




# The Simons Observatory: Pipeline comparison and validation for large-scale $B$ -modes

Kevin Wolz<sup>1,2</sup> , Susanna Azzoni<sup>3,4</sup>, Carlos Hervías-Caimapo<sup>5,6</sup>, Josquin Errard<sup>7</sup>, Nicoletta Krachmalnicoff<sup>1,2,8</sup>, David Alonso<sup>3</sup>, Carlo Baccigalupi<sup>1,2,8</sup>, Antón Baleato Lizancos<sup>9,10</sup> , Michael L. Brown<sup>11</sup>, Erminia Calabrese<sup>12</sup>, Jens Chluba<sup>11</sup>, Jo Dunkley<sup>17,18</sup>, Giulio Fabbian<sup>12,13</sup>, Nicholas Galitzki<sup>14</sup>, Baptiste Jost<sup>7,15</sup> , Magdy Morshed<sup>7,19</sup>, and Federico Nati<sup>16</sup>

<sup>1</sup> International School for Advanced Studies (SISSA), Via Bonomea 265, 34136 Trieste, Italy  
e-mail: kevin.wolz93@gmail.com

<sup>2</sup> National Institute for Nuclear Physics (INFN) – Sezione di Trieste, Via Valerio 2, 34127 Trieste, Italy

<sup>3</sup> Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK

<sup>4</sup> Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU, WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan

<sup>5</sup> Department of Physics, Florida State University, Tallahassee, Florida 32306, USA

<sup>6</sup> Instituto de Astrofísica and Centro de Astro-Ingeniería, Facultad de Física, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, 7820436 Macul, Santiago, Chile

<sup>7</sup> Université Paris Cité, CNRS, Astroparticule et Cosmologie, 75013 Paris, France

<sup>8</sup> Institute for Fundamental Physics of the Universe (IFPU), Via Beirut 2, 34151 Grignano (TS), Italy

<sup>9</sup> Berkeley Center for Cosmological Physics, Department of Physics, University of California, Berkeley, CA 94720, USA

<sup>10</sup> Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

<sup>11</sup> Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, The University of Manchester, Oxford Road, Manchester M20 4PE, UK

<sup>12</sup> School of Physics and Astronomy, Cardiff University, The Parade, Cardiff, Wales CF24 3AA, UK

<sup>13</sup> Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, 10010 New York, NY, USA

<sup>14</sup> Department of Physics, University of Texas at Austin, Austin, Texas 78722, USA

<sup>15</sup> Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU, WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan

<sup>16</sup> Department of Physics, University of Milano-Bicocca, Piazza della Scienza 3, 20126 Milan, Italy

<sup>17</sup> Joseph Henry Laboratories of Physics, Jadwin Hall, Princeton University, Princeton, NJ 08544, USA

<sup>18</sup> Department of Astrophysical Sciences, Peyton Hall, Princeton University, Princeton, NJ 08544, USA

<sup>19</sup> CNRS-UCB International Research Laboratory, Centre Pierre Binétruy, IRL2007, CPB-IN2P3, Berkeley, USA

Received 8 February 2023 / Accepted 5 March 2024

## ABSTRACT

**Context.** The upcoming Simons Observatory Small Aperture Telescopes aim at achieving a constraint on the primordial tensor-to-scalar ratio  $r$  at the level of  $\sigma(r=0) \lesssim 0.003$ , observing the polarized CMB in the presence of partial sky coverage, cosmic variance, inhomogeneous non-white noise, and Galactic foregrounds.

**Aims.** We present three different analysis pipelines able to constrain  $r$  given the latest available instrument performance, and compare their predictions on a set of sky simulations that allow us to explore a number of Galactic foreground models and elements of instrumental noise, relevant for the Simons Observatory.

**Methods.** The three pipelines employ different combinations of parametric and non-parametric component separation at the map and power spectrum levels, and use  $B$ -mode purification to estimate the CMB  $B$ -mode power spectrum. We applied them to a common set of simulated realistic frequency maps, and compared and validated them with focus on their ability to extract robust constraints on the tensor-to-scalar ratio  $r$ . We evaluated their performance in terms of bias and statistical uncertainty on this parameter.

**Results.** In most of the scenarios the three methodologies achieve similar performance. Nevertheless, several simulations with complex foreground signals lead to a  $>2\sigma$  bias on  $r$  if analyzed with the default versions of these pipelines, highlighting the need for more sophisticated pipeline components that marginalize over foreground residuals. We show two such extensions, using power-spectrum-based and map-based methods, that are able to fully reduce the bias on  $r$  below the statistical uncertainties in all foreground models explored, at a moderate cost in terms of  $\sigma(r)$ .

**Key words.** methods: data analysis – methods: statistical – cosmic background radiation – cosmological parameters – early Universe – inflation

## 1. Introduction

One of the next frontiers in cosmological science using the cosmic microwave background (CMB) is the observation of

large-scale  $B$ -mode polarization, and the consequent potential detection of primordial gravitational waves. Such a detection would grant us a glance into the infant Universe and its high-energy physics, at scales unattainable by any other experiment.

Primordial tensor perturbations, which would constitute a stochastic background of primordial gravitational waves, would source a parity-odd  $B$ -mode component in the polarization of the CMB (Kamionkowski et al. 1997; Seljak 1997; Seljak & Zaldarriaga 1997; Zaldarriaga & Seljak 1997). The ratio between the amplitudes of the primordial power spectrum of these tensor perturbations and the primordial spectrum of the scalar perturbations is referred to as the tensor-to-scalar ratio  $r$ . This ratio covers a broad class of models of the early Universe, allowing us to test and discriminate between models that predict a wide range of values of  $r$ . These include vanishingly small values, as resulting from models of quantum gravity (e.g. Ijjas & Steinhardt 2018, 2019), as well as those expected to soon enter the detectable range, predicted by models of inflation (Starobinskiĭ 1979; Abbott & Wise 1984; Martin et al. 2014a,b; Planck Collaboration X 2020). An unequivocal measurement of  $r$ , or a stringent upper bound, would thus greatly constrain the landscape of theories of the early Universe.

Although there is no evidence of primordial  $B$ -modes yet, current CMB experiments place stringent constraints on their amplitude, finding  $r < 0.036$  at 95% confidence (BICEP/Keck Collaboration 2021) when evaluated at a pivot scale of  $0.05 \text{ Mpc}^{-1}$ . At the same time, these experiments firmly establish that the power spectrum of primordial scalar perturbations is not exactly scale-independent, with the scalar spectral index  $n_s - 1 \sim 0.03$  (e.g. Planck Collaboration VI 2020). Given this measurement, several classes of inflationary models predict  $r$  to be in the  $\sim 10^{-3}$  range (see Kamionkowski & Kovetz 2016, and references therein).

Even though the only source of primordial large-scale  $B$ -modes at linear order are tensor fluctuations, in practice, a measurement is complicated by several factors: first, the gravitational deflection of the background CMB photons by the cosmic large-scale structure creates coherent sub-degree distortions in the CMB, known as CMB lensing (Lewis & Challinor 2006). Through this mechanism, the nonlinear scalar perturbations from the late Universe transform a fraction of the parity-even  $E$ -modes into  $B$ -modes at intermediate and small scales (Zaldarriaga & Seljak 1998). Second, diffuse Galactic foregrounds have significant polarized emission, and in particular foreground components such as synchrotron radiation and thermal emission from dust produce  $B$ -modes with a significant amplitude. Component separation methods, which exploit the different spectral energy distributions (SED) of the CMB and foregrounds to separate the different components, are thus of vital importance (Delabrouille & Cardoso 2007; Leach et al. 2008). Practical implementations of these methods must also be able to carry out this separation in the presence of instrumental noise and systematic effects (e.g. Natoli et al. 2018; Abitbol et al. 2021).

Polarized Galactic foregrounds pose a formidable obstacle when attempting to measure primordial  $B$ -modes at the level of  $r \sim 10^{-3}$ . Current measurements of Galactic emission demonstrate that at the relevant scales, the Galactic  $B$ -mode signal would dominate any existing primordial signal (Planck Collaboration X 2016; Planck Collaboration Int. XXX 2016; Planck Collaboration IV 2020; Planck Collaboration XI 2020). At the minimum of polarized Galactic thermal dust and synchrotron emission, around 80 GHz, their  $B$ -mode signal represents an effective tensor-to-scalar ratio with amplitude larger than the target CMB signal, even in the cleanest regions of the sky (Kraichmalnicoff et al. 2016). Component separation methods are able to clean most of this, but small residuals left after the cleaning could be comparable to the primordial

$B$ -mode signal we want to measure. Many recent works analyze this problem and make forecasts on how well we could potentially measure  $r$  with different ground-based and satellite experiments (e.g. Betoule et al. 2009; Bonaldi & Ricciardi 2011; Katayama & Komatsu 2011; Armitage-Caplan et al. 2012; Errard & Stompor 2012, 2019; Remazeilles et al. 2016, 2018a,b, 2021; Stompor et al. 2016; Errard et al. 2016; Hervías-Caimapo et al. 2017, 2022; Alonso et al. 2017; Thorne et al. 2019; Azzoni et al. 2021; CMB-S4 Collaboration 2022; Vacher et al. 2022; LiteBIRD Collaboration 2022). These works highlight that, if left untreated, systematic residuals from an overly simplistic characterization of foregrounds will bias an  $r \sim 10^{-3}$  measurement by several  $\sigma$ . Thus, it is of vital importance to model the required foreground complexity when cleaning the multi-frequency CMB observations, and to keep a tight control over systematics without introducing significant bias.

Multiple upcoming CMB experiments rank the detection of large-scale primordial  $B$ -modes among their primary science targets. Near-future experiments such as the BICEP Array (Hui et al. 2018) target a detection at the level of  $r \sim 0.01$ , while in the following decade, next-generation projects, such as LiteBIRD (Hazumi et al. 2019) and CMB-S4 (Abazajian et al. 2016), will aim at  $r \sim 0.001$ .

The Simons Observatory (SO), like the BICEP Array, targets the detection of primordial gravitational waves at the level of  $r \sim 0.01$  (see “The Simons Observatory: science goals and forecasts”, SO Collaboration 2019), and its performance at realizing this goal is the main focus of this paper. SO is a ground-based experiment, located at the Cerro Toco site in the Chilean Atacama desert, which observes the microwave sky in six frequency channels, from 27 to 280 GHz, with full science observations scheduled to start in 2024. SO consists of two main instruments. On the one hand, a Large Aperture Telescope (LAT) with a 6m diameter aperture targets small-scale CMB physics, secondary anisotropies, and the CMB lensing signal. Measurements of the latter will serve to subtract lensing-induced  $B$ -modes from the CMB signal to retrieve primordial  $B$ -modes (using a technique known as “delensing”, see Namikawa et al. 2022). On the other hand, multiple Small Aperture Telescopes (SATs) with 0.4m diameter apertures will make large-scale, deep observations of  $\sim 10\%$  of the sky, with the main aim of constraining the primordial  $B$ -mode signal, peaking on scales  $\ell \sim 80$  (the so-called “recombination bump”). We refer to SO Collaboration (2019) for an extended discussion on experimental capabilities.

In this paper, we aim at validating three independent  $B$ -mode analysis pipelines. We compare their performance regarding a potential  $r$  measurement by the SO SATs, and evaluate the capability of the survey to constrain  $\sigma(r=0) \leq 0.003$  in the presence of foreground contamination and instrumental noise. To that end, we produce sky simulations encompassing different levels of foreground complexity, CMB with different values of  $r$  and different amounts of residual lensing contamination, and various levels of the latest available instrumental noise<sup>1</sup>, calculated from the parametric models presented in SO Collaboration (2019).

We feed these simulations through the analysis pipelines and test their performance, quantifying the bias and statistical uncertainty on  $r$  as a function of foreground and noise

<sup>1</sup> We note that the pipelines are still agnostic to some aspects of the instrumental noise such as filtering, which may impact the overall forecasted scientific performance. We anticipate studying these in detail in future work.

**Table 1.** Overview of the component separation pipelines used to infer  $r$ .

Pipeline	Method	Data space	Blind/parametric	$r$ inference step
A	Cross- $C_\ell$ cleaning	Harmonic (power spectra)	Parametric	Multi-frequency $C_\ell$ likelihood
B	NILC cleaning	Needlets (maps)	Blind	CMB-only $C_\ell$ likelihood
C	Map-based cleaning	Pixels (maps)	Parametric	CMB-only $C_\ell$ likelihood

complexity. The three pipelines are described in detail in Sect. 2. Section 3 presents the simulations used in the analysis, including the models used to produce CMB and foreground sky maps, as well as instrumental noise. In Sect. 4, we present our forecasts for  $r$ , the power spectrum products, and a comparison of the relative weights assigned to the individual frequency channels when recovering the cleaned CMB. Section 4.4 shows preliminary results on a set of new, complex foreground simulations. In Sect. 5 we summarize and draw our conclusions. Appendix A summarizes the  $\chi^2$  analysis performed on the cross- $C_\ell$  cleaning pipeline, while Appendix B discusses biases on Gaussian simulations observed with the NILC cleaning pipeline.

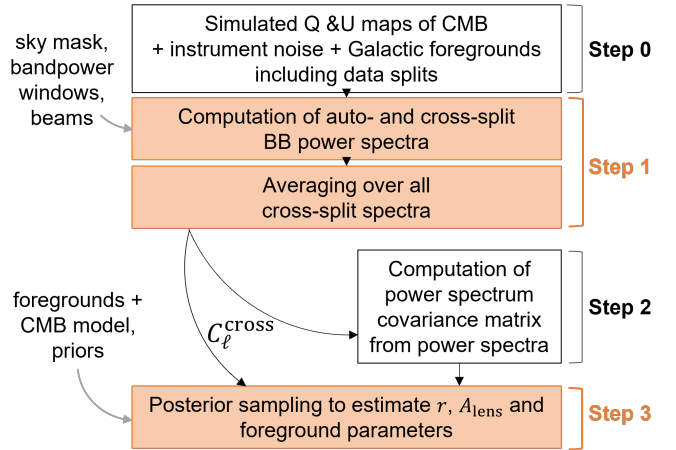
## 2. Methods, pipelines

In this section we present our three component separation pipelines, that adopt complementary approaches widely used in the literature: power-spectrum-based parametric cleaning (BICEP2 Collaboration & Keck Array Collaboration 2016, 2018), needlet internal linear combination (NILC) blind cleaning (Delabrouille et al. 2009; Basak & Delabrouille 2012, 2013), and map-based parametric cleaning (Poletti & Errard in prep.). In the following, these are denominated pipelines A, B, and C, respectively. The cleaning algorithms operate on different data spaces (harmonic, needlet, and pixel space) and vary in their cleaning strategy (parametric, meaning that we assume an explicit model for the frequency spectrum of the foreground components, or blind, meaning that we do not model the foregrounds or make any assumptions on what their frequency spectrum should be). Hence, they do not share the same set of method-induced systematic errors. This will serve as an important argument in favor of claiming robustness of our inference results.

Table 1 lists the three pipelines and their main properties. Although there are some similarities between these analysis pipelines and the forecasting frameworks that were exploited in SO Collaboration (2019), the tools developed for this paper are novel implementations designed to deal with realistic SO data-like inputs, including complex noise and more exotic foreground simulations compared to what was considered in the previous work. We stress again that no filtering or other systematic effects were included in the noise maps.

