

# A Bayesian Calibration Framework for EDGES

Steven G. Murray,<sup>1</sup>★ Judd D. Bowman,<sup>1</sup> Peter H. Sims,<sup>2</sup> Nivedita Mahesh,<sup>1</sup> Alan E. E. Rogers,<sup>3</sup> Raul A. Monsalve,<sup>4,1,5</sup> Titu Samson,<sup>1</sup> Akshatha Konakondula Vydula,<sup>1</sup>

<sup>1</sup>*School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85287, USA*

<sup>2</sup>*Department of Physics and McGill Space Institute, McGill University, Montréal, QC H3A 2T8, Canada*

<sup>3</sup>*Haystack Observatory, Massachusetts Institute of Technology, MA 01886, USA*

<sup>4</sup>*Space Sciences Laboratory, University of California Berkeley, Berkeley, CA 94720, USA*

<sup>5</sup>*Facultad de Ingeniería, Universidad Católica de la Santísima Concepción, Alonso de Ribera 2850, Concepción, Chile*

arXiv:2209.03459v1 [astro-ph.CO] 7 Sep 2022

Accepted XXX. Received YYY; in original form ZZZ

## ABSTRACT

We develop a Bayesian model that jointly constrains receiver calibration, foregrounds and cosmic 21 cm signal for the EDGES global 21 cm experiment. This model simultaneously describes calibration data taken in the lab along with sky-data taken with the EDGES low-band antenna. We apply our model to the same data (both sky and calibration) used to report evidence for the first star formation in 2018. We find that receiver calibration does not contribute a significant uncertainty to the inferred cosmic signal (< 1%), though our joint model is able to more robustly estimate the cosmic signal for foreground models that are otherwise too inflexible to describe the sky data. We identify the presence of a significant systematic in the calibration data, which is largely avoided in our analysis, but must be examined more closely in future work. Our likelihood provides a foundation for future analyses in which other instrumental systematics, such as beam corrections and reflection parameters, may be added in a modular manner.

**Key words:** cosmology: observations – methods: statistical – dark ages, reionization, first stars

## 1 INTRODUCTION

The globally-averaged brightness temperature of the hyperfine spin-flip transition of neutral hydrogen (the 21 cm line) is a powerful probe of the thermal history of the early Universe ( $z \sim 6 - 30$ ; for reviews, see eg. Furlanetto et al. 2006; Pritchard & Loeb 2012; Furlanetto 2016). Accurately observing this brightness temperature, and separating it from the bright foreground emission of our Galaxy, have proven to be an exceptional challenge. Several instruments have taken up this challenge, including EDGES (Bowman et al. 2008; Rogers & Bowman 2012), LEDA (Bernardi et al. 2016), BIGHORNS (Sokolowski et al. 2015), and SARAS (Girish et al. 2020). Since the publication of the first evidence for star formation in Cosmic Dawn by the EDGES collaboration (Bowman et al. 2018, hereafter B18), there has been an increased interest in independent verification, resulting in several new and upcoming experiments, eg. SARAS3 (Nambissan et al. 2021), ASSASSIN (McKinley et al. 2020) and REACH (eg. Anstey et al. 2020). Importantly, the recent results of SARAS3 (Singh et al. 2022) appear inconsistent with the inferred cosmic signal of B18, suggesting that either measurement (or both) may be contaminated by systematics.

Despite the overwhelming magnitude of the foregrounds ( $\sim 10^5$  times the signal, eg. Shaver et al. 1999), they are not the primary challenge *in isolation*. Indeed, physical models of foreground spectra are incredibly smooth, defined by relatively low-order deviations from a power-law (Jelić et al. 2010). Conversely, most cosmological

signals have more rapid spectral structure, much of which is not captured by these same low-order basis-sets (Bevins et al. 2021). Rather, the primary challenge arises via the multiplication of these bright smooth foregrounds by relatively small spectral structures induced by the instrument. These come from a number of physically distinct mechanisms, including *beam chromaticity* (in which angular structure in the sky and beam are translated to frequency structure via the frequency-dependence of the beam-shape primarily due to reflections from the edges of the ground plane and nearby objects (Rogers et al., 2022, *submitted*), reflection parameters of the signal chain, and receiver gains. Structures created by these instrumental systematics that happen to share similar scales as the expected signal must thus be avoided, or calibrated to a precision of about  $10^{-5}$  in order to provide minimal contamination to the estimated cosmic signal.

The results of B18 were carefully calibrated, and are expected to have residual systematics that are subdominant to the (surprisingly strong) cosmological absorption feature. Nevertheless, the analysis was performed in a way that obscures the relationship between the uncertainties on the known systematics and the final uncertainties on the cosmological estimate. That is, the reported error bars were obtained via independent estimates of the propagated uncertainties of various systematics, added in quadrature. This is a crude estimate, not accounting for correlations in the effects of different unknown parameters, nor properly accounting for our prior knowledge of these parameters.

This is made compelling by Sims & Pober (2020), who use the Bayes Factor – a rigorous metric of the comparative evidence for one

★ E-mail: steven.g.murray@asu.edu

model over another – to argue that a simple unmodeled systematic that exhibits as a damