Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **523,** 2453–2477 (2023) https://doi.org/10.1093/mnras/stad1567
Advance Access publication 2023 May 24

The foreground transfer function for H I intensity mapping signal reconstruction: MeerKLASS and precision cosmology applications

Steven Cunnington [1]★† Laura Wolz,[1] Philip Bull,[1,2] Isabella P. Carucci,[3,4] Keith Grainge,[1] Melis O. Irfan [2,5] Yichao Li,[6,2] Alkistis Pourtsidou,[7,8,2] Mario G. Santos,[2,9] Marta Spinelli [10,2] and Jingying Wang [11,2]

[1] *Jodrell Bank Centre for Astrophysics, Department of Physics & Astronomy, The University of Manchester, Manchester M13 9PL, UK*
[2] *Department of Physics and Astronomy, University of the Western Cape, Robert Sobukwe Road, Cape Town 7535, South Africa*
[3] *Dipartimento di Fisica, Università degli Studi di Torino, via P. Giuria 1, I-10125 Torino, Italy*
[4] *INFN – Istituto Nazionale di Fisica Nucleare, Sezione di Torino, via P. Giuria 1, I-10125 Torino, Italy*
[5] *Department of Physics and Astronomy, Queen Mary University of London, London E1 4NS, UK*
[6] *Department of Physics, College of Sciences, Northeastern University, Wenhua Road, Shenyang 11089, China*
[7] *Institute for Astronomy, The University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK*
[8] *Higgs Centre for Theoretical Physics, School of Physics and Astronomy, The University of Edinburgh, Edinburgh EH9 3FD, UK*
[9] *South African Radio Astronomy Observatory (SARAO), 2 Fir Street, Cape Town 7925, South Africa*
[10] *Department of Physics, Institute of Particle Physics & Astrophysics, ETH Zurich 8093, Switzerland*
[11] *Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, China*

Accepted 2023 May 20. Received 2023 April 24; in original form 2023 February 14

ABSTRACT

Blind cleaning methods are currently the preferred strategy for handling foreground contamination in single-dish H I intensity mapping surveys. Despite the increasing sophistication of blind techniques, some signal loss will be inevitable across all scales. Constructing a corrective transfer function using mock signal injection into the contaminated data has been a practice relied on for H I intensity mapping experiments. However, assessing whether this approach is viable for future intensity mapping surveys, where precision cosmology is the aim, remains unexplored. In this work, using simulations, we validate for the first time the use of a foreground transfer function to reconstruct power spectra of foreground-cleaned low-redshift intensity maps and look to expose any limitations. We reveal that even when aggressive foreground cleaning is required, which causes > 50 per cent negative bias on the largest scales, the power spectrum can be reconstructed using a transfer function to within sub-per cent accuracy. We specifically outline the recipe for constructing an unbiased transfer function, highlighting the pitfalls if one deviates from this recipe, and also correctly identify how a transfer function should be applied in an autocorrelation power spectrum. We validate a method that utilizes the transfer function variance for error estimation in foreground-cleaned power spectra. Finally, we demonstrate how incorrect fiducial parameter assumptions (up to ± 100 per cent bias) in the generation of mocks, used in the construction of the transfer function, do not significantly bias signal reconstruction or parameter inference (inducing < 5 per cent bias in recovered values).

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

1 INTRODUCTION

Probing fluctuations in the Universe's density field is an excellent tool for furthering precision cosmology. A number of large sky surveys have been commissioned with this aim and have contributed towards constraining parameters in the standard cosmological model (eBOSS Collaboration 2021; Heymans et al. 2021; DES Collaboration 2022). Despite cosmic microwave background (CMB) experiments leading the way with constraints (Planck Collaboration VI 2020), it is expected that large-scale structure maps will soon be the leading resource given the three-dimensional information they provide (Slosar et al. 2008; Giannantonio et al. 2012; Alonso et al. 2015b). Increases in survey volume will allow fluctuations across the largest scales to be probed, improving constraints. It is within the relatively unexplored *ultra*-large scales where novel tests of general relativity will be possible and where there will be increased sensitivity to new physics such as non-Gaussian fluctuations in the Universe's primordial density field (Camera et al. 2013; Alvarez et al. 2014; Baker & Bull 2015; Camera, Maartens & Santos 2015; Fonseca et al. 2015; Bull 2016; Cunnington 2022). Ultra-large scales are also

★ E-mail: steven.cunnington@manchester.ac.uk
† On behalf of the MeerKLASS Collaboration.

highly linear, avoiding the complex modelling challenges facing surveys attempting to exploit non-linear regimes (D'Amico et al. 2020; Martinelli et al. 2021; Pourtsidou 2023).

An efficient method for surveying large volumes is using radio telescopes to map the redshifted 21cm emission from neutral hydrogen (H I). The H I, which mostly resides in galaxies in the post-reionization Universe, traces the underlying dark matter, thus allowing the Universe's large-scale structure to be probed. By rapidly scanning the sky and recording all radiation as unresolved diffuse emission, the faint 21cm signals are integrated allowing for a comprehensive survey of H I density in 3D to be obtained. This process is known as H I intensity mapping (Bharadwaj et al. 2001; Battye, Davies & Weller 2004; Chang et al. 2008; Wyithe, Loeb & Geil 2008).

The H I power spectrum has been recently detected on small Mpc scales (Paul et al. 2023) with intensity maps from the 64-dish MeerKAT array, a pathfinder telescope for the Square Kilometre Array Observatory (SKAO) (SKA Cosmology SWG 2020). This detection used MeerKAT as an interferometer which has higher sensitivity on small scales. However, within the next few years the MeerKAT Large Area Synoptic Survey (MeerKLASS) plans to conduct a wide ($\gtrsim 4{,}000\,\mathrm{deg}^2$) H I intensity mapping survey, potentially spanning $0.4 < z < 1.45$ in redshift if performed using the UHF band (Santos et al. 2017). Since the MeerKAT interferometer does not have sufficiently short baselines to achieve such a field of view, the observations will be gathered using the single-dish data from each element of the array. This autocorrelation mode of observation, often referred to as *single-dish mode*, is also planned for the full SKAO in order to probe large-scale cosmology, which has been identified as a top priority science objective (Weltman et al. 2020). In the pre-SKA era however, MeerKAT will pursue low-redshift H I intensity mapping and has already demonstrated calibration and map-making from single-dish mode observations with a small pilot survey (Wang et al. 2021). This same pilot survey was also used to achieve the first single-dish mode cosmological detection with a multidish array in cross-correlation with an overlapping galaxy survey (Cunnington et al. 2022).

Since H I intensity mapping records all emission in the frequency range of the instrument, the major challenge is removing any signals which are not cosmological H I. This can include radio frequency interference (RFI) and astrophysical foregrounds, both of which can dominate by orders of magnitude over the weak H I signal.[1] In principle, RFI should be time-varying and can be flagged when it enters the observations. However, foregrounds will consistently enter the observations due to their fixed sky coordinates, therefore a process for separating them from the H I is required. The dominant sources producing foregrounds in the low-redshift H I frequency range ($\sim 300 < \nu < 1420\,\mathrm{MHz}$) are synchrotron and free–free radiation from within our own galaxy, along with extra-galactic concentrated emission from strong point sources such as active galactic nuclei (Oh & Mack 2003; Santos, Cooray & Knox 2005; Alonso, Ferreira & Santos 2014).

To date, *blind* foreground cleaning techniques have been the only approach that has led to cross-correlation detections of a cosmological power spectrum in single-dish intensity mapping (Masui et al. 2013; Wolz et al. 2017, 2022; Anderson et al. 2018; Cunnington et al. 2022). Blind techniques exploit the robust assumption that foregrounds are a dominant and correlated contribution to the observations and can be statistically reduced into a few components which are removed. This requires little prior knowledge of the foregrounds which is a huge advantage since it is challenging to obtain a detailed understanding of the foreground's precise amplitude through frequency, or how they respond to instrumental systematics. Blind foreground removal performed at the map level has proven to be the most successful approach and these have been validated and refined in simulations (Wolz et al. 2014; Alonso et al. 2015a; Carucci, Irfan & Bobin 2020; Cunnington et al. 2021a; Spinelli et al. 2021). Interferometric intensity mapping can to some extent adopt a foreground avoidance strategy, which assumes they are isolated in a foreground wedge region in $(k_\perp, k_\parallel)$-Fourier space (Liu, Parsons & Trott 2014; Paul et al. 2023). The foreground avoidance technique has the advantage of being immune to signal loss from foreground cleaning. However, recent studies have shown some component separation improves foreground mitigation relative to foreground avoidance alone in interferometric surveys (Chen, Wolz & Battye 2023). Hence blind foreground removal is likely to be an adopted technique beyond single-dish mode experiments.

While blind foreground cleaning algorithms themselves have been well studied, the precise effects they cause on the underlying H I field have not been to the same extent. Broadly speaking there are two unfavourable consequences from blind foreground cleaning, and both will occur simultaneously to some extent. The first is *foreground residuals*, i.e. foreground contamination not removed from the data resulting from undercleaning. The second is *signal loss* i.e. the reduction in the H I power spectrum amplitude resulting from the foreground clean. It is this second issue that is the main focus of this paper. While foreground residuals are of course important, their influence on the data is similar to RFI and instrumental noise, that is, they cause an additive bias to the estimated H I power spectrum. However, for foregrounds, it is expected that their residuals should be reducible to sub-dominant levels relative to the H I (Cunnington et al. 2021a). Furthermore, in cross-correlation with a foreground-free tracer such as a galaxy survey, any additive bias from foreground residuals will be absent and the only impact will be on the error budget.

For signal loss, it has been shown that even in ideal simulations, some loss is always inevitable across all scales when blind foreground removal methods are applied, and this is not mitigated in cross-correlations (Cunnington et al. 2021a). Ignoring or incorrectly estimating signal loss, unsurprisingly, leads to a biased recovery of the H I power spectrum. Thus signal loss is a crucial concept to understand exhaustively for precision cosmology to be possible with H I intensity mapping. The necessary process of signal reconstruction, i.e. correcting for the signal loss, is where there is little dedicated study. A foreground transfer function $\mathcal{T}$ can be simply defined as the object which delivers a reconstructed signal power spectrum that is unbiased to the underlying truth i.e. $\langle P_{\rm rec}(k)\rangle = P_{\rm true}(k)$, where $\langle P_{\rm rec}(k)\rangle \equiv P_{\rm clean}(k)/\mathcal{T}(k)$. Previous intensity mapping detections (Masui et al. 2013; Anderson et al. 2018; Cunnington et al. 2022; Wolz et al. 2022) have all relied on a process of mock signal injection to estimate the foreground transfer function. By subjecting the injected mocks to the same foreground cleaning process as the observations, we can use the drop in the measured mock power spectra to estimate the transfer function. This method was first extensively analysed in Switzer et al. (2015) in the context of the Green Bank Telescope (GBT) H I intensity maps (Masui et al. 2013; Switzer et al. 2013). There have been similar methods of signal injection to correct for signal loss implemented for epoch of reionization experiments where past analyses underestimated signal

[1]Additional contaminants come from atmospheric emission and ground pickup which can be approximately modelled as constant over time when a constant elevation scanning strategy is adopted.

loss leading to biased results (Cheng et al. 2018), highlighting the importance of correctly understanding this issue. To date, there has been no dedicated simulations-based investigation into the reliability of the transfer function for low-redshift H I intensity maps blindly cleaned at the map level, despite its clear importance.

In this work, we use various H I intensity mapping simulations to validate the reliability of the transfer function for signal reconstruction in foreground-cleaned maps. We explore how signal loss arises in a foreground clean and illustrate the subtleties of this both analytically and empirically in simulation tests, demonstrating how a transfer function can be correctly estimated to account for all these subtleties. We focus exclusively on Principal Component Analysis (PCA)-based foreground cleaning but much of the formalism and results presented will be transferable to other foreground cleaning techniques. Furthermore, while our focus is on single-dish intensity mapping, the conclusions will also be applicable to interferometers. We demonstrate how a foreground transfer function is a reliable tool for small pilot surveys, validating it on simulations constructed using empirical MeerKAT 2019 observations aiming to realistically emulate current MeerKAT intensity maps. Lastly, we look to the future and pursue to what extent the transfer function can be relied on for conducting precision cosmology with intensity mapping where sub-per cent accuracy on parameter estimation is the aim.

This paper is structured as follows; in Section 2 we present an overview of the formalism for a blind PCA-based foreground clean, explicitly highlighting where signal loss arises. Section 3 presents how one should construct a foreground transfer function to correct for signal loss. In Section 4 we test the transfer function on a simulation of a MeerKAT-like intensity mapping pilot survey, validating the transfer function in this low signal-to-noise regime. Section 5 focuses on how suited the transfer function is for the purposes of precision cosmology, showcasing the robustness of the transfer function even where the fiducial cosmology assumed for its construction disagrees with the truth in the observations. Finally, we conclude in Section 6.

2 SIGNAL LOSS FROM FOREGROUND CLEANING

We begin with a pedagogical introduction to the formalism describing blind foreground cleaning and with the aid of simulations demonstrate some key concepts of signal loss induced by the foreground clean. For consistency, we largely follow the notation in Switzer et al. (2015). While we focus on a PCA-based method, the formalism we present is in principle transferable to more sophisticated blind foreground removal techniques when they are used as linear filters (e.g. Bobin et al. 2007; Chapman et al. 2012; Alonso et al. 2015a; Carucci et al. 2020; Cunnington et al. 2021a; Irfan & Bull 2021; Spinelli et al. 2021). We use a set of simulations with separable H I signal and foreground contributions, allowing us to provide examples of the claims made in certain derivations. We begin by using some generic simulations which are similar to that used in Cunnington et al. (2021a). The exact details of these simulations are outlined in Appendix A1 but we include a short summary of points below.

(i) The 1 $(\mathrm{Gpc}/h)^3$ MULTIDARK (Klypin et al. 2016) *N*-body semi-analytical simulation with approximate cold gas masses is used for the single realization of the underlying *true* H I signal at a snapshot redshift of $z = 0.39$, gridded into $n_\mathrm{x}, n_\mathrm{y}, n_\mathrm{z} = 256, 256, 256$ voxels. We include redshift-space distortions (RSD) to provide the H I signal with an anisotropic signature. A frequency range of $900 < \nu < 1156$ MHz with resolution $\delta\nu = 1$ MHz is assumed which is consistent with the snapshot redshift and is reasonably consistent with MeerKAT L-band intensity mapping observations.

(ii) We simulate galactic synchrotron, galactic free–free, and bright point source emission at these frequencies to provide a foreground sky. We use the Planck Legacy Archive[2] FFP10 simulations for the synchrotron and free–free emission. The point sources catalogue is produced following the same approach as in Battye et al. (2013).

(iii) We cut a patch of sky consistent with the 1 $(\mathrm{Gpc}/h)^2$ H I survey size and chose this to be centred on the Stripe 82 region of the sky, where a real survey could be targeted. The foreground component with $n_\theta = n_\mathrm{x} \times n_\mathrm{y}$ angular pixels and $n_\nu \equiv n_\mathrm{z}$ frequency channels is added on to the H I signal simulation.

(iv) To increase the complexity of the foreground clean, we simulate instrumental polarization leakage (Carucci et al. 2020; Cunnington et al. 2021a) which disrupts the smooth frequency spectra of the foreground simulations, requiring a clean which is more aggressive and consistent with real data. For this we used the CRIME[3] software (Alonso et al. 2014). This is used by default and we highlight any cases where this has not been used.

(v) By default we add no further instrumental effects, but in some cases we introduce instrumental noise and smoothing perpendicular to the line of sight to emulate the telescope beam. For the noise, we assume isotropic Gaussian white noise with $\sigma_\mathrm{n} = 1$ mK, approximately corresponding to 30 h of observation time on a MeerKAT-like survey of $\sim 3000\,\mathrm{deg}^2$ sky (see equations A3 and A4 for more details). This noise dominates over the H I signal which has an rms of $\sigma_{\mathrm{H\,I}} \sim 0.14$ mK. The beam we approximate as a Gaussian with comoving transverse length-scale $R_\mathrm{beam} = 10\,h^{-1}$Mpc (see equation A2 for a definition). We clearly indicate cases where noise or a beam has been added. We discuss the limitations of these approximations and also explore some more realistic systematics in Section 4 based on real MeerKAT pathfinder data, which we introduce there.

Throughout we will refer to these as the MD1GPC simulations. Later in the paper, we use some more specific simulations to explore different scenarios which we will introduce then, but for the majority of our results, we use the MD1GPC by default unless otherwise clearly stated.

To begin a PCA-based clean of the H I + foregrounds combination, we first calculate the ν, ν' covariance of the foreground contaminated data X_obs, where the data matrices X have dimensions $[n_\nu, n_\theta]$. The covariance is estimated by $\mathsf{C} = (n_\theta - 1)^{-1}\mathsf{X}_\mathrm{obs}^\mathrm{T}\mathsf{X}_\mathrm{obs} = \mathsf{U}\mathbf{\Lambda}\mathsf{U}^\mathrm{T}$, where the last equality is the eigen-decomposition (or diagonalization) of the covariance matrix, with U representing a matrix with the n_ν spectral eigenvectors U_i and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. Neglecting noise contributions,[4] i.e. $\mathsf{X}_\mathrm{obs} = \mathsf{X}_\mathrm{f+s} \equiv \mathsf{X}_\mathrm{f} + \mathsf{X}_\mathrm{s}$, we can write this as

$$\mathsf{C} = (n_\theta - 1)^{-1}\,(\mathsf{X}_\mathrm{f} + \mathsf{X}_\mathrm{s})^\mathrm{T}\,(\mathsf{X}_\mathrm{f} + \mathsf{X}_\mathrm{s})\,, \tag{1}$$

which expands to

$$\begin{aligned}\mathsf{C}_\mathrm{f+s} &= (n_\theta - 1)^{-1}\left(\mathsf{X}_\mathrm{f}\mathsf{X}_\mathrm{f}^\mathrm{T} + \mathsf{X}_\mathrm{f}\mathsf{X}_\mathrm{s}^\mathrm{T} + \mathsf{X}_\mathrm{s}\mathsf{X}_\mathrm{f}^\mathrm{T} + \mathsf{X}_\mathrm{s}\mathsf{X}_\mathrm{s}^\mathrm{T}\right)\\ &= \mathsf{C}_\mathrm{f} + \mathsf{C}_\Delta\,,\end{aligned} \tag{2}$$

[2] pla.esac.esa.int/pla

[3] intensitymapping.physics.ox.ac.uk/CRIME.html

[4] This is mainly done for brevity and incorporating noise into this formalism is not overly complicated. It acts as additional perturbations to the pure foreground modes in a similar way to the H I signal, as we later discuss.

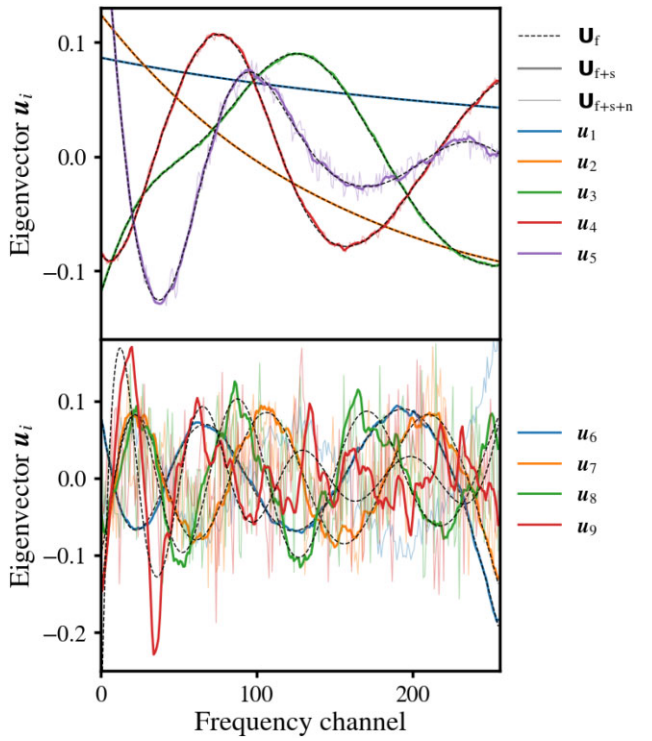

Figure 1. The first nine eigenvectors (first five in the top panel and next four in the bottom) from the PCA on the MD1GPC foreground contaminated simulations. The dashed black lines show purely foreground modes taken from the PCA on foreground-only sims. The solid colour lines show the eigenvectors perturbed by the inclusion of the H I signal in the data. These perturbations are the origins of signal loss in foreground cleaning. The thin faint solid colour lines show eigenvectors perturbed by the inclusion of the H I signal + noise.

where $\mathsf{C}_\Delta = (n_\theta - 1)^{-1}(\mathsf{X}_f\mathsf{X}_s^T + \mathsf{X}_s\mathsf{X}_f^T + \mathsf{X}_s\mathsf{X}_s^T)$ are residual contributions to the estimate of the foreground covariance. The estimate of the foregrounds, which is to be removed from the observations, is then given by

$$\hat{\mathsf{X}}_f = \mathsf{U}\mathsf{S}\mathsf{U}^T\mathsf{X}_{obs}, \tag{3}$$

where, following Switzer et al. (2015), we introduce the selection matrix S which is zero everywhere except for the first N_{fg} elements along the diagonal which are set equal to one, assigning the number of contaminated modes projected out from each line of sight.