### 2.1. Pipeline A: Cross- $C_\ell$ cleaning

Pipeline A is based on a multi-frequency power-spectrum-based component separation method, similar to that used in the latest analysis carried out by the BICEP/Keck collaboration (BICEP2 Collaboration & Keck Array Collaboration 2016, 2018). The data vector is the full set of cross-power spectra between all frequency maps,  $C_\ell^{vv'}$ . The likelihood compares this against a theoretical prediction that propagates the map-level sky and instrument model to the corresponding power spectra. The

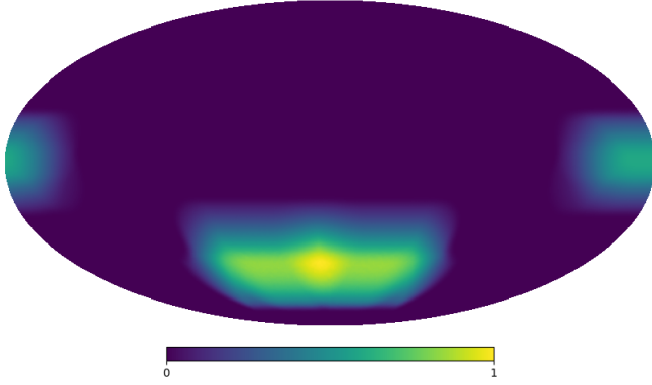
**Fig. 1.** Schematic of pipeline A. Orange colors mark steps that are repeated 500 times, once for each simulation.

full pipeline is publicly available<sup>2</sup>, and a schematic overview is provided in Fig. 1.

In step 1, power spectra are measured using a pseudo- $C_\ell$  approach with  $B$ -mode purification as implemented in NaMaster (Alonso et al. 2019), accounting for the leakage of  $E$ -mode power into  $B$ -mode power caused by incomplete sky coverage. As described in Smith & Zaldarriaga (2007) the presence of a sky mask leads to the presence of ambiguous modes contaminated by full-sky  $E$ -modes. These must be removed at the map level to avoid the contribution to the power spectrum uncertainties from the leaked  $E$ -modes. The mask used for this analysis traces the hits count map released in SO Collaboration (2019) (see Fig. 2), and its edges are apodized using a C1-type kernel (see Grain et al. 2009) with an apodization scale of 10 degrees, yielding an effective sky coverage of  $f_{\text{sky}} \sim 10\%$ . Each power spectrum is calculated in bandpower windows with constant bin width  $\Delta\ell = 10$ , of which we only keep the range  $30 \leq \ell \leq 300$ . Our assumption is that on real data, larger scales are contaminated by atmospheric noise and filtering, whereas smaller scales, targeted by the SO-LAT and useful for constraining lensing  $B$ -modes, do not contain any significant primordial  $B$ -mode contribution. To avoid a significant bias in the auto-correlations when removing the impact of instrumental noise, a precise noise model is required that may not be available in practice. We address this issue by using data splits, which in the case of real data may be formed by subdividing data among observation periods, sets of detectors, sky patches, or by other means, while in this paper, we resort to simulations.

We construct simulated observations for each sky realization comprising  $S = 4$  independent splits with the same sky but different noise realizations (each with a commensurately larger noise amplitude). We compute  $BB$  power spectra from pairs of maps, each associated with a given data split and a given

<sup>2</sup> See [github.com/simonsobs/BBPower](https://github.com/simonsobs/BBPower)



**Fig. 2.** Apodized SAT hits map with effective  $f_{\text{sky}} = 10\%$  used in this paper, shown in equatorial projection. Mask edges are apodized using a C1-type kernel with an apodization scale of 10 degrees.

frequency channel. For any fixed channel pair combination, we average over the corresponding set of  $S(S - 1)/2 = 6$  power spectra with unequal split pairings. For  $N = 6$  SAT frequency channels, this results in a collection of  $N(N + 1)/2 = 21$  noise-debiased multi-frequency power spectra, shown in Fig. 3. We note that, in principle, we could model and subtract the noise bias explicitly, since we have full control over the noise properties in our simulations. In realistic settings, however, the accuracy of an assumed noise model may be limited. While inaccurate noise modeling would affect the statistical uncertainty  $\sigma(r)$  through the covariance matrix calculated from simulations, the cross-split approach ensures robustness of the inferred value of  $r$  against noise-induced bias.

In step 2, we estimate the bandpower covariance matrix from simulations, assuming no correlations between different multipole windows. We note that, since our realistic foreground templates cannot be used as statistical samples, the covariance computation assumes Gaussian foreground simulations (see Sect. 3) to include foreground signal variance in the budget. As we show in Appendix A, this covariance matrix is indeed appropriate, as it leads to the theoretically expected empirical distribution of the  $\chi^2_{\text{min}}$  statistic not only in the case of Gaussian foregrounds, but also for the non-Gaussian foreground simulations. Inaccurate covariance estimates would make this statistic peak at higher or lower values, which we do not observe.

Step 3 is the parameter inference stage. We use a Gaussian likelihood when comparing the multi-frequency power spectra with their theoretical prediction. We note that, in general, the power spectrum likelihood is non-Gaussian, and Hamimeche & Lewis (2008) provide an approximate likelihood that is able to account for this non-Gaussianity. We explicitly verified that both likelihoods lead to equivalent parameter constraints, and thus choose the simpler Gaussian option. The validity of the Gaussian approximation is a consequence of the central limit theorem, since each measured bandpower consists of effectively averaging over  $N_{\text{modes}} \simeq \Delta\ell \times f_{\text{sky}} \times (2\ell + 1) > 61$  independent squared modes on the scales used here. We note that this assumption is valid thanks to the relatively large SO-SAT sky patch and would not hold any longer for the smaller BICEP/Keck-like patch sizes. The default sky model is the same as that described in Abitbol et al. (2021). We model the angular power spectra of dust and synchrotron as power laws of the form  $D_\ell = A_c(\ell/\ell_0)^{\alpha_c}$ , with  $\ell_0 = 80$ , and  $c = d$  or  $s$  for dust and synchrotron, respectively. The dust SED is modeled as a modified black-body spectrum with spectral index  $\beta_d$  and temperature  $\Theta_d$ ,

which we fix to  $\Theta_d = 20$  K. The synchrotron SED is modeled as a power law with spectral index  $\beta_s$ . Finally, we consider a dust-synchrotron correlation parameter  $\epsilon_{ds}$ . Including the tensor-to-scalar ratio  $r$  and a free lensing  $B$ -mode amplitude  $A_{\text{lens}}$ , this fiducial model has nine free parameters:

$$\{A_{\text{lens}}, r, A_d, \alpha_d, \beta_d, A_s, \alpha_s, \beta_s, \epsilon_{ds}\}. \quad (1)$$

We refer to this method as “ $C_\ell$ -fiducial”. Table 2 lists the priors on its parameters.

The main drawback of power-spectrum-based pipelines in their simplest incarnation, is their inability to account for spatial variation in the foreground spectra. If ignored, this spatial variation can lead to biases at the level of  $r \sim O(10^{-3})$ , which are significant for the SO target. At the power spectrum level, spatially-varying SEDs give rise to frequency decorrelation, which can be included in the model. In this work, we show results for an extended model that uses the moment-based<sup>3</sup> parameterization of Azzoni et al. (2021) to describe the spatial variation of  $\beta_d$  and  $\beta_s$ . The model introduces four new parameters

$$\{B_s, \gamma_s, B_d, \gamma_d\}, \quad (2)$$

where  $B_c$  parameterizes the amplitude of the spatial variations in the spectral index of component  $c$ , and  $\gamma_c$  is their power spectrum slope (see Azzoni et al. 2021, for further details). We refer to results using this method as “ $C_\ell$ -moments”, or “A + moments”. The priors in the shared parameter space are the same as for  $C_\ell$ -fiducial. Table 2 lists the priors on its additional four parameters. For both methods, we sample posteriors using the emcee code (Foreman-Mackey et al. 2013). It should be noted that we assume a top-hat prior on  $r$  in the range  $[-0.1, 0.1]$  instead of imposing  $r > 0$ . The reason is that we would like to remain sensitive to potential negative biases on  $r$ . While negative  $r$  values do not make sense physically, they may result from volume effects caused by choosing specific priors on other parameters that we marginalize over. Opening the prior on  $r$  to negative values allows us to monitor these unwanted effects, offering a simple robustness check. On real data, this will be replaced by a positivity prior  $r > 0$ , but only after ensuring that our specific prior choices on the other parameters do not bias  $r$ , which is the focus of future work.

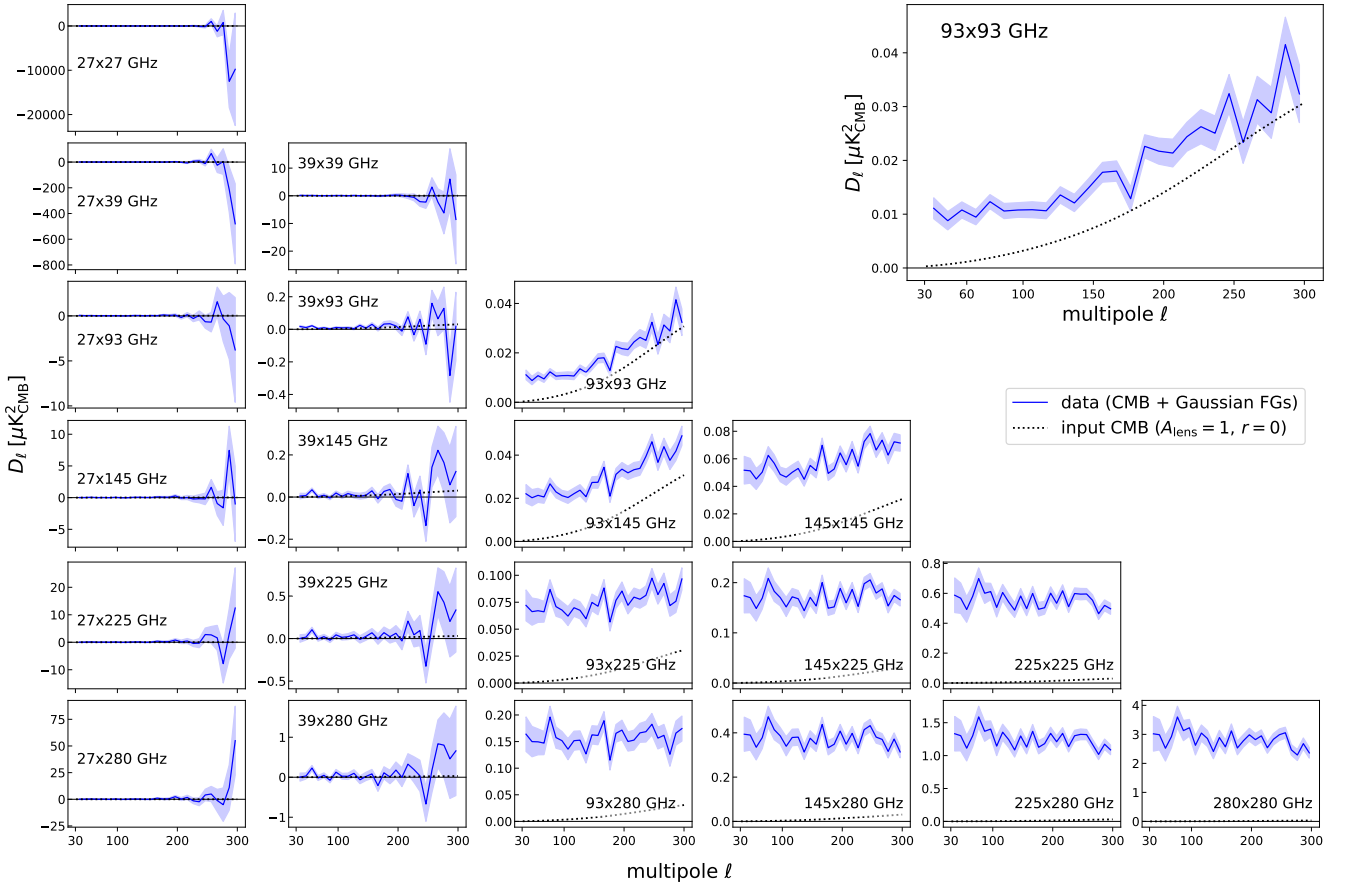
## 2.2. Pipeline B: NILC cleaning

Our second pipeline is based on the blind Internal Linear Combination (ILC) method, which assumes no information on foregrounds whatsoever, and instead only assumes that the observed data contains one signal of interest (the CMB), plus noise and contaminants (Bennett et al. 2003). The method assumes a simple model for the observed multi-frequency maps  $\mathbf{d}_\nu$  at  $N_\nu$  frequency channels (in either pixel or harmonic space)

$$\mathbf{d}_\nu = a_\nu \mathbf{s} + \mathbf{n}_\nu, \quad (3)$$

where  $a_\nu$  is the black-body spectrum of the CMB,  $\mathbf{s}$  is the amplitude of the true CMB signal, and  $\mathbf{n}_\nu$  is the contamination in channel  $\nu$ , which includes the foregrounds and instrumental noise. ILC exploits the difference between the black-body spectrum of the CMB and the SED(s) of other components that may be

<sup>3</sup> See Tegmark (1998), Chluba et al. (2017), Vacher et al. (2023) for more details on the moment-expansion formalism in the context of CMB foregrounds, and Mangilli et al. (2021) as an alternative power-spectrum-based description.



**Fig. 3.** Simulated power spectrum input data analyzed by pipeline A. We show a single realization of CMB and Gaussian foregrounds. Blue shaded areas quantify the  $1\sigma$  Gaussian standard deviation calculated from simulations of CMB, noise, and Gaussian foregrounds. We note that negative auto-spectra can occur at noise-dominated scales as a result of cross-correlating data splits.

**Table 2.** Parameter priors for pipeline A, considering both the  $C_\ell$ -fiducial model and the  $C_\ell$ -moments model.

