SKAO H i intensity mapping: blind foreground subtraction challenge

Marta Spinelli, Isabella P. Carucci, Steven Cunnington, Stuart E. Harper, Melis O. Irfan, José Fonseca, Alkistis Pourtsidou, Laura Wolz

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **509,** 2048–2074 (2022) https://doi.org/10.1093/mnras/stab3064
Advance Access publication 2021 October 22

SKAO H I intensity mapping: blind foreground subtraction challenge

Marta Spinelli [1,2,3,4★] Isabella P. Carucci [5,6,7] Steven Cunnington [8] Stuart E. Harper,[9] Melis O. Irfan,[4,8] José Fonseca [4,8,10,11] Alkistis Pourtsidou [4,8] and Laura Wolz[9]

[1]*Institute for Particle Physics and Astrophysics, ETH Zürich, Wolfgang Pauli Strasse 27, CH-8093 Zürich, Switzerland*
[2]*INAF – Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, I-34143 Trieste, Italy*
[3]*IFPU – Institute for Fundamental Physics of the Universe, Via Beirut 2, I-34014 Trieste, Italy*
[4]*Department of Physics and Astronomy, University of the Western Cape, Robert Sobukhwe Road, Bellville 7535, South Africa*
[5]*Dipartimento di Fisica, Università degli Studi di Torino, Via P. Giuria 1, I-10125 Torino, Italy*
[6]*INFN – Istituto Nazionale di Fisica Nucleare, Sezione di Torino, Via P. Giuria 1, I-10125 Torino, Italy*
[7]*AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France*
[8]*Astronomy Unit, School of Physics and Astronomy, Queen Mary University of London, Mile End Road, London E1 4NS, UK*
[9]*Jodrell Bank Centre for Astrophysics, Alan Turing Building, Department of Physics & Astronomy, School of Natural Sciences, The University of Manchester, Oxford Road, Manchester M13 9PL, UK*
[10]*Dipartimento di Fisica 'G. Galilei', Università degli Studi di Padova, Via Marzolo 8, I-35131 Padova, Italy*
[11]*INFN – Istituto Nazionale di Fisica Nucleare, Sezione di Padova, Via Marzolo 8, I-35131 Padova, Italy*

Accepted 2021 October 18. Received 2021 October 17; in original form 2021 July 29

ABSTRACT

Neutral Hydrogen Intensity Mapping (H I IM) surveys will be a powerful new probe of cosmology. However, strong astrophysical foregrounds contaminate the signal and their coupling with instrumental systematics further increases the data cleaning complexity. In this work, we simulate a realistic single-dish H I IM survey of a 5000 deg^2 patch in the 950–1400 MHz range, with both the MID telescope of the SKA Observatory (SKAO) and MeerKAT, its precursor. We include a state-of-the-art H I simulation and explore different foreground models and instrumental effects such as non-homogeneous thermal noise and beam side lobes. We perform the first Blind Foreground Subtraction Challenge for H I IM on these synthetic data cubes, aiming to characterize the performance of available foreground cleaning methods with no prior knowledge of the sky components and noise level. Nine foreground cleaning pipelines joined the challenge, based on statistical source separation algorithms, blind polynomial fitting, and an astrophysical-informed parametric fit to foregrounds. We devise metrics to compare the pipeline performances quantitatively. In general, they can recover the input maps' two-point statistics within 20 per cent in the range of scales least affected by the telescope beam. However, spurious artefacts appear in the cleaned maps due to interactions between the foreground structure and the beam side lobes. We conclude that it is fundamental to develop accurate beam deconvolution algorithms and test data post-processing steps carefully before cleaning. This study was performed as part of SKAO preparatory work by the H I IM Focus Group of the SKA Cosmology Science Working Group.

Key words: large-scale structure of Universe – radio lines: galaxies.

1 INTRODUCTION

Over the next few decades, a new generation of radio telescopes will revolutionize our understanding of cosmology via observations of the radio continuum emission and the 21-cm line emission from neutral hydrogen gas (H I). Most notably, the telescope arrays of the SKA Observatory (SKAO) will conduct large radio surveys in order to test the standard cosmological model (SKA Cosmology SWG 2020).

With the combined frequency range of both the SKAO-LOW and SKAO-MID telescopes, from 1.4 GHz down to 50 MHz, the H I surveys can trace the matter distribution from the present time to the epoch of reionization and beyond. H I gas, as the first and most abundant element in the Universe, is an excellent tracer of the large-scale structure and its evolution. However, due to the weakness of its emission, the highest redshift at which a galaxy has been observed thanks to its 21-cm line is $z \sim 0.376$ (Fernández et al. 2016), and even the forthcoming SKAO surveys can only detect statistically significant samples for cosmology up $z \sim 0.4$ (SKA Cosmology SWG 2020).

Intensity Mapping (IM) is a relatively recent technique to circumvent detection limitations by observing the integrated H I line emission from unresolved sources in large volume elements of the sky (Bharadwaj, Nath & Sethi 2001; Battye, Davies & Weller 2004; Chang et al. 2008; Peterson et al. 2009; Wyithe & Loeb 2009; Seo et al. 2010). Neutral Hydrogen Intensity Mapping (H I IM) surveys are very time efficient compared to traditional galaxy surveys as the low spatial resolution and large redshift range allow us to observe immense cosmic volumes within relatively short observation times. The resulting H I maps trace the largest scales of the matter distribution of the underlying dark matter field with excellent redshift resolution due to the telescope's fine frequency channelization. Even though the original idea for H I IM stems from using large single-dish telescopes such as the Green Bank Telescope (GBT; Chang

★ E-mail: spinemart@gmail.com

et al. 2010; Masui et al. 2013; Switzer et al. 2013), surveys can be conducted by a range of instrumental settings such as compact interferometric arrays or arrays of smaller dish telescopes.

For the SKAO Project, the planned cosmological IM surveys will be conducted in the so-called single-dish mode: Each dish operates as a single telescope, and maps are co-added (Battye et al. 2013; Bull et al. 2015). The resulting angular resolution of about 1 degree at $z \sim 0.4$ is very low, but the scanning is fast, and a large area coverage of $\sim$30 000 deg^2 can be achieved using a few thousand hours (SKA Cosmology SWG 2020). The single-dish observations can be complemented by deep interferometric surveys that access the smaller scales beyond the primary beam of the telescope. Additionally, it has been shown that a large amount of small-scale information can be retrieved from the line-of-sight modes with its high redshift resolution (Villaescusa-Navarro, Alonso & Viel 2017). Present and forthcoming instruments with planned H I IM surveys are BINGO (Battye et al. 2016), CHIME (Bandura et al. 2014), FAST (Hu et al. 2020), HIRAX (Newburgh et al. 2016), Tianlai (Das et al. 2018), and uGMRT (Chakraborty et al. 2021). Most importantly for this work, the SKAO pathfinder MeerKAT in South Africa is already taking pilot data for its MeerKLASS survey (Santos et al. 2017; Wang et al. 2021) and the 64 MeerKAT dishes will eventually be incorporated into the SKAO-MID telescope array when it will commence operation in the late 2020's.

However, the detection of the H I IM has proven to be observationally challenging. Since its first application by Chang et al. (2010) with the GBT data more than a decade ago, few other studies have claimed the detection of the signal, and always in cross-correlation with a galaxy catalogue (Masui et al. 2013; Anderson et al. 2018; Wolz et al. 2021). The main obstacle to detecting the H I signal comes from the presence of astrophysical foregrounds orders of magnitude stronger than the H I signal. While astrophysical foregrounds, predominantly due to synchrotron and free–free emission at the relevant (around 1 GHz) frequencies, have a known spatial distribution and frequency correlation, their convolution with instrumental systematics and other observational effects can render signal separation a very challenging task.

In recent years, many studies have addressed the problem in the context of single-dish H I IM and investigated the quality of foreground removal methods on data (Switzer et al. 2015; Wolz et al. 2017) as well as simulations (e.g. Ansari et al. 2012; Wolz et al. 2014; Alonso et al. 2015; Shaw et al. 2015; Olivari, Remazeilles & Dickinson 2016; Carucci, Irfan & Bobin 2020; Fonseca & Liguori 2021; Makinen et al. 2021; Soares et al. 2021; Yohana et al. 2021), where blind and non-parametric methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Generalized Morphological Component Analysis (GMCA) have proven most powerful. In addition, many studies set particular focus on individual observational systematics such as primary beam effects (Matshawule et al. 2021), polarization leakage (Shaw et al. 2015; Spinelli, Bernardi & Santos 2018; Carucci et al. 2020; Cunnington et al. 2021a), 1/*f* noise (Harper et al. 2018; Chen et al. 2020; Li et al. 2021), and radio frequency interference due to satellites (Harper & Dickinson 2018). Findings of these studies point to the fact that all observational effects sensitively depend on the individual instrument and survey design, making end-to-end simulations a crucial requirement towards a valid detection of the H I IM signal in autocorrelation.

In this study, we present a detailed study of foreground removal methods for MeerKAT and future SKAO-MID H I IM surveys implemented by the H I Intensity Mapping Focus Group of the SKA Cosmology Science Working Group (SWG). For the first time, we conduct a Blind Foreground Subtraction Challenge where participants are presented with simulated data cubes of unknown foregrounds, H I signal, and instrumental specifics such as the beam and noise level. We implement a realistic scanning strategy for an $\sim$5000 deg^2 survey resulting in anisotropic noise, as well as a more sophisticated beam model with chromatic side lobes for the SKAO-MID and MeerKAT dishes in addition to the conventional Gaussian beam approximation. We use two different implementations of the astrophysical foregrounds in order to investigate the impact of foreground models on the separation techniques. The true level of H I signal and noise in the mocks was not known to the participants of the Blind Challenge and the submitted results have not been adjusted or modified after unblinding the submissions. The participants used a total of nine different pipelines to clean the data cubes, ranging from different kinds of blind (PCA, FASTICA, and GMCA) to non-blind (parametric fitting) source separation algorithms. We stress that, although the cleaning techniques employed in this work have been proven powerful when applied to less realistic simulations, we do not expect them to perform perfectly facing these new complexities. We are thus equally interested in their absolute and relative performances to understand weaknesses and strengths.

The paper is structured as follows: In Section 2, we describe the end-to-end simulation and properties of the mock data cubes. In Section 3, we outline the cleaning methods and in Section 4 the statistical estimators we use for the comparison among cleaned residuals and input maps. In Section 5, we describe how we run the Challenge. Results are presented in Section 6, followed by a broad discussion in Section 7. We draw our conclusions and give future perspectives in Section 8.

2 SIMULATIONS

In this section, we describe the various ingredients of our mock data. The sky simulations of the signal and the foregrounds are presented in Sections 2.1 and 2.2, respectively. We consider two different foreground models: a simplistic one based on Santos et al. (2005) (MS_{05}, Section 2.2.1) and a more realistic and physically motivated one based on available data and the Planck Sky Model (Delabrouille et al. 2013) (PSM, Section 2.2.2). All components of the sky simulation are summarized in Table 1. In Section 2.3, we describe the instrumental simulations, detailing the assumed beam model (Section 2.3.1) and the observing strategy and noise (Section 2.3.3). We focus on both an SKAO-MID telescope-like survey and a MeerKAT-like IM survey, considering for the former case a smaller beam and a lower noise level. We also explore two different beam models, a standard Gaussian beam and a more realistic beam model that includes side lobes based on the apertures of the MeerKAT/SKAO-MID dishes, to which we will be referring to as the Airy beam. The different telescope and survey specifications are reported in Table 2. We focus on a frequency range covering 950–1400 MHz binned into 512 observational channels, similar to the MeerKAT's *L* band. Our sky maps are created using the HEALPix format (Górski et al. 2005), at $N_{\rm side} = 512$, providing $\sim$7 arcmin resolution. Fig. 1 illustrates the footprint considered in this work.

We model the observed sky temperature, $T_{\rm obs}$, in the direction $\hat{\mathbf{n}}$ and as a function of frequency ν as

$$T_{\rm obs}(\nu, \hat{\mathbf{n}}) = \int {\rm d}\Omega\, B(\nu, \hat{\mathbf{n}}, \Omega) \left[T_{\rm fg}(\nu, \hat{\mathbf{n}}) + T_{\rm HI}(\nu, \hat{\mathbf{n}})\right] + T_{\rm noise}(\nu, \hat{\mathbf{n}}), \quad (1)$$

where $T_{\rm fg}(\nu, \hat{\mathbf{n}})$ is the astrophysical foreground emission and $T_{\rm HI}(\nu, \hat{\mathbf{n}})$ is the 21-cm signal from cosmic H I. Both are convolved with the telescope beam $B(\nu, \hat{\mathbf{n}}, \Omega)$, pointing in the direction $\hat{\mathbf{n}}$ and

Table 1. A brief description of the components of the sky simulations. The cosmic neutral hydrogen signal (H I) is presented in more detail in Section 2.1, while the foreground components are described in Section 2.2. We consider two different sets of foreground simulations: the MS_{05} and the PSM model (see the main text for details).

Sky component		Description	
H I		PINOCCHIO LPT light-cone haloes painted with an $M_{\rm H\,I}$–$M_{\rm halo}$ relation extrapolated from the GAEA semi-analytical model (Spinelli et al. 2020)	
Foregrounds		MS_{05} model (as in Santos, Cooray & Knox 2005), parameters of equation (5) $\{A[\mathrm{mK}^2]; \beta; \alpha\}$	PSM model (as in Carucci et al. 2020)
	Galactic synchrotron	{700; 2.4; 2.80}	Haslam 408 MHz, (Remazeilles et al. 2015) with spatially varying synchrotron spectral index (Miville-Deschênes et al. 2008)
	Free–free	{0.088; 3.0; 2.15}	Hα template (Finkbeiner 2003)
	Extragalactic free–free	{0.014; 1.0; 2.10}	–
	Point sources	{57; 1.1; 2.07}	Source count model with flux cut at 0.1 Jy (Olivari et al. 2018)

Table 2. Simulated telescope and survey parameters.

Parameter	SKAO	MeerKAT
$N_{\rm dish}$	133	64
T_{rx}	7.5 K	9.8 K
$T_{\rm spill}$	4 K	4 K
$\Delta\nu$	1 MHz	1 MHz
ρ_ν	20	∞
Strip declinations	$-45, -30, -15, 0$	
Strip width	15 deg	
Scan speed	1 deg s^{-1}	
$\Omega_{\rm sky}$	$\approx$5000 deg^2	

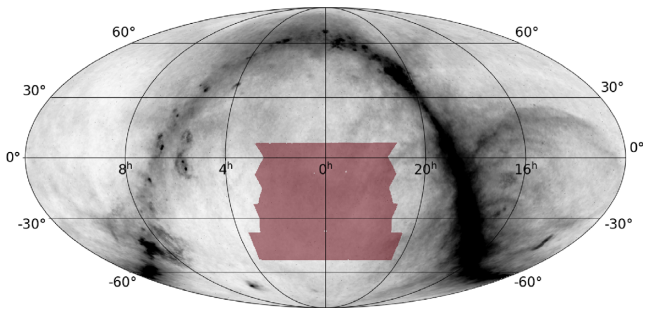

Figure 1. An illustration of the footprint considered in this work (maroon shaded region) and its position with respect to the Galactic plane (in equatorial coordinates). The foreground emission model used for this plot is described in Section 2.2.2 and a more accurate description of the shape of the footprint can be found in Section 2.3.

covering the solid angle $\mathbf{\Omega}$. The response of the telescope also adds a thermal noise component $T_{\rm noise}(\nu, \hat{\mathbf{n}})$ that varies with frequency and also with direction since we take into account a scanning strategy.

Our simulations could be made more complex adding other systematics such as missing channels due to RFI, $1/f$ noise, or satellite contamination. In this work, we focus on the inclusion of realistic beam modelling and non-homogeneous noise in order to first establish their impact on the cleaning methods, leaving further systematics to future studies.

2.1 Cosmological simulation

Since the quality of the foreground cleaning procedure for IM experiments will inevitably depend on the properties of the H I signal, having a realistic description of its large-scale distribution and evolution with redshift is crucial. At low redshifts, neutral hydrogen is expected to be hosted only in high-density regions where, shielded from UV radiation, it has survived the reionization process. Given the relatively poor spatial resolution of single-dish experiments, each voxel in the sky is expected to host a large number of galaxies. This implies that it is possible to simulate the H I clustering without describing the single galaxies but by considering the total amount of neutral hydrogen mass $M_{\rm H\,I}$ hosted by a halo with mass $M_{\rm halo}$, i.e. the $M_{\rm H\,I}$–$M_{\rm halo}$ relation (e.g. Bagla, Khandai & Datta 2010; Carucci et al. 2015; Carucci, Corasaniti & Viel 2017; Modi et al. 2019; Asorey et al. 2020; Zhang et al. 2021). In this work, we use the H I Probe Populator (`HIP-POP`[1]) that combines a full-sky halo light-cone with information on the baryonic content extrapolated from a semi-analytical model of galaxy formation and evolution. `HIP-POP` uses the PINOCCHIO code (Monaco, Theuns & Taffoni 2002; Taffoni, Monaco & Theuns 2002; Monaco et al. 2013; Munari et al. 2017) to generate catalogues of cosmological dark matter haloes with a known mass, position, velocity, and merger history. PINOCCHIO is based on the Lagrangian Perturbation Theory (LPT) and is able to reproduce, with very good accuracy, the hierarchical formation of dark matter haloes. We produce 1 Gpc h^{-1} boxes using 2048^3 particles to reach a minimum halo mass of $\lesssim 10^{11}$ $\mathrm{M}_\odot$ h^{-1} and construct a full-sky light-cone. On the largest scale, there will be repetitions due to the limited size of the box that we replicate to fill the light-cone. This is not a problem in our case since we will select a relatively small patch at low redshift.

We populate each halo following Spinelli et al. (2020), who used the outputs of the semi-analytical model GAEA (De Lucia, Kauffmann & White 2004; De Lucia et al. 2014; Hirschmann, De Lucia & Fontanot 2016; Zoldan et al. 2017). Specifically, we use the version of the code described in Xie et al. (2017), run on the merger trees of the Millennium II simulation (MII; Boylan-Kolchin et al. 2009). With 2160^3 particles in a 100 Mpc h^{-1} box, it can describe galaxies down to H I masses of 10^7 $\mathrm{M}_\odot$ h^{-1}. MII is based on a WMAP1 cosmological model (Spergel et al. 2003) with $\Omega_{\rm m} = 0.25$, $\Omega_{\rm b} = 0.045$, $h = 0.73$, and $\sigma_8 = 0.9$. For consistency, our PINOCCHIO light-cone assumes the same cosmology.

[1] Spinelli et al., *in preparation.*

For each available GAEA snapshot relevant for our purposes, we measure the $M_{\rm H\,I}$ as a function of $M_{\rm halo}$ and model it using the $M_{\rm H\,I}$–$M_{\rm halo}$ relation:

$$M_{\rm H\,I}(M_{\rm halo}) = M_{\rm halo}\left[a_1\left(\frac{M_{\rm halo}}{10^{10}}\right)^{\beta} {\rm e}^{-\left(\frac{M_{\rm halo}}{M_{\rm break}}\right)^{\alpha}} + a_2\right] {\rm e}^{-\left(\frac{M_{\rm min}}{M_{\rm halo}}\right)^{0.5}}, \tag{2}$$

where $a_1, \beta, \alpha, M_{\rm break}, a_2$, and $M_{\rm min}$ are free parameters (Spinelli et al. 2020). We construct a Gaussian likelihood for these parameters and, assuming large flat priors, we reconstruct their posteriors through the MULTINEST sampler (Feroz & Hobson 2008; Feroz, Hobson & Bridges 2009) using an MPI-enabled PYTHON wrapper (Zwart, Price & Bernardi 2016). We thus obtain a trend in redshift for each of the $M_{\rm H\,I}(M_{\rm halo})$ parameters that we interpolate with a spline. A similar procedure is followed for the scatter of the $M_{\rm H\,I}$–$M_{\rm halo}$ relation. In this way, we have a prescription to populate haloes with H I at each needed redshift that we use for the full light-cone.

Since the H I signal will be measured in redshift space, we use the plane-parallel approximation to displace the real-space halo positions using their peculiar velocities.

We construct a HEALPix map with $N_{\rm side} = 512$ for each of the 512 frequencies of interest binning the redshift space positions of the halo centres in slices of $\Delta\nu$ (see Table 2). Given the volume of each such defined portion of the light-cone and its total $M_{\rm H\,I}$ mass, one can compute the H I density $\rho_{\rm H\,I}$ and estimate the brightness temperature fluctuation in each pixel $\hat{\mathbf{n}}$ (Mao et al. 2012):

$$\delta T_{\rm H\,I}(\nu, \hat{\mathbf{n}}) = \overline{\delta T_{\rm H\,I}}(z)\left[\frac{\rho_{\rm H\,I}(\hat{\mathbf{n}})}{\overline{\rho_{\rm H\,I}}(z)}\right]. \tag{3}$$

The mean H I brightness temperature at a given redshift z can be computed following Furlanetto, Oh & Briggs (2006)

$$\overline{\delta T_{\rm H\,I}}(z) = 23.88 x_{\rm H\,I}\left(\frac{\Omega_{\rm b}h^2}{0.02}\right)\sqrt{\frac{0.15}{\Omega_{\rm m}h^2}\frac{(1+z)}{10}}\ {\rm mK}, \tag{4}$$

where $x_{\rm H\,I} \equiv \Omega_{\rm H\,I}/\Omega_{\rm H}$ is the fraction of neutral atomic hydrogen and $\Omega_{\rm H\,I}(z) = 8\pi G\overline{\rho_{\rm H\,I}}(z)/(3H_0^2)$. The highest frequencies considered correspond to a very local universe and, in this case, the virial radius of the most massive haloes can be comparable to the size of a voxel. To avoid such spurious overdensities, when in this regime, we do not assign all the H I mass to the halo centre, but we distribute the H I mass according to a NFW profile (Navarro, Frenk & White 1996), thus spreading the H I to neighbouring voxels.

2.2 Foreground models

The dominant foregrounds present between 950 and 1400 MHz are diffuse synchrotron emission, diffuse free–free emission, and extragalactic point sources. In this work, we explore two established models of IM foregrounds: a Gaussian realization of the components based on Santos et al. (2005), referred to as $\rm MS_{05}$ in this work, and a PSM-based simulation (Delabrouille et al. 2013), referred to as PSM.

2.2.1 $\rm MS_{05}$

Santos et al. (2005) constructed Gaussian realizations of extragalactic and diffuse Galactic emissions to investigate their effect on the extraction of cosmological information from 21-cm IM data. The angular power spectrum for each foreground component takes the form

$$C_\ell(\nu_i, \nu_j) = A\left(\frac{1000}{\ell}\right)^{\beta}\left(\frac{\nu_{\rm ref}^2}{\nu_i\nu_j}\right)^{\alpha} I_\ell^{ij}, \tag{5}$$

where the reference frequency used is $\nu_{\rm ref} = 130$ MHz, A is the power spectrum amplitude, β controls the angular scaling, and α is the spectral index across the frequency range. Each of the foreground components is parametrized with a different set of $\{A, \beta, \alpha\}$ values, reported in Table 1. The term I_ℓ^{ij} encodes the frequency coherency of the foreground and is expected to be unity for complete correlation. Santos et al. (2005) considered departures from the complete correlation adding a decorrelation term. However, for simplicity, and to keep this as the most idealized of foreground models, we choose to omit this and assume $I_\ell^{ij} = 1$.

2.2.2 PSM

With the aim of testing cleaning on realistic foreground contamination, we take advantage of the FFP10 (Full Focal Plane) all-sky simulations from the *Planck legacy archive*.[2] We use the high frequency versions of the FFP10 simulations as these are available at $N_{\rm side} = 2048$, allowing us to downgrade the maps to our desired $N_{\rm side} = 512$. The FFP10 simulations take their foreground contributions from the PSM (Delabrouille et al. 2013), which in turn uses empirical data sets to inform its estimates. While these models only hold true under specific assumptions, which will be discussed, it is worthwhile to include foreground simulations that are (1) not Gaussian in nature and (2) have the possibility of being correlated with each other. We outline the main features of this set of foregrounds; for more details, we refer the reader to Carucci et al. (2020), where this model was first assembled for the frequencies of interest.

Galactic synchrotron emission: We use the FFP10 synchrotron simulation at 217 GHz, based on the source-subtracted and destriped version of the Haslam 408 MHz map (Remazeilles et al. 2015), and scale it across frequencies using the synchrotron spectral index map of Miville-Deschênes et al. (2008). The Haslam 408 MHz map is assumed to contain negligible amounts of Galactic free–free emission at high Galactic latitudes. The synchrotron spectral index map used has been formed from 408 MHz and 23 GHz data and so may in fact be slightly steeper than the true synchrotron spectral indices at MHz frequencies. For our study, however, we only require spatially varying spectral indices within the physically expected range for synchrotron emission. The spectral index map is at a lower resolution than our intended simulation resolution. We add in detail below the resolution threshold of 5 deg of the spectral index map using the following Gaussian realization:

$$C_\ell = A\left(\frac{1000}{\ell}\right)^{2.4}, \tag{6}$$

where the amplitude (A) is set using the angular power spectrum of the 5 deg spectral index map.

Galactic free–free emission: We scale the FFP10 free–free simulation at 217 GHz, based on the all-sky Hα template of Finkbeiner (2003), down to our MHz frequency range using a spatially constant spectral index of -2.1. It should be noted that free–free emission is the least dominant foreground component across our frequency and Galactic latitude range.

Extragalactic point sources: This contribution is the only component not taken from the FFP10 simulations; for this, we use the

[2] https://wiki.cosmos.esa.int/planck-legacy-archive/

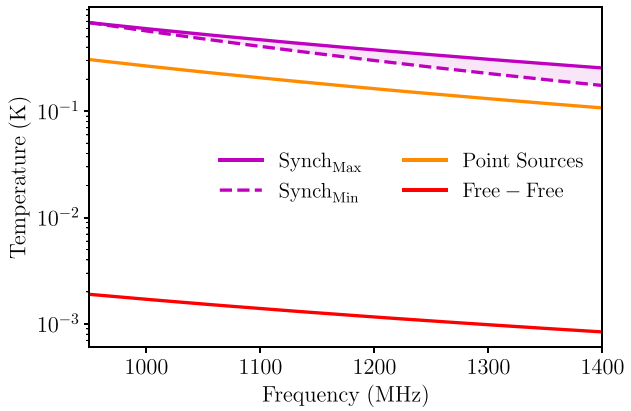

Figure 2. The spectral forms of each PSM emission component normalized using the mean emission temperature within the sky region investigated by this work; synchrotron emission displayed as a shaded area between the steepest and shallowest spectral index present in the simulation.

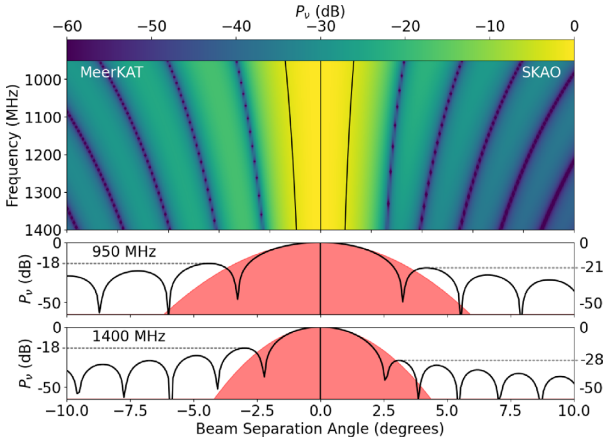

Figure 3. Simulated beam models for MeerKAT (left) and SKAO-MID (right) dishes. The top panels show the change in the beam pattern with frequency out to a beam separation angle of 10 deg. The black lines in the top panel show the FWHM of each beam model. The *centre* and *bottom* panels show cuts through the MeerKAT (*left*) and SKA-MID (*right*) beam patterns at 950 and 1400 MHz (i.e. the top and bottom of the simulated band), respectively. The red shaded region shows the equivalent Gaussian beam of each instrument. The *dashed grey* lines mark the amplitude of the first side lobe for the Airy beam model used for each instrument.

prescription outlined in Olivari et al. (2018), which expands on the empirical 1.4 GHz source count model of Battye et al. (2013). The model requires three selection criteria: (1) the cut-off flux, i.e. the value above which we assume that point sources are bright enough to be identified and removed (e.g. Wang et al. 2010; Matshawule et al. 2021); (2) the average point source spectral index; and (3) the distribution of this spectral index across the map. We use a flux cut-off of 0.1 Jy and a Gaussian distribution for the source spectral index centred at -2.7 with $\sigma = 0.2$.

The spectral forms of all components of our PSM-based foreground simulations are shown in Fig. 2.

2.3 Telescope simulation

2.3.1 Beam models

We aim to test how well component separation methods work in the presence of a beam model that includes not just the main lobe but also a side-lobe structure that changes with frequency. For these simulations, we are considering the dishes used in the MeerKAT array and the SKAO-MID array. For the MeerKAT dishes, we assume unobstructed 13.5 m apertures, and 15 m unobstructed apertures with underilluminated primaries to reduce the side-lobe amplitude for the SKAO-MID dishes. We do not model the final SKAO-MID survey that will include observations from both 13.5 and 15 m dishes (integrating the MeerKAT dishes); instead, we focus on two separate surveys with different dish properties to analyse the effect of these characteristics distinctly.

We generate the beam models for both dish types using modified Airy beam functions that allow for Gaussian tapered aperture distributions defined as (Wilson, Rohlfs & Huttemeister 2009)

$$E(\rho_\nu) = \mathrm{e}^{-0.5(\rho_\nu/\sigma_\rho)^2}, \tag{7}$$

where σ_ρ defines the width of a Gaussian taper and ρ_ν is the number wavelengths across the dish at a given frequency defined as

$$\rho_\nu = \frac{D\nu}{2c}, \tag{8}$$

where D is the dish diameter (either 13.5 or 15 m), ν is the observing frequency, and c is the speed of light. For the MeerKAT dishes, we find that side-lobe structure is best represented by setting the Gaussian taper width in equation (7) to be $\sigma_\rho = \infty$, which describes a dish that is being uniformly illuminated (i.e. the MeerKAT beam model is represented by an Airy beam). For the SKAO-MID dishes, we expect the larger dishes to be underilluminated to improve the side-lobe response, and we adopt a value of $\sigma_\rho = 20$.

To generate the beam pattern at each frequency, we integrate over the aperture distribution frequency for each beam separation angle (θ) as

$$B(\nu, \theta) = \left| \frac{\int E(\rho_\nu) j_0(\rho_\nu \sin(\theta)) \rho_\nu \mathrm{d}\rho_\nu}{\int E(\rho_\nu) \rho_\nu \mathrm{d}\rho_\nu} \right|, \tag{9}$$

where j_0 is the zeroth-order Bessel function. The resulting beam patterns for the MeerKAT and SKAO-MID dishes are shown side by side in Fig. 3 for the frequency range 950–1400 MHz, and beam separation angle out to 40 deg from the main lobe. The upper panel in Fig. 3 shows how the beam models evolve with frequency. The marked black lines in the upper panel show how the full width at half-maximum (FWHM) of each beam model changes with frequency, changing by just 0.44 and 0.35 deg FWHM for the MeerKAT and SKAO-MID dish models, respectively. The lower panel shows a slice of the beam model at 1175 MHz. Here, we can see that the first side lobe in the MeerKAT model (-18 dB) is 4 times larger than the first SKAO-MID side lobe at the same frequency (-24 dB). For the MeerKAT model, the first side-lobe response is close to constant, while the SKAO-MID first side lobe changes by a factor of 5 from -21 to -28 dB across the band. In the rest of the text and figures, we will refer to these beam models as Airy beam models.

We also produce a Gaussian beam model for each dish model that is representative of the main beam response of the telescope. We define these more approximate beam models as

$$G_\nu(\theta) = \mathrm{e}^{-4\log(2)(\theta/\theta_{\rm FWHM})^2}, \tag{10}$$

where the $\theta_{\rm FWHM}$ evolves with frequency as it is forced to match the measured FWHM of the Airy beam models.

Finally, we convolve the sky models described in Section 2.2 with each beam model. The convolution is performed by transforming the map and the beam model into the spherical harmonic domain. For

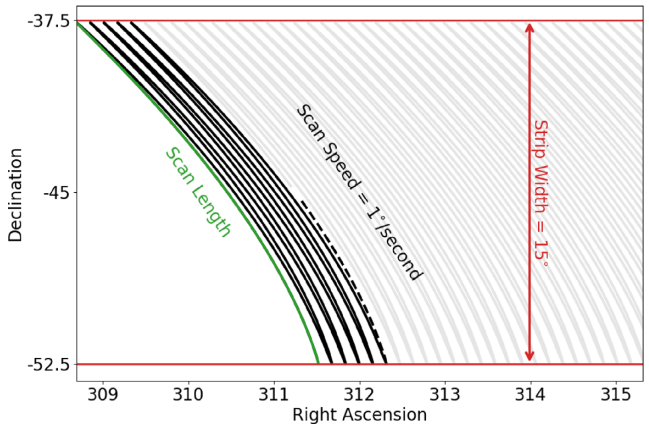

Figure 4. Example of the constant elevation scanning strategy used to map out the strip centred on −45 deg declination. For visualization purposes, we highlight the first scan in this example in *green*, which defines a single scan length, the *black* lines represent the area already scanned, while the *faint black* lines represent the upcoming scans. The *red* lines mark the declination boundaries of the strip, which has a width of 15°. Each strip has the same width and scanning speed.

a radially symmetric beam model, the spherical harmonic transform of the beam pattern is defined as

$$B_\ell(\nu) = 2\pi \int B(\nu,\theta) P_\ell(\cos(\theta)) \sin(\theta) \mathrm{d}\theta \,, \tag{11}$$

where P_ℓ are Legendre polynomials.

2.3.2 *Observing strategy*

For our simulations, we use a simulated observatory to create a mock IM data set that is closely representative of both the proposed SKAO-MID Band 2 survey set out in SKA Cosmology SWG (2020) and the ongoing MeerKLASS survey (Santos et al. 2017; Pourtsidou 2018; Wang et al. 2021). The purpose of the simulations is to create inhomogeneities in the noise distribution around the map and create a patch shape that is representative of realistic observations. Both the realistic noise distribution and patch shape of the simulations will enable us to test the component separation methods on a quasi-realistic data set.

The simulated survey scanning strategy used constant elevation azimuth scans to map out four strips in declination. Fig. 4 shows an example of how the strip centred on −45 deg declination was mapped out. We observed each strip at half of the maximum elevation as seen from the centre of the SKAO-MID array, observing each strip both when it was rising and setting. The observation time for the simulated survey is approximately 40 h per dish, equating to a total observing time of approximately 5000 h for the SKAO-MID and 2500 h for the MeerKAT array. The total sky area mapped is approximately 5000 deg^2 spanning between -52.5 deg $< \delta <$ 7.5 deg in declination, and each strip is 70 deg long centred at 0 deg in right ascension. The choice of patch location was made to match preliminary H I IM observations from MeerKAT (Wang et al. 2021), and also to have minimal galactic foreground contributions. However, we are aware that due to strong satellite RFI any real ground-based survey would not choose to observer near $\delta \sim 0$ deg (e.g. Harper & Dickinson 2018); however, this is not an issue here as we are not including RFI within the simulation and the exact declination of the patch will not significantly change the results.

The fixed elevation azimuth scanning strategy was simulated using a simple sine-wave model of the telescope motion described as

$$A = \frac{\Delta A}{2} \sin(2\pi t/T) + A_0 \,, \tag{12}$$

where A is the telescope azimuth, A_0 is the central azimuth corresponding to the declination of each strip, T is the time to complete a single scan defined as $T = \Delta A / v_{\rm scan}$, where $v_{\rm scan}$ is the scan speed of the telescope, and ΔA is the scan length that is dependent on the strip width, the scanning speed, and the elevation that we calculate numerically for each scan. The choice of a sine function to model the telescope azimuth motion as opposed to a triangular waveform was to also include the effect of the telescope turnaround time. The elevation is modelled as a constant value for each strip and ranges between 25 and 30 deg. For a summary of the simulation parameters, see Table 2.

2.3.3 *Noise model*

For both the SKAO-MID and MeerKAT receiver noise models, we assume the noise to be Gaussian and white. The noise per pixel is calculated by

$$\sigma = \frac{T_{\rm sys}}{\sqrt{N_{\rm dish}\tau\,\Delta\nu}} \,, \tag{13}$$

where τ is the integration time in seconds per pixel that is defined by the observing strategy described in Section 2.3.2, $\Delta\nu$ is the bandwidth of each frequency channel, $N_{\rm dish}$ is the number of dishes in the array, and $T_{\rm sys}$ is the system temperature.[3]

We define the system temperature for both receiver types as

$$T_{\rm sys}(\hat{\mathbf{n}}) = T_{\rm rx} + T_{\rm CMB} + T_{\rm spill} + T_{\rm sky}(\hat{\mathbf{n}}) \,, \tag{14}$$

where $T_{\rm CMB} = 2.73$ K is the CMB monopole contribution, $T_{\rm spill} = 3$ K is the approximate contribution to spillover, $T_{\rm sky}(\hat{\mathbf{n}})$ is the brightness of the sky along line-of-sight $\hat{\mathbf{n}}$, and $T_{\rm rx}$ is the receiver temperature.[4] For the receiver temperature of the SKAO-MID dishes, we used the band 2 receiver temperature prediction given in SKA Cosmology SWG (2020) that gives $T_{\rm rx}^{\rm SKAO} = 7.5$ K. For MeerKAT, we use the mean of the measured receiver temperature response defined as (Braun et al. 2019)

$$T_{\rm rx}^{\rm MeerKAT} = \langle 7.5 + 6.8|\nu_{\rm GHz} - 1.65|^{1.5} \rangle \ {\rm K} \,, \tag{15}$$

which gives $T_{\rm rx}^{\rm MeerKAT} = 9.8$ K.

2.4 Final combined data product

Our final data sets for the Blind Challenge are composed of the cosmological H I (Section 2.1) added to the foreground model (either $\rm MS_{05}$ or PSM – Section 2.2). The maps are then processed by the telescope simulation outlined in Section 2.3, which emulates the effects from the particular beam. We add in some instrumental noise specified by the type of telescope and the scanning strategy. The procedure is schematically summarized in Fig. 5.

Since our PSM model is based on empirical data, it inherently includes a zero-point (or monopole) signal. On the other hand, for

[3]Equation (13) is strictly for a single polarization receiver. In principle, two polarizations would be available but the resulting factor 2 in the equation is within our uncertainty in the total system temperature budget. We have thus ignored it.

[4]For these simulations, we do not include any atmospheric contribution but it is expected to be only a few K at 1 GHz (Bigot-Sazy et al. 2015).

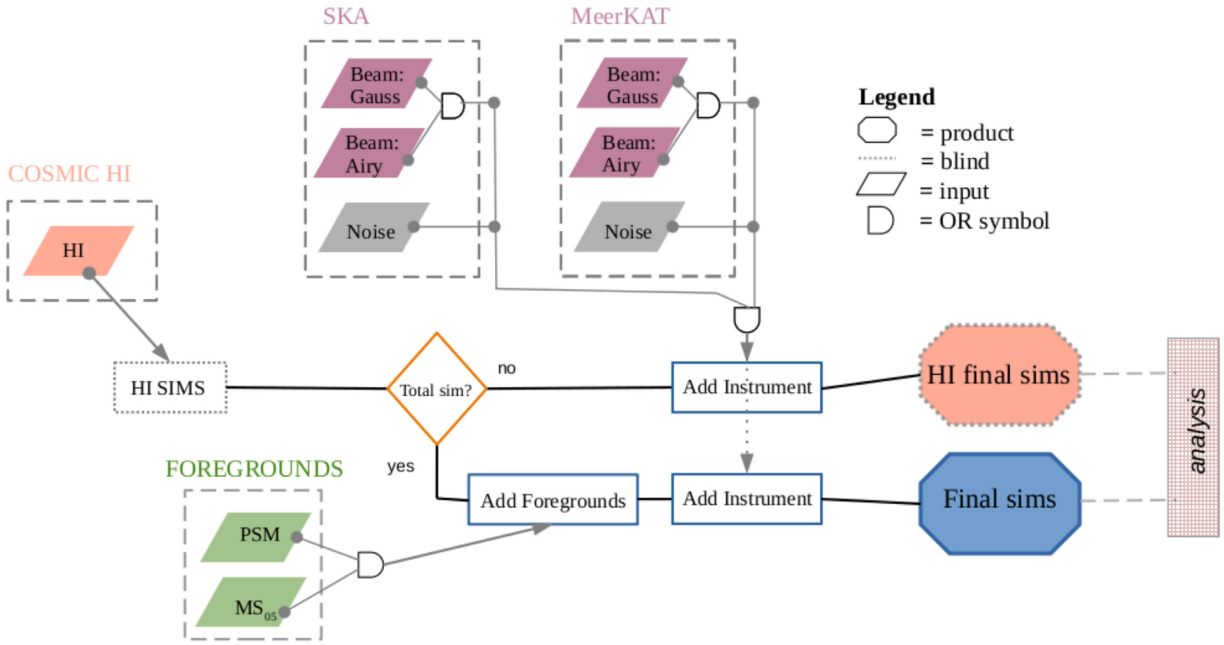

Figure 5. Flowchart of the construction of the simulations described in Section 2, highlighting the various options for foreground modelling and instrumental effects. We recall that the H I level is not known to the participants of the challenge (dotted frames indicate the blind components).

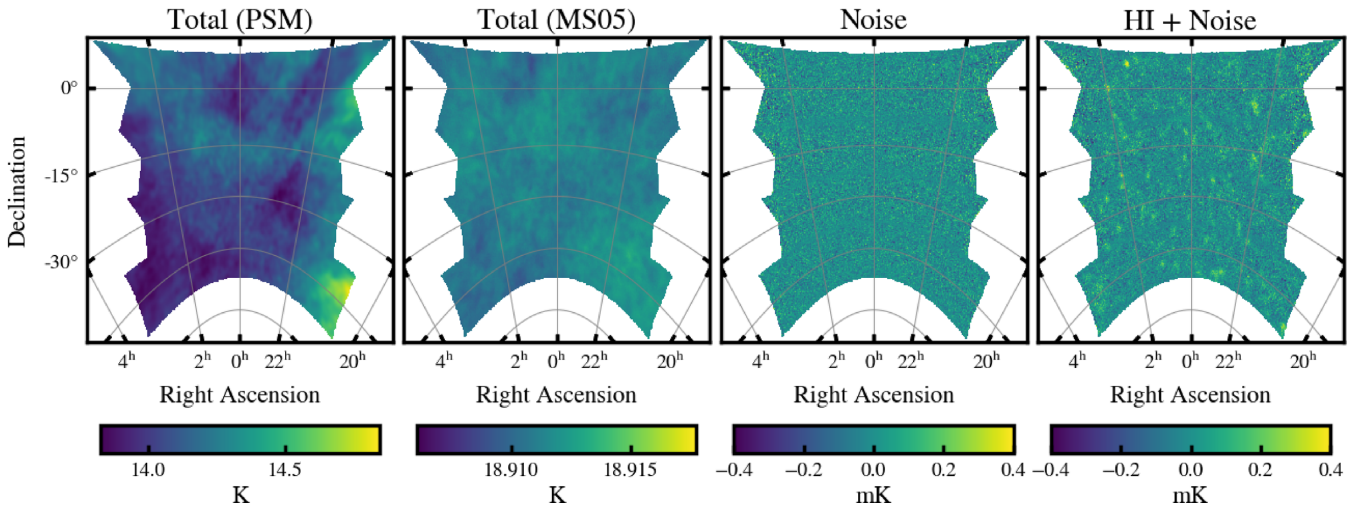

Figure 6. Maps of the contributing components to the final simulated signal at 1000 MHz for the SKAO-MID Airy beam case. The first two panels show the two different foreground simulations we implement: the PSM (Delabrouille et al. 2013) and the Gaussian realizations presented in Santos et al. (2005) (MS_{05}). The last two panels show the simulated thermal noise alone and with the cosmological H I signal that we aim to recover, i.e. **R** in equation (16). The noise in these examples is generated using the PSM as the input for T_{sky} in equation (14).

the MS_{05} model we have Gaussian realizations with mean zero amplitude. To balance this effect, we add an artificial monopole to the MS_{05} model, which is given by the T_{CMB} and T_{sky} components in equation (14). For the latter, the offset is derived by roughly scaling the mean value of the sky at 408 MHz (Wehus et al. 2017) with the expected synchrotron spectral index at low frequencies (e.g. Platania et al. 1998, 2003). While this results in some differences between the MS_{05} and PSM models for the monopole amplitude, it is not overly important for our investigation since the monopole is used mostly to fix the total system temperature at each frequency and most of the foreground cleaning methods are not concerned with the monopole level.

Fig. 6 shows the final simulated maps of the different components in our combined data product for a frequency of 1000 MHz. The first two panels show the two different foreground models, PSM and MS_{05}, respectively. There is more spatial structure in the PSM model as expected, whereas the MS_{05} is uniformly Gaussian distributed. The third panel shows the noise, in this example for the SKAO-MID instrument. Close inspection reveals subtle horizontal stripes of lower noise due to our scanning strategy (i.e. regions observed more often have lower noise). The final panel shows the addition of the cosmological H I signal: The noise floor is quite high and dominates the small scales, yet we can notice by eye the large-scale features of the H I.

Table 3. Summary of the nine foreground cleaning pipelines used in this work. See Section 3 for details.

Method	Assumption on foreground components	Pipeline	Brief description and references
Principal component analysis	Statistically uncorrelated	PCA(a)	Classical PCA with no weighting (see Cunnington et al. 2021a)
		PCA(b)	*fg_rm* code (Alonso et al. 2015), with inverse *rms* weighting
		PCAwls	Classical PCA applied on the wavelet-transformed data
Independent component analysis	Non-Gaussian	FASTICA(a)	Based on `Scikit-learn` package
		FASTICA(b)	*fg_rm* code (Alonso et al. 2015)
Generalized morphological component analysis	Sparse in a given domain and morphologically diverse	GMCA	Sparsity enforced in the wavelet domain (see Carucci et al. 2020)
		mixGMCA	PCA on the coarse scale + GMCA on small scales
Polynomial fitting	Smooth in frequency	poLOG	In log–log space (*fg_rm* code; Alonso et al. 2015)
Parametric fitting	Assumptions on spectral indices	LSQ	Fit to individual foregrounds

3 FOREGROUND SUBTRACTION METHODS

The observed temperature maps defined in equation (1) can be represented by two-dimensional (frequency and pixel) data cubes, $\mathbf{X} \equiv T_{\rm obs}(\nu, \hat{\mathbf{n}})$. Most of the cleaning algorithms we use assume that we can linearly decompose the matrix $\mathbf{X}$ in a set of $N_{\rm fg}$ sources in pixel space $\mathbf{S}$ modulated in frequency through a mixing matrix $\mathbf{A}$ plus some residuals $\mathbf{R}$ that should in principle contain most of the cosmological signal that we aim to recover together with the white instrumental noise:

$$\mathbf{X} = \mathbf{AS} + \mathbf{R}\,. \tag{16}$$

In practice, we do not expect the above decomposition to hold perfectly for a data cube, as leakage between the frequency-correlated and -uncorrelated parts is unavoidable. The foreground cleaning process boils down to solving equation (16) to find $\mathbf{R}$, once the number of sources (foregrounds) is set to $N_{\rm fg}$. More explicitly, using equation (1), for each frequency channel i and pixel n_p, we relate the cleaned residual $\mathbf{R}$ to the input signal through

$$R_{ip} = \int {\rm d}\Omega \left[B(\nu, n_p, \Omega) T_{\rm H\,I}(\nu_i, n_p)\right] + T_{\rm noise}(\nu_i, n_p)\,. \tag{17}$$

The assumptions to be made in order to find the matrix $\mathbf{A}$ and components $\mathbf{S}$ that satisfy equation (16) vary from method to method.

The methods used in this challenge are summarized in Table 3; we describe them in more detail in the following sections.

3.1 PCA

PCA can be used to identify an estimate for the mixing matrix $\mathbf{A}$, the columns of which will be given by the first $N_{\rm fg}$ principal components. The principal components are essentially the eigenvectors of the mean-centred data $\nu\nu'$ covariance matrix $\mathbf{C}$, given by

$$C_{ij} = \frac{1}{N_{\hat{n}}} \sum_{p=1}^{N_{\hat{n}}} w_i \Delta T(\nu_i, n_p)\, w_j \Delta T(\nu_j, n_p)\,, \tag{18}$$

where $\Delta T(\nu_i, n_p) = T(\nu_i, n_p) - \bar{T}(\nu_i)$ and the summation is over all $N_{\hat{n}}$ pixels. The w factors provide an optional map weighting. The eigendecomposition is given by $\mathbf{CV} = \mathbf{V}\Lambda$, where Λ is the diagonal matrix of N_ν eigenvalues. The first $N_{\rm fg}$ columns from the eigenvector matrix $\mathbf{V}$ represent the entries for the mixing matrix. Computing the covariance matrix is useful since the magnitudes of its eigenvalues Λ offer some guidance on how many principal components to include, i.e. the choice of $N_{\rm fg}$ (see, for example, Fig. 22 that we will comment in the discussion in Section 7). In brief, since we know that foregrounds have undoubtedly higher amplitude and higher variance than the cosmological signal, we expect them to be well characterized by the first few principal eigenvalues and eigenvectors.

As summarized in Table 3, in this Challenge we use three PCA implementations: PCA(a), PCA(b), and PCAwls. PCA(a) uses a straightforward implementation of the process described in this section, with no weighting ($w_i = w_j = 1$), replicating the pipeline used in Cunnington et al. (2021a). PCA(b) uses the publicly available code *fg_rm*[5] (see Alonso et al. 2015) with the implemented inverse noise weighting; i.e. we use the root mean square (rms) of the map at each frequency:

$$w_i = \frac{1}{\sigma_i}\,, \quad \sigma_i = \sqrt{\frac{1}{N_{\hat{n}}} \sum_{p=1}^{N_{\hat{n}}} \Delta T(\nu_i, n_p)^2}\,, \tag{19}$$

designed to minimize the influence of noise on the identification of dominant foreground modes. Lastly, PCAwls is an implementation of PCA on wavelet-transformed data with no weights and will be described in Section 3.3.

3.2 FASTICA

Fast Independent Component Analysis (FASTICA) is a widely used method developed in Hyvärinen (1999) and employed for foreground cleaning on simulated H I data (Chapman et al. 2012; Wolz et al. 2014; Cunnington et al. 2019; Carucci et al. 2020) as well as real data (Wolz et al. 2017, 2021; Hothi et al. 2021).

FASTICA estimates the mixing matrix $\mathbf{A}$ by assuming that the sources are statistically independent of each other. The method therefore aims to maximize statistical independence that can be assessed using the central limit theorem, which states that the greater the number of independent variables in a distribution, the more Gaussian that distribution will be (i.e. the probability density function of several independent variables is always more Gaussian than that of a single variable). Hence, by maximizing any statistical quantity that measures non-Gaussianity, we can identify statistical independence.

Before assessing non-Gaussianity, FASTICA begins by mean-centring the data, and then carries out a *whitening* step that aims to achieve a covariance matrix equal to the identity matrix for these whitened data (i.e. the components will be uncorrelated and their variances normalized to unity). Since this whitening step can be achieved with a PCA analysis, FASTICA is essentially an extension

[5] https://github.com/damonge/fg_rm

of PCA, and hence in most cases in the context of foreground cleaning, will provide very similar results.

For maximizing non-Gaussianity, an approximation of the negentropy is used. In the context of 21-cm foreground cleaning, the approximation of negentropy uses a set of optimally chosen non-quadratic functions that are applied to the data and averaged over for all available pixels. The maximization of negentropy by averaging over angular pixels means that for purely Gaussian sources, FASTICA will be unable to improve upon the initial PCA step carried out in the whitening step due to Gaussian sources having an equivalent zero negentropy. This explains the similarity in results often found between PCA and FASTICA when most of the simulated components are Gaussian fields (Alonso et al. 2015; Cunnington et al. 2021a).

As summarized in Table 3, in this Challenge we use two FASTICA implementations: FASTICA(a) and FASTICA(b). The FASTICA(a) pipeline uses the FASTICA module in `Scikit-learn`[6] (Pedregosa et al. 2011). FASTICA(b) uses the public *fg_rm* code (Alonso et al. 2015). Despite the fact that the two implementations use different codes to apply the same FASTICA methodology, their differences lie on pre-processing choices of input data and the choice of the number of modes to remove (see Fig. 9).

3.3 GMCA and wavelet decomposition

GMCA is a blind component separation method based on sparsity (Bobin et al. 2007). It assumes that the $N_{\rm fg}$ foreground components verify two hypotheses: they are sparse in a given transformed domain (i.e. most samples are zero-valued) and their supports are disjoint; in other words, the foreground components are *morphologically* diverse (i.e. their non-zero samples appear at different locations). GMCA has been successfully applied in various astrophysical contexts [e.g. cosmic microwave background data (Bobin et al. 2013, 2014), high-redshift 21-cm interferometry (Chapman et al. 2013; Patil et al. 2017), X-ray images of supernova remnants (Picquenot et al. 2019), and gravitational waves (Blelly, Moutarde & Bobin 2020)].

Carucci et al. (2020) showed the wavelet domain to be optimal to sparsely describe foregrounds and contaminants in the low-z H I IM context. First, we project the data $\mathbf{X}$ on to wavelet space. The GMCA algorithm aims at minimizing the following cost function:

$$\min_{\mathbf{A},\mathbf{S}} \sum_{i=1}^{N_{\rm fg}} \lambda_i \, ||S_i||_1 + \left|\left|\mathbf{X} - \mathbf{AS}\right|\right|_2 , \tag{20}$$

where the first term is the ℓ_1 norm, i.e. $\sum_{j,k}|S_{j,k}|$: This constitutes a constraint for sparsity, mediated by the regularization coefficients λ_i. The second term is a usual data-fidelity ℓ_2 norm term. We find solutions for $\mathbf{A}$ and $\mathbf{S}$ by iterating a projected alternate least-squares procedure: We fix $\mathbf{A}$ and perform a least-squares update to determine $\mathbf{S}$, we compute the thresholds λ_i via mean absolute deviation of S_i, we update $\mathbf{A}$ with $\mathbf{S}$ fixed, and so on. The key point is the thresholding: It allows us to keep the samples with the highest amplitudes, which are the most informative to retrieve the mixing matrix $\mathbf{A}$ (i.e. they most likely belong to the foreground components and are the least likely to be contaminated by the cosmological signal and noise), and it provides robustness in terms of convergence since the thresholds decrease with the progressive iterations.

[6]https://scikit-learn.org/

In the Challenge described in this work, we decided to test three different cleaning methods based on wavelet decomposition and GMCA.

(i) *PCAwls*: We perform a PCA decomposition as described in Section 3.1 on the wavelet-transformed data. We expect it to be equivalent to PCA in standard pixel space as the PCA algorithm does not depend on the domain in which data are described. The purpose of using PCAwls has been to add an extra set of solutions with the PCA method, i.e. a different participant using a different pipeline and choosing a different number of components $N_{\rm fg}$ to remove (see later Fig. 9 for a summary of the $N_{\rm fg}$ choices).

(ii) *GMCA*: We apply GMCA as it is described above and by Carucci et al. (2020).

(iii) *mixGMCA*: We apply PCA on the largest scale of the wavelet-transformed data and GMCA on the remaining scales. By largest scale, we mean the coarse approximation of the maps resulting from the initial low-pass filtering of the wavelet decomposition (Starck, Murtagh & Fadili 2010). We assemble the two solutions back together before re-transforming the maps into pixel space. This allows us to have two different mixing matrices $\mathbf{A}$ and two different numbers $N_{\rm fg}$ of components for the small and large spatial scales of maps.

Ongoing work on optimizing the GMCA method in the H I IM context resulted in the development of mixGMCA, which we use here for the first time in the literature. Carucci et al. (2020) highlighted the need of having a different number of components for different spatial scales, and Cunnington et al. (2021a) highlighted how, in the IM context, the sparse assumption might not suit the largest scales, yet it holds well in the small ones. Analysis of LOFAR observations also supports the idea of having $N_{\rm fg}$ dependent on scale (Hothi et al. 2021). The wavelet decomposition offers a straightforward framework for analysing multiscale data. With mixGMCA, we further developed this idea by allowing different mixing matrices to describe the data cube at different scales.

3.4 Logarithmic polynomial fitting

One of the first approaches to foreground subtraction methods is to come up with a base of smooth functions in frequency that we can then use to model the foregrounds. This has been extensively used (Wang et al. 2006, 2013; Ghosh et al. 2011a; Ansari et al. 2012), and here we follow the approach of Alonso et al. (2015) and perform a power-law base expansion in log–log space. In particular, we will use polynomials of the logarithm of the frequency, i.e.

$$\log T_{\rm fg}(\nu, \hat{\mathbf{n}}) = \sum_{n=1}^{N_{\rm fg}} \alpha_n(\hat{\mathbf{n}})[\log(\nu)]^{n-1} . \tag{21}$$

We then solve the log–log space equation equivalent to equation (16). For this purpose, we used the code *fg_rm* (Alonso et al. 2015) with the frequency logarithm polynomials, and we also weight the data using the rms (see equation 19) translated into logarithmic space, $\sigma_{i,\log T} \simeq \sigma_i/T$. In this Challenge, we set $N_{\rm fg} = 6$. We refer to this pipeline as poLOG.

3.5 Parametric fitting

Parametric methods, unlike blind component separation, assume that a considerable portion of the measured total signal is well known due to prior empirical knowledge. Specifically, we could make the following assumptions:

(i) Diffuse synchrotron and free–free emission are non-negligible at MHz frequencies, with synchrotron emission dominating at high Galactic latitudes, as indicated by the numerous ground-based surveys collated for use by the Global Sky Model (Zheng et al. 2017).

(ii) Diffuse free–free emission has a spatially constant spectral index that can also be considered spectrally constant over our 500 MHz frequency range (Bennett et al. 1992).

(iii) Free–free emission, synchrotron emission, and the extragalactic point source contributions are heavily degenerate with each other due to their similar spectral forms (power laws with similar indices as in Fig. 2) (Planck Collaboration XXV 2016).

Here, we only fit for the diffuse emissions: free–free and synchrotron. We attempt to use the foreground degeneracy to our advantage by trialling the assumption that the extragalactic temperature contribution will be absorbed into either our estimate of Galactic synchrotron emission or our estimate of Galactic free–free emission or both.

For our parametric fit, we require the zero-level at each frequency map to be set solely by the diffuse foreground emissions we intend to fit: No additional temperature contributions can be present. Zero-level contributions can include (1) the CMB monopole, which is both spatially constant and constant across frequency and hence easy to subtract; (2) the receiver temperature, which we subtract under the assumption that this component can be measured by each experiment (e.g. Wang et al. 2021); and (3) the average temperature of all the unresolved extragalactic point sources. Regarding the latter contribution, values for these averages at various frequencies are available in the literature (e.g. Gervasi et al. 2008; Mauch et al. 2020); hence, we decide to subtract the true value for this average (i.e. the fiducial value used in our simulation) from the total temperatures at each frequency before beginning our fit.

We aim to determine the true $\mathbf{A}$ in equation (16) for the combination of free–free and synchrotron emission; for this, we require both the synchrotron and free–free emission spectral index per pixel. For free–free emission, we use the true (i.e. the fiducial value used in our simulation) value of -2.1 at each map pixel.

We find that it is optimum to first obtain the synchrotron spectral index from the total temperature data assuming that the free–free contribution is negligible. This works in practice by performing a least-squares fit at each pixel using the PYTHON module `lmfit` with two free parameters: the amplitude and spectral index of synchrotron emission. The parameter space of the synchrotron spectral index is restricted to within 10 per cent of the total temperature spectral index across the first three frequencies. We weight our fit using the FFP10 free–free emission map smoothed to 1.5 deg and scaled to each frequency as an estimate for noise. Having fitted for the synchrotron spectral index at each pixel $\beta_{\rm sy}(\hat{\mathbf{n}})$ our mixing matrix estimate can then be expressed as

$$\tilde{\mathbf{A}} = \begin{pmatrix} (\nu/\nu_0)^{\beta_{\rm sy}(\hat{\mathbf{n}})} \\ (\nu/\nu_0)^{-2.1} \end{pmatrix} . \tag{22}$$

The matrix of emission amplitudes ($\mathbf{S}$) is computed by again minimizing the standard least-squares problem:

$$\mathbf{S} = (\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})^{-1} \tilde{\mathbf{A}}^T \mathbf{X} , \tag{23}$$

where $\mathbf{X}$ are the total temperature data. Any components of the total data that can be characterized by a power law with spectral indices similar to the range of the indices within our mixing matrix estimate will be grouped together as foregrounds. We present the residual between the total data and our estimated combined foregrounds at each frequency as an estimate for H I emission plus noise. We refer to this pipeline as LSQ.

4 SUMMARY STATISTICS

The analysis of the simulations and the quality assessment for the residual maps after cleaning require estimators to compress the three-dimensional information contained in the data cubes. Although some studies have started to explore observational effects in higher order statistics (Cunnington, Watkinson & Pourtsidou 2021b; Jolicoeur et al. 2021), in this work, we focus on two-point summary statistics, looking at both the angular and line-of-sight directions, keeping the two separated to distinguish features that could show up independently in each direction. In particular, we compute the angular power spectrum as a function of frequency $C_\ell(\nu)$ (Section 4.1) and the one-dimensional line-of-sight power spectrum $P_{\rm los}(k_\nu)$ (Section 4.2). These choices relieve us from making extra assumptions (e.g. flat-sky approximation and thin-channel assumption to translate observed frequencies into distances). The same summary statistics are computed for the residual maps and the input signal plus noise, allowing a straightforward quantitative estimation of the performance of the various methods. We acknowledge that, by comparing the statistics, we cannot properly discriminate between a true reconstructed signal or a contribution of leaked foregrounds with a resulting power spectrum similar to the input signal. To this end, other strategies – although with different caveats – could be used, such as the cross-correlation of the residuals maps with the input signal and, for cleaning methods that involve the construction of a mixing matrix, the estimation of the leakage through appropriate projections of such matrix (e.g. Carucci et al. 2020). The direct comparison of summary statistics offers a simple and efficient way to test all different cleaning methods; moreover, the autospectra of the recovered maps represent the final product of observations before the cosmological analysis. Therefore, in this work, we rely on these statistics and their comparison with the input counterparts.

For future work, where we plan to assess the cosmological content of the cleaned maps, a proper error estimation of the reconstructed two-point statistics and covariance analysis will be crucial. In this analysis, we have roughly estimated the uncertainties on these statistics using both jackknife and theoretical errors and found that prominent features of the various methods persist even considering these estimated uncertainties. This implies that enough meaningful comparison of the cleaning methods can be achieved even without the errors, and we thus postpone a detailed analysis of uncertainties to a follow-up project.

4.1 Angular power spectrum

At a given frequency, the simulated sky patch has been constructed as a HEALPix map and can be decomposed in spherical harmonics. For the full sky case, the angular power spectrum can be estimated from the spherical harmonic coefficient $a_{\ell m}(\nu)$ of this decomposition,

$$\hat{C}_\ell(\nu) \equiv \frac{1}{2\ell+1} \sum_{m=-\ell}^{m=+\ell} |a_{\ell m}|^2 . \tag{24}$$

This estimator is no longer valid for sky patches, but can be corrected, in first approximation, by dividing by the sky fraction covered by the patch.

However, in the presence of sharp edges, such as the ones caused by the single-dish scanning strategy assumed here (see Fig. 6), the coupling induced by the mask can be important, and should

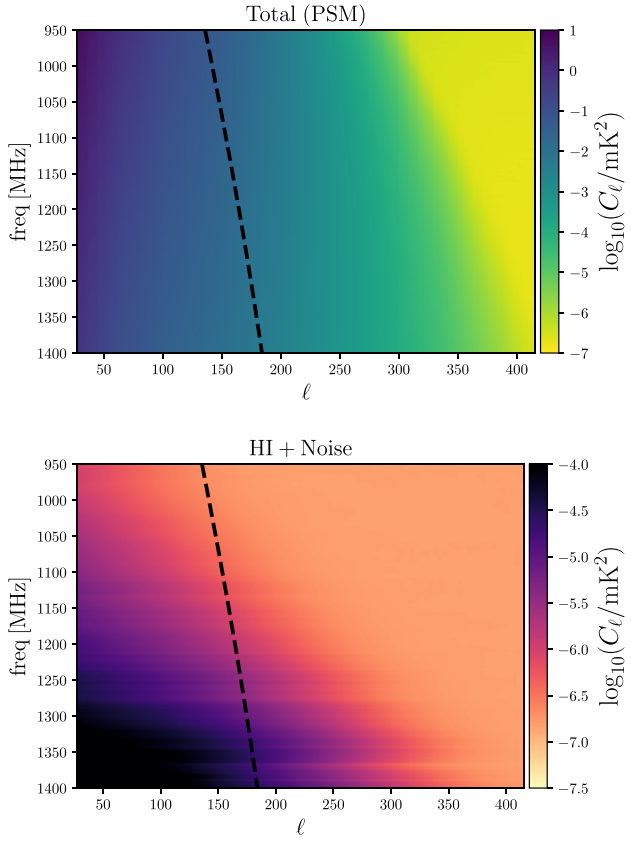

Figure 7. The angular power spectrum C_ℓ of the total sky emission considering the PSM foregrounds (upper panel) and the H I signal plus noise (lower panel) as a function of frequency, considering an SKAO-MID instrument and the Airy beam model. The black dashed line in both panels traces the evolution with frequency of the angular scale of the FWHM of the telescope beam.

be corrected for. One efficient and commonly used solution is the Monte Carlo Apodized Spherical Transform Estimator (MASTER; Hivon et al. 2002). In this work, we compute this correction using the NAMASTER software[7] (Alonso et al. 2019). Although NAMASTER has been optimized to deal with partial sky coverage, a complete validation of the result would require further studies and possibly a refinement of the final patch footprint. For the purpose of this work, whose main intent is to compare performances of different foreground methods, there is no such concern.

We present in the upper panel of Fig. 7 the evolution of the angular power spectrum as a function of frequency for one of the final data cubes, where the more realistic PSM foregrounds are considered. Dominated by the smooth foreground emission, the C_ℓ shows the effect of the beam suppressing signal at progressively larger scales going to lower frequencies (for reference, the black dashed line corresponds to the beam FWHM). As expected, the emission is stronger at lower frequencies. We show results for the SKAO-MID case, which we find similar to the MeerKAT case.

The lower panel of Fig. 7 shows the H I cosmic signal plus noise we are aiming to recover, orders of magnitude fainter than the astrophysical foreground emission. In this case, it is not only the beam effect that dictates the amplitude of the power spectrum, but also the interplay between the structured H I signal (whose intensity fades at lower frequencies) and the instrumental noise (that increases at lower frequencies). Indeed, at small scales, we can recognize the (quasi)scale-invariant noise floor at $\sim 10^{-7}\,\mathrm{mK}^{-2}$ covering the structure of the signal (differently at different channels), confirming what is shown in Fig. 6.

[7] https://github.com/LSSTDESC/NaMaster

4.2 Radial power spectrum

We compute the one-dimensional power spectrum directly in frequency space, $P_{\rm los}(k_\nu)$ with $k_\nu = 1/\nu$. It is the most straightforward choice to investigate how well the radial information is recovered (Alonso et al. 2015; Villaescusa-Navarro et al. 2017). Here, we follow the procedure described in Carucci et al. (2020). In short, for each pixel – i.e. line of sight – we Fourier transform the temperature along the frequency direction $\tilde{T} = F[T(\nu)]$, and we compute $P_{\rm los}(k_\nu)$ by averaging over the power spectra from each pixel n_p:

$$P_{\rm los}(k_\nu) = \Delta\nu \langle |\tilde{T}(k_\nu, n_p)|^2 \rangle_{N_{\hat{n}}} \,. \tag{25}$$

We expect the smooth foregrounds, which are strongly correlated in frequency, to display more power at small k_ν. We can see from Fig. 8 that this is indeed the case, for both the more realistic PSM foregrounds and for the MS_{05} model. The effect of the different instrumental response is very small when looking at the total sky signal in the first two panels, whereas we can clearly see the offset between the SKAO-MID and MeerKAT cases for the cosmic signal plus noise $P_{\rm los}$, caused by the different noise levels and beam models in the right-hand panel. The amplitude of $P_{\rm los}$ for the MS_{05} model is lower than the PSM one that is intrinsic to the MS_{05} model construction that simply adds a small mean-centred, Gaussian oscillation on top of $T_{\rm sys}$ (see Section 2.2.1).

5 THE BLIND CHALLENGE

In this section, we describe the procedure of the Blind Foreground Subtraction Challenge. This type of approach is increasingly adopted in cosmological studies (e.g. Kitching et al. 2013; Nishimichi et al. 2020) and is a useful and transparent test for the maturity of analysis pipelines (Prat et al. 2021). In this work, both the simulation of the H I signal and the details of the assembly of the components' maps (including beam convolution and addition of instrumental noise) have been kept *blind* to the participants that attempted the foreground cleaning.

The final data cubes, summarized in Section 2.4, can thus be effectively treated as mock observations. A common pre-cleaning processing is described in Section 5.1, while the details of the blind challenge procedure are presented in Section 5.2.

5.1 Common pre-processing

For a diffraction-limited antenna, the FWHM of the beam pattern is proportional to the dish size and the observing frequency, resulting in a variable resolution in frequency across the data cubes. Real data analyses have found it useful to counteract this effect by *resmoothing* the maps (Switzer et al. 2015; Wolz et al. 2021), i.e. by convolving them to a common FWHM (often 10–20 per cent lower than the one of the lowest frequency). To test the advantage of the resmoothing, we opt for two approaches: (1) cleaning the data cubes at the native channel-dependent resolutions; and (2) resmoothing all maps of the data cube to a common resolution. We thus created an extra set of resmoothed data cubes where all maps have been deconvolved to a Gaussian beam with FWHM equal to 1.05 times the FWHM of

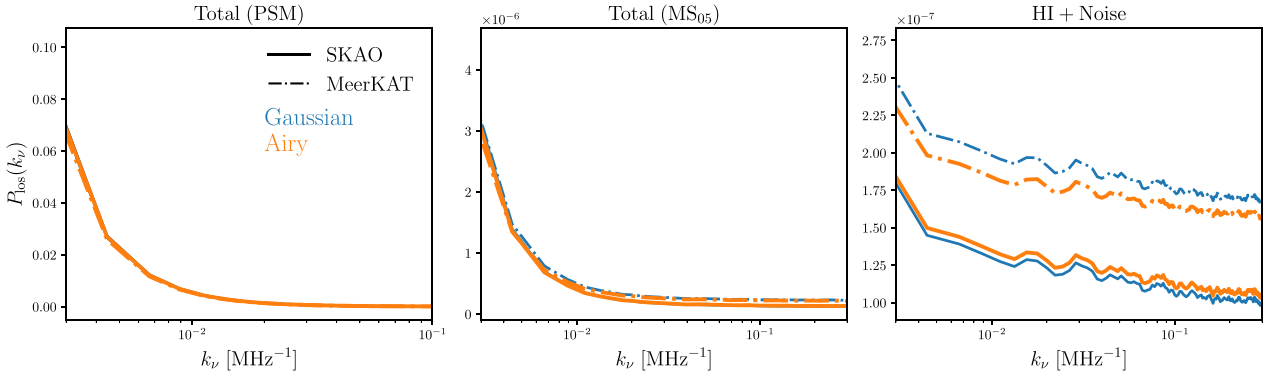

Figure 8. The line-of-sight power spectrum $P_{\rm los}$ for the total sky emission considering the PSM foregrounds (left) or the MS$_{05}$ model (centre), compared with the H I signal plus the noise (right), for both an SKAO-MID-like survey (solid lines) and a MeerKAT-like survey (dot–dashed lines). We show results after convolution with a Gaussian beam model (in blue) and the Airy beam model (in orange).

the lowest frequency channel. The resmoothing Gaussian kernel is defined as

$$a_{\ell m}^{\rm Res}(\nu) = a_{\ell m}^{Or}(\nu) e^{-(\ell(\ell+1))/(16\ln(2))\left[\theta_R^2 - \theta^2(\nu)\right]}, \quad (26)$$

where θ_R is the FWHM to convolve the data to, and $\theta(\nu)$ is the FWHM at frequency ν (see Section 2.3.1).

Because of the border effects of the Gaussian smoothing, we had to define a new (smaller) footprint, going roughly from a coverage of 11 per cent of the sky to 10 per cent. Moreover, because the SKAO-MID and MeerKAT beams are different, these new footprints are also (slightly) different for the two instrumental set-ups. To avoid the inclusion of a different footprint in the comparison of the results, the final footprint created for the resmoothed case has been used on the original data cubes too.

The combinations of two foreground models, two beam models, two instrumental set-ups, and frequency-dependent versus constant angular resolution, resulted in having a total of 16 different input data cubes to analyse.

5.2 Blind cleaning

The cleaning of the various data cubes has been performed with the nine pipelines summarized in Table 3. As discussed in Section 5.1, for each pipeline, 16 residual data cubes (expected to contain only the H I signal and the noise) have been submitted. Most of the used cleaning methods are blind source separation techniques (PCA, FASTICA, GMCA, and mixGMCA), where the only assumption is related to a statistical property of foregrounds (e.g. non-Gaussianity and sparsity); poLOG explicitly assumes frequency smoothness of the foreground emission, while LSQ tries to reconstruct known properties of their emission.[8]

For the blind methods, each participant was free to choose the number of components $N_{\rm fg}$ to subtract. The variety of choices made for the different data cubes by the various participants are presented in Fig. 9 and reported further in Table A1. We separate the SKAO-MID and the MeerKAT cases, and the type of data cubes, specifying the foreground and the beam models used and the original or resmoothed scenarios. Fig. 9 highlights the difficulty and subjectivity in choosing $N_{\rm fg}$, especially facing increasingly realistic sky mocks.

[8]Unlike the other methods, LSQ requires prior information on the map monopole and it could only be run on the PSM foreground model cases as it relies on the foreground spectral forms each being known well enough to be parametrized.

A summary diagram of the procedure is reported in Fig. 10, together with the subsequent steps for the analysis and comparison of the results, detailed in the next section.

6 RESULTS

In this section, we report the results of the Blind Challenge. We present a qualitative overview of the results for the original data cubes in Section 6.1 and for the resmoothed data cubes in Section 6.2. A comprehensive discussion on the relative performances of the various cleaning methods via quantitative metrics is in Section 6.3.

6.1 Original data cube

Gaussian beam: In the top panel of Fig. 11, we show the angular power spectrum C_ℓ for a given frequency (1225 MHz as an example) for the Gaussian beam case and focusing on the more realistic PSM foreground model. The reconstructed signal is consistent across pipelines, at least at large scales ($\ell \lesssim 250$), and comparable with the expected input signal. As the beam model starts suppressing the signal, small differences among methods are visible. The effect is slightly stronger for the MeerKAT case, where the noise level and beam suppression are higher.

In the lower panel of Fig. 11, we plot the line-of-sight power spectrum $P_{\rm los}$, again for the Gaussian beam case and PSM foreground model. At high and intermediate values of k_ν, the cleaned $P_{\rm los}$ show behaviours in good agreement with the true H I signal. At closer inspection, we can see that some of the methods tend to underestimate the signal's amplitude while others slightly overpredict it. At low k_ν, where most of the foreground power is, all blind methods show some level of overcleaning, as it is extremely difficult to separate foregrounds from the signal in this region. On the contrary, the LSQ method overpredicts the signal, probably due to the leakage of foreground emission, which is not well isolated and removed by the method, into the H I plus noise part.

Airy beam: The Airy beam model case shown in Fig. 12 presents a more complex scenario. At the angular power spectrum level (top panels), there is qualitative agreement among pipelines, except for the LSQ method going astray after $\ell \gtrsim 250$. Most notably, all cleaning methods consistently display a peak in $P_{\rm los}$ around $k_\nu \sim 0.045\,{\rm MHz}^{-1}$ (bottom panels). Matshawule et al. (2021) have identified and analysed a similar effect in their simulations. They attributed it to the presence of the 20 MHz oscillation in the beamwidth as a function of frequency, which is enforced in their standard

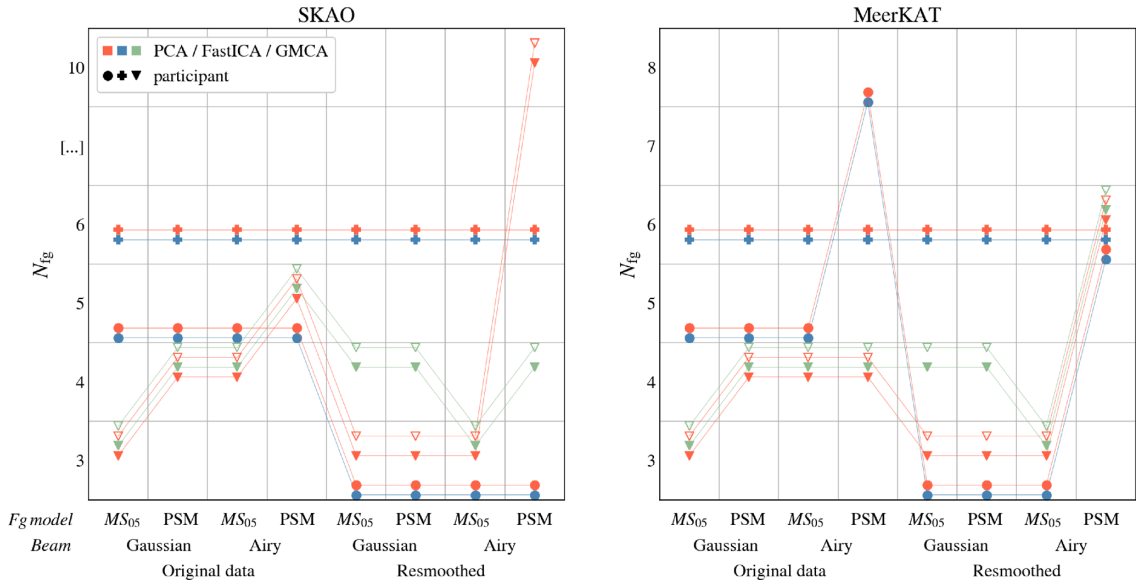

Figure 9. Graphical representation of N_{fg} used for the various cleaning methods as reported in Table A1, shown separately for the case of SKAO (left-hand panel) and MeerKAT (right). Each row represents the integer number N_{fg} used in each method, with the lines offset to facilitate reading. Each column refers to a specific data set as described by the legend below the *x*-axis. Different colours correspond to different cleaning methods and different symbols to the different participants who performed the cleaning (i.e. different pipelines too). In the case of mixGMCA, two values of N_{fg} need to be set: for a PCA run at the large scale and a GMCA run at small scales; we decompose the mixGMCA information into two PCA/GMCA cases that we highlight using empty symbols.

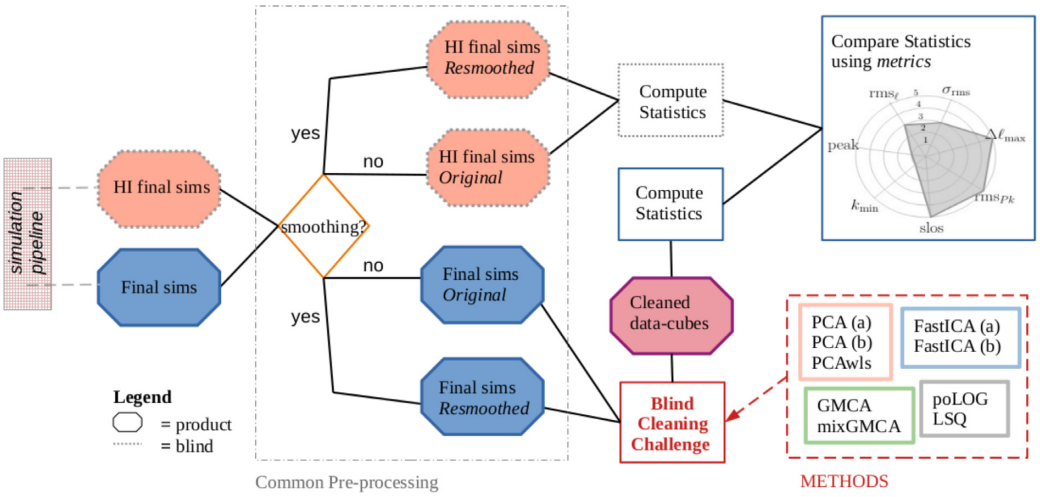

Figure 10. Flowchart of the analysis procedure. The input simulations (see Fig. 5) are pre-processed as described in Section 5.1; we then compute the summary statistics described in Section 4 for all data cubes and compare the maps recovered by different pipelines through the metrics defined in Section 6.3. Finally, results are compressed in the radar charts of Figs 20 and 21.

modelling of the FWHM of the main lobe in order to reproduce the holographic measurements of the MeerKAT beam by Asad et al. (2021). In our case, the feature is caused by the (oscillating) changing positions of the side lobes across the frequency band. We believe that the fact that both works find the oscillations at ∼20 MHz is a coincidence since both oscillations have different origins.

Beam and foreground structure interaction: In Fig. 13, we show the estimator $(P_{clean} - P_{true})/P_{true}$, where P_{clean} is the power spectrum of the residual maps for a specific cleaning method, while P_{true} is the input signal and noise. The *peak* feature in the P_{los} of the cleaned data completely disappears using the MS_{05} foreground model, i.e. when the foregrounds are Gaussian. It implies that the more realistic Airy beam alone is not the cause of this effect, but it is instead its combination with the more structured PSM foreground emissions. The latter finding agrees with Matshawule et al. (2021), and we qualitatively interpret it as follows. Since the sky temperature varies for different lines of sight, the frequency behaviour caused by the Airy beam gives rise to oscillations with slightly different amplitude as a function of direction. The different cleaning methods can spot the beam oscillations at the map level, but they tend to miss its exact amplitude in all lines of sight. If the sky is just a Gaussian realization of a foreground-like power spectrum, as for the MS_{05} model, these line-of-sight differences statistically cancel out, and no peak appears in the P_{los}. We verified that the above conclusion holds even if the MS_{05} model fluctuations are enhanced by two orders of magnitude. Indeed, running PCA cleaning on these artificial models with strong

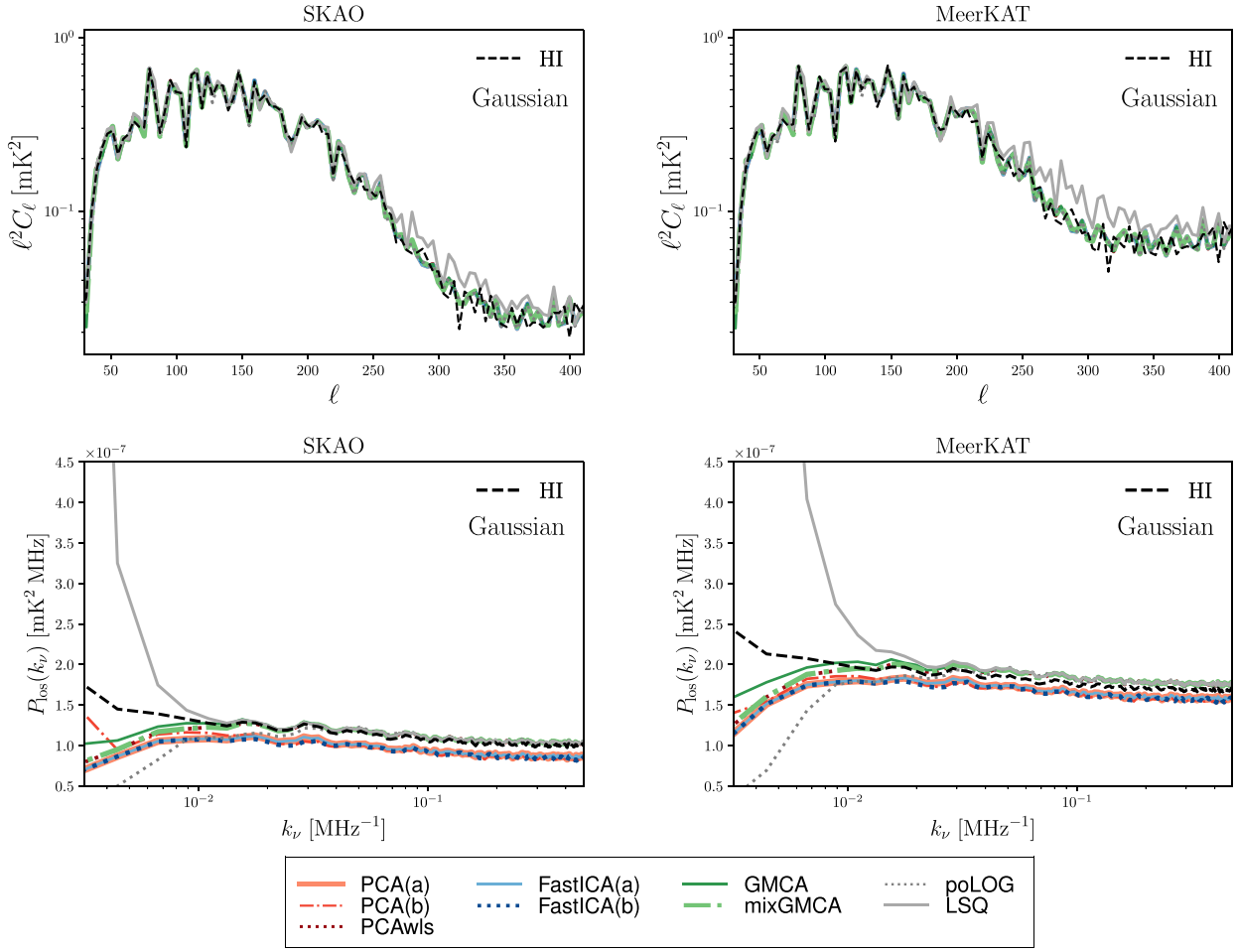

Figure 11. Angular power spectrum C_ℓ at 1225 MHz (top) and line-of-sight power spectrum $P_{\rm los}$ (bottom) of the residual maps for both an SKAO-like (left) survey or a MeerKAT-like (right) survey. We focus on the realistic foreground (PSM) case. The results are colour coded to distinguish the various cleaning methods. Similar techniques and/or different implementations of the same algorithm are grouped together: PCA in reds, FASTICA in blues, GMCA in greens, and non-blind method in grey. All panels show results when the adopted beam model is Gaussian. The true H I signal, convolved with the appropriate beam, is shown as a black dashed line.

Gaussian foreground fluctuations ($\rm MS_{05}$ x100), we still find no excess of power in the $P_{\rm los}$ at the scale corresponding to the oscillation in the beam side lobe.

On the contrary, due to the realistic sky structures of the PSM foreground model, there is no averaging effect and the $P_{\rm los}$ shows the clear excess at $k_\nu \sim 0.045\,\rm MHz^{-1}$. From Fig. 12, we see that the strongest peak feature in the $P_{\rm los}$ appears for the LSQ and poLOG pipelines, which have less line-of-sight freedom in adapting to the foregrounds. Indeed, even if the peak feature is observed in all methods, they experience it with different severity (see again bottom panels of Fig. 12). These considerations become important when one tries to mitigate the *peak* after the cleaning. For instance, if the contamination is limited to few channels one could flag and remove them from the analysis; on the contrary, artefacts affecting a larger k_ν range will be harder to handle.

We analyse in more detail the effect of the beam on the angular power spectrum. We plot in Fig. 14 the C_ℓ of the cleaned residuals as a function of frequency on the vertical axis. On the left-hand panels, we report the PCA(b) method, showing that the cleaning performs differently going from the Gaussian beam case (top) to the Airy beam model (bottom). For comparison, we also plot the C_ℓ of the residuals in the Airy beam case for the FASTICA(a) (top right panel) showing a similar effect to the PCA(b). The interaction of the Airy beam with the spatial structure of the PSM foregrounds results in an excess of power at small scales in the residuals that evolves with frequency. As for the peak in the $P_{\rm los}$, the effect in the C_ℓ is present only in the cleaned maps and not in the original H I convolved with the Airy beam (see the lower panel of Fig. 7). The poLOG method (lower right panel), which enforces smoothness by construction, is instead free of this small-scale frequency feature.

We present in Fig. 15 the angular power spectrum residuals for the GMCA method, where $C_\ell^{\rm clean}$ is the angular power spectrum of the cleaned maps (as in Fig. 14) and $C_\ell^{\rm true}$ is the one for the original foreground free H I plus noise, convolved with the same beam model. The reconstruction is easier in the presence of the simpler $\rm MS_{05}$ foreground model and we find that this conclusion generally holds for all the cleaning methods. As expected from the results of Fig. 13, the fringe pattern at small scales appears only for the combined presence of the Airy beam model and the PSM foreground model.

6.2 Resmoothed data cube

As introduced in Section 5.1, the foreground cleaning pipelines have been tested also on pre-processed data cubes that have been

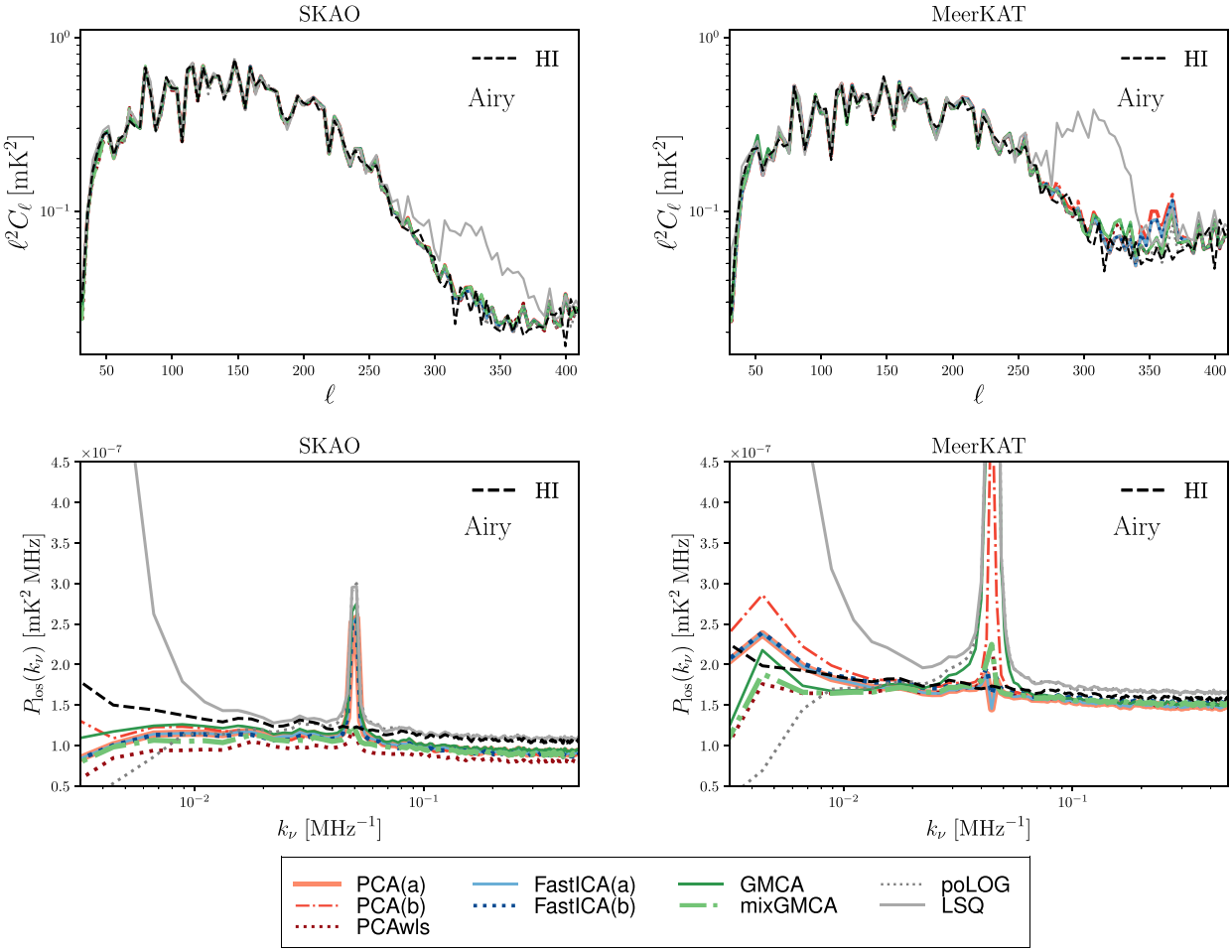

Figure 12. Same as Fig. 11 but for the Airy beam model. PCA(a), FASTICA(a), and FASTICA(b) often overlap.

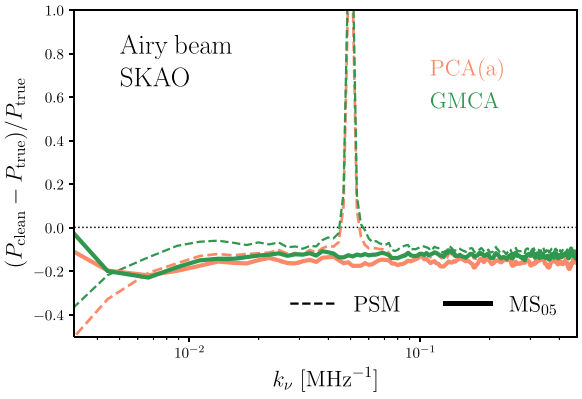

Figure 13. The different effect of the Airy beam model on the cleaning when the sky model is constructed with the MS$_{05}$ foreground (solid lines) or the PSM model (dashed lines). The line-of-sight power spectrum is shown for the SKAO case and for two different cleaning methods as an example: PCA(a) in orange and GMCA in green.

resmoothed with a Gaussian beam to a common lower resolution. Even if this practice has been generally adopted in single-dish experiments to reduce the impact of instrumental systematics contaminating the data (e.g. polarization leakage as in Switzer et al. 2013), in our particular (and more idealized) simulated setup, we instead generally conclude that a simple Gaussian resmoothing does not ease the blind source separation process (especially forcing a simple Gaussian resmoothing in the Airy beam scenarios), although it partially reduces residual foreground contamination for the LSQ method. We now discuss this point in more detail.

Fig. 16 compares residuals looking at the line-of-sight power spectrum for the SKAO case. We show again the estimator $(P_{\rm clean} - P_{\rm true})/P_{\rm true}$, where now $P_{\rm true}$ is the power spectrum of the resmoothed input signal and noise.

Top panels refer to the Gaussian beam and bottom to the Airy beam; on the right are the resmoothed cases. The Gaussian and Airy cases have been already shown in Figs 11 and 12 but the *relative* estimator allows a better quantification of the differences between the original and resmoothed scenarios. When maps have been resmoothed, we generally find more signal loss (i.e. a tendency to overclean) for the blind source separation methods, and a slightly larger k_ν interval affected by the peak feature in the Airy beam case. The parametric LSQ is an exception and we find that resmoothing helps the reconstruction of the signal. Indeed, the LSQ method performs power-law fits per pixel across frequency and so relies upon a single pixel to represent the same area of sky across the frequency range. Although not leading to signal loss, the resmoothing procedure does slightly enhance the few percent oscillatory pattern arising for the poLOG method, which is probably linked to the specific polynomial truncation.

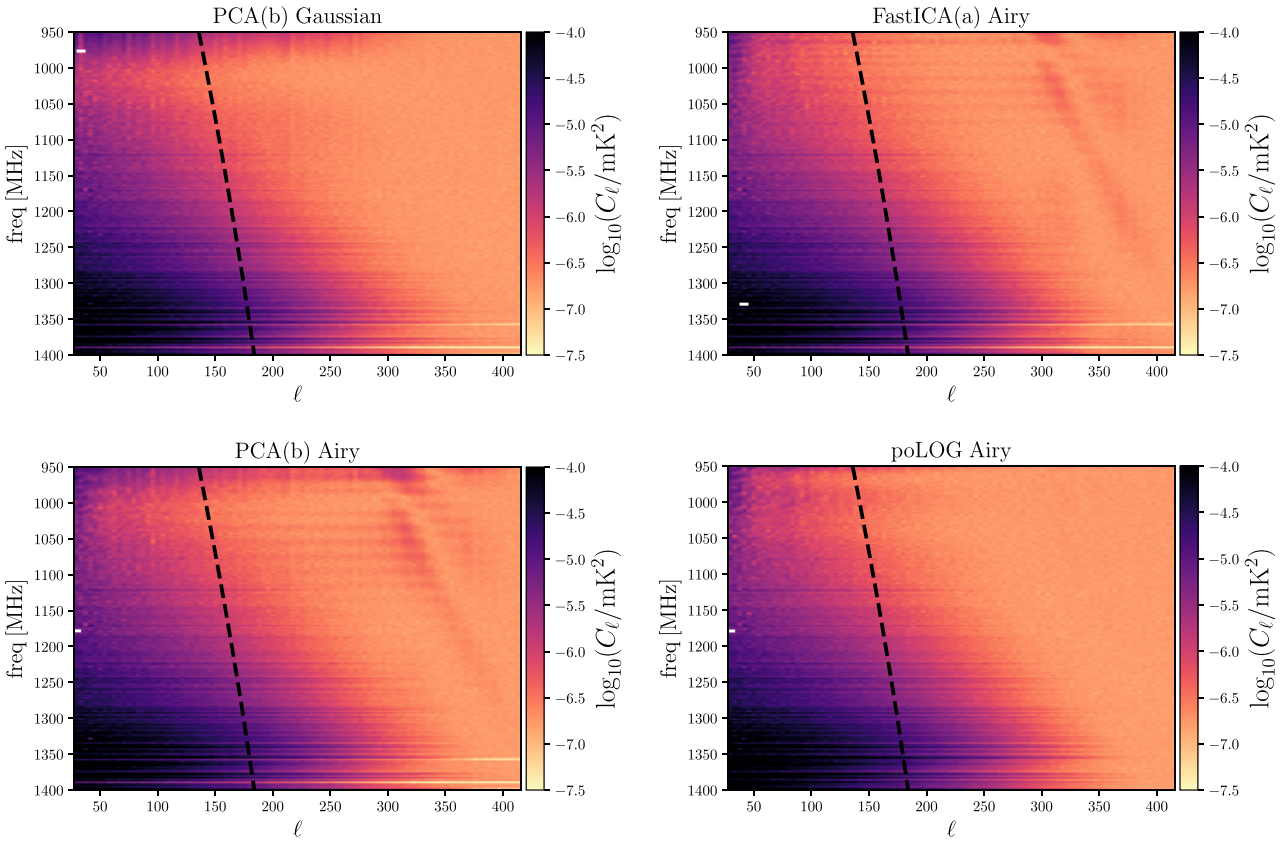

Figure 14. The angular power spectrum C_ℓ as a function of frequency of the residual maps (SKAO case) after cleaning with one of the implementations of the PCA method (left-hand panels). The more realistic case where the sky data cube has been convolved with the Airy beam model (lower left panel) presents a complex frequency behaviour that is not present in the Gaussian case (upper left panel). The right-hand panels show how this frequency feature, induced in the cleaning by the presence of the Airy beam, affects the cleaning by other methods. We show FASTICA(a) (upper right panel) that has a similar shape as most of the other methods, and the poLOG (lower right panel) that, enforcing smoothness by construction, does not display any frequency features in the angular power spectra. The black dashed line in all panels traces the evolution with frequency of the angular scale of the FWHM of the telescope beam.

We show the effect of resmoothing on the angular power spectrum in Fig. 17. We present results for mixGMCA while noting that all methods (excluded the particular LSQ case) behave similarly. Resmoothing the data cube seems to slightly improve the recovery of the C_ℓ for the Gaussian beam case, whereas, in the Airy beam case, it only enhances the fringe patterns in the $(C_\ell^{\rm clean} - C_\ell^{\rm true})/C_\ell^{\rm true}$ behaviour (at $\ell \gtrsim 250$).

We report also that, coherently with the line-of-sight power spectrum, resmoothing improves the reconstruction of the angular power spectrum for LSQ method in the Gaussian cases and slightly in the Airy beam case, as shown in Fig. 18.

Summarizing, we find that for all pipelines the cleaning becomes more difficult in the presence of a more realistic telescope beam. Our simple Gaussian resmoothing does not amend this challenge. In the transverse direction, cleaning methods struggle where the signal clustering is subdominant compared to the noise. In the radial direction, the intermediate range in k_ν is the less compromised; however, when the spatially structured PSM foregrounds are coupled to the Airy beam, a *peak* feature appears for almost all cleaned data cubes (see Fig. 13).

6.3 Quantitative comparison

In this section, we present a set of metrics to allow a *relative* comparison between the various cleaning methods in producing cleaned residual data cubes whose power spectra reproduce those of the true cosmological signal plus noise. A comparison in terms of preserved cosmological information is left for future work, whereas these power-spectrum-based metrics allow for an immediate and comprehensive view of the quality of the cleaning.

6.3.1 *Performance metrics*

Angular power spectrum: The estimator for the accuracy of the recovered angular power spectrum is defined as

$$\eta_C(\ell, \nu) \equiv \left((C_\ell^{\rm clean} - C_\ell^{\rm true})/C_\ell^{\rm true}\right)(\nu)$$

and varies substantially[9] across ℓ and frequency ν, as can be seen in e.g. Fig. 17. We characterize its overall behaviour with the following metrics:

1 rms_ℓ: To have a first estimate of the quality of the cleaning, we compute the root-mean-square (rms) value of $\eta_C(\ell, \nu)$ for every

[9]The irregularity of our footprint makes the estimation of the angular power spectra also quite noisy, although this is not an issue for the relative comparison among pipelines' results. We leave studies on patch optimization for future work.

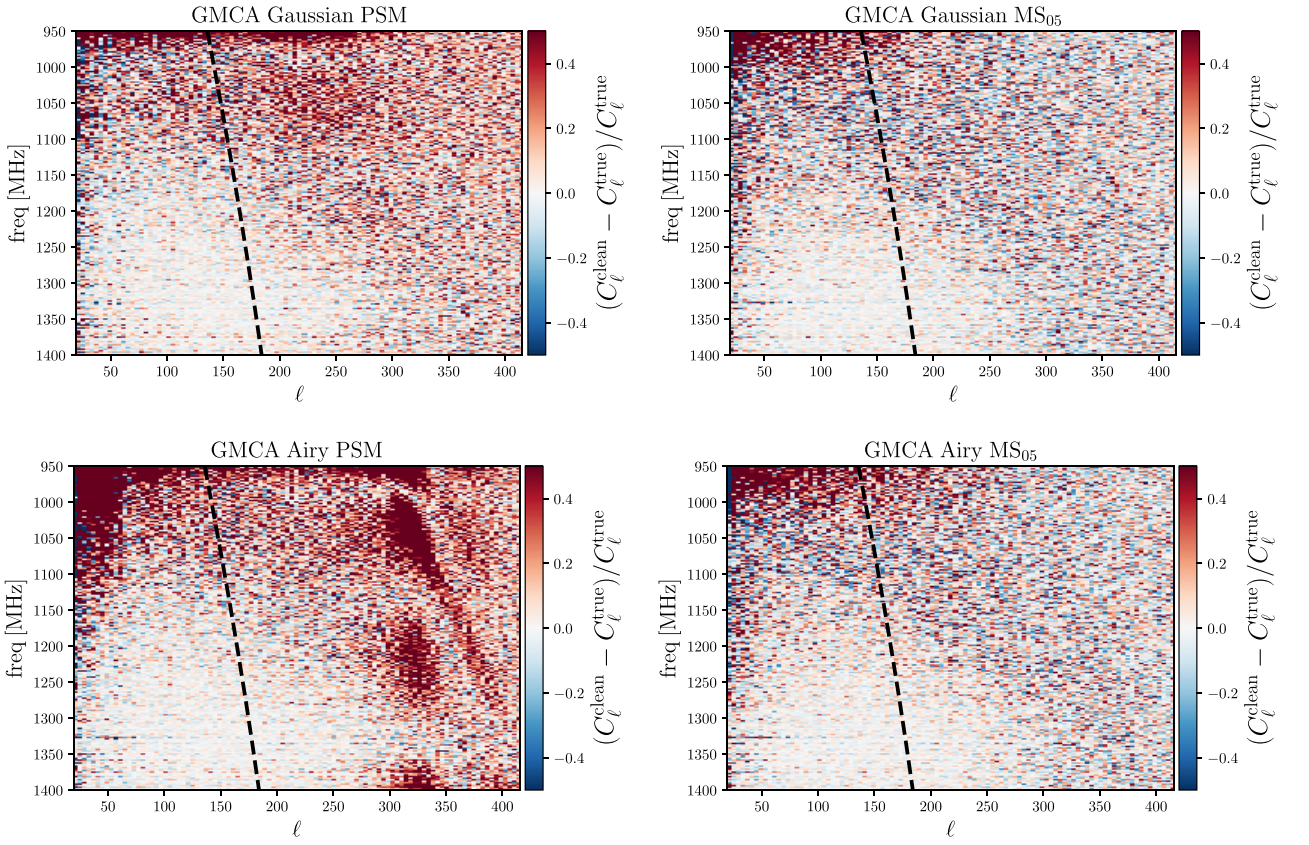

Figure 15. The estimator $(C_\ell^{\rm clean} - C_\ell^{\rm true})/C_\ell^{\rm true}$ for one of the cleaning methods (GMCA in this example), where $C_\ell^{\rm clean}$ is the angular power spectrum of the cleaned maps, while $C_\ell^{\rm true}$ is the angular power spectrum for the input signal and noise, as a function of frequency. We present results for the original data cubes. The more realistic PSM foreground model (left-hand panels) is compared with the $\rm MS_{05}$ foreground model (right-hand panels). We consider both the Gaussian and the Airy beam and focus on the SKAO case. Results are qualitatively similar in all cases, showing that cleaning is easier in the $\rm MS_{05}$ case and that in absence of foreground structure, there are no frequency features induced by the Airy beam. For reference, the black dashed line in all panels traces the evolution with frequency of the angular scale of the FWHM of the telescope beam.

frequency ν

$$\mathrm{rms}_{C\ell}(\nu) = \left(\frac{1}{(\ell_{\rm max} - \ell_{\rm min})} \sum_{\ell=\ell_{\rm min}}^{\ell_{\rm max}} \eta_C(\ell, \nu)^2 \right)^{1/2}, \tag{27}$$

and define rms_ℓ, its mean value across the $N_\nu = 512$ channels of our cleaned data cubes,

$$\mathrm{rms}_\ell = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \mathrm{rms}_{C\ell}(\nu_i)\,. \tag{28}$$

We exclude scales larger than $\ell_{\rm min} = 15$ and smaller than $\ell_{\rm max} = 500$ to reduce contamination from the mask and the noise, respectively. In general, the lower the value of rms_ℓ, the better the cleaning, with caveats that we try to track down with the next metrics.

2 $\sigma_{\rm rms}$: In order to capture the channel-to-channel variability of the rms value, we also compute its scatter across frequencies $\sigma_{\rm rms}$

$$\sigma_{\rm rms} = \left(\frac{1}{N_\nu} \sum_{i=1}^{N_\nu} (\mathrm{rms}_{C\ell}(\nu_i) - \mathrm{rms}_\ell)^2 \right)^{1/2}. \tag{29}$$

A smaller value of $\sigma_{\rm rms}$ indicates that a certain method is more consistent in the reconstruction across channels (although it could indicate a consistently biased reconstruction).

3 $\Delta\ell_{\rm max}$: A method that perfectly reconstructs large and intermediate scales while getting the noise floor wrong can have an rms higher than a method that is consistently biased at all ℓ. We thus opt for a third metric that quantifies the cumulative number of ℓ-bins across channels for which the agreement with the expected input signal is better than 30 per cent, i.e.

$$\Delta\ell_{\rm max} = \sum_{i=1}^{N_\nu} \left(\sum_{\ell=\ell_{\rm min}}^{\ell_{\rm max}} f_i(\ell) \right), \tag{30}$$

with

$$f_i(\ell) = \begin{cases} 1 & \text{if } |\eta_C(\ell, \nu_i)| < 30 \text{ per cent} \\ 0 & \text{else} \end{cases}. \tag{31}$$

Line-of-sight power spectrum: We now consider the radial direction and define

$$\eta_P(k) \equiv (P_{\rm clean}(k) - P_{\rm true}(k))/P_{\rm true}(k).$$

Its generic behaviour is more consistent among methods and overall less noisy than the one for the angular power spectrum, as we can see in e.g. Fig. 16, also due to the large number of pixels available in our patch. To characterize $\eta_P(k)$, we define the following metrics, also sketched in Fig. 19:

1) *rms*$_{Pk}$: As for the angular estimator, the first quantity to assess is the distance of the recovered signal from the input one, through the rms value of $\eta_P(k)$,

$$\mathrm{rms}_{Pk} = \left(\frac{1}{N_k} \sum_{i=1}^{N_k} \eta_P(k_i)^2 \right)^{1/2}, \tag{32}$$

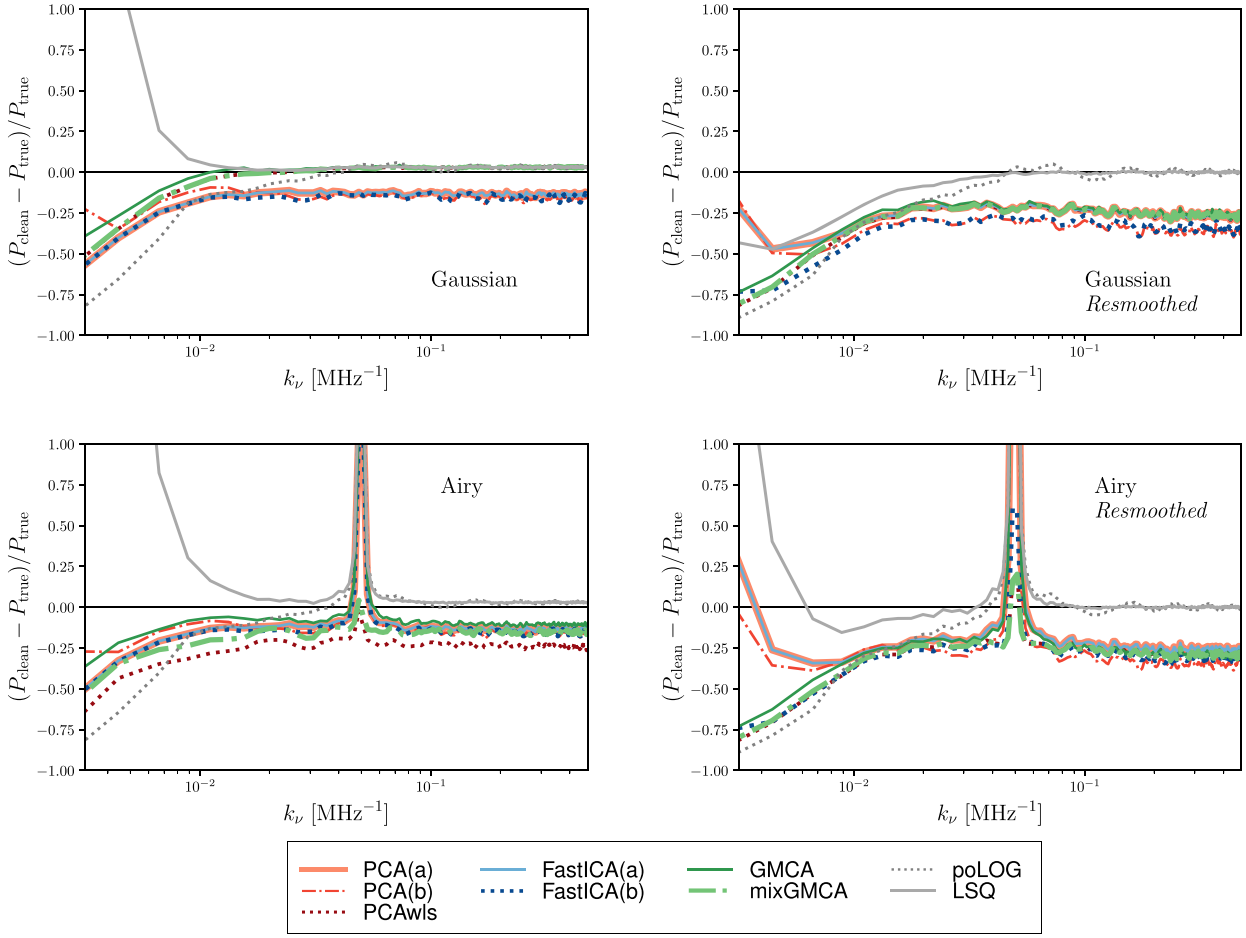

Figure 16. Comparison of the estimator $(P_{\rm clean} - P_{\rm true})/P_{\rm true}$, where $P_{\rm clean}$ is the power spectrum of the residual maps for the various cleaning methods, while $P_{\rm true}$ is the input signal and noise original or resmoothed depending on the specific case. The left-hand panels show the difference between the Gaussian and Airy beam model for the original data (as in the lower panels of Figs 11 and 12). The right-hand panels present the same cases but for the resmoothed data. The resmoothing procedure affects the line-of-sight power spectrum as the Airy beam *peak* broads and the reconstructed signal has an enhanced offset with respect to the true one. All panels refer to SKAO–PSM cases.

where N_k is the number of k-bins. The lower the rms, the more successful the cleaning.

2) *slos*: As noted already in Fig. 16, the estimator $\eta_P(k)$ shows a roughly constant bias at small scales for most cleaning methods. Due to overcleaning, this bias is often negative and an indicator for cosmological signal loss. We thus define slos as the mean value of the estimator for $k_\nu > 0.1$ MHz^{-1}. Despite the name, residual foregrounds in the cleaned maps may give rise to a positive value for slos. In general, the higher the absolute value for slos, the worse the cleaning performance.

3) $k_{\rm min}$: Due to their coherence in frequency, the foreground emission has power predominantly at small k_ν, making these scales the most difficult to recover. To characterize the extent of this contamination, we define $k_{\rm min}$ as the smallest k_ν at which the residual $P_{\rm los}$ starts deviating more than ± 30 per cent from the expected cosmological signal [i.e. $|\eta_P(k_{\rm min})| = 0.3$]. A smaller $k_{\rm min}$ indicates a more successful cleaning that extends on a larger range of scales.

4) *peak*: As already mentioned, in the Airy beam case coupled to the PSM foreground model, we observe a spiky feature in the $P_{\rm clean}$ and thus also in η_P (see Fig. 16 and Section 6.1 for a more thorough discussion). The height of this peak is proportional to the extent in k_ν range in which the spurious artefact appears, and we decide to include it in our set of metrics for the cleaning quality (the higher the peak value, the worse the cleaning performance).

6.3.2 *Method performance ranking*

We evaluate the metrics described above for all submitted residual data cubes; for each of the 16 set-ups and each of the 7 metrics, we have a distribution of values (one for each pipeline; see Tables B1 and B2). We mark each pipeline from 1 to 5 depending on their *relative* performance: The best method scores 5 and the worst method 1; the other marks are assigned binning the interval defined by the two extremes. The binning allows multiple methods to score the same value, including the two extreme ones.

To visualize the seven metrics together (three for the angular plus four for the line-of-sight power spectra), we compile a radar chart for each submission, where the area covered by the chart relates to the cleaning performance: the larger the area, the more accurate the cleaning. We focus on the more realistic PSM foreground model to draw conclusions and present the SKAO cases in Fig. 20 and the MeerKAT cases in Fig. 21. Each figure consists of four quadrants: The left column refers to the Gaussian beam cases, and the right to the Airy beam, with the corresponding resmoothed scenarios on the

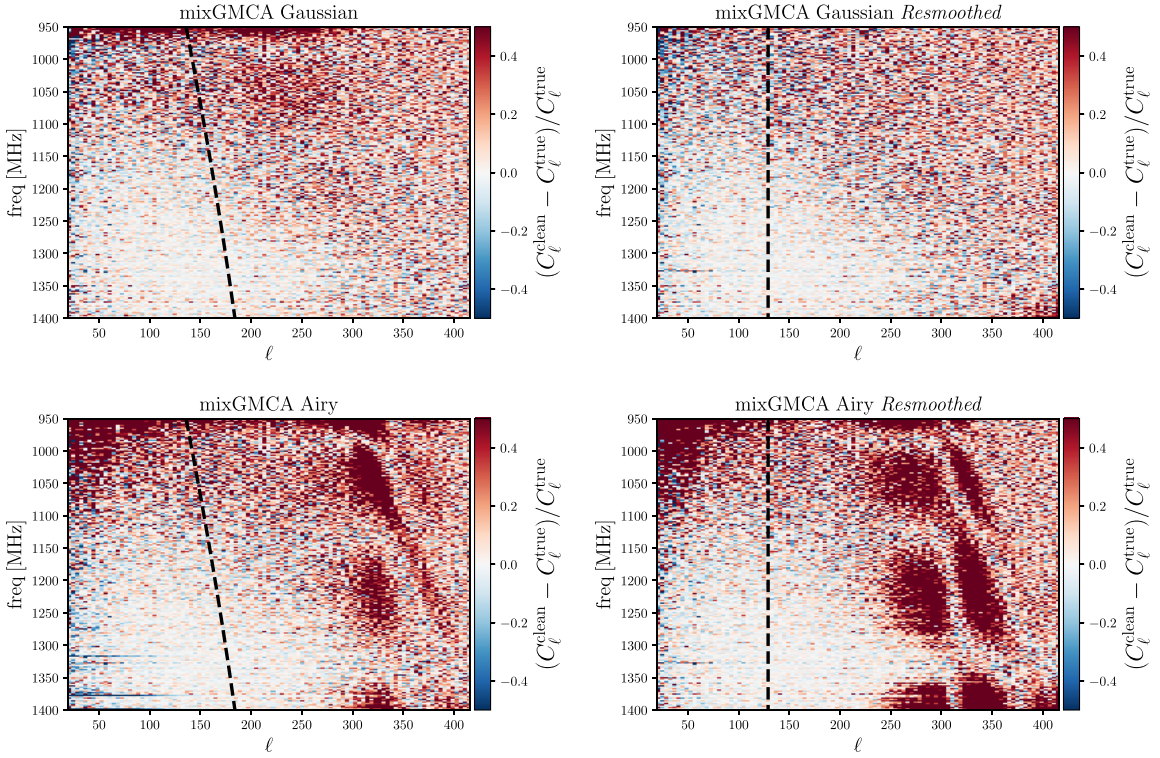

Figure 17. The estimator $(C_\ell^{\rm clean} - C_\ell^{\rm true})/C_\ell^{\rm true}$ for one of the cleaning methods (mixGMCA in this example) in the PSM and SKAO scenarios, where $C_\ell^{\rm clean}$ is the angular power spectrum of the cleaned maps, while $C_\ell^{\rm true}$ is the angular power spectrum for the input signal and noise, as a function of frequency. From top to bottom, the beam model changes from Gaussian to Airy, while from left to right is shown the effect of resmoothing. For reference, the black dashed line in all panels indicates the angular scale of the FWHM of the telescope beam. After resmoothing, this scale is constant with frequency.

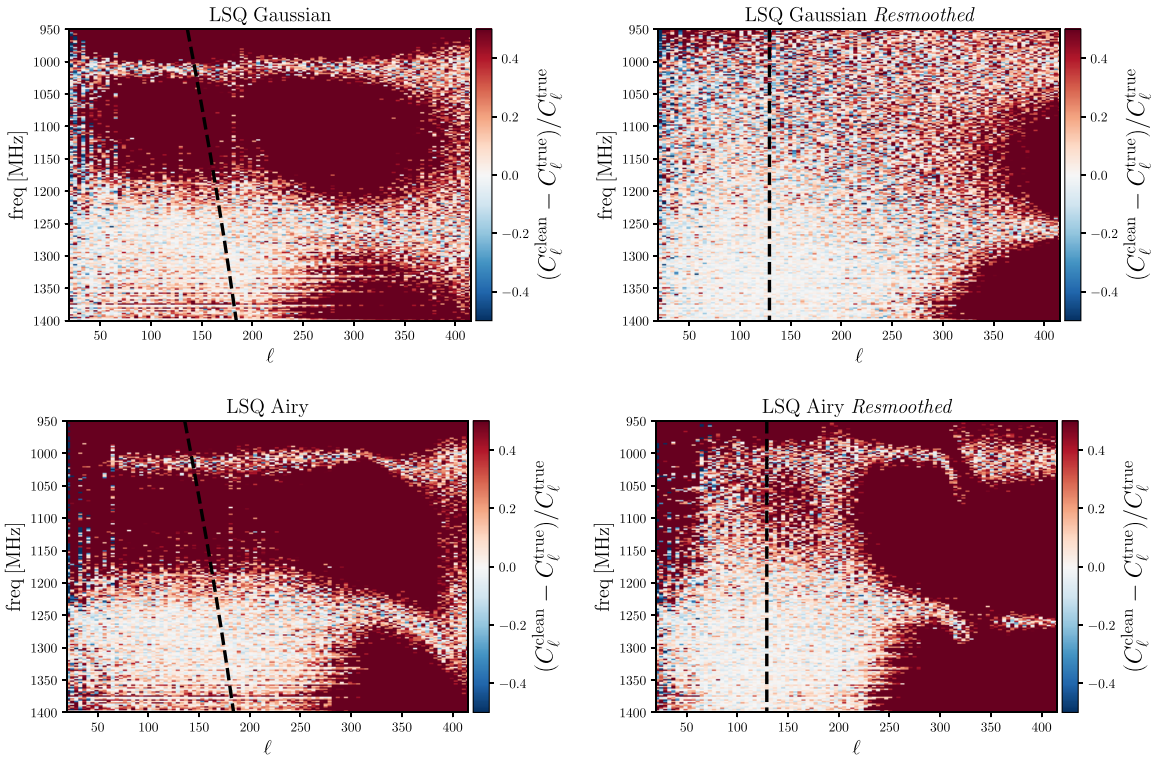

Figure 18. Same as Fig. 17, but here we consider the LSQ method.

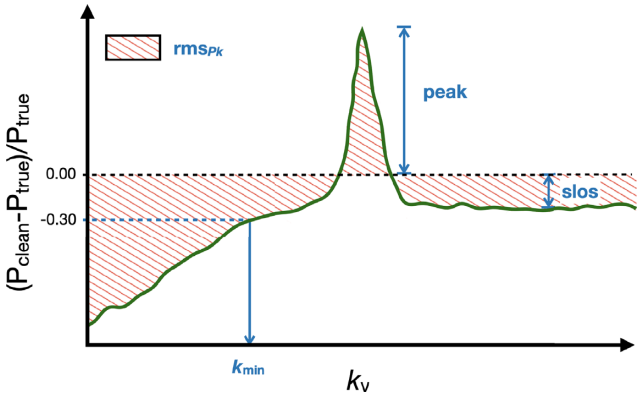

Figure 19. Sketch summarizing the four metrics used to compare the submitted cleaned residuals against ground truth in terms of the radial $P_{los}(k_\nu)$ power spectrum. The curve is a representative one as in Fig. 16. We remind the reader that the *peak* feature appears only for the Airy beam model in combination with the non-trivial PSM foregrounds. See the main text for definitions.

second row. We display nine radar charts in each quadrant, one for each method that joined the Challenge. The title of each chart details (1) the method it refers to and (2) the number of sources removed, N_{fg} (where applicable). Both are also present in the colour coding: green for PCA, blue for FASTICA, green for GMCA, and grey for poLOG and LSQ, and the intensity of the colour is proportional to N_{fg} . We can think of N_{fg} as an extra parameter and dimension of the radar charts, since when interpreting the performances of the methods, one should take N_{fg} into account. For instance, it is generically true that, in a given observational set-up and cleaning method, the higher N_{fg} , the more the loss of cosmic signal (e.g. see bottom panel of fig. 4 in Cunnington et al. 2021a, and discussion therein).

To preserve the same seven-edge structure for all the radar charts, we decide to show a *peak* rating for the cases with no peak (i.e. Gaussian beam), assigning a 5 to all methods.

Looking at Figs 20 and 21, we can generically conclude that no method clearly outperforms the others and that the efficiency of a given method can vary when facing different types of dirty datacubes. Nevertheless, the richness of these results allows us to understand and highlight different issues related to the contaminant cleaning problem and the methods used to face it. In the next section, we discuss the latter and attempt a comprehensive comparison of the methods' performances for all simulation set-ups involved in the Challenge.

7 DISCUSSION

All pipelines show some strengths at different observables and metrics. Here, we discuss results by dividing them into component separation method employed.

PCA performance: PCA(a) reconstructs well the angular power spectrum for the Gaussian beam model, in both the SKAO and MeerKAT cases (upper left panel of Figs 20 and 21); interestingly, PCA(b), adopting a similar N_{fg}, seems to struggle more in the reconstruction of C_ℓ. This is particularly true at low frequency, as we notice comparing the upper left panel of Fig. 14 with Fig. 7. This behaviour is due to the inverse rms weighting used in PCA(b). We have checked this hypothesis after analysing the performance of the submissions and unblinding the results. We reran the PCA(b) pipeline (with same parameters) removing the weights when computing the data covariance in equation (18). Doing so, and after having the non-weighted PCA(b) go through our performance pipeline, we conclude that the chosen weighting was indeed the reason for a bad reconstruction of the low frequencies. Indeed, the data cubes are characterized by an rms inversely proportional to frequency. That is, the lower frequency channels are less taken into account by PCA(b); therefore, the highest eigenvalues come mainly from the higher frequencies at a fixed value of N_{fg}, which forces the shape of the more structured residuals of the higher frequencies to the whole channel range. The weighting scheme used in PCA(b) was intended to minimize the influence of noise in the component separation, but down-weighting the lower frequencies actually has proven to be detrimental for the cleaning process. Although this inverse frequency band rms weighting is non-beneficial with these realistic simulations, we cannot discard weighting schemes in general – as for example pixel rms weighting schemes. We will implement these schemes in future work.

PCAwls shows good performances across all set-ups and, interestingly, C_ℓ are even better reconstructed in the presence of the Airy beam model, contrary to what we observe for PCA(a). For the original data and in the presence of the Gaussian beam model, PCAwls, together with GMCA and mixGMCA, seems to also recover very accurately the radial power spectrum signal (see also the upper left panel of Fig. 16). The results slightly worsen for the radial power spectrum metrics when moving away from the original data cube with the Gaussian beam model. This can be seen also in Fig. 16, where we notice an increment in the signal loss for all blind methods: The bias at small scales changes from less than ∼20 per cent to more than ∼25 per cent. Despite this, PCAwls shows consistently high performances across all metrics and cases. Results are similar for both the original and resmoothed cases, while PCA(a) and (b) typically worsen the quality of the cleaning in the latter case.

FASTICAperformance: Figs 20 and 21 show that FASTICA does not improve on PCA, as already discussed in the context of simulations in e.g. Alonso et al. (2015), Matshawule et al. (2021), and Cunnington et al. (2021a). We recall instead that the application of these two techniques on real data suggests an interesting complementarity and more conservative cleaning results for FASTICA (Wolz et al. 2017). FASTICA(b) is more robust than PCA(b) in the low-frequency reconstruction since only the former has been run adopting an rms weighting of the frequency channels. Moreover, we found good agreement between the two implementations of FASTICA, especially for the original data cubes, where also the N_{fg} chosen is similar.

On N_{fg} and resmoothing: The results for the PCA and FASTICA residuals presented above highlight two important points: (1) Even when using the same N_{fg}, the specific implementation of a method and the pre-processing choices (e.g. mean-centring the maps and weighting scheme) play a non-negligible role; and (2) the resmoothing of the maps with an extra Gaussian kernel may redistribute information among eigenvectors, suppressing the number of relevant eigenvalues of the frequency–frequency covariance matrix of the data cube. This may mislead the N_{fg} choice.

