Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **524,** 3724–3740 (2023) https://doi.org/10.1093/mnras/stad2102
Advance Access publication 2023 July 13

# Detecting the H I power spectrum in the post-reionization Universe with SKA-Low

Zhaoting Chen [1]★ Emma Chapman [2] Laura Wolz[1] and Aishrila Mazumder[1]

[1] *Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, The University of Manchester, Manchester M13 9PL, UK*
[2] *School of Physics and Astronomy, The University of Nottingham, Nottingham, NG7 2RD, UK*

Accepted 2023 July 6. Received 2023 July 6; in original form 2023 February 22

## ABSTRACT

We present a survey strategy to detect the neutral hydrogen (H I) power spectrum at $5 < z < 6$ using the SKA-Low radio telescope in presence of foregrounds and instrumental effects. We simulate observations of the inherently weak H I signal post-reionization with varying levels of noise and contamination with foreground amplitudes equivalent to residuals after sky model subtraction. We find that blind signal separation methods on imaged data are required in order to recover the H I signal at large cosmological scales. Comparing different methods of foreground cleaning, we find that Gaussian Process Regression (GPR) performs better than Principle Component Analysis (PCA), with the key difference being that GPR uses smooth kernels for the total data covariance. The integration time of one field needs to be larger than ∼250 h to provide large enough signal-to-noise ratio (SNR) to accurately model the data covariance for foreground cleaning. Images within the primary beam field-of-view give measurements of the H I power spectrum at scales $k \sim 0.02\,\mathrm{Mpc}^{-1} - 0.3\,\mathrm{Mpc}^{-1}$ with SNR ∼2–5 in $\Delta[\log(k/\mathrm{Mpc}^{-1})] = 0.25$ bins assuming an integration time of 600 h. Systematic effects, which introduce small-scale fluctuations across frequency channels, need to be $\lesssim 5 \times 10^{-5}$ to enable unbiased measurements outside the foreground wedge. Our results provide an important validation towards using the SKA-Low array for measuring the H I power spectrum in the post-reionization Universe.

**Key words:** techniques: interferometric – (*cosmology*:) large-scale structure of Universe – radio lines: general.

## 1 INTRODUCTION

The standard model of cosmology, the Λ cold dark matter (ΛCDM) model, helps us describe and understand the observed Universe. In particular, measurements of the cosmic microwave background (CMB; e.g. Planck Collaboration 2020a) and the large-scale structure (LSS; e.g. Alam et al. 2021) can be well fitted by the ΛCDM model, producing precise, per-cent level constraints on the model parameters. However, as we reach further into the realm of precision cosmology, potential inconsistency between different probes arises in the form of cosmological tensions. Namely, measurements of the Hubble parameter in the local Universe using tip of the red-giant branch and Type Ia Supernovae (e.g. Riess et al. 2022) have significant discrepancies ∼5$\sigma$ with the measurements made using the CMB (e.g. Planck Collaboration 2020b). There also exists a tension of ∼2.7$\sigma$ between the measurements of the amplitude of the dark matter clustering $S_8$ from the CMB and from the LSS (e.g. Amon et al. 2022).

The disagreements between different cosmological observations highlight the need for understanding the evolutionary history of the Universe. The CMB captures the cosmic structure at the last scattering surface $z \sim 1100$ (Dodelson & Schmidt 2020) while the local measurements are made at $z \lesssim 2.0$, missing a large part of the observable Universe in between. One promising approach to fill the gap is neutral hydrogen (H I) intensity mapping (e.g. Battye, Davies & Weller 2004; Chang et al. 2008; Mao et al. 2008; Wyithe & Loeb 2009; Battye et al. 2013; Kovetz et al. 2017). It uses the emission line of the H I atoms, at the rest wavelength of ∼ 21 cm, as a tracer of the underlying dark matter distribution. Neutral hydrogen is the most abundant element in the Universe after recombination as predicted by the Big Bang nucleosynthesis (Alpher, Bethe & Gamow 1948; Dodelson & Schmidt 2020). The formation of dark matter structures, i.e. dark matter halos, attracts baryonic matter to fall into the halos and produces luminous stars and galaxies during the cosmic dawn (Schaerer 2002). The ultra-violet radiation produced by these objects ionized the initially neutral inter-galactic medium (IGM), a process known as the cosmic reionization (Furlanetto, Oh & Briggs 2006). The 21-cm emission is dominated by the H I inside the IGM during the cosmic reionization, after which the majority of the remaining H I resides in the dark matter halos (Rahmati et al. 2013). Therefore, the H I signal traces different cosmic structures during different epochs and can be used to probe cosmology across a wide range of redshifts.

The spectroscopic nature of the 21-cm line allows the measurement of the matter clustering across the history of structure formation from the cosmic Dark Ages, to the Epoch of Reionization (EoR), and all the way to the low-redshift Universe. However, the H I signal is inherently weak, and resolving H I sources requires deep integration time even for observing the H I galaxies in the local Universe (e.g. Haynes et al. 2018). Without the need to resolve individual sources

★ E-mail: zhaoting.chen@manchester.ac.uk

of the H I emission, intensity mapping is a technique that maps the 21-cm emission across a large area of the sky with relatively coarse angular resolution, allowing efficient surveys of large cosmological volumes suitable for testing the ΛCDM model. Ongoing experiments targeting different redshifts include MeerKAT (Santos et al. 2016), Canadian Hydrogen Intensity Mapping Experiment (CHIME; CHIME Collaboration 2022), Tianlai (Xu, Wang & Chen 2015), Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017), Low-Frequency Array (LOFAR; Patil et al. 2017), Murchison Widefield Array (MWA; Tingay et al. 2013), and more, covering $z \sim 0.0$–$10.0$. In the future, the Square Kilometre Array Observatory (SKAO) will further enable detections of the neutral hydrogen clustering, with SKA-Low observing at 50–350 MHz, covering the redshift range from the cosmic Dark Ages $z \sim 27$ down to the post-EoR Universe $z \sim 3.0$ (Koopmans et al. 2015), and SKA-Mid observing at 350 MHz to 15.4 GHz covering $z \lesssim 3$ (Square Kilometre Array Cosmology Science Working Group et al. 2020).

The biggest challenge of H I intensity mapping is measuring the signal against the foregrounds which are several orders of magnitude brighter than the H I. In order to measure the H I signal, extreme accuracy in the instrument calibration is necessary to model the foregrounds (Barry et al. 2016). The desired calibration accuracy calls for a thorough understanding of the sky (e.g. Trott & Wayth 2017; Murray & Trott 2018), the beam (e.g. Thyagarajan et al. 2015; Ewall-Wice et al. 2016b), and the systematics (e.g. Trott et al. 2018). Techniques of foreground mitigation can then be utilized to extract the H I signal. The spectral smoothness of the foregrounds contrasts with the H I which is discretely structured in frequency since, for the H I, different frequencies correspond to different redshifts, and therefore different line-of-sight (LOS) distances. Fourier transformation along the frequency direction to the delay time space for individual baselines, a technique called the 'delay transform', can thus be used to isolate modes of the power spectrum where the H I dominates (Morales & Hewitt 2004; Parsons et al. 2012a, b). The region of the wavenumber $k$-space where H I signal can be measured is called the 'observation window' whereas the region dominated by the foregrounds is the 'foreground wedge' (Datta, Bowman & Carilli 2010; Morales et al. 2012; Liu, Parsons & Trott 2014). Measuring the H I power spectrum in the observation window is usually referred to as 'foreground avoidance', which is one approach among ongoing efforts of measuring the EoR signal. Alternatively and/or additionally, blind signal separation (BSS) techniques can also be applied on the foregrounds, or the residuals of them after sky model subtraction. These techniques work mostly on the frequency–frequency covariance of the data, such as fast independent component analysis (fastICA, Chapman et al. 2012; Wolz et al. 2014), generalized morphological component analysis (GMCA; Chapman et al. 2013), correlated component analysis (CCA; Bonaldi & Brown 2015), Gaussian process regression (GPR; Mertens, Ghosh & Koopmans 2018), and more (see Chapman & Jelić 2019 for a review). For H I observations targeting the post-reionization Universe, foreground removal using BSS methods is typically used to recover the H I signal, with transfer function corrections of signal loss (e.g. Switzer et al. 2015; Cunnington et al. 2023a).

Using the methods mentioned above, progress has been made at different redshifts towards the detection of the H I power spectrum. For single dish experiments targeting the low-redshift Universe, cross-correlation detections of the H I signal with optical galaxies have been made by the *Green Bank Telescope* (Masui et al. 2013; Switzer et al. 2013; Wolz et al. 2022), the *Parkes* telescope (Anderson et al. 2018), and the *MeerKAT* telescope (Cunnington et al. 2023b). A similar cross-correlation measurement has also been made by the CHIME telescope using stacking (CHIME Collaboration 2023). The first auto-correlation detection has been made using the *MeerKAT* telescope as a radio interferometer (Paul et al. 2023). For experiments targeting EoR, upper limits on the H I power spectrum have been found by the MWA (Ewall-Wice et al. 2016a; Trott et al. 2020) and HERA (The HERA Collaboration 2022) using the delay transform and foreground avoidance, and by LOFAR using map making with GMCA and GPR foreground removal (Patil et al. 2017; Mertens et al. 2020).

In light of the recent progress, in this paper we explore the possibility of measuring the H I power spectrum at $5 < z < 6$ using SKA-Low. While this redshift range is within the frequency coverage of the instrument, it has been largely neglected since it is not in the interests of the primary goal of H I science for SKA-Low, which mainly focuses on the EoR (Koopmans et al. 2015). Despite probing different physics, observations of the post-reionization Universe can benefit significantly from the wide frequency range of the SKA-Low telescope, as the deep observations of the EoR fields will provide accurate modelling of the radio continuum and the instrument. Furthermore, it has been suggested that the Universe may still be partially ionized at $z \sim 5.5$ (Bosman et al. 2022), in contrast with conventional constraints on the end of reionization to be at $z \sim 6$ (e.g. Fan et al. 2006). Using the H I power spectrum at $5 < z < 6$ provides a unique method of constraining the end of reionization. However, measuring the H I clustering at these redshifts has its own challenges. The H I signal at the quasi-linear scales probed at $5 < z < 6$ will be lower than the signal at the EoR. Meanwhile, the low-frequency band contains more foreground contamination than the L-band typically used for intensity mapping at lower redshifts. It is important to quantify the signal and foreground level at these frequencies as well as the instrument effects, to verify if these redshifts can be used for cosmology.

In this paper, we present an end-to-end pipeline including simulations of the sky signals and the interferometric observations, the foreground mitigation, and the power spectrum estimation to provide a proof-of-concept study for measuring the H I power spectrum at $5 < z < 6$ using SKA-Low. Using the simulation pipeline with different settings, we explore different levels of foreground residual and noise level to find the requirements on integration time and foreground modelling needed. Methods for residual foreground removal are investigated focusing on the comparison between Principle Component Analysis (PCA) and GPR, with quantitative investigations into the differences in the performance of these two methods. We present our forecasts for future SKA-Low surveys on the power spectrum measurements. Impacts of systematics are also briefly discussed to provide an estimation of the requirements on levels of the systematics.

The paper is organized as follows: The simulation of the sky signal is described in Section 2. Simulations of the interferometric observations to get the images and subsequent power spectrum estimation from the images are discussed in Section 3. The presence and the structure of the foreground wedge, with foreground mitigation methods applied, are quantified in Section 4. The robustness of the foreground mitigation methods is tested in the presence of thermal noise and systematic effects in Section 5. We present the concluding remarks in Section 6. Throughout this paper, we assume the ΛCDM cosmology from Planck Collaboration (2020b).

## 2 SIMULATIONS OF THE RADIO SKY

In this section, we outline the simulations of the sky signal which consist of the H I signal and the foregrounds at $5 < z < 6$,

corresponding to ∼200–240 MHz. The SKA-Low instrument is designed to have a maximum channel resolution of 5.4 kHz (Braun et al. 2019). Since we are only interested in the H I intensity mapping which uses large voxels to map the distribution of the H I emission, we reduce the simulated data volume by assuming the redshift bin is covered by 66 frequency channels with a channel bandwidth of 510 kHz. The coarser frequency resolution of 510 kHz corresponds to $k_{\parallel} \sim 0.4\,\mathrm{Mpc}^{-1}$. While increasing the frequency resolution gives access to higher $k_{\parallel}$ where the foregrounds are weaker, the small scales beyond BAO wiggles are difficult to model for cosmological inferences. We leave simulations with the full frequency resolution for future work.

The primary beam field-of-view (FoV) for SKA-Low at these frequencies is ∼3 degrees (Braun et al. 2019). We simulate $(10.5\,\mathrm{deg})^2$ sky areas around the pointing centre for all the components of the sky signal. While the sky area only extends to the −20 dB (1 per cent) sidelobes of the primary beam, we find that there is no sharp features in the cylindrical power spectrum from simulated foreground residuals (see Appendix B). As discussed later in Section 3, we perform the power spectrum estimation using only the centre $(1.5\,\mathrm{deg})^2$ and therefore the $(10.5\,\mathrm{deg})^2$ sky area is sufficient. The pointing centre is at the EoR0 field (Lynch et al. 2021) at RA = 0 h, Dec = −27 deg. The methods for generating the components are described as follows.

### 2.1 Diffuse Galactic radiation

The diffuse Galactic radiation at these scales is dominated by the synchrotron radiation. We use the all-sky 'Haslam map' of synchrotron radiation at 408 MHz (Haslam et al. 1981, 1982) with the updated version described in Remazeilles et al. (2015). The map is then extrapolated to the frequencies of interest using the Global Sky Model (Zheng et al. 2017) at 1.4 and 2.3 GHz to calculate the spectral indices of the map pixels. The curvature of the spectral indices (see e.g. Irfan et al. 2022) is neglected for simplicity.

The pixel size of the input Haslam map is $(1.72\,\mathrm{arcmin})^2$, corresponding to HEALPIX (Górski et al. 2005; Zonca et al. 2019) NSIDE = 2048. An image of $(10.5\,\mathrm{deg})^2$ around the pointing centre is created with a pixel size of 21 arcsec. The image is then Gaussian smoothed with a resolution of 1.75 arcmin. The input synchrotron radiation at the central frequency of our simulation 220 MHz is shown in Fig. 1.

Free–free emission from the Galactic electrons also contributes to the diffuse Galactic radiation. Following Lian et al. (2020), we use FG21SIM [1] to simulate the Galactic free–free emission. It is based on the Hα intensity map in Finkbeiner (2003). The free–free emission in the frequency range of our interest is several orders of magnitude smaller than the synchrotron as shown in Fig. 1.