Model	$C_\ell$ -fiducial and $C_\ell$ -moments									$C_\ell$ -moments only			
Parameter	$A_{\text{lens}}$	$r$	$A_d$	$\alpha_d$	$\beta_d$	$A_s$	$\alpha_s$	$\beta_s$	$\epsilon_{ds}$	$B_s$	$\gamma_s$	$B_d$	$\gamma_d$
Prior type	TH	TH	TH	TH	G	TH	TH	G	TH	TH	TH	TH	TH
Center value	1.0	0.0	25	0.0	1.54	2.0	-1.0	-3.0	0.0	0.0	-4.0	5.0	-4.0
Half width	1.0	0.1	25	0.5	0.11	2.0	1.0	0.3	1.0	10.0	2.0	5.0	2.0

**Notes.** Prior types are either Gaussian (G) or top-hat (TH), considered distributed symmetrically around the center value with half width meaning the standard deviation (Gaussian) or the half width (top-hat).

present in the data. The method aims at reconstructing a map of the CMB component  $\tilde{\mathbf{s}}$  as a linear combination of the data with a set of weights  $w_\nu$ , allowed to vary across the map,

$$\tilde{\mathbf{s}} = \sum_\nu w_\nu \mathbf{d}_\nu = \mathbf{w}^T \hat{\mathbf{d}}, \quad (4)$$

where both  $\mathbf{w}$  and  $\hat{\mathbf{d}}$  are  $N_\nu \times N_{\text{pix}}$  matrices, with  $N_{\text{pix}}$  being the number of pixels. We optimize the weights by minimizing the variance of  $\tilde{\mathbf{s}}$  and find

$$\mathbf{w}^T = \frac{\mathbf{a}^T \hat{\mathbf{C}}^{-1}}{\mathbf{a}^T \hat{\mathbf{C}}^{-1} \mathbf{a}}, \quad (5)$$

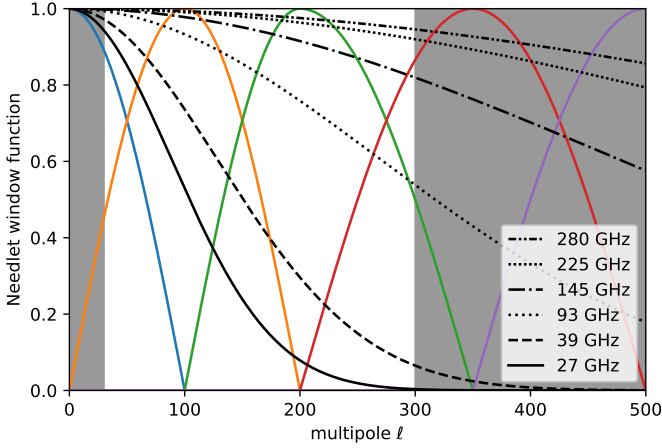
where  $\mathbf{a}$  is the black-body spectrum of the CMB (i.e., a vector filled with ones if maps are in thermodynamic temperature units), and  $\hat{\mathbf{C}} = \langle \hat{\mathbf{d}} \hat{\mathbf{d}}^T \rangle$  is the frequency-frequency covariance

matrix per pixel of the observed data. We do not assume any correlation between pixels in this work.

In our particular implementation, we use the NILC method (Delabrouille et al. 2009; Basak & Delabrouille 2012, 2013). NILC uses localization in pixel and harmonic space by finding different weights  $\mathbf{w}$  for a set of harmonic filters, called “needlet windows”. These windows are defined in harmonic space  $h_i(\ell)$  for  $i = 0, \dots, n_{\text{windows}} - 1$  and must satisfy the constraint  $\sum_{i=0}^{n_{\text{windows}}-1} h_i(\ell)^2 = 1$  in order to preserve the power of the reconstructed CMB. We use  $n_{\text{windows}} = 5$  needlet windows shown in Fig. 4, and defined by

$$h_i(\ell) = \begin{cases} \cos(\frac{\pi}{2}(\ell_i^{\text{peak}} - \ell)/(\ell_i^{\text{peak}} - \ell_i^{\text{min}})) & \text{if } \ell_i^{\text{min}} \leq \ell < \ell_i^{\text{peak}} \\ 1 & \text{if } \ell = \ell_i^{\text{peak}} \\ \cos(\frac{\pi}{2}(\ell - \ell_i^{\text{peak}})/(\ell_i^{\text{max}} - \ell_i^{\text{peak}})) & \text{if } \ell_i^{\text{peak}} < \ell \leq \ell_i^{\text{max}} \end{cases}, \quad (6)$$

with  $\ell_{\text{min}} = \{0, 0, 100, 200, 350\}$ ,  $\ell_{\text{max}} = \{100, 200, 350, 500, 500\}$ , and  $\ell_{\text{peak}} = \{0, 100, 200, 350, 500\}$  for the corresponding



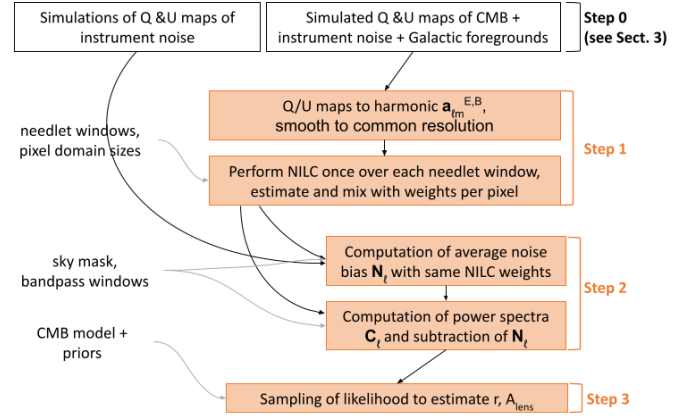
**Fig. 4.** Five needle windows and beam transfer functions as used in pipeline B. Colored lines are needle windows in harmonic space, black dashed lines are the transfer functions  $b_\ell^r$  for the six SO-SAT frequency channels. The FWHM of every beam is listed in Table 3. Gray shaded areas denote multipole ranges that are omitted during  $r$  inference with pipeline B.

five needle windows. Even though we do not use the full 500- $\ell$  range where windows are defined for the likelihood sampling, we still perform the component separation on all five windows up to multipoles beyond our upper limit of  $\ell = 300$ , in order to avoid edge effects on the smaller scales.

Let us now describe the NILC procedure as illustrated in Fig. 5. In step 1, we perform our CMB reconstruction in the  $E$  and  $B$  field instead of  $Q$  and  $U$ . We transform the observed maps to  $a_{\ell m}^X$  with  $X \in E, B$ . All frequency channels are then brought to a common beam resolution by rescaling the harmonic coefficients with an appropriate harmonic beam window function. The common beam we adopt is the one from the third frequency channel at 93 GHz, which corresponds to a FWHM of 30 arcmin. For each needle window index  $i$ , we multiply  $a_{\ell m}^X$  by  $h_i(\ell)$  as a harmonic filter. Since different frequency channels have different limiting resolutions, we do not use all channels in every needle window. The first two windows use all six frequency channels, the third window does not use the 27 GHz channel, and the last two needle windows do not use the 27 and 39 GHz channels. The covariance matrix  $\hat{C}$  has dimensions  $N_\nu \times N_\nu \times N_{\text{pix}}$ . For each pixel  $p$ , its corresponding  $N_\nu \times N_\nu$  elements are computed directly from the data, averaging over the pixels inside a given pixel domain  $\mathcal{D}(p, i)$  around each pixel. In practice, the element  $\nu, \nu'$  of the covariance matrix is calculated by multiplying the two filtered maps at channels  $\nu$  and  $\nu'$ , then smoothing that map with a Gaussian kernel with FWHM equal to the size of the pixel domain  $\mathcal{D}(p, i)$ . The FWHMs for the pixel domain size are 185, 72, 44, 31, and 39 degrees for each needle window, respectively<sup>4</sup>.

We then proceed to calculate the weights  $\mathbf{w}^T$  (see Eq. (5)) for window  $i$ , which is an array with shape  $(2, N_\nu, N_{\text{pixels}}^i)$ , with the first dimension corresponding to the  $E$  and  $B$  fields. We note that the number of pixels  $N_{\text{pixels}}^i$  is different for each needle window, since we use different pixel resolutions that depend on the small-

<sup>4</sup> The domain sizes are estimated directly from the needle window scale (see details in the Appendix A of Delabrouille et al. 2009). The ILC bias can be minimized by enlarging the pixel domains to be big enough to include a higher number of modes. We choose the resulting ILC bias to not exceed 0.2%, for which we need pixel domain sizes large enough so that each needle window contains at least 2500 modes.



**Fig. 5.** Schematic of pipeline B. Orange colors mark steps that are repeated 500 times, once for each simulation.

est scale covered by the respective window. Finally, we apply Eq. (4) to obtain an ILC-reconstructed CMB map for window  $i$ . The final step is to filter this map in harmonic space for a second time with the  $h_i(\ell)$  window. The final reconstructed CMB map is the sum of these maps for all five needle windows.

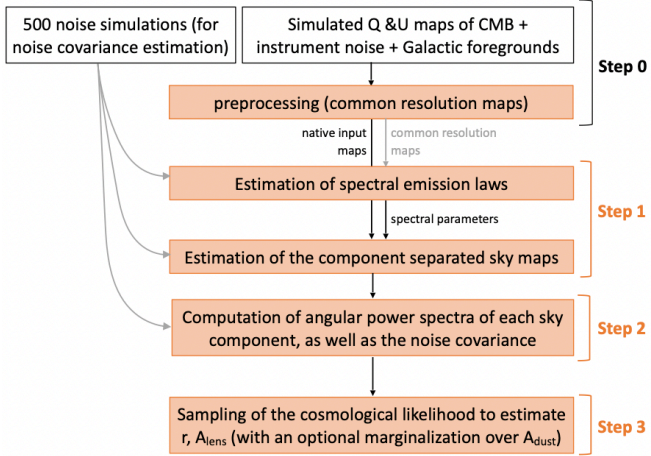
In step 2, the reconstructed CMB maps are compressed into power spectra using NaMaster and deconvolved to the common beam resolution. We use  $B$ -mode purification as implemented in the software and the mask shown in Fig. 2. We estimate the noise bias  $N_\ell$  in the final map by computing the power spectrum of noise-only simulations processed with the needle weights and windows obtained from the simulated data as described above.  $N_\ell$  is averaged over simulations and subtracted from the  $C_\ell$  of the reconstructed maps.

Finally, in step 3, we run a Monte Carlo Markov Chain (MCMC) over the reconstructed  $BB$  spectrum (we ignore  $EE$  and  $EB$ ) with two free parameters, the tensor-to-scalar ratio  $r$  and the amplitude of the  $BB$  lensing spectrum  $A_{\text{lens}}$ . For the posterior sampling, we use the Python package emcee. Both parameters have a top hat prior (between 0 and 2 for  $A_{\text{lens}}$ , and between  $-0.013$  and infinity for  $r$ ). The covariance matrix is calculated directly over 500 simulations with the same setup but with Gaussian foregrounds. As likelihood, we use the same Gaussian likelihood used in pipeline A and restrict the inference to a multipole range  $30 < \ell \leq 300$ .

While the NILC implementation described above is blind, it can be extended to a semi-blind approach that introduces a certain level of foreground modeling. For example, constrained ILC (cILC, Remazeilles et al. 2011) explicitly nullifies one or more contaminants (such as thermal dust) in observed maps, by including their modeled SED in the variance minimization that results in the ILC weights (see Eq. (5)). This foreground modeling can be further extended to include the moment expansion of the SED described in Sect. 2.1. This method, known as constrained moment ILC (cMILC, Remazeilles et al. 2021), has proven effective at cleaning the large-scale  $B$ -mode contamination for space experiments such as LiteBIRD. While not used in this work, these extensions and others will be considered in future analyses with more complex foregrounds and systematics.

### 2.3. Pipeline C: map-based cleaning

Our third pipeline is a map-based parametric pipeline based on the fgbuster code (Poletti & Errard in prep.). This approach is



**Fig. 6.** Schematic of pipeline C. Orange colors indicate repetition for each simulation.

based on the data model

$$\mathbf{d} = \hat{\mathbf{A}}\mathbf{s} + \mathbf{n}, \quad (7)$$

where  $\mathbf{d}$  is a vector containing the polarized frequency maps,  $\mathbf{s}$  is a vector containing the  $Q$  and  $U$  amplitudes of the sky signals (CMB, foregrounds) and  $\mathbf{n}$  is the noise contained in each frequency map. The matrix  $\hat{\mathbf{A}} = \hat{\mathbf{A}}(\boldsymbol{\beta})$  is the so-called mixing matrix, assumed to be parameterized by a set of spectral indices  $\boldsymbol{\beta}$ . Starting from the observed (mock) input data  $\mathbf{d}$ , Fig. 6 shows a schematic of the pipeline, comprising four steps.

Step 0 is the preprocessing of input simulations. For each simulation, we combine the simulated noise maps, the foreground, and CMB maps and save them on disk. We create a new set of frequency maps,  $\tilde{\mathbf{d}}$ , smoothed with a common Gaussian kernel of 100' FWHM.

Step 1 is the actual component separation stage. We optimize the spectral likelihood, defined as (Stompor et al. 2009):

$$-2 \log(\mathcal{L}_{\text{spec}}(\boldsymbol{\beta})) = \left( \hat{\mathbf{A}}^T \hat{\mathbf{N}}^{-1} \tilde{\mathbf{d}} \right)^T \left( \hat{\mathbf{A}}^T \hat{\mathbf{N}}^{-1} \hat{\mathbf{A}} \right)^{-1} \left( \hat{\mathbf{A}}^T \hat{\mathbf{N}}^{-1} \tilde{\mathbf{d}} \right) \quad (8)$$

which uses the common resolution frequency maps,  $\tilde{\mathbf{d}}$ , built during step 0. The right hand side of Eq. (8) contains a sum over the observed sky pixels, assumed to have uncorrelated noise. The diagonal noise covariance matrix  $\hat{\mathbf{N}}$  is computed from 500 noise-only simulations. Although, in principle,  $\hat{\mathbf{N}}$  can be non-diagonal, we do not observe any significant bias of the spectral likelihood due to this approximation in this study. By minimizing Eq. (8) we estimate the best-fit spectral indices  $\tilde{\boldsymbol{\beta}}$  and the corresponding mixing matrix  $\tilde{\mathbf{A}} \equiv \hat{\mathbf{A}}(\tilde{\boldsymbol{\beta}})$ . We also estimate the uncertainties on the recovered spectral indices as provided by the minimizer, a truncated Newton algorithm (Nash 1984) as implemented in `scipy` (Virtanen et al. 2020). Having thus obtained estimators of the foreground SEDs, we can recover the sky component maps with the generalized least-square equation

$$\tilde{\mathbf{s}} = \left( \tilde{\mathbf{A}}^T \hat{\mathbf{N}}^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}^T \hat{\mathbf{N}}^{-1} \tilde{\mathbf{d}} \equiv \hat{\mathbf{W}}\tilde{\mathbf{d}}, \quad (9)$$

where  $\tilde{\mathbf{d}}$  is the input raw data, and not the common resolution maps. In steps 1 and 2, we have the possibility to use an inhomogeneous noise covariance matrix  $\hat{\mathbf{N}} = \hat{\mathbf{N}}(\hat{\mathbf{n}})$  and, although this is not exploited in this work, a spatially varying mixing matrix

$\boldsymbol{\beta} = \boldsymbol{\beta}(\hat{\mathbf{n}})$ . For the latter, one can use the multi-patch or clustering methods implemented in `fgbuster` (Errard & Stompor 2019; Puglisi et al. 2022).

Step 2 comprises the calculation of angular power spectra. The recovered CMB polarization map is transformed to harmonic space using `NaMaster`. We estimate an effective transfer function,  $\mathbf{B}_\ell^{\text{eff}} = \hat{\mathbf{W}}\mathbf{B}_\ell$ , associated with the reconstructed components  $\tilde{\mathbf{s}}$ , from the channel-specific beams  $\mathbf{B}_\ell$ . Correcting for the impact of this effective beam is vital to obtain an unbiased  $BB$  power spectrum of the foreground-cleaned CMB,  $\tilde{C}_\ell^{\text{CMB}}$ . In the second step, we use noise simulations to estimate the noise bias

$$\tilde{N}_\ell = \frac{1}{N_{\text{sim}}} \sum_{\text{sims}} \sum_{m=-\ell}^{\ell} \frac{\tilde{\mathbf{n}}_{\ell,m} \tilde{\mathbf{n}}_{\ell,m}^\dagger}{2\ell + 1}. \quad (10)$$

where  $\tilde{\mathbf{n}} = \hat{\mathbf{W}}\mathbf{n}^{\text{sim}}$  is the noise in the recovered component-separated sky maps. We consider 500 simulations to estimate the noise bias.

Step 3 is the cosmological analysis stage. We model the angular power spectrum of the component-separated CMB map, including the noise contribution, as

$$C_\ell^{\text{CMB}}(r, A_{\text{lens}}) \equiv C_\ell^{\text{prim}}(r) + C_\ell^{\text{lens}}(A_{\text{lens}}) + \tilde{N}_\ell^{\text{CMB}} \quad (11)$$

and compare data and model with the cosmological likelihood

$$-2 \log \mathcal{L}^{\text{cosmo}} = \sum_{\ell} (2\ell + 1) f_{\text{sky}} \left( \frac{\tilde{C}_\ell^{\text{CMB}}}{C_\ell^{\text{CMB}}} + \log(C_\ell^{\text{CMB}}) \right). \quad (12)$$

It is worth noting that this is only an approximation to the true map-level Gaussian likelihood which approximates the effective number of modes in each multipole after masking and purification as  $f_{\text{sky}}(2\ell + 1)$ , thus neglecting any mode-coupling effects induced by the survey footprint. We grid the likelihood above along the two dimensions  $r$  and  $A_{\text{lens}}$ . For each simulation we then estimate the maximum-likelihood values and 68% credible intervals from the marginal distributions of  $r$  and  $A_{\text{lens}}$ . We verified that the distributions of recovered  $\{r, A_{\text{lens}}\}$  across simulations are well described by a Gaussian, hence supporting the Gaussian likelihood in Eq. (12).

Pipeline C also offers the option to marginalize over a dust template. The recovered components in  $\tilde{\mathbf{s}}$ , Eq. (9), include the dust  $Q$  and  $U$  maps which are typically recovered with high signal-to-noise. In the same way that we compute  $\tilde{C}_\ell^{\text{CMB}}$  in step 2, we compute the  $BB$  component of the recovered dust map,  $\tilde{C}_\ell^{\text{dust}}$ . We then update our cosmological likelihood, Eq. (11), by adding a dust term:

$$C_\ell^{\text{CMB}} = C_\ell^{\text{CMB}}(r, A_{\text{lens}}) + A_{\text{dust}} \tilde{C}_\ell^{\text{dust}}. \quad (13)$$

This is a similar approach to earlier methods (Errard & Stompor 2019; LiteBIRD Collaboration 2022). When choosing this approach, the inference of  $r$  during step 3 therefore involves the marginalization over both parameters  $A_{\text{lens}}$  and  $A_{\text{dust}}$ . In principle one could add synchrotron or other terms in Eq. (13) but we limit ourselves to dust as it turns out to be the largest contamination, and, in practice, marginalizing over it allows us to get unbiased estimates of cosmological parameters. In the remainder of this paper, we refer to this method as ‘‘C + dust marginalization’’.

### 3. Description of input simulations

We built a set of dedicated simulations against which to test our data analysis pipelines and compare results. The simulated maps include cosmological CMB signal, Galactic foreground emission as well as instrumental noise.

**Table 3.** Instrument and noise specifications used to produce the simulations in this work.

Frequency [GHz]	$FWHM$ [arcmin]	Baseline		Goal	Pessimistic	Optimistic	
		Noise [ $\mu\text{K}\cdot\text{arcmin}$ ]	Noise [ $\mu\text{K}\cdot\text{arcmin}$ ]	Noise [ $\mu\text{K}\cdot\text{arcmin}$ ]	$\ell_{\text{knee}}$	$\ell_{\text{knee}}$	$\alpha_{\text{knee}}$
27	91	46	33	30	15	-2.4	
39	63	28	22	30	15	-2.4	
93	30	3.5	2.5	50	25	-2.5	
145	17	4.4	2.8	50	25	-3.0	
225	11	8.4	5.5	70	35	-3.0	
280	9	21	14	100	40	-3.0	

**Notes.** It should be stressed that these levels correspond to homogeneous noise, while our default analysis assumes noise maps weighted according to the SAT hits map.

### 3.1. Instrumental specifications and noise

We simulate polarized Stokes  $Q$  and  $U$  sky maps as observed by the SO-SAT telescopes. All maps are simulated using the HEALPix pixelation scheme (Górski et al. 2005) with resolution parameter  $N_{\text{side}} = 512$ .

We model the SO-SAT noise power spectra as

$$N_\ell = N_{\text{white}} \left[ 1 + \left( \frac{\ell}{\ell_{\text{knee}}} \right)^{\alpha_{\text{knee}}} \right], \quad (14)$$

where  $N_{\text{white}}$  is the white noise component while  $\ell_{\text{knee}}$ , and  $\alpha_{\text{knee}}$  describe the contribution from  $1/f$  noise. Following SO Collaboration (2019; hereinafter SO2019), we consider four scenarios: “baseline” and “goal” levels for the white noise component, and “pessimistic” and “optimistic” correlated noise. The empirical  $1/f$  scenarios are based on measurements from recent experiments and consider polarization modulation, filtering, and atmospheric transmission corresponding to the conditions at the Atacama site. The values of white noise,  $\ell_{\text{knee}}$ , and  $\alpha_{\text{knee}}$  associated with the different cases are reported in Table 3. We note that noise levels correspond to a sky fraction of  $f_{\text{sky}} = 10\%$  and five years of observation time, as in SO2019. Differently from SO2019, we cite polarization noise levels at a uniform map coverage, accounting for the factor of  $\sim 1.3$  difference compared to Table 1 in SO2019<sup>5</sup>. We simulate noise maps as Gaussian realizations of the  $N_\ell$  power spectra. In our main analysis, we use noise maps with pixel weights computed from the SO-SAT hits map (see Fig. 2) and refer to this as “inhomogeneous noise”. In Sect. 4.2, we briefly present results obtained from equally weighted noise pixels, which we refer to as “homogeneous noise”. Otherwise, all results in this paper assume inhomogeneous noise. We note that, although inhomogeneous, the noise realizations used here lack some of the important anisotropic properties of realistic  $1/f$  noise, such as stripes due to the scanning strategy. Thus, together with the impact of other time-domain effects (e.g. filtering), we leave a more thorough study of the impact of instrumental noise properties for future work.

### 3.2. CMB

We simulate the CMB signal as isotropic Gaussian random realizations following a power spectrum computed at the Planck 2018 best-fit  $\Lambda\text{CDM}$  cosmology. Our baseline model does not include any primordial tensor signal ( $r = 0$ ) but incorporates lensing power in the  $BB$  spectra ( $A_{\text{lens}} = 1$ ). We consider also two modifications of this model: (i) primordial tensor signal with  $r = 0.01$ , representing a  $\gtrsim 3\sigma$  target detection for SO with

<sup>5</sup> Polarization noise accounts for a factor of  $\sqrt{2}$  and homogeneous noise for a factor of  $\sqrt{0.85}$  compared to Table 1 in SO2019.

$\sigma(r) = 0.003$ , as forecasted by SO2019; (ii) reduced lensing power with  $A_{\text{lens}} = 0.5$ , corresponding to a 50% delensing efficiency, achievable for SO as shown in Namikawa et al. (2022).

For every scenario, we simulated 500 realizations of the CMB signal, convolved with Gaussian beams for each frequency channel, with FWHMs as reported in Table 3.

### 3.3. Foregrounds

Thermal emission from Galactic dust grains and synchrotron radiation are known to be the two main contaminants to CMB observations in polarization, at intermediate and large angular scales, impacting therefore measurements of the primordial  $BB$  signal. The past years have seen many studies on the characterization of polarized Galactic foreground emission, thanks to the analysis of WMAP and Planck data, as well as low frequency surveys (Harper et al. 2022; Krachmalnicoff et al. 2018). However, many aspects of their emission remain unconstrained, including, in particular, the characterization of their SEDs and their corresponding variation across the sky. To properly assess the impact of foreground emission on component separation and  $r$  constraints, we therefore use four sets of sky emission models. As specified in the following, we use the Python sky model (PYSM) package (Thorne et al. 2017) to simulate polarized foreground components, with some additional modifications:

**Gaussian foregrounds.** We simulate thermal dust emission and synchrotron radiation as Gaussian realizations of power law  $EE$  and  $BB$  power spectra. Although inaccurate, since foregrounds are highly non-Gaussian, this idealistic model was used to validate the different pipelines and to build approximate signal covariance matrices from 500 random realizations. In particular, we estimate the amplitudes of the polarized foreground signal (evaluated for  $D_\ell = \ell(\ell + 1)C_\ell/2\pi$  at  $\ell = 80$ ) and the slope of angular power spectra from the PYSM synchrotron and thermal dust templates, evaluated at the SO-SAT sky patch. We obtain the following values ( $d$ : thermal dust at 353 GHz;  $s$ : synchrotron at 23 GHz):  $A_{EE}^d = 56 \mu\text{K}_{\text{CMB}}^2$ ,  $A_{BB}^d = 28 \mu\text{K}_{\text{CMB}}^2$ ,  $\alpha_{EE}^d = -0.32$ ,  $\alpha_{BB}^d = -0.16$ ;  $A_{EE}^s = 9 \mu\text{K}_{\text{CMB}}^2$ ,  $A_{BB}^s = 1.6 \mu\text{K}_{\text{CMB}}^2$ ,  $\alpha_{EE}^s = -0.7$ ,  $\alpha_{BB}^s = -0.93$ . This model assumes the frequency scaling of the maps across the SO channels to be a modified black body for thermal dust emission, with fixed spectral parameters  $\beta_d = 1.54$  and  $T_d = 20$  K, and a power law for synchrotron with fixed  $\beta_s = -3$  (in antenna temperature units).

**d0s0 model.** In this case, multi-frequency maps are taken from the d0s0 PYSM model. This model includes templates for thermal dust emission coming from Planck high frequency observations and from WMAP 23 GHz maps from synchrotron



**Table 4.** Mean  $r$  and (16, 84)% credible interval from 500 simulations, as inferred by three pipelines (and two extensions) on four foreground models and four noise cases, two of which are shown in Fig. 8 (no delensing is assumed in these fiducial results).

Noise	FG model	$10^3 \times (r \pm \sigma(r))$				
		Pipeline A	+ moments	Pipeline B	Pipeline C	+dust marg.
Goal rms, optimistic $1/f$	Gaussian	$-0.1 \pm 2.1$	$0.0 \pm 2.8$	$0.6 \pm 2.6^{(\dagger)}$	$1.6 \pm 2.7$	$-1.8 \pm 4.4$
	d0s0	$-0.4 \pm 2.1$	$-0.5 \pm 2.7$	$-0.1 \pm 2.1$	$0.5 \pm 2.3$	$-1.7 \pm 3.5$
	d1s1	$1.8 \pm 2.1$	$-0.2 \pm 2.8$	$2.1 \pm 2.1$	$2.6 \pm 2.3$	$0.0 \pm 3.3$
	dmsm	$3.9 \pm 2.1$	$0.3 \pm 2.7$	$3.8 \pm 2.1$	$5.3 \pm 2.4$	$0.2 \pm 3.0$
Goal rms, pessimistic $1/f$	Gaussian	$-0.2 \pm 2.5$	$-0.1 \pm 2.7$	$1.1 \pm 3.3^{(\dagger)}$	$0.9 \pm 2.1$	$-0.9 \pm 5.3$
	d0s0	$-0.6 \pm 2.5$	$-0.5 \pm 2.8$	$-0.5 \pm 2.8$	$0.1 \pm 2.5$	$-0.9 \pm 4.0$
	d1s1	$1.3 \pm 2.5$	$0.1 \pm 3.0$	$1.2 \pm 2.8$	$3.4 \pm 3.1$	$-0.0 \pm 3.9$
	dmsm	$3.2 \pm 2.6$	$0.3 \pm 3.9$	$2.1 \pm 2.8$	$5.5 \pm 2.4$	$0.6 \pm 4.2$
Baseline rms, optimistic $1/f$	Gaussian	$-0.1 \pm 2.6$	$-0.3 \pm 3.3$	$0.5 \pm 3.3^{(\dagger)}$	$0.5 \pm 3.2$	$-1.9 \pm 5.9$
	d0s0	$-0.4 \pm 2.6$	$-0.3 \pm 3.3$	$-0.9 \pm 2.7$	$0.7 \pm 2.9$	$-1.8 \pm 4.4$
	d1s1	$1.7 \pm 2.6$	$-0.2 \pm 3.4$	$1.0 \pm 2.7$	$1.8 \pm 2.7$	$-0.8 \pm 4.8$
	dmsm	$3.9 \pm 2.6$	$0.3 \pm 3.5$	$2.5 \pm 2.7$	$5.5 \pm 3.2$	$0.4 \pm 5.0$
Baseline rms, pessimistic $1/f$	Gaussian	$-0.3 \pm 3.4$	$-0.3 \pm 3.8$	$1.6 \pm 4.1^{(\dagger)}$	$0.0 \pm 2.9$	$-1.6 \pm 5.3$
	d0s0	$-0.7 \pm 3.4$	$-0.06 \pm 3.9$	$-0.6 \pm 3.6$	$1.1 \pm 3.2$	$-1.1 \pm 5.3$
	d1s1	$1.1 \pm 3.4$	$-0.6 \pm 4.0$	$0.5 \pm 3.6$	$3.8 \pm 3.2$	$-1.2 \pm 5.3$
	dmsm	$2.8 \pm 3.4$	$-0.6 \pm 4.0$	$1.2 \pm 3.6$	$6.0 \pm 3.1$	$-0.5 \pm 5.1$

**Notes.** <sup>(†)</sup>These results are calculated on a smaller, more homogeneous mask, shown in Fig. B.1. This is explained in Appendix B.

radiation. SEDs are considered to be uniform across the sky with the same values of the spectral parameters used for the Gaussian simulations.

**d1s1 model.** This model uses the same foreground amplitude templates as d0s0, but with the inclusion of spatial variability for spectral parameters, as described in Thorne et al. (2017).

**dmsm model.** This model represents a modification of the d1s1 spatial variation of spectral parameters. For thermal dust we smoothed the  $\beta_d$  and  $T_d$  templates at an angular resolution of 2 degrees, in order to down-weight the contribution of instrumental noise fluctuations in the original PYSM maps. For synchrotron emission we modified the  $\beta_s$  PYSM in order to account for the additional information coming from the analysis of S-PASS data at 2.3 GHz (see Krachmalnicoff et al. 2018). In particular S-PASS data show that the synchrotron spectral index presents enhanced variations with respect to the PYSM template. We therefore multiplied the fluctuations in the  $\beta_s$  map by a factor 1.6 to take into consideration larger variations. Moreover, we added small scale fluctuations (with a minimum angular resolution of 2 degrees), as Gaussian realization of a power-law power spectrum with slope  $-2.6$  (see Fig. 11 in Krachmalnicoff et al. 2018).

We note that this set of foreground models generalizes the what was done SO Collaboration (2019), since it includes the d1s1 model, used for large-scale  $B$ -mode forecasts in that earlier analysis. As for the CMB simulations, the multi-frequency foreground maps at the SO reference frequencies were convolved with Gaussian beams, to reach the expected angular resolution. We assumed delta-like frequency bandpasses in order to accelerate the production of these simulations, although all pipelines are able to handle finite bandpasses. Therefore this approximation should not impact the performance of any of the pipelines presented here.