We show in Fig. 22 the ordered eigenvalues of the frequency–frequency covariance of the SKAO–PSM foreground model data cubes corresponding to different beam models and resmoothed or not scenarios. As discussed in Section 3.1, one criterion for determining N_{fg} is to recognize the number of clearly dominant eigenvalues, as the dominant modes are expected to contain most of the foregrounds. Resmoothing redistributes the power of these modes, potentially suggesting a lower N_{fg}. However, despite the effect on the eigenvalue spectra, our analysis indicates that keeping the same or decreasing N_{fg} in the resmoothed cases does not lead to a good cleaning

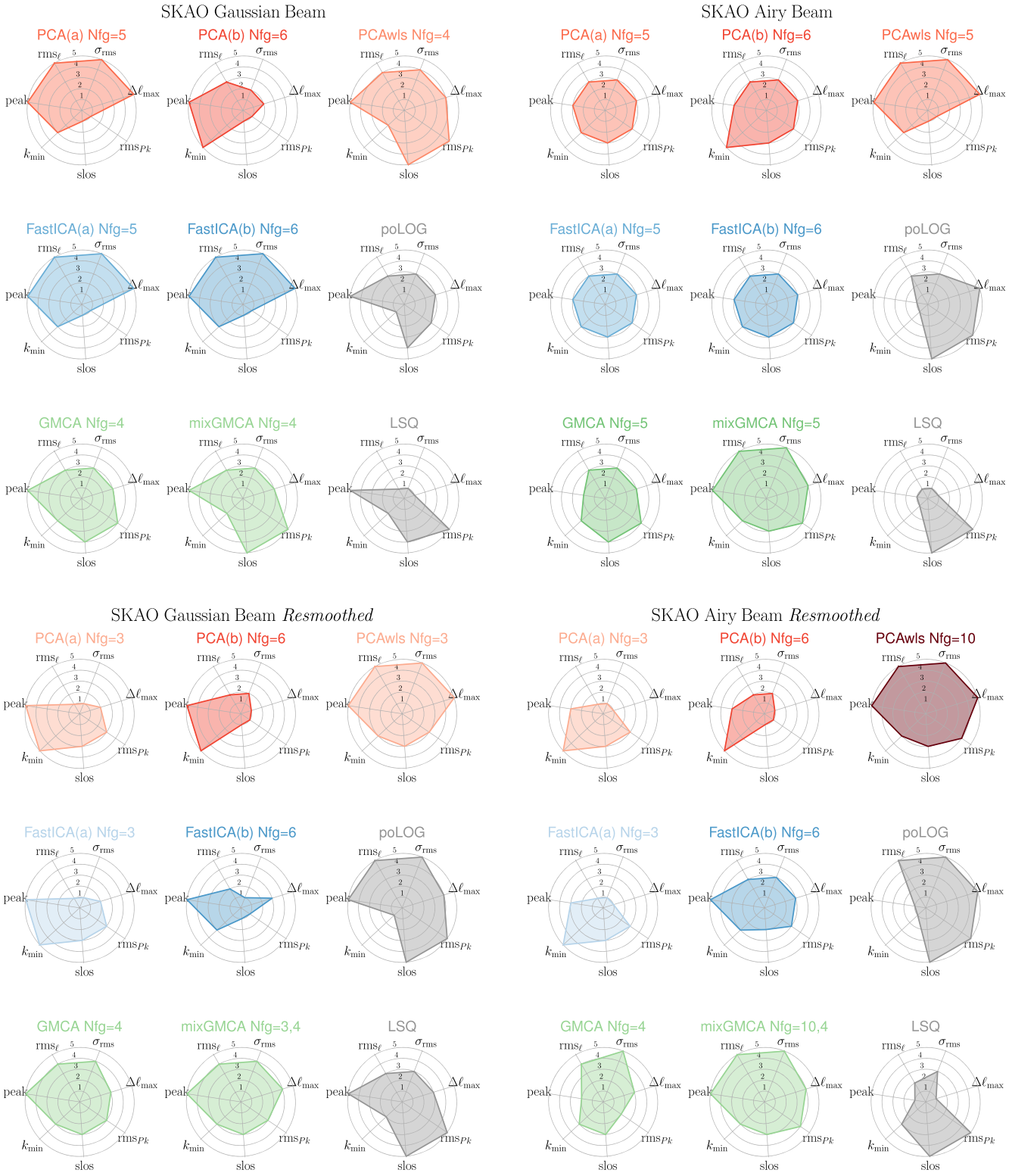

Figure 20. Radar charts showing the performance of the various methods on the different metrics defined in Section 6.3 for an SKAO-MID IM survey and divided in four different panels, one for each of the Gaussian/Airy beam models or original/resmoothed combination. For a given metric, we marked each method from 1 to 5, depending on the *relative* quality in the cleaning (1 = worst, 5 = best); hence, the bigger the area covered by the chart, the better the overall performance. In the scenarios with no *peak* feature (i.e. Gaussian Beam), we assign a 5 to all pipelines to keep the seven-edge structure for the radar charts. Methods are colour coded: PCA pipelines in red, FASTICA in blue, GMCA in green, and non-blind methods in grey. For each blind method, we also report the number of subtracted components $N_{\rm fg}$, and the intensity of the colour is scaled proportionally (darker colour corresponds to higher $N_{\rm fg}$) to help the reading. mixGMCA is associated with two different $N_{\rm fg}$: the first for the largest scale PCA and the second for smaller scales GMCA (see also Table A1).

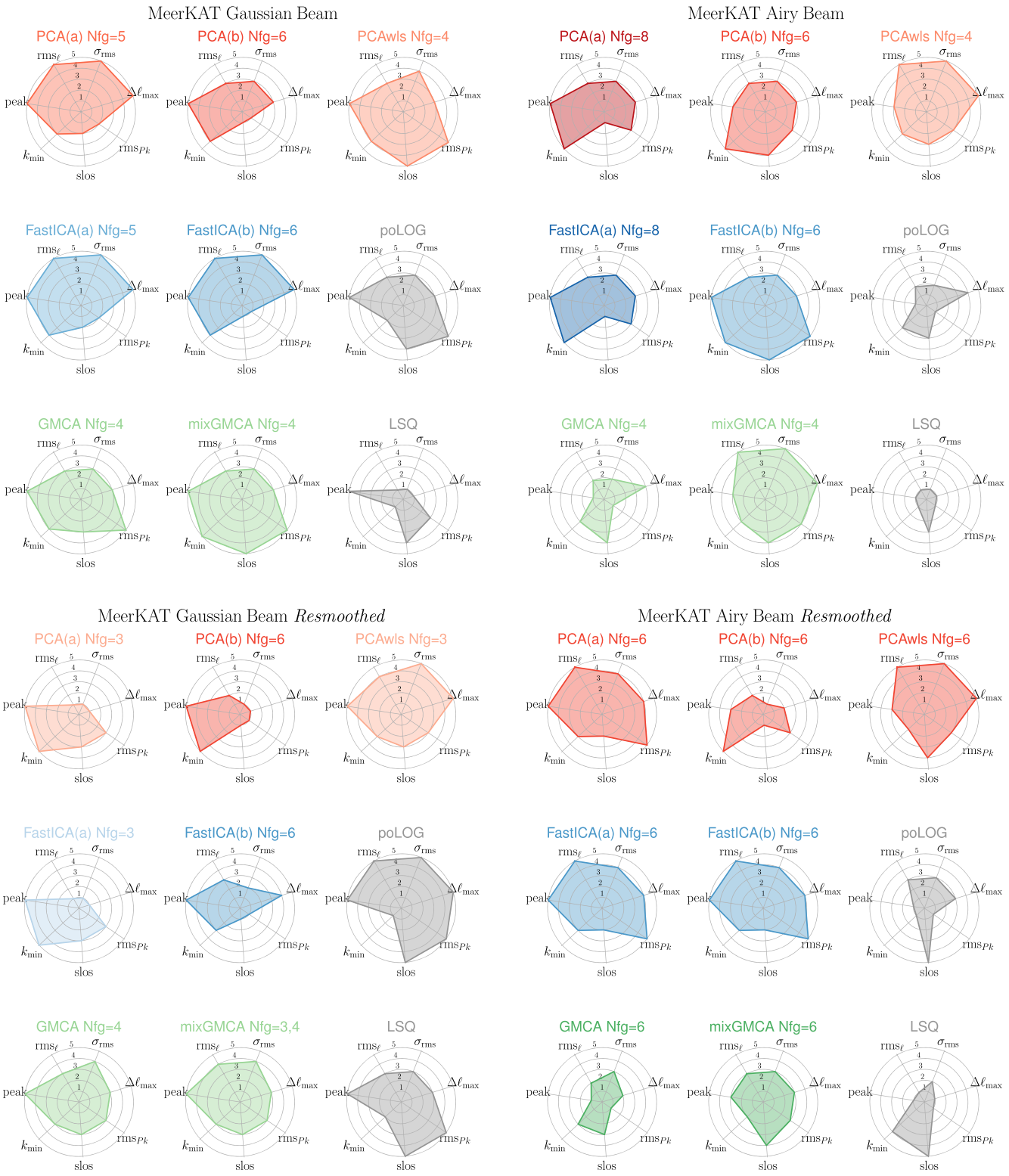

Figure 21. Same as Fig. 20 but for the MeerKAT scenario instead of the SKAO one.

performance: In most cases, it led to more undercleaning in the C_ℓ and more overcleaning in the $P_{\rm los}$. We stress again that the poor performance of the resmoothing depends on the simulation specifics. Different, more subtle, systematics and real observation contaminants could instead benefit from this type of pre-processing. Moreover, a Gaussian deconvolution was possibly not accurate enough for the Airy beam case (see also Matshawule et al. 2021). We postpone a more detailed study of resmoothing to future works.

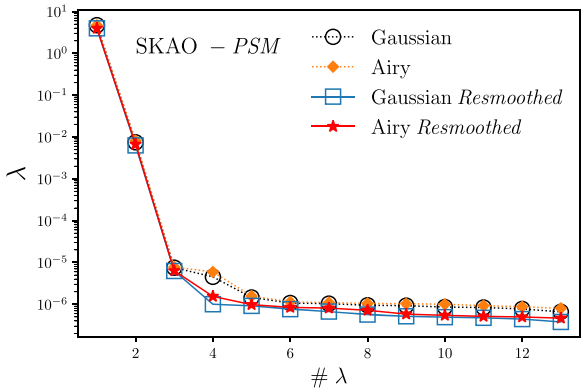

Figure 22. Ordered eigenvalues of the frequency–frequency covariance of the SKAO–PSM foreground model data cubes for different beam model and pre-processing options.

GMCA performance: GMCA and mixGMCA perform similarly in the case of the Gaussian beam model (for both the original and resmoothed cases); indeed, they reconstruct relatively well the radial power spectrum for the original data set (see in particular Fig. 20 and the upper left panel of Fig. 16) and the angular power spectrum in the resmoothed cases. Looking at mixGMCA in Fig. 17, we can also get an idea of the absolute performance of the cleaning on C_ℓ: The reconstruction agrees with the input signal better than few per cent for a large range of ℓ and frequencies. As expected, the reconstruction is more difficult for the scales and frequencies more affected by the beam.

In the more realistic case of the Airy beam model, mixGMCA is better than the GMCA cleaning. Our interpretation is the following. With the Airy beam at play, the morphology of the maps becomes more complex, especially at small scales. The GMCA algorithm seeks and catches those new features and decomposes the signal accordingly, paying most care on those small-scale structures that well satisfy the sparsity assumption. In other words, while performing the source separation process, GMCA decomposes the data cube in the $N_{\rm fg}$ sources that best characterize the small scales, while neglecting the larger (smoother and less sparse) scales. mixGMCA overcomes this problem treating separately the large coarse scale (as decomposed by the wavelet transform) with a PCA cleaning and, moreover, disentangling the $N_{\rm fg}$ needed.

poLOG performance: The poLOG method performs well in presence of a Gaussian beam, and in particular on the resmoothed data cubes. The mean level of the radial information is correctly reconstructed. However, it is possible that the metrics do not penalize enough the small but present oscillatory pattern shown in Fig. 16 (see also Ghosh et al. 2011a, b). The results are almost equally good in the presence of the Airy beam for the SKAO case, although the quality of the cleaning lowers for the MeerKAT case, possibly due to the more prominent side lobes of its beam. Interestingly, when the Airy beam is considered, while the peak is clearly visible in the reconstructed radial power spectrum, the angular power spectra, due to the smoothness assumption of poLOG, do not present the typical fringe pattern at high ℓ (see Fig. 14).

LSQ performance: The LSQ method relies on more physical modelling of the foreground and assumes an a priori knowledge of the monopole of the maps. Unsurprisingly, the mean level of the radial information is always correctly reconstructed given the perfect knowledge of the monopole. LSQ performs satisfactorily in the case of a Gaussian beam model for the resmoothed case, while struggling in the case of the more complex Airy beam model. Overall, we see that this parametric method is not sophisticated enough to deal with realistic data cubes. A way forward could be to upgrade it and include a modelling of the specific instrument beam and noise properties.

On the low-frequency channels: From all the angular power spectrum figures presented so far, it is evident that all pipelines find more challenging the recovery of the input signal and noise in the lowest frequency channels considered. We believe that this is due to a combination of effects including (1) the stronger beam suppression and (2) the relative lower intensity of the H I signal with respect to instrumental noise level (see Fig. 7). The latter point depends both on the specific H I model and on the intrinsic channelization of the IM experiments (constant channel width corresponds to thicker redshift slices at low frequency, where the clustering of the cosmological signal gets averaged more).

8 CONCLUSIONS

8.1 Summary of the challenge

In this work, we presented a Blind Foreground Cleaning Challenge on a realistic set of low-redshift H I IM simulations for an ∼5000 deg^2 single-dish survey with MeerKAT or the SKAO-MID telescope. The simulations, covering the 950–1400 MHz range, include an H I signal generated by combining a semi-analytical galaxy formation model with a cosmological halo simulation, and astrophysical foregrounds, generated using two alternative models: a Gaussian realization of the foreground two-point statistic and frequency scaling properties, and a more empirically informed one, based on the PSM. We simulated instrumental effects through a commonly used Gaussian beam and an Airy beam model that includes side lobes. We modelled a fixed-elevation scanning strategy resulting into a non-homogeneous noise level.

In summary, the various set-up combinations resulted in 16 *dirty* data cubes to be cleaned, resulting in increasingly realistic scenarios that allow a gradual understanding of the role of individual observational features in the cleaning process. Nine foreground cleaning pipelines joined this first Blind Challenge, i.e. without prior knowledge of H I signal, foregrounds, beam model, and noise level. Seven of the pipelines (versions of PCA, FASTICA, and GMCA) linearly decompose the given data-cube leveraging statistical properties of the foreground components such as non-Gaussianity or sparsity. The other two methods either impose the foreground smoothness in frequency (polynomial fitting) or make physical assumptions on the foreground properties (least-squares fitting). Testing many different methods on the same simulation allowed us to quantify their *relative* accuracy on cleaning. We devised a set of criteria to describe the quality of the cleaned residuals in terms of their angular and line-of-sight power spectra and presented their relative performance using radar charts (see Figs 20 and 21).

8.2 Lessons learned

Our results suggest that, even among similar methods, subtleties related to each specific implementation can lead to substantial differences in the cleaning performance, and that the choice of $N_{\rm fg}$ is not easily deducible and objective without extra prior information of the signal. Nevertheless, in the presence of a Gaussian beam, all pipelines (with the exception of least-squares fitting) are capable of recovering within 20 per cent the input power spectra in the frequency range and spatial scales with the least beam suppression.

Interestingly, when the more realistic Airy beam model is considered in combination with the non-Gaussian PSM foregrounds, the cleaning is more complicated and the residuals show (1) a clear spike-like feature in the line-of-sight power spectrum and (2) a fringe pattern in C_ℓ at small angular scales, caused by an oscillation in frequency of the beam side-lobe positions. By enforcing smoothness, the polynomial fitting method is the only exception, not inducing the latter effect on the angular power spectrum. These systematics are caused by the interaction of spatially structured foregrounds with the far side lobes of the primary beam. We expect these effects to worsen in the presence of stronger point sources (Matshawule et al. 2021) or for observations closer to the Galactic plane. In general, also strong Galactic emission at more than 30 deg from the line of sight could play a role, implying that accurate measurements of the *full* primary beam response will be critical for the success of SKAO, MeerKAT, or any single-dish H I intensity experiment.

We found that resmoothing with a Gaussian kernel does not improve the absolute performance of the cleaning (with the exception of the least-squares fitting method). However, (1) our simulation does not include some challenging systematics – such as polarization leakage – that could be mitigated by an aggressive resmoothing[10] and (2) a more accurate deconvolution model including side-lobe structure should be used. Most existing cleaning methods do not directly use any beam information during the component separation process, while our results highlight the need for a more accurate treatment of the beam. More sophisticated strategies are possible, for example performing component separation and deconvolution simultaneously (e.g. Carloni Gertosio & Bobin 2021).

In general, we conclude that methods based on statistical properties of the data (PCA, FASTICA, GMCA, and mixGMCA) should be generally preferred to parametric ones, given the current knowledge of foregrounds at the relevant frequencies combined with the systematic effects.

We find that implementing the cleaning in parallel with more than one method is an excellent practice to unveil different data characteristics. Indeed, a source separation method is more efficient than another if its assumptions suit the data better, helping develop ad hoc cleaning strategies. For instance, we report that mixGMCA, a hybrid PCA-GMCA algorithm, has shown overall improvement compared to its parent methods and the best consistency among all scenarios (i.e. its performance is satisfactory in all cases). Hybrid approaches have the potential to retain the advantages of each of the methods that compose it. In particular, mixGMCA removes the brightest diffuse astrophysical contamination with PCA on large scales while carefully handling the small-scale instrument-driven defects in the maps with GMCA.

8.3 Perspectives

In this work, we explored several methods available in the literature, making it the most comprehensive study so far for post-reionization H I IM foreground cleaning. Nevertheless, more methods could be tested on our end-to-end simulations [e.g. GNILC (Olivari et al. 2016; Fornazier et al. 2021), GPR (Mertens, Ghosh & Koopmans 2018; Soares et al. 2021), and KPCA (Irfan & Bull 2021)]. Known systematics could also be included, such as polarization leakage (Alonso, Ferreira & Santos 2014; Shaw et al. 2015; Spinelli et al. 2018), satellite contamination (Harper & Dickinson 2018), strong RFI flagging (Carucci et al. 2020), 1/*f* noise (Harper et al. 2018; Chen et al. 2020; Li et al. 2021), point source masking (Switzer et al. 2019), and a more realistic description of the system temperature (Wang et al. 2021). As more IM data will be available, it will be possible to understand new observational effects and systematics and include them in the modelling, also paving the way to simulation-based learning algorithms for addressing foreground cleaning.

This first Challenge is designed as the baseline case to test the ability to recover the H I cosmological signal, including realistic observational effects. These simulations lay the ground for developing more complex and detailed end-to-end simulations necessary to improve foreground cleaning pipelines leading to robust H I signal detection in the forthcoming MeerKAT/SKAO era.

[10]McCallum et al. (2021) have recently proposed to suppress polarization leakage at the map-making stage.

ACKNOWLEDGEMENTS

We thank the anonymous referee for useful suggestions that improved the readability of this manuscript. We warmly thank Jingying Wang, Phil Bull, and Keith Grainge for valuable feedback. MS would like to thank Tiago Castro for valuable help with the PINOCCHIO simulations, and Siyambonga Matshawule and Mario Santos for useful discussions. IPC thanks Jérôme Bobin for feedback on the GMCA and mixGMCA implementations. LW would like to thank Clive Dickinson and Keith Grainge for useful discussion in the Challenge set-up. MS acknowledges funding from the INAF PRIN-SKA 2017 project 1.05.01.88.04 (FORECaST) and support from the INFN INDARK PD51 grant. IPC acknowledges support from the 'Departments of Excellence 2018–2022' Grant (L. 232/2016) awarded by the Italian Ministry of University and Research (MUR), from the 'Ministero degli Affari Esteri della Cooperazione Internazionale – Direzione Generale per la Promozione del Sistema Paese Progetto di Grande Rilevanza ZA18GR02', and, at the early stage of this work, from the European Union through the grant LENA (ERC StG no. 678282). SC is supported by STFC grant ST/S000437/1. MI acknowledges support from the South African Radio Observatory, National Research Foundation (Grant No. 84156) and, at the early stage of this work, from the European Union through the grant LENA (ERC StG no. 678282). JF was supported by the University of Padova under the STARS Grants programme *CoGITO: Cosmology beyond Gaussianity, Inference, Theory, and Observations* and by the UK Science & Technology Facilities Council (STFC) Consolidated Grant ST/P000592/1. AP is a UK Research and Innovation Future Leaders Fellow, grant MR/S016066/1, and also acknowledges support by STFC grant ST/S000437/1.

This research utilized Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT http://doi.org/10.5281/zenodo.438045. This work made use of the South African Centre for High-Performance Computing, under the project *Cosmology with Radio Telescopes*, ASTRO-0945. This research made use of `Numpy` (Harris et al. 2020), `Astropy` (The Astropy Collaboration 2013), `Scipy` (Virtanen et al. 2020), `healpy` (Zonca et al. 2019), and the *HEALPix* (Górski et al. 2005) package.

Author contribution: All authors contributed to the design of the Blind Challenge and composing of the article. MS led the analysis and presentation of the results, performed together with IPC. MS and IPC drafted the first version of the manuscript. Simulations: MS (H I distribution), SC (MS_{05} foregrounds), SH (instrumental effects), and MI (PSM foregrounds). Participants of the Blind Challenge: IPC (PCAwls, GMCA, mixGMCA), SC [PCA(a), FASTICA(a)], MI (LSQ), and JF [PCA(b), FASTICA(b), poLOG]. The project was initiated and coordinated by LW and AP as co-chairs of the H I Intensity Mapping Focus Group of the SKA Cosmology SWG.

DATA AVAILABILITY

Simulated data cubes have been produced in support of this research. They are publicly available at the UWC-CRC Repository. The code used to simulate the instrumental effects can be found on `github`: https://github.com/SharperJBCA/SWGSimulator.