While we expect the foregrounds to be orders of magnitude larger than the H I signal, and C_f to be the dominant term in equation (2), it is important for the additional perturbations from the signal through C_Δ to be considered. Due to this mix of foregrounds and signal, the eigenvectors we obtain are

$$\mathsf{U} \equiv \mathsf{U}_{f+s} = \mathsf{U}_f + \mathbf{\Delta}, \tag{4}$$

where U_f are pure foreground modes and $\mathbf{\Delta}$ are the perturbed contributions caused by the signal. We show the difference between the unperturbed (U_f) and perturbed (U_{f+s}) eigenvectors in Fig. 1. This shows the first nine most dominant eigenvectors from the MD1GPC simulations. For the unperturbed, pure foreground eigenvectors (dashed black lines), the modes are smooth in frequency, only showing longer wavelength oscillations caused by the polarization leakage. However, for the eigenvectors estimated from the foreground and signal mix (solid coloured lines), perturbations to the modes caused by the signal start to arise. These perturbations are more severe the higher the order of the eigenvector. This is because the eigenvectors become increasingly less dominant and are more easily perturbed by the presence of the signal whose contributions remain fairly consistent through all the eigenvectors due to its high-rank properties.

Fig. 1 demonstrates how signal loss can enter a foreground clean. Using the eigenvectors U_{f+s} as the basis functions which are projected out in the PCA clean, this will project out modes that have some H I structure shown by the perturbations on the lower modes. It is tempting to try and address this issue at this early stage and use a low-pass filter or smooth the perturbed modes to correct the perturbations from the signal. However, we briefly experimented with various smoothing routines with this aim and in all cases, the results were made worse. Since the aim of this work is not to enhance foreground cleaning efficiency, but instead to ensure we can control signal loss, we defer any investigations into cleaning optimization to future work.

We also show in Fig. 1 the impact on the eigenmodes by including the dominant instrumental noise (U_{f+s+n} shown by faint coloured lines). These create much larger perturbations to the pure foreground modes, hence large noise can impact foreground cleaning. This is an important issue for early pilot surveys where observation time is low since in these cases, the noise will dominate over the H I and will be the main source of perturbations to the eigenvectors. We will discuss this in more detail later and demonstrate how intensity maps with a high level of noise, and other additive systematics like residual RFI, can still undergo signal reconstruction using a transfer function. For the remainder of this section, we omit the instrumental noise for simplicity.

The perturbations in Fig. 1 are dependent on the ratio between the foreground amplitude and the other components e.g. the H I. While the H I signal amplitude will be consistently uniform due to the cosmological principle, the foreground emission can vary with the choice of sky patch, e.g. being orders of magnitude higher near the galactic plane relative to the South Celestial Pole. This was explored in Cunnington et al. (2021a) where different foreground regions were tested. For the remainder of this work, however, we will stick to one region as most of the conclusions we draw are generic regardless of how strong the foregrounds are, within a physically reasonable range.

2.1 Toy model foreground cleaning

Here, we investigate some idealized foreground cleaning scenarios to demonstrate the nature of signal loss in blind foreground cleaning. We begin with the most ideal toy case scenario where we project out pure foreground modes from pure foreground-only data and subtract this from the observed combination X_{f+s}. Of course, if we could access perfect foreground-only data X_f this could be simply subtracted from the observed foreground and signal mix, so there would be no need for any mode projection cleaning process. However, we proceed with this example since it provides valuable insight from which we can add complication. This first toy-case is given by

$$\mathsf{X}_{toy:clean1} = \mathsf{X}_{f+s} - \mathsf{U}_f\mathsf{S}\mathsf{U}_f^T\mathsf{X}_f, \tag{5}$$

In this ideal case, since we are projecting out perfect foreground modes from pure foreground data, the optimal selection matrix S would have the diagonals filled with ones ($N_{fg} = n_\nu$) i.e. the identity matrix, to remove all foreground without the consequence of signal loss. In reality, this is not possible and a balance is sought between projecting out enough modes to remove foregrounds but not so many that large signal loss is sustained.

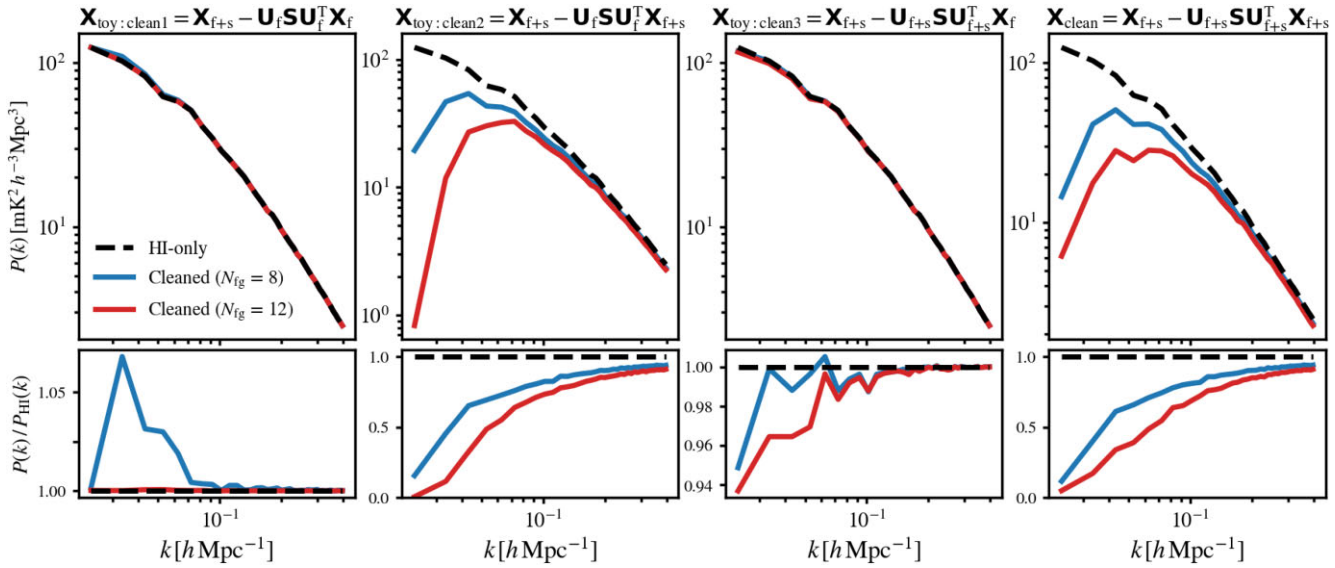

Figure 2. Autopower spectra for different foreground cleaning cases in comparison with the true H I-only simulations (black dashed line). The first three panels represent idealized scenarios. *Far-left* shows projecting pure foreground modes ($\mathsf{U}_{\rm f}$) out from pure foreground data ($\mathsf{X}_{\rm f}$). *The second panel* shows projecting pure foreground modes out from foreground data mixed with signal ($\mathsf{X}_{\rm f+s}$). *The third panel* shows projecting modes perturbed by the signal ($\mathsf{U}_{\rm f+s}$) out from pure foreground data. *Far-right* shows the realistic scenario where modes are perturbed by the signal and these are then projected out from foreground data mixed with signal.

In Fig. 2 we show the measured power spectrum for the idealized toy clean in equation (5) (far-left panel). We also show other cases of foreground clean in the other panels which we will discuss shortly. Details on the power spectrum estimation formalism used throughout the paper are presented in Appendix B. We show the power spectra in comparison with the original H I-only (foreground-free) data, which we are aiming to agree with. In all cases we show two cleaned examples with 8 and 12 modes projected out, i.e. the first $N_{\rm fg} = 8$ and $N_{\rm fg} = 12$ elements in the diagonal of S are set to 1. This is why we do not reach perfect agreement in the ideal top-left panel, because we are not projecting out all pure foreground modes, leaving some foreground residual in the remaining modes. Thus, there is slight disagreement with the H I-only, albeit at a sub-per cent level for the $N_{\rm fg} = 12$ case. The values of $N_{\rm fg} = 8,\ 12$ are chosen to sufficiently suppress the polarized foregrounds in the increasingly more realistic cases shown by the other panels which we discuss next.

In reality, the situation in equation (5) is not possible, because we cannot project out modes from foreground-only data $\mathsf{X}_{\rm f}$ because the observed data we have, $\mathsf{X}_{\rm f+s}$, are inherently mixed with signal. Signal loss begins to manifest in the case where we project out foreground modes from the observed data mix

$$\mathsf{X}_{\rm toy:clean2} = \mathsf{X}_{\rm f+s} - \mathsf{U}_{\rm f}\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm f+s}. \tag{6}$$

This is still an idealized scenario since we are projecting out *pure* foreground modes from the data. In reality, a further complication arises since the modes identified in the PCA will be perturbed by the presence of signal i.e. $\mathsf{U}_{\rm f} \rightarrow \mathsf{U}_{\rm f+s}$, which we will discuss shortly. Comparing the first panel with the second, where the difference is that in the former case, pure foreground modes are now being projected out from the mix of foreground and signal (equation 6), evidence of signal loss in the cleaned power spectrum begins to show. The signal loss is clearly dominating over any foreground residuals remaining in the data from only projecting out a finite number of modes. In other words, the small ~ 5 per cent additive bias in the far-left panel caused by foreground residuals is not seen in the second panel, due to a more dominant impact from signal loss. Of course if a smaller number for $N_{\rm fg}$ were chosen, foreground residuals would cause more of a problem.

The second panel of Fig. 2 confirms that signal loss begins to manifest when modes (even purely foreground ones) are projected out of the data, which is a combination of foregrounds and signal. The reason for this is because signal will unavoidably have degeneracies with some foreground structure. Thus when a set of foreground functions are projected out of the data, some signal will leak into this subtraction, mainly large-scale (small-k) modes since these are most degenerate with the foreground structure which is highly correlated through frequency.

The third panel of Fig. 2 shows a final idealized toy case where we are only projecting out modes from the pure foreground map, but the modes are now perturbed by the presence of signal, $\mathsf{U}_{\rm f+s}$ (see equation 4). This is something we have to deal with in reality where we cannot form a perfect estimation for the foreground-only eigenmodes $\mathsf{U}_{\rm f}$, from the true observed data where foreground and signal are mixed. In this case, the eigenvectors themselves are perturbed and it is these perturbed modes that we project along the data;

$$\mathsf{X}_{\rm toy:clean3} = \mathsf{X}_{\rm f+s} - \mathsf{U}_{\rm f+s}\mathsf{S}\mathsf{U}_{\rm f+s}^{\rm T}\mathsf{X}_{\rm f}. \tag{7}$$

This provides an interesting result with signal loss again appearing to be the more dominant effect with little evidence of additive bias from foreground residuals. However, we are only projecting out modes from pure foreground data, so it seems counter-intuitive that there is signal loss. As we will explicitly show in the following section, this is caused by the perturbation to the modes from the presence of signal ($\mathsf{U}_{\rm f+s}$), which creates a complicated mix of subtracted terms that can have signal correlating and anticorrelating contributions, as identified in Switzer et al. (2015).

2.2 The origins and subtleties of signal loss

Despite seeing signal loss in the second and third panels of Fig. 2, both cases still represent unrealistic scenarios. In reality, we see a

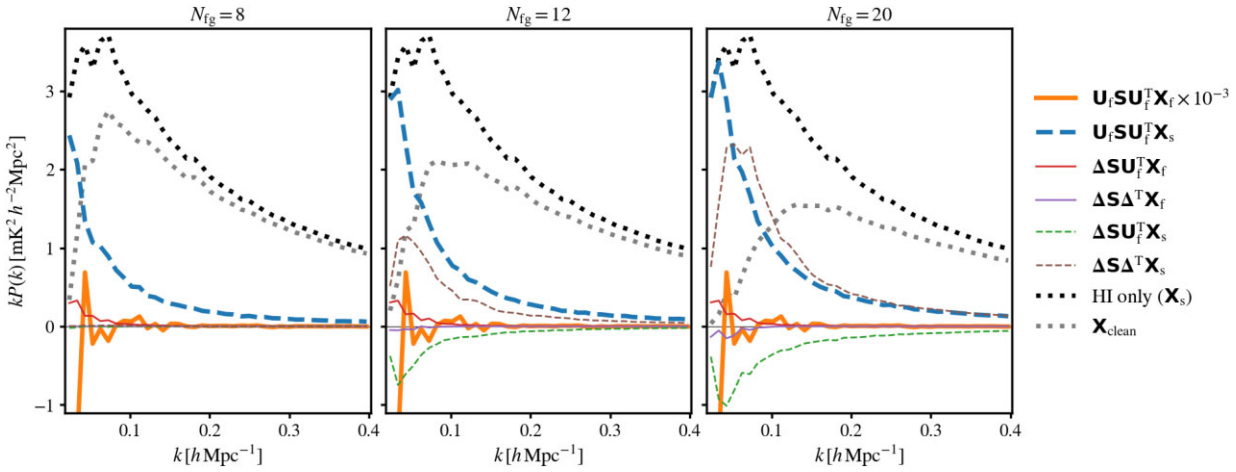

Figure 3. Contributions to signal loss from the decomposed terms in equation (9). Each power spectra shows the cross-correlation between the original-H I and a subtracted term in the foreground clean. The different panels represent different numbers of PCA modes removed ($N_{\rm fg}$). The solid lines indicate residual foreground contributions, whereas the dashed lines indicate H I signal contributions. The thin lines indicate perturbed contributions from $\mathbf{\Delta}$ caused by the presence of signal in the eigenmode estimation. For reference, we also include the pure H I-power (black dotted) along with the total cleaned result $\mathsf{X}_{\rm clean}$ (grey dotted).

combination of both where the presence of signal perturbs the eigenmodes ($\mathsf{U}_{\rm f+s}$) as well as complicating the clean since information is projected out from data which contain not just foregrounds, but signal too ($\mathsf{U}_{\rm f+s}\mathsf{S}\mathsf{U}_{\rm f+s}^{\rm T}\mathsf{X}_{\rm f+s}$). Thus, the true resulting cleaned data are given by

$$\mathsf{X}_{\rm clean} = \mathsf{X}_{\rm f+s} - \mathsf{U}_{\rm f+s}\mathsf{S}\mathsf{U}_{\rm f+s}^{\rm T}\mathsf{X}_{\rm f+s}. \tag{8}$$

The result from this foreground clean is shown in the final far-right panel of Fig. 2. Results appear similar to $\mathsf{X}_{\rm toy:clean2}$ but some differences can be seen on large scales caused by the increased complexity of having perturbed eigenmodes ($\mathsf{U}_{\rm f+s}$).

To understand the complexity of foreground cleaning, we expand the above equation (8) into all its terms, giving

$$\begin{aligned}\mathsf{X}_{\rm clean} &= \left[1 - (\mathsf{U}_{\rm f} + \mathbf{\Delta})\,\mathsf{S}\,(\mathsf{U}_{\rm f} + \mathbf{\Delta})^{\rm T}\right](\mathsf{X}_{\rm f} + \mathsf{X}_{\rm s}) \\ &= \mathsf{X}_{\rm f} + \mathsf{X}_{\rm s} - \mathsf{U}_{\rm f}\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm f} - \mathsf{U}_{\rm f}\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm s} - \Delta\mathsf{S}\mathsf{U}_{\rm f}\mathsf{X}_{\rm f} - \mathsf{U}_{\rm f}\mathsf{S}\Delta^{\rm T}\mathsf{X}_{\rm f} \\ &\quad - \mathbf{\Delta}\mathsf{S}\mathbf{\Delta}^{\rm T}\mathsf{X}_{\rm s} - \mathbf{\Delta}\mathsf{S}\mathbf{\Delta}^{\rm T}\mathsf{X}_{\rm f} - \mathbf{\Delta}\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm s} - \mathsf{U}_{\rm f}\mathsf{S}\mathbf{\Delta}^{\rm T}\mathsf{X}_{\rm s}. \end{aligned} \tag{9}$$

In Fig. 3 we show power spectra for the subtracted decomposed terms in equation (9), plotting their cross-correlation with the pure H I signal to demonstrate where signal loss originates. Since these are the subtracted terms, the higher their cross-power with pure-H I, the more they are contributing to signal loss. For reference, we also show the pure-H I (i.e. the H I autocorrelation) as the black dotted line, and the fully cleaned result (all terms from equation 9 combined) as the grey dotted line.

As expected, a large bulk of signal loss is caused by the subtraction of the $\mathsf{U}_{\rm f}\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm s}$ term (blue dashed line). This *direct* signal loss will clearly increase for higher $N_{\rm fg}$ i.e. more ones along the diagonal of S, and this is demonstrated by the growing amplitude of the blue dashed line mostly at small-k, going left to right in the panels. The projected out foregrounds $\mathsf{U}_{\rm f}\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm f}$ are entirely uncorrelated with the H I signal as shown by the consistent with zero power spectrum (orange line). However, we still decrease the amplitude of $\mathsf{U}_{\rm f}\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm f}$ by three orders of magnitude (indicated in the legend) since this dominant term still has large purely statistical fluctuations in the H I cross-correlation dependent only on the foreground realization. The thinner lines represent terms including perturbative contributions from the H I signal $\mathbf{\Delta}$. This is where the issue of signal loss begins to complicate. As $N_{\rm fg}$ increases, the contribution to signal loss from $\Delta\mathsf{S}\Delta^{\rm T}\mathsf{X}_{\rm s}$ (brown dashed line) becomes non-negligible. Complicating matters further is the removal of the anticorrelating contribution in $\Delta\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm s}$. Lastly, there are also noticeable correlations in the perturbed foreground removed terms. The thin red line shows how the removed term $\Delta\mathsf{S}\mathsf{U}_{\rm f}^{\rm T}\mathsf{X}_{\rm f}$ will also introduce a contribution to signal loss. This explains the previous result presented in the bottom-left panel of Fig. 2 where, despite only projecting out modes from pure foreground data $\mathsf{X}_{\rm f}$, signal loss still arose in the cleaned power spectrum, albeit at a ~ 5 per cent level in the largest scales. This will be caused by the signal perturbations $\mathbf{\Delta}$ which introduce a small correlation with the H I signal, shown by the red line in Fig. 3.

The complex mix of signal correlation and anticorrelation caused by the perturbations $\mathbf{\Delta}$ from non-foreground modes can clearly affect the overall signal loss. The impact of signal perturbations on the foreground modes becomes increasingly more important the more aggressive the foreground clean due to $\mathbf{\Delta}$ having more influence over less dominant foreground modes, as seen in Fig. 1. It is therefore crucial to model or emulate all these contributions in any signal reconstruction to avoid unbiased results as highlighted in previous literature (Switzer et al. 2015; Cheng et al. 2018). In Section 3 we will explore how signal injection can be used to construct a foreground transfer function and validate with simulations how it is able to successfully emulate all the subtle contributions to signal loss. We will also explicitly highlight cases where a transfer function can be incorrectly assembled such that some of the contributions demonstrated by Fig. 3 are not accounted for, leading to incorrect estimations of signal loss.

3 VALIDATING THE TRANSFER FUNCTION WITH SIMULATIONS

As demonstrated in the previous section, signal loss from blind foreground cleaning is complicated by the subtraction of subsets of data that have spurious correlations and anticorrelations with the H I signal. The spurious correlations arise because the estimated modes projected out in the blind foreground clean are perturbed by the presence of the H I signal itself. Thus the signal loss is dependent on the specific realization of the signal, foregrounds,

and their combination. Modelling the signal loss, or measuring it in pure simulations, is therefore potentially problematic and could lead to biased results. In this section, we explore how we can utilize the observed contaminated data itself to emulate the complex spurious correlations in injected mock data and use the signal loss experienced in the mocks to construct a foreground transfer function. This data-driven approach has been extensively used in single-dish intensity mapping detections (Masui et al. 2013; Anderson et al. 2018; Cunnington et al. 2022; Wolz et al. 2022). Here, we present the formalism for an unbiased application of the transfer function and for the first time validate its performance on simulations while also trying to expose any limitations.

Throughout this section, we treat the MD1GPC simulation as the *observed* intensity mapping data, and use lognormal mocks as completely separate simulations in the construction of the transfer functions. This maintains a certain independence between the injected mocks and the simulated signal that we are trying to recover, as would be the case in real observations. There is also the option of generating more complex mocks to inject into the data, for example using field level forward modelling (Obuljen et al. 2022) or an H I-halo prescription as trialled in Wolz et al. (2022); however, we found lognormal mocks sufficient for our purposes.

In this section, in cases where we are investigating the cleaned or reconstructed power spectrum, unless stated otherwise, we will use the cross-correlation power spectrum between the foreground-cleaned MD1GPC map and the H I-only (foreground-free) MD1GPC map. This is so that foreground residuals will be less of an issue and their additive bias does not confuse the investigation of signal loss and reconstruction accuracy. In cross-correlation with the H I-only map, any difference relative to the original H I-only autopower spectrum will be caused solely by signal loss.

