad (27)$$

In this equation,  $\delta_{ik}$  is a Kronecker delta. A derivation of equation (27) can be found in Appendix A. We emphasize that this is a rough estimate, but it does provide some insight into the statistics of NKL. There are four important features about equation (27) that we would like to emphasize.

(i) Signal-dominated modes ( $\lambda_i < 1$ ) will be negatively biased, as they only pick up contributions from terms where  $\lambda_j > \lambda_i$ .

(ii) The  $\lambda_i - \lambda_j$  in the denominator ensures that the bias will be small for modes  $\lambda_i < 1$ . However, the bias will be worse for modes with  $\lambda_i$  closer to 1.

(iii) This bias scales like  $\frac{1}{N_p}$ , meaning that it will become less severe when larger chunks are used in the analysis.

(iv) The bias includes no dependence on the number of frequency channels used in the analysis. This is different from ILC methods in which the bias becomes more severe as more frequency channels are included.

A test of this equation will be presented in Section 5.4.

### 3.6 MGNILC, a corollary

One can also use the foregrounds approximation methods of Section 3.3 to modify the GNILC technique described in Olivari et al.

(2015). In GNILC, one generates an estimate for the HI signal plus noise by applying a filter to the data:

$$\hat{s} = \hat{\mathbf{S}}(\hat{\mathbf{S}}^T \hat{\mathbf{R}}^{-1} \hat{\mathbf{S}})^{-1} \hat{\mathbf{S}}^T \hat{\mathbf{R}}^{-1} x. \quad (28)$$

In this formula,  $x$  is the data vector (all frequency channels for a particular needlet coefficient),  $\hat{s}$  is an estimate of the HI plus noise for a particular line of sight,  $\hat{\mathbf{R}}$  is an estimate of the covariance of the data, and  $\hat{\mathbf{S}}$  is given by

$$\hat{\mathbf{S}} = \hat{\mathbf{R}}_{\text{HI}+N}^{\frac{1}{2}} \mathbf{U}_s. \quad (29)$$

Here, we have that  $\hat{\mathbf{R}}_{\text{HI}+N}$  is a prior of signal plus noise covariance, and  $\mathbf{U}_s$  is a matrix containing the subset of eigenvectors of  $\hat{\mathbf{R}}$  that are dominated by the signal plus noise. In this formula, raising a matrix to the 1/2 power refers to taking the Hermitian square root. When using an ILC technique such as GNILC, one must consider errors in  $\hat{\mathbf{R}}$ . In particular,  $\hat{\mathbf{R}}$  will have some error due to the finite sample size used in the estimation:

$$\hat{\mathbf{R}} = \mathbf{R} + \Delta. \quad (30)$$

In this formula,  $\mathbf{R}$  is the true covariance of the data and  $\Delta$  is an error term given by

$$\Delta = (\hat{\mathbf{R}}_{\text{HI}+N} - \mathbf{R}_{\text{HI}+N}) + (\hat{\mathbf{R}}_{\text{FG}} - \mathbf{R}_{\text{FG}}) + \hat{\mathbf{C}}. \quad (31)$$

In this formula, the first two terms correspond to errors in the foregrounds and HI +  $n$ , respectively. The third term corresponds to a spurious correlation that appears between the foregrounds and signal plus noise due to the finite sample size used. This is given by the equation

$$\hat{\mathbf{C}} = \frac{1}{N_p} \sum_q (s_q f_q^T + f_q s_q^T). \quad (32)$$

In this formula,  $N_p$  is the number of needlet coefficients in the neighbourhood being considered and  $q$  is a variable that indices lines of sight in the needlet coefficient neighbourhood. We also introduce the variable  $s_q = h_q + n_q$ .

Similarly to what was seen in NKL,  $\hat{\mathbf{C}}$  causes residual foregrounds to have a negative correlation with the HI plus noise. This leads to the negative ‘ILC bias’ described in Delabrouille et al. (2009). This bias scales like  $\frac{n_{\text{ch}}}{N_p}$ . Such a bias is troublesome at larger scales, for which  $N_p$  is restricted by the number of needlet coefficients available.

To mitigate this problem, we propose MGNILC. In this approach, we use a modified covariance estimate

$$\hat{\mathbf{R}}_m = \hat{\mathbf{R}}_{\text{FG}} + \hat{\mathbf{R}}_{\text{HI}+N}. \quad (33)$$

In this formula,  $\hat{\mathbf{R}}_{\text{FG}}$  is an estimate of the foreground covariance in the needlet neighbourhood of interest using the approach described in Section 3.3. In this case, our error term will look like

$$\hat{\mathbf{C}}_m = \frac{1}{N_p} (s'_q f_q^T + f_q s_q'^T). \quad (34)$$

In this formula,  $s'_q$  is the HI left over after performing the foreground approximation. Whether using PCA or DAYENU for the foreground approximation, we should find that  $s'_q$  is essentially a version of  $s_q$  that has been low-pass filtered in delay space. This reduces the ILC bias in two ways. For one, the fact that the HI plus noise has been low-pass filtered prevents the negative bias from affecting larger delays. Moreover, it also prevents any spurious correlations between high delay components of  $s$  with the low delay components of  $f$ , providing some additional mitigation of the ILC bias. However, the ILC bias problem will not be fixed completely as there will of course still be correlations between  $f_q$  and the low delay content of  $s_q$ .



## 4 SIMULATED MAPS FOR TESTING

This section provides a description of the simulated maps used for testing the various foreground removal techniques described earlier in this paper.

### 4.1 Cosmological signal

We generated simulated H I maps using the CORA package.<sup>1</sup> The CORA software assumes that the 21 cm signal is Gaussian and isotropic. For computational convenience, the covariance is estimated using the flat sky approximation. One then simulates the 21 cm fluctuations using equation C5 from Shaw et al. (2015). The mean H I temperature is then supplied using equation C4 from that same paper. The CORA software uses cosmological parameters from *Planck* 2018 (Planck Collaboration 2020). Moreover, we used the default CORA setting, which is to estimate  $\Omega_{\text{HI}}$  using the model given in Crighton et al. (2015). The CORA software package also incorporates redshift-space distortions, assuming a constant H I bias of  $b(z) = 1$  by default.

### 4.2 Foregrounds

In this paper, we consider only foregrounds from Galactic synchrotron radiation. For the synchrotron radiation, both unpolarized and polarized contributions are included. For the polarized foregrounds, 1 percent leakage of both the  $Q$  and  $U$  components was assumed. This value was chosen to be consistent with a typical level of polarization leakage for H I IM telescopes. These foregrounds are simulated using the CRIME software package (Alonso et al. 2014). CRIME models unpolarized Galactic synchrotron emission via

$$T(\hat{n}, \nu) = T_{\text{Haslam}}(\hat{n}) \left( \frac{\nu}{\nu_H} \right)^{\beta(\hat{n})} + \delta T. \quad (35)$$

In this formula,  $T_{\text{Haslam}}$  is the Haslam map temperature,  $\beta$  is a spectral index generated from the *Planck* sky model (Delabrouille et al. 2013), and  $\nu_H$  is 408 MHz, the frequency of the Haslam map. The  $\delta T$  term accounts for angular scales  $l \gtrsim 200$ , which are smaller than the Haslam map's resolution. This term is generated assuming that the foregrounds are Gaussian and isotropic at these scales. The power spectrum assumed can be found in Alonso et al. (2014).

The polarized synchrotron emission is modelled using the measured Faraday depth through the entire Milky Way presented in Oppermann et al. (2012). CRIME then assumes, among other things, that the number of emitting regions follows a Gaussian distribution as a function of Faraday depth.

CORA also can be used to model the polarized synchrotron radiation. The model used by CORA is essentially the same as that used by CRIME. The main theoretical difference is that CORA assumes that emitting regions at the same Faraday depth for a particular line of sight are independent. This leads to a slightly different dependence on the emission at Faraday depth  $\psi$  for a particular line of sight. The Faraday depth coherence lengths are also slightly different, being  $0.5 \text{ rad m}^{-2}$  for CRIME and  $1 \text{ rad m}^{-2}$  for CORA. The value for CRIME was chosen to provide results consistent with the Hammurabi simulation package (Jaffe et al. 2012), while CORA's value was chosen to be consistent with observations at 1.4 GHz (Wolleben et al. 2006). In practice, we found that the foregrounds provided by CRIME appeared to be more chromatic than those from CORA. Thus, to be conservative, we used CRIME to simulate the foregrounds.

<sup>1</sup><https://github.com/radiocosmology/cora>

**Table 1.** Parameters describing the hypothetical telescope used in this paper.

Parameter	Value
D	13.5 m
$T_{\text{inst}}$	20 K
$f_{\text{sky}}$	1
$t_{\text{obs}}$	40 000 h
$N_{\text{dishes}}$	64
$n_{\text{ch}}$	256
$[\nu_{\text{min}}, \nu_{\text{max}}]$	[980, 1080 MHz] and [400, 500 MHz].
$\Delta \nu$	0.390625 MHz

### 4.3 Instrumental effects and noise

In this paper, tests of various foreground removal techniques are performed assuming an instrument similar to the MeerKAT array operating in single-dish mode. The parameters of this telescope are described in Table 1. One may notice that the observing time of 40 000 h is quite large. This value is not particularly realistic, but was chosen to ensure that the noise would not overpower the H I signal. In future telescopes such as the Packed Ultra-wideband Mapping Array (PUMA), adequate noise levels can be achieved with larger numbers of antennas and less integration time. The value  $f_{\text{sky}} = 1$  was chosen since we are imagining an instrument that observes the whole sky. Any masking is then applied after observations have taken place.

This hypothetical instrument was chosen for realism and simplicity. Since it consists of an array of dishes operating in single-dish mode, we can model the beams by convolving with a Gaussian. This is much easier than trying to derive maps from simulated visibilities. Moreover, the relatively low angular resolution of the instrument ( $\approx 1^\circ$ ) ensures that large values of  $\ell$  will not be required. This relaxes computational requirements when working with needlets. We approximate beam effects in a way that is identical to Carucci et al. (2020). In particular, we convolve the simulated maps with a frequency-dependent Gaussian beam. We then reconvolve all maps to give them the same angular resolution.