As discussed later in Section 3.1, we make image cubes of the observations to perform residual foreground removal and power spectrum estimation. In interferometric observations, during the calibration and imaging process, the diffuse emission is largely subtracted and no visible structure is left in the image cube (see e.g. Rajohnson et al. 2022). Therefore, in our work, we assume the majority of diffuse emission has been removed and model the diffuse foreground residual amplitude as 0.1 per cent of the original emission of our simulation. Although this approach will require accurate modelling of the sky signal, it is fully within the power of SKA-Low. Note that while we are only simulating 66 frequency channels from 200 to 240 MHz, a much wider frequency range, from

[1] https://github.com/ChenxiSSS/FG21SimPlus

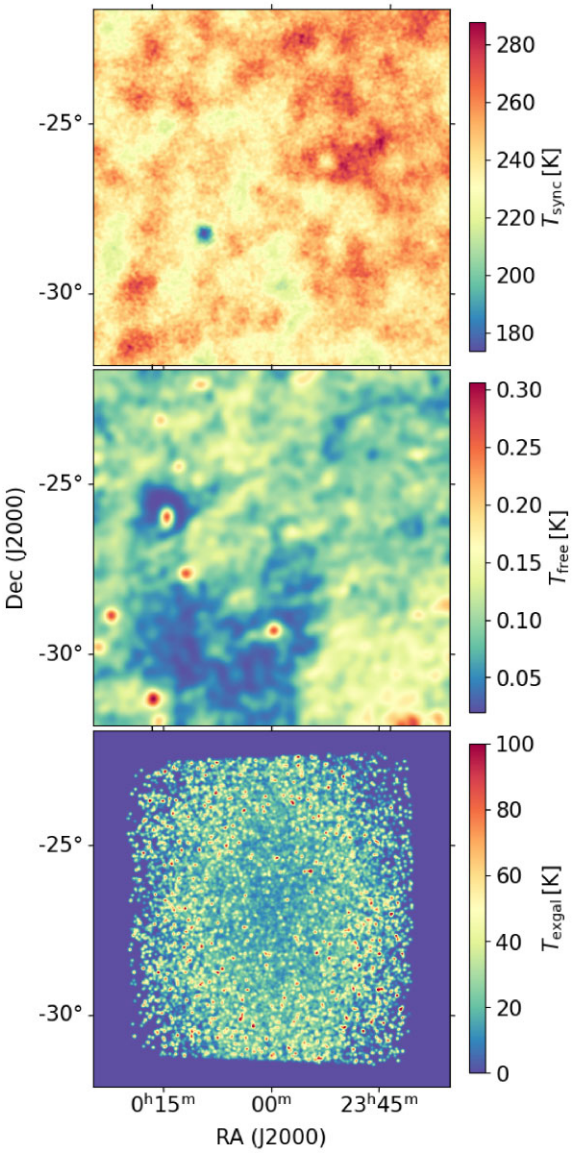


**Figure 1.** The input sky simulations of different foreground components at 220 MHz as described in Section 2. The simulation of the synchrotron radiation is shown in the top panel. The simulation of the free–free emission is shown in the middle panel. The simulation of the extragalactic radio sources are shown in the bottom panel. The pixel size of the figure is $(21\,\mathrm{arcsecond})^2$ and the total size of the signal simulation is $(10.5\,\mathrm{deg})^2$. Note that the extragalactic signal shown in the bottom panel is simply for illustration with the sources plotted as point sources. Values larger than 100 K are masked for better presentation. When simulating the observations, the radio sources are directly put in as a source catalogue instead of a map, as described in Section 2.2.

50–350 MHz, will be utilized in future SKA-Low observations to provide accurately modelling of the continuum emission. As we show later in Section 3.1, the output foreground image cube fluctuates on the scale of ∼2 mJy per point spread function (PSF), corresponding to the overall fluctuation of roughly 80 mJy, consistent with the flux density level of residual image cubes from existing EoR observations (see e.g. fig. 2 of Mertens et al. 2020). Thus, the assumption for the level of foreground residual is representative for SKA-Low. It is beyond the interest of this preliminary work to simulate the entire frequency range and produce the sky model for visibility subtraction.

Note that there are other sources of foregrounds that are of Galactic origins, such as supernovae remnants (Wang et al. 2015). Since the dominant component of the foregrounds is the synchrotron, we expect that the Galactic foreground simulated in our work is enough to capture the amplitude and the structure of the diffuse emission and leave other components of the Galactic foregrounds for future study.

### 2.2 Extragalactic radio sources

Apart from the Galactic diffuse emission, extragalactic radio sources also contribute to the overall foreground emission. While the Galactic foregrounds are mostly diffuse, the extragalactic foregrounds are typically individual sources of finite size. Understanding the properties of the radio galaxies is a major scientific goal for radio surveys. For example, both continuum and H I science results have been produced using the same fields of the MIGHTEE survey (Heywood et al. 2022; Sinigaglia et al. 2022); Observations of EoR0 field from the MWA are used to produce both the upper limits on the reionization power spectrum and the source catalogue (Beardsley et al. 2016; Trott et al. 2020; Lynch et al. 2021).

For future observations using SKA-Low, we expect a good understanding of the radio sources in the fields which will be iteratively improved as the observations themselves will further help build more complete catalogues. Here we use the source catalogue from the LOFAR Two-meter Sky Survey observations of the ELAIS-N1 (EN1) field (Sabater et al. 2021) and rotate the centre of the field to our pointing centre as shown in Fig. 1. The EN1 catalogue covers slightly less than the $(10.5\,\mathrm{deg})^2$ sky area used for simulating the diffuse foregrounds. As discussed later in Section 3, we only image the central $(1.5\,\mathrm{deg})^2$ fields so the smaller input sky area for the radio sources has negligible impacts on the intensity of the foreground emission in our image cubes. In real observations, the bright sources in the beam sidelobe pose challenges to the data calibration which we do not consider in this work. These issues can be mitigated by techniques such as secondary and direction-dependent calibrations (see e.g. Patil et al. 2017; Mertens et al. 2018; Heywood et al. 2022).

In the source catalogue, we impose a flux density cut of 10 mJy assuming all sources above this flux density can be perfectly peeled. The 10 mJy limit is fairly conservative and can be set lower given the high sensitivity of SKA-Low. For example, using 12 nights of LOFAR-EoR data observing the North Celestial Pole (NCP), Mertens et al. (2020) produced source-subtracted images with fluctuations at 50 mJy level. The source model of the NCP field has also been built iteratively over the years down to sources with flux density down to ∼3 mJy (Yatawatta et al. 2013). The depth of the sky model for the EoR0 field simulated in this work can also be expected to reach mJy level. Furthermore, we expect the sources below this flux density to be modelled with 90 per cent accuracy. This is again a conservative estimate, as relatively short observations of only 13 h used in Patil et al. (2017) reports ∼ 5 per cent error in recovering the flux density of a known bright source. As we discuss later, we focus on deep observations with ≥300 h of observation and therefore it is expected that the flux of the sources around 1 mJy can be accurately modelled with below 10 per cent errors. We assume no position errors for the sky modelling.

### 2.3 The H I signal

H I resides mostly inside the dark matter halos after the EoR at $z \lesssim 6$. The collapse of the cold gas leads to star formation, creating strong correlations between the star formation rate and the molecular ($H_2$) gas content of the galaxies (Leroy et al. 2008). Therefore, the clustering of H I can be related to the star forming properties of the galaxies and can be used to constrain the galaxy astrophysics (e.g. Wolz et al. 2016; Chen et al. 2021). At higher redshifts beyond cosmic noon $z > 2$, the fraction of H I within galaxies start to drop (Villaescusa-Navarro et al. 2018) and the distribution of the H I tilts more towards the massive halos (Spinelli et al. 2020). Due to the lack of direct observations on these H I emission sources at higher redshifts, the properties of the H I within halos are not well understood, which can be dramatically improved by future H I intensity mapping experiments.

The large sky area of $(10.5\,\mathrm{deg})^2$, and the $5 < z < 6$ redshift bin, result in a light cone of ∼1500 Mpc in the transverse direction and ∼500 Mpc in the los direction. For our purposes of exploring the detectability of the signals, instead of using a full hydrodynamical simulation, we use semi-analytical simulations based on dark matter simulations and H I Halo Occupation Distribution (HOD; Cooray & Sheth 2002). It allows us to efficiently simulate the large volume required. The detailed steps of our H I simulation are as follows:

(i) Assuming the Planck18 cosmology (Planck Collaboration 2020b), we use PINOCCHIO[2] (Monaco, Theuns & Taffoni 2002; Monaco et al. 2013) to simulate nine boxes of dark matter distributions, each with a volume of $(620\,\mathrm{Mpc})^3$. The total volume of $9 \times (620\,\mathrm{Mpc})^3$ is to ensure that the lightcone falls well within the simulated volume, avoiding edge effects. The total volume is divided into nine sub-boxes to avoid computational difficulties.

(ii) Each sub-box has 1850 grid points per side, resulting in a mass resolution of $\sim 3.25 \times 10^9\,\mathrm{M_\odot\,h^{-1}}$. Note that this mass resolution is likely not enough to resolve all the H I-rich halos (see e.g. Villaescusa-Navarro et al. 2018). However, it is enough to capture the bias of the H I clustering which is sufficient for our purposes.

(iii) Each sub-box is simulated across the $5 < z < 6$ redshift bin with a snapshot taken at each observing frequency channel, equalling a total of 66 snapshots (see Section 3 for specifications of the observations). The halo positions relative to the centre of the box in comoving space, the velocities, and the mass of the haloes are taken.

(iv) The nine sub-boxes are then put together onto $3 \times 3$ grids with the centres of the boxes re-positioned. We take the observer to be at (0,0,0) and the centre of the 5th box is at $(0,0,\mathrm{X_{cen}})$ where $\mathrm{X_{cen}}$ is the comoving distance at the centre of the $5 < z < 6$ redshift bin. The halos are re-positioned accordingly.

(v) The peculiar velocities of the halos are calculated given the 3D halo velocities and the position vectors. The halo positions are modified to redshift space according to the Kaiser effect (Kaiser 1987).

(vi) Each halo is assigned an H I mass according to the H I HOD of the IllustrisTNG simulation in Villaescusa-Navarro et al. (2018). The H I HOD follows $M_{\rm HI} = M_0(M_{\rm h}/M_{\rm min})^\alpha \exp(-(M_{\rm min}/M_{\rm h})^{0.35})$ with $M_{\rm h}$ the halo mass. We adopt the parameter values at $z = 5$, with $M_0 = 1.9 \times 10^9\,h^{-1}\,\mathrm{M_\odot}$, $M_{\rm min} = 2.0 \times 10^{10}\,h^{-1}\,\mathrm{M_\odot}$, and $\alpha = 0.74$. All H I masses are put into the halo centres, since we are only interested in large scales $k < 0.5\,\mathrm{Mpc^{-1}}$ and hence the halos are unresolved. The H I masses are then multiplied by a constant factor so that at each redshift the H I mass density, $\Omega_{\rm HI}$, equals to $10^{-3}$. This is consistent with the observation of Crighton et al. (2015) and ensures that the clustering amplitude is realistic.

(vii) The distances between the halos and the observer are calculated. For snapshot $i$ corresponding to frequency channel $i$, the los

[2]https://github.com/pigimonaco/Pinocchio

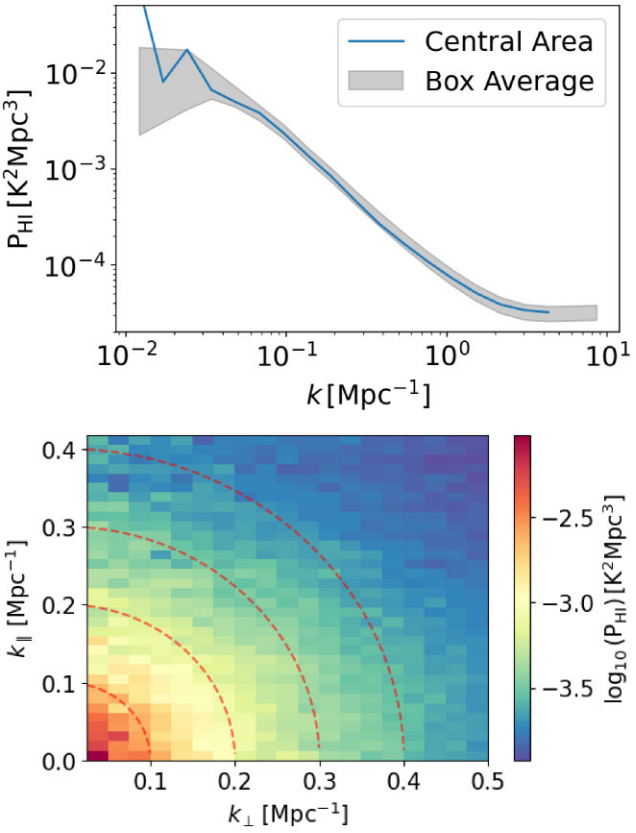


**Figure 2.** The brightness temperature power spectrum of the H I simulation described in Section 2.3. In the top panel the blue solid line shows the 1D H I power spectra for the central area of $(1.5\,\mathrm{deg})^2$ in the simulated lightcone. The shaded area shows the one standard deviation range of the input H I power spectrum where the standard deviation is calculated from all the snapshots of all the sub-boxes. The bottom panel shows the cylindrical power spectrum of the central area of $(1.5\,\mathrm{deg})^2$ in the simulated lightcone. The red dashed line denotes the $k = \{0.1, 0.2, 0.3, 0.4\}$ $\mathrm{Mpc}^{-1}$ contours for reference. The H I power spectrum of the central area agrees tightly with the H I power spectrum of the entire box, and is largely isotropic.

comoving distance range $[X^i_{min}, X^i_{max}]$ is calculated according to the channel bandwidth and central frequency. Only halos in the distance range are selected.

(viii) A rotational matrix along $y$-axis to rotate the $x$–$z$ plane is applied to the halos so that the centre of the simulation corresponds to the pointing centre RA = 0 h and Dec = −27 deg. The halo positions are converted to angular coordinates. The H I mass is converted to the flux density assuming that the flux is distributed as a step function across the frequency channel. This is a reasonable assumption given that the velocity resolution of SKA-Low at $5 < z < 6$ is not high enough to resolve the emission profiles of the H I sources.

To validate our H I simulation, we compute the H I power spectra for the nine sub-boxes, and compare to the central $(1.5\,\mathrm{deg})^2$ area of the light cone which we will use for imaging later. The resulting average H I power spectra for the boxes and for the central input image is shown in Fig. 2.

We emphasize that the variance of the H I signal, shown as the shaded area in Fig. 2, is underestimated. This is due to the fact that we assume a deterministic relation between the H I and halo mass, ignoring the scatter of the relation (see e.g. fig. 4 of Villaescusa-Navarro et al. 2018). The scatter comes from the

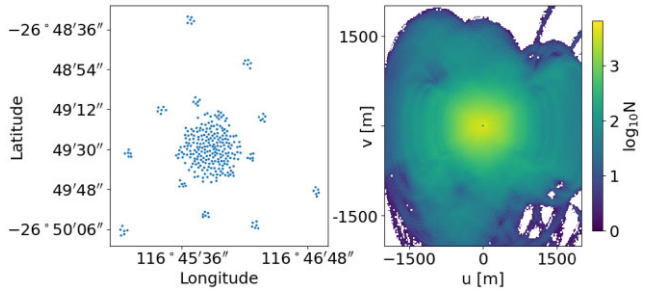


**Figure 3.** Left-hand panel: The station layout used in our simulation. Each dot denotes one station. Right-hand panel: The u–v distribution of the simulation for a 12-h tracking of the EoR0 field with a time resolution of 180 s. The colours denote the number of instantaneous baselines in one u–v grid. Each u–v grid has a size of $16 \times 16\,\mathrm{m}^2$.

assembly bias of halos, which can be introduced by the inhomogenous reionization history (e.g. Long et al. 2022). In our case of investigating the detectability in thermal noise dominated case, this effect is negligible and we leave more realistic simulations for future work.

Note that the $(1.5\,\mathrm{deg})^2$ image size corresponds to a maximum length scale equivalent to $k \sim 0.03\,\mathrm{Mpc}^{-1}$. Scales larger than this can not be probed by the image, as one can see from the top panel of Fig. 2. At smaller scales, $k > 1\,\mathrm{Mpc}^{-1}$, the H I power spectrum hits the shot-noise plateau. This is not accurate and the actual shot noise should be much lower. In our simulation, the H I is directly put as point sources in the halo centres, so that the number density of H I sources is underestimated (see Spinelli et al. 2020 for a discussion of this). The actual shot noise should be much lower and requires more in-depth modeling of the H I halo model (Wolz et al. 2019; Chen et al. 2021). As we will discuss in Section 3, the minimum $k$-scale probed in our simulation is $k \sim 0.3\,\mathrm{Mpc}^{-1}$ and therefore we are not affected by this insufficient modelling. The cylindrical power spectrum shown in the bottom panel of Fig. 2 indicates that the H I power spectrum from our simulation gives the correct isotropic features, and therefore can be reliably used to study the detectability of the H I power spectrum in the presence of the foreground wedge.

## 3 SIMULATIONS OF OBSERVATIONS

In this section, we describe the simulation of the SKA-Low interferometer to observe the input sky signal discussed in Section 2, the imaging routine to produce the image cube within the primary beam FoV, and the power spectrum estimation.

### 3.1 From sky signal to image product

The SKA-Low array will consist of 131,072 log-period dipole antennas within 512 stations covering the southern sky from 50 to 350 MHz. Since the specific station layout and specifications are not finalized, we use the v3 station layout (de Lera Acedo et al. 2020) assuming a frequency channel bandwidth of 510 kHz. We only take the central area with 296 stations with a maximum baseline length of 3.15 km. The longest baselines are not of cosmological interest and are thus neglected to reduce data volume. The frequency range we simulate is from 202.56 to 235.76 MHz, covering redshift 5–6 with 66 frequency channels. The station layout is shown at the left-hand panel of Fig 3.

The visibility data are simulated to represent one night of observation at the EoR0 field. We assume a total integration of 12 h

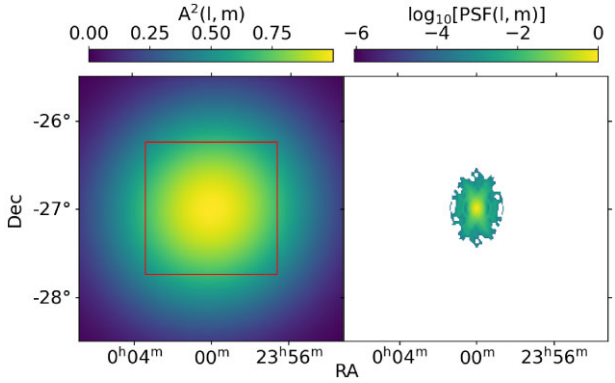


**Figure 4.** Left-hand panel: The power-square beam $A^2(l, m)$ around the pointing centre in our simulation. The primary beam is averaged across all stations. The red square shows the $(1.5\,\mathrm{deg})^2$ area within which the image cube is produced. Right-hand panel: The PSF corresponding to the u–v coverage of our simulation using natural weighting. Pixels with values ≈0 are left blank. Both figures have a size of $(3\,\mathrm{deg})^2$ with 512 × 512 pixels. Note that both the primary beam and the PSF are frequency-dependent and we show the values at central frequency 220 MHz here for presentation.

with a time-resolution of 180 s in one tracking. The resulting u–v coverage of the baselines is shown in the right-hand panel of Fig 3. The u–v coverage shown is dense within $|u| < 1000$m (which corresponds to the physical scale $k \lesssim 0.5\,\mathrm{Mpc}^{-1}$). Choosing the u–v grid length to correspond to our image size, we find no loss of u–v grid sampling, justifying the usage of a relatively coarse time resolution.

Following the observational specifications discussed above, we use the OSKAR[3] package (Mort et al. 2010) to generate the visibility data. OSKAR takes in the telescope specifications, sky model and observation strategy to simulate the primary beam, the u–v coverage and the visibility data. It can also be used to generate dirty images, which we use to produce the image cube. The sky area for the imaging output is determined by the primary beam size. In the calculation of the power spectrum, the primary beam attenuation is squared since the power spectrum is the Fourier density field squared (see e.g. Parsons et al. 2014). To image within the primary beam field-of-view, we take the limit where the power-square beam attenuation reaches ∼0.5. The primary beam is largely Gaussian near the pointing centre as shown in Fig. 4, resulting in power-square beam having half the full width at half-maximum (FWHM) comparing to the actual beam. The image size is accordingly set to be $(1.5\,\mathrm{deg})^2$ and we choose the pixel size to be $(0.45\,\mathrm{arcmin})^2$ with 200 × 200 grids. We apply the W-projection algorithm (Cornwell, Golap & Bhatnagar 2008) with natural weighting to the baselines to produce the image cube. The power-square beam and the synthesized beam (PSF) are shown in Fig 4. The PSF in Fourier space has a FWHM of $k \sim 0.3\,\mathrm{Mpc}^{-1}$.

Gaussian random noise are added to the visibility data to simulate the thermal noise. The amplitude of the thermal noise is determined by the radiometer equation (Wilson, Rohlfs & Hüttemeister 2013),

$$\sigma_\mathrm{N} = \frac{2\mathrm{k_B}T_\mathrm{sys}}{A_\mathrm{e}\sqrt{\delta f\,\delta t}}, \tag{1}$$

where $\mathrm{k_B}$ is the Boltzmann constant, $T_\mathrm{sys}$ is the system temperature, $A_\mathrm{e}$ is the effective collecting area, $\delta f$ is the frequency channel bandwidth, $\delta t$ is the time resolution. We follow Braun et al. (2019)

[3]https://github.com/OxfordSKA/OSKAR

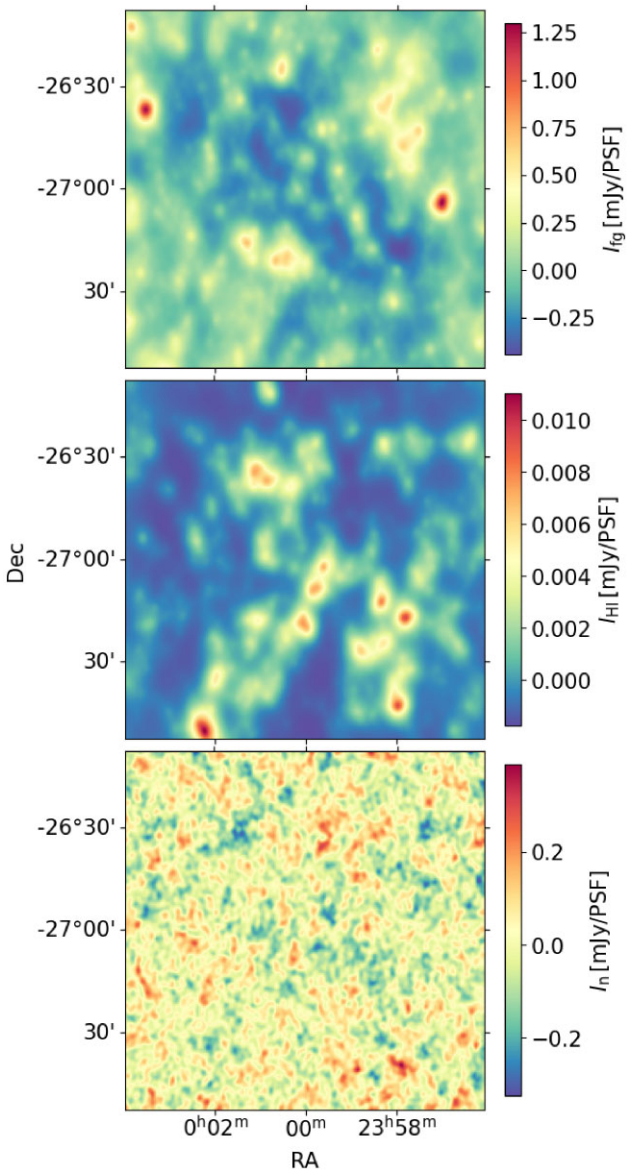


**Figure 5.** The output dirty images of the simulations at the central 220 MHz frequency channel. The top panel shows the output dirty image of the foregrounds. The central panel shows the H I image. The bottom panel shows the dirty image of the thermal noise. The images have a size of $(1.5\,\mathrm{deg})^2$ with 200 × 200 pixels. The images are shown in the units of mJy per PSF.

and set the natural sensitivity $A_\mathrm{e}/T_\mathrm{sys} = 1.235\,\mathrm{m^2\,K^{-1}}$ to generate random complex Gaussian on every baseline. The images at the central frequency for the foregrounds, the H I, and the thermal noise are shown in Fig. 5. All images are dirty images with no cleaning routine applied. Throughout this paper, we use 'Jy per PSF' and 'kelvin per PSF' units for the images before deconvolution with the PSF. The 'PSF' refers to the integrated PSF area in steradian $\int \mathrm{d}l\,\mathrm{d}m\,\mathrm{PSF}(l, m)$. 'Jy per PSF' is more commonly referred to as 'Jy per beam'. We use 'Jy per PSF' to avoid confusion with the primary beam.

In Section 5.1 when we discuss residual foreground removal, the thermal noise is rescaled by a factor of $\sqrt{t_\mathrm{sim}/t_\mathrm{int}}$, where $t_\mathrm{sim} = 12$ h is the observation time for the simulated one tracking and $t_\mathrm{int}$ is the total integration time set to 360, 480, and 600 h for different scenarios. The rescaling mimics coherent averaging of the visibility data over

multiple nights. The thermal noise power spectrum is ∼4 orders of magnitude larger than the H I power spectrum as we show in Fig. A1 in Appendix A.

### 3.2 Simulating systematics

Real observations will contain a wealth of systematics, including the radio frequency interference (RFI), gain instabilities, calibration errors, and more. While it is beyond the scope of this paper to properly take into account all of the systematics, we aim to simulate the effect of systematics that can lead to spectral instability in a simplistic way. The systematics are simulated using

$$V^{\rm i}_{\rm obs}(u^{\rm i}, v^{\rm i}, f^{\rm i}) = (1 + \delta e_f)V^{\rm i}_{\rm true} + V_{\rm TN}, \quad (2)$$

where $V^{\rm i}_{\rm true}$ is the visibility data of the i$^{\rm th}$ baseline without the systematics and the thermal noise. $V_{\rm TN}$ is the thermal noise visibility. $\delta e_f$ follows a Gaussian distribution with zero mean and only depends on the frequency channel. We simulate $\delta e_f$ with different standard deviations from $10^{-5}$ up to $10^{-4}$. The systematic errors are multiplied to the full visibility data before the assumed sky model subtraction. This effect is a crude approximation for bandpass calibration error averaged across all timesteps, creating fluctuations on small frequency scales which will leak foreground power into the observation window and bias the foreground removal techniques as we discuss in Section 5. Note that the calibration errors are complex and have smooth structures in frequency for H I observations (see e.g. figs 2 and 3 of Byrne et al. 2019). In our case, we focus on the blind removal of residual foreground after calibration and choose Gaussian errors so that the foreground scatter is present across the delay space (see Appendix B).

It is worth pointing out that the 200–240 MHz frequency range hosts several prominent sources of RFI. Around 220 MHz there are the RF11 and RF12 bands of digital TV (see e.g. fig. 2 of Offringa et al. 2015), which can be identified through flagging algorithms (e.g. Offringa et al. 2010; Wilensky et al. 2019). The larger end of the frequency range ∼240 MHz sits right next to military satellite band (242–272 MHz) which may cause complete data loss of the entire frequency range (see fig. 4 of Sokolowski, Wayth & Lewis 2015). The presence of this RFI forbids us to go below redshifts $z < 5$. Overall we expect that the 200–240 MHz frequency range can be observed without substantial loss of data.

### 3.3 H I power spectrum from the imaging route

The image cube can be used to estimate the H I power spectrum. We compute the H I power spectrum from the imaged data instead of measuring the delay power spectrum directly from the visibilities for two reasons. First, the cosmological quantities such as the Hubble parameter and the comoving distance have significant evolution across the large redshift bin $\Delta z = 1$, making the delay power spectrum estimation very difficult especially with regards to deconvolving w-projection kernel and primary beam attenuation. Secondly, if we can verify the detectability of one field in image space, we can probe larger cosmological scales through image mosaicing of overlapping fields.

To calculate the H I power spectrum, we first transform the flux density $I(l, m, f)$ in the image cube into Fourier space brightness temperature

$$\tilde{T}(\boldsymbol{k}_\perp, k_\parallel) = \int \frac{{\rm d}^3x}{\mathbf{V}} \exp\big[-i\boldsymbol{k}\cdot\boldsymbol{x}\big] \left(\frac{\lambda^2}{2k_{\rm B}}\right)^2 \frac{I(\boldsymbol{x})}{A(\boldsymbol{x})}, \quad (3)$$

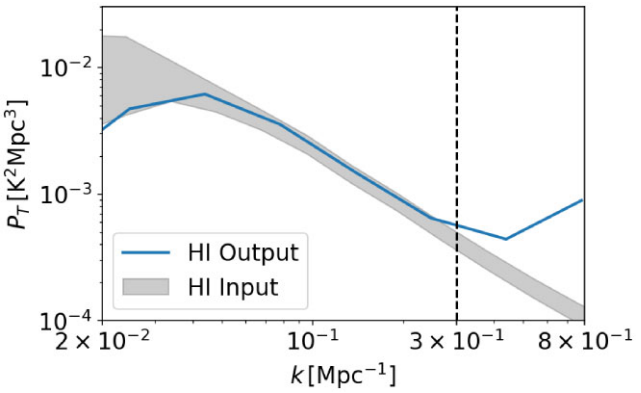


**Figure 6.** The H I power spectrum estimated from the image cube using H I-only visibility data ('H I output'), compared against the input H I power spectrum described in Section 2.3 ('H I input'). The vertical black dashed line corresponds to $k = 0.3\,{\rm Mpc}^{-1}$ where the effects of the PSF starts to dominate.

where $\boldsymbol{x} = [l \cdot D_c(z_{\rm c}), m \cdot D_c(z_{\rm c}), D_c(z_f)]$ is the physical coordinate corresponding to the sky coordinate $(l, m)$ and observing frequency $f$. $\mathbf{V}$ is the comoving volume of the image cube. $D_c(z)$ is the comoving distance at redshift $z$. $z_{\rm c}$ is the centre of the redshift bin and $z_f = f_{21}/f - 1$ is the redshift corresponding to the frequency $f$ where $f_{21}$ is the rest frequency of the 21-cm line. $A(\boldsymbol{x})$ is the primary beam attenuation. $\lambda$ is the observing wavelength. The transverse coordinates for each voxel are assigned assuming an effective comoving distance, which is important to ensure that the operators for residual foreground removal and Fourier transformation are commutable as we discuss in Appendix A. $\tilde{T}(\boldsymbol{k}_\perp, k_\parallel)$ is in the units of kelvin per PSF. The H I power spectrum in 3D $k$-space is

$$P_{\rm HI}(\boldsymbol{k}_\perp, k_\parallel) = \frac{|\tilde{T}(\boldsymbol{k}_\perp, k_\parallel)|^2}{|\widetilde{\rm PSF}(\boldsymbol{k}_\perp, f_{\rm c})|^2}, \quad (4)$$

where $\widetilde{\rm PSF}(\boldsymbol{k}_\perp, f_{\rm c})$ is the 2D Fourier transform of the PSF at the central frequency $f_{\rm c}$

$$\widetilde{\rm PSF}(\boldsymbol{k}_\perp, f_{\rm c}) = \int {\rm d}l\,{\rm d}m \exp\big[-2\pi i(lu + mv)\big] {\rm PSF}(l, m, f_{\rm c}). \quad (5)$$

In the calculations above, several approximations have been made. The frequency evolution of the PSF is assumed to be negligible over the frequency bandwidth of the simulated observation. The physical coordinates of the voxels are assigned assuming an effective comoving distance. The flat-sky approximation is also used. While these assumptions may not be accurate enough for precision cosmology, as we show in Fig. 6, it can be seen that the output H I power spectrum is within the $1\sigma$ region of the input. It is sufficiently accurate for studying the detectability of the signal. The scales probed are from $k \sim 0.03\,{\rm Mpc}^{-1}$, limited by the size of the image, to $k \sim 0.3\,{\rm Mpc}^{-1}$, limited by the image resolution due to the PSF. In the power spectrum results shown hereafter, a Blackman–Harris frequency taper is also applied to minimize potential leakage of foregrounds and systematics, with the details discussed in Appendix A.

## 4 QUANTIFYING THE FOREGROUND WEDGE

In this section, we use the image cube from H I and foreground visibility data without the thermal noise to explore the limits of reducing foreground contamination. Without the thermal noise and any systematics, the H I and foreground-only case showcases the best possible scenario for residual foreground removal. It helps us

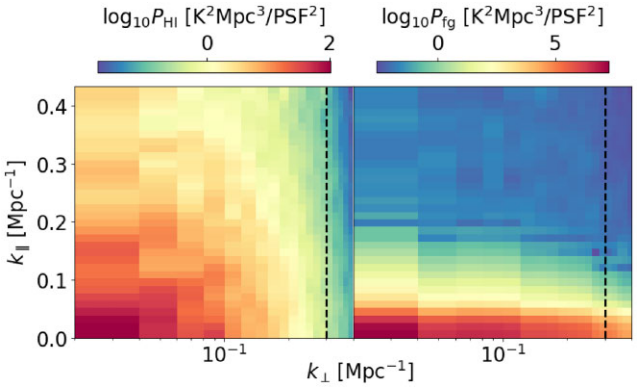


**Figure 7.** The cylindrical power spectra for the H I (left) and the foregrounds (right) estimated from the output image cubes. Note that the PSF is not deconvolved from the power spectra and the power spectra are in the units of $K^2Mpc^3/PSF^2$.

understand the requirements for sky modelling to enable detection and locate the observation window in the $k_\perp$–$k_\parallel$ plane. We particularly focus on scales of cosmological interest $k < 0.2\,\mathrm{Mpc}^{-1}$, especially the largest scale that can be probed using our image cube $k \sim 0.03\,\mathrm{Mpc}^{-1}$. If these scales can be probed with little foreground contamination, future surveys using wide-field imaging and mosaicing can further extend the scales larger than the first baryon acoustic oscillation (BAO; Eisenstein & Hu 1998) peak at $k \sim 0.04\,\mathrm{Mpc}^{-1}$ to the linear scales for cosmological analysis.

### 4.1 Observation window using only foreground avoidance

We first use the H I-only image cube and foreground-only image cube to estimate the power spectra for the H I and the foregrounds to compare them in cylindrical $k$-space. The cylindrical power spectra for the H I and the foregrounds are shown in Fig. 7. Comparing the ratio between the H I power spectrum and the foreground power spectrum as shown in the top left-hand panel of Fig. 8, the foreground power spectrum is larger than the H I power spectrum at $k_\parallel \lesssim 0.12\,\mathrm{Mpc}^{-1}$, leaving no observation window at linear and BAO scales. In Fig. 8, the region where foreground power dominates does not have a clear wedge structure. This is due to the fact that the bright sources in the primary beam side-lobes, which contributes mostly to the wedge structure at high $k_\perp$, are assumed to be already removed in our simulation. Without the strong foreground emissions coming from large angular extent (high delay time), the wedge structure at high $k_\perp$ no longer exists. The lack of wedge feature can also be seen from observations (e.g. LOFAR observaions shown in Mertens et al. 2020; Hothi et al. 2021). As we demonstrate in Section 4.2, the wedge structure reappears after foreground cleaning is applied to the data. This is due to the fact that removing residual foregrounds reduces the foreground power near the pointing centre, making the foreground emission at larger angular distance comparatively brighter.

If we relax the 10 per cent modelling residual as described in Section 2.2 to an extreme 0.1 per cent, $k_\parallel \lesssim 0.05\,\mathrm{Mpc}^{-1}$ scales are still lost as shown in the top right-hand panel of Fig. 8, which invalidates the usage of the observations for cosmology. The result suggests that even with extreme level of calibration and sky modelling accuracy, it is unlikely that foreground avoidance can be used to measure the H I power spectrum at cosmological scales at $5 < z < 6$ due to the weakness of the H I signal at these redshifts. We can use foreground removal methods to mitigate the contamination at large scales as we show in the following sections.

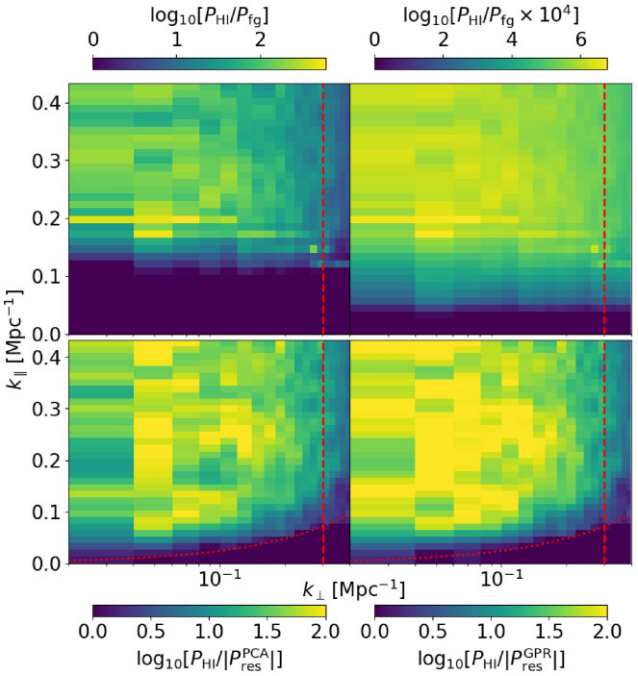


**Figure 8.** Top left-hand panel: The ratio between the H I power spectrum and the foreground power spectrum in cylindrical $k_\perp$–$k_\parallel$ space. The vertical red dashed line denotes the $k = 0.3\,\mathrm{Mpc}^{-1}$ line where the effects of the PSF start to dominate. Top right-hand panel: The ratio between the H I power spectrum and the foreground power spectrum, with the foreground power suppressed by a factor of $10^4$. Bottom left-hand panel: The ratio between the H I power spectrum and the residual power spectrum after PCA cleaning. The red dotted line denotes the 'horizon limit' $k_\parallel = 0.24k_\perp$ calculated according to equation (11). Bottom right-hand panel: The ratio between the H I power spectrum and the residual power spectrum after GPR cleaning. In all the panels shown, values below 1 are set to 1 for better presentation. The darkest end of the colour scale corresponding to the region where the foreground power is larger than the H I.

### 4.2 Residual foreground removal

In order to suppress foreground contamination down to the wedge and create an observation window at large scales, we explore methods of blind source subtraction to remove the residual foregrounds. We focus on two methods commonly used, namely the PCA (e.g. Spinelli et al. 2022) and GPR (e.g. Mertens et al. 2018; Soares et al. 2022). Following Chen, Wolz & Battye (2023), with the observation window enlarged due to the foreground cleaning we can choose a criteria for the power spectrum estimation in 1D $k$-space

$$k_\parallel > c_k k_\perp, \tag{6}$$

where $c_k$ is a constant to be set. The value for $c_k$ can be found by iteratively testing with larger values to the point where the 1D power spectrum results converge.

We write out the general formalism for frequency–frequency covariance based foreground removal methods

$$\hat{\mathbf{X}}_{\mathrm{fg}} = \hat{\mathbf{C}}_{\mathrm{fg}}\hat{\mathbf{C}}^{-1}\mathbf{X}, \tag{7}$$

where $\mathbf{X}$ is the mean-centred image cube which has dimensions of $(N_f, N_p)$ with $N_f$ the number of frequency channels and $N_p$ the number of pixels in one frequency channel. $\hat{\mathbf{C}}_{\mathrm{fg}}$ is an estimation of the covariance matrix for the foregrounds and $\hat{\mathbf{C}}^{-1}$ is the inverse of the estimation of the total data covariance. For different methods such as PCA and GPR, different choices of $\hat{\mathbf{C}}_{\mathrm{fg}}$ and $\hat{\mathbf{C}}$ are used,

producing different reconstructed foregrounds which we discuss in detail in Section 5.1.

#### 4.2.1 Foreground Removal using PCA

The PCA method separates the foregrounds by using the eigenvalue decomposition of the frequency–frequency data covariance matrix (e.g. Cunnington et al. 2021)

$$\hat{\mathbf{C}}_{\rm d} = \mathbf{X}\mathbf{X}^{\rm T}/(\mathrm{N_p} - 1), \tag{8}$$

The eigenvalues and eigenvectors of the covariance matrix are then calculated. An estimation of the foregrounds can be extracted from the data matrix using

$$\hat{\mathbf{X}}_{\rm fg}^{\rm PCA} = \mathbf{A}\mathbf{A}^{\rm T}\mathbf{X},\ \mathbf{A} = [\mathbf{v}_1, ..., \mathbf{v}_{\mathrm{N_{fg}}}]. \tag{9}$$

Here, $\mathbf{v}_{\rm i}$ is the eigenvector corresponding to the $\mathrm{i^{th}}$ largest eigenvalue and a total of $N_{\rm fg}$ modes are removed. To link it to equation (7), we can rewrite equation (9) as

$$\hat{\mathbf{X}}_{\rm fg}^{\rm PCA} = \left(\mathbf{A}\mathbf{A}^{\rm T}\hat{\mathbf{C}}_{\rm d}\right)\left(\hat{\mathbf{C}}_{\rm d}\right)^{-1}\mathbf{X}, \tag{10}$$

where it is straightforward to see that, in the case of PCA, $\hat{\mathbf{C}}_{\rm PCA} = \hat{\mathbf{C}}_{\rm d}$ and $\hat{\mathbf{C}}_{\rm fg}^{\rm PCA} = \mathbf{A}\mathbf{A}^{\rm T}\hat{\mathbf{C}}_{\rm d}$.

In our case, the eigenvalues of the data covariance reach a plateau after the third eigenvalue, suggesting that $\mathrm{N_{fg}} = 3$ is a good choice for cleaning the foregrounds and avoiding overcleaning the signal. The ratio between the H I power spectrum and the residual power spectrum after cleaning is shown in the bottom left-hand panel of Fig. 8. Throughout the paper, the residual power spectrum is defined as the power spectrum of the residual foreground image $\mathbf{X}_{\rm res} = \mathbf{X}_{\rm fg} - \hat{\mathbf{X}}_{\rm fg}$, where $\mathbf{X}_{\rm fg}$ is the image of the input foregrounds and $\hat{\mathbf{X}}_{\rm fg}$ is the removed foreground by either PCA or GPR.

Comparing Figs 7 and 8, the cleaning efficiently enlarges the observation window at small $k_\perp$. If the foreground contamination is optimally mitigated, the foreground wedge can be located using the 'horizon limit' (Liu et al. 2014)

$$c_k^{\rm h} = \frac{H(z)D_c(z)\theta_0}{c(1+z)}, \tag{11}$$

where $H(z)$ is the Hubble parameter, $c$ is the speed of light and $\theta_0$ is the angular extent of the instrument beam. As a crude approximation we choose $\theta_0 = 2\sqrt{\Omega_{\rm beam}}/\pi$ where $\Omega_{\rm beam}$ is the integrated primary beam which gives $c_k^{\rm h} = 0.24$. From the bottom left-hand panel of Fig. 8, we can see that the foreground wedge is close to the horizon limit which is marked by the red dotted line, showing that the foreground cleaning is efficient. Iteratively increasing the threshold we find that the 1D power spectrum converges at $c_k = 0.3$, which we use from now on in this paper.

#### 4.2.2 Foreground removal using GPR

GPR constructs the foreground component by fitting parameterized kernels to the data covariance. Suppose we have the H I kernel $\mathbf{K}_{\rm HI}$, the foreground kernel $\mathbf{K}_{\rm fg}$ and the thermal noise kernels $\mathbf{K}_{\rm n}$ fitted, then the estimated foreground can be written as (e.g. Mertens et al. 2018)

$$\hat{\mathbf{X}}_{\rm fg}^{\rm GPR} = \mathbf{K}_{\rm fg}\left(\mathbf{K}_{\rm fg} + \mathbf{K}_{\rm n} + \mathbf{K}_{\rm HI}\right)^{-1}\mathbf{X}. \tag{12}$$

It is straightforward to see that, in the case of GPR, $\hat{\mathbf{C}}_{\rm GPR} = \mathbf{K}_{\rm fg} + \mathbf{K}_{\rm n} + \mathbf{K}_{\rm HI}$ and $\hat{\mathbf{C}}_{\rm fg}^{\rm GPR} = \mathbf{K}_{\rm fg}$.

The H I and the foreground covariance matrices are shown in Fig. 9 for reference. The H I covariance is highly diagonal, due to the discrete and uncorrelated nature of the H I along the los. On the other hand, the foreground covariance is smooth and shows a clear spectral feature along the frequency direction, corresponding to the negative spectral indices of the radio sources. Due to the spectral evolution of the foreground covariance, the conventional choice of a Matérn kernel (Matérn 1966) does not describe the foreground covariance well. Instead, we use Markov chain Monte Carlo (MCMC) to fit the kernels using the following steps:

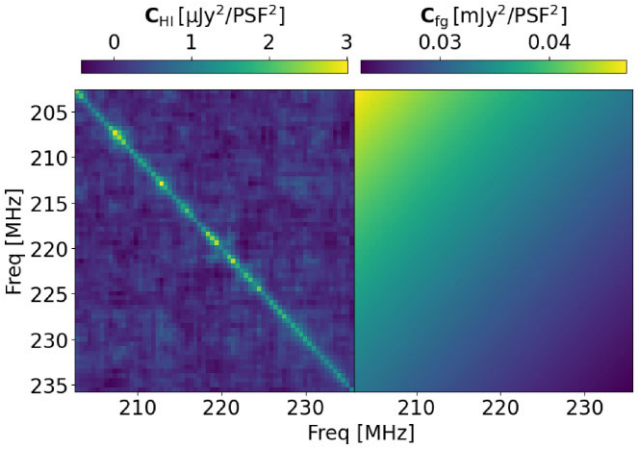


**Figure 9.** The frequency–frequency covariance matrices for the H I ($\mathbf{C}_{\rm HI}$) on the left and the foregrounds ($\mathbf{C}_{\rm fg}$) on the right. The covariance matrices are calculated from H I-only and foreground-only image cubes following equation (8).

(i) In each step, a random value $\sigma_{\rm n}$ is sampled and a diagonal kernel $\mathbf{K}_{\rm n} = \sigma_{\rm n}^2\delta_{\rm ij}^{\rm K}$ is calculated where $\delta^{\rm K}$ is the Kronecker delta. In this section, $\mathbf{K}_{\rm n}$ is the H I kernel. Following Soares et al. (2022), in Section 5 when thermal noise is included, $\mathbf{K}_{\rm n}$ is the sum of the H I and the thermal noise covariance matrices.

(ii) The total data covariance matrix is then subtracted by the diagonal kernel $\mathbf{K}_{\rm n}$. A third-order polynomial fitting is then performed on every row of the subtracted result, creating a fitted kernel $\mathbf{K}_{\rm fit}$. The kernel is then symmetrized to get the foreground kernel $\mathbf{K}_{\rm fg} = (\mathbf{K}_{\rm fit} + \mathbf{K}_{\rm fit}^{\rm T})/2$.

(iii) The parameters for the kernels are then fitted by maximizing the log-marginal likelihood $\log p = -(\mathbf{X}^{\rm T}\mathbf{K}^{-1}\mathbf{X} + \log|\mathbf{K}| + n\log 2\pi)/2$, where $n$ is the number of data points sampled and $\mathbf{K}$ is the sum of the kernels $\mathbf{K} = \mathbf{K}_{\rm fg} + \mathbf{K}_{\rm n}$.

(iv) The MCMC fitting is then performed with 20 random walkers with 2000 iterations to make sure the chains converge. The initial guess of $\sigma_{\rm n}$ is taken to be the square root of the trace of the data covariance. The final kernels are the 50 per cent percentile of the $\mathbf{K}_{\rm n}$ and the $\mathbf{K}_{\rm fg}$ samples in the chains excluding the first 100 steps.

Note that after GPR cleaning, a bias correction can be applied as shown in Mertens et al. (2018). We follow the quadratic estimator formalism of Kern & Liu (2021) and show in Appendix A that the bias correction term in our case is negligible. The resulting foreground residual power spectrum compared to the H I power spectrum is shown in the bottom right-hand panel of Fig. 8. Comparing the foreground wedge in the GPR case with the horizon limit and with the PCA case, we can see that in the absence of thermal noise, GPR is slightly more efficient in cleaning the foregrounds and both methods do well enough to enable the detection of the H I at large scales $k < 0.1\,\mathrm{Mpc}^{-1}$. At the largest spatial scales of the image, there is negative residual power from overcleaning. The differences between these two methods are discussed later in Section 5.1.

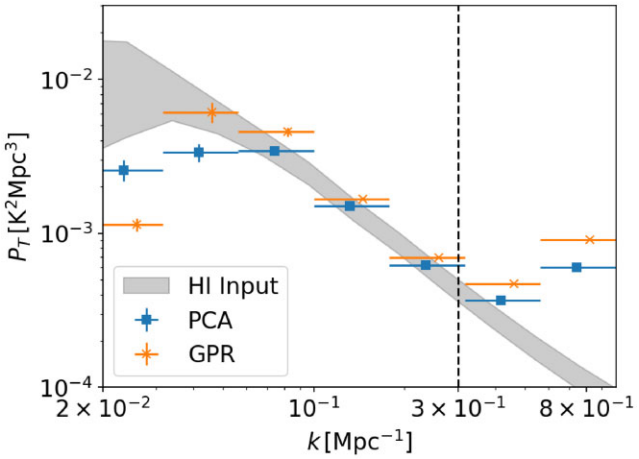


**Figure 10.** The 1D H I power spectrum results from image cubes of H I and foreground using foreground removal methods measured in the $k_\parallel > 0.3k_\perp$ regions. The blue data points shows the results from PCA and the yellow points show the results from GPR. The vertical dashed line denotes $k = 0.3\,\mathrm{Mpc}^{-1}$ where the effects of PSF start to dominate. The centres of the $k$-bins are misplaced by 5 per cent for presentation.

The success of the methods in cleaning the foregrounds indicates that we can measure the H I power spectrum from the SKA-Low observation at $5 < z < 6$, as we show using the 1D power spectrum in Fig. 10. As mentioned, both methods can enable the measurements of the H I power spectrum from $k \sim 0.05\,\mathrm{Mpc}^{-1}$ up to $k \sim 0.3\,\mathrm{Mpc}^{-1}$.

## 5 FORECASTS FOR SKA-LOW

In this section, we further explore the detectability of the H I power spectrum for SKA-Low observations by including different levels of thermal noise in the simulation. In particular, to enable the measurement of the H I power spectrum, the robustness of the foreground removal methods in the presence of the thermal noise must be tested. Furthermore, we simulate systematics by generating stochastic errors along the frequency direction to test the limits of level of systematics allowed.

### 5.1 Robust foreground cleaning with low SNR

In Section 4.2, we show that the foreground removal methods can suppress the foreground wedge to the horizon limit. However, this result is based on the fact that the empirical data covariance is 'clean', i.e. the covariance is purely a combination of the H I and the foregrounds. Therefore, the distinctive features of the H I can be extracted from the signal using PCA and GPR. In reality, the data covariance is likely to contain a high level of thermal noise as well as systematics, making it difficult to construct the covariance of the foregrounds. We test PCA and GPR in the presence of different levels of thermal noise. As described in equation (1) in Section 3.1, we simulate the thermal noise for the 12h tracking and rescale it to match 360, 480, and 600 h of integration time.

We first show the results for the 360 h case and compare the effects of foreground removal methods. The PCA and GPR routines are kept the same as in Section 4.2 with the observation window $c_k = 0.3$. The ratio between the underlying H I power spectrum and the foreground residual after removal in cylindrical $k$-space is shown in Fig. 11. In contrast with the results shown in Fig. 8, the amplitude of the residual power increases significantly. For the

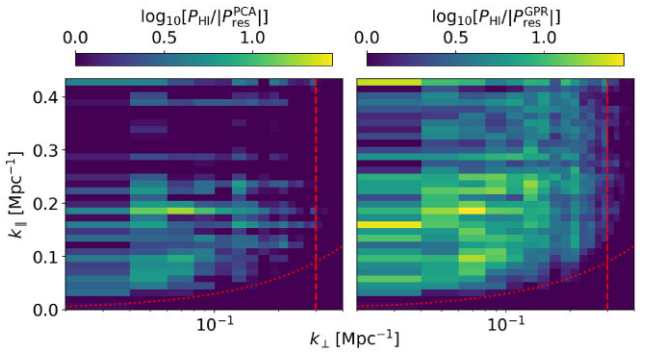


**Figure 11.** Left-hand panel: The ratio between the H I power spectrum and the residual foreground power spectrum in cylindrical $k_\perp$–$k_\parallel$ space using the PCA cleaning. Right-hand panel: The same with the left-hand panel except the residual is obtained using the GPR cleaning. All panels shown have values below 1 set to 1 to separate the observation window from the foreground wedge. The vertical red dashed line denotes the $k = 0.3\,\mathrm{Mpc}^{-1}$ line where the effects of the PSF start to dominate. The red dotted line denotes the boundary for the observation window $k_\parallel = 0.3k_\perp$.

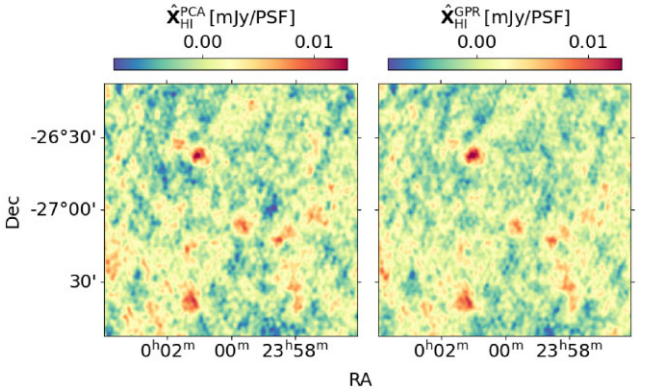


**Figure 12.** Left-hand panel: The estimated H I image $\hat{\mathbf{X}}_{\rm HI}$ defined in equation (13) using the PCA cleaning at the central 220 MHz frequency channel. Right-hand panel: The same with the left-hand panel except the residual is obtained using the GPR cleaning. The colour scales of the images are set to range from −0.11 to 0.13 mJy per PSF for fair comparisons. The residual images obtained from PCA and GPR are similar with each other, yet the level of foreground leakage differs significantly as shown in Fig. 11.

PCA case, the observation window is heavily contaminated by the foregrounds while the contamination is less severe in GPR. Note that this difference is not visible in the residual image cube as we show in Fig 12. The amplitude of the fluctuation of the residual is roughly the same with no indications of the different levels of foreground contamination.

The difference between PCA and GPR can be seen using the formalism in Section 4.2. Comparing equations (10) and (12), we can see that GPR uses the fitting result to obtain smooth kernels of the H I and the foregrounds for cleaning. On the other hand, PCA directly operates on the total data covariance, which contains a fluctuation around zero in the non-diagonal elements because of the thermal noise. The fluctuation of the thermal noise leads to small-scale oscillations in the residual covariance. For comparison, we calculate the covariance of the 'estimated' H I, i.e. the total image subtracted by the removed foreground and the noise component

$$\hat{\mathbf{X}}_{\rm HI} = \mathbf{X}_{\rm d} - \mathbf{X}_{\rm n} - \hat{\mathbf{X}}_{\rm fg}. \quad (13)$$

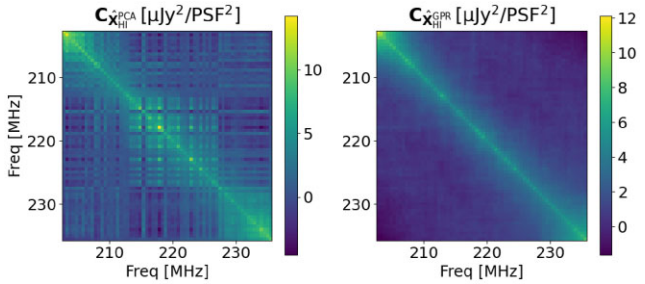


**Figure 13.** Left-hand panel: The frequency–frequency covariance of the 'estimated' H I image $\hat{\mathbf{X}}_{\rm HI}$ obtained using the PCA cleaning. Right-hand panel: The same with the left-hand panel except the residual is obtained using the GPR cleaning.

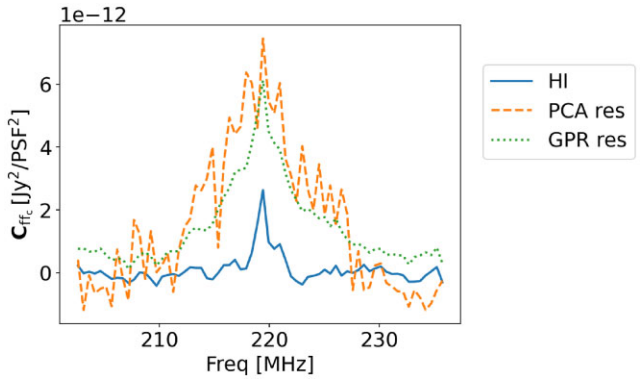


**Figure 14.** The frequency–frequency covariance of the residual foreground image $\hat{\mathbf{X}}_{\rm res}$ obtained using the PCA cleaning ('PCA res') and the GPR cleaning ('GPR res') for the central row $\mathbf{C}_{\rm ff_c}$ with $\rm f_c = 220$ MHz. The H I covariance is shown in blue solid line ('H I') for reference.

Comparing the covariance of $\hat{\mathbf{X}}_{\rm HI}^{\rm PCA}$ and $\hat{\mathbf{X}}_{\rm HI}^{\rm GPR}$ as shown in Fig. 13, we can see that while the amplitude of the covariance is roughly the same and close to the true H I shown in the left-hand panel of Fig. 9, the PCA case has large fluctuations across the frequency channel, leading to the stripe-like features in the frequency–frequency covariance matrix. While this fluctuation is also present in GPR, its amplitude is much smaller and the dominating component is still the diagonal H I covariance. For PCA, however, this fluctuation introduces a small-scale fluctuation that spills foreground power into the observation window, resulting in severe signal loss at all scales including scales where the foreground power is originally already lower than the H I as shown in the upper left-hand panel of Fig. 8. To further illustrate the small-scale contamination, we calculate the covariance matrices for the residual foreground $\hat{\mathbf{X}}_{\rm res}$ for PCA and GPR and compare them with the H I covariance as shown in Fig. 14. Both methods have clear foreground residual structure over large frequency scales. However, the PCA residual has a much larger small-scale fluctuation with the amplitude larger than the diagonal H I. The small-scale fluctuation results in severe contamination in high $k_\parallel$ modes inside the observation window as shown in Fig. 11.

When comparing PCA and GPR, we assume full knowledge of the true H I, thermal noise, and foregrounds in our simulation to perform quality checks on the foreground removal methods. It is important to note that the foreground removal and power spectrum estimation routines do not rely on knowing the underlying components. The foreground removal is performed blindly and the H I power spectrum is estimated by subtracting a thermal noise covariance as discussed in

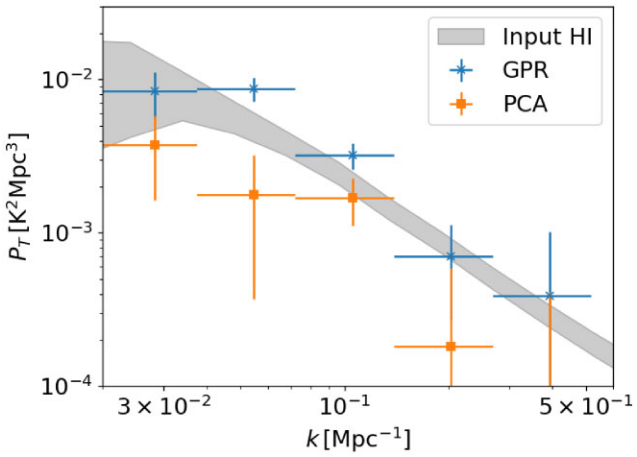


**Figure 15.** The 1D H I power spectrum measurements with 360 h of integration time after residual foreground cleaning. The error bars on the horizontal axis denote the width of the $k$-bins and the error bars on the vertical axis denote the errors of the bandpower estimation. The results for GPR are shown in the shape of cross ('GPR') and the results for PCA are shown in the shape of squares ('PCA'). The shaded region denotes the input H I power spectrum ('Input H I').

Appendix A. We choose logarithmically distributed k-bins from 0.01 to 1 $\rm Mpc^{-1}$ with $\Delta[\log(k/{\rm Mpc}^{-1})] = 0.25$ and show the resulting 1D power spectrum for 360 h of integration time for both PCA and GPR in Fig. 15. Throughout this paper, the error bars on the 1D power spectrum are calculated by calculating the sampling variance of the 3D powers that fall into the 1D $k$-bins. The resulting measurement errors on the power spectrum are

$$\Delta P(k_i) = \frac{\rm std[P(k \in k_i)]}{\sqrt{N^i_{\rm modes}}}, \tag{14}$$

where std$[P(k \in k_i)]$ denotes the standard deviation among the 3D powers that belongs in the $i$th $k$-bin and $N^i_{\rm modes}$ denotes the number of $k$-points in the $i$th $k$-bin.

As shown in Fig. 15, the foreground contamination leads to overestimation for the GPR case from $k \sim 0.03$–$0.3\,\rm Mpc^{-1}$. The severe contamination of foregrounds results in signal loss on most scales for the PCA case. In conclusion, we find that in the presence of high thermal noise, PCA induces foreground contamination into the observation window due to the small-scale fluctuation in the data covariance matrix. On the other hand, GPR does not introduce sizable foreground leakage into the observation window and mitigates the foregrounds to enable the measurements of the H I power spectrum at large scales.

Using GPR, we present our forecasts for the H I power spectrum measurement for SKA-Low observations of the EoR0 field assuming 360, 480, and 600 h of integration time in Fig. 16. The power spectrum results converge to the input H I as the noise level decreases. For 360 h of integration time, all bandpower measurements are within the $1\sigma$ error of the input H I with the bandpower at $k \sim 0.05\,\rm Mpc^{-1}$ slightly overestimated due to foreground contamination. While not shown, we also tested that decreasing the integration time to 250 h results in the bias exceeding the $1\sigma$ error. We conclude that the integration time of one field needs to be greater than 250 h to enable unbiased measurements of the H I power spectrum. In the case of 480 h, the H I power spectrum can be measured with ∼3 signal-to-noise ratio (SNR) from $k \sim 0.03$ to $0.3\,\rm Mpc^{-1}$. Further increasing the integration time to 600 h, we find that the bias further decreases

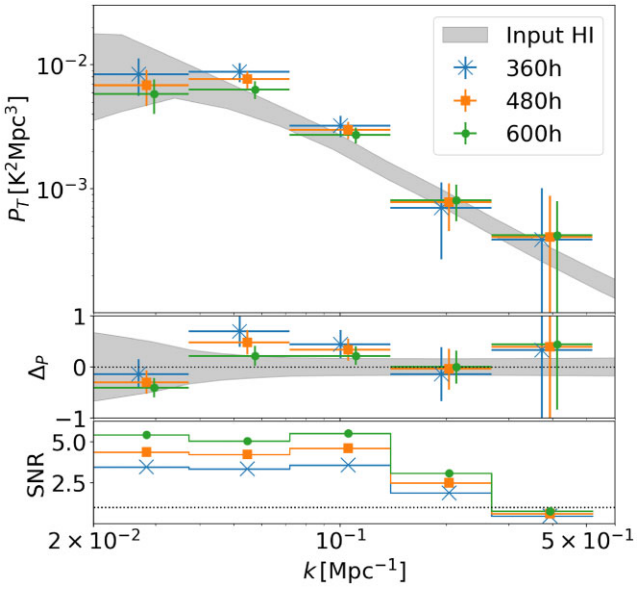


**Figure 16.** Top panel: The 1D H I power spectrum measurements with 360, 480, and 600 h of integration time after GPR foreground cleaning. The error bars on the horizontal axis denote the width of the $k$-bins and the error bars on the vertical axis denote the errors of the bandpower estimation. The shaded region denotes the input H I power spectrum ('Input H I'). The centres of the k-bins for the 360 and 600 h cases are misplaced by 5 per cent in $k$-direction for better presentation. Central panel: The fractional difference between the estimated H I power spectrum and the underlying H I input $\Delta_P = (\hat{P}_{\rm HI} - P_{\rm HI})/P_{\rm HI}$. The black dotted line denotes $\Delta_P = 0$. Bottom panel: The signal-to-noise ratio (SNR) of the measurements. The black dotted line denotes SNR = 1.

and the error bar scales as $\sqrt{t_{\rm int}}$, suggesting that the thermal noise is the dominant source of the measurement errors.

### 5.2 Impact of systematics on foreground removal

Interferometric observations contain various systematics, such as RFI, gain fluctuations, calibration errors, etc. These systematics impact the H I power spectrum measurement in various ways. For example, the data loss coming from RFI requires inpainting or novel Fourier transform methods which leaves residuals in the power spectrum (Trott et al. 2016; Pagano et al. 2022). Imperfect calibrations leaks the foreground power into the observation window (Barry et al. 2016). Gain and phase errors contribute to the foreground contamination in the H I power spectrum (Mazumder et al. 2022). As a proof of concept, we are aiming to give a qualitative assessment of the impact of the systematics. Following equation (2), we set std($\delta e_f$) to $10^{-5}$, $5 \times 10^{-5}$, and $10^{-4}$ to check the resulting power spectrum estimation. All foreground removal and power spectrum estimation steps are kept the same as Section 5.1 and we choose the integration time to be 600 h to isolate the impact of the systematics. Previous literature suggest that $<10^{-5}$ level of systematic error is needed for the measurement (Barry et al. 2016; Mazumder et al. 2022). Note that, however, as we simulate the systematics as a random error on the true signal, it acts as a small frequency-scale fluctuation on the residual foregrounds which can be partially mitigated. Therefore, using methods such as GPR, we can still recover the observation window with the level of systematics higher than $10^{-5}$.

In Fig. 17, we show the H I power spectrum measured from the $k_\parallel > 0.3k_\perp$ window with the signal perturbed by the $10^{-5}$, $5 \times 10^{-5}$, and $10^{-4}$ systematic error. For a very small level of $10^{-5}$ systematic effects, the GPR foreground removal method is unaffected by the systematics and removes the residual foreground sufficiently. As we increase the level of systematics to $5 \times 10^{-5}$, the foreground starts to leak into the observation window especially at small scales, leading to overestimation of the H I power spectrum. Increasing the systematics to $10^{-4}$ the contamination becomes severe and leads to biased estimation of the H I power spectrum at all scales.

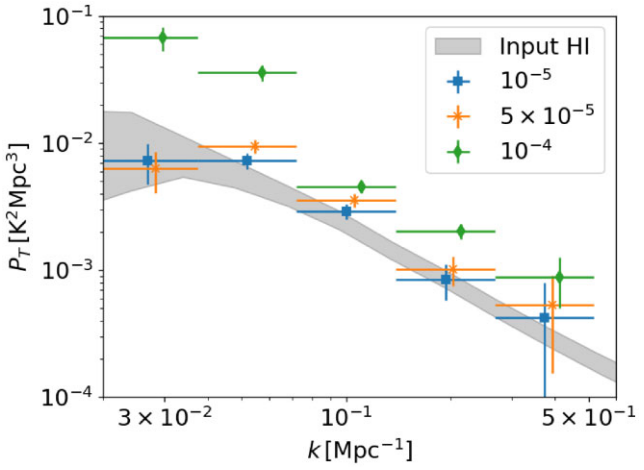


**Figure 17.** The H I power spectrum measured from the residual foreground removed image cube as described in Section 5.2 for different levels of systematics ($10^{-5}$, $5 \times 10^{-5}$, and $10^{-4}$). The error bars on the horizontal axis denote the width of the $k$-bins and the error bars on the vertical axis denote the errors of the bandpower estimation. The shaded region denotes the input H I power spectrum ('Input H I'). The centres of the k-bins for the $10^{-5}$ and $10^{-4}$ systematic effect cases are misplaced by 5 per cent in $k$-direction for better presentation.

Similar to Section 5.1, we can use the covariance matrices to show that the systematic effects break the frequency smoothness of the foreground, leading to biased foreground removal results. The covariance matrices of the 'estimated' H I in presence of different levels of systematics are shown in Fig. 18. When no significant systematic effects are included as shown in the top panels, the reconstructed H I covariance matrix is largely diagonal, suggesting that no sizeable foreground leakage is present. However, as we increase the level of systematics to $5 \times 10^{-5}$, the small-scale stripes similar to the ones discussed in Section 5.1 appear. For the $5 \times 10^{-5}$ case, we can see that the diagonal component is still dominant and indeed as shown in Fig. 17 the H I power spectrum is still accurately measured. Increasing the level of systematics to $10^{-4}$, we can see that the covariance is completely dominated by the contamination from the systematics, leaving no observation window for the H I at all. In conclusion, the level of residual systematics needs to be contained at $<10^{-4}$ and ideally $\lesssim 5 \times 10^{-5}$ for accurate measurement of the H I power spectrum.

## 6 CONCLUSIONS

In this paper, we present the first proof of concept for measuring the H I power spectrum at $5 < z < 6$ using SKA-Low. We have presented an end-to-end simulation and data analysis pipeline, generating the sky signal, the interferometric observation, performing the imaging and the power spectrum estimation. We use the pipeline to generate realistic simulations consistent with deep observations of the EoR0

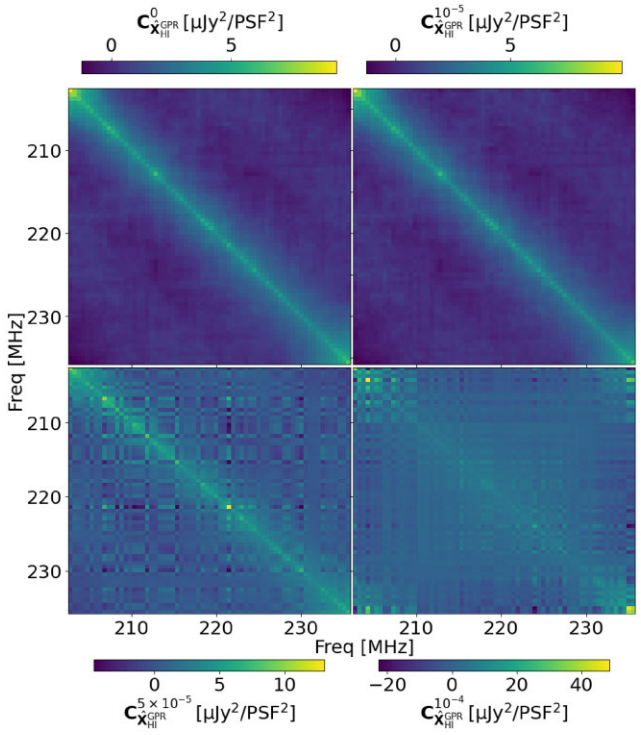


**Figure 18.** Top left-hand panel: The frequency–frequency covariance matrix $\mathbf{C}_{\hat{\mathbf{X}}^{\rm GPR}_{\rm HI}}$ of the 'estimated' H I image $\hat{\mathbf{X}}_{\rm HI}$ obtained using the GPR cleaning with no systematics. Top right-hand panel: The same covariance matrix with the left-hand panel except the simulation includes systematic effects with $10^{-5}$ fluctuations. Bottom left-hand panel: The same covariance matrix with $5 \times 10^{-5}$ systematic effects. Bottom right-hand panel: The same covariance matrix with $10^{-4}$ systematic effects.

field using SKA-Low and test foreground mitigation methods to present our forecasts for future SKA-Low observations.

We start by simulating the input sky signal including the H I and the foregrounds. Galactic foregrounds are generated based on templates from observed maps of the radio sky and extrapolated to the frequency range of our interests. We use a realistic radio source catalogue to simulate the extragalactic radio sources. The H I clustering signal is generated by using large-volume, realistic dark matter halo simulations with an H I HOD inpainting. Generating the sky signal with different levels of foreground residuals compared with the underlying H I signal, we find that:

(i) Assuming a realistic amplitude for foreground residuals after sky model subtraction to be at ∼80 mJy in the image cube, the foregrounds reside mainly at low $k_\parallel \lesssim 0.1\,{\rm Mpc}^{-1}$, leaving an observation window at high $k_\parallel$ for estimation of the H I power spectrum. Residual foregrounds need to be subtracted using blind source separation methods to enable the measurement of the H I power spectrum at large cosmological scales $k < 0.1\,{\rm Mpc}^{-1}$.

(ii) Testing PCA and GPR to remove the residual foregrounds, we find that if bright sources with flux density > 10 mJy are subtracted with the rest of the sources being modelled to 90 per cent accuracy, removing the residual foreground can enable detections of the H I power spectrum at large scales. The foreground wedge is consistent with the intrinsic foreground power coupled with the instrument chromaticity, with the wedge corresponding to the primary beam size.

(iii) Assuming no contribution from thermal noise and systematic effects, the empirical data covariance matrix calculated from the image cube reflects the true underlying covariance of the sky signal. Therefore, PCA and GPR can both sufficiently remove the foregrounds with trivial differences between these two methods.

(iv) From the image cube with $(1.5\,{\rm deg})^2$ sky area within the primary beam FoV, we can measure the H I power spectrum from $k \sim 0.02\,{\rm Mpc}^{-1}$ to $k \sim 0.3\,{\rm Mpc}^{-1}$.

The results suggest that measuring the H I power spectrum at $5 < z < 6$ for cosmological analysis using SKA-Low is viable and will open up a new window for cosmology in the near future. Using wide-field imaging and/or mosaicing, we can probe linear cosmological scales $k \sim 0.01\,{\rm Mpc}^{-1}$ to quasi-linear scales $k \sim 0.3\,{\rm Mpc}^{-1}$. The wide range of clustering scales probed can be used to constrain cosmology (Pourtsidou 2023).

The detection of the H I signal at large cosmological scales depends heavily on the robustness of foreground mitigation strategies. Simulating different level of depths for the observation, we find that:

(i) In general, future observations using SKA-Low contain a high level of thermal noise fluctuations. The effects of the thermal noise on the data covariance are visible even for deep observations >250 h.

(ii) The thermal noise fluctuations in the empirical data covariance matrix make residual foreground removal more difficult. Thermal noise creates numerical features on the foreground-removed image cube on small frequency scales, breaking the spectral smoothness of the data covariance.

(iii) As a result of the spectral fluctuations, foreground removal methods induce numerical artefacts on small frequency scales. The numerical artefacts leak power into the observation window which leads to significant bias on the H I power spectrum estimation. Even scales $k_\parallel > 0.1\,{\rm Mpc}^{-1}$ which can be probed with just foreground avoidance can be contaminated.

(iv) Comparing PCA and GPR, we find that GPR performs much better in the presence of thermal noise. The key factor is that GPR uses smooth kernels to model the signal and apply the fitted kernels instead of the actual data covariance matrix for the foreground removal. For observation with integration time >250 h, GPR can sufficiently remove the foregrounds and allow unbiased estimation of the H I power spectrum for $k_\parallel > 0.3k_\perp$ regions.

(v) For the integration time of 600 h, SKA-Low will be able to measure the H I power spectrum in the $5 < z < 6$ bin from 0.03 to $0.3\,{\rm Mpc}^{-1}$ with a SNR of ∼5 across the scales.

In conclusion, the viability of detecting the cosmological H I power spectrum at $5 < z < 6$ using SKA-Low depends on deep observations to preserve the spectral smoothness of the data covariance to facilitate sufficient foreground removal. It will allow accurate measurement of the H I power spectrum, on the premise that deep fields with effective integration time $\gtrsim$ 300 h are observed. Our results not only solidify the science case of measuring post-reionization cosmology with SKA-Low, but also provides insights into survey design for maximizing the scientific output of the instrument.

Finally, we provide a qualitative study into the systematic effects by introducing spectral fluctuations that can originate from bandpass instabilities and calibration errors. Testing the data analysis pipeline for different levels of systematics we find that:

(i) Systematic effects such as bandpass instabilities will introduce fluctuations in the small frequency interval, breaking the spectral smoothness of the foregrounds. It leads to spillover of the foreground power into the observation window outside the foreground wedge.

(ii) In the image cube averaged across all timesteps, the effective systematic errors across the frequency channels need to be small to suppress the contamination. If the level of the systematics is above $10^{-4}$, the power spectrum measurement will be biased across all scales of interests.

(iii) For systematic errors $\lesssim 5 \times 10^{-5}$, we find that using GPR to perform foreground removal gives unbiased estimation of the H I power spectrum.

The requirements on containing the systematic errors below 1 per cent level again highlight the need for deep observations with good understanding of the sky model and the instrument. With the unprecedented power of the SKA-Low array, we expect that future surveys will be sufficiently systematic-mitigated to enable the detection of the H I power spectrum for the high redshift, post-reionization Universe.

Our work strongly favours using the future SKA-Low data for H I science after cosmic reionization. We have demonstrated that the H I power spectrum can be measured with statistical significance using observational depth that can easily be reached using SKA-Low. Furthermore, we have showcased residual foreground removal using GPR that suppresses the foreground wedge to probe cosmological scales, which is robust in the presence of a reasonable level of systematic effects. The tools presented in this paper can be further used for more realistic simulations of SKA-Low observations to develop the data analysis pipeline towards future detections.

## ACKNOWLEDGEMENTS

We thank Keith Grainge and Mike Wilensky for discussions. EC acknowledges the support of a Royal Society Dorothy Hodgkin Fellowship and a Royal Society Enhancement Award. LW is a UK Research and Innovation Future Leaders Fellow (grant MR/V026437/1). AM acknowledges the support of a UK Research and Innovation Future Leaders Fellowship (grant MR/V026437/1). Apart from aforementioned packages, this work also uses PYTORCH (Paszke et al. 2019), NUMPY (Harris et al. 2020), SCIPY (Virtanen et al. 2020), ASTROPY (Astropy Collaboration 2018), CAMB (Lewis, Challinor & Lasenby 2000), EMCEE (Foreman-Mackey et al. 2013), and MATPLOTLIB (Hunter 2007).

## DATA AVAILABILITY

Data underlying this paper will be shared on reasonable request to the corresponding author.

## REFERENCES

Alam S. et al., 2021, Phys. Rev. D, 103, 83533
Alpher R. A., Bethe H., Gamow G., 1948, Phys. Rev., 73, 803
Amon A. et al., 2022, Phys. Rev. D, 105, 23514
Anderson C. J. et al., 2018, MNRAS, 476, 3382
Astropy Collaboration, 2018, AJ, 156, 123
Barry N., Hazelton B., Sullivan I., Morales M. F., Pober J. C., 2016, MNRAS, 461, 3135
Battye R. A., Browne I. W. A., Dickinson C., Heron G., Maffei B., Pourtsidou A., 2013, MNRAS, 434, 1239
Battye R. A., Davies R. D., Weller J., 2004, MNRAS, 355, 1339
Beardsley A. P. et al., 2016, ApJ, 833, 102
Bonaldi A., Brown M. L., 2015, MNRAS, 447, 1973
Bosman S. E. I. et al., 2022, MNRAS, 514, 55
Braun R., Bonaldi A., Bourke T., Keane E., Wagg J., 2019, preprint (arXiv:1912.12699)
Byrne R. et al., 2019, ApJ, 875, 70
Chang T.-C., Pen U.-L., Peterson J. B., McDonald P., 2008, Phys. Rev. Lett., 100, 91303
Chapman E. et al., 2012, MNRAS, 423, 2518
Chapman E. et al., 2013, MNRAS, 429, 165
Chapman E., Jelić V., 2019, in Mesinger A., ed, Foregrounds and their mitigation. The Cosmic 21-cm Revolution; Charting the first billion years of our universe, preprint (arXiv:1909.12369)
Chen Z., Wolz L., Battye R., 2023, MNRAS, 518, 2971
Chen Z., Wolz L., Spinelli M., Murray S. G., 2021, MNRAS, 502, 5259
CHIME Collaboration, 2022, ApJS, 261, 29
CHIME Collaboration, 2023, ApJ, 947, 16
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
Cornwell T. J., Golap K., Bhatnagar S., 2008, IEEE J. Sel. Top. Signal Process., 2, 647
Crighton N. H. M. et al., 2015, MNRAS, 452, 217
Cunnington S. et al., 2023a, MNRAS, 523, 2453
Cunnington S. et al., 2023b, MNRAS, 518, 6262
Cunnington S., Irfan M. O., Carucci I. P., Pourtsidou A., Bobin J., 2021, MNRAS, 504, 208
Datta A., Bowman J. D., Carilli C. L., 2010, ApJ, 724, 526
de Lera Acedo E. et al., 2020, preprint (arXiv:2003.12744)
DeBoer D. R. et al., 2017, PASP, 129, 45001
Dodelson S., Schmidt F., 2020, Modern Cosmology. Academic Press, Cambridge, Massachusetts
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
Ewall-Wice A. et al., 2016a, MNRAS, 460, 4320
Ewall-Wice A. et al., 2016b, ApJ, 831, 196
Fan X. et al., 2006, AJ, 131, 1203
Finkbeiner D. P., 2003, ApJS, 146, 407
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Harris C. R. et al., 2020, Nature, 585, 357
Haslam C. G. T., Klein U., Salter C. J., Stoffel H., Wilson W. E., Cleary M. N., Cooke D. J., Thomasson P., 1981, A&A, 100, 209
Haslam C. G. T., Salter C. J., Stoffel H., Wilson W. E., 1982, A&AS, 47, 1
Haynes M. P. et al., 2018, ApJ, 861, 49
Heywood I. et al., 2022, MNRAS, 509, 2150
Hothi I. et al., 2021, MNRAS, 500, 2264
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Irfan M. O. et al., 2022, MNRAS, 509, 4923
Kaiser N., 1987, MNRAS, 227, 1
Kern N. S., Liu A., 2021, MNRAS, 501, 1463
Koopmans L. et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 1 , preprint (arXiv:1505.07568)
Kovetz E. D. et al., 2017, preprint (arXiv:1709.09066)
Leroy A. K., Walter F., Brinks E., Bigiel F., de Blok W. J. G., Madore B., Thornley M. D., 2008, AJ, 136, 2782
Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
Lian X., Xu H., Zhu Z., Hu D., 2020, MNRAS, 496, 1232
Liu A., Parsons A. R., Trott C. M., 2014, Phys. Rev. D, 90, 23018
Long H., Morales-Gutiérrez C., Montero-Camacho P., Hirata C. M., 2022, preprint (arXiv:2210.02385)
Lynch C. R. et al., 2021, PASA, 38, e057
Mao Y., Tegmark M., McQuinn M., Zaldarriaga M., Zahn O., 2008, Phys. Rev. D, 78, 23529
Masui K. W. et al., 2013, ApJ, 763, L20
Matérn B., 1966, Spatial Variation; Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations. Stockholm. Statens Skogsforskningsinstitut. University of Sweden, Meddelanden, https://books.google.co.uk/books?id=HTWTwgEACAAJ
Mazumder A., Datta A., Chakraborty A., Majumdar S., 2022, MNRAS, 515, 4020
Mertens F. G. et al., 2020, MNRAS, 493, 1662
Mertens F. G., Ghosh A., Koopmans L. V. E., 2018, MNRAS, 478, 3640

Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Monaco P., Theuns T., Taffoni G., 2002, MNRAS, 331, 587
Morales M. F., Hazelton B., Sullivan I., Beardsley A., 2012, ApJ, 752, 137
Morales M. F., Hewitt J., 2004, ApJ, 615, 7
Mort B. J., Dulwich F., Salvini S., Adami K. Z., Jones M. E., 2010, in 2010 IEEE International Symposium on Phased Array Systems and Technology. p. 690
Murray S. G., Trott C. M., 2018, ApJ, 869, 25
Offringa A. R. et al., 2015, PASA, 32, e008
Offringa A. R., de Bruyn A. G., Biehl M., Zaroubi S., Bernardi G., Pandey V. N., 2010, MNRAS, 405, 155
Pagano M. et al., 2023, MNRAS, 520, 5552
Parsons A. R. et al., 2014, ApJ, 788, 106
Parsons A. R., Pober J. C., Aguirre J. E., Carilli C. L., Jacobs D. C., Moore D. F., 2012b, ApJ, 756, 165
Parsons A., Pober J., McQuinn M., Jacobs D., Aguirre J., 2012a, ApJ, 753, 81
Paszke A. et al., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, Advances in Neural Information Processing Systems 32. Curran Associates, Inc., New York (US), p. 8024
Patil A. H. et al., 2017, ApJ, 838, 65
Paul S., Santos M. G., Townsend J., Jarvis M. J., Maddox N., Collier J. D., Frank B. S., Taylor R., 2021, MNRAS, 505, 2039
Paul S., Santos M. G., Chen Z., Wolz L., 2023, preprint (arXiv:2301.11943)
Planck Collaboration, 2020a, A&A, 641, A1
Planck Collaboration, 2020b, A&A, 641, A6
Pourtsidou A., 2023, MNRAS, 519, 6246
Rahmati A., Pawlik A. H., Raičević M., Schaye J., 2013, MNRAS, 430, 2427
Rajohnson S. H. A. et al., 2022, MNRAS, 512, 2697
Remazeilles M., Dickinson C., Banday A. J., Bigot-Sazy M. A., Ghosh T., 2015, MNRAS, 451, 4311
Riess A. G. et al., 2022, ApJ, 934, L7
Sabater J. et al., 2021, A&A, 648, A2
Santos M. et al., 2016, in MeerKAT Science: On the Pathway to the SKA. p. 32, preprint (arXiv:1709.06099)
Schaerer D., 2002, A&A, 382, 28
Sinigaglia F. et al., 2022, ApJ, 935, L13
Soares P. S., Watkinson C. A., Cunnington S., Pourtsidou A., 2022, MNRAS, 510, 5872
Sokolowski M., Wayth R. B., Lewis M., 2015, in 2015 IEEE Global Electromagnetic Compatibility Conference (GEMCCON). p. 1
Spinelli M., Carucci I. P., Cunnington S., Harper S. E., Irfan M. O., Fonseca J., Pourtsidou A., Wolz L., 2022, MNRAS, 509, 2048
Spinelli M., Zoldan A., De Lucia G., Xie L., Viel M., 2020, MNRAS, 493, 5434
Square Kilometre Array Cosmology Science Working Group et al., 2020, PASA, 37, e007
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
Tegmark M., Hamilton A. J. S., Xu Y., 2002, MNRAS, 335, 887
The HERA Collaboration, 2023, ApJ, 945, 124
Thyagarajan N. et al., 2015, ApJ, 804, 14
Tingay S. J. et al., 2013, PASA, 30, e007
Trott C. M. et al., 2016, ApJ, 818, 139
Trott C. M. et al., 2018, ApJ, 867, 15
Trott C. M. et al., 2020, MNRAS, 493, 4711
Trott C. M., Wayth R. B., 2017, PASA, 34, e061
Villaescusa-Navarro F. et al., 2018, ApJ, 866, 135
Virtanen P. et al., 2020, Nature Methods, 17, 261
Wang L., Cui X., Zhu H., Tian W., 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14). p. 64, preprint (arXiv:1501.04645)
Wilensky M. J., Morales M. F., Hazelton B. J., Barry N., Byrne R., Roy S., 2019, PASP, 131, 114507
Wilson T. L., Rohlfs K., Hüttemeister S., 2013, Tools of Radio Astronomy, Springer, Berlin, Heidelberg
Wolz L. et al., 2022, MNRAS, 510, 3495
Wolz L., Abdalla F. B., Blake C., Shaw J. R., Chapman E., Rawlings S., 2014, MNRAS, 441, 3271
Wolz L., Murray S. G., Blake C., Wyithe J. S., 2019, MNRAS, 484, 1007
Wolz L., Tonini C., Blake C., Wyithe J. S. B., 2016, MNRAS, 458, 3399
Wyithe J. S. B., Loeb A., 2009, MNRAS, 397, 1926
Xu Y., Wang X., Chen X., 2015, ApJ, 798, 40
Yatawatta S. et al., 2013, A&A, 550, A136
Zheng H. et al., 2017, MNRAS, 464, 3486
Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, J. Open Source Softw., 4, 1298

## APPENDIX A: QUADRATIC ESTIMATOR FOR POWER SPECTRUM ESTIMATION

We present the quadratic estimator for the power spectrum estimation used in this paper, following equations (3) and (4). The aim of using the quadratic estimator formalism is to incorporate renormalization of the estimator after the operations of foreground cleaning and frequency tapering. It also performs bias correction to remove the thermal noise power spectrum and potentially some bias from the GPR cleaning. Our formalism follows closely the work of Kern & Liu (2021) and Chen et al. (2023). Note that we are not aiming to construct the covariance for the total data vector with the number of elements being the number of pixels times the number of frequency channels $N_{\rm pix} \times N_{\rm f}$. The resulting covariance matrix of size $(N_{\rm pix} \times N_{\rm f})^2$ is too large and therefore not of our interests for a preliminary study. Instead, we construct the estimator for each pixel across the $k_\parallel$ direction, so that we are only dealing with one pixel at a time with a covariance matrix of size $N_{\rm f} \times N_{\rm f}$.