3.1 Summarized recipe for the transfer function and its unbiased results

We begin by providing a summary of how a transfer function can be used for correcting signal loss from foreground cleaning, and validate the performance of the process. We then go into more detail in the remainder of this section, clarifying exactly how the transfer function can be constructed and used for various scenarios. In short, the foreground transfer function is constructed by injecting mock data into the observed maps. Then, by running the same foreground removal routine, one will subject the mock data to a similar signal loss that is experienced by the true underlying H I signal, thus giving an estimate of the true signal loss.

Below is the step-by-step recipe for how to construct and apply an unbiased transfer function;

(i) First, foreground clean the observed data $\mathsf{X}_{\rm obs}$ by projecting out $N_{\rm fg}$ PCA modes i.e. $\mathsf{X}_{\rm clean} = \mathsf{X}_{\rm obs} - \mathsf{U}\mathsf{S}\mathsf{U}^{\rm T}\mathsf{X}_{\rm obs}$, where U is a matrix of eigenvectors from the diagonalization of the ν, ν' covariance matrix estimated empirically from the data, and S is the selection matrix with ones along the first $N_{\rm fg}$ diagonal elements and zeros elsewhere.

(ii) Compute the power spectrum for the foreground-cleaned data, which will be negatively biased due to the signal loss from the foreground clean. The power spectrum is given here by $P_{\rm clean}(k) = \mathcal{P}(\mathsf{X}_{\rm clean}, \mathsf{X}_{\rm tr})$, where $\mathcal{P}(\mathsf{X}_{\rm clean}, \mathsf{X}_{\rm tr})$ is an operator which measures the cross-power spectrum between data sets $\mathsf{X}_{\rm clean}$ and $\mathsf{X}_{\rm tr}$, then reduces the power into the spherically averaged k-bins. Here, $\mathsf{X}_{\rm tr}$ is any overlapping tracer, which can be the intensity map itself for an autocorrelation survey, or a galaxy survey as a common example of cross-correlation.

(iii) Generate mock H I signal maps $\mathsf{X}_{\rm m}$ with the same dimensions as the observed data. In this work we use a fast lognormal transform process to generate mocks from the H I power spectrum model given in equation (B4). We investigate the consequences of variation in the input mocks in Section 5.

(iv) Emulate signal loss in the mock by injecting it into the real data and foreground cleaning the observed data and mock combination,

$$\mathsf{X}^{\rm m}_{\rm clean} = (\mathsf{X}_{\rm obs} + \mathsf{X}_{\rm m}) - \mathsf{U}_{\rm obs+m}\mathsf{S}\mathsf{U}^{\rm T}_{\rm obs+m}(\mathsf{X}_{\rm obs} + \mathsf{X}_{\rm m}) - [\mathsf{X}_{\rm clean}]. \quad (10)$$

The term in the square brackets is subtracting the cleaned observed data without mocks to remove contributions in the map uncorrelated to the mock signal thus reducing the variance of the transfer function, as we will explicitly demonstrate.

(v) The foreground transfer function is then given by

$$\mathcal{T}(k) = \left\langle \frac{\mathcal{P}(\mathsf{X}^{\rm m}_{\rm clean}, \mathsf{X}_{\rm m})}{\mathcal{P}(\mathsf{X}_{\rm m}, \mathsf{X}_{\rm m})} \right\rangle_{N_{\rm mock}}, \quad (11)$$

where the angled brackets denote an averaging over iterations of a suitably large number of mocks ($N_{\rm mock}$) until a converged transfer function is achieved. We use $N_{\rm mock} = 100$ by default unless otherwise mentioned.

(vi) De-bias the cleaned power spectrum using the transfer function to reconstruct the signal loss with $P_{\rm rec}(k) = P_{\rm clean}(k)[\mathcal{T}(k)]^{-1}$. Note the index of -1 should also be used in autocorrelation i.e. an autocorrelation of an intensity map should *not* have signal loss corrected for twice, as we will demonstrate in Section 3.4.

(vii) The covariance of the reconstructed power spectrum can also be extracted from the mocks used in the transfer function calculation. While the mean of $P_{\rm clean}\mathcal{T}_i^{-1}$ over all $N_{\rm mock}$ iterations provides the reconstructed power spectrum, the covariance estimates the errors inclusive of signal loss uncertainty. However, as we will show, it is crucial not to subtract the square-bracket $\mathsf{X}_{\rm clean}$ term in equation (10) when estimating the covariance, as this will include foreground residuals, instrumental noise, etc. all of which should contribute to the error budget. We discuss this in detail in Section 3.3.

The numerator in equation (11) is taking the cross-correlation between the cleaned mock $\mathsf{X}^{\rm m}_{\rm clean}$ and original mock $\mathsf{X}_{\rm m}$ with no cleaning effects. This should therefore not be overly influenced by foreground residuals and differences between this cross-correlation and the autocorrelation in the denominator should only be caused by signal loss from the foreground clean, thus their ratio provides the level of the original signal remaining in the power spectrum of the cleaned mock $\mathsf{X}^{\rm m}_{\rm clean}$. The crucial part for equation (11) is having a process for obtaining $\mathsf{X}^{\rm m}_{\rm clean}$ such that the signal loss it experiences across all scales is the same as the signal loss in the actual data $\mathsf{X}_{\rm s}$. To achieve this we inject mock signal $\mathsf{X}_{\rm m}$ into the observed data with foregrounds and true H I signal ($\mathsf{X}_{\rm f+s+m} \equiv \mathsf{X}_{\rm f} + \mathsf{X}_{\rm s} + \mathsf{X}_{\rm m}$) then project out the same number of modes as in the original foreground clean of the observations i.e.;

$$\mathsf{X}^{\rm m}_{\rm clean} = \mathsf{X}_{\rm f+s+m} - \mathsf{U}_{\rm f+s+m}\mathsf{S}\mathsf{U}^{\rm T}_{\rm f+s+m}\mathsf{X}_{\rm f+s+m} - [\mathsf{X}_{\rm clean}]. \quad (12)$$

This is equivalent to what we presented in the summarized recipe in equation (10). As discussed, the term in the square bracket is subtracting the cleaned observed data (with no mock injection) to reduce the transfer function variance, which we discuss in more detail later. The presence of mock signal will cause perturbations to the eigenmodes, and will emulate the signal loss coming from both projecting out the modes with signal perturbations and the complex correlations between all the cross terms discussed in Section 2.2 and equation (9).

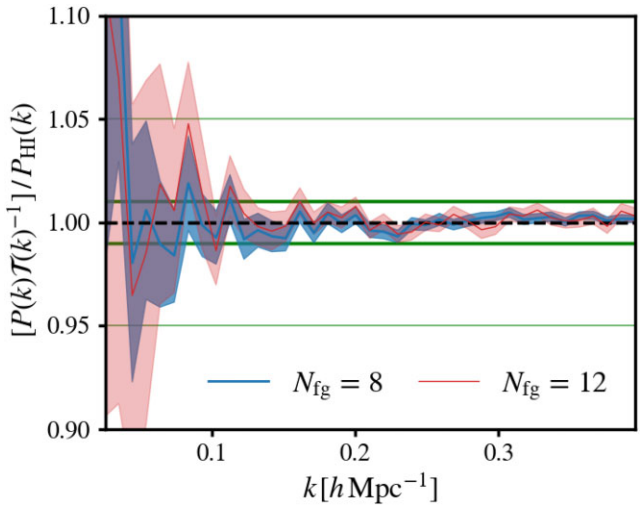

Figure 4. Accuracy of reconstructed foreground-cleaned power spectra relative to the foreground-free H I-only data ($P_{\rm HI}$) from simulated intensity maps. The foreground-cleaned power spectrum has been reconstructed using the transfer function to correct for the signal loss from foreground cleaning. The transfer functions are calculated using equations (11) and (12) and averaging over 100 lognormal mocks. The shaded bands show the rms over these 100 mocks. Results for a mild ($N_{\rm fg} = 8$, blue lines) and more aggressive ($N_{\rm fg} = 12$, red lines) foreground cleans are shown. The dark thick (light-thin) green horizontal lines indicate sub 1 per cent (5 per cent) accuracy regions of the reconstructed power spectrum.

The presence of the true observed H I signal $\mathsf{X}_{\rm s}$ in $\mathsf{X}_{\rm f+s+m} \equiv \mathsf{X}_{\rm f} + \mathsf{X}_{\rm s} + \mathsf{X}_{\rm m}$ in equation (12) creates unwanted complications to the transfer function construction which should ideally only be concerned with the mock signal $\mathsf{X}_{\rm m}$ and its relationship with $\mathsf{X}_{\rm f}$. The presence of $\mathsf{X}_{\rm s}$ will not matter from a *direct* signal loss perspective, since this will not affect the cross-correlation with $\mathsf{X}_{\rm m}$ in equation (11). However, $\mathsf{X}_{\rm s}$ will perturb the estimation of the eigenvectors. This is unwanted because we ideally only want the mock signal to perturb the eigenvectors and just produce $\mathsf{U}_{\rm f+m}$, but by injecting mocks into the true signal we will actually measure

$$\mathsf{U}_{\rm f+s+m} = \mathsf{U}_{\rm f} + \mathbf{\Delta}_{\rm s} + \mathbf{\Delta}_{\rm m}, \tag{13}$$

where we have introduced the subscripts m and s to the perturbations $\mathbf{\Delta}$ to distinguish perturbations from mock signal and true H I signal, respectively. The two sources of perturbation is not seen in the foreground clean of just the observations ($\mathsf{X}_{\rm f} + \mathsf{X}_{\rm s}$). In other words the eigenvectors are now being perturbed twice. As we will show from our results shortly, this appears to have little impact and we still obtain an unbiased transfer function. We tested the transfer function in an idealized case where mock signal was injected into just pure foreground ($\mathsf{X}_{\rm f} + \mathsf{X}_{\rm m}$) and found little difference in performance compared to the realistic case where true signal is present ($\mathsf{X}_{\rm f} + \mathsf{X}_{\rm s} + \mathsf{X}_{\rm m}$).

The accuracy of the reconstructed power spectra is demonstrated in Fig. 4. The simulated observations are cleaned by removing either $N_{\rm fg} = 8$ or 12 PCA modes, then the transfer function is used to correct for the signal loss, with the reconstructed result being divided by the original foreground-free simulation $P_{\rm HI}(k)$. Thus, a perfect reconstruction would give unity across all scales. We see excellent performance with sub-per cent accuracy achieved across most scales above $k > 0.1\,h\,{\rm Mpc}^{-1}$ for the $N_{\rm fg} = 8$ case. Performance is still good for the $N_{\rm fg} = 12$ case, albeit with a noticeable drop in accuracy relative to $N_{\rm fg} = 8$ mostly at large scales (small-k). This will be caused by the increased effect from spurious correlations between foregrounds and mock signal, which, as we demonstrated in Fig. 3, increases for more aggressive (higher $N_{\rm fg}$) foreground cleans. This will not necessarily bias the results since the variance in the transfer function also increases for higher $N_{\rm fg}$, as shown by the shaded regions, thus can be reflected in the error estimations (discussed in a later section).

In general on large ($k < 0.1\,h\,{\rm Mpc}^{-1}$) scales, we see a less reliable result in terms of pure accuracy, but this is also accounted for by the transfer function variance, which can reach $\gtrsim 5$ per cent on these scales. The performance at large scales is however dependent on the size of the intensity mapping survey. The depth of the $1\,({\rm Gpc}/h)^3$ MD1GPC simulation at $z \sim 0.39$ is reasonably consistent with a MeerKAT L-band survey, assuming it uses the complete band range ($0.2 < z < 0.58$). However, future surveys in UHF band, and then eventually the SKAO, will cover much wider frequency ranges. This will mean reconstructed modes at $k < 0.1\,h\,{\rm Mpc}^{-1}$ become more reliable due to a suppression of sample variance and less signal loss which will now be contained to even larger scales. We will demonstrate this point later in Section 5 with some additional simulations which cover a larger volume.

The presence of the true observed H I signal in the transfer function calculation will increase its variance because there will be residual true signal after the foreground clean. This will be uncorrelated from the mocks and act like noise and increase the variance across all of the mocks being averaged over. This is why we subtract the cleaned data (the $\mathsf{X}_{\rm clean}$ term in the square brackets of equation 12), since this is only contributing variance to the result. We will revisit this point in the next section where we will demonstrate that the increased variance coming from $\mathsf{X}_{\rm clean}$ can be utilized for error estimation. With the $\mathsf{X}_{\rm clean}$ subtraction, this version of the transfer function is not only achieving a good accuracy but also a good uncertainty on most scales, shown by the shaded region.

The validation of the transfer function demonstrated by Fig. 4 is an important result. This is a method for reconstructing signal loss which is applicable on real data and delivers unbiased results across all scales and within sub-per cent precision across smaller scales where the particular survey volume allows those modes to be well sampled ($k > 0.1\,h\,{\rm Mpc}^{-1}$ for the case of the MD1GPC simulations). The compromise of having to inject mock signal into a combination of both foreground and true H I signal is an unavoidable complication; however, there is no evidence that this causes any bias in the reconstructed power spectrum. Furthermore, by subtracting the cleaned data ($\mathsf{X}_{\rm clean}$ term in the square brackets of equation 10), we found the increase in variance relative to an ideal case where no true signal ($\mathsf{X}_{\rm s}$) is present in the transfer function calculation was only ~ 20 per cent. It is crucial that the form of the transfer function as defined by equations (11) and (10) be followed and in Appendix C we explicitly highlight the consequences of deviating from this prescription, demonstrating the significant biases caused when different definitions of the transfer function are used.

In Figs 5 and 6 we demonstrate the *shape* of the transfer function in k-space and in doing so analyse where signal loss is most severe. Fig. 5 shows transfer functions for different PCA modes removed (given by $N_{\rm fg}$). This confirms that signal loss increases with $N_{\rm fg}$ and is higher at smaller-k, both as expected. We also show the impact from adding the dominant instrumental noise. Perhaps counter-intuitively, this causes less signal loss. This is because the noise is the dominant source of perturbations to the pure foreground modes (as shown by Fig. 1), hence these noise-dominant modes will have less contribution from the H I signal and removing them causes less signal loss. However, this will result in a poorer overall foreground clean. Thus

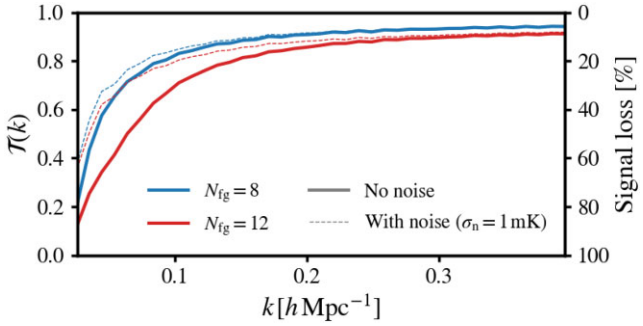

Figure 5. Foreground transfer functions $\mathcal{T}(k)$ constructed using equations (11) and (10) for the MD1GPC simulations for different numbers of PCA modes removed (given by $N_{\rm fg}$). The solid lines indicate noise-free simulations; the thin dashed lines are for cases where white noise with rms $\sigma_{\rm n} = 1$ mK, which dominates over the H I, is added to the simulated observations.

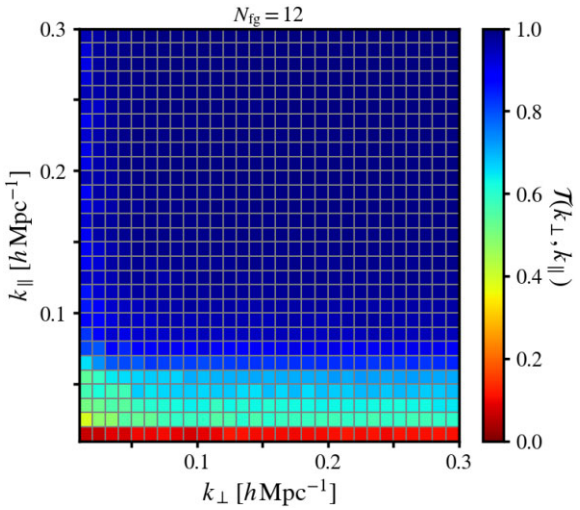

Figure 6. Similar to Fig. 5, this shows the foreground transfer function but now in cylindrical $k_\perp$, $k_\parallel$ space where $N_{\rm fg} = 12$ PCA modes have been removed. This is for the noise-free case.

foreground residuals and the high-level noise already present will cause problems for additive biases in autocorrelation and would also lead to higher errors in cross-correlation. So the presence of high noise is not beneficial as Fig. 5 alone may appear to suggest. We explore the high noise scenario further in Section 4 where we use simulations which emulate an early pathfinder-like intensity mapping survey.

Similarly to Fig. 5, an example transfer function is also shown in Fig. 6 but now decomposed into cylindrical contributions in $k_\perp$, $k_\parallel$. As already well established, signal loss is overwhelmingly a function of small-$k_\parallel$. However, there is also some slight $k_\perp$ dependence with signal loss being slightly higher at small-$k_\perp$ caused by the large angular structures in the foregrounds. It is important to highlight that the nature of signal loss will vary depending on not just the foreground's strength and spectral smoothness, but also on the depth of the survey in frequency. This means that the signal loss presented in Figs 5 and 6 is specific to the MD1GPC simulation. However, the conclusions we have drawn from this are still mostly generic. For example, signal loss is still contained in small-$k_\parallel$ modes in more systematic dominated intensity maps as shown in recent MeerKAT analysis (Cunnington et al. 2022) and as we will show later in the simulations designed to emulate a small MeerKAT pilot intensity mapping survey. While signal loss appears widespread throughout all modes in the spherically averaged Fig. 5, where >5 per cent signal loss is evident even on small scales, it is clear from Fig. 6 that signal loss does tend to zero (where $\mathcal{T}(k_\perp, k_\parallel) \sim 1$) above some $k_\parallel$ cut. This raises an intriguing possibility of adopting a hybrid foreground cleaning/avoidance strategy where a blind foreground clean is run on the full data, but then an avoidance strategy is used where only modes above some $k_\parallel$ are kept for further analysis. At the very least this would limit the dependence on the transfer function but would unlikely be reliable enough to completely avoid any use of signal reconstruction. Furthermore, the methodology of the transfer function would still be a required tool for robustly assessing where an optimum cut in $k_\parallel$ should be made. Scale cuts will also limit the scope and constraints possible with the experiment. We defer any investigations of $k_\parallel$ cuts to future work and continue to test the transfer function on small-$k_\parallel$, especially since these are the scales that will test the performance of the transfer function most stringently.

3.2 1D versus 2D bandpowers in the transfer function

Our results so far have reduced all power directly into 1D spherically averaged k-bins and the results in Fig. 4 suggest this can be sufficient. Provided the same k-bins are used in the 1D spherical averaging for the transfer function construction and cleaned power spectrum, the transfer function should encapsulate the same anisotropic signal loss in each k-bin as inflicted on the cleaned data. Furthermore, going straight to 1D k-bins avoids extra compression steps which could potentially lead to results being lossier. However, it has been proposed in the literature that because the signal loss is anisotropic (demonstrated by Fig. 6) the transfer function should be estimated and applied in 2D cylindrical $k_\perp$, $k_\parallel$ space, with these bandpowers then re-binned to provide the final spherically averaged 1D power spectrum (Masui et al. 2013; Switzer et al. 2015).

We tested the 2D-cylindrical transfer function approach and found evidence of higher variance in the standard case where complicated polarized foregrounds are present in the observations. We demonstrate this in Fig. 7. Here, the 1D reconstruction (blue shading) shows our default set-up used everywhere else in the paper, averaging straight into 1D spherical k-bins. The 2D reconstruction refers to a case where we average all power into $100\,k_\perp \times 100\,k_\parallel$ cylindrical linear-spaced bins, with $0 < k_\perp, k_\parallel < 0.4\,h\,{\rm Mpc}^{-1}$, in the transfer function construction. The measured power for the cleaned observations is also reduced into the same 2D bins and a reconstructed power spectrum for a single ith mock iteration is given by $P_{{\rm rec},i}(k_\perp, k_\parallel) = P_{\rm clean}(k_\perp, k_\parallel)/\mathcal{T}_i(k_\perp, k_\parallel)$. The 2D powers then undergo a weighted average into 1D k-bins to give the rebinned 1D power spectrum, defined by

$$P_{{\rm rec},i}(k) = \frac{\sum_\alpha N_\alpha P_{{\rm rec},i}(k_{\perp,\alpha}, k_{\parallel,\alpha})}{\sum_\alpha N_\alpha}, \tag{14}$$

where all unique 2D $k_\perp$, $k_\parallel$ bandpowers are indexed by α and the summation is over all 2D powers contained in the 1D bin $k \equiv \sqrt{k_\perp^2 + k_\parallel^2} \in (k - \Delta k/2, k + \Delta k/2)$. N_α is the number of 3D Fourier modes contained in the particular 2D $k_{\perp,\alpha}$, $k_{\parallel,\alpha}$ bandpower. Similar to the direct 1D reconstruction, the mean over all ith mocks in $P_{{\rm rec},i}(k)$ gives the final estimated reconstructed power spectrum, and the variance provides an estimate of the expected errors. The results for the 2D rebinned power are shown by the red-hatched shading in Fig. 7, where the large increase in variance is clear. We found by switching off the complexity caused to the foregrounds by the simulated polarization leakage made the 2D transfer function more

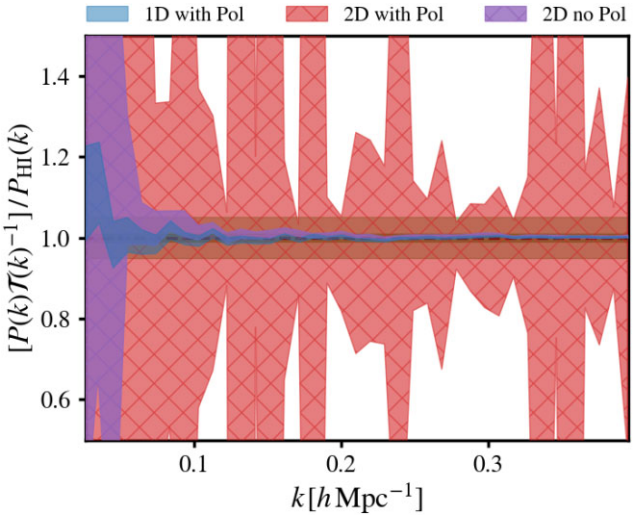

Figure 7. Demonstration of the large increase in the variance of the reconstructed power spectrum, when the transfer function is constructed and applied in 2D $k_\perp$, $k_\parallel$ space. The blue results show the standard 1D cases (same as Fig. 4) with $N_{\rm fg} = 8$. The red-hatched results show the 2D construction following steps in Section 3.2. The purple-hatched results also show results with a 2D construction, but with simplified foregrounds excluding the polarization leakage.

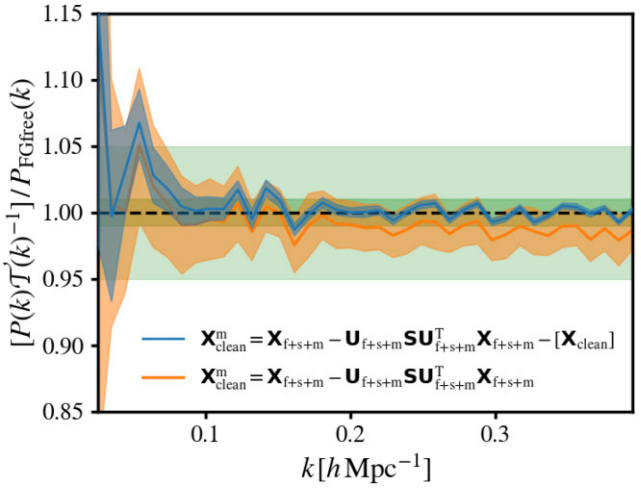

Figure 8. Impact from subtracting the cleaned observational data $\mathsf{X}_{\rm clean}$ in equation (10). The shaded regions represent the rms over the 100 mocks used to construct the transfer function. The rms is significantly reduced when the cleaned data are subtracted. This test was done using a foreground clean with $N_{\rm fg} = 8$ on simulations with high instrumental noise. Here, the reconstructed result is divided by the foreground-free power spectrum $P_{\rm FGfree}$ which contains identical instrumental noise.

reliable (purple hatched results), although still a higher variance is returned in these results, particularly at small-k, relative to the 1D case.

It is likely that outliers in isolated iterations are causing the large variance in Fig. 7. These outliers would then be suppressed in the simpler case where there is no polarization leakage, thus less complex residuals or signal loss to cause extreme spurious correlations in the mocks. In an attempt to suppress the variance, we increased the number of mocks used in the 2D polarization leakage construction to 500, but this yielded no improvement. An extension that may be necessary to avoid the blow-up in variance is a more detailed weighting to the rebinning procedure we perform in equation (14). In Switzer et al. (2015) they discuss applying an inverse covariance weighting to maximize the 1D signal-to-noise ratio. This could down-weight some of the outliers in our iterations and suppress the large variance in the 2D polarized results. We defer this extension to future work, ideally with even more realistic sims where a conclusive study can be performed into whether the 2D transfer function construction can be more optimal. Since our results suggest that direct 1D construction is better performing, at least in the case of the spherically averaged power spectrum, we use this approach for the rest of the paper, unless presenting a 2D cylindrical transfer functions which we now only do for demonstration purposes.

3.3 Error estimation for power spectra reconstructed with a foreground transfer function

Evaluating how to correctly estimate the contributions to the error from foreground contamination and signal loss uncertainty will be crucial for future precision cosmology with H I intensity mapping. This is the focus of this section. An approach taken in some previous intensity mapping detections (Anderson et al. 2018) has been to use the variance over the mock simulations used in the transfer function construction for the error estimate on cross-correlation measurements. It is possible to capture this uncertainty from the variance in the transfer function i.e. the errors can be estimated using

$$\hat{\sigma}_{P_{\rm c}} = \sigma\{P_{\rm clean}\mathcal{T}_i^{-1}\}, \tag{15}$$

where $\mathcal{T}_i$ is the transfer function from the ith mock in the construction, and $\sigma\{\}$ is taking the rms over all i iterations. The rms over all transfer function iterations which include the injected true observations should provide an error estimate which incorporates thermal noise, foreground residuals, residual RFI, sample variance, and signal loss from foreground cleaning. However, crucially this approach relies on modifying the transfer function definition so that the $\mathsf{X}_{\rm clean}$ term in equation (10) is not subtracted.

We begin by demonstrating the impact subtracting the cleaned observed data $\mathsf{X}_{\rm clean}$ has on the transfer function. Since we are investigating error estimation, for this section it is helpful to use data with noticeable error-bar size, so we therefore add the dominant $\sigma_{\rm n} = 1$ mK noise to the MD1GPC simulations. Fig. 8 shows the performance of the transfer function for the high-noise simulations, calculated using equations (11) and (10), both with and without the $\mathsf{X}_{\rm clean}$ subtraction. It is encouraging to see that the addition of noise is not majorly affecting the performance of the transfer function. For the case where $\mathsf{X}_{\rm clean}$ has been subtracted (blue results), the accuracy is only mildly affected relative to the noise-free results in Fig. 4. For the noise-inclusive results of Fig. 8, we divide by $P_{\rm FGfree}$ in the y-axis which contains the same noise as the reconstructed power. This is to divide out the fluctuations caused by the presence of the dominant noise, allowing analysis into the performance of the reconstruction alone. For the case where $\mathsf{X}_{\rm clean}$ has not been subtracted (orange results), there is a slight drop in accuracy at small scales. This small bias is absent when we use an H I signal without RSD thus it appears to be caused by the addition of uniform noise in the presence of an anisotropic H I signal. We also found this small bias is decreased when we use the 2D transfer function construction outlined in Section 3.2, although this relied on there being no polarization leakage which otherwise causes the variance of the result to blow up, as we showed. We discuss the performance of the transfer function in the presence of anisotropic phenomena later, but given this is a small bias and is absent in the subtracted

$\mathsf{X}_{\rm clean}$ case, it is not overly important for this discussion on error covariance estimation.

Fig. 8 suggests that subtracting $\mathsf{X}_{\rm clean}$ is the favourable strategy if one is purely pursuing the most accurate transfer function possible. However, the main point of this plot is the difference in variance between the two cases. If one is using the variance on the transfer function as a basis for the error estimation, the reduced variance caused by subtracting $\mathsf{X}_{\rm clean}$ has implications and can lead to underestimated errors on the reconstructed power spectrum, as we will now demonstrate.

To quantitatively evaluate error estimation performance, it is helpful to analyse how errors for a power spectrum measurement are analytically derived in a foreground-free case. The variance on a cross-correlation power spectrum $P_{\rm c}$ between two tracers, 1 and 2, can be estimated as (Feldman, Kaiser & Peacock 1994)

$$\sigma^2_{\rm theory} = \frac{1}{2N_{\rm m}}\left[P_{\rm c}^2(k) + \left(P_1 + \frac{V\sigma_1^2}{\langle f_1\rangle^2}\right)\left(P_2 + \frac{V\sigma_2^2}{\langle f_2\rangle^2}\right)\right], \qquad (16)$$

where $N_{\rm m}$ is the number of modes spherically averaged in each k-bin, V is the volume of a single voxel on the Fourier grid, uncorrelated noise in the field is represented by the variance σ^2, which for an ideal intensity map would be the variance of the instrumental noise, and $\langle f\rangle$ is the background mean for the field e.g. the mean brightness temperature for intensity mapping. For galaxy surveys, the noise component given by the second terms in the curved brackets will reduce to shot-noise i.e. $V\sigma^2/\langle f\rangle^2 = 1/\bar{n}_{\rm g}$, where $\bar{n}_{\rm g}$ is the galaxy number density. We refer the reader to Blake (2019) where a detailed derivation of the above is provided with applications to intensity mapping and its cross-correlations with galaxy surveys.

Extending equation (16) to incorporate contributions from foreground contamination is challenging. Foreground residuals could presumably be estimated and would provide an additive variance, or be assumed sub-dominant enough not to warrant inclusion. However, the uncertainty from the transfer function, which for high signal loss is non-negligible at large scales, requires careful inclusion. The uncertainty in the transfer function can be estimated from the variance over the mocks used to construct it as we have shown; however, analytically adding this into the error budget of equation (16) is non-trivial since this will not necessarily be a contribution entirely independent of the noise and cosmic variance already being factored for in equation (16). This is why some previous analyses have used the transfer function variance as a basis for overall error estimation. To evaluate whether this is a robust method, we can use the analytical errors as a benchmark. Errors estimated based on the variance in the transfer function should approximately agree on small scales with the analytical ones where foreground contamination and signal loss are minimal, but the large noise still dominates.

We validated that the analytical errors (equation 16) are a good estimate for the foreground-free case. Using the MD1GPC simulations with the large $\sigma_{\rm n} = 1$ mK Gaussian white-noise but with no foregrounds, we measure the cross-correlation power spectrum with a noise-free equivalent, then estimate the errors using equation (16) and evaluate the $\chi^2_{\rm dof}$ given by

$$\chi^2_{\rm dof} = \sum_k \frac{P_{\rm data}(k) - P_{\rm mod}(k)}{\sigma_P(k)}\Big/(N_{\rm k} - 1), \qquad (17)$$

where $N_{\rm k}$ is the number of k-bins. $P_{\rm mod}$ is the model defined in Appendix B1 with parameters matched to the MULTI-DARK inputs used in the MD1GPC simulation. The analytical errors return a $\chi^2_{\rm dof} \sim 1$ as expected, evidence that the errors are a reasonable size, given the reliable model.

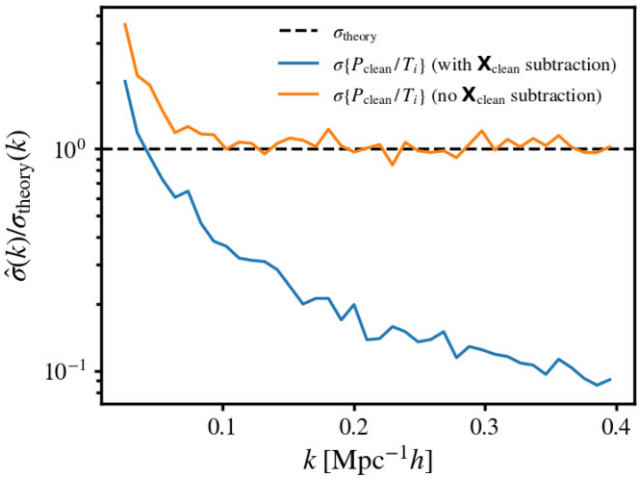

Figure 9. Comparison between methods of error estimation for simulations with signal loss reconstructed by a transfer function. The black dashed line shows an analytical error estimate given by equation (16) which we have shown to be reliable for foreground-free data. The coloured lines show error estimation based on the variance of the transfer function, under two difference scenarios, with $\mathsf{X}_{\rm clean}$ subtraction (see equation 10) and without. This is with the MD1GPC simulation where $N_{\rm fg} = 8$ PCA modes removed.

Using the analytical errors $\sigma_{\rm theory}$ as a validated benchmark, Fig. 9 shows how the error estimation based on the variance in the transfer function compares for the same simulations but with cleaned foregrounds. The blue line shows the case using a transfer function defined by equations (11) and (10), where the cleaned observed data ($\mathsf{X}_{\rm clean}$) have been subtracted to reduce the variance. This is underestimating the errors relative to the analytical ones, likely caused by subtracting the $\mathsf{X}_{\rm clean}$ term which will contain noise and foreground residuals, which contribute to the error budget. Fig. 9 therefore shows that subtracting the cleaned data $\mathsf{X}_{\rm clean}$ in the transfer function construction results in the transfer function variance no longer being a reliable means for estimating the errors. However, it is clear from the orange line, which is equivalent to the blue but without the $\mathsf{X}_{\rm clean}$ subtraction, that the variance from this version of the transfer function leads to an excellent agreement with the analytical errors at high-k as required. Furthermore, it also shows an increase in error at small-k as one would expect where signal loss is highest, thus it is incorporating the increased uncertainty from signal reconstruction, not accounted for in the analytical errors.

Using the variance in the transfer function mocks for analysing uncertainties also has the advantage of being able to examine off-diagonal covariance of the data, something not trivially possible with the analytical approach. In Fig. 10 we show the $k_i k_j$ covariance matrix C_{ij} (top row) as well as the normalized correlation matrix defined by $R_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$. The left column shows the foreground-free scenario where we inject mocks into the H I + high-noise simulations to get an estimate of the covariance without a foreground cleaning step. The right column is equivalent but with cleaned foregrounds and a transfer function constructed without the subtraction of $\mathsf{X}_{\rm clean}$. The covariance over all iterations in the reconstructed power spectra (equation 15, but now including off-diagonal elements $i \neq j$) estimates the covariance of the observed data. As expected, and consistent with the orange line of Fig. 9, the cleaned foregrounds are increasing covariance on large scales, but encouragingly they do not appear to increase off-diagonal correlations between k-bins.

We conclude from this investigation that the variance in the transfer function is a reliable tool for error estimation in the final

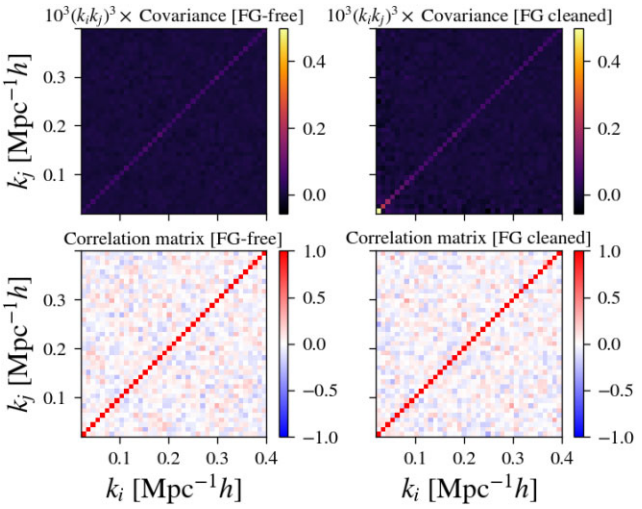

Figure 10. Covariance and correlation matrices for foreground-free (left column) and foreground cleaned with transfer function reconstructed signal loss (right column), without the $\mathsf{X}_{\rm clean}$ subtraction in the transfer function. Produced using the MD1GPC simulations with $N_{\rm fg} = 8$ PCA modes removed, as in Fig. 9. The covariance matrices have been multiplied by the product of the two-k-bins cubed times 10^3 for demonstration purposes so high-k covariance can be seen.

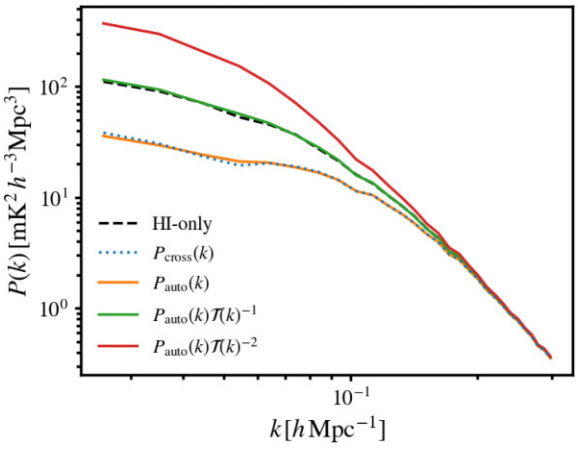

Figure 11. Demonstrating the correct application of the transfer function in an autocorrelation analysis. The black dashed line shows the original foreground-free power spectrum. The blue dotted line shows the cross-correlation with the foreground-free. The orange line shows the autocorrelation. For this aggressive ($N_{\rm fg} = 12$) foreground clean, foreground residuals should be minimal and the similar amplitude between $P_{\rm auto}$ and $P_{\rm cross}$ suggests the signal loss is similar in both. The green line shows the correct application of the transfer function and the red line shows the overcorrection where $\mathcal{T}^{-2}$ is used.

reconstructed power spectrum, provided the cleaned data $\mathsf{X}_{\rm clean}$ term is not subtracted in its construction. This becomes similar in approach to galaxy surveys which use vast suites of mocks as their primary method for estimating the covariance in their data (e.g. Zhao et al. 2021). If opting to use the transfer function variance for error estimation, it becomes important to ensure that all aspects of the error budget are emulated in the mocks used in the transfer function construction. An example of this would be in a galaxy survey cross-correlation where the galaxy shot noise would need to be captured using galaxy mocks with the correct number densities and survey coverage. With the observational data injected into the mocks, we are also including variance from signal loss, foreground residuals, residual RFI, instrumental noise, etc. all of which are currently not well enough understood to reliably emulate in mock intensity maps.[5]

3.4 Autocorrelation and cross-correlation applications

So far in this section, we have considered the cross-power spectrum between a foreground-cleaned H I intensity map and the original foreground-free, H I-only map. This is so that any foreground residuals or noise in the cleaned maps do not complicate the analysis of H I signal loss in the power spectra. Since foreground residuals and noise will not correlate with the H I-only maps, the additive biases they cause are avoided in cross-correlation; thus the only departure from the true-H I power should be just from signal loss. However, H I intensity mapping surveys will also aim to conduct analysis in autocorrelation and we need to consider how signal loss behaves in this scenario.

It has been previously suggested that there would be twice as much signal lost in the autocorrelation power spectrum because its effects are present twice in the map product $P_{\rm HI}(\boldsymbol{k}) \propto |\tilde{\mathsf{X}}(\boldsymbol{k})|^2$. This would mean a $\mathcal{T}(k)^{-2}$ correction factor is needed to reconstruct the power spectrum. However, we found from our simulations that this is not the case, and the same degree of signal loss is also present in an autocorrelation as is in cross-correlation. In other words, the same correction of $\mathcal{T}(k)^{-1}$ is also needed in the autocorrelation as well as in cross-correlation. Fig. 11 demonstrates this finding showing how the cross-correlation (blue dotted line) and autocorrelation (orange-solid) appear to have approximately equivalent levels of signal loss. For this test, we return to the noise-free simulations and use an aggressive $N_{\rm fg} = 12$ PCA clean which will suppress foreground residuals significantly, making it reasonable to ignore their influence on the results. The green line shows what appears to be the correct application of the $\mathcal{T}(k)^{-1}$ transfer function, whereas the red line shows the consequential overcorrection from applying the $\mathcal{T}(k)^{-2}$ to the autocorrelation.

To provide a deeper understanding for why signal loss to the power spectrum is the same for autocorrelation and cross-correlations, we present in Fig. 12 the amplitude of all 3D-Fourier mode products for a randomly chosen k-bin. The spherically averaged $P(k)$ value for the chosen $0.0590 < k < 0.0687$ h Mpc^{-1} bin is then simply the average of all these amplitudes, which is stated in the legend for each scenario. The top panel shows the comparison between an autocorrelation with foreground cleaning and the cross-correlation between foreground-cleaned and foreground-free (H I-only) data. As can be seen, the average of the modes is approximately the same in both cases and thus consistent with Fig. 11, demonstrating that signal loss is equivalent in autocorrelation and cross-correlation. The reason for this is related to the fact that the same modes are projected out of the analysis in both cases and signal loss does not compound when two maps with the same removed modes are combined in an autocorrelation. We confirm this to be the case in the bottom panel of Fig. 12 where we use simulations with a foreground from a different region of sky so that we produce a cleaned map $\mathsf{X}^{\rm clean}_{\rm FG2}$ which will have different foreground modes removed compared to the original $\mathsf{X}^{\rm clean}_{\rm FG1}$ used in the rest of the paper for the MD1GPC simulation. The exact regions are not overly important just the fact that they will generate a different set of modes which are projected out in

[5]Jackknife resampling will also be a useful tool when unknown systematics are present.

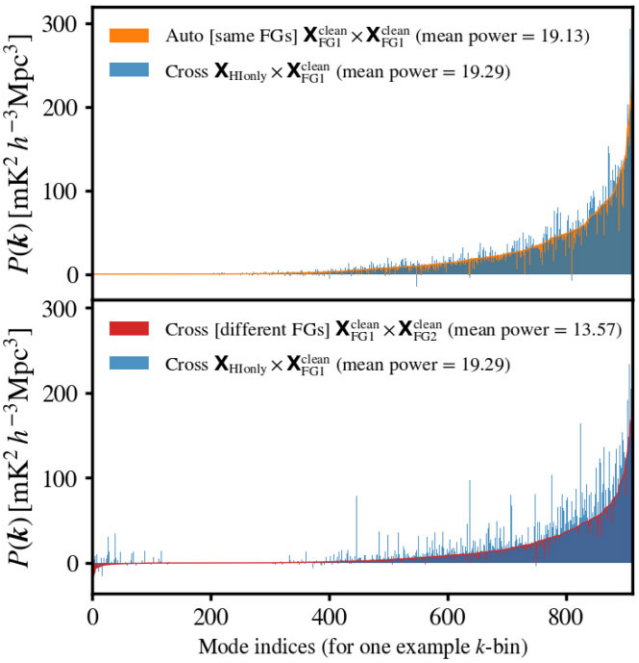

Figure 12. Amplitude of all Fourier modes $P(\boldsymbol{k}) \propto \mathrm{Re}\left\{\tilde{\mathrm{X}}_{\rm A}(\boldsymbol{k}) \cdot \tilde{\mathrm{X}}_{\rm B}^{*}(\boldsymbol{k})\right\}$ in the range $0.0590 < k < 0.0687$ h Mpc^{-1}, which corresponds to one chosen k-bin in the spherically averaged power spectrum. The average of these modes, stated in the legend for each scenario, will represent the spherically averaged power spectrum value for the particular k-bin. The top panel shows the equivalence in signal-loss between autocorrelation and cross-correlation. The lower panel shows how signal loss can be larger where a different set of modes has been projected out in the foreground clean shown by the red results which is the cross-correlation between two different simulated foreground regions.

the foreground clean. When these two foreground-cleaned maps are cross-correlated (red results) we now get a drop in power relative to the cross-correlation between $\mathrm{X}_{\rm FG1}^{\rm clean}$ and the H I-only map (see the mean power in the legend) showing that the signal loss is related to the modes being projected out in the foreground clean, and it is only a difference in these which will create further signal loss in an H I autocorrelation.

The results demonstrated by Fig. 12 have consequences for autocorrelation analyses with H I intensity mapping. Not just because it further confirms that signal loss should be the same in autocorrelation and cross-correlation where the same foreground modes are projected out, but also because the results in the bottom panel show when two differently cleaned maps are cross-correlated, the signal loss becomes more complex to estimate. This is relatable to a method that is likely to be pursued when attempting an autocorrelation detection whereby cross-correlations are measured between different subsets of the observations, created either by splitting data into different time-blocks (sub-seasons) as done in GBT analysis (Masui et al. 2013; Wolz et al. 2022), or by splitting different dishes as is possible with a multidish telescope such as MeerKAT. This is pursued in order to avoid the additive biases from noise and time- or dish-dependent systematics. While this method would still observe the same foreground, the response to systematics may be different in each subset creating a scenario similar to the red results in Fig. 12. We leave further investigation into this specific form of autocorrelation method for future dedicated studies.

4 APPLICATIONS TO PATHFINDER INTENSITY MAPPING WITH MEERKLASS

In the previous sections, the MD1GPC simulations used have been deliberately kept free of further observational effects (except for a couple of identified cases) besides the foreground contamination which included simulated polarization leakage. However, a relevant question for current pathfinder single-dish experiments is whether the early pilot surveys, which typically have low signal-to-noise and additional systematic observational effects, can also rely on the foreground transfer function to correct for signal loss. In these pathfinder surveys (e.g. Cunnington et al. 2022; Wolz et al. 2022) foreground cleaning is typically aggressive and signal loss can reach high levels, thus one could argue that we are more reliant on signal reconstruction in these early surveys, compared with future surveys where foreground cleaning and systematics will be more controlled.

The cosmological detection in Cunnington et al. (2022) (like all other intensity mapping detections preceding it) relied on a foreground transfer function to reconstruct the signal loss from foreground cleaning. The 7.7σ cross-correlation detection significance fell to $\sim 4\sigma$ without signal reconstruction. The survey covered just 200 deg^2 and only the frequency channels spanning 1015–973 MHz ($0.4 < z < 0.46$) were used to avoid the worst RFI. Furthermore, the observation gathered just 10.5 hours of data per dish. This means sky coverage and signal-to-noise were low and foregrounds could be easily impacted by systematics, rendering signal loss more complex and widespread than the examples we have investigated so far.

The reliability of the transfer function was validated for the results in Cunnington et al. (2022) and here we demonstrate these validation tests by utilizing a different set of simulations which we refer to as the MDMK simulations, which aim to emulate the MeerKAT 2019 pilot intensity mapping survey (Wang et al. 2021). We use the survey's non-uniform mask and use fluctuations in the uncleaned foreground sky to generate systematic perturbations in the MDMK foreground simulations creating a demanding cleaning requirement. We outline the details of how this is achieved in the following section.

4.1 MDMK MeerKLASS simulations

To emulate current MeerKAT pathfinder data and investigate the performance of the transfer function when signal loss is spread across a wide range of scales, we utilize the MeerKAT 2019 pilot survey data (Wang et al. 2021; Cunnington et al. 2022). The survey targeted a single patch of ~ 200 deg^2 in the WiggleZ 11hr field, covering $153^\circ < \mathrm{RA} < 172^\circ$ and $-1^\circ < \mathrm{Dec.} < 8^\circ$. The telescope observed at constant elevation, scanning back and forth through azimuth taking 1.5 h to complete one time-block. Seven time-blocks were obtained with a mix of rising and setting scans, creating offset coverage providing the footprint which can be seen in the MDMK simulation maps in Fig. 13. Holes in the footprint are evident and are caused from the multistage RFI flagging which can leave gaps in the scanning. For the MeerKAT H I simulation (top-panel of Fig. 13), we use the same MULTI-DARK simulation as in MK1GPC (Section A1) but calculate the physical volume covered by the MeerKAT 2019 data and cut a volume of this size from the $1\,(\mathrm{Gpc}/h)^3$ cube. We use a similar pixelization as the 2019 data ($n_{\rm x}, n_{\rm y}, n_{\rm z} = 133, 41, 250$), and then apply the exact same footprint mask. The MeerKAT 2019 observations were performed in L band but to avoid dominant RFI, only 199 channels with a 973.2–1014.6 MHz frequency range ($0.400 < z < 0.459$) were used. Additionally, of the 199 channels selected to use, a further 32 are removed due to evidence of RFI contribution in their eigenmodes.

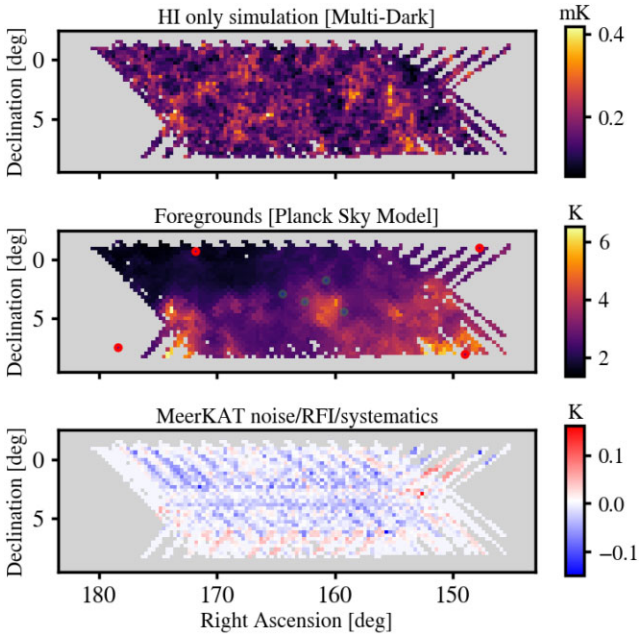

Figure 13. MDMK simulated maps aiming to emulate MeerKAT 2019 intensity mapping data, averaged along the 973.2−1014.6 MHz frequency range. The top panel shows H I only, produced using the MULTI-DARK *N*-body semi-analytical simulation. The middle panel shows the foregrounds simulated using a perturbed version from the Planck Sky Model. The points on the foreground map indicate the positions of the example spectra plotted in Fig. 14, with the green and red points representing pixels near the centre or edge of the MeerKAT footprint respectively. The bottom panel shows an estimate for some MeerKAT noise and residual systematics obtained by subtracting data observed at different times (see Section 4.1.2).

We also replicated this exact channel flagging in the MDMK simulations.

4.1.1 *Frequency perturbed foregrounds*

Evidence of systematics was seen in the MeerKAT 2019 data, which was expected given the low amount of observational time and it being a first of its kind pilot survey. One way systematics were evident was in the perturbations to what should be smooth spectra in the raw foreground sky. The exact cause of these systematic perturbations is beyond the aim of this paper but we can still use the distorted spectra from the real data to perturb an idealized foreground simulation, emulating the main impact from these systematics on the foreground clean and signal reconstruction.

To create foregrounds for the MDMK simulations we begin by using the Planck Sky Model to generate synchrotron and free–free emission at the relevant frequencies and sky position, as with the MD1GPC simulation. The bottom panel of Fig. 13 shows the foreground simulation for the MeerKAT 2019 footprint. Unlike the MD1GPC simulation, we do not include any simulated polarization leakage and instead use the MeerKAT 2019 data itself, aiming to create more realistic systematic perturbations to the foregrounds. This is done by fitting a smooth polynomial to each line of sight in the 2019 data, then the systematic perturbations to the foreground spectra can be approximated by the ratio between the data and the polynomial i.e.

$$T_{\text{perturbation}}(\boldsymbol{x}, \nu) = \frac{T_{\text{2019-data}}(\boldsymbol{x}, \nu)}{T_{\text{smooth-poly}}(\boldsymbol{x}, \nu)}. \quad (18)$$

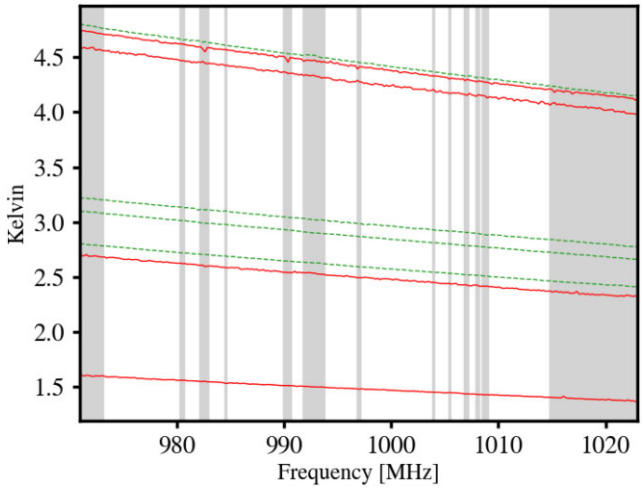

Figure 14. Example spectra from the MDMK foreground simulations aiming to emulate MeerKAT pilot survey data. The perturbations to the spectra have been produced using the MeerKAT 2019 data to create realistic foreground simulations with systematic effects (see Section 4.1.1). The green dashed lines represent pixels taken from a central position in the MeerKAT footprint (corresponding to green points in Fig. 13). The red solid lines represent pixels near from the edge of the footprint (corresponding to red points in Fig. 13) and are more vulnerable to systematics, hence the more noticeable perturbations. The grey-shaded regions represent channels which were flagged in the cross-correlation analysis (Cunnington et al. 2022), which we also flag in this simulation for consistency.

We found on average these perturbations were small sub-per cent values; however, they could be as high as 3.7 per cent. We multiply these perturbations with the PSM foreground and Fig. 14 shows some example perturbed spectra from the final simulation. The perturbations are worse near the edge of the map (shown as red solid lines) where due to the lower coverage, systematics can have more impact. This is consistent with what was found in the actual data and in the cross-correlation analysis these edge pixels are down-weighted (Cunnington et al. 2022). Fig. 14 also shows the flagged channels (grey regions) used in the cross-correlation analysis which we also adopt in the MDMK simulation.

4.1.2 *Anisotropic systematics and RFI residuals*

The analysis of the autocorrelation power spectrum for the 2019 MeerKAT survey in Cunnington et al. (2022) showed evidence of additive biases most likely from instrumental noise, residual RFI, or other systematics. Their contribution appears to dominate over the H I signal because the autopower spectra amplitude was larger than one would expect from H I only power. We also include a contribution to the MDMK simulation which attempts to emulate these types of additional components. Again, we use the real MeerKAT 2019 survey itself to produce a map of time-varying anisotropic contributions and add these on to the MDMK H I and perturbed foreground maps. This is achieved by taking the residuals from different time blocks in the MeerKAT 2019 survey. We take the difference between the first four time-blocks and the last three where these residuals will represent components that vary in time. Therefore, in principle, this should not include the H I signal or the foregrounds since these would be consistent in time, but instead only include time-varying systematic contributions, which is what we are aiming to emulate.

The map of the MeerKAT time-block residuals is shown in the bottom panel of Fig. 13. In some instances, there are no shared

pixels in both time-block groups so the residual is undefined. For these pixels we instead resort to adding a large level of Gaussian random noise whose variance dominates the H I signal by one order of magnitude. However, when these are plotted in Fig. 13, which is the average along the line of sight, their contribution is averaged down which is why the amplitude appears relatively low around the edges where most of the missing pixels between time-blocks are. The pixels from the actual MeerKAT residuals, which are most concentrated in the centre where shared coverage is better, appear higher relative to the Gaussian noise. This will be due to the residuals being more correlated in frequency, thus do not average down in the plot. The frequency correlation and apparent anisotropies of the residuals as evident in Fig. 13 suggest they are contributions beyond simple instrumental noise. While this is a complication for the pilot survey analysis, it is useful for our purposes, providing additional complications to foreground cleaning and signal loss in the MDMK simulations.

In these simulations, while we do not explicitly include the effects from a realistic telescope beam, some of the impacts it has on the foregrounds will be included in the perturbations we add to the simulations from equation (18). A simple Gaussian beam is trivial to include and assuming it is approximately matched in the transfer function construction, it makes no difference to the performance of the transfer function as we will explicitly show in Section 5.1. However, in reality, the MeerKAT beam will be more complex with wide-reaching frequency-dependent side lobes which could complicate foreground cleaning (Matshawule et al. 2021; Spinelli et al. 2021). Trying to replicate this in the mocks in the transfer function construction may be difficult and an investigation into what level this needs to be considered should be pursued. This requires a more detailed simulation which we will pursue in follow-up work.

4.2 Correcting signal loss in pathfinder data

Fig. 15 shows power spectra for the MDMK (MeerKAT-based) simulations presented in the previous sub-section. The black dashed line shows the foreground-free H I-only result, and the red solid line shows the result from adding the perturbed foregrounds and residual time-varying systematics based on real MeerKAT data, then performing a $N_{\rm fg} = 10$ PCA clean. We find that using the perturbed foregrounds (described in Section 4.1.1) is the main cause for requiring an aggressive foreground clean, highlighting the importance of instrument calibration so that smooth spectra are maintained in the observed data. Similar to MeerKAT data (Cunnington et al. 2022), signal loss in the foreground-cleaned data is widespread throughout all scales, with noticeable signal loss occurring even in the highest k. The main reason for this is due to the decreased depth of the frequency/redshift range. We tested this with the main MD1GPC simulation. Reducing the number of pixels along the LoS by a factor of 4 to 64 pixels with 0.25 Gpc h^{-1} of depth produced >75 per cent signal loss in the smallest four k-bins, and still 13 per cent in the highest k-bins. This was even without the polarization leakage and just removing four PCA modes. For an equivalent scenario but using the full 1 Gpc h^{-1} depth, the signal loss is never greater than 30 per cent and only 3 per cent at the highest k. This can be understood by considering that modes projected out of narrow frequency-range data will be confined to a higher $k_\parallel$ space. Thus the signal loss will also spread into higher k-modes. Encouragingly this means that signal loss should be naturally mitigated in future observations using a larger frequency range. This will be possible with MeerKAT UHF-band observations which will probe lower frequencies where RFI is expected to be less dominant, thus a more complete frequency range can be used.

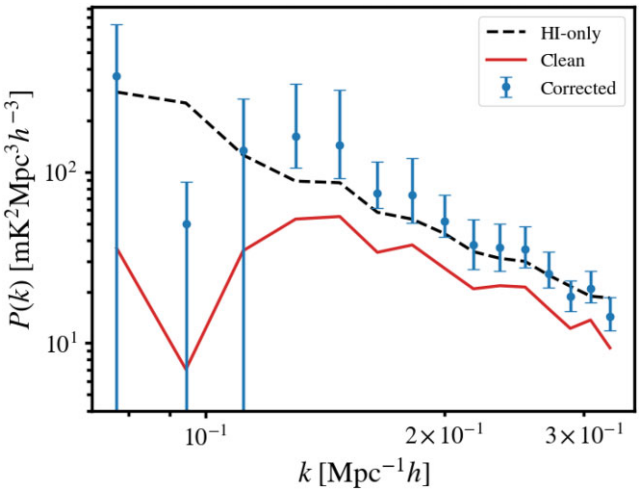

Figure 15. Power spectra for MDMK simulations which are small in volume, have a more complex non-continuous survey footprint, high anisotropic noise and residual systematics, as well as systematically perturbed foregrounds, all to emulate actual MeerKAT pathfinder data. We show the cross-correlation with the H I-only maps. Signal loss from the $N_{\rm fg} = 10$ PCA clean (red line) is larger and more widespread into high k compared with previous results in the more idealized MD1GPC simulation. Despite this the transfer function is encouragingly still able to reconstruct a reasonably unbiased result shown by the blue data points. Error bars are given by the limits of the central 68th percentile region from the distribution of reconstructed power spectra using the transfer function mocks.

Despite the more complex and widespread signal loss in Fig. 15 (red line), when we construct a transfer function using the process summarized in Section 3.1, we are able to reconstruct the correct H I power spectrum. As in previous tests, the clean and corrected power spectra are cross-correlations with the original-H I to avoid any issues with residual foreground contamination confusing the assessment of signal loss. The power spectra are naturally more noisy than the previous MD1GPC simulations due to the decreased volume, the systematically perturbed foregrounds (see Fig. 14) and the large time-varying systematics (Fig. 13 bottom panel) inserted into the simulation. The instrumental noise and additive systematics is an important consideration that we did not include in the default MD1GPC simulations of previous sections. This additional noise will introduce extra perturbations to the foreground modes, in the same way the signal introduces perturbations (see Fig. 1). If the noise is large, as is the case for pathfinder observations with low observational time, these perturbations will be large. Encouragingly, this does not appear to cause noticeable problems for the transfer function, evidenced by the corrected result in Fig. 15, which includes large additional contributions that dominate over the H I signal.

Given the more complex nature of the pilot survey simulation, some reconstructed power spectra from the transfer function mocks produced outliers and returned non-Gaussian distributions for each k-mode. In this scenario, using the rms over the mocks for the errors would be a poor estimation and would be overly distorted by the outliers. We therefore instead use the 68th percentiles limits to provide the asymmetric error bars, which is what are presented in Fig. 15. To obtain the converged distribution, we used 1000 mocks in the transfer function calculation. This presents a further advantage of using the transfer function mocks for error estimation, providing more options to handle non-Gaussian uncertainties. This of course

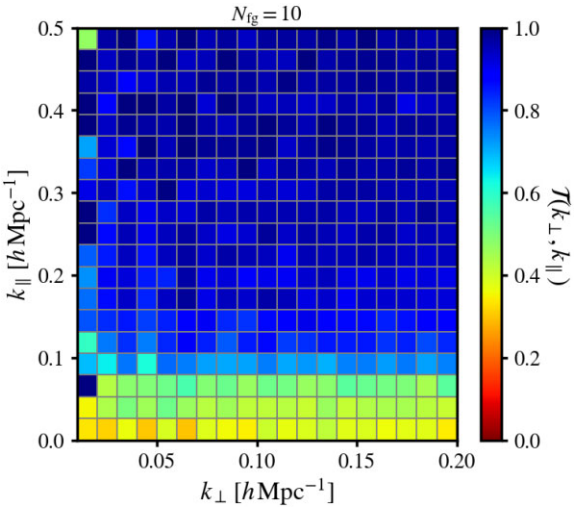

Figure 16. Foreground transfer function in cylindrical $k_\perp$, $k_\parallel$ space for the MDMK simulations emulating MeerKAT pathfinder data. $N_{\rm fg} = 10$ PCA modes have been removed for the foreground clean. The 1D spherically averaged version of this is used for the corrected results in Fig. 15.

would have complex implications for further analysis and parameter estimation, which we do not investigate here. However, errors should naturally become more Gaussian as intensity map quality improves and noise, systematics, etc. are reduced.

Fig. 16 shows the computed transfer function in cylindrical $k_\perp$, $k_\parallel$ space for the MDMK simulations. It is interesting to analyse the differences between this more realistic case and that from the more idealized MD1GPC simulations in Fig. 6. This again reveals that signal loss is more widespread into larger $k_\parallel$ modes, which is consistent with the more widespread signal loss evident in Fig. 15. We still see large regions where $\mathcal{T} \sim 1$ suggesting that the approach of discarding some regions of k-space to massively reduce the dependence on the transfer function (as we discussed in Section 3) could still be pursued even with small intensity mapping pilot surveys.

The results from the MDMK simulations provide validation for using the foreground transfer function with pathfinder survey intensity maps. We have used MeerKAT data to attempt to complicate the foreground clean and signal loss to stress-test the current signal reconstruction process. The simulations produced make no attempt to understand the source of the perturbations to foreground spectra or additive time-varying systematics, simply emulating them as realistically as possible to mimic the challenge they pose. While the success is encouraging, the investigation would be completed further by including specific simulations of known observational effects such as directly simulating RFI, $1/f$ correlated noise, non-uniform noise, a non-Gaussian beam etc. This would allow more analysis into exactly what observational effects are most troublesome for foreground cleaning and signal loss. The development of a robust simulation pipeline for MeerKAT single-dish intensity maps including a realistic beam and all these observational effects is outside the aims of this paper but is being pursued in other MeerKLASS collaboration projects (e.g. Irfan et al. 2023).

5 PRECISION COSMOLOGY SUITABILITY

In this section, we look to future H I intensity mapping observations and test how reliable a transfer function would be where sub-per cent accuracy on parameter estimates is required. We return to the more generic simulations of MD1GPC to avoid the investigation being confused by the large statistical noise present in the previous section's realistic simulations of a MeerKAT pilot survey.

5.1 Mock parameter dependence

Up until now, it has not been investigated how robust the accuracy of the transfer function is when there are discrepancies between parameters used in the transfer function construction and their true fiducial values in the real observed data. If the parameter assumptions used for the generation of 100 mocks strongly influence the final accuracy of the reconstructed power spectrum then this is a large concern for precision cosmology since this would lead to biased cosmological parameter estimates.

In this section, we demonstrate how mild the mock parameter dependency is and show how large $+/-$ 100 per cent discrepancies between the assumed parameters in the mocks used for transfer construction and the underlying truth in the data, mostly only yield small $\lesssim 1$ per cent inaccuracies in the recovered parameter estimations. We test this by treating the MD1GPC as the observed data with the underlying 'true' parameters, then vary some of the values $\{\Omega_{\rm HI} \propto \overline{T}_{\rm HI}, f, \sigma_{\rm v}, R_{\rm beam}\}$ in the lognormal mocks which are used to construct the transfer functions. The model power spectrum we use to generate the lognormal mocks is described in more detail in Appendix B1 but we repeat it here for convenience

$$P_{\rm mod}(k,\mu) = \overline{T}^2_{\rm HI} \frac{\left(b^2_{\rm HI} + f\mu^2\right)^2}{1 + (k\mu\sigma_{\rm v}/H_0)^2} P_{\rm m}(k) \times \exp\left[-(1-\mu^2)k^2 R^2_{\rm beam}\right]. \quad (19)$$

For this test we used the MD1GPC simulations with a $\sigma_{\rm n} = 1$ mK dominant white noise and a default Gaussian beam where $R_{\rm beam} = 10\,h^{-1}$Mpc.

Fig. 17 shows how subtle the impact on parameter estimation is when parameters used in the mock's model power spectra, given by the panel titles, are biased by an amount indicated by the x-axis. The y-axis shows the percentage bias relative to the foreground-free parameter estimation. In all cases, we sample the parameter posterior distribution in a Bayesian MCMC only varying one parameter at a time, fixing all other parameters in the model (equation 19) to fiducial values fitted to the foreground-free MD1GPC simulation. The error bars represent the 68 per cent confidence regions in the posterior distributions and they are plotted relative to the median of the foreground-free posterior. To avoid non-linear complications in the modelling we only use modes where $k < 0.3\ h\,{\rm Mpc}^{-1}$. While these scales are still quite non-linear, our model worked reasonably well on these scales and since we are testing its performance relative to the foreground-free case, any shortcomings caused by non-linear effects will be present in both.

$\Omega_{\rm HI}$, which is proportional to $\overline{T}_{\rm HI}$ (see equation B5), will only change the amplitude of the H I power spectrum in our simple linear model, and the transfer function appears extremely robust to these scale-independent amplitude changes, with no > 1 per cent bias being induced. We even tested going to 4 $\times$ the fiducial truth on $\Omega_{\rm HI}$, since this is a very unconstrained parameter, but this still yielded a sub-per cent bias. Note that the red-cross indicates an undefined result for the $\Omega_{\rm HI} - 100$ per cent case since the intensity maps are zero when $\overline{T}_{\rm HI} \propto \Omega_{\rm HI} = 0$. The increase in uncertainties with a decreasing $\Omega_{\rm HI}$ mock parameter input is caused by the white noise having a more dominant impact relative to these lower amplitude mocks in the transfer function construction.

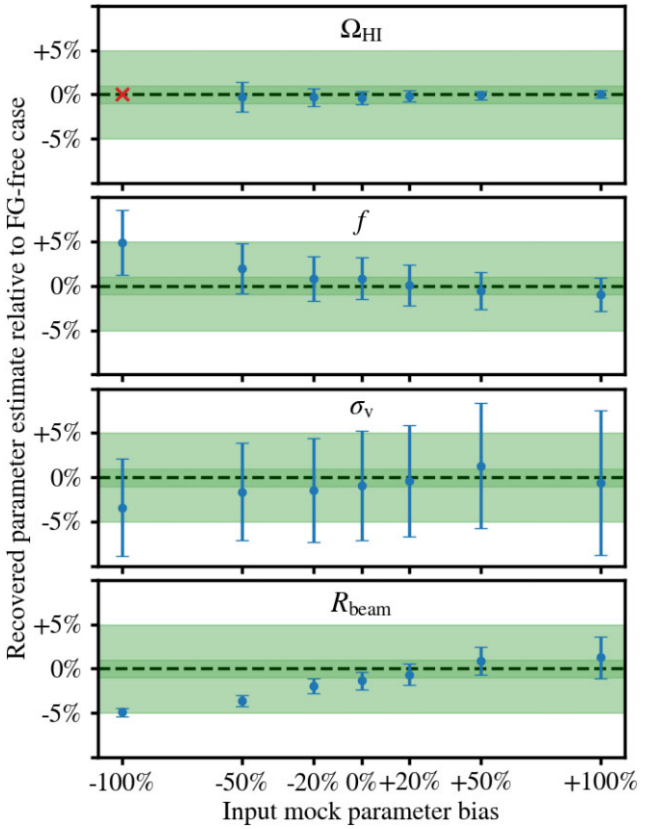

Figure 17. Robustness of the transfer function in response to biased parameter assumptions. Each panel shows a different parameter used in the mocks which construct the transfer function. The x-axis shows different per cent-biases relative to the correct fiducial values. The y-axis shows the estimated parameter posterior from an MCMC on the reconstructed data using a transfer function. The error bars represent the 68 per cent confidence region in the estimated posteriors. These are plotted relative to the foreground-free parameter estimates to demonstrate that only small biases are induced from incorrect parameter assumptions in the transfer function construction. These results are produced with the MD1GPC simulation with $\sigma_{\rm n} = 1$ mK dominant noise and a Gaussian beam with $R_{\rm beam} = 10\,h^{-1}$Mpc. All results are for a $N_{\rm fg} = 12$ PCA clean.

The growth rate f is included to model the RSD which are present in the MD1GPC simulations. f will be an anisotropic parameter since RSD are a line-of-sight-only effect. This appeared to induce more of a bias in the reconstructed power spectra, but it still only caused < 2 per cent in most cases tested. For certain cosmological parameters such as the growth rate, much tighter priors will be applicable such that the biased parameter values we have used for f will be unrealistic. It is interesting to see that failing to include a linear RSD model in the mocks at all, as shown by the -100 per cent result (i.e. $f = 0$), induces a ~ 5 per cent bias which suggests that it is important to include some basic anisotropic RSD in the mocks for reconstruction accuracy.

$\sigma_{\rm v}$ is again used to model RSD but as a phenomenological attempt to model fingers-of-god (FoG) on mildly non-linear scales. This will also be an anisotropic parameter but is also directly scale-dependent too, having a greater influence at high-k. This shows ~ 3 per cent bias if the parameter is set to zero. Even though this is a phenomenological parameter without a physically defined fiducial value, it is still unlikely that the parameter will be completely omitted without a suitable method for replacing its modelling effects. We highlight that the increase in error bars for the f and $\sigma_{\rm v}$ parameters is not necessarily foreground-related. Some parameters will be constrained better than others and $\sigma_{\rm v}$ relies more on small scales which are damped by the beam. Furthermore, both parameters would be more suitably constrained by including the quadrupole as opposed to the spherically averaged monopole we are using here.

Lastly, we introduce the size of the beam $R_{\rm beam}$ as a varying nuisance parameter. This is defined as the rms of the beam profile in comoving units at the probed redshift. Similarly to $\sigma_{\rm v}$ this is another scale-dependent anisotropic parameter, although this time affecting high-$k_\perp$ modes. This can reach a ~ 5 per cent negative bias if completely unaccounted for, shown by the -100 per cent result. For nuisance parameters linked to the instrument such as $R_{\rm beam}$, we should have a much tighter prior on its value effectively ruling out a >20 per cent incorrect assumption. The small ~ 1 per cent bias relative to the truth in the case where the correct fiducial beam has been used (see 0 per cent input mock parameter bias result), suggests there could be some discrepancy in high-$k_\perp$ modes between foreground-free and reconstructed foreground-cleaned data, where $R_{\rm beam}$ has the most impact, leading to inaccuracies in its estimation. Given this is only a very small bias and is only in a nuisance parameter, we defer this to future work, where we will investigate the impact on signal reconstruction in the presence of more complex beams.

Given that the small biases appear to be caused by anisotropic divergences between mocks and observations, we investigated whether the 2D transfer function (discussed in Fig. 7) could yield improvements. We found that when there is no polarization leakage in the simulations, the average bias on the reconstructed power spectrum from using $f = 0$ in the mocks is 3.8 per cent relative to the foreground-free power spectrum. Interestingly though, when we construct and apply the transfer function in 2D, then rebin into 1D following the same procedure in equation (14), the bias is only 1.9 per cent. However, as we found from the results in Fig. 7, the variance blows up when we reintroduce the more complex foreground with polarization leakage. We thus leave further investigation to future work with more realistic simulations where we will definitely test if the large variance in the 2D transfer function can be reliably reduced, and if it then still decreases the small biases from anisotropic inconsistencies we see in Fig. 17.

There are some important conclusions to draw from the results in Fig. 17. First, the results demonstrate how the transfer function works. It is not a process where an exact replica of the real data is required to measure the precise impact of signal loss. Rather the mocks injected act as a test field to construct a response function caused by the foreground cleaning. Broadly speaking, it appears that the parameters used in the mocks for the transfer function do not have a strong influence on the final accuracy of the reconstructed power spectrum. However, where sub-per cent accuracy is the aim, it is clearly beneficial to have the mocks attempt emulation of some of the features inherent in the observational data. For example, completely neglecting the telescope beam or linear RSD in the mocks can have a noticeable impact on the reconstruction accuracy. This lays the foundation for many further inquiries into this topic. For example; will a more realistic frequency-dependent beam with a side-lobe structure be sufficiently emulated by a Gaussian beam in the mocks for the purposes of the foreground transfer function? Do any of the other multitude of parameters that we want to probe or are forced to include in our model as nuisance parameters, have a stronger influence on reconstruction accuracy? What happens when we allow multiple parameters to vary simultaneously as opposed to varying one parameter at a time as done for the results in Fig. 17?

While these results are strong evidence that a transfer function will not bias parameter inference, the most robust way to confirm this would be with more realistic simulated observational effects e.g. a MeerKAT model of the beam, and to perform detailed modelling with a multiparameter MCMC fit to the reconstructed power spectrum, including *shape* parameters e.g. h, n_s, ω_c, ω_b, under a range of different scenarios. This is beyond the scope of this work but is something we will aim to showcase in a follow-up study.

Since it seems the fiducial mock parameters have a sub-dominant influence over the reconstruction accuracy, one prospect to consider is a process whereby the mock parameters are updated based on the parameter inference from the real data. In this way an iterative transfer function could be developed whereby as the parameter posteriors for the observed data are estimated and converge on a final parameter estimate, these values can be used to update the transfer function calculation and avoid any possibility of it biasing the parameter inference. We discuss some of the early investigations for this possibility in Appendix D, but also largely leave this to further dedicated investigation.

5.2 Probing exotic physics on ultra-large scales

As a final test of the transfer function's performance, we examine its ability to reconstruct signal on the largest scales, even when the underlying true signal has some unknown 'non-standard' properties. For this test, we focus on primordial non-Gaussianity (PNG) which can be probed on the largest scales in galaxy surveys (Mueller et al. 2022) and soon in H I intensity maps (Li & Ma 2017; Witzemann et al. 2019; Karagiannis et al. 2021).

The nature of the fluctuations in the primordial Universe which arise during inflation carry a wealth of information regarding the physical mechanisms that shaped the early Universe. The parameter f_{NL} quantifies the departure from Gaussianity in the primordial Universe (Komatsu & Spergel 2001) and for the so-called local-type of PNG, $f_{NL} \neq 0$ would be evidence of non-Gaussian fluctuations, ruling out slow roll, single-field inflation in favour of more exotic multifield models (Creminelli & Zaldarriaga 2004). Constraints on PNG so far come from CMB anisotropies and results are consistent with Gaussian fluctuations with $f_{NL} = 0.9 \pm 5.1$ (Planck Collaboration VI 2020). However, large-scale structure surveys, in particular intensity mapping, are expected to soon lead the way in improving PNG precision. Evidence for PNG in large-scale structure surveys will manifest as a scale-dependent correction to the linear bias (Dalal et al. 2008). This correction scales as k^{-2} thus it is at *ultra*-large scales where sensitivity to f_{NL} becomes most prominent.

In this section, we generate a new underlying H I intensity map simulation, no longer using the MD1GPC simulation. The reason for this is first because we wish to add a clear signature of PNG into the field, and secondly, because we need to cover much larger scales where we will be able to probe the scales that are sensitive to the f_{NL} parameter. We use the *N*-body COmoving Lagrangian Acceleration (COLA)[6] (Tassev, Zaldarriaga & Eisenstein 2013; Tassev et al. 2015) code to generate a fast *N*-body simulation on a $(8{,}000\,h^{-1}\mathrm{Mpc})^3$ grid with 256^3 pixels which approximately represents a wide and deep H I intensity mapping survey with something like SKAO. We seed the simulation with an H I power spectrum given by

$$P_{HI}(k,\mu,z) = \overline{T}^2_{HI}(z)\left[b_{HI}(z) + \Delta b_{HI}(k,z) f_{NL} + f(z)\mu^2\right]^2 P_m(k,z), \quad (20)$$

where

$$\Delta b_{HI}(k,z) = [b_{HI}(z) - 1]\frac{3\Omega_m H_0^2 \delta_c}{c^2 k^2 T(k) D(z)} \quad (21)$$

and $\delta_c = 1.686$ is the critical matter density contrast for spherical collapse, $T(k)$ is the *matter* (not foreground) transfer function, and lastly the growth function can be defined by

$$D(z) = \frac{5}{2}\Omega_m H_0^2 H(z)\int_z^\infty \frac{1+z'}{H^3(z')}\mathrm{d}z'. \quad (22)$$

We set $f_{NL} = 100$ in equation (20) to provide a clear scale-dependent bias on large scales in the new observed data simulation. We add foreground contamination to this following the same steps as the MD1GPC foreground model, run the PCA foreground clean, then construct the foreground transfer function following the same methods in the rest of this section. The lognormal mocks which construct the transfer function will assume $f_{NL} = 0$ and we will be testing if this transfer function can still recover the true underlying H I power spectrum with a scale-dependent bias induced by the $f_{NL} = 100$ PNG signature. Despite $f_{NL} = 100$ being confidently ruled out by Planck18 observations, we still use this to set up a highly diverged underlying cosmology from that assumed in the transfer function construction, similar to the extreme biases we tested in Fig. 17.

Fig. 18 shows the results for the large-scale PNG COLA simulation. The black dashed line shows the foreground-free result where the impact from the $f_{NL} = 100$ input is clearly evident on small-k. For reference we also plot the $f_{NL} = 0$ equivalent case shown by the purple dotted line. Adding foreground contamination and cleaning is enough to greatly distort the PNG signature (shown by the red-solid line). For this simulation, we do not add polarized foregrounds, therefore only a $N_{fg} = 4$ PCA clean is required. This is more representative of a future large sky survey where it is safer to assume an enhanced level of calibration has been achieved. The blue data points show the result where we have used the transfer function to reconstruct the signal loss from the foreground clean. The errors are estimated from the variance of the transfer function as outlined in Section 3.3. There is good agreement between the reconstructed power spectrum and the true underlying power spectrum (black dashed) where $f_{NL} = 100$, shown in more detail by the residuals in the bottom panel. We emphasize that this is an extreme case where the underlying true cosmology is very different from that assumed by the mocks in the transfer function, and despite this only three of the measured modes in the reconstructed power spectrum are more than 1σ beyond sub-per cent accuracy.

Previous work in Cunnington, Camera & Pourtsidou (2020) investigated signal loss from foreground cleaning in the context of PNG and demonstrated how a phenomenological model could be implemented to account for the signal loss. While this delivered successful simulation-based results, there was a worrying degeneracy between the parameters in the signal loss model and f_{NL}, thus tight priors would be needed for such a method to allow good constraints on f_{NL}. These tight priors would only come from a very good understanding of foreground contamination and signal loss. The advantage behind the foreground transfer function approach is that no phenomenological model is required and the signal loss is reconstructed using the observed data itself without the need for additional nuisance parameters. We emphasize that this is a preliminary investigation into PNG with a foreground transfer function and a more complete study with robust large-scale simulations is required, including more detailed modelling and MCMC parameter estimation, which we defer to future work.

[6] pypi.org/project/pycola3

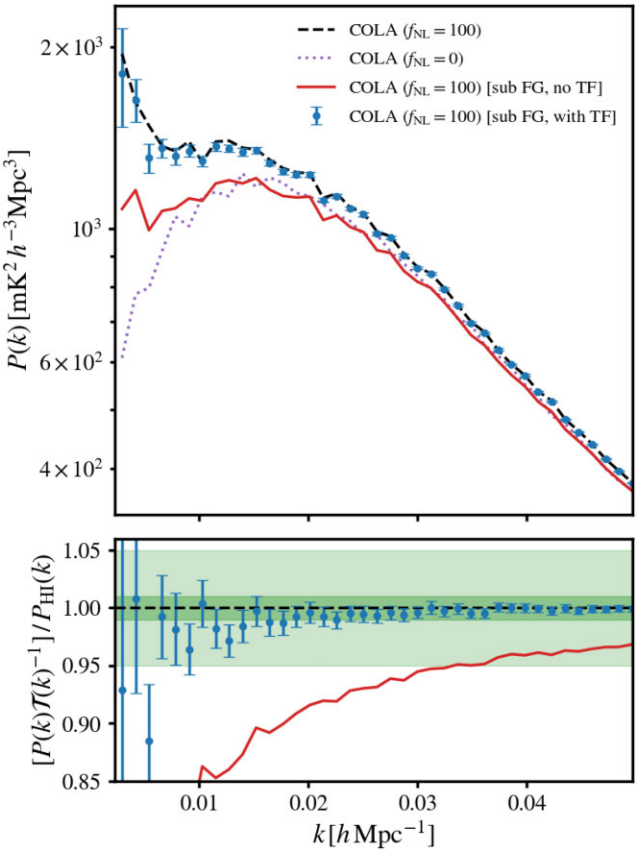

Figure 18. Capability to detect 'non-standard' cosmology on ultra-large scales when the fiducial cosmology used in the transfer function construction assumed standard. For this test, we use the PNG parameter f_{NL}. We let the underlying truth have an extreme $f_{NL} = 100$, produced using COLA simulations (black dashed line). Despite an incorrect $f_{NL} = 0$ assumption used in the mocks for the transfer function, the correct f_{NL} signature is still recovered in the reconstructed power spectrum (blue data points) across most scales.

The previous results in Fig. 17 supported the claim that a discrepancy between parameters assumed for the transfer function and the underlying truth in the data, only has a mild influence on results. Fig. 18 extends this conclusion to the largest scales where foreground contamination is most troublesome and thus is a very encouraging result. A slight negative bias in the reconstructed power spectrum is apparent, likely caused by the lower f_{NL} in the mocks, thus we cannot claim complete independence from the mocks. Sensible assumptions should therefore still be made when fixing a fiducial cosmology for the transfer function construction. However, there is potential for this small bias to be eradicated by adopting the iterative approach to transfer function calculation as discussed in Appendix D. In this case, the recovered f_{NL} inferred from the reconstructed power spectrum, could be used again to seed new mocks used in a new transfer function calculation, thus achieving enhanced accuracy. This is another item we leave for follow-up work, extending beyond the initial discussion and tests in Appendix D.

6 CONCLUSIONS

H I intensity mapping has the potential to be a leading resource for precision cosmology. A major challenge involves removing astrophysical foregrounds that dominate the underlying H I cosmological signal by several orders of magnitude. Simulations and real observations are providing evidence that foregrounds can be sufficiently cleaned using blind separation techniques. However, quantifying the precise impact foreground cleaning has on the H I power spectrum is crucial for avoiding biased analyses. In this work, we validate a method involving mock signal injection into the observed data as a means for accurately estimating the signal loss induced in the H I as a consequence of the foreground clean. This method, referred to as the foreground transfer function, has been used in real data analysis before and its accuracy was validated for the results in Cunnington et al. (2022), but its reliability for the purposes of precision cosmology has not been studied until now. For the first time, we present simulation-based tests demonstrating the foreground transfer function's accuracy as a tool for reconstructing estimated H I power spectra to sub-per cent accuracies.