We added noise to the simulated maps in the same way as Carucci. In particular, the noise per pixel follows a Gaussian distribution with a standard deviation given by

$$\sigma_N(\nu) = T_{\text{sys}}(\nu) \sqrt{\frac{4\pi f_{\text{sky}}}{\Delta \nu t_{\text{obs}} N_{\text{dishes}} \Omega_{\text{beam}}}}. \quad (36)$$

In this formula,  $T_{\text{sys}}$  is the system temperature,  $f_{\text{sky}}$  is the fraction of the sky observed in the survey,  $\Delta \nu$  is the frequency resolution,  $t_{\text{obs}}$  is observing time, and  $\Omega_{\text{beam}}$  is the beam solid angle. Similarly, we estimate the system temperature in the same way as Carucci, assuming

$$T_{\text{sys}}(\nu) = T_{\text{instr}}[\text{K}] + 66 \left( \frac{\nu}{300[\text{MHz}]} \right)^{2.55}. \quad (37)$$

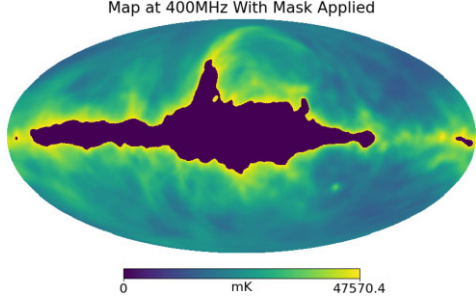
Once noise and Gaussian beam effects are accounted for, the maps used in testing can be described schematically via the equation

$$\mathbf{X} = (B_{\text{low}} - B) * (B * (\mathbf{f} + \mathbf{h}) + \mathbf{n}). \quad (38)$$

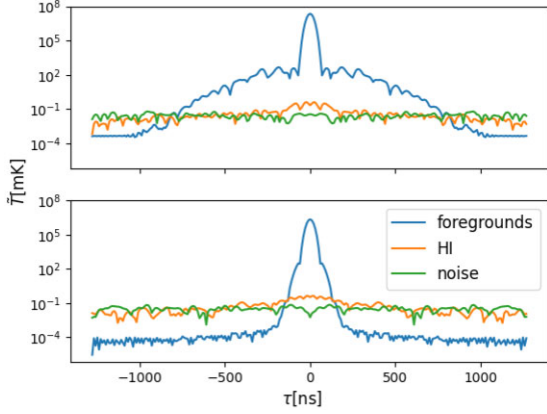
In this schematic,  $B*$  represents convolution with the beam and  $B_{\text{low}}$  is the beam at lowest frequency channel in the band.

## 5 TESTS OF METHODS

In this section, we test the performance of various foreground removal methods and compare the results. In particular, we test GMCA, PCA, ICA, GNILC, and NKL. It should be noted that the severity of



**Figure 2.** An illustration of the mask used for tests of the foreground removal methods in the high-redshift case. The mask has been applied to this map, which includes simulated foregrounds, signal and noise.



**Figure 3.** The top plot shows delay spectra at the brightest pixel in the unmasked map in the higher redshift case. The bottom plot shows delay spectra at the brightest pixel remaining after the mask has been applied.

polarized foregrounds is expected to vary as a function of redshift. In particular, the chromaticity of polarized foregrounds will become more severe at lower frequencies due to Faraday rotation. Therefore, we conduct tests in two different redshift regimes corresponding to low- and high-frequency bands of MeerKAT.

### 5.1 Additional map preparation steps

After performing the steps described in Section 4.3, we took some additional steps to aid the cleaning process. These were done both to exclude highly contaminated pixels and to lower computation times.

As expected, polarized foreground chromaticity is quite severe in the Galactic plane. We applied a mask to the data to exclude the brightest 15 percent of pixels. Such masks were created for both the low- and high- $z$  test cases. The mask created for the high- $z$  test case is shown in Fig. 2. The chromaticity problem is illustrated in Fig. 3. The top plot in the figure shows delay spectra for the brightest line of sight in the high- $z$  test case. Note here that the foregrounds dominate for almost all delays. The bottom plot shows delay spectra at the brightest unmasked line of sight. In this case, we find that the foregrounds occupy a relatively small region of delay space, allowing for more effective cleaning.

It was noted that the Gaussian half-power beamwidths in both the low- and high- $z$  test cases are greater than  $1^\circ$ . As such, we truncated maps in both test cases at  $\ell = 255$  for computational simplicity. Moreover, we performed needlelet transformations using  $B = 2$  with scales running from  $j = 4$  up to 8. We provide details of how the data are represented in the needlelet domain in Table 2.

**Table 2.** A summary of the needlelet coefficients in these tests. We use the same scales,  $l$  ranges, and  $n_{\text{side}}$  values for both the low- and higher- $z$  tests.

Scale	$l$ range	$n_{\text{side}}$
Scaling function	[0,16]	16
$j = 4$	[8,32]	32
$j = 5$	[16,64]	64
$j = 6$	[32,128]	128
$j = 7$	[64,256]	256
$j = 8$	[128,512]	512

It is worth discussing here the impacts that  $B$  and the range of  $j$  can have on one’s analysis. Increasing the value of  $B$  will decrease the spread of the needlelet coefficients in real space, while increasing their spread in harmonic space. So, choosing a value of  $B$  greater than 2 would in principle allow for finer chunking of the maps but would also lead to coarser partitioning in  $\ell$ -space.

Next, let us consider the range of  $j$  chosen. Since the maps were truncated at  $\ell = 255$ , scales with  $j > 8$  would contain no information and thus were not included. The minimum scale value  $j_{\text{min}} = 4$  was chosen to ensure that the scaling function would have enough statistics to be cleanable by NKL.

The value of  $B$  and range of  $j$  chosen for this paper do work. However, other values were not tested, so it is unknown whether these are the optimal choices one could make.

### 5.2 Implementation of cleaning techniques

When cleaning with blind methods like PCA, ICA, and GMCA, we use priors on the H I signal and noise to select the appropriate number of components to be removed. In particular, we generated additional sets of H I maps using CORA and noise maps using equation (36). We then applied the beam convolution described in Section 4.3 and produced covariance estimates from these processed maps using equation (12). In this paper, we chose to use  $N_{\text{maps}} = 10$  mock maps for estimating the covariances. We then used our models of the H I and noise covariances to estimate the appropriate number of modes to remove using the AIC prescription described in Olivari et al. (2015).

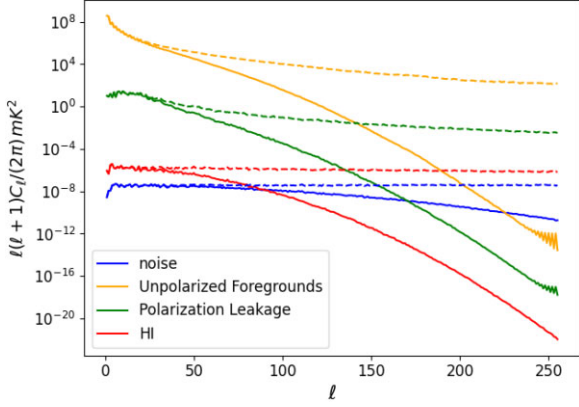
For the cases of GNILC, MGNILC, and NKL, we used the same sets of mock H I and noise maps as above to estimate the needlelet coefficient covariances.

For GNILC, we selected the appropriate number of coefficients using the Olivari AIC prescription. We also chose to use windows with a size of at least  $10^5$  coefficients. This seemed to be the window size required to minimize the ILC bias described in Delabrouille et al. (2009). There were of course exceptions to this rule for scales up to  $j = 6$ , since needlelet maps at lower  $j$  did not have enough coefficients. These windows were generated simply by choosing the closest  $N$  coefficients to the coefficient of interest.

### 5.3 Power spectrum estimation

In order to estimate the power spectrum, we used the method described in Liu, Zhang & Parsons (2016). In this approach, one computes Bessel-spherical harmonic modes of the sky

$$T_{\ell m}(k) = \sqrt{\frac{2}{\pi}} \int d\Omega dr r^2 j_\ell(kr) Y_{\ell m}^*(\hat{n}) \phi(\mathbf{r}) T(\mathbf{r}), \quad (39)$$



**Figure 4.** Angular power spectra estimated from unmasked maps using HEALPY at 450 MHz. Solid lines correspond to maps which have undergone beam convolution. Dashed lines correspond to maps that have not undergone any convolution steps.

where  $\phi(\mathbf{r})$  is a window function representing the survey volume. One then computes a ‘spherical harmonic power spectrum’

$$S_\ell(k) = 2\pi^2 \left[ \int d^3r \phi(\mathbf{r})^2 j_\ell(kr)^2 \right]^{-1} \frac{\sum_m |T_{\ell m}(k)|^2}{2\ell + 1}. \quad (40)$$

This spherical harmonic power spectrum is a spherical analogue to the commonly seen cylindrical power spectrum. It is also a useful tool in that it allows for different  $\ell$  modes to be checked individually. Then, in the case of a translation-invariant sky, one can form the power spectrum estimator

$$\hat{P}(k) = \sum_\ell w_\ell S_\ell(k). \quad (41)$$

In this formula, the  $w_\ell$  are weights which depend on the survey volume used. These weights account for the fact that sensitivity to a certain  $k$  mode may vary with angular scale  $\ell$ . In this case of a translation-invariant sky,  $\langle \hat{P}(k) \rangle = P(k)$ . It should be mentioned that the maps presented in this paper do not strictly satisfy translation invariance. Foregrounds vary significantly with line of sight and the HI is subject to redshift-space distortions and cosmic evolution. Even the noise breaks translation invariance, as its amplitude varies with frequency. Despite this, we still present  $\hat{P}(k)$  as a ‘power spectrum’, since it provides us with a weighted average of fluctuations at length-scale  $k$ , even when translation invariance is broken.

#### 5.4 Test at higher redshifts

For this test, we used the hypothetical instrument described in Table 1 and assumed it to have 256 evenly spaced frequency channels running from 400 to 500 MHz.

In Fig. 4, we provide angular power spectra for the 450 MHz frequency channel used in this test. The maps used to produce these power spectra have not been masked. We chose to use unmasked maps here in order to avoid edge effects at the mask boundaries. In this case, the beam convolution causes significant loss of signal above  $\ell \approx 50$ . Due to this, any power spectra produced by our instrument will be missing significant amounts of small length-scale information. For this reason, we will evaluate the effectiveness of the cleaning techniques based on how well they recover the power spectra of convolved maps. It should also be noted that the noise suffers less loss at  $\ell > 50$  than the other components. This is due to the fact that the foreground and HI have been convolved with a beam twice while the noise has only been convolved once. This causes the

**Table 3.** A summary of the NKL cleaning parameters used in the high- $z$  test.

Scale	Number of chunks	SNR
Scaling function	3	1
$j = 4$	12	4
$j = 5$	32	4
$j = 6$	64	4
$j = 7$	96	4
$j = 8$	96	4

noise to be the dominant component of the maps at those smaller angular scales.