## 4. Results and discussion

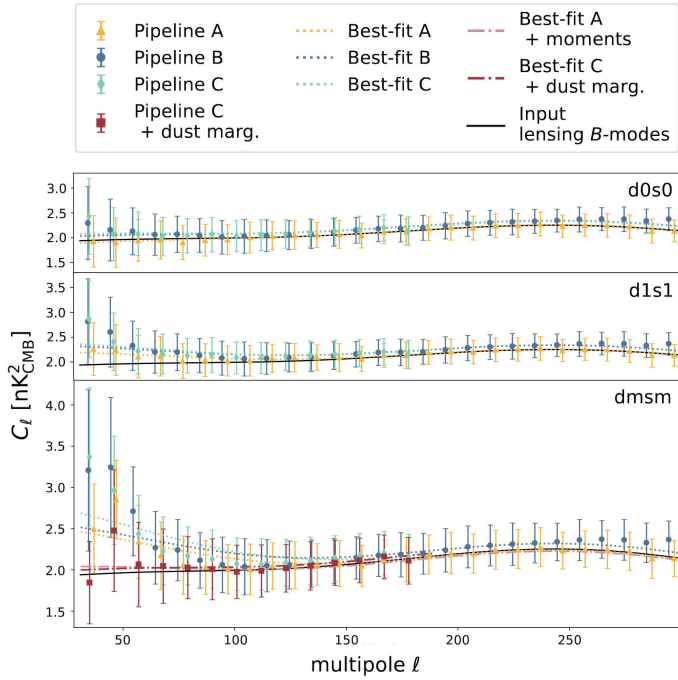
Simulations were generated for four different noise and foreground models, respectively, (see Sect. 3), for a total of 16 dif-

ferent foreground-noise combinations. For the main analysis, we consider a fiducial CMB model with  $A_{\text{lens}} = 1$  (no delensing) and  $r = 0$  (no primordial tensor fluctuations). In addition, we explored three departures from the fiducial CMB model, with input parameters ( $A_{\text{lens}} = 0.5, r = 0$ ), ( $A_{\text{lens}} = 1, r = 0.01$ ), and ( $A_{\text{lens}} = 0.5, r = 0.01$ ). Here we report the results found for all these cases.

### 4.1. Power spectra

Let us start by examining the CMB power spectrum products. Pipelines B and C produce CMB-only maps and base their inference of  $r$  on the resulting power spectra, whereas pipeline A works directly with the cross-frequency power spectra of the original multi-frequency maps. Nevertheless, CMB power spectra are an important data product that every pipeline should be able to provide. Following the methods presented in Dunkley et al. (2013) and Planck Collaboration XI (2016), Planck Collaboration V (2020), we use a modified version of pipeline A that retrieves CMB-only bandpowers from multi-frequency power spectra, marginalizing over foregrounds with an MCMC sampler as presented in Sect. 2.1. We note that this method, originally developed for high- $\ell$  CMB science, is applicable since we are in the Gaussian likelihood regime. By reinserting this cleaned CMB spectrum into a Gaussian likelihood with parameters ( $r, A_{\text{lens}}$ ), we obtain constraints that are consistent with the results shown in Table 4.

Figure 7 shows the CMB power spectra for the three complex foreground simulations d0s0, d1s1, and dmsm (upper, middle, and lower panel, respectively) while considering the goal-optimistic noise scenario. The various markers with error bars denote the measured CMB power spectra and their  $1\sigma$  standard deviation across 500 simulations, while the black solid line denotes the input CMB power spectrum. Results are shown in gold triangles, blue circles, turquoise diamonds for pipeline A, B, and C respectively. The dotted lines show the best-fit CMB model for the three nominal pipelines (using the same color scheme). Only in the dmsm foreground scenario, which is the most complex considered here, we also show the results from

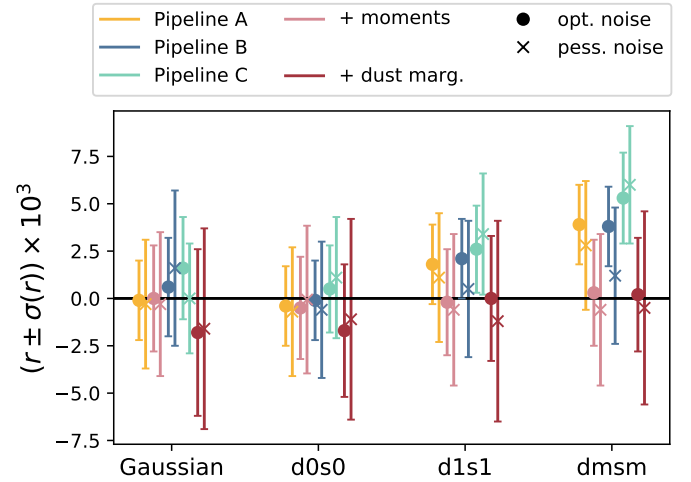


**Fig. 7.** CMB-only power spectra resulting from component separation with pipelines A, B, and C. We show non-Gaussian foreground scenarios d0s0 (top panel), d1s1 (middle panel), and dmsm (bottom panel) and consider the goal-optimistic noise scenario. The different colored markers with error bars show the mean of 500 simulations and the scatter between them (corresponding to the statistical uncertainties of a single realization). The dotted lines in the corresponding colors indicate the best-fit power spectrum model. In the dmsm case, we show the extended pipeline results from A + moments and C + dust marginalization with the best-fit models shown as dot-dashed lines. The black solid line is the input CMB model containing lensing  $B$ -modes only. We stress that pipeline C only considers multipoles up to  $\ell = 180$  in the power spectrum likelihood.

pipeline C + dust marginalization (dark red squares with error bars), and the best-fit CMB power spectrum from A + moments (pink dot-dashed line) and C + dust marginalization (dark red dot-dashed line).

For the nominal pipelines (A, B, and C) without extensions, the measured power spectra display a deviation from the input CMB at low multipoles, increasing with rising foreground complexity. For dmsm at multipoles  $\lesssim 50$ , this bias amounts to about  $1.5\sigma$  and goes down to less than  $0.5\sigma$  at  $80 \lesssim \ell \lesssim 250$ . The three pipelines agree reasonably well, while pipeline A appears slightly less biased for the lowest multipoles. Pipelines B and C show an additional mild excess of power in their highest multipole bins, with a  $< 0.3\sigma$  increase in pipeline C for  $130 \lesssim \ell \lesssim 170$  and up to  $1\sigma$  for the highest multipole ( $\ell = 297$ ) in pipeline B. This might indicate power leakage from the multiple operations on map resolutions implemented in pipelines B and C. In pipeline B, these systematics could come from first deconvolving the multi-frequency maps and then convolving them with a common beam in order to bring them to a common resolution, whereas in pipeline C, the leakage is likely due to the linear combination of the multi-resolution frequency maps following Eq. (9). Other multipole powers lie within the  $1\sigma$  standard deviation from simulations for all three pipelines.

Both extensions, A + moments and C + dust marginalization, lead to an unbiased CMB power spectrum model, as shown by the pink and dark red dot-dashed lines and the square mark-



**Fig. 8.** Mean  $r$  with (16, 84)% credible interval from 500 simulations. We apply the three nominal component separation pipelines (plus extensions) to simulations with four foreground scenarios of increasing complexity. We assume a fiducial cosmology with  $r = 0$  and  $A_{\text{lens}} = 1$ , inhomogeneous noise with goal sensitivity and optimistic  $1/f$  noise component (dot markers), and inhomogeneous noise with baseline sensitivity and pessimistic  $1/f$  noise component (cross markers). We note that the NILC results for Gaussian foregrounds are based on a smaller sky mask, see Appendix B.

ers in the lower panel of Fig. 7. In the case of pipelines B and C, comparing the best-fit models obtained from the measured power spectra to the input CMB model, we find sub-sigma bias for all bins with  $\ell > 100$ . We show, however, that the ability to marginalize over additional foreground residuals (e.g. the dust-template marginalization in pipeline C) is able to reduce this bias on all scales, at the cost of increased uncertainties. Implementing this capability in the blind NILC pipeline B would likely allow to reduce the bias that we see.

The SO-SATs are expected to constrain the amplitude of CMB lensing  $B$ -modes to an unprecedented precision. As can be seen from Fig. 7, individual, cleaned CMB bandpowers without delensing at multipoles  $\ell \gtrsim 150$  achieve a signal-to-noise ratio of about 10, accounting for a combined precision on the lensing amplitude of  $\sigma(A_{\text{lens}}) \lesssim 0.03$  when considering multipoles up to  $\ell_{\text{max}} = 300$ . As we show in the following section, this is consistent with the inference results obtained by pipelines A and B.

#### 4.2. Constraints on $r$

Having presented the results on the CMB power spectra, let us now examine the final constraints on  $r$  obtained by each pipeline applied to 500 simulations. These results are summarized in Fig. 8 and Table 4. Figure 8 shows the mean  $r$  and (16, 84)% credible intervals found by each pipeline as a function of the input foreground model (labels on the  $x$  axis). Results are shown for five pipeline setups: pipeline A using the  $C_\ell$ -fiducial model (red), pipeline A using the  $C_\ell$ -moments model (yellow), pipeline B (blue), pipeline C (green), and pipeline C including the marginalization over the dust amplitude parameter (cyan). For each pipeline, we show two points with error bars. The dot markers and smaller error bars correspond to the results found in the best-case instrument scenario (goal noise level, optimistic  $1/f$  component), while the cross markers and larger error bars correspond to the baseline noise level and pessimistic  $1/f$  component. The quantitative results are reported in Table 4.

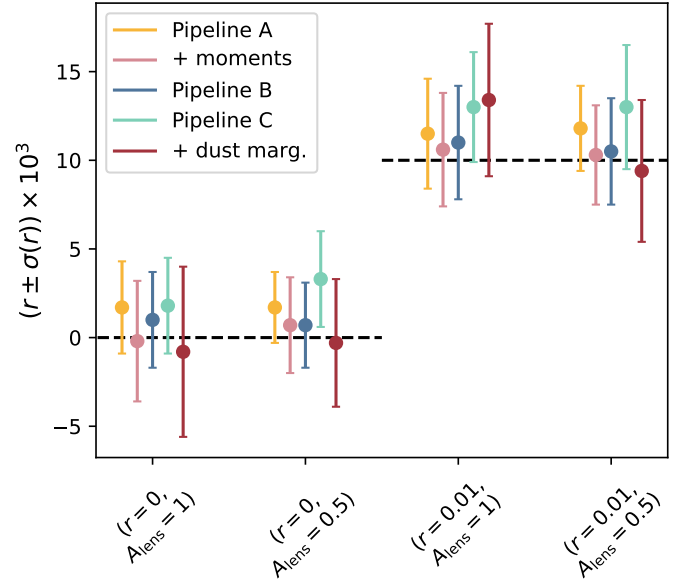
We start by discussing the nominal pipelines A, B, and C without considering any extensions. We find that for the simpler Gaussian and `d0s0` foregrounds, the nominal pipelines obtain unbiased results, as expected. Pipeline B shows a slight positive bias for Gaussian foregrounds, in combination with inhomogeneous noise only. This bias is absent for homogeneous noise and can be traced back to the pixel covariance matrix used to construct the NILC weights. We discuss this in more detail in Appendix B. For now, we show the results using a smaller, more homogeneously weighted mask. We stress that these results, marked with a †, are not comparable to the rest in Table 4, since they are calculated on a different mask. The more complex `d1s1` foregrounds lead to a  $\sim 1\sigma$  bias in the goal and optimistic noise scenario. The `dmsm` foregrounds lead to a noticeable increase of the bias of up to  $\sim 2\sigma$ , seen with pipeline C in all noise scenarios and with pipeline A in the goal-optimistic case, and slightly less with pipeline B. The modifications introduced in the `dmsm` foreground model include a larger spatial variation in the synchrotron spectral index  $\beta_s$  with respect to `d1s1`, and are a plausible reason for the increased bias on  $r$ .

Remarkably, we find that, in their simplest incarnation, all pipelines achieve comparable statistical uncertainty on  $r$ , ranging from  $\sigma(r) \simeq 2.1 \times 10^{-3}$  to  $\sigma(r) \simeq 3.6 \times 10^{-3}$  (a 70% increase), depending on the noise model. Changing between the goal and baseline white noise levels results in an increase of  $\sigma(r)$  of  $\sim 20$ – $30\%$ . Changing between the optimistic and pessimistic  $1/f$  noise has a similar effect on the results from pipelines A and B, although  $\sigma(r)$  does not increase by more than 10% when changing to pessimistic  $1/f$  noise for pipeline C. These results are in reasonable agreement with the forecasts presented in SO Collaboration (2019).

Let us now discuss the pipeline extensions A + moments and C + dust marginalization. Notably, in all noise and foreground scenarios, the two extensions are able to reduce the bias on  $r$  to below  $1\sigma$ . For the Gaussian and `d0s0` foregrounds, we consistently observe a small negative bias (at the  $\sim 0.1\sigma$  level for A + moments and  $<0.5\sigma$  for C + dust marginalization). This bias may be caused by the introduction of extra parameters that are prior dominated, like the dust template’s amplitude in the absence of residual dust contamination, or the moment parameters in the absence of varying spectral indices of foregrounds. If those extra parameters are weakly degenerate with the tensor-to-scalar ratio, the marginal  $r$  posterior will shift according to the choice of the prior on the extra parameters. The observed shifts in the tensor-to-scalar ratio and their possible relation with these volume effects will be investigated in a future work. For the more complex `d1s1` and `dmsm`, both pipeline extensions effectively remove the bias observed in the nominal pipelines, achieving a  $\sim 0.5\sigma$  bias and lower.

The statistical uncertainty  $\sigma(r)$  increases for both pipeline extensions, although by largely different factors. While C + dust marginalization yields  $\sigma(r)$  between  $3.0 \times 10^{-3}$  and  $5.9 \times 10^{-3}$ , the loss in precision for A + moments is significantly smaller, with  $\sigma(r)$  varying between  $2.7 \times 10^{-3}$  and  $4.0 \times 10^{-3}$  depending on the noise scenario, an average increase of  $\sim 25\%$  compared to pipeline A. In any case, within the assumptions made regarding the SO noise properties, it should be possible to detect a primordial  $B$ -mode signal with  $r = 0.01$  at the  $2$ – $3\sigma$  level with no delensing. The impact of other effects, such as time domain filtering or anisotropic noise may affect these forecasts, and will be studied in more detail in the future.

We repeated this analysis for input CMB maps generated assuming either  $r = 0$  or  $0.01$ , and either  $A_{\text{lens}} = 0.5$  or  $1$ . For simplicity, in these cases we considered only the baseline



**Fig. 9.** Mean  $r$  and (16, 84)% credible interval from 500 simulations, using the three nominal pipelines plus extensions. We assume input models including primordial  $B$ -modes and 50% delensing efficiency, the SO baseline noise level with optimistic  $1/f$  component, and the `d1s1` foreground template.

white noise level with optimistic  $1/f$  noise and the moderately complex `d1s1` foreground model. We show results in Fig. 9 and Table 5. A 50% delensing efficiency results in a reduction in the final  $\sigma(r)$  by 25–30% for pipelines A and B,  $\sim 10$ – $20\%$  for A + moments, and 0–33% for C + dust marginalization. The presence of primordial  $B$ -modes with a detectable amplitude increases the contribution from cosmic variance to the error budget, with  $\sigma(r)$  growing by up to 40% if  $r = 0.01$ , in agreement with theoretical expectations. Using C + dust marginalization and considering no delensing, we even find  $\sigma(r)$  decreasing, hinting at the possible breaking of the degeneracy between  $r$  and  $A_{\text{dust}}$ . We conclude that all pipelines are able to detect the  $r = 0.01$  signal at the level of  $\sim 3\sigma$ . As before, we observe a  $0.5$ – $1.2\sigma$  bias on the recovered  $r$  that is eliminated by both the moment expansion method and the dust marginalization method.

Finally, we explored how cosmological constraints and the pipelines’ performances are affected by noise inhomogeneity resulting from weighting the noise pixels according to the SO-SAT hits map. The geographical location of SO and the size of the SAT field of view constrain possible scanning strategies. In particular, SO must target a patch that has a relatively large sky fraction  $f_{\text{sky}} \sim 0.15$  and is surrounded by a  $\sim 10$  degree wide boundary with significantly higher noise (see hits map in Fig. 2). The lower panel of Fig. 10 shows the ratio between the values of  $\sigma(r)$  found using inhomogeneous noise realizations and those with homogeneous noise in the baseline-optimistic noise model with `d0s0` foregrounds, averaged over 500 simulations. We see that for all pipeline scenarios,  $\sigma(r)$  increases by  $\sim 30\%$  due to the noise inhomogeneity.

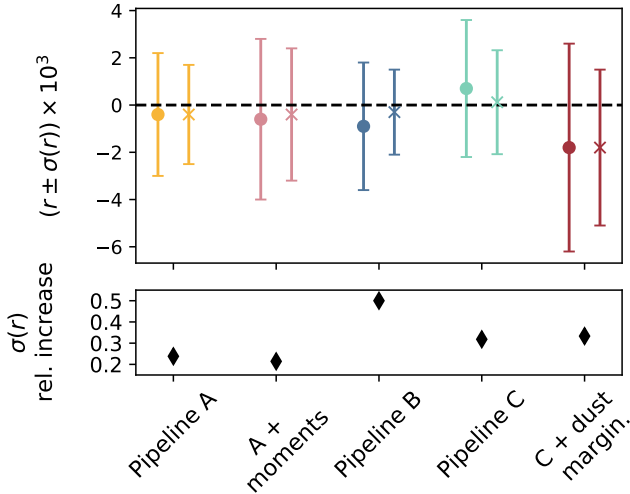
### 4.3. Channel weights

Our three baseline pipelines differ fundamentally in how they separate the sky components. One common feature among all pipelines is the use of six frequency channels to distinguish

**Table 5.** Mean  $r$  with (16, 84)% credible interval from 500 simulations, using the three nominal pipelines with extensions.

Input CMB	$10^3 \times (r \pm \sigma(r))$				
	Pipeline A	+ moments	Pipeline B	Pipeline C	+dust marg.
$(r = 0, A_{\text{lens}} = 1)$	$1.7 \pm 2.6$	$-0.2 \pm 3.4$	$1.0 \pm 2.7$	$1.8 \pm 2.7$	$-0.8 \pm 4.8$
$(r = 0, A_{\text{lens}} = 0.5)$	$1.7 \pm 2.0$	$0.7 \pm 2.7$	$0.7 \pm 2.4$	$3.3 \pm 2.7$	$-0.3 \pm 3.6$
$(r = 0.01, A_{\text{lens}} = 1)$	$11.5 \pm 3.1$	$10.6 \pm 3.2$	$11.0 \pm 3.2$	$13.0 \pm 3.1$	$13.4 \pm 4.3$
$(r = 0.01, A_{\text{lens}} = 0.5)$	$11.8 \pm 2.4$	$10.3 \pm 2.8$	$10.5 \pm 3.0$	$13.0 \pm 3.5$	$9.4 \pm 4.0$

**Notes.** Our input models contain primordial  $B$ -modes of an amplitude  $r = 0.01$  and 50% delensing efficiency. We assume the SO baseline noise level with optimistic  $1/f$  component and d1s1 foregrounds, see Fig. 9.



**Fig. 10.** Mean  $r$  with (16, 84)% credible intervals from 500 simulations, applying the three nominal component separation pipelines plus extensions. We assume the d0s0 foreground scenario with baseline white noise level and optimistic  $1/f$  component. Cross markers with smaller error bars correspond to homogeneous noise across the SAT field of view and dot markers with larger error bars correspond to inhomogeneous noise. The relative increase in  $\sigma(r)$  between both is shown in the bottom panel.

components by means of their different SEDs. In Fig. 11 we visualize the channel weights as a function of the band center frequency, showing the pipelines in three vertically stacked panels. In the upper panel, we show the effective weights for the CMB applied to the noise-debiased raw power spectra used by pipeline A, distinguishing between weights for each harmonic bin:

$$\mathbf{w}_\ell^T = \frac{\mathbf{a}^T \hat{\mathbf{C}}_\ell^{-1}}{\mathbf{a}^T \hat{\mathbf{C}}_\ell^{-1} \mathbf{a}}. \quad (15)$$

Here,  $\hat{\mathbf{C}}_\ell$  is the  $6 \times 6$  matrix of raw cross-frequency power spectra from noisy sky maps,  $\mathbf{a}$  is a vector of length six, filled with ones. This is equivalent to the weights employed by the SMICA component separation method (Cardoso et al. 2008) and by ILC as explained in Sect. 2.2. The middle panel shows the pixel-averaged NILC weights for the five needlet windows (Fig. 4) used in pipeline B. In the lower panel, we show the CMB weights calculated with the map-based component separation, pipeline C, averaged over the observed pixels to yield an array of six numbers. We averaged all channel weights over 100 simulations containing CMB ( $r = 0, A_{\text{lens}} = 1$ ), d1s1 foregrounds, and one of two noise models: goal white noise with optimistic  $1/f$  noise is shown as dashed lines, whereas baseline white noise

with pessimistic  $1/f$  noise is shown as solid lines. Moreover, the gray shaded areas quantify the  $1-\sigma$  uncertainty region of these weights in the baseline-pessimistic case estimated from 100 simulations. We see from Fig. 11 that the average channel weights agree well between pipelines A, B, and C. Mid-frequency channels at 93 and 145 GHz are assigned positive CMB weights throughout all pipelines, while high- and low-frequency channels tend to be suppressed owing to a larger dust and synchrotron contamination, respectively. More specifically, the 280 GHz channel is given negative weight in all pipelines, while average weights at 27, 39, and 225 GHz are negative with pipeline C and either positive or negative in pipelines A and B, depending on the angular scale. The CMB channel weight tends to consistently increase for pipelines A and B as a function of multipole, a fact well exemplified by NILC at 225 GHz, matching the expectation that the CMB lensing signal becomes more important at high  $\ell$ . Overall, Fig. 11 illustrates that foregrounds at low and high frequencies are consistently subtracted by the three component separation pipelines, with the expected scale dependence in pipelines A and B. Moreover, at every frequency, the channel weights are non-negligible and of similar size across the pipelines, meaning that all channels give a relevant contribution to component separation for all pipelines.

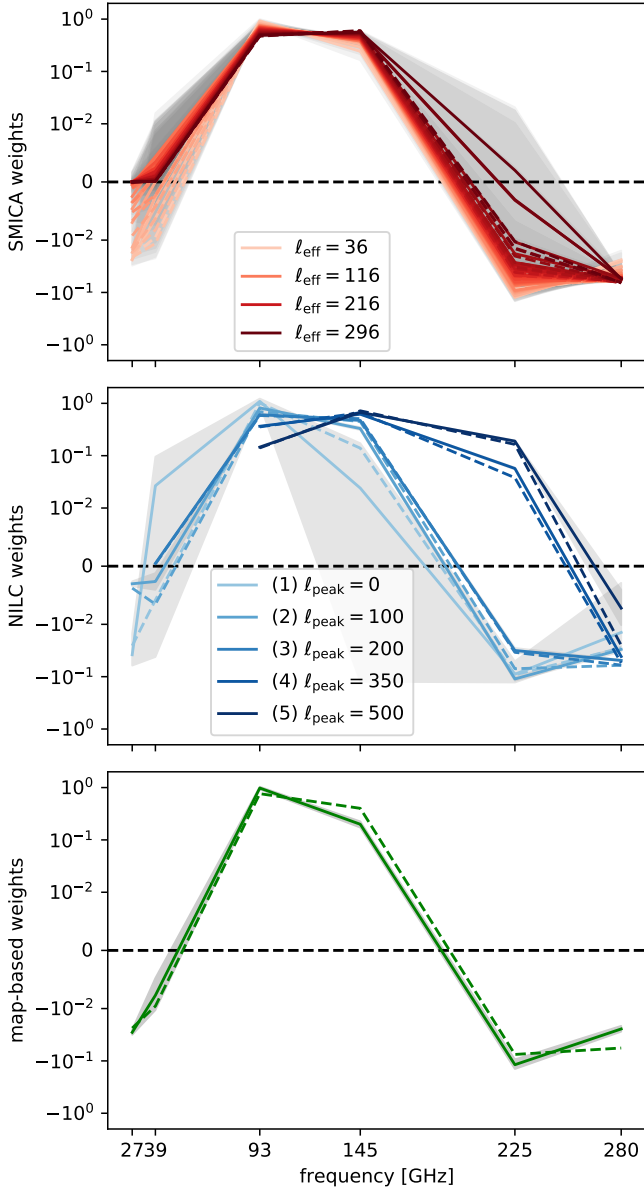
#### 4.4. More complex foregrounds: d10s5

During the completion of this paper, the new PYSM3 Galactic foreground models<sup>6</sup> were made publicly available. In particular, in these new models, templates for polarized thermal dust and synchrotron radiation were updated including the following changes:

1. Large-scale thermal dust emission is based on the GNILC maps (Planck Collaboration IV 2020), which present a lower contamination from CIB emission with respect to the d1 model, based on Commander templates.
2. For both thermal dust and synchrotron radiation, small scale structures are added by modifying the logarithm of the polarization fraction tensor<sup>7</sup>.
3. Thermal dust spectral parameters are based on GNILC products, with larger variation of  $\beta_d$  and  $T_d$  parameter at low resolution compared to the d1 model. Small-scale structure is also added as Gaussian realizations of power-law power spectra.
4. The new template for  $\beta_s$ , includes information from the analysis of S-PASS data (Krachmalnicoff et al. 2018), in a similar way as the one of the sm model adopted in this work. In addition, small-scale structures are present at sub-degree angular scales.

<sup>6</sup> See [pysm3.readthedocs.io/en/latest](https://pysm3.readthedocs.io/en/latest)

<sup>7</sup> See [pysm3.readthedocs.io/en/latest/preprocess-templates](https://pysm3.readthedocs.io/en/latest/preprocess-templates)



**Fig. 11.** Channel-specific weights associated with the component-separated CMB for the three nominal pipelines. We show the SMICA weights for 27 different  $\ell$ -bins calculated from raw, noisy  $C_\ell$ s (pipeline A, upper panel), pixel-averaged NILC weights for five needlet windows (pipeline B, middle panel), and pixel-averaged weights from parametric map-based component separation (pipeline C, lower panel). Weights are averaged over 100 simulations, shown are goal white + optimistic  $1/f$  noise (dashed lines) as well as baseline white + pessimistic  $1/f$  noise (solid lines). The semitransparent gray areas represent the channel weights'  $1\text{-}\sigma$  standard deviation across 100 simulations, covering baseline + pessimistic noise.

These modifications are encoded in the models called **d10** and **s5** in the updated version of **PySM**. Although these models are still to be considered preliminary, both in terms of their implementation details in **PySM**<sup>8</sup> and in general, being based on datasets that may not fully involve the unknown level of foreground complexity, we decided to dedicate an extra section

<sup>8</sup> The **PySM** library is currently under development, with beta versions including minor modifications of the foreground templates being realized regularly. In this part of our analysis we make use of **PySM** v3.4.0B3.

to their analysis. For computational speed, we ran the five pipeline set-ups on a reduced set of 100 simulations containing the new **d10s5** foregrounds template, CMB with a standard cosmology ( $r = 0$ ,  $A_{\text{lens}} = 1$ ) and inhomogeneous noise in the goal-optimistic scenario. The resulting marginalized posterior mean and (16, 84)% credible intervals on  $r$ , averaged over 100 simulations, are:

$$\begin{aligned}
 r \times 10^3 &= 19.2 \pm 1.9 && \text{(pipeline A)} \\
 r \times 10^3 &= 2.7 \pm 2.8 && \text{(A + moments)} \\
 r \times 10^3 &= 15.8 \pm 1.9 && \text{(pipeline B)} \\
 r \times 10^3 &= 22.0 \pm 2.6 && \text{(pipeline C)} \\
 r \times 10^3 &= -1.5 \pm 5.1 && \text{(C + dust marg.)}
 \end{aligned} \tag{16}$$

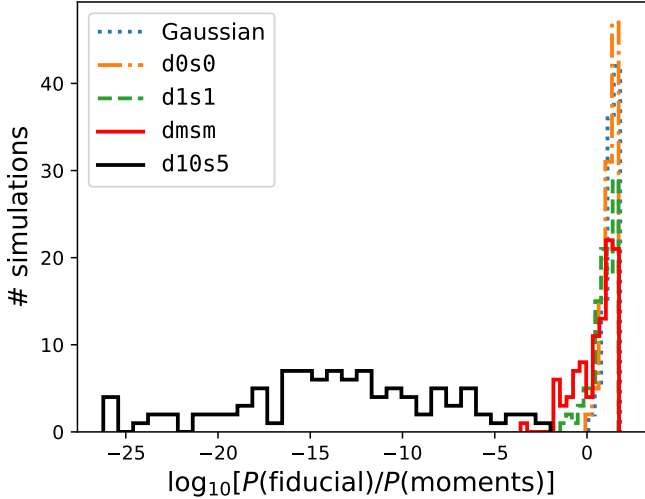
We note that the respective bias obtained with pipelines A, B, and C are at 10, 8, and  $8\sigma$ , at least quadrupling the bias of the **dmsm** foreground model. Crucially, this bias is reduced to less than  $1\sigma$  with the A + moments pipeline, with 45% increase in  $\sigma(r)$  compared to pipeline A, and  $0.3\sigma$  with the C + dust-marginalization pipeline, with a 95% increase in  $\sigma(r)$  compared to pipeline C. This makes A + moments the unbiased method with the lowest statistical error.

The  $C_\ell$ -fiducial model achieves minimum  $\chi^2$  values of  $601 \pm 41$ . Although this is an increase of  $\Delta\chi^2 \sim 30$  with respect to the less complex foreground simulations (see Appendix A), the associated probability to exceed (PTE) is 0.10 (assuming our null distribution is a  $\chi^2$  with  $N_{\text{data}} - N_{\text{parameters}} = 558$  degrees of freedom), and therefore it would not be possible to identify the presence of a foreground bias by virtue of the model providing a bad fit to the data. The minimum  $\chi^2$  values we find also confirm that the covariance matrix calculated from Gaussian simulations is still appropriate for the non-Gaussian **d10s5** template. On the other hand, A + moments achieves minimum  $\chi^2$  values of  $537 \pm 33$ , which is about 4% lower than for less complex foreground simulations, indicating an improved fitting accuracy.

As shown in Fig. 12, the relative model odds between  $C_\ell$ -fiducial and  $C_\ell$ -moments (see Appendix A for more details) vary between  $10^{-26}$  and  $10^{-2}$ , clearly favoring  $C_\ell$ -moments. Out of 100 **d10s5** simulations, 99 yield model odds below 1% and 78 below  $10^{-5}$ . As opposed to the less complex foreground simulations (**d0s0**, **d1s1**, and **dmsm**), **d10s5** gives strong preference to using the moment expansion in the power spectrum model. We note that the AIC-based model odds are computed from the differences of  $\chi^2$  values that stem from the same simulation seed and are therefore insensitive to bias from noise and cosmic variance. This explains why AIC odds are the more powerful model comparison tool when compared with the  $\chi^2$  analysis presented above.