This work used a selection of simulations to enable tests on a range of scenarios. In all cases we used an underlying H I intensity map generated from an *N*-body simulation, from which we could measure the 'true' signal. Simulated foreground maps were then added and a PCA-based foreground clean was performed, the consequences of which were analysed relative to the truth. We varied the complexity of the foregrounds and observational effects providing scenarios with differing demands from the foreground clean, hence presenting a range in signal loss. Foreground transfer functions were constructed using lognormal H I intensity mapping mocks and their ability to recover the true signal could be scrutinized. In this work, our focus was on *signal loss* from overcleaning, as opposed to the other undesirable consequence of *foreground residuals* caused by undercleaning. Foreground residuals should be reducible to sub-dominant contributions or circumvented in cross-correlation with e.g. galaxy surveys, hence are less problematic than signal loss for precision cosmology. In the majority of tests, we therefore inspected the reconstructed cross-correlation of the cleaned intensity map with the original foreground-free (H I-only) map. This way, only the effects from signal loss would manifest and the impact from foreground residuals would be mitigated, which cause a positive bias in the H I autocorrelation.

We summarize our main conclusions below;

(i) We summarized the recipe for estimating a foreground transfer function (Section 3.1) using mock signal injection into the observed data, which delivers an unbiased reconstructed power spectrum (Fig. 4). These results included simulated polarization leakage which demanded an aggressive foreground clean, resulting in > 50 per cent signal loss on the largest scales. In Fig. C1 we demonstrated the potential consequences of deviating from this unbiased recipe resulting in underestimated signal loss of up to ~ 30 per cent.

(ii) We validated a technique for estimating the covariance in reconstructed power spectra which involves calculating the covariance across the reconstructed power spectra from all injected mock realizations (Fig. 9). Crucially, when adopting this approach, the cleaned observed data $\mathsf{X}_{\rm clean}$ must not be subtracted when calculating the transfer function, which is otherwise removed to reduce the variance in the transfer function and optimize its accuracy.

(iii) It has been previously assumed that the transfer function should be applied twice to correct for autocorrelation signal-loss i.e. $P^{\rm auto}_{\rm rec} = P^{\rm auto}_{\rm clean}\mathcal{T}^{-2}$. However, our simulations tests show that this is not the case (Fig. 11). At the fundamental 3D Fourier transform level, the average suppression in signal is the same in autocorrelation and cross-correlations with a foreground-free tracer (Fig. 12). Thus a $\mathcal{T}^{-1}$ transfer function is the correct reconstruction in both cross- and autocorrelation power spectra.

(iv) When calculating the transfer function, we estimated power directly in 1D spherically averaged k-bins, finding this performed well on all our simulations. This deviates from some previous approaches which instead calculate and apply a transfer function in 2D cylindrically averaged $k_\perp$, $k_\parallel$-bins, then re-projects these bandpowers into 1D k-bins. However, we found evidence that this made results prone to a higher variance (Fig. 7), which may require a more detailed weighted average in the 2D to 1D projection to mitigate the issue.

(v) To 'stress-test' the transfer function performance, we applied it on a pathfinder-like intensity map where the survey volume is relatively small, systematic effects are present, and signal-to-noise is low. To do this we produced the MDMK simulation that emulated the MeerKAT 2019 pilot survey using the same footprint, non-uniform frequency coverage, and perturbed foreground spectra, produced using the MeerKAT data itself (Section 4.1). Despite these increased challenges which result in more widespread signal loss, the transfer function was still able to reconstruct an unbiased power spectrum (Fig. 15), validating the approach in Cunnington et al. (2022).

(vi) Finally, we confirmed how the transfer function accuracy has a relatively low dependency on the input parameters used for the mock generation in the transfer function calculation. To demonstrate this we chose some example parameters and biased them relative to their true fiducial values. Even going to extreme $+/-$ 100 per cent biases, yielded small biases in the reconstructed power spectra and < 5 per cent bias in the recovered parameter estimates relative to the foreground-free case (Fig. 17). We also ran an extreme test where the underlying fiducial cosmology had an $f_{\rm NL} = 100$ value producing a scale-dependent bias on the largest scales. Despite a foreground clean heavily distorting this PNG signature, we demonstrated that an unbiased recovery is still obtained even if assuming $f_{\rm NL} = 0$ in the transfer function calculation (Fig. 18).

These results place increased confidence in using a foreground transfer function as H I intensity mapping ventures into precision cosmology. There is also some flexibility in regard to the dependency on reconstruction. Where levels of signal loss are high (e.g. > 50 per cent on large scales), post-cleaning scale cuts can be imposed to limit the dependency on the reconstruction. We discussed this in Section 3.1 and argued that excluding small-$k_\parallel$ modes from the analysis will massively reduce signal loss. A transfer function is still an essential tool in these scenarios since some reconstruction across all scales will be required. Importantly, the transfer function will also estimate the regions where signal loss is high, informing scale cut locations.

We have strived to demonstrate our results on a range of simulations with differing observational effects and survey sizes. However, simulations can have approximations not present in reality, thus validating the transfer function for signal loss reconstruction will remain an ongoing pursuit with more specific tests. One relevant extension we discussed will be to use simulations with a complex beam pattern, as opposed to the Gaussian beam assumed in this work. An incorrect beam model can have a high impact beyond the transfer function hence, we decided to leave this to a future more dedicated study, extending on the work of Matshawule et al. (2021) and Spinelli et al. (2021). We will also aim to include 1/f noise in the simulations which can impact foreground cleaning (Harper et al. 2018; Li et al. 2021; Irfan et al. 2023). Investigating how the transfer function can be constructed and applied in other clustering estimators will also be crucial. For example, extending the transfer function formalism to higher order multipoles, applying it to correct the quadrupole and hexadecapole. There also needs to be investigation into whether it can be applied in configuration space to the correlation function or whether it can be used for correcting foreground cleaning effects in n-point statistics such as the bispectrum (Cunnington, Watkinson & Pourtsidou 2021b).

There is continual improvement in foreground removal (Irfan & Bull 2021; Makinen et al. 2021; Gao et al. 2022; Soares et al. 2022) mostly owing to machine learning and it may be that blind routines and the signal loss they cause become obsolete. Forward modelling frameworks have also been proposed as a means for reconstructing foreground contaminated modes (Modi et al. 2019). Furthermore, there is a possibility that future surveys and understanding will be sufficiently sophisticated to allow precise modelling of the foregrounds without requiring removal (Fonseca & Liguori 2021). However, the reliability of these methods on real data is yet to be showcased and it is likely that blind foreground removal techniques will be the preferred method for some time. It is therefore crucial that we continue to understand how to correct for the signal loss they cause.

ACKNOWLEDGEMENTS

The authors would like to thank José Luis Bernal, Zhaoting Chen, Aishrila Mazumder, Azadeh Moradinezhad Dizgah, Paula S. Soares, and Amadeus Wild for helpful discussions. SC also would like to thank Isabelle Ye, Georgia Kiddier, and Dounia Lacroze all of whom pursued masters projects testing the foreground transfer function, supplying plenty of thought-provoking results. We also thank Matilde Barberi Squarotti for noticing a typo in the pre-print relating to the power spectrum model equation.

SC is supported by a UK Research and Innovation Future Leaders Fellowship grant [MR/V026437/1]. LW is a UK Research and Innovation Future Leaders Fellow [MR/V026437/1]. This result is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 948764; PB). PB acknowledges support from STFC Grant ST/T000341/1. IPC acknowledges support from the 'Departments of Excellence 2018–2022' Grant (L. 232/2016) awarded by the Italian Ministry of University and Research (MIUR) and from the 'Ministero degli Affari Esteri della Cooperazione Internazionale (MAECI) – Direzione Generale per la Promozione del Sistema Paese Progetto di Grande Rilevanza ZA18GR02. AP is a UK Research and Innovation Future Leaders Fellow [grant MR/S016066/2]. MS acknowledges support from the AstroSignals Synergia grant CRSII5_193826 from the Swiss National Science Foundation. MGS and MI acknowledge support from the South African Radio Astronomy Observatory and National Research Foundation (Grant No. 84156) We acknowledge the use of the Ilifu cloud computing facility, through the Inter-University Institute for Data Intensive Astronomy (IDIA). The MeerKAT telescope is operated by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation.

DATA AVAILABILITY

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

REFERENCES

Alonso D., Ferreira P. G., Santos M. G., 2014, MNRAS, 444, 3183
Alonso D., Bull P., Ferreira P. G., Santos M. G., 2015a, MNRAS, 447, 400
Alonso D., Bull P., Ferreira P. G., Maartens R., Santos M., 2015b, ApJ, 814, 145
Alvarez M. et al., 2014, preprint (arXiv:1412.4671)
Anderson C. J. et al., 2018, MNRAS, 476, 3382
Baker T., Bull P., 2015, ApJ, 811, 116
Battye R. A., Davies R. D., Weller J., 2004, MNRAS, 355, 1339
Battye R. A., Browne I. W. A., Dickinson C., Heron G., Maffei B., Pourtsidou A., 2013, MNRAS, 434, 1239
Bharadwaj S., Nath B., Nath B. B., Sethi S. K., 2001, J. Astrophys. Astron., 22, 21
Blake C., 2019, MNRAS, 489, 153
Blitz L., Rosolowsky E., 2006, ApJ, 650, 933
Bobin J., Starck J.-L., Fadili J., Moudden Y., 2007, IEEE Trans. Image Process., 16, 2662
Bull P., 2016, ApJ, 817, 26
Camera S., Santos M. G., Ferreira P. G., Ferramacho L., 2013, Phys. Rev. Lett., 111, 171302
Camera S., Maartens R., Santos M. G., 2015, MNRAS, 451, L80
Carucci I. P., Irfan M. O., Bobin J., 2020, MNRAS, 499, 304
Chang T.-C., Pen U.-L., Peterson J. B., McDonald P., 2008, Phys. Rev. Lett., 100, 091303
Chapman E. et al., 2012, MNRAS, 423, 2518
Chen Z., Wolz L., Battye R., 2023, MNRAS, 518, 2971
Cheng C. et al., 2018, ApJ, 868, 26
Creminelli P., Zaldarriaga M., 2004, J. Cosmol. Astropart. Phys., 10, 006
Croton D. J. et al., 2016, ApJS, 222, 22
Cunnington S., 2022, MNRAS, 512, 2408
Cunnington S., Camera S., Pourtsidou A., 2020, MNRAS, 499, 4054
Cunnington S., Irfan M. O., Carucci I. P., Pourtsidou A., Bobin J., 2021a, MNRAS, 504, 208
Cunnington S., Watkinson C., Pourtsidou A., 2021b, MNRAS, 507, 1623
Cunnington S. et al., 2022, MNRAS, 518, 6262
D'Amico G., Gleyzes J., Kokron N., Markovic K., Senatore L., Zhang P., Beutler F., Gil-Marín H., 2020, J. Cosmol. Astropart. Phys., 05, 005
DES Collaboration, 2022, Phys. Rev. D, 105, 023520
Dalal N., Dore O., Huterer D., Shirokov A., 2008, Phys. Rev. D, 77, 123514
Dickinson C., Davies R. D., Davis R. J., 2003, MNRAS, 341, 369
eBOSS Collaboration, 2021, Phys. Rev. D, 103, 083533
Feldman H. A., Kaiser N., Peacock J. A., 1994, ApJ, 426, 23
Fonseca J., Liguori M., 2021, MNRAS, 504, 267
Fonseca J., Camera S., Santos M., Maartens R., 2015, ApJ, 812, L22
Gao L.-Y., Li Y., Ni S., Zhang X., 2022, preprint (arXiv:2212.08773)
Giannantonio T., Porciani C., Carron J., Amara A., Pillepich A., 2012, MNRAS, 422, 2854
Harper S., Dickinson C., Battye R., Roychowdhury S., Browne I., Ma Y.-Z., Olivari L., Chen T., 2018, MNRAS, 478, 2416
Heymans C. et al., 2021, A&A, 646, A140
Irfan M. O., Bull P., 2021, MNRAS, 508, 3551
Irfan M. O., Li Y., Santos M. G., Bull P., Gu J., Cunnington S., Grainge K., Wang J., 2023, preprint (arXiv:2302.02683)
Kaiser N., 1987, MNRAS, 227, 1
Karagiannis D., Fonseca J., Maartens R., Camera S., 2021, Phys. Dark Universe, 32, 100821
Klypin A., Yepes G., Gottlober S., Prada F., Hess S., 2016, MNRAS, 457, 4340
Knebe A. et al., 2018, MNRAS, 474, 5206
Komatsu E., Spergel D. N., 2001, Phys. Rev. D, 63, 063002
Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
Li Y.-C., Ma Y.-Z., 2017, Phys. Rev. D, 96, 063525
Li Y., Santos M. G., Grainge K., Harper S., Wang J., 2021, MNRAS, 501, 4344
Liu A., Parsons A. R., Trott C. M., 2014, Phys. Rev. D, 90, 023019
Makinen T. L., Lancaster L., Villaescusa-Navarro F., Melchior P., Ho S., Perreault-Levasseur L., Spergel D. N., 2021, J. Cosmol. Astropart. Phys., 04, 081
Martinelli M. et al., 2021, A&A, 649, A100
Masui K. W. et al., 2013, ApJ, 763, L20
Matshawule S. D., Spinelli M., Santos M. G., Ngobese S., 2021, MNRAS, 506, 5075
Modi C., White M., Slosar A., Castorina E., 2019, J. Cosmol. Astropart. Phys., 11, 023
Mueller E.-M. et al., 2022, MNRAS, 514, 3396
Obuljen A., Simonović M., Schneider A., Feldmann R., 2022, preprint (arXiv:2207.12398)
Oh S. P., Mack K. J., 2003, MNRAS, 346, 871
Paul S., Santos M. G., Chen Z., Wolz L., 2023, preprint (arXiv:2301.11943)
Planck Collaboration XIII, 2016, A&A, 594, A13
Planck Collaboration VI, 2020, A&A, 641, A6
Pourtsidou A., 2023, MNRAS, 519, 6246
SKA Cosmology SWG, 2020, Publ. Astron. Soc. Aust., 37, e007
Santos M. G., Cooray A., Knox L., 2005, ApJ, 625, 575
Santos M. G. et al., 2017, Proc. Sci, MeerKAT Science: On the Pathway to the SKA. SISSA, Trieste, PoS#32
Slosar A., Hirata C., Seljak U., Ho S., Padmanabhan N., 2008, J. Cosmol. Astropart. Phys., 08, 031
Soares P. S., Watkinson C. A., Cunnington S., Pourtsidou A., 2022, MNRAS, 510, 5872
Spinelli M., Carucci I. P., Cunnington S., Harper S. E., Irfan M. O., Fonseca J., Pourtsidou A., Wolz L., 2021, MNRAS, 509, 2048
Switzer E. R. et al., 2013, MNRAS, 434, L46
Switzer E. R., Chang T.-C., Masui K. W., Pen U.-L., Voytek T. C., 2015, ApJ, 815, 51
Tassev S., Zaldarriaga M., Eisenstein D., 2013, J. Cosmol. Astropart. Phys., 06, 036
Tassev S., Eisenstein D. J., Wandelt B. D., Zaldarriaga M., 2015, preprint (arXiv:1502.07751)
Wang J. et al., 2021, MNRAS, 505, 3698
Weltman A. et al., 2020, Publ. Astron. Soc. Austral., 37, e002
Wilson T. L., Rohlfs K., Hüttemeister S., 2009, Tools of Radio Astronomy. Springer-Verlag
Witzemann A., Alonso D., Fonseca J., Santos M. G., 2019, MNRAS, 485, 5519
Wolz L., Abdalla F., Blake C., Shaw J., Chapman E., Rawlings S., 2014, MNRAS, 441, 3271
Wolz L. et al., 2017, MNRAS, 464, 4938
Wolz L. et al., 2022, MNRAS, 510, 3495
Wyithe S., Loeb A., Geil P., 2008, MNRAS, 383, 1195
Zhao C. et al., 2021, MNRAS, 503, 1149
Zoldan A., De Lucia G., Xie L., Fontanot F., Hirschmann M., 2017, MNRAS, 465, 2236

APPENDIX A: SIMULATED INTENSITY MAP DATA

In this work, we use three different simulations to allow testing of the transfer function under different scenarios. These three simulations are referred to as

(1) **MD1GPC**: The standard MULTI-DARK 1 Gpc$^3h^{-3}$ simulation we used as the default in the majority of the paper unless otherwise mentioned.

(2) **MDMK**: MULTI-DARK simulations but applied to a MeerKAT pilot-survey footprint to test applications of the transfer function in data representative of current single-dish intensity maps. Used for the investigation in Section 4 and the details of its construction are explained there.

(3) **COLA**: *N*-body simulation which can be run for large physical dimensions to allow investigation into the robustness of transfer

function on ultra-large-scales in future data sets. Used only for the $f_{\rm NL}$ investigation in Section 5.2.

The MDMK and COLA simulations are mostly extensions of the MD1GPC simulations, explained in the relevant sections (Section 4 and Section 5.2 respectively). The details of the default MDMK simulation are outlined below.

A1 Multi-Dark H I simulation (MD1GPC)

For our main simulated H I intensity maps, which are used in all parts of the paper unless clearly stated, we use the same simulations as those adopted in (Cunnington et al. 2021a; Cunnington et al. 2021b). These used the MultiDark-Galaxies *N*-body simulation data (Knebe et al. 2018) and the catalogue produced from the SAGE (Croton et al. 2016) semi-analytical model application. These galaxies were produced from the dark matter cosmological simulation MultiDark-Planck (MDPL2) (Klypin et al. 2016), which follows the evolution of 3840^3 particles in a cubical volume of $1\,(\mathrm{Gpc}/h)^3$ with mass resolution of $1.51 \times 10^9 h^{-1}\mathrm{M}_\odot$ per dark matter particle. The cosmology adopted for this simulation is based on Planck15 cosmological parameters (Planck Collaboration XIII 2016), with $\Omega_{\rm m} = 0.307$, $\Omega_{\rm b} = 0.048$, $\Omega_\Lambda = 0.693$, $\sigma_8 = 0.823$, $n_{\rm s} = 0.96$ and Hubble parameter $h = 0.678$. The catalogues are split into 126 snapshots between redshifts $z = 17$ and $z = 0$. In this work we chose low-redshift, post-reionization data to test the transfer function and use the snapshot at $z = 0.39$ to emulate a MeerKAT-like survey performed in the L-band ($899 < \nu < 1184$ MHz, or equivalently $0.2 < z < 0.58$). Although there is no reason to suspect conclusions will be any different for any reasonable redshift choice between $0 < z < 3$. We obtained this publicly available data from the Skies & Universes web page.[7]

We used each galaxies (x, y and z) coordinates and placed them on to a grid with $n_{\rm x}, n_{\rm y}, n_{\rm z} = 256, 256, 256$ pixels and $1\,(\mathrm{Gpc}/h)^3$ in physical size. To simulate observations in redshift space inclusive of RSD, we used the peculiar velocities of the galaxies. Assuming the LoS is along the z-dimension and given the plane-parallel approximation is exact for this Cartesian data, RSD can be simulated by displacing each galaxy's position to a new coordinate $z_{\rm RSD}$ given by $\mathrm{z_{RSD}} = \mathrm{z} + v_\parallel(1+z)h/H(z)$, where $v_\parallel$ is the galaxy's peculiar velocity along the LoS (z-dimension) which is given as an output of the simulation in units of $\mathrm{km\,s^{-1}}$.

To simulate the contribution to the signal from each galaxy, we used the cold gas mass $M_{\rm cgm}$ output from the MultiDark data and from this we can infer an H I mass with $M_{\rm HI} = f_{\rm H} M_{\rm cgm}(1 - f_{\rm mol})$ where $f_{\rm H} = 0.75$ represents the fraction of hydrogen present in the cold gas mass and the molecular fraction is given by $f_{\rm mol} = R_{\rm mol}/(R_{\rm mol} + 1)$ (Blitz & Rosolowsky 2006), with $R_{\rm mol} \equiv M_{H_2}/M_{\rm HI} = 0.4$ (Zoldan et al. 2017). It is this H I mass that we binned into each voxel with position $\boldsymbol{x}$, to generate a data cube of H I masses $M_{\rm HI}(\boldsymbol{x})$, which should trace the underlying matter density generated by the catalogue's *N*-body simulation for the snapshot redshift z. These H I masses are converted into an H I brightness temperature for a frequency width of $\delta\nu$ subtending a solid angle $\delta\Omega$ given by

$$T_{\rm HI}(\boldsymbol{x}, z) = \frac{3 h_{\rm P} c^2 A_{12}}{32\pi m_{\rm h} k_{\rm B} \nu_{21}} \frac{1}{[(1+z) r(z)]^2} \frac{M_{\rm HI}(\boldsymbol{x})}{\delta\nu\, \delta\Omega}, \tag{A1}$$

where $h_{\rm P}$ is the Planck constant, A_{12} the Einstein coefficient that quantifies the rate of spontaneous photon emission by the hydrogen atom, $m_{\rm h}$ is the mass of the hydrogen atom, $k_{\rm B}$ is Boltzmann's constant, ν_{21} the rest frequency of the 21cm emission and $r(z)$ is the comoving distance out to redshift z (we will assume a flat universe). Since H I simulations on this scale have a finite halo-mass resolution, there will be some contribution from the H I within the lowest mass host haloes which is not included in the final $T_{\rm HI}$ signal. To account for this, it is typical for a rescaling of the final $T_{\rm HI}$ to be performed to bring the mean H I temperature, $\overline{T}_{\rm HI}$, in agreement with the modest data constraints we have for this value. For the effective redshift of our data, $z = 0.39$, we used a fiducial value of $\overline{T}_{\rm HI} = 0.0743$ mK which our maps were re-scaled to.