For the case of NKL, the chunk size, and desired signal-to-noise ratio (SNR) of the KL cleaning vary with needlet scale. We summarize this in Table 3. The values shown in this table were chosen through trial and error. Ideally, one would like to clean with as many chunks as possible, since this would provide the best statistics. However, we are limited by two factors. For one, making the chunks too small leads to an inaccurate  $\hat{C}_{\text{FG}}$  and incomplete cleaning of the foregrounds. We also found that using small chunks also leads to inaccuracies in  $C_{\text{HI}}$ . However, this is not a fundamental issue and would be fixed by simply using more mock maps in our estimates. Trying to keep our models accurate is what motivated our choices for the number of chunks for scales up through  $j = 6$ . For scales  $j = 7$  and 8, we used 96 chunks as this was the largest value for which our computer was able to compute eigenvectors.

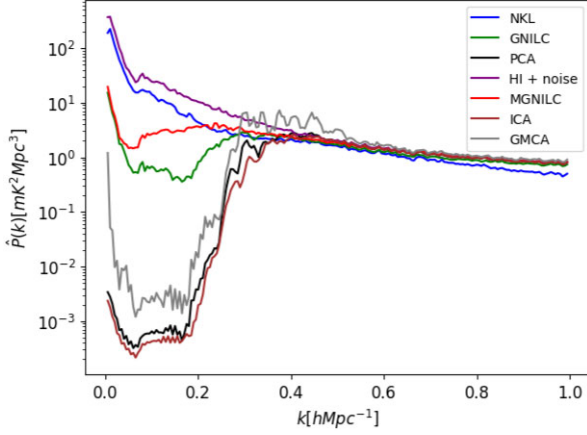
In practice, an analyst using NKL would probably decide on a chunking scheme by trial and error using simulations. However, it may be worth providing some intuition on how these parameters may depend on the survey in question. We should emphasize here that this intuition is really guess work. More precise characterization will need to be determined by trying out NKL in different scenarios.

Suppose one decreased the resolution of one’s instrument. This would lead to the angular power spectra like those shown in Fig. 4 dropping off more quickly as a function of  $\ell$ . For adequately large (i.e. small  $j$ ) needlet scales, the coefficients would not change much and our approach would stay more or less the same. For smaller needlet scales, noise contamination would become more severe and one may find that larger chunks are required to control errors in  $\hat{C}_{\text{FG}}$  due to noise leakage.

Now, suppose that one decreased the survey area. Naturally, one would mask out adequately small needlet coefficients and decrease the number of chunks accordingly. It is possible that one could decrease the size of chunks used, as the small and hard to model correlations of distant chunks is now gone. However, this could still pose problems as we expect the anticorrelation between residual foregrounds and signal to scale like  $1/N_p$ .

In the case of an especially small survey area, one may also find that there are not enough non-zero coefficients to adequately estimate  $\hat{C}_{\text{FG}}$  at lower values of  $j$  and in the scaling function. In such a case, one would likely have to increase either the minimum value of  $j$  or the value  $B$ , increasing the number of coefficients per map but providing a coarser partitioning of  $\ell$ -space.

We found for this high- $z$  case that approximating the foregrounds with DAYENU worked better than with PCA. We believe DAYENU’s superior performance in this scenario is likely due to the fact that it treats each pixel in the needlet map individually. On the other hand, PCA gives each pixel the same treatment. This fine-grained approach is useful at higher redshifts where foreground chromaticity varies quite significantly with line of sight.



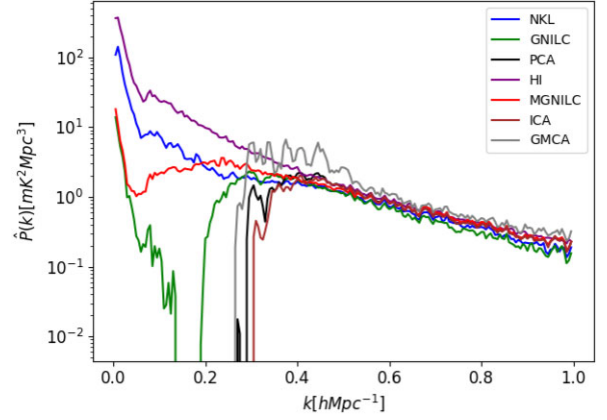
**Figure 5.** Power spectra estimated from the cleaned high-redshift maps. The ‘HI + noise’ curve shows the power spectrum of the masked, beam-convolved maps. NKL was tested on a second realization of signal plus noise, producing similar results.

For the DAYENU foreground approximation, we chose the SVD cut-off to be a factor of  $10^{-14}$  below the largest singular value of  $\mathbf{C}^{\top}$ . This value ensured that all foregrounds were captured in the approximation.

The MGNILC cleaning was performed using needlet neighbourhoods of size 1500 coefficients. MGNILC was also performed using the same foreground approximation as for NKL.

For all blind methods, 78 components were used to model the foregrounds. This value was selected by AIC.

In Fig. 5, we present the 3D power spectra recovered by the methods described above using equation (41). One may note that these results look more pessimistic than what is seen in other papers on this topic. This is because we are including polarized foregrounds at a higher redshift, which is uncommon. Moreover, most of the sky is kept, including areas with relatively bright foregrounds. One will also notice that all methods tested behaved similarly at small spatial scales ( $k \gtrsim 0.6 \text{ hMpc}^{-1}$ ). One exception to this is the NKL curve, which is lower than the others at such scales. This is due to the fact that NKL is the only method tested here that makes a distinction between HI and noise. These small scales carry noise comparable to or greater than the HI signal. As such, NKL looks different than the other methods in this regime. It should be noted that this feature goes away when we set  $\mathbf{C}_S = \mathbf{C}_{\text{HI}} + \mathbf{C}_N$ . This feature is also not a problem, as the debiased data will show that the HI is still well preserved by NKL at these small scales. At larger spatial scales, there is a clear bifurcation between the blind and non-blind methods. The blind methods clean extremely aggressively at such scales, leaving residuals up to 5 orders of magnitude below the signal plus noise power spectrum. This makes sense, as these large scales will correspond to the foreground-dominated delays such as those seen in Fig. 3. There is also a notable trough for certain techniques below  $k \approx 0.20 \text{ hMpc}^{-1}$ . These scales should correspond roughly to the regime in delay space shown in Fig. 3 where the foregrounds and HI plus noise are approaching the same magnitude. On the other hand, the non-blind methods seem to be able to preserve some information from within this foreground-dominated region of delay space, having residuals several orders of magnitude closer to that of the true signal plus noise power spectrum. Moreover, NKL seems to provide roughly a factor of 10–50 improvement over GNILC at these larger scales. However, it should be noted that all methods incur significant signal loss at the large spatial scales. Even NKL is



**Figure 6.** Power spectrum estimates which have been debiased. The ‘HI’ curve provides the power spectrum of the HI maps after beam convolution and masking.

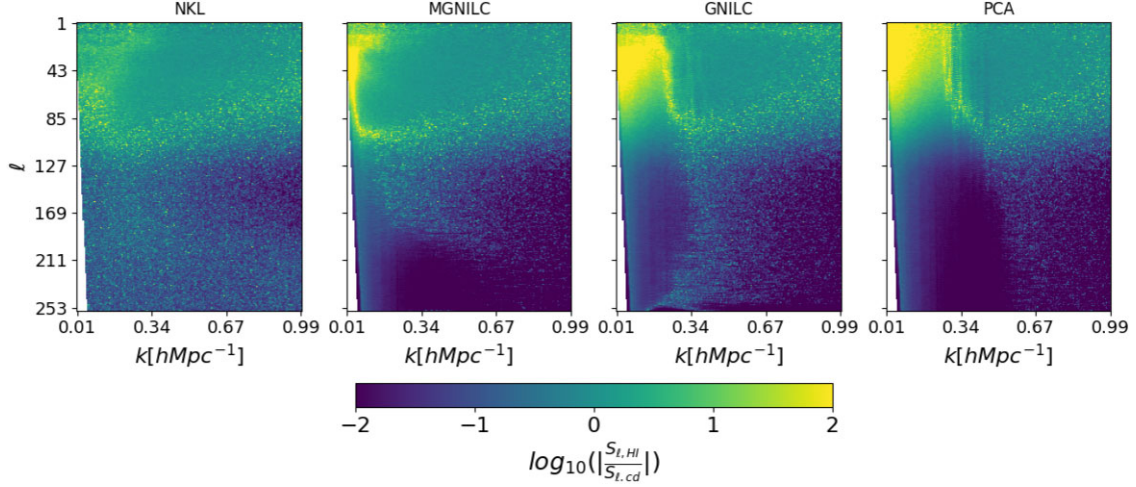
a factor of a few below the desired level. This is perhaps something that could be compensated for through the use of a transfer function, but we leave such an analysis to future work.

In Fig. 6, we present power spectrum curves which have been debiased to remove the noise. The power spectra are debiased according to the equation

$$\hat{P}_{\text{debiased}}(k) = \hat{P}(k) - \hat{P}_N(k). \quad (42)$$

In this formula,  $\hat{P}(k)$  is the power spectrum estimate obtained from the cleaned maps and  $\hat{P}_N(k)$  is an estimate of the noise power spectrum obtained from our model of the noise. Note that NKL was debiased slightly differently, with the noise model used being one taking into account the effects of the KL transform on the noise. There are a few features in this plot worth noting. First, it should be noted that the debiased power spectra for GNILC, ICA, GMCA, and PCA become negative at lower  $k$ -values. This is due to the loss in both the signal and noise incurred at large scales by these methods, combined with the fact that the noise power spectrum estimate  $\hat{P}_N$  does not account for this loss. It should be noted that these power spectrum values would likely not become negative had we accounted for this loss. However, even with a more sophisticated debiasing, one would still find these methods being outperformed by MGNILC and NKL. The MGNILC power spectrum estimate never becomes negative. This is because it suffers from less signal loss than the previously mentioned methods. NKL was debiased somewhat differently than the other methods, with the noise power spectrum being estimated from maps cleaned with NKL. It was found that NKL increased the power of the noise at large length-scales  $k \lesssim 0.35 \text{ hMpc}^{-1}$ , and decreased the noise power at scales  $k \gtrsim 0.35 \text{ hMpc}^{-1}$ . The use of a modified noise power spectrum resolved these issues in the case of NKL. Moreover, GNILC seems to underestimate the HI power spectrum at higher values of  $k$ . This is due to the ILC bias described in Delabrouille et al. (2009).

In Fig. 7, we present graphs illustrating the ratio of  $|S_{l, \text{HI}}|$  to  $|S_{l, \text{cd}}|$  as a function of  $k$ . The variable  $S_{l, \text{HI}}$  represents the spherical harmonic power spectrum of the HI plus noise. On the other hand,  $S_{l, \text{cd}}$  represents the spherical harmonic power spectrum obtained from cleaned maps. Note that we also debias  $S_{l, \text{cd}}$ . In this figure, one will notice that there are common features between techniques. First, one will notice larger values of  $\log_{10}(|\frac{S_{l, \text{HI}}}{S_{l, \text{cd}}}|)$  in the top left corner of each plot. This region corresponds to large angular scales and low delays. Since such modes carry the most severe foreground



**Figure 7.** The log-scaled absolute value ratio of the true high- $z$  spherical harmonic power spectrum ( $S_{l, \text{HI}}$ ) over the spectrum recovered by cleaning and debiasing ( $S_{l, \text{cd}}$ ).

contamination, one will find them being cleaned most aggressively. This effect is most severe for PCA and least severe for NKL. Next, one will notice the plots becoming darker at higher  $\ell$ . This is due to the fact that the beam convolution applied causes the noise to overpower the HI at small angular scales. The strength of the noise relative to the HI at these scales leads to inaccuracy in the debiasing process. Overall, it appears that NKL has the lowest errors at large spatial scales, as expected from Fig. 5.

It is also important to consider residual foregrounds and any bias that these residuals may add to the estimated the HI signal. We will investigate these foreground residuals for GNILC, MGNILC, NKL, and PCA. Each of these methods cleans the data by first producing a matrix (or matrices) used for cleaning, followed by application of the matrix (matrices) to the data. For instance, when using PCA, one generates a cleaning matrix  $\mathbf{R}$  from a covariance estimate of the data, then cleans the data as  $\mathbf{d}_{\text{cleaned}} = \mathbf{R}(\mathbf{d} - \bar{\mathbf{d}})$ . In order to produce foreground residual estimates, we apply these cleaning matrices to data containing only foregrounds. So, for instance, the foreground residuals for PCA will be

$$\mathbf{f}_{\text{residual}} = \mathbf{R}(\mathbf{f} - \bar{\mathbf{f}}). \quad (43)$$

A similar approach can be taken to obtain the foreground residuals for GNILC, MGNILC, and NKL. We also produce an estimate of the HI plus noise present in the cleaned maps in a similar way. In the case of PCA, we would have

$$(\mathbf{h} + \mathbf{n})_{\text{cleaned}} = \mathbf{R}((\mathbf{h} - \bar{\mathbf{h}}) + (\mathbf{n} - \bar{\mathbf{n}})). \quad (44)$$

Fig. 8 illustrates the effect that these foreground residuals have on the resulting power spectra. First, notice that the dotted lines, representing the power spectra of the signal plus noise present in the clean maps, tend to be higher than the power spectra of the cleaned maps for the NKL, GNILC, and MGNILC cases. This is indicative of the presence of a negative bias in all three techniques. Moreover, this feature implies that the negative bias present in these techniques decreases the power spectrum estimates, worsening agreement with the true HI power spectrum. The negative bias seen in GNILC and MGNILC is predicted by the ILC bias described in Delabrouille et al. (2009). The negative bias in NKL was of course predicted in the analysis performed in Section 3.5.

We also explicitly tested equation (27) on the results of NKL at needlet scale  $j = 4$ . In this case, we have  $N_p = 1024$ . The results of

this are shown in Fig. 9. For this figure, we estimate  $E[\mathcal{F}'_i \mathcal{H}'_i]$  by averaging the KL coefficients across all columns of  $\mathbf{h}$  and  $\mathbf{f}$ . So, we have

$$\hat{E}[\mathcal{F}'_i \mathcal{H}'_i] = \frac{1}{N_p} (\phi_i^T \mathbf{h}) (\mathbf{f}^T \phi_i). \quad (45)$$

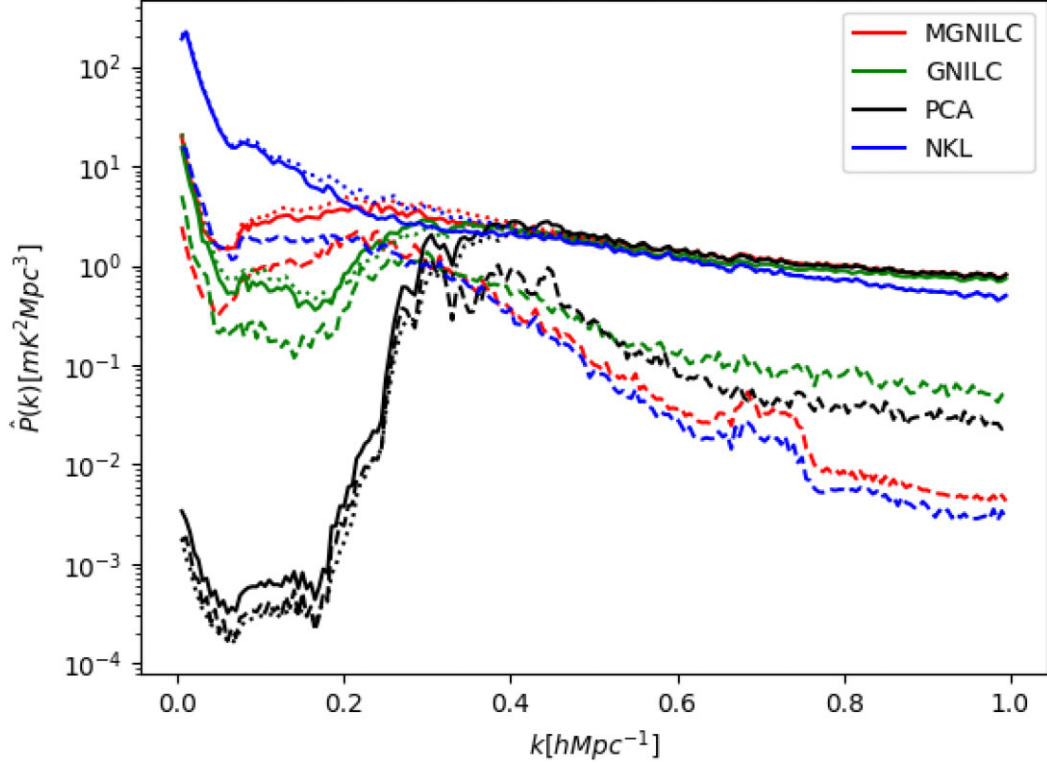
Since we are less interested in foreground-dominated modes, we only present results up to  $\lambda = 2$  for clarity. In this plot, one can see that equation (27) underestimates the severity of the bias, but is able to roughly capture what is happening. In particular, it predicts that the bias is negative and that it becomes more negative as it one approaches  $\lambda' = 1$ , the eigenvalue at which the foregrounds and signal plus noise are expected to have equal variance.

Another aspect of equation (27) to consider is the claim that  $E[\mathcal{H}'_i \mathcal{F}'_k] = 0$  for  $i \neq k$ . We present a test of this in Fig. 10. In this figure, the curve shows estimated values of  $E[\mathcal{H}'_i \mathcal{F}'_k]$  for  $\lambda'_i = 0.15$ . We see that this correlation does not strictly come out to be 0 as predicted by equation (27). Rather, it seems to oscillate about 0, with 0 correlation being outside the error bars of many eigenvalues. In this case, the average of these off-diagonal terms (up to  $\lambda' = 0.25$ ) was slightly negative (order  $10^{-3}$  and within a standard deviation of 0). This is much smaller than the diagonal terms  $E[\mathcal{H}'_i \mathcal{H}'_i]$ , which are of order 1. As such, these off-diagonal terms will increase errors in the power spectrum estimates, but will not contribute much of a net bias.

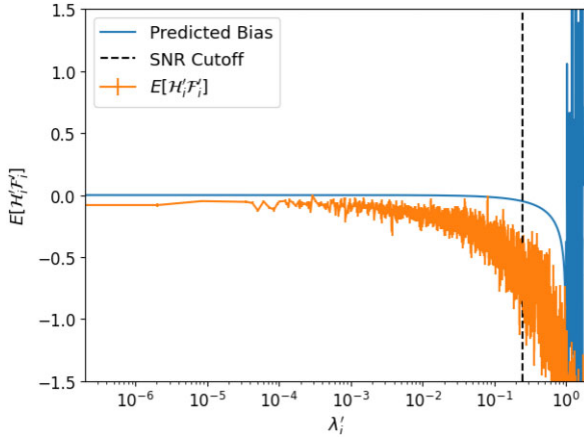
### 5.5 Test at lower redshifts

In this subsection, we test these same foreground removal methods at lower redshifts than in Section 5.4. In particular, we used 256 evenly spaced frequency channels from 980 up to 1080 MHz. This test was motivated by the fact that the characteristics of the foregrounds change with redshift. In particular, models predict that polarized foregrounds ought to be less severe in this regime. Note that we have again chosen to mask the brightest 15 percent of pixels.

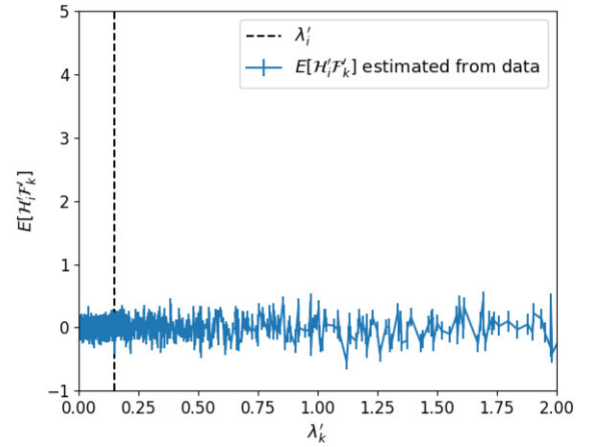
In Fig. 11, we present the angular power spectra estimated at the 1030 MHz frequency channel. In this case, for the convolved maps, the HI has a much stronger contribution to the maps than the noise for scales up to  $\ell \approx 200$ . Although not as severe as in the higher



**Figure 8.** Power spectra illustrating the effects of foreground residuals in the high-redshift case. Solid lines correspond to the power spectra of the cleaned maps. Dashed lines correspond to spectra obtained from foreground residuals. Dotted lines correspond to the power spectrum of the H I plus noise present in the cleaned maps.



**Figure 9.** A test of equation (27), which modelled the negative bias incurred by NKL, performed on the  $j = 4$  needlet scale.



**Figure 10.** Estimated values of  $E[\mathcal{H}_i^j \mathcal{F}_k^j]$  for  $\lambda_i^j = 0.15$ .

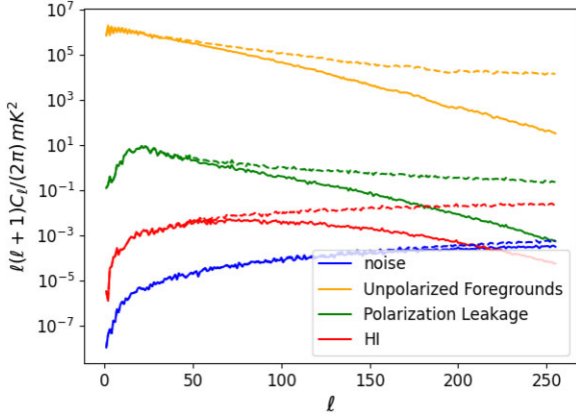
redshift case, the angular resolution of the antenna beams have once again caused significant signal loss at the smaller angular scales.

For the blind methods, we tried to use AIC to estimate the optimal number of modes to remove. In particular, AIC predicted four modes as the best option. However, preliminary checks using the radial power spectrum estimation code from Carucci<sup>2</sup> showed a large spike in the radial power spectrum at low  $k_\nu$ . This spike disappeared when the number of modes was increased to 5, resulting in better agreement between the radial power spectra of the H I plus noise and the cleaned

maps. As a result, we chose to remove five modes for all blind methods in this test case.

As for GNILC, NKL, and MGNILC, we once again used the needlet domain scheme described in Table 2. In this case, we again only compute needlets for modes up to  $\ell = 255$ , as the beams used in this regime have widths  $\theta \approx 1^\circ$ . GNILC was performed in the same way as in Section 5.4, where we sought to use windows of size  $10^5$  coefficients. For MGNILC and GNILC, we used neighbourhoods of size 3000 coefficients as AIC struggled to select the correct number of modes when using neighbourhoods of size 1500. For NKL, we describe the cleaning parameters in Table 4. The number of chunks used has remained more or less the same as in Section 5.4. We have

<sup>2</sup><https://github.com/isab3lla/gmca4im>



**Figure 11.** Angular power spectra estimated from unmasked maps using HEALPY at 1030 MHz. Solid lines correspond to maps that have undergone beam convolution. Dashed lines correspond to maps that have undergone no beam convolution.

**Table 4.** A summary of the number of chunks and SNR values chosen for each scale in the low- $z$  test.

Scale	Number of chunks	SNR
Scaling function	6	1
$j = 4$	12	1
$j = 5$	32	1
$j = 6$	64	1
$j = 7$	96	1
$j = 8$	96	1

however, increased the number of chunks for the scaling function and the  $j = 4$  scale. In addition, note that the same SNR is used for all needlet scales. For this particular test, we found that approximating the foregrounds using the PCA approach was more effective than using DAYENU. That is what we will present in this section. It seems likely that PCA provides an advantage over DAYENU in this scenario for two reasons. For one, the foregrounds in this scenario were quite narrow in delay space. The delay spectra of these foregrounds were particularly narrow compared to those of the higher redshift test case. The fact that the foregrounds vary so rapidly at delays close to 0 lead to our DAYENU-based scheme struggling to accurately estimate the width of the foregrounds. In particular, this approach was usually overestimating the width of the foregrounds in delay space. On the other hand, the fine-grained nature of the DAYENU-based approach was less important here than in the higher redshift case. This is due to the fact that foreground chromaticity did not vary as much as function of the line of sight in the lower redshift test case. It should be noted however, that AIC struggled with selecting the correct number of modes during the foreground approximation step. We fixed this simply by choosing to remove  $m + 2$  modes, where  $m$  was the number of modes selected by AIC. It should also be noted that we chose  $\mathbf{C}_S = \mathbf{C}_{H1} + \mathbf{C}_N$  for this particular test case. This was done since our computer program interpreted  $\mathbf{C}_{H1}$  as being singular, preventing the KL transform from being performed. Tacking on  $\mathbf{C}_N$  was a convenient workaround since the noise is very small in this test and the computations for the  $\mathbf{C}_S = \mathbf{C}_N + \mathbf{C}_{H1}$  test had already been performed.

In Fig. 12, we present power spectra estimated from cleaned maps. It should be noted that the noise is quite low in this test case. As such, the debiasing process of equation (42) makes an imperceptible

difference. Because of this, we only present the debiased results. Note that all methods in this figure were debiased using the same noise power spectrum estimate  $\hat{P}_N(k)$ . In this figure, one will notice first that GMCA and PCA produce essentially identical results, and that they seem to provide the best match to the H I plus noise power spectrum. However, this is somewhat misleading, as the foreground residuals at low  $k$  roughly match the H I plus noise power spectrum. This is made clear in Fig. 13, where one can see that the power spectrum at these scales is mostly due to foreground residuals. The GNILC and MGNILC power spectra also receive a boost from their foreground residuals, but the effect is not as dramatic as in the case of GMCA and PCA. It also appears that the ILC bias in this case is not as severe as in the higher redshift case. This is likely due to the fact that the larger neighbourhoods of needlet space were used. As in the higher redshift test case, we find that the presence of foreground residuals decreases the NKL power spectrum at low  $k$ . However, the effect on NKL is smaller in this case compared to the one at higher redshift.

Interestingly, ICA has performed differently from GMCA and PCA in this context. It appears that ICA has smaller foreground residuals than either GMCA or PCA in this case. Next, one will notice that ICA, MGNILC, GNILC, and NKL give very similar results down to  $k \approx 0.15 \text{ hMpc}^{-1}$ . For  $k < 0.15 \text{ hMpc}^{-1}$ , the non-blind methods once again provide significantly improved results. For scales  $0.03 \text{ hMpc}^{-1} < k < 0.15 \text{ hMpc}^{-1}$ , we find that NKL, MGNILC, and GNILC provide similar results. At the very lowest scale ( $k < 0.03 \text{ hMpc}^{-1}$ ), we find NKL provides a factor of a few improvement over GNILC and MGNILC.

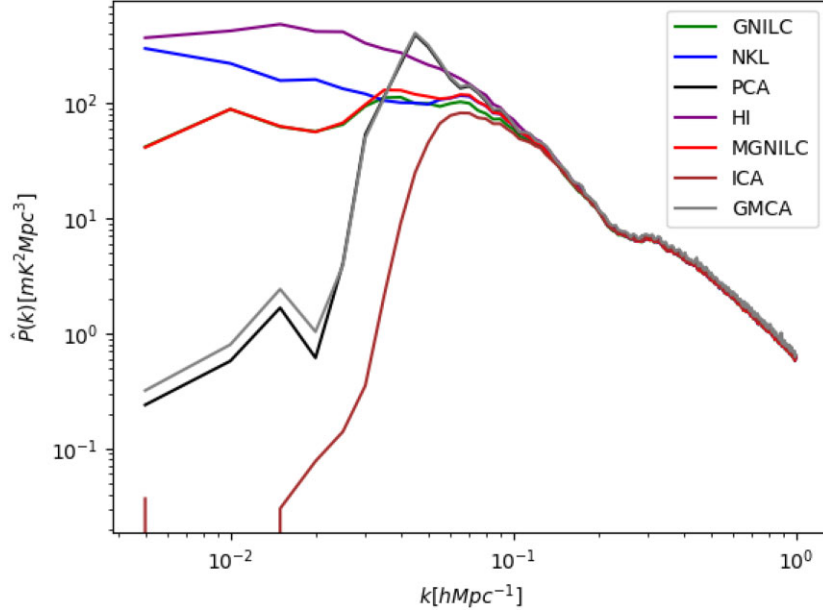
Consider again Fig. 13, which shows the foreground residuals present in this test case. As mentioned earlier, PCA and GMCA are dominated by foregrounds at  $k$ -values less than about  $0.2 \text{ hMpc}^{-1}$ . For the non-blind methods, we find that foreground residuals are strongest at low values of  $k$ . This is expected, as such scales will be dominated by contributions from low delays where foregrounds are strongest. However, foreground residuals are not quite as severe as in the higher redshift test case.

In Fig. 14, we show the spherical harmonic power spectra for various cleaning methods. This figure presents the same metric as was presented in Fig. 7. In this plot, notice first that PCA performs poorly at low  $k$  for the lower  $\ell$  modes. At these values of  $\ell$ , the very lowest  $k$  modes were cleaned too aggressively, while the slightly higher  $k$  modes were not cleaned enough, leaving behind significant foreground residuals. PCA did not suffer from such under-cleaning at the higher  $\ell$  modes, however. For  $\ell \gtrsim 200$ , MGNILC and NKL provide very similar results, with both leaving behind significant foreground residuals. This is due to the fact that PCA was used to generate the foreground approximation in this test case. When using PCA to generate the foreground approximation, AIC chose to include too few modes. As a result, not all of the foregrounds were included in  $\hat{f}$ . Any foregrounds not included in  $\hat{f}$  will not be cleaned by NKL or MGNILC, leading to the undercleaning observed here. At lower  $\ell$ , however, it appears that NKL performs best, providing the least signal loss at the lowest  $k$  modes.

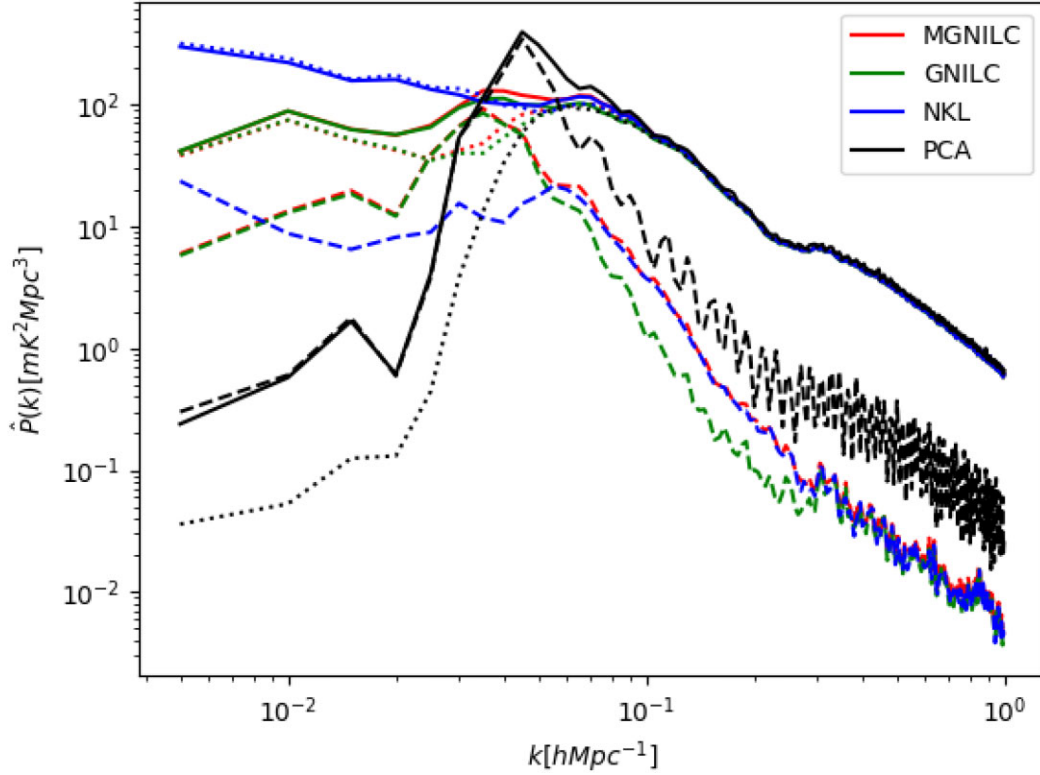
## 5.6 Performance of NKL given modified H I priors

One source of concern with respect to non-blind methods is how their performance may change subject to systematic effects or modified priors. We will leave systematics testing for future work. For now we will test the performance of NKL subject to modified H I priors.