REFERENCES

Alonso D., Ferreira P. G., Santos M. G., 2014, MNRAS, 444, 3183
Alonso D., Bull P., Ferreira P. G., Santos M. G., 2015, MNRAS, 447, 400
Alonso D., Sanchez J., Slosar A., LSST Dark Energy Science Collaboration, 2019, MNRAS, 484, 4127
Anderson C. J. et al., 2018, MNRAS, 476, 3382
Ansari R. et al., 2012, A&A, 540, A129
Asad K. M. B. et al., 2021, MNRAS, 502, 2970
Asorey J. et al., 2020, MNRAS, 495, 1788
Bagla J. S., Khandai N., Datta K. K., 2010, MNRAS, 407, 567
Bandura K. et al., 2014, in Stepp L. M., Gilmozzi R., Hall H. J., eds, Proc. SPIE Conf. Ser. Vol. 9145, Ground-Based and Airborne Telescopes V. SPIE, Bellingham, p. 914522
Battye R. A., Davies R. D., Weller J., 2004, MNRAS, 355, 1339
Battye R. A., Browne I. W. A., Dickinson C., Heron G., Maffei B., Pourtsidou A., 2013, MNRAS, 434, 1239
Battye R. et al., 2016, preprint (arXiv:1610.06826)
Bennett C. L. et al., 1992, ApJ, 396, L7
Bharadwaj S., Nath B. B., Sethi S. K., 2001, J. Astrophys. Astron., 22, 21
Bigot-Sazy M. A. et al., 2015, MNRAS, 454, 3240
Blelly A., Moutarde H., Bobin J., 2020, Phys. Rev. D, 102, 104053
Bobin J., Starck J.-L., Fadili J., Moudden Y., 2007, IEEE Trans. Image Process., 16, 2662
Bobin J., Starck J. L., Sureau F., Basak S., 2013, A&A, 550, A73
Bobin J., Sureau F., Starck J. L., Rassat A., Paykari P., 2014, A&A, 563, A105
Boylan-Kolchin M., Springel V., White S. D. M., Jenkins A., Lemson G., 2009, MNRAS, 398, 1150
Braun R., Bonaldi A., Bourke T., Keane E., Wagg J., 2019, preprint (arXiv:1912.12699)
Bull P., Ferreira P. G., Patel P., Santos M. G., 2015, ApJ, 803, 21
Carloni Gertosio R., Bobin J., 2021, Digit. Signal Process., 110, 102946
Carucci I. P., Villaescusa-Navarro F., Viel M., Lapi A., 2015, J. Cosmol. Astropart. Phys., 2015, 047
Carucci I. P., Corasaniti P.-S., Viel M., 2017, J. Cosmol. Astropart. Phys., 2017, 018
Carucci I. P., Irfan M. O., Bobin J., 2020, MNRAS , 499, 304
Chakraborty A. et al., 2021, ApJ, 907, L7
Chang T.-C., Pen U.-L., Peterson J. B., McDonald P., 2008, Phys. Rev. Lett., 100, 091303
Chang T.-C., Pen U.-L., Bandura K., Peterson J. B., 2010, Nature, 466, 463
Chapman E. et al., 2012, MNRAS, 423, 2518
Chapman E. et al., 2013, MNRAS, 429, 165
Chen T., Battye R. A., Costa A. A., Dickinson C., Harper S. E., 2020, MNRAS, 491, 4254
Cunnington S., Wolz L., Pourtsidou A., Bacon D., 2019, MNRAS, 488, 5452
Cunnington S., Irfan M. O., Carucci I. P., Pourtsidou A., Bobin J., 2021a, MNRAS, 504, 208
Cunnington S., Watkinson C., Pourtsidou A., 2021b, MNRAS, 507, 1623
Das S. et al., 2018, in Zmuidzinas J., Gao J.-R., eds, Proc. SPIE Conf. Ser. Vol. 10708, Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy IX. SPIE, Bellingham, p. 1070836
De Lucia G., Kauffmann G., White S. D. M., 2004, MNRAS, 349, 1101
De Lucia G., Tornatore L., Frenk C. S., Helmi A., Navarro J. F., White S. D. M., 2014, MNRAS, 445, 970
Delabrouille J. et al., 2013, A&A, 553, A96
Fernández X. et al., 2016, ApJ, 824, L1
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Finkbeiner D. P., 2003, ApJS, 146, 407
Fonseca J., Liguori M., 2021, MNRAS, 504, 267
Fornazier K. S. F. et al., 2021, preprint (arXiv:2107.01637)
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Gervasi M., Tartari A., Zannoni M., Boella G., Sironi G., 2008, ApJ, 682, 223
Ghosh A., Bharadwaj S., Ali S. S., Chengalur J. N., 2011a, MNRAS, 411, 2426
Ghosh A., Bharadwaj S., Ali S. S., Chengalur J. N., 2011b, MNRAS, 418, 2584
Górski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, ApJ, 622, 759
Harper S. E., Dickinson C., 2018, MNRAS, 479, 2024
Harper S. E., Dickinson C., Battye R. A., Roychowdhury S., Browne I. W. A., Ma Y. Z., Olivari L. C., Chen T., 2018, MNRAS, 478, 2416
Harris C. R. et al., 2020, Nature, 585, 357
Hirschmann M., De Lucia G., Fontanot F., 2016, MNRAS, 461, 1760
Hivon E., Górski K. M., Netterfield C. B., Crill B. P., Prunet S., Hansen F., 2002, ApJ, 567, 2
Hothi I. et al., 2021, MNRAS, 500, 2264
Hu W., Wang X., Wu F., Wang Y., Zhang P., Chen X., 2020, MNRAS, 493, 5854
Hyvärinen A., 1999, IEEE Trans. Neural Netw., 10, 626
Irfan M. O., Bull P., 2021, MNRAS, 508, 3551
Jolicoeur S., Maartens R., De Weerd E. M., Umeh O., Clarkson C., Camera S., 2021, J. Cosmol. Astropart. Phys., 2021, 039
Kitching T. D. et al., 2013, ApJS, 205, 12
Li Y., Santos M. G., Grainge K., Harper S., Wang J., 2021, MNRAS, 501, 4344
McCallum N., Thomas D. B., Bull P., Brown M. L., 2021, MNRAS, 508, 5556
Makinen T. L., Lancaster L., Villaescusa-Navarro F., Melchior P., Ho S., Perreault-Levasseur L., Spergel D. N., 2021, J. Cosmol. Astropart. Phys., 2021, 081
Mao Y., Shapiro P. R., Mellema G., Iliev I. T., Koda J., Ahn K., 2012, MNRAS, 422, 926
Masui K. W. et al., 2013, ApJ, 763, L20
Matshawule S. D., Spinelli M., Santos M. G., Ngobese S., 2021, MNRAS, 506, 5075
Mauch T. et al., 2020, ApJ, 888, 61
Mertens F. G., Ghosh A., Koopmans L. V. E., 2018, MNRAS, 478, 3640
Miville-Deschênes M. A., Ysard N., Lavabre A., Ponthieu N., Macías-Pérez J. F., Aumont J., Bernard J. P., 2008, A&A, 490, 1093
Modi C., Castorina E., Feng Y., White M., 2019, J. Cosmol. Astropart. Phys., 2019, 024
Monaco P., Theuns T., Taffoni G., 2002, MNRAS, 331, 587
Monaco P., Sefusatti E., Borgani S., Crocce M., Fosalba P., Sheth R. K., Theuns T., 2013, MNRAS, 433, 2389
Munari E., Monaco P., Sefusatti E., Castorina E., Mohammad F. G., Anselmi S., Borgani S., 2017, MNRAS, 465, 4658
Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563
Newburgh L. B. et al., 2016, in Hall H. J., Gilmozzi R., Marshall H. K., eds, Proc. SPIE Conf. Ser. Vol. 9906, Ground-Based and Airborne Telescopes VI. SPIE, Bellingham, p. 99065X
Nishimichi T., D'Amico G., Ivanov M. M., Senatore L., Simonovic M., Takada M., Zaldarriaga M., Zhang P., 2020, Phys. Rev. D, 102, 123541
Olivari L. C., Remazeilles M., Dickinson C., 2016, MNRAS, 456, 2749
Olivari L. C., Dickinson C., Battye R. A., Ma Y. Z., Costa A. A., Remazeilles M., Harper S., 2018, MNRAS, 473, 4242
Patil A. H. et al., 2017, ApJ, 838, 65
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Peterson J. B. et al., 2009, Astro2010: The Astronomy and Astrophysics Decadal Survey, Science White Papers. p. 234
Picquenot A., Acero F., Bobin J., Maggi P., Ballet J., Pratt G. W., 2019, A&A, 627, A139
Planck Collaboration XXV, 2016, A&A, 594, A25
Platania P., Bensadoun M., Bersanelli M., De Amici G., Kogut A., Levin S., Maino D., Smoot G. F., 1998, ApJ, 505, 473

Platania P., Burigana C., Maino D., Caserini E., Bersanelli M., Cappellini B., Mennella A., 2003, A&A, 410, 847
Pourtsidou A., 2018, Proc. Sci., HI Intensity Mapping with MeerKAT. SISSA, Trieste, PoS#037
Prat J. et al., 2021, preprint (arXiv:2105.13541)
Remazeilles M., Dickinson C., Banday A. J., Bigot-Sazy M. A., Ghosh T., 2015, MNRAS, 451, 4311
Santos M. G., Cooray A., Knox L., 2005, ApJ, 625, 575
Santos M. G. et al., 2017, Proc. Sci., A Large Sky Survey with MeerKAT. SISSA, Trieste, PoS#032
Seo H.-J., Dodelson S., Marriner J., Mcginnis D., Stebbins A., Stoughton C., Vallinotto A., 2010, ApJ, 721, 164
Shaw J. R., Sigurdson K., Sitwell M., Stebbins A., Pen U.-L., 2015, Phys. Rev. D, 91, 083514
SKA Cosmology SWG, 2020, Publ. Astron. Soc. Aust., 37, e007
Soares P. S., Watkinson C. A., Cunnington S., Pourtsidou A., 2021, MNRAS
Spergel D. N. et al., 2003, ApJS, 148, 175
Spinelli M., Bernardi G., Santos M. G., 2018, MNRAS, 479, 275
Spinelli M., Zoldan A., De Lucia G., Xie L., Viel M., 2020, MNRAS, 493, 5434
Starck J., Murtagh F., Fadili J., 2010, Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity. Cambridge Univ. Press, Cambridge
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
Switzer E. R., Anderson C. J., Pullen A. R., Yang S., 2019, ApJ, 872, 82
Taffoni G., Monaco P., Theuns T., 2002, MNRAS, 333, 623
The Astropy Collaboration, 2013, A&A, 558, A33
Villaescusa-Navarro F., Alonso D., Viel M., 2017, MNRAS, 466, 2736
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Wang X., Tegmark M., Santos M. G., Knox L., 2006, ApJ, 650, 529
Wang J. et al., 2010, ApJ, 723, 620
Wang J. et al., 2013, ApJ, 763, 90
Wang J. et al., 2021, MNRAS, 505, 3698
Wehus I. K. et al., 2017, A&A, 597, A131
Wilson T. L., Rohlfs K., Huttemeister S., 2009, Tools of Radio Astronomy. Springer, Berlin
Wolz L., Abdalla F. B., Blake C., Shaw J. R., Chapman E., Rawlings S., 2014, MNRAS, 441, 3271
Wolz L. et al., 2017, MNRAS, 464, 4938
Wolz L. et al., 2021, preprint (arXiv:2102.04946)
Wyithe J. S. B., Loeb A., 2009, MNRAS, 397, 1926
Xie L., De Lucia G., Hirschmann M., Fontanot F., Zoldan A., 2017, MNRAS, 469, 968
Yohana E., Ma Y.-Z., Li D., Chen X., Dai W.-M., 2021, MNRAS, 504, 5231
Zhang J. et al., 2021, preprint (arXiv:2107.01638)
Zheng H. et al., 2017, MNRAS, 464, 3486
Zoldan A., De Lucia G., Xie L., Fontanot F., Hirschmann M., 2017, MNRAS, 465, 2236
Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, J. Open Source Softw., 4, 1298
Zwart J. T. L., Price D., Bernardi G., 2016, Astrophysics Source Code Library, record ascl:1606.004

APPENDIX A: PIPELINE ASSUMPTIONS ON THE NUMBER OF FOREGROUND COMPONENTS

We present in Table A1 the choices made for the various pipelines on the number of foreground sources to subtract in order to clean the different data cubes (also shown in Fig. 9). In order to assess the most appropriate value of $N_{\rm fg}$ to use, one can look for convergence in the power spectra of residuals while increasing the number of removed components (e.g. fig. 10 in Carucci et al. 2020). $N_{\rm fg}$ can also be estimated by looking at the behaviour of the eigenvalues of the frequency–frequency covariance of data (e.g. fig. 4 in Cunnington et al. 2021a). In the work of Olivari et al. (2016), an automatized choice of $N_{\rm fg}$ is attempted, although highly dependent on prior knowledge of the level of the cosmological signal. The values reported in Table A1 do not show a strong consistency across the different methods, neither a clear trend as a function of the cases studied. Subjectivity seems to have played a major role. Although not reported in the table, the number of foreground components is necessary and crucial for the poLOG method too, since one needs to fix the order of the polynomial that properly describes

Table A1. The chosen values of the number of subtracted components $N_{\rm fg}$ for the different blind cleaning algorithms, as a function of experiment (SKAO/MeerKAT), beam type (Airy/Gaussian), foreground model (PSM/$\rm MS_{05}$), and pre-processing of the data (original data or resmoothed). We remind that the mixGMCA method has two $N_{\rm fg}$, for the large and small scales; here, we report both unless the two coincide.

	Original data				Resmoothed			
Beam:	Gaussian		Airy		Gaussian		Airy	
Fg model:	$\rm MS_{05}$	PSM	$\rm MS_{05}$	PSM	$\rm MS_{05}$	PSM	$\rm MS_{05}$	PSM
				SKAO				
PCA(a)	5	5	5	5	3	3	3	3
PCA(b)	6	6	6	6	6	6	6	6
PCAwls	3	4	4	5	3	3	3	10
FASTICA(a)	5	5	5	5	3	3	3	3
FASTICA(b)	6	6	6	6	6	6	6	6
GMCA	3	4	4	5	4	4	3	4
mixGMCA	3	4	4	5	3/4	3/4	3	10/4
				MeerKAT				
PCA(a)	5	5	5	8	3	3	3	6
PCA(b)	6	6	6	6	6	6	6	6
PCAwls	3	4	4	4	3	3	3	6
FASTICA(a)	5	5	5	8	3	3	3	6
FASTICA(b)	6	6	6	6	6	6	6	6
GMCA	3	4	4	4	4	4	3	6
mixGMCA	3	4	4	4	3/4	3/4	3	6

the foregrounds. Wang et al. (2006) explored different values and concluded that $N_{\rm fg} = 4$ was sufficient for their $z > 6$ simulation. Ansari et al. (2012) considered lower redshifts and truncated their number of components at $N_{\rm fg} = 2$. On the other hand, Alonso et al. (2015) are more conservative as they concluded that $N_{\rm fg} = 7$ are needed. As a compromise between these previous works, here the poLOG method has always been used with $N_{\rm fg} = 6$.

APPENDIX B: PERFORMANCE METRICS VALUES

For completeness, in Tables B1 and B2 we report the values computed for the metrics described in Section 6.3.1, for SKAO and MeerKAT, respectively. In Section 6.3.2, for a given case study (i.e. for a given experiment, a particular beam type, and post-processing choice), the performances of the nine different cleaning methods have been ranked and a *relative* mark between 1 and 5 has been assigned (see the radar charts of Figs 20 and 21). The values reported in Tables B1 and B2 carry further information. For example, it is possible to see the (negative) effect of resmoothing on both the slos and $\Delta\ell_{\rm max}$ (expressed as the percentage of reconstructed C_ℓ values with a precision better than 30 per cent). Moreover, while for the Gaussian beam both the SKAO-MID and MeerKAT set-ups lead to similar results, the smaller side lobes of the SKAO-MID dishes ease the cleaning performances.

Table B1. The values of the seven metrics described in Section 6.3 used for ranking the various cleaning methods for the SKAO case. For simplicity, we express $\Delta\ell_{\rm max}$ as a percentage. The metric $k_{\rm min}$ is expressed in MHz^{-1}. The peak feature is present only when considering the Airy beam and is thus not reported for the Gaussian beam case.

Method	rms$_\ell$	$\sigma_{\rm rms}$	$\Delta\ell_{\rm max}$ (per cent)	rms$_{Pk}$	slos	$k_{\rm min}$ (MHz^{-1})	Peak
			Gaussian				
PCA(a)	0.15	0.12	94.5	0.14	0.14	0.0044	–
PCA(b)	0.54	1.34	83.6	0.16	0.16	0.0022	–
PCAwls	0.15	0.13	93.7	0.03	0.03	0.0067	–
FastICA(a)	0.15	0.12	94.5	0.14	0.14	0.0044	–
FastICA(b)	0.15	0.12	94.3	0.16	0.16	0.0044	–
GMCA	0.18	0.23	92.7	0.03	0.03	0.0044	–
mixGMCA	0.16	0.14	93.4	0.03	0.03	0.0067	–
poLOG	0.19	0.32	92.9	0.04	0.03	0.0089	–
LSQ	2.23	4.87	54.4	0.03	0.03	0.0067	–
			Airy				
PCA(a)	0.67	1.14	82.5	0.16	0.13	0.0044	1.09
PCA(b)	0.89	1.74	76.5	0.18	0.16	0.0022	1.10
PCAwls	0.28	0.38	88.3	0.23	0.24	0.0044	– 0.02
FastICA(a)	0.67	1.14	82.5	0.16	0.13	0.0044	1.09
FastICA(b)	0.66	1.11	82.4	0.18	0.15	0.0044	1.08
GMCA	0.73	1.24	83.0	0.15	0.11	0.0044	1.24
mixGMCA	0.34	0.42	84.7	0.15	0.15	0.0044	0.04
poLOG	0.69	1.28	88.8	0.14	0.03	0.0089	1.44
LSQ	4.15	8.72	49.5	0.14	0.03	0.0089	1.44
			Gaussian *resmoothed*				
PCA(a)	36.16	88.13	56.8	0.26	0.26	0.0044	–
PCA(b)	24.64	55.49	52.2	0.35	0.36	0.0044	–
PCAwls	10.72	34.88	67.8	0.26	0.26	0.0067	–
FastICA(a)	36.16	88.13	56.8	0.26	0.26	0.0044	–
FastICA(b)	19.39	95.85	64.7	0.34	0.34	0.0067	–
GMCA	12.32	36.37	64.8	0.25	0.26	0.0067	–
mixGMCA	11.58	36.04	66.0	0.26	0.26	0.0067	–
poLOG	10.22	33.90	66.6	0.05	0.001	0.0133	–
LSQ	13.23	40.25	63.9	0.03	0.0004	0.0089	–
			Airy *resmoothed*				
PCA(a)	41.74	94.85	47.3	0.32	0.26	0.0044	2.35
PCA(b)	26.23	79.55	46.8	0.39	0.35	0.0044	2.22
PCAwls	12.62	40.10	61.5	0.28	0.29	0.0067	0.20
FastICA(a)	41.74	94.85	47.3	0.32	0.26	0.0044	2.35
FastICA(b)	17.16	57.05	54.0	0.31	0.31	0.0067	0.61
GMCA	15.56	42.33	56.0	0.36	0.29	0.0067	2.56
mixGMCA	13.59	41.29	58.2	0.28	0.29	0.0067	0.20
poLOG	13.50	41.23	62.5	0.27	0.001	0.0111	2.96
LSQ	28.78	62.10	46.8	0.27	0.001	0.0067	2.93

Table B2. Same as Table B1 but for the MeerKAT case. For simplicity, we express $\Delta\ell_{\rm max}$ as a percentage. The metric $k_{\rm min}$ is expressed in MHz^{-1}. The peak feature is present only when considering the Airy beam and is thus not reported for the Gaussian beam case.

Method	rms$_\ell$	$\sigma_{\rm rms}$	$\Delta\ell_{\rm max}$ (per cent)	rms$_{Pk}$	slos	$k_{\rm min}$ (MHz^{-1})	Peak
			Gaussian				
PCA(a)	0.14	0.09	95.0	0.06	0.06	0.0044	–
PCA(b)	0.24	0.39	90.3	0.07	0.07	0.0044	–
PCAwls	0.15	0.10	94.2	0.04	0.04	0.0044	–
FastICA(a)	0.14	0.09	95.0	0.06	0.06	0.0044	–
FastICA(b)	0.15	0.09	94.8	0.07	0.07	0.0044	–
GMCA	0.19	0.24	93.3	0.05	0.05	0.0044	–
mixGMCA	0.15	0.11	94.0	0.04	0.04	0.0044	–
poLOG	0.18	0.22	93.9	0.05	0.04	0.0089	–
LSQ	2.27	3.82	50.2	0.05	0.04	0.0111	–
			Airy				
PCA(a)	0.53	0.43	75.6	0.06	0.06	0.0022	0.10
PCA(b)	1.86	3.36	70.2	0.16	0.04	0.0022	2.27
PCAwls	0.34	0.36	84.4	0.06	0.05	0.0044	0.29
FastICA(a)	0.53	0.43	75.6	0.06	0.06	0.0022	0.10
FastICA(b)	0.54	0.45	75.3	0.03	0.03	0.0022	0.15
GMCA	4.40	7.98	78.7	1.14	0.04	0.0044	17.50
mixGMCA	0.34	0.36	83.8	0.05	0.04	0.0044	0.30
poLOG	4.46	8.20	81.7	1.19	0.05	0.0067	18.24
LSQ	8.47	12.84	47.0	1.21	0.05	0.0133	18.42
			Gaussian *resmoothed*				
PCA(a)	102.29	255.73	56.5	0.26	0.27	0.0022	–
PCA(b)	54.23	214.19	53.8	0.35	0.36	0.0022	–
PCAwls	20.79	66.59	67.1	0.26	0.27	0.0067	–
FastICA(a)	102.29	255.73	56.5	0.26	0.27	0.0022	–
FastICA(b)	29.72	134.30	66.4	0.34	0.34	0.0067	–
GMCA	25.89	71.83	64.0	0.25	0.26	0.0067	–
mixGMCA	23.29	69.53	65.3	0.26	0.27	0.0067	–
poLOG	18.82	63.22	66.9	0.06	0.002	0.0155	–
LSQ	25.95	79.45	64.8	0.03	0.001	0.0111	–
			Airy *resmoothed*				
PCA(a)	25.07	67.71	52.3	0.27	0.28	0.0067	0.26
PCA(b)	48.93	223.23	43.7	0.60	0.32	0.0022	8.00
PCAwls	24.87	65.96	53.0	0.58	0.13	0.0089	8.74
FastICA(a)	25.07	67.71	52.3	0.27	0.28	0.0067	0.26
FastICA(b)	25.07	67.71	52.3	0.27	0.28	0.0067	0.26
GMCA	40.85	84.91	45.1	4.30	0.26	0.0067	65.60
mixGMCA	28.53	70.24	49.8	0.58	0.14	0.0089	8.76
poLOG	32.42	77.99	51.1	4.41	0.004	0.0133	67.36
LSQ	62.52	131.25	38.1	4.44	0.003	0.0044	67.90

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