MD1GPC foreground simulations

To produce the foreground contamination we simulated different foreground processes, including galactic synchrotron, free–free emission and point source emission. We also included the effects of polarization leakage which will act as an extra component of foreground with non-smooth spectra (Cunnington et al. 2021a), thus posing an increased challenge for the foreground clean. The foregrounds we used can thus be decomposed as $T_{\rm fg} = T_{\rm sync} + T_{\rm free} + T_{\rm point} + T_{\rm pol}$, which represent the synchrotron, free–free, point sources and polarization leakage.

We briefly summarize the simulation technique for these components but for a full outline we refer the reader to Cunnington et al. (2021a) and Carucci et al. (2020) where they were also used. The synchrotron emission is based on Planck Legacy Archive[8] FFP10 simulations of synchrotron emission at 217 and 353 GHz formed from the source-subtracted and destriped 0.408 GHz map. The free–free simulation is from the FFP10 217 GHz free–free simulation which is a composite of the Dickinson, Davies & Davis (2003) free–free template and the WMAP MEM free–free templates. The point sources are based on the empirical model of Battye et al. (2013) and makes the assumption that point sources over 10 mJy will be identifiable and thus can be removed. Lastly, we simulated polarization leakage with the use of the `CRIME`[9] software (Alonso et al. 2014), which provides maps of Stokes Q emission at each frequency and we fix the polarization leakage to 0.5 per cent of the Stokes Q signal.

For the foregrounds we assumed they have been observed in a frequency range of $900 < \nu < 1156$ MHz, consistent with the $z = 0.39$ redshift for the cosmological simulation. Each of the 256 map slices along the z-direction acts as an observation in a frequency channel giving a channel width of $\delta\nu = 1$ MHz. This therefore emulates the spectral distinction between the cosmological H I and foregrounds utilized in the foreground clean. From the full-sky foreground map we cut a region of sky centred on the Stripe 82, a field well observed by surveys. The size of this sky region is $54.1 \times 54.1\,\mathrm{deg}^2$ which corresponds to the size of a $1\,(\mathrm{Gpc}/h)^2$ patch at the $z = 0.39$ snapshot redshift of our cosmological simulation.

A2 Instrumental effects

Here, we outline some of the additional observational effects which we switch on and off for different scenarios throughout the paper. Unless clearly mentioned, the reader can assume these effects have not been included for simplicity.

[7] www.skiesanduniverses.org

[8] pla.esac.esa.int/pla

[9] intensitymapping.physics.ox.ac.uk/CRIME.html

Telescope beam

The effect from the telescope beam is a smoothing to the temperature field in directions perpendicular to the LoS. A simple, and often sufficient, method to simulate these beam effects is to convolve the density field with a Gaussian kernel whose FWHM ($\theta_{\rm FWHM}$) is chosen to match the model of the radio telescope one is trying to emulate. We can define this Gaussian smoothing kernel with

$$\mathcal{B}_{\rm G}(\nu, \boldsymbol{s}_\perp) = \exp\left[-4\ln 2\left(\frac{\boldsymbol{s}_\perp}{r(\nu)\,\theta_{\rm FWHM}(\nu)}\right)^2\right] = \exp\left[\frac{1}{2}\left(\frac{\boldsymbol{s}_\perp}{R_{\rm beam}}\right)^2\right], \tag{A2}$$

where $\boldsymbol{s}_\perp = \sqrt{\Delta x^2 + \Delta y^2}$ is the perpendicular spatial separation from the centre of the beam. $R_{\rm beam} = r(z)\,\sigma_{\rm beam}$ defines the physical size of the beam's central lobe in Mpc/h, where $\sigma_{\rm beam} = \theta_{\rm FWHM}/(2\sqrt{2\ln 2})$ represents the standard deviation of the Gaussian kernel in radians. $R_{\rm beam}$ is dependent on frequency through the comoving distance out to the density fluctuations which changes with frequency ($r(\nu)$). It also has a further frequency dependence from the intrinsic beam size of the instrument, which is itself a function of frequency, generically given by $\theta_{\rm FWHM} \approx c/\nu D_{\rm dish}$, where $D_{\rm dish}$ is the diameter of the radio telescope dish.

As we discussed in the main body of the paper, including a more sophisticated model of the beam is worth investigating since this could have implications for foreground removal, signal loss, and thus signal reconstruction. A more complex beam with far-reaching side lobes and frequency dependence has been shown to create additional challenges for foreground cleaning (Matshawule et al. 2021; Spinelli et al. 2021). This is an upgrade which will be targeted in future work.

Instrumental noise

An unavoidable source of noise in intensity mapping comes from the thermal motion of electrons inside the electronics of the instrument which produce Gaussian-like fluctuating currents, with a mean current of zero but a non-zero rms. The consequence from this is a component of white-noise contained in the maps. From the radiometer equation, the rms of the thermal noise contained in time-ordered data for an instrument with system temperature $T_{\rm sys}$, frequency resolution $\delta\nu$ and time per pointing $t_{\rm p}$, will be given by (Wilson, Rohlfs & Hüttemeister 2009)

$$\sigma_{\rm n} = T_{\rm sys}\Big/\sqrt{2\,\delta\nu\,t_{\rm p}}. \tag{A3}$$

At map level this will create a field of white noise added into the data, with rms $\sigma_{\rm n}$. In the case of the power spectrum this produces an additive component; $P_{\rm HI} \to P_{\rm HI} + P_{\rm N}$ where $P_{\rm N} = \sigma_{\rm n}^2/V_{\rm cell}$. Since this should be uncorrelated and independent at different observation times and for different dishes, this thermal noise can be averaged down as survey time increases. Thus, it is not seen as a major problem for future intensity mapping surveys where long observation campaigns will be conducted. In all cases where we include noise in the simulations we use Gaussian white noise with a value of $\sigma_{\rm n} = 1$ mK. This is designed to dominate over the H I which has an rms of $\sigma_{\rm HI} \sim 0.14$ mK. The time per pointing is defined as

$$t_{\rm p} = N_{\rm dish} t_{\rm obs}\,(\theta_{\rm FWHM}/3)^2\,/A_{\rm sky}, \tag{A4}$$

where we have assumed the pixel size will be 1/3 of the beam size. For a MeerKAT-like $A_{\rm sky} \sim 3{,}000$ deg^2 survey with $N_{\rm dish} = 64$ dishes, $\delta\nu = 0.2$ MHz frequency resolution and $T_{\rm sys} = 16$ K (Wang et al. 2021), the $\sigma_{\rm n} = 1$ mK dominant noise will correspond to $t_{\rm obs} \sim 30$ hrs of observation time.

APPENDIX B: POWER SPECTRUM ESTIMATION

Here, we briefly outline the method for measuring power spectra, used throughout the paper. This follows the same methodology as the MeerKAT intensity mapping pipeline (Cunnington et al. 2022).

We define the Fourier transform of the H I intensity maps $\delta T_{\rm HI}$ as

$$\tilde{F}_{\rm HI}(\boldsymbol{k}) = \sum_{\boldsymbol{x}} \delta T_{\rm HI}(\boldsymbol{x}) w_{\rm HI}(\boldsymbol{x}) \exp(i\boldsymbol{k}\cdot\boldsymbol{x}), \tag{B1}$$

where $w_{\rm HI}$ are the weights that can be applied to optimize the power spectrum measurement. For our simulations we simply assume $w_{\rm HI} = 1$ everywhere. The H I power spectrum is then estimated by

$$\hat{P}_{\rm HI}(\boldsymbol{k}) = \frac{V_{\rm cell}}{\sum_{\boldsymbol{x}} w_{\rm HI}^2(\boldsymbol{x})} |\tilde{F}_{\rm HI}(\boldsymbol{k})|^2. \tag{B2}$$

In many of the results, we use the cross-correlations between an H I-only (foreground-free) simulation $\tilde{F}_{\rm HI}$ and a foreground-cleaned one $\tilde{F}_{\rm clean}$. In this case the cross-correlation is similarly defined by

$$\hat{P}_{\rm X}(\boldsymbol{k}) = \frac{V_{\rm cell}}{\sum_{\boldsymbol{x}} w_{\rm HI}^2(\boldsymbol{x})} {\rm Re}\left\{\tilde{F}_{\rm HI}(\boldsymbol{k})\cdot\tilde{F}^*_{\rm clean}(\boldsymbol{k})\right\}. \tag{B3}$$

These power spectra are either spherically averaged into bandpowers $|\boldsymbol{k}| \equiv k$ to provide the 1D power spectra results, or cylindrically averaged into $k_\perp$, $k_\parallel$ bins to produce the demonstrative 2D power spectra plots.

B1 Modelling the H I intensity mapping power spectrum

Wherever we require a model for the observational simulations, we use the below prescription;

$$P_{\rm mod}(k,\mu) = \overline{T}_{\rm HI}^2 \frac{\left(b_{\rm HI}^2 + f\mu^2\right)^2}{1 + (k\mu\sigma_{\rm v}/H_0)^2} P_{\rm m}(k) \times \exp\left[-(1-\mu^2)k^2 R_{\rm beam}^2\right], \tag{B4}$$

where $b_{\rm HI}$ is the linear bias for the H I field and $\overline{T}_{\rm HI}$ is the mean H I temperature in mK and approximately related to the H I density fraction by

$$\overline{T}_{\rm HI}(z) = 180\,\Omega_{\rm HI}(z)\,h\,\frac{(1+z)^2}{\sqrt{\Omega_{\rm m}(1+z)^3 + \Omega_\Lambda}}\ {\rm mK}, \tag{B5}$$

where $\Omega_{\rm m}$ and Ω_Λ are the density fractions for matter and the cosmological constant, respectively. The linear RSD are accounted for in equation (B4) by the $f\mu^2$ term (Kaiser 1987), where f is the growth rate of structure and μ is the cosine of the angle from the line of sight. In the denominator, we approximately account for the non-linear effects of RSD, commonly referred to as Fingers-of-God, and $\sigma_{\rm v}$ is the velocity dispersion of the H I, with H_0 as the Hubble constant. $P_{\rm m}$ is the matter power spectrum produced using `CAMB`[10] (Lewis, Challinor & Lasenby 2000) with a Planck18 (Planck Collaboration VI 2020) cosmology. The exponential factor accounts for the smoothing of perpendicular modes due to the beam, where $R_{\rm beam}$ is the standard deviation of the Gaussian beam profile in comoving units, as explained in Appendix A.

[10] camb.readthedocs.io/en/latest/

APPENDIX C: EXAMPLES OF BIASED TRANSFER FUNCTIONS

As demonstrated in Section 2 and explicitly highlighted in Switzer et al. (2015), when estimating the impact from blind foreground cleaning on the signal, it is insufficient to consider only the direct loss to the signal in the removed $\mathsf{U}_\mathrm{f}\mathsf{S}\mathsf{U}_\mathrm{f}^\mathrm{T}\mathsf{X}_\mathrm{s}$ piece. One must also include the impact of spurious correlations caused by the presence of non-foreground data such as the H I signal itself and the perturbations $\mathbf{\Delta}$ these cause to the estimated eigenmodes. This is particularly important when considering how to estimate signal loss using mock data. For example, adopting an approach which simply looks to project out the observed data modes $\mathsf{U}_\mathrm{f+s}$ from realizations of mock data X_m, will neglect the signal loss from the perturbed terms. Even though the foreground modes are perturbed by the true signal X_s, cross-terms from these perturbations will be uncorrelated with the mock signal X_m, which is what matters in the construction of the transfer function where signal loss to the mocks is being evaluated. As an example, if we assume the cleaned mocks can be defined by

$$\mathsf{X}^\mathrm{m}_\mathrm{clean:bias1} = \mathsf{X}_\mathrm{m} - \mathsf{U}_\mathrm{f+s}\mathsf{S}\mathsf{U}^\mathrm{T}_\mathrm{f+s}\mathsf{X}_\mathrm{m}, \tag{C1}$$

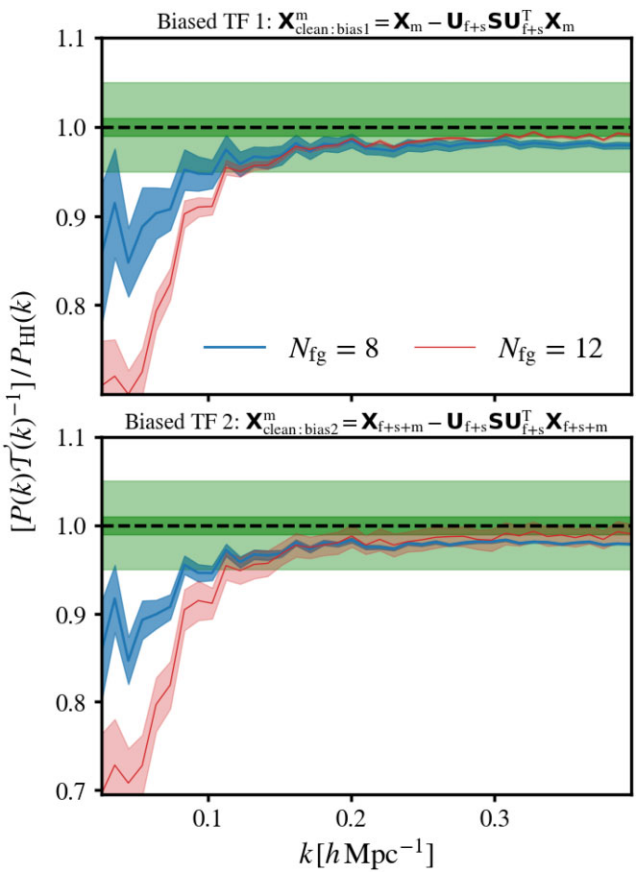

Figure C1. Same as Fig. 4 but for versions of the transfer function that deliver biased results. The transfer functions versions vary based on their definition of $\mathsf{X}^\mathrm{m}_\mathrm{clean}$, which are defined by the panel titles in each version and are discussed further in the appendix text. All transfer functions are calculated by averaging over 100 lognormal mocks and the shaded bands show the rms over these 100 mocks. For each version, results for a mild ($N_\mathrm{fg} = 8$, blue lines) and more aggressive ($N_\mathrm{fg} = 12$, red lines) foreground cleans are shown. Dark (light) green regions indicate sub 1 per cent (5 per cent) accuracy of the reconstructed power spectrum.

then use this in the transfer function (equation 11), the resulting reconstructed power spectrum will be slightly biased (> 1 per cent) on small scales (as shown by the top panel of Fig. C1), and more extremely biased on larger scales, reaching 10 per cent departure from the truth at $k \sim 0.1\,h\,\mathrm{Mpc}^{-1}$ for the $N_\mathrm{fg} = 12$ case. This is because the source of the perturbations to the eigenmodes, which in this case is only the true signal $\mathsf{U}_\mathrm{f+s} = \mathsf{U}_\mathrm{f} + \Delta_\mathrm{s}$, is being projected out of data (X_m) which will have no correlation with these perturbations, thus this contribution is neglected, hence the bias. This also explains why the bias is worse for higher N_fg, because the neglected correlations are larger for higher N_fg (shown by Fig. 3).

A slight improvement can be attempted on equation (C1) by projecting out the data modes $\mathsf{U}_\mathrm{f+s}$ over the true data with mock signal injected $\mathsf{X}_\mathrm{f+s+m}$

$$\mathsf{X}^\mathrm{m}_\mathrm{clean:bias2} = \mathsf{X}_\mathrm{f+s+m} - \mathsf{U}_\mathrm{f+s}\mathsf{S}\mathsf{U}^\mathrm{T}_\mathrm{f+s}\mathsf{X}_\mathrm{f+s+m}. \tag{C2}$$

However, as the bottom panel of Fig. C1 shows, the bias is still present since it still lacks any correlation in perturbed modes and mock signal, thus fails to emulate the correlations between signal and foregrounds.

These results from the biased versions of the transfer function thus highlight the importance of emulating the spurious correlations between foregrounds and signal in the construction of the transfer function. The correct approach from equation (10) should thus always be adopted. Along with the discussion of this point in Switzer et al. (2015), it has also been investigated in epoch of reionization studies (Cheng et al. 2018) where it was acknowledged how neglecting these additional complications leads to an underestimation of the signal loss.

APPENDIX D: ITERATIVE FOREGROUND TRANSFER FUNCTION & A BAYESIAN APPROACH?

Our investigation which varied the input parameters for the mocks used in the transfer function construction provided encouraging results (demonstrated by Figs 17 and 18). This showed that the transfer function only has a very mild dependence on the mock input parameters. For the parameters we tested, we found they can be highly biased relative to the truth in observed data, but this does not have a significant impact on the reconstructed power spectra. However, some dependence was still noticed, and if one is striving to maximize accuracy in parameter inference then this dependence may be enough to cause concern.

We raised the idea in the main text of an *iterative* transfer function. Since the reconstructed power spectra show good agreement with the truth despite large mock parameter biases, there is a strong possibility that the parameters inferred from a reconstructed power spectrum will be much closer to the truth. Using these updated parameter estimates to seed new mocks and construct a new transfer function, it is highly likely the new reconstructed power spectrum based on the updated transfer function will have an even better agreement with the truth. This process could then be repeated indefinitely, iteratively improving the accuracy of the transfer function until convergence on all inferred parameters is achieved.

A challenge for an iterative transfer function is the computational demand required to repeatedly calculate one. The larger surveys become, the larger the mocks need to be thus computational expense will only increase. Thus, time wasted on mock generation when convergence has already been reached would be an unnecessary bottleneck. To provide some guidance on the issue we consider how many mocks are required for a stable transfer function. In Fig. D1

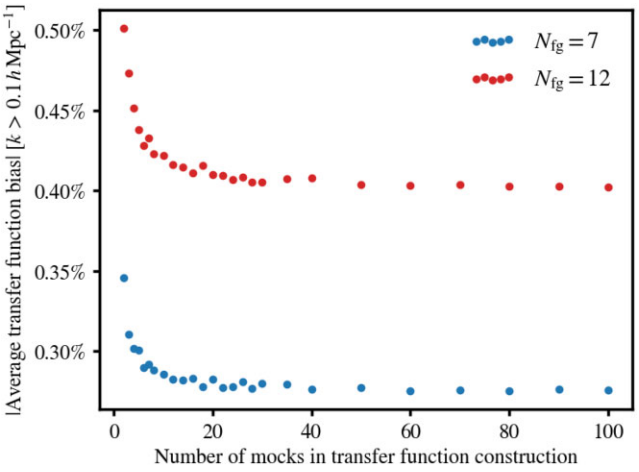

Figure D1. Number of mocks required in transfer function computation to reach a converged level of accuracy in the reconstructed power spectrum for two levels of foreground cleaning given by N_{fg}. The y-axis shows the mean bias for each number of mocks in the reconstructed power spectrum at $k > 0.1\,h\,\mathrm{Mpc}^{-1}$ values relative to the truth (H I-only power). We average the *absolute values* of the the power spectrum biases to avoid potential cancellation to zero in highly fluctuating results around zero. For each number of mocks tested, we average over 100 realized combinations to get a stable accuracy level.

we investigate how many mocks are required before the accuracy of the transfer function reaches a converged level. We define the accuracy of the transfer function by the average bias across k-modes for the reconstructed power spectrum relative to the original H I-only power i.e. $[P(k)\mathcal{T}(k)^{-1}]/P_{\mathrm{H\,I}}(k)$. In this accuracy calculation, we ignore large scales ($k < 0.1\,h\,\mathrm{Mpc}^{-1}$) where accuracy fluctuations can be quite large. We calculate the transfer function using different numbers of mocks and for each number tested we average over 100 iterations so that the returned accuracy is stable.

Fig. D1 suggests that when using just two mocks to construct the transfer function, excellent accuracy is already achievable, although these results would be prone to fluctuating performance based on the two mocks used each time. We see that accuracy can be improved by increasing the number of mocks but convergence is quickly reached at around 20. For the different levels of foreground clean, shown by the different N_{fg} results, the accuracy levels differ with the higher N_{fg} returning poorer accuracy as expected due to the increased contribution from the spurious foreground and signal correlations for higher N_{fg}. However, we see that convergence is reached at a consistent number of mocks for both N_{fg} cases. Despite the evidence suggesting that only 20 mocks would be needed for any reasonably foreground clean, this would need to be tested further in real cases where the data could be more contaminated and thus may require a higher number of mocks for convergence. However, even in reality, the convergence level could be tested to ensure a Bayesian analysis is not computing an unnecessary number of mocks. What is encouraging from Fig. D1 is how quick convergence is reached even for our fairly complex foreground simulations. We remind the reader that these mocks are only lognormal mocks and thus can be generated rapidly.

This paper has been typeset from a TEX/LATEX file prepared by the author.