In Shaw et al. (2015), it is mentioned that the CORA software package, which was used for this paper, assumes the H I power



**Figure 12.** 3D power spectra computed for various techniques in the lower redshift case. The ‘HI’ curve was estimated from maps that had undergone beam convolution and masking. We also tested NKL on a second realization of the signal plus noise, which produced similar results.



**Figure 13.** Power spectra illustrating the effects of residual foregrounds for various techniques in the low-redshift case. These curves are organized in the same manner as Fig. 8.

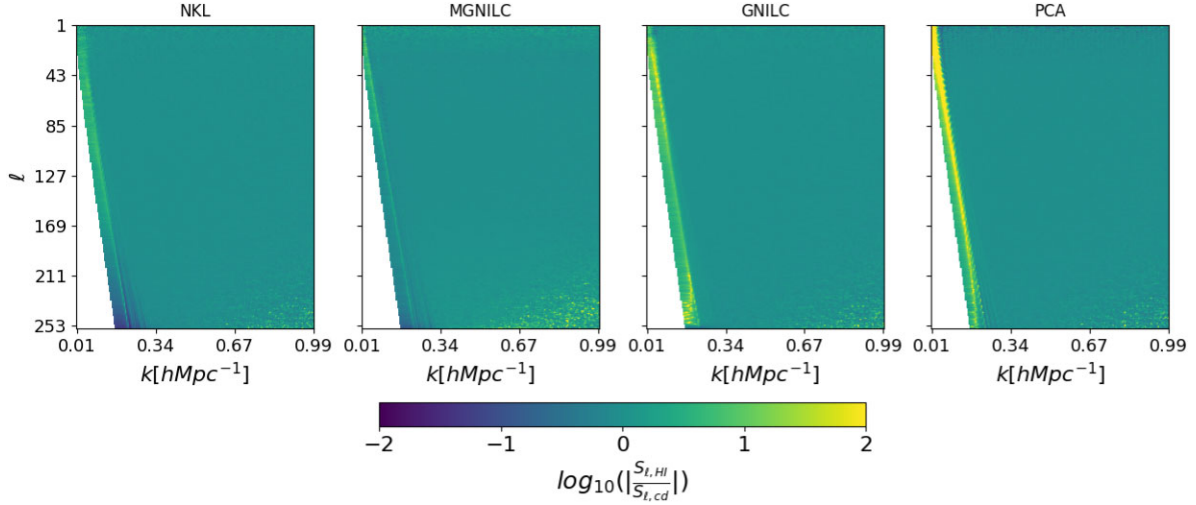
spectrum in the flat sky limit is equal to

$$P_{T_b}(\mathbf{k}, z, z') = \overline{T_b(z)}\overline{T_b(z')} (b + f\mu^2)^2 D_+(z)D_+(z')P_m(k). \quad (46)$$

In this formula,  $\overline{T_b}$  is proportional to the HI abundance  $\Omega_{\text{HI}}(z)$  and  $b(z)$  is the HI bias.

Although modelling and measurement work has been done, the values of these parameters remain uncertain, particularly at higher redshifts. For the past two decades, much work has been done to derive estimates of  $\Omega_{\text{HI}}$  from measurements. Fig. 14 of Hu et al. (2020) conveniently provides a summary of  $\Omega_{\text{HI}}(z)$  estimates out to  $z = 5$ . This figure shows increasing error bars and even





**Figure 14.** The ratio of the true spherical harmonic power spectrum ( $S_{l, \text{HI}}$ ) to the spectrum recovered by cleaning and debiasing ( $S_{l, \text{cd}}$ ) for the low- $z$  test.

disagreement between measurements as redshift increases. This becomes particularly severe for  $z \gtrsim 4$ .

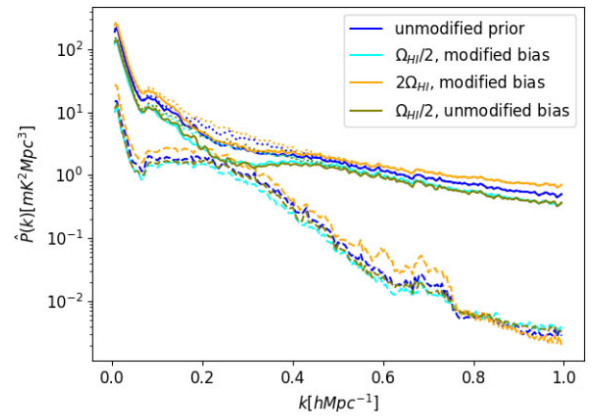
When searching in the literature for information about the hydrogen bias, one will typically find constraints coming from cross-correlation studies. These cross-correlation studies will present estimates of the product  $\Omega_{\text{HI}} b r_{\text{HI}}$ , where  $r_{\text{HI}}$  is a correlation coefficient. As a result, most of the information one finds in the literature about  $b$  by itself comes from modelling work. The model described in Castorina & Villaescusa-Navarro (2017) was presented at redshifts  $0.8 < z < 5$  and showed significant uncertainties. Possible values of  $b(z)$  spanned over a factor of  $\approx 2$  at any given redshift. This model however, did predict a smoothly increasing  $b(z)$  as a function of redshift. As a simple way to test the effect of errors in the HI prior, we repeated NKL subject to modified HI priors. In particular, NKL was performed on the same maps as before, but provided with modified estimates of the HI covariance matrices. In these test cases, we vary both the bias  $b$  and the abundance  $\Omega_{\text{HI}}$  assumed by the prior. In particular, we scale  $\Omega_{\text{HI}}$  by a constant factor and change the bias from a constant to a linear function. The coefficients of this linear curve are chosen to ensure  $b(z_{\text{min}}) = 1$  and  $b(z_{\text{max}}) = 2$ . To summarize, we modify the priors in three different ways:

- (i)  $\Omega_{\text{HI}} \rightarrow \Omega_{\text{HI}}/2$  and  $b(z) = 1 \rightarrow b(z) = az + c$
- (ii)  $\Omega_{\text{HI}} \rightarrow 2\Omega_{\text{HI}}$  and  $b(z) = 1 \rightarrow b(z) = az + c$
- (iii)  $\Omega_{\text{HI}} \rightarrow \Omega_{\text{HI}}/2$ .

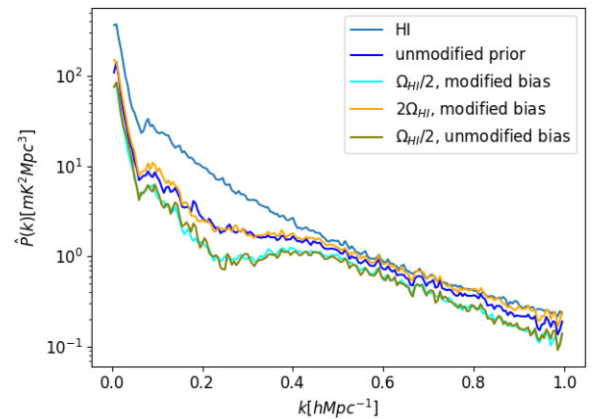
These robustness tests were performed both in the higher redshift scenario and the lower redshift scenario. The coefficients chosen for the bias were adjusted in each case to ensure that the bias varied from 1 to 2 over the redshift range.

In Fig. 15, we present results from the robustness check at higher redshifts. As can be seen, NKL is fairly robust to errors in the prior in this scenario. All priors have provided roughly similar levels of foreground residuals. One will also notice that the  $\Omega_{\text{HI}}/2$  curves are lower and the  $2\Omega_{\text{HI}}$  curve is higher. The difference in these two curves is due to the fact that the KL step of NKL chooses the number of modes to discard based on the expected signal to ratio. This SNR estimate is of course dependent upon the assumed HI amplitude. It should also be noted here that the two  $\Omega_{\text{HI}}/2$  curves look quite similar. This would imply that errors in the bias prior did not make much of a difference in this case.

In Fig. 16, we present debiased power spectra for the high-redshift test case. As one could predict from Fig. 15, we see that the  $2\Omega_{\text{HI}}$



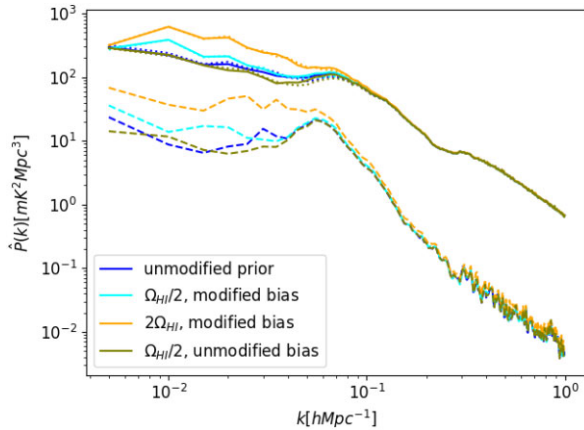
**Figure 15.** A summary of NKL performance subject to modified priors in the higher redshift scenario. The curves shown here should be interpreted in the same as Fig. 8.



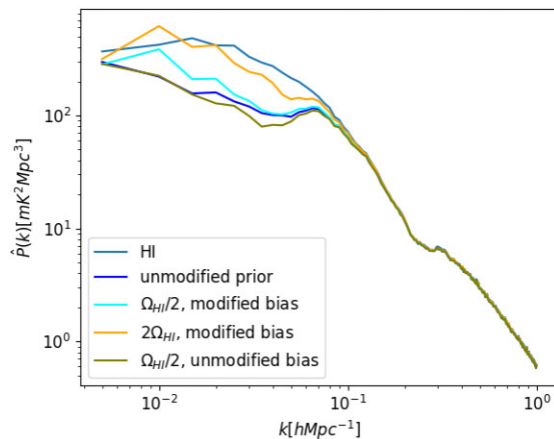
**Figure 16.** Debiased NKL power spectra in the high-redshift test case, subject to modified priors.

provides roughly similar performance to the correct prior case upon debiasing. On the other hand, the cases with  $\Omega_{\text{HI}}/2$  suffer from some additional signal loss.