These results consider only the most optimistic noise scenario. Other cases would likely lead to larger uncertainty and, as a consequence, lower relative biases. In this regard, it is highly encouraging to see two pipeline extensions being able to robustly separate the cosmological signal from Galactic synchrotron and dust emission with this high-level complexity. This highlights the importance of accounting for and marginalizing over residual foreground contamination due to frequency decorrelation for the level of sensitivity that SO and other next-generation observatories will achieve.

The contrast between the results obtained on the **dmsm** and **d10s5** simulations gives us an opportunity to reflect on the strategy one should follow when determining the fiducial component separation method to use in primordial  $B$ -mode searches.



**Fig. 12.** Empirical distribution of the AIC-based relative model odds between the  $C_\ell$ -fiducial and the  $C_\ell$ -moments model from 100 simulations. We compare five different Galactic foreground templates, including the PYSM foreground model d10s5. Negative values indicate preference for the moments model. We find strong preference for the  $C_\ell$ -moments model in the d10s5 foreground scenario, and only then.

Although the dmsm model leads to a  $\sim 2\sigma$  bias on  $r$  under the simplest component separation algorithms, simple model-selection metrics are not able to provide significant evidence that a more sophisticated modeling of foregrounds is needed. The situation changes with d10s5. A conservative approach is therefore to select the level of complexity needed for component separation by ensuring that unbiased constraints are obtained for all existing foreground models consistent with currently available data. The analysis methods passing this test can then form the basis for the fiducial  $B$ -mode constraints. Alternative results can then be obtained with less conservative component separation techniques, but their goodness of fit (or any similar model selection metric) should be compared with that of the fiducial methods. These results should also be accompanied by a comprehensive set of robustness tests able to identify signatures of foreground contamination in the data. This will form the basis of a future work. In a follow-up paper, we will also explore the new set of complex PYSM3 foreground templates in more detail.

## 5. Conclusions

In this paper, we present three different component separation pipelines designed to place constraints on the amplitude of cosmological  $B$ -modes on polarized maps of the SO Small Aperture Telescopes. The pipelines are based on multi-frequency  $C_\ell$  parametric cleaning (Pipeline A), blind Needlet ILC cleaning (Pipeline B), and map-based parametric cleaning (Pipeline C). We also introduce extensions of pipelines A and C that marginalize over additional residual foreground contamination, using a moment expansion or a dust power spectrum template, respectively. We tested and compared their performance on a set of simulated maps containing lensing  $B$ -modes with different scenarios of instrumental noise and Galactic foreground complexity. The presence of additional instrumental complexity, such as time-domain filtering, or anisotropic noise, are likely to affect our results. The impact of these effects will be more thoroughly studied in future work.

We find the inferred uncertainty on the tensor-to-scalar ratio  $\sigma(r)$  to be compatible between the three pipelines. While the simpler foreground scenarios (Gaussian, d0s0) do not bias  $r$ , spectral index variations can cause an increased bias of  $1-2\sigma$  if left untreated, as seen with more complex foreground scenarios (d1s1, dmsm). Modeling and marginalizing over the spectral residuals is vital to obtain unbiased  $B$ -mode estimates. The extensions to pipelines A and C are able to yield unbiased estimates on all foreground scenarios, albeit with a respective increase in  $\sigma(r)$  by  $\sim 20\%$  (A + moments) and  $>30\%$  (C + dust marginalization). These results are in good agreement with the forecasts presented in SO Collaboration (2019).

After testing on simulations with an  $r = 0.01$  cosmology, we conclude that under realistic conditions and if the forecasted map noise levels and characteristics are achieved, SO should be able to detect a  $r = 0.01$  signal at  $\sim 2-3\sigma$  after five years of observation. Inhomogeneous noise from the SAT map-making scanning strategy brings about 30% increase in  $\sigma(r)$  as compared to homogeneous noise. Analyzing the per-channel weights for our pipelines, we find all frequency channels to be relevant for the CMB signal extraction and all pipelines to be in good agreement. These forecasts cover the nominal SO survey, and can be considered pessimistic in the light of prospective additional SATs that will further improve the sensitivity on large angular scales.

We also carried out a preliminary analysis of new, more complex, foreground models recently implemented in PYSM3, in particular the d10s5 foreground template. The much higher level of spatial SED variation allowed by this model leads to a drastic increase in the bias on  $r$  by up to  $10\sigma$ , when analyzed with the nominal pipelines A, B, and C. Fortunately, this bias can be reduced to below  $1\sigma$  when using A + moments and C + dust marginalization. These extensions lead to a 45% and 95% degradation of the error bars, respectively. Our results highlight the importance of marginalizing over residuals caused by frequency decorrelation for SO-like sensitivities. Although our analysis of d10s5 is less exhaustive than that of the other foreground models presented here, it is encouraging to confirm that we have the tools at hand to obtain robust, unbiased constraints on the tensor-to-scalar ratio in the presence of such complex Galactic foregrounds. In addition to the algorithmic improvements presented in this paper, the inclusion of external data sets such as FYST/CCAT-Prime (CCAT-Prime Collaboration 2022) may prove helpful at mitigating foregrounds.

In preparation for the data collected by SO in the near future, we will continue our investigations into Galactic foreground models with other levels of complexity as the field progresses. Nevertheless, the current work shows that the analysis pipelines in place for SO are able to obtain robust constraints on the amplitude of primordial  $B$  modes in the presence of Galactic foregrounds covering the full range of complexity envisaged by current, state-of-the-art models.

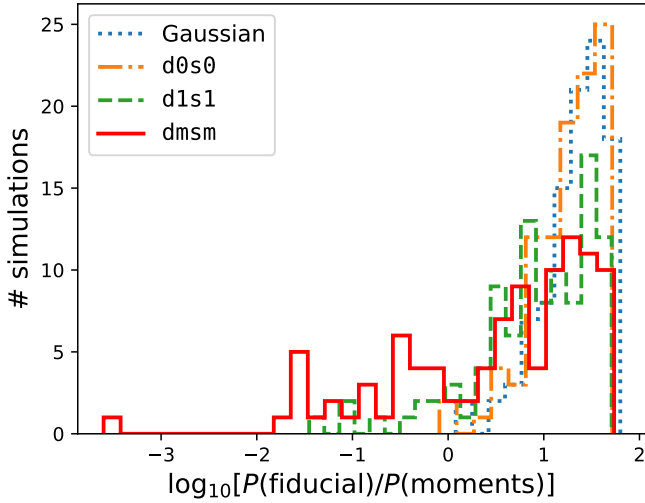
*Acknowledgements.* The authors would like to thank Ken Ganga, Arthur Kosowsky, and the anonymous referee for useful feedback. The group at SISSA acknowledges support from the COSMOS Network of the Italian Space Agency and the InDark Initiative of the National Institute for Nuclear Physics (INFN). KW is funded by a SISSA PhD fellowship. SA is funded by a Kavli/IPMU doctoral studentship. CHC acknowledges NSF award 1815887 and FONDECYT Postdoc fellowship 3220255. DA is supported by the Science and Technology Facilities Council through an Ernest Rutherford Fellowship, grant reference ST/P004474. This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (PI: Josquin Errard, Grant agreement No. 101044073). ABL is a BCCP fellow at UC Berkeley and Lawrence Berkeley National Laboratory. MLB acknowledges funding from UKRI and STFC (Grant awards ST/X006344/1 and ST/X006336/1). EC acknowledges

support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 849169). JC was furthermore supported by the ERC Consolidator Grant CMBSPEC (No. 725456) and the Royal Society as a Royal Society University Research Fellow at the University of Manchester, UK (No. URF/R/191023). GF acknowledges the support of the European Research Council under the Marie Skłodowska Curie actions through the Individual Global Fellowship No. 892401 PiCOGAMBAS. We acknowledge the use of CAMB (Lewis et al. 2000), healpy (Zonca et al. 2019), numpy (Harris 2020), matplotlib (Hunter 2007), emcee (Foreman-Mackey et al. 2013), and fgbuster (Errard & Stompor 2019; Puglisi et al. 2022) software packages.

## References

- Abazajian, K. N., Adshead, P., Ahmed, Z., et al. 2016, arXiv e-prints [arXiv:1610.02743]
- Abbott, L. F., & Wise, M. B. 1984, *Nucl. Phys. B*, 244, 541
- Abitbol, M. H., Alonso, D., Simon, S. M., et al. 2021, *JCAP*, 2021, 032
- Akaike, H. 1974, *IEEE Trans. Autom. Control*, 19, 716
- Alonso, D., Dunkley, J., Thorne, B., & Naess, S. 2017, *Phys. Rev. D*, 95, 043504
- Alonso, D., Sanchez, J., Slosar, A., & LSST Dark Energy Science Collaboration 2019, *MNRAS*, 484, 4127
- Armitage-Caplan, C., Dunkley, J., Eriksen, H. K., & Dickinson, C. 2012, *MNRAS*, 424, 1914
- Azzoni, S., Abitbol, M. H., Alonso, D., et al. 2021, *JCAP*, 2021, 047
- Basak, S., & Delabrouille, J. 2012, *MNRAS*, 419, 1163
- Basak, S., & Delabrouille, J. 2013, *MNRAS*, 435, 18
- Bennett, C. L., Hill, R. S., Hinshaw, G., et al. 2003, *ApJS*, 148, 97
- Betoule, M., Pierpaoli, E., Delabrouille, J., Le Jeune, M., & Cardoso, J. F. 2009, *A&A*, 503, 691
- BICEP2 Collaboration & Keck Array Collaboration 2016, *Phys. Rev. Lett.*, 116, 031302
- BICEP2 Collaboration & Keck Array Collaboration 2018, *Phys. Rev. Lett.*, 121, 221301
- BICEP/Keck Collaboration 2021, *Phys. Rev. Lett.*, 127, 151301
- Bonaldi, A., & Ricciardi, S. 2011, *MNRAS*, 414, 615
- Cardoso, J. F., Martin, M., Delabrouille, J., Betoule, M., & Patanchon, G. 2008, arXiv e-prints [arXiv:0803.1814]
- CCAT-Prime Collaboration (Aravena, M., et al.) 2022, *ApJS*, 264, 7
- Chluba, J., Hill, J. C., & Abitbol, M. H. 2017, *MNRAS*, 472, 1195
- CMB-S4 Collaboration 2022, *ApJ*, 926, 54
- Delabrouille, J., & Cardoso, J. F. 2007, International Summer School on Data Analysis in Cosmology (Valencia, Spain, cel-00162531.), 47
- Delabrouille, J., Cardoso, J. F., Le Jeune, M., et al. 2009, *A&A*, 493, 835
- Dunkley, J., Calabrese, E., Sievers, J., et al. 2013, *JCAP*, 07, 025
- Errard, J., & Stompor, R. 2012, *Phys. Rev. D*, 85, 083006
- Errard, J., & Stompor, R. 2019, *Phys. Rev. D*, 99, 043529
- Errard, J., Feeney, S. M., Peiris, H. V., & Jaffe, A. H. 2016, *JCAP*, 2016, 052
- Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, *PASP*, 125, 306
- Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, *ApJ*, 622, 759
- Grain, J., Tristram, M., & Stompor, R. 2009, *Phys. Rev. D*, 79, 123515
- Hamimeche, S., & Lewis, A. 2008, *Phys. Rev. D*, 77, 103013
- Harper, S. E., Dickinson, C., Barr, A., et al. 2022, *MNRAS*, 513, 5900
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Nature*, 585, 357
- Hazumi, M., Ade, P. A. R., Akiba, Y., et al. 2019, *J. Low Temp. Phys.*, 194, 443
- Hervías-Caimapo, C., Bonaldi, A., & Brown, M. L. 2017, *MNRAS*, 468, 4408
- Hervías-Caimapo, C., Bonaldi, A., Brown, M. L., & Huffenberger, K. M. 2022, *ApJ*, 924, 11
- Hui, H., Ade, P. A. R., Ahmed, Z., et al. 2018, *Proc. SPIE Int. Soc. Opt. Eng.*, 10708
- Hunter, J. D. 2007, *Comput. Sci. Eng.*, 9, 90
- Ijjas, A., & Steinhardt, P. J. 2018, *CQG*, 35, 135004
- Ijjas, A., & Steinhardt, P. J. 2019, *Phys. Lett. B*, 795, 666
- Kamionkowski, M., & Kovetz, E. D. 2016, *ARA&A*, 54, 227
- Kamionkowski, M., Kosowsky, A., & Stebbins, A. 1997, *Phys. Rev. Lett.*, 78, 2058
- Katayama, N., & Komatsu, E. 2011, *ApJ*, 737, 78
- Krachmalnicoff, N., Baccigalupi, C., Aumont, J., Bersanelli, M., & Mennella, A. 2016, *A&A*, 588, A65
- Krachmalnicoff, N., Carretti, E., Baccigalupi, C., et al. 2018, *A&A*, 618, A166
- Leach, S. M., Cardoso, J. F., Baccigalupi, C., et al. 2008, *A&A*, 491, 597
- Lewis, A., & Challinor, A. 2006, *Phys. Rep.*, 429, 1
- Lewis, A., Challinor, A., & Lasenby, A. 2000, *ApJ*, 538, 473
- LiteBIRD Collaboration 2022, *Prog. Theor. Exp. Phys.*, ptac150
- Maltoni, M., & Schwetz, T. 2003, *Phys. Rev. D*, 68, 033020
- Mangilli, A., Aumont, J., Rotti, A., et al. 2021, *A&A*, 647, A52
- Martin, J., Ringeval, C., Trotta, R., & Vennin, V. 2014a, *JCAP*, 2014, 039
- Martin, J., Ringeval, C., & Vennin, V. 2014b, *Phys. Dark Universe*, 5, 75
- Namikawa, T., Baleato Lizancos, A., Robertson, N., et al. 2022, *Phys. Rev. D*, 105, 023511
- Nash, S. G. 1984, *SIAM J. Numer. Anal.*, 21, 770
- Natoli, P., Ashdown, M., Banerji, R., et al. 2018, *JCAP*, 2018, 022
- Pearson, K. 1900, *London Edinburgh Dublin Philos. Mag. J. Sci.*, 50, 157
- Planck Collaboration X. 2016, *A&A*, 594, A10
- Planck Collaboration XI. 2016, *A&A*, 594, A11
- Planck Collaboration IV. 2020, *A&A*, 641, A4
- Planck Collaboration V. 2020, *A&A*, 641, A5
- Planck Collaboration VI. 2020, *A&A*, 641, A6
- Planck Collaboration X. 2020, *A&A*, 641, A10
- Planck Collaboration XI. 2020, *A&A*, 641, A11
- Planck Collaboration Int. XXX. 2016, *A&A*, 586, A133
- Puglisi, G., Mihaylov, G., Panopoulou, G. V., et al. 2022, *MNRAS*, 511, 2052
- Remazeilles, M., Delabrouille, J., & Cardoso, J.-F. 2011, *MNRAS*, 410, 2481
- Remazeilles, M., Dickinson, C., Eriksen, H. K. K., & Wehus, I. K. 2016, *MNRAS*, 458, 2032
- Remazeilles, M., Banday, A. J., Baccigalupi, C., et al. 2018a, *JCAP*, 2018, 023
- Remazeilles, M., Dickinson, C., Eriksen, H. K., & Wehus, I. K. 2018b, *MNRAS*, 474, 3889
- Remazeilles, M., Rotti, A., & Chluba, J. 2021, *MNRAS*, 503, 2478
- Seljak, U. 1997, *ApJ*, 482, 6
- Seljak, U., & Zaldarriaga, M. 1997, *Phys. Rev. Lett.*, 78, 2054
- Smith, K. M., & Zaldarriaga, M. 2007, *Phys. Rev. D*, 76, 043001
- SO Collaboration 2019, *JCAP*, 2019, 056
- Starobinskiĭ, A. A. 1979, *Sov. J. Exp. Theor. Phys. Lett.*, 30, 682
- Stompor, R., Errard, J., & Poletti, D. 2016, *Phys. Rev. D*, 94, 083526
- Stompor, R., Leach, S., Stivoli, F., & Baccigalupi, C. 2009, *MNRAS*, 392, 216
- Tegmark, M. 1998, *ApJ*, 502, 1
- Thorne, B., Dunkley, J., Alonso, D., & Naess, S. 2017, *MNRAS*, 469, 2821
- Thorne, B., Dunkley, J., Alonso, D., et al. 2019, arXiv e-prints [arXiv:1905.08888]
- Vacher, L., Aumont, J., Montier, L., et al. 2022, *A&A*, 660, A111
- Vacher, L., Chluba, J., Aumont, J., Rotti, A., & Montier, L. 2023, *A&A*, 669, A5
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nat. Methods*, 17, 261
- Wagenmakers, E., & Farrell, S. 2004, *Psychon Bull. Rev.*, 11, 192
- Zaldarriaga, M., & Seljak, U. 1997, *Phys. Rev. D*, 55, 1830
- Zaldarriaga, M., & Seljak, U. 1998, *Phys. Rev. D*, 58, 023003
- Zonca, A., Singer, L., Lenz, D., et al. 2019, *J. Open Source Software*, 4, 1298

## Appendix A: Validation of power spectra and posteriors



**Fig. A.1.** Empirical distribution of the AIC-based relative model odds between the  $C_\ell$ -fiducial and the  $C_\ell$ -moments model from 100 simulations. Same as Fig. 12 but only considering the four less complex Galactic foreground templates. Note the different  $x$  scale. We do not see strong preference for  $C_\ell$ -moments (indicated by large negative values) in any of the cases considered.

The goodness of fit of the power spectrum likelihood can be assessed by the well-known minimum- $\chi^2$  statistic  $\hat{\chi}_{\min}^2 = (\mathbf{d} - \hat{\mathbf{t}})^T \mathbf{C}^{-1} (\mathbf{d} - \hat{\mathbf{t}})$ , with  $\mathbf{d}$ ,  $\hat{\mathbf{t}}$ , and  $\mathbf{C}$  denoting data, best-fit theory model, and covariance, respectively. We perform this validation for pipeline A only, since it includes the foreground model at the likelihood level. Under the hypothesis of Gaussian data,  $\hat{\chi}_{\min}^2$  is expected to follow a  $\chi^2$  distribution with  $N - P = 558$  degrees of freedom (Pearson 1900; Maltoni & Schwetz 2003). When using Gaussian foregrounds, we know that our data are sufficiently Gaussian and the covariance is exact within the MCMC noise level. We compute the  $\hat{\chi}_{\min}^2$  for 100 simulated data realizations containing standard cosmology ( $r = 0$ ,  $A_{\text{lens}} = 1$ ), inhomogeneous noise in the goal-optimistic scenario and in four foreground cases (Gaussian, d0s0, d1s1, dmsm), considering both the  $C_\ell$ -fiducial and the  $C_\ell$ -moments model. The empirical distri-

butions match the theoretical expectation in all cases, showing that the covariance matrix is appropriate, even for non-Gaussian foreground input templates.

We also assess, for each simulation seed, which foreground model is preferred by the simulated data, using the *Akaike Information Criterion* (AIC, Akaike 1974). We compute the difference  $\Delta\text{AIC} = 2\Delta k + \hat{\chi}_{\min}^2(\text{moments}) - \hat{\chi}_{\min}^2(\text{fiducial})$ , where  $\Delta k = 4$  is the number of excess parameters of the  $C_\ell$ -moments over  $C_\ell$ -fiducial. Following Wagenmakers & Farrell (2004), the number  $\exp(\Delta\text{AIC}/2)$  can be interpreted as the relative model odds  $P(C_\ell\text{-fiducial})/P(C_\ell\text{-moments})$ .

The results are shown in Fig. A.1. The AIC test detects no clear preference for  $C_\ell$ -moments over  $C_\ell$ -fiducial. Among 100 simulations containing dmsm foregrounds, 30 have model odds below 1, and only a single simulation below  $10^{-2}$ . In case of the dmsm simulation set, this might come as a surprise, considering that the  $C_\ell$ -moments model allows to mitigate a  $1\text{-}2\sigma$  bias. We conclude that one must be careful when interpreting  $< 2\sigma$  detections in marginal posterior distributions, as they may be difficult to distinguish from subdominant residual foreground contamination using standard model-comparison techniques. As shown in Sect. 4.4, this situation changes in the presence of input foregrounds that induce a large bias.

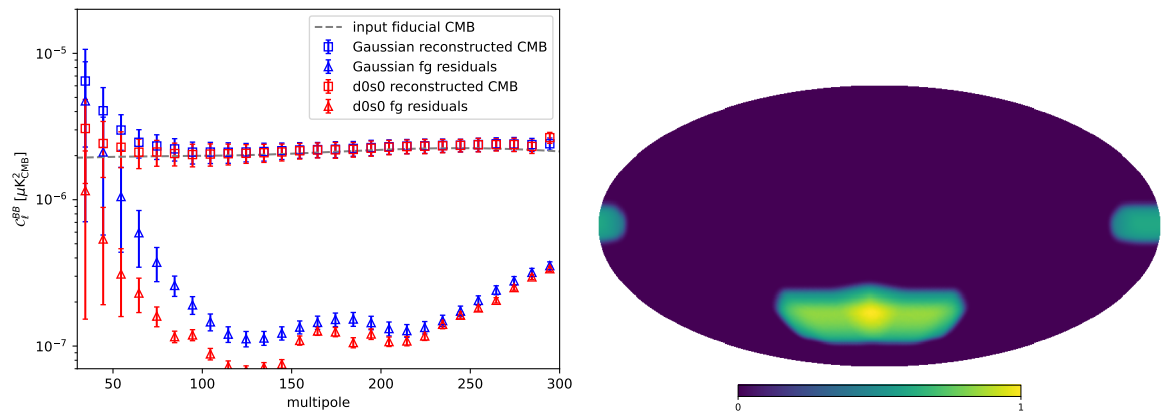
Lastly, we tested the robustness of the statistical results on  $r$  quoted in Sect. 4. Using pipelines A, A+moments, and B, we calculated the mean and maximum a-posteriori (MAP) value of the marginal  $r$  posterior for each of 500 simulations containing CMB realized with the fiducial cosmology ( $r = 0$ ,  $A_{\text{lens}} = 1$ ), dmsm foregrounds, and inhomogeneous noise in the baseline-optimistic scenario. We computed the average and standard deviation over the 500 simulations, and repeated the procedure for the lensing amplitude  $A_{\text{lens}}$ . Table A.1 shows the results. We find that for both parameters and all three pipelines, the sample average of the mean and the MAP agree at the  $0.1\sigma$  level. We also find consistency between the standard deviation of the marginal posteriors averaged over 500 simulations, the sample scatter of the marginal posterior mean values, and the sample scatter of the MAP values computed from 500 simulations. Table A.1 also lists the Gaussian error on the average mean and MAP values, corresponding to the sample scatter divided by the square root of the number of simulations,  $\sqrt{500}$ . While this test would result in a bias on  $r$  for all component separation pipelines (yet below  $2\sigma$  in the case of the A + moments pipeline), passing it is far beyond the requirements set by the statistical sensitivity of SO.



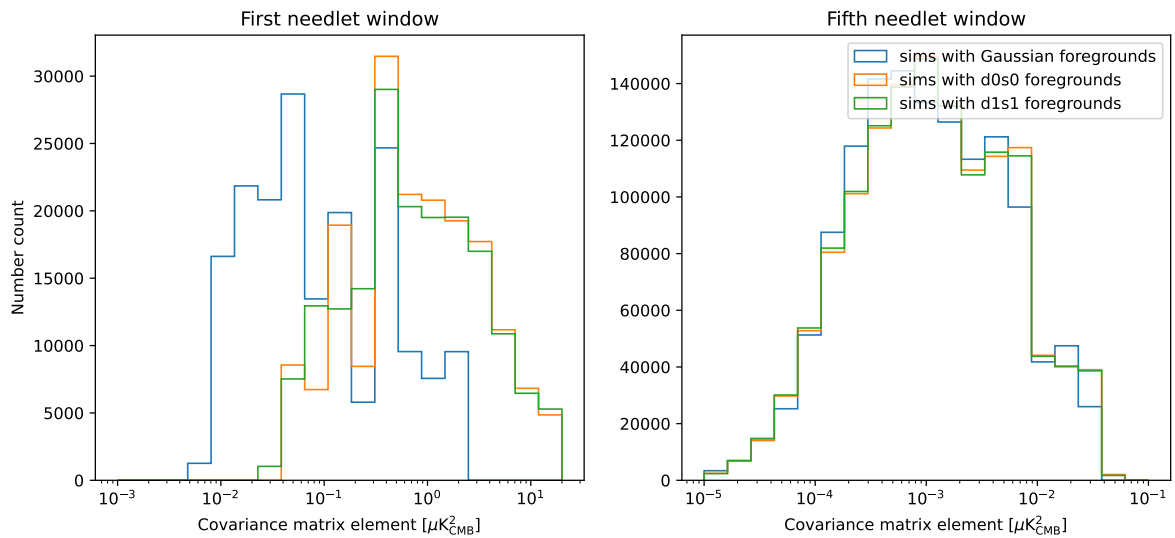
**Table A.1.** Posterior statistics on  $r$  and  $A_{\text{lens}} - 1$  from 500 simulations. We assume `dmsm` foregrounds and baseline-optimistic noise, as inferred by the three nominal component separation pipelines. No delensing or primordial gravitational waves are assumed.

Pipeline	parameter	posterior mean			posterior maximum			posterior standard deviation
		mean	$\sigma$	$\Delta$	mean	$\sigma$	$\Delta$	
A	$r \times 10^3$	$3.9 \pm 0.1$	2.6	+1.5	$4.0 \pm 0.1$	2.6	+1.6	2.6
	$(A_{\text{lens}} - 1) \times 10^2$	$-0.6 \pm 0.1$	3.3	-0.2	$-0.7 \pm 0.1$	3.3	-0.2	3.5
A+moments	$r \times 10^3$	$0.3 \pm 0.1$	4.0	+0.1	$-0.1 \pm 0.2$	4.0	0.0	3.5
	$(A_{\text{lens}} - 1) \times 10^2$	$0.0 \pm 0.1$	3.4	0.0	$0.0 \pm 0.1$	3.4	0.0	3.6
B	$r \times 10^3$	$2.5 \pm 0.1$	2.7	+0.9	$2.5 \pm 0.1$	2.7	+0.9	2.7
	$(A_{\text{lens}} - 1) \times 10^2$	$6.2 \pm 0.1$	3.1	+2.0	$6.2 \pm 0.1$	3.1	+2.0	3.2

## Appendix B: Bias on $r$ for Gaussian foregrounds and pipeline B



**Fig. B.1.** Foreground bias and custom analysis mask related to needlet-based component separation. *Left panel:* Foreground residuals in the reconstructed CMB power spectra in the inhomogeneous goal-pessimistic noise scenario, with a fiducial input cosmology of  $r = 0$ ,  $A_{\text{lens}} = 1$ , and Gaussian and `d0s0` foregrounds. We show the reconstructed CMB and foreground power spectrum residuals  $C_l^{BB}$ . The residual is calculated by mixing the pure foreground maps with the NILC weights. *Right panel:* constrained analysis mask used by pipeline B in the case of Gaussian foregrounds, built from the fiducial mask shown in Fig. 2.



**Fig. B.2.** Distribution of covariance matrix elements used to build the ILC weights (see Eq. 5) for pipeline B. We show values for a single simulation seed used for B-mode reconstruction, and include all cross-frequency combinations. We assume inhomogeneous goal-pessimistic noise, with a fiducial ( $r = 0$ ,  $A_{\text{lens}} = 1$ ) CMB, with Gaussian, `d0s0`, and `d1s1` foregrounds. *Left panel:* first needlet window (largest scales), *right panel:* fifth needlet window (smallest scales).

In the results of pipeline B, we note a consistent bias on  $r$  for simulations that include Gaussian foregrounds in combination with inhomogeneous noise. This bias seems unreasonable, considering the results that use more complex foregrounds, such as **d0s0**, but the same noise and CMB simulations, lead to considerably less bias. The cosmological bias is directly caused by a bias at the large angular scales of the  $BB$  power spectrum, which is easily visible when plotting the spectra. We confirm that this bias corresponds to foreground bias. In Fig. B.1 (left panel), we show the reconstructed CMB  $C_\ell^{BB}$  spectrum for the standard cosmology ( $r = 0$ ,  $A_{\text{lens}} = 1$ ) with inhomogeneous goal-pessimistic noise, for both the Gaussian (blue squares) and **d0s0** (red squares) foregrounds. The marker and error bar shows the mean and  $1-\sigma$  standard deviation across the 500 simulations. The excess bias at large scales is evident as the reconstructed CMB clearly surpasses the fiducial  $BB$  spectrum. We can calculate the exact foreground bias present on each reconstructed CMB, by mixing the pure foregrounds maps with the same NILC weights (Eq. 5), and taking the power spectra of that foreground bias reconstruction. This is shown as the blue triangles for Gaussian and red triangles for **d0s0** foregrounds. In the Gaussian case, we see that the large-scale bias is almost entirely caused by foregrounds.

We know this bias is directly related to the NILC weights that mix the frequency maps (Eq. 5). The weights are calcu-

lated directly from the frequency-frequency covariance matrix, which in turn is calculated from the input frequency maps themselves. If, for instance, we wish to test if the weights calculated with the simulations including Gaussian foregrounds are incorrect, we can do the cross check by using other weights to mix those same frequency maps. We take the NILC weights calculated for the simulations with **d0s0** foregrounds (shown in Fig. B.1) and use them to mix the frequency maps that include the Gaussian foregrounds. In this case, the mean best fit is  $r = -0.0001 \pm 0.0026$ , which is comparable to  $r = -0.0005 \pm 0.0028$  for the simulations including **d0s0** foregrounds (shown in Table 4).

Any incorrect weights must originate from the NILC covariance matrix. The combination of directly calculating the covariance matrix over maps that have inhomogeneous noise and Gaussian foregrounds creates the observed bias. We can alleviate this problem by spatially constraining the hits mask (Fig. 2) to a more homogeneous area, where the noise map realization will also be more homogeneous. The mask we use to do this is shown in Fig. B.1 (right panel). This custom mask has a smaller sky fraction than the fiducial analysis mask, which increases the statistical uncertainty on the inferred value of  $r$ . The measurements using this more constrained and more homogeneous mask are marked with a  $\dagger$  in Table 4.