In this section, we use $i$ to denote the $i$th pixel in the Fourier transformed image cube. For each pixel, the Fourier density gives a bandpower vector $(\hat{\mathbf{p}}^i_{\rm T})_\alpha$, with the $\alpha$th element being the power spectrum at $(\boldsymbol{k}^i_\perp, k^\alpha_\parallel)$. The quadratic estimator can be written as

$$(\hat{\mathbf{p}}^i_{\rm T})_\alpha = \left(\tilde{\mathbf{d}}^i\right)^\dagger \mathbf{E}^i_\alpha \tilde{\mathbf{d}}^i - \hat{\mathbf{b}}_\alpha, \tag{A1}$$

where $\mathbf{E}^i_\alpha$ and $\hat{\mathbf{b}}_\alpha$ are the estimation matrix and bias correction respectively. Here, $\tilde{\mathbf{d}}^i$ is the data vector along the frequency direction for the $i$th pixel. We collapse the Fourier transform along the transverse directions and the PSF deconvolution in this data vector so that for the $j$th frequency channel

$$(\tilde{\mathbf{d}}^i)_j = \int \frac{{\rm d}^2 x_\perp}{\mathbf{V}} \exp\left[-i\boldsymbol{k}^i_\perp \cdot \boldsymbol{x}_\perp\right] \times \left(\frac{\lambda^2}{2k_{\rm B}}\right)^2 \frac{I(\boldsymbol{x}_\perp, x^j_\parallel)}{A^2(\boldsymbol{x}_\perp, x^j_\parallel)} \Big/ \widetilde{\rm PSF}(\boldsymbol{k}^i_\perp, f_{\rm c}). \tag{A2}$$

The estimation matrix $\mathbf{E}^i_\alpha$ can be written as

$$(\mathbf{E}^i_\alpha)_\alpha = \sum_\beta \mathbf{M}_{\alpha\beta}\mathbf{R}^\dagger\mathbf{T}^\dagger\mathbf{F}^\dagger w_\beta \mathbf{F}\,\mathbf{T}\,\mathbf{R} = \sum_\beta \mathbf{M}_{\alpha\beta}\mathbf{R}^\dagger\mathbf{T}^\dagger\mathbf{C}_{,\beta}\mathbf{T}\mathbf{R}, \tag{A3}$$

where $\mathbf{M}_{\alpha\beta}$ is the renormalization matrix, $\mathbf{T}$ is the frequency taper, $w_\beta$ is the selection matrix with all elements being zero except the $\beta$th diagonal element and $\mathbf{F}$ is the 1D discrete Fourier transform kernel along the frequency direction. $\mathbf{C}_{,\beta} = \mathbf{F}^\dagger w_\beta \mathbf{F}$ is the Fourier operator. $\mathbf{R}$ is the foreground removal operation. For PCA as described in equation (9), the removal matrix is $\mathbf{R} = \mathbf{I} - \mathbf{A}\mathbf{A}^{\rm T}$ where $\mathbf{I}$ is the identity matrix. For GPR as described in equation (12), the removal matrix is $\mathbf{R} = \mathbf{I} - \mathbf{K}_{\rm fg}\left(\mathbf{K}_{\rm fg} + \mathbf{K}_{\rm n} + \mathbf{K}_{\rm HI}\right)^{-1}$.

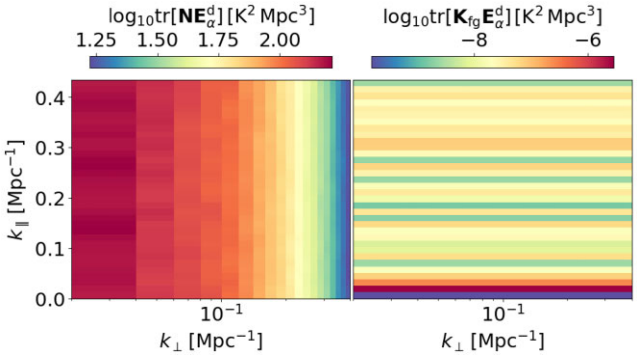


**Figure A1.** Left-hand panel: The cylindrical power spectrum of the thermal noise calculated using tr[$\mathbf{NE}_\alpha^{\rm d}$] for the case of 360 h of integration time. Right-hand panel: The cylindrical power spectrum of the GPR bias correction using tr[$\mathbf{K}_{\rm fg}\mathbf{E}_\alpha^{\rm d}$] for the case of 360 h of integration time.

The renormalization matrix can be calculated by taking the expectation value of equation (A1)

$$\langle(\hat{\mathbf{p}}_{\rm T}^i)_\alpha\rangle = \sum_\beta {\rm tr}\left[\mathbf{C}_{,\beta}\mathbf{E}_\alpha^i\right](\mathbf{p}_{\rm T}^i)_\beta + {\rm tr}\left[\left(\mathbf{N}+\mathbf{C}_{\rm fg}\right)\mathbf{E}_\alpha^{\rm d}\right] - \hat{b}_\alpha^{\rm d}. \tag{A4}$$

Following Kern & Liu (2021), we can form the quantity

$$H_{\alpha\beta} = {\rm tr}\left[\mathbf{R}^\dagger\mathbf{T}^\dagger\mathbf{C}_{,\alpha}\mathbf{T}\,\mathbf{R}\mathbf{C}_{,\beta}\right], \tag{A5}$$

and choose $\mathbf{M} = \mathbf{H}^{-1/2}$ (Tegmark, Hamilton & Xu 2002) to renormalize the estimator.

In order to remove the bias in the power spectrum estimation from the foregrounds and the thermal noise, from equation (A4) we can choose

$$\hat{b}_\alpha^{\rm d} = {\rm tr}\left[\left(\mathbf{N}+\mathbf{C}_{\rm fg}\right)\mathbf{E}_\alpha^{\rm d}\right] \tag{A6}$$

to remove the bias. In reality though, we do not know the underlying thermal noise and the foregrounds. In order to remove the noise bias, we calculate $\mathbf{N}$ by simulating 1000 realizations of the thermal noise using the same $\sigma_{\rm N}$. Here, $\sigma_{\rm N}$ is assumed to be a known quantity, which is the case for our simulation. In real observations, a good estimation of $\sigma_{\rm N}$ can be obtained by calculating the fluctuations of the Stokes V visibility data on long baselines (e.g. Paul et al. 2021). For each realization, we pass the visibility data to the same imaging pipeline to generate the image cubes. For each pixel in the image cube, we then calculate the average frequency–frequency correlation across all realizations to obtain an estimation of $\mathbf{N}$. We bin the resulting noise bias tr[$\mathbf{NE}_\alpha^{\rm d}$] into cylindrical $k$-space and show the thermal noise power spectrum in Fig. A1 for the case with 360 h of integration time. The vertical stripes follow the baseline densities on different scales.

The covariance of the foregrounds can be extracted from the GPR fitted kernel $\mathbf{K}_{\rm fg}$. Note that, since we work on the frequency–frequency covariance, $\mathbf{K}_{\rm fg}\mathbf{E}_\alpha^{\rm d}$ is the same for each pixel, and therefore there is no $k_\perp$ dependence of the bias term as shown in the right-hand panel of Fig. A1. We note that this is not a result of GPR but the result of our simplified quadratic estimator formalism. Nevertheless, it gives us a good estimation of the order of magnitude of the GPR bias correction. As one can see, the correction is at least two orders of magnitude smaller than the H I signal shown in Fig. 15, and therefore this bias correction is negligible in our case.

Finally, we comment on the fact that in the power spectrum estimation, the 2D Fourier transform shown in equation (A2) is applied to the data before the GPR removal $\mathbf{R}$, while the GPR fitting for the kernels are done on the original data vector before the transform. These two operations are commutable, as the GPR removal only operates along the frequency direction, independent of the 2D Fourier transform on the transverse plane. We verified that there is no visible difference in the resulting power spectrum if these two operations are swapped. Performing the 2D Fourier transform first allows us to only construct the estimator one pixel at a time, providing massive speed-up.

## APPENDIX B: CAVEATS OF THE SIMULATIONS

We discuss the limitations of our simulation settings. Specifically, we quantify the effects of limited $(10.5\,{\rm deg})^2$ sky area for the input signal, coupled with the instrument beam which gets cut off at 1 per cent at the 10.5 deg angular extent. Furthermore, we discuss the Gaussian calibration errors simulated in terms of its structure in frequency.

The primary beam of the instrument is shown in Fig. B1. Around the centre $(10.5\,{\rm deg})^2$ region, the beam only goes down to $10^{-2}$, introducing sharp features in the simulation. We first note that, as discussed in Section 4, the image power spectrum does not show a clear wedge structure due to the small image size. To investigate the chromatic structure of the data, we instead calculate the delay power spectrum directly from visibility and present it in Fig. B2. As shown in the top panel, the full foreground delay power spectrum

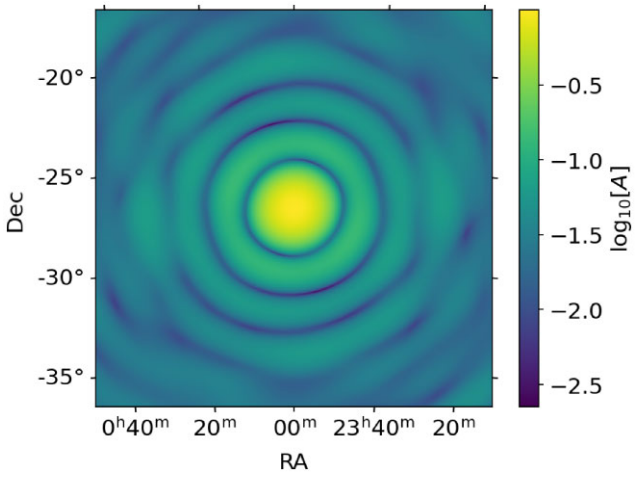


**Figure B1.** The primary beam of the instrument around the pointing centre in our simulation. The image size is $(20\,{\rm deg})^2$. The beam is simulated at the central frequency 220 MHz and averaged over all time steps for one station.

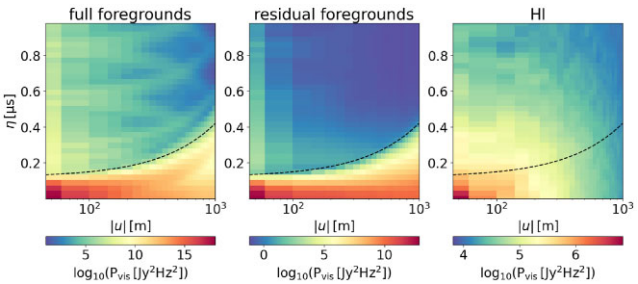


**Figure B2.** The delay power spectra of the visibility data in our simulation. Left-hand panel: The delay power spectrum of the full foreground signal. Central panel: The residual foregrounds. Right-hand panel: the H I signal. The $|u|$ and $\eta$ range correspond roughly to the $k$-range of the cylindrical power spectra shown in the paper. The black dashed line denotes the foreground wedge.

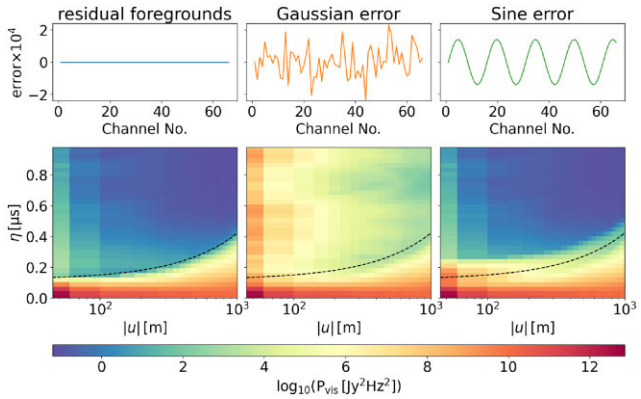


**Figure B3.** The delay power spectra of the visibility data after applying gain errors. Left-hand panel: The delay power spectrum of the residual foreground signal with no error applied. Central panel: The residual foregrounds with Gaussian errors as shown in the top and the delay power spectrum shown in the bottom. Right-hand panel: The residual foregrounds with sine errors.

shows a clear wedge structure. Above the wedge, the effect of sky signal getting cut off at 10.5 deg can be seen as the diamond-shape structures. Assuming the bright sources are removed as described in Section 2, we calculate the delay power spectrum of the residual foregrounds shown in the centre panel of Fig. B2. The chromatic features disappear as there is no bright emission coming from the beam sidelobes. Finally, we also present the delay power spectrum of the H I signal in Fig. B2 to show that the sky cut-off does not affect the H I simulation.

The calibration errors for SKA-Low observations are likely smooth in frequency (Byrne et al. 2019), which are not the Gaussian random fluctuations we use. Using the delay power spectra, we then investigate the assumption of the Gaussian gain errors described in Section 3. For comparison, we simulate another type of error that follows the sine function with a period of 15 frequency channels. The errors are then rescaled so that the standard deviation across the channels is $10^{-4}$. The sine errors are then compared against the Gaussian errors as shown in Fig. B3.

For the sine error case shown in the right-hand panel, the foreground wedge gets lifted into higher delay. Comparing to the Gaussian error case in the central panel, the leakage still concentrates around relatively low delay. This means that the foreground contamination can be easier to remove for GPR, as its structure has large frequency intervals. In the Gaussian case, however, the scatter of the foreground power into higher delay is visible across all scales. The foreground contamination is at the smallest frequency interval, which is difficult to distinguish from the H I signal. Therefore, we conclude that the conclusions reached in Section 5.2 are robust, as the foreground contamination is not an optimistic case.

We emphasize that the smooth frequency structures of the gain errors pose other challenges in sky modelling and continuum subtraction, which are beyond the scope of this paper and left for future work.

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