In Fig. 17, we present a summary of results with varied priors in the low- $z$  case. As in the higher redshift scenario, decreasing  $\Omega_{\text{HI}}$



**Figure 17.** A summary of NKL performance subject to modified priors in the lower redshift scenario. The curves shown here should be interpreted in the same way as Fig. 8.



**Figure 18.** Debiased NKL power spectra in the lower redshift test case, subject to modified priors.

without adjusting the bias leads to roughly similar results. However, adjusting the bias seems to have more of an impact, with both the  $\Omega_{\text{HI}}/2$  and  $2\Omega_{\text{HI}}$  test cases showing larger power spectrum values at low  $k$  compared to the case with correct priors. Further testing found that NKL discarded the same number of modes in both the correct prior and  $\Omega_{\text{HI}}/2$  with linear bias test cases. This would imply that the performance difference between the  $\Omega_{\text{HI}}/2$  and  $2\Omega_{\text{HI}}$  curves of Fig. 17 is due to differences in the eigenmodes generated for the KL transform, rather than NKL discarding fewer modes due to the change in prior.

In Fig. 18, we present debiased power spectra for the lower redshift scenario. In this figure, one sees that decreasing  $\Omega_{\text{HI}}$  without adjusting the bias leads to similar results, with the cleaning being conducted slightly more aggressively. The test cases with modified bias assumptions oddly lead to better agreement, with some over estimation at the lowest values of  $k$  in the  $2\Omega_{\text{HI}}$  case.

## 6 DISCUSSION AND CONCLUSION

In this paper, we introduced the NKL and MGNILC techniques for cleaning foregrounds from HI IMs. Moreover we tested these

methods, and various others, on full sky maps in two different bands of observations ([980 MHz, 1080 MHz] and [400 MHz, 500 MHz]). These tests were conducted assuming a hypothetical telescope similar to MeerKAT operating in single-dish mode. Instrumental effects were modelled simply, with noise being drawn from a Gaussian distribution and beam effects modelled by convolution with Gaussians of frequency-dependent width. Moreover, we assumed full sky coverage and masked out the brightest 15 per cent of pixels.

At higher redshifts, where foregrounds are more severe, we found that non-blind methods such as GNILC and NKL outperform blind methods by several orders of magnitude at large spatial scales. However, all methods suffered significant signal loss. The most accurate method was NKL, which was still off by an order of magnitude in the higher redshift test and a factor of 2 or 3 in the lower redshift case. It may be possible to compensate for such signal loss using a transfer function method. For instance, the work presented in Cunningham et al. (2023) seems to show promising results when signal loss is at the 50 per cent level.

We also tested the robustness of NKL against modified priors. We found that the performance of NKL remains fairly stable while varying  $\Omega_{\text{HI}}$  and the bias  $b$ . It was found that varying  $\Omega_{\text{HI}}$  changed how aggressively NKL cleaned, while varying the bias lead to different behaviour at low  $k$  in the lower redshift scenario. However, even when provided with modified priors, NKL still provided better performance at large length-scales than the other map-space cleaning methods described in this paper.

We must emphasize that these tests do not tell the entire story of foreground removal. For instance, convolution with a Gaussian beam is not quite realistic. A more realistic beam would be more complicated than this, including sidelobes and asymmetries. Moreover, a realistic map-making process is not so simple as just convolving with a beam. One has to perform some sort of maximum-likelihood estimate based on the raw data coming from the antennas. Such a process may lead to map-making artefacts that are not so easily described. In addition, one must also consider inevitable imperfections in beam calibration. Such imperfections would likely provide their own artefacts and possibly add additional chromaticity to the foregrounds. We plan to address these issues in future work.

For perspective, we would like to recall the visibility-based work presented in Shaw et al. (2015). In that paper, polarized foregrounds were removed by projecting the visibilities onto the null space of the polarized beam matrix. A KL transform was then used to clean the unpolarized foregrounds from the visibilities. That approach recovered the HI signal effectively down to  $k \approx 0.02 \text{ hMpc}^{-1}$  for a simplified version of the CHIME telescope operating between 400 and 500 MHz. One conclusion, however, was that the main beamwidths needed to be understood to an accuracy of 0.1 per cent for this method to work.

In this paper, our motivation for further investigation of map-based techniques was that they may be more robust to beam miscalibration than are visibility-based methods. For instance, Shaw’s methods would not work in a situation where beamwidth errors are at the 1 per cent level. This leads to an incorrect model of  $C_{\text{FG}}$ , preventing the KL transform from being effective. We mention  $C_{\text{FG}}$  in specific because it is more sensitive to errors than  $C_{\text{HI}}$ . Having 1 per cent errors in  $C_{\text{FG}}$  can be hazardous when dealing with such a large dynamic range as in HI IM. However, NKL and MGNILC estimate  $C_{\text{FG}}$  directly from the data, hopefully providing it with some robustness against calibration issues.

As a next step, we intend to use the driftscan package, which employs the m-mode formalism of Shaw et al. (2015), to produce more realistic maps. This approach creates visibilities from a beam

model. We can then use a modified version of the package to generate Wiener-filtered maps from these visibilities. We will then introduce beamwidth errors to test how robust these methods are. We also intend to introduce a method similar to NKL which acts in the visibility domain. Such a method may be robust to errors in the beamwidths, as the foreground covariance would be estimated from the data and not from a beam model.

## ACKNOWLEDGEMENTS

We thank Le Zhang, Ted Bunn, Calvin Osinga, and our anonymous reviewer for providing helpful feedback on this work. This work was partially supported by NSF Award aAST-1616554, the University of Wisconsin Graduate School, the Thomas G. Rosenmeyer Cosmology Fund, and by a student award from the Wisconsin Space Grant Consortium.

## 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  
 Amiri M. et al., 2023, *ApJ*, 947, 16  
 Anderson C. J. et al., 2018, *MNRAS*, 476, 3382  
 Bandura K. et al., 2014, *Ground-based and Airborne Telescopes V*, 9145, 914522  
 Battye R. A., Davies R. D., Weller J., 2004, *MNRAS*, 355, 1339  
 Bharadwaj S., Sethi S. K., 2001, *J. Astrophys. Astron.*, 22, 293  
 Bigot-Sazy M.-A. et al., 2015, ASP Conf. Ser., 502, 41  
 Bobin J., 2017, Astrophysics Source Code Library, record ascl:1710.016  
 Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, *ApJ*, 803, 21  
 Carucci I. P., Irfan M. O., Bobin J., 2020, *MNRAS*, 499, 304  
 Castorina E., Villaescusa-Navarro F., 2017, *MNRAS*, 471, 1788  
 Chapman E. et al., 2012a, *MNRAS*, 423, 2518  
 Chapman E. et al., 2012b, *MNRAS*, 429, 165  
 Chen X., 2012, *Int. J. Mod. Phys. Conf. Ser.*, 12, 256  
 Crighton N. H. M. et al., 2015, *MNRAS*, 452, 217  
 Cunnington S., Irfan M. O., Carucci I. P., Pourtsidou A., Bobin J., 2021, *MNRAS*, 504, 208  
 Cunnington S. et al., 2022, *MNRAS*, 518, 6262  
 Cunnington S. et al., 2023, *MNRAS*, 523, 2453  
 Datta A., Bowman J. D., Carilli C. L., 2010, *ApJ*, 724, 526  
 DeBoer D. R., HERA, 2015, in American Astronomical Society Meeting Abstracts #225, p. 328.03  
 Delabrouille J., Cardoso J.-F., Le Jeune M., Betoule M., Fay G., Guilloux F., 2009, *A&A*, 493, 835  
 Delabrouille J. et al., 2013, *A&A*, 553, A96  
 Ewall-Wice A. et al., 2016, *ApJ*, 831, 196  
 Ewall-Wice A. et al., 2020, *MNRAS*, 500, 5195  
 Ghoghogh B., Karray F., Crowley M., 2023, Eigenvalue and Generalized Eigenvalue Problems: Tutorial, preprint (arXiv:1903.11240)  
 Hothi I. et al., 2020, *MNRAS*, 500, 2264  
 Hu W., Wang X., Wu F., Wang Y., Zhang P., Chen X., 2020, *MNRAS*, 493, 5854  
 Jaffe T., Waelkens A., Reinecke M., Kitaura F. S., Ensslin T. A., 2012, Astrophysics Source Code Library, record ascl:1201.014  
 Leistedt B., McEwen J. D., Vanderghaynst P., Wiaux Y., 2013, *A&A*, 558, A128  
 Liu A., Shaw J. R., 2020, *PASP*, 132, 062001  
 Liu A., Zhang Y., Parsons A. R., 2016, *ApJ*, 833, 242  
 Madau P., Meiksin A., Rees M. J., 1997, *ApJ*, 475, 429

- Marins A. et al., 2022, Foreground removal and 21 cm signal estimates: comparing different blind methods for the BINGO Telescope, preprint (arXiv:2209.11701)  
 Marinucci D. et al., 2007, *MNRAS*, 383, 539  
 Mertens F. G., Ghosh A., Koopmans L. V. E., 2018, *MNRAS*, 478, 3640  
 Moore D. F., Aguirre J. E., Parsons A. R., Jacobs D. C., Pober J. C., 2013, *ApJ*, 769, 154  
 Moore D. F. et al., 2017, *ApJ*, 836, 154  
 Morales M. F., 2005, in Kassim N., Perez M., Junor W., Henning Peds, ASP Conf. Ser. Vol. 345, From Clark Lake to the Long Wavelength Array: Bill Erickson's Radio Science. Astron. Soc. Pac., San Francisco, p. 452  
 Narowich F. J., Petrushev P., Ward J. D., 2006, *SIAM J. Math. Anal.*, 38, 574  
 Olivari L. C., Remazeilles M., Dickinson C., 2015, *MNRAS*, 456, 2749  
 Oppermann N. et al., 2012, *A&A*, 542, A93  
 Paul S., Santos M. G., Chen Z., Wolz L., 2023, A first detection of neutral hydrogen intensity mapping on Mpc scales at  $z \approx 0.32$  and  $z \approx 0.44$ , preprint (arXiv:2301.11943)  
 Planck Collaboration VI, 2020, *A&A*, 641, A6  
 Santos M. et al., 2017, in MeerKATScience: On the Pathway to the SKA, preprint (arXiv:1709.06099)  
 Shaw J. R., Sigurdson K., Sitwell M., Stebbins A., Pen U.-L., 2015, *Phys. Rev. D*, 91, 083514  
 Slepian D., 1978, *AT T Tech. J.*, 57, 1371  
 Soares P. S., Watkinson C. A., Cunnington S., Pourtsidou A., 2021, *MNRAS*, 510, 5872  
 Switzer E. R. et al., 2013, *MNRAS*, 434, L46  
 Tegmark M., Taylor A. N., Heavens A. F., 1997, *ApJ*, 480, 22  
 Wolleben M., Landecker T. L., Reich W., Wielebinski R., 2006, *A&A*, 448, 411  
 Wolz L., Abdalla F. B., Blake C., Shaw J. R., Chapman E., Rawlings S., 2014, *MNRAS*, 441, 3271  
 Wolz L. et al., 2021, *MNRAS*, 510, 3495  
 Zhang L., Bunn E. F., Karakci A., Korotkov A., Sutter P. M., Timbie P. T., Tucker G. S., Wandelt B. D., 2016, *ApJS*, 222, 3  
 Zuo S., Chen X., Mao Y., 2023, *ApJ*, 945, 38

## APPENDIX A: DETAILS OF THE ERROR ANALYSIS

In this appendix, we provide derivations of the expressions provided in Section 3.5. We derive the formulae in this subsection using first order perturbation theory. To begin, let us consider the generalized eigenvalue problem at hand. It turns out that equation (23) can be massaged into

$$\mathbf{C}_S^{-1/2} \hat{\mathbf{C}}_{\text{FG}} \mathbf{C}_S^{-1/2} \phi_i'' = \lambda_i \phi_i'' \quad (\text{A1})$$

In this formula,  $\phi_i'' = \mathbf{C}_S^{1/2} \phi_i'$  and  $\mathbf{C}_S^{1/2}$  is the Hermitian square root of  $\mathbf{C}_S$ . Note that this is the same as the equation we would get had we set up the generalized eigenvalue equation using variables whitened with  $\mathbf{C}_S^{-1/2}$ .

It turns out that the math involved in this problem is easier when whitened variables are used. As such, any symbols used during the rest of this derivation will be used to represent whitened variables. So, for instance,  $f_p$  will refer to row  $p$  of  $\mathbf{C}_S^{-1/2} \mathbf{f}$ . Moreover,  $\phi_i'$  will refer to the whitened eigenvector  $\phi_i''$ . This will not change our results, since it can be easily shown that  $E[\mathcal{F}_i' \mathcal{H}_k']$  does not change upon whitening of the variables.

To first order, the perturbed eigenvector  $\phi_i'$  will look like

$$\phi_i' = \phi_i + \sum_{j \neq i} \frac{\phi_j^T \Delta \phi_i}{\lambda_i - \lambda_j} \phi_j \quad (\text{A2})$$

In this formula,

$$\Delta = \Delta_{\text{FG}} + \mathbf{C}_{h'h'} + \mathbf{C}_{f'h'} + \mathbf{C}_{n'h'} + \mathbf{C}_{n'n'} + \mathbf{C}_{fn'}. \quad (\text{A3})$$

Refer to Section 3.5 for definitions of the terms on the right of equation (A3). Let us now consider the correlation between residual foregrounds and HI. We will find that

$$E[\mathcal{F}'_i \mathcal{H}'_k] = E[(\phi_i + \Delta\phi_i)^T f_p h_p^T (\phi_k + \Delta\phi_k)]. \quad (\text{A4})$$

Refer to Section 3.5 for definitions of  $\mathcal{F}'_i$  and  $\mathcal{H}'_i$ . Note also that  $a_p$  refers to column  $p$  of matrix  $a$ . This expectation value ought to be equal to 0 in the unperturbed case. We will also discard the second-order term, giving

$$E[\Delta\phi_i^T f_p h_p^T \phi_k] + E[\phi_i^T f_p h_p^T \Delta\phi_k]. \quad (\text{A5})$$

Recall that  $\Delta\phi_x$  really consists of several contributions

$$\Delta\phi_x = \sum_{j \neq x} \frac{\phi_j^T (\Delta_{\text{FG}} + \mathbf{C}_{h'h'} + \mathbf{C}_{fh'} + \mathbf{C}_{n'h'} + \mathbf{C}_{n'n'} + \mathbf{C}_{fn'}) \phi_x}{\lambda_x - \lambda_j} \phi_j. \quad (\text{A6})$$

The term involving  $\Delta_{\text{FG}}$  will not contribute to  $E[\mathcal{F}'_k \mathcal{H}'_i]$  as it will consist of the expectation of products of three terms depending on the foregrounds and one term depending on the HI. Since the HI and foregrounds are uncorrelated, this should result in an expectation value of 0. Similar arguments can be made for  $\mathbf{C}_{h'h'}$ ,  $\mathbf{C}_{n'h'}$ ,  $\mathbf{C}_{n'n'}$ , and  $\mathbf{C}_{fn'}$ , which will each involve odd degrees of foregrounds, noise or HI, resulting in 0 (or at least small) expectation. Small expectation is mentioned since terms like  $h'_p$  and  $f_p$  will be correlated in principle, since  $h'_p$  does depend on the foregrounds. However, we will assume that such correlation are small.

It will be the case however, that  $\mathbf{C}_{fh'}$  will create a non-negligible correlation between the foreground residuals and HI. We will have

$$E[\mathcal{F}'_i \mathcal{H}'_k] = \sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} (E[\phi_j^T \mathbf{C}_{fh'} \phi_i \phi_j^T f_p h_p^T \phi_k] + \sum_{j \neq k} \frac{1}{\lambda_k - \lambda_j} E[\phi_j^T f_p h_p^T \phi_j \phi_j^T \mathbf{C}_{fh'} \phi_k]). \quad (\text{A7})$$

Let us start by considering the first term of the right side of equation (A7). Applying the definition of  $\mathbf{C}_{fh'}$ , we obtain

$$\sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \sum_q \frac{1}{N_p} E[\phi_j^T (h'_q f_q^T + f_q h_q^T) \phi_i \phi_j^T f_p h_p^T \phi_k]. \quad (\text{A8})$$

For first term of equation (A8), we will have

$$\sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \sum_q \frac{1}{N_p} E[\phi_j^T h'_q f_q^T \phi_i \phi_j^T f_p h_p^T \phi_k]. \quad (\text{A9})$$

Let us now assume that  $h'$  and  $f$  are uncorrelated. This assumption is not strictly true, as the filter used to generate  $h'$  will be a function of the foregrounds. However, we found in numerical tests that correlations between  $h'$  and  $f$  were small. So, this approximation is safe to make. Next, we re-arrange terms inside the expectation value and using our recent assumption, we can express equation (A9) as

$$\sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \sum_q \frac{1}{N_p} E[\phi_j^T h'_q h_p^T \phi_k] E[\phi_i^T f_q f_p^T \phi_j]. \quad (\text{A10})$$

Next, note that the rapid angular variation of the signal implies that terms of in the sum of equation (A10) will be suppressed when  $q$

differs from  $p$ . As such, we will approximate equation (A10) as

$$\sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \frac{1}{N_p} E[\phi_j^T h'_p h_p^T \phi_k] E[\phi_i^T f_p f_p^T \phi_j]. \quad (\text{A11})$$

Assuming that  $E[f_p f_p^T]$  varies little as a function of  $p$ , then we can say  $E[\phi_i^T f_p f_p^T \phi_j] \approx 0$ . This assumption should be a good one when chunk sizes are small enough that the foreground temperature varies little within each chunk. On other hand, consider the case where one has a single chunk covering the whole map. In this case, the columns of  $f$  will correspond to individual lines of sight. In such a case,  $E[f_p f_p^T]$  will vary significantly with  $p$  and our assumption is no longer valid.

Thus, we find that the first term on the right side of equation (A8) is approximately 0, as long as the chunk sizes are chosen to be adequately small. As for the second term of equation (A8), we can do some rearranging and apply our previous assumptions to say that

$$\sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \sum_q \frac{1}{N_p} E[\phi_j^T f_q h_q^T \phi_i \phi_j^T f_p h_p^T \phi_k] \approx \frac{1}{N_p} \sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} E[\phi_j^T f_p f_p^T \phi_j] E[\phi_i^T h'_p h_p^T \phi_k]. \quad (\text{A12})$$

Let us define the matrix  $\alpha_{ik} = E[\phi_i^T h'_p h_p^T \phi_k]$ . In a numerical test, we found that  $\alpha_{ik}$  roughly follows the relation

$$\alpha_{ik} \approx \begin{cases} 0 & \lambda_i < 1 \\ \delta_{ik} & \lambda_i > 1. \end{cases} \quad (\text{A13})$$

This makes sense, as  $h'_p$  is just a version of  $h_p$  that has been smoothed using some filter (either PCA or DAYENU). We know that modes with large  $\lambda_i$  will couple strongly to the foregrounds, and thus to the smooth parts of  $h_p$ . As such, it should be the case that

$$E[\phi_i^T h'_p h_p^T \phi_k] \approx E[\phi_i^T h_p h_p^T \phi_k] = \delta_{ik}, \quad (\text{A14})$$

when  $\lambda_i$  is large. On the other hand, for smaller  $\lambda_i$ , one would expect that the mode would couple most strongly to higher delay parts of  $h_p$ . Since  $h'_p$  has been smoothed, it should be the case that

$$|E[\phi_i^T h'_p h_p^T \phi_k]| \ll 1, \quad (\text{A15})$$

for modes with small  $\lambda_i$ . Using this result, we obtain

$$\sum_{j \neq i} \frac{1}{\lambda_i - \lambda_j} \sum_q \frac{1}{N_p} E[\phi_j^T f_q h_q^T \phi_i \phi_j^T f_p h_p^T \phi_i] \approx \frac{\delta_{ik}}{N_p} \sum_{j \neq i} \frac{\alpha_{ij} \lambda_j}{\lambda_i - \lambda_j}. \quad (\text{A16})$$

We can go through the same process for the other term involved in equation (A7). This results in

$$\frac{\delta_{ik}}{N_p} \sum_{j \neq i} \frac{\alpha_{jj} \lambda_i}{\lambda_i - \lambda_j}, \quad (\text{A17})$$

where  $\delta_{ik}$  is a Kronecker delta function. Combining these two terms, we find

$$E[\mathcal{F}'_i \mathcal{H}'_k] \approx \frac{\delta_{ik}}{N_p} \sum_{j \neq i} \frac{(\alpha_{ii} \lambda_j + \alpha_{jj} \lambda_i)}{\lambda_i - \lambda_j}. \quad (\text{A18})$$

This paper has been typeset from a  $\text{\LaTeX}$  file prepared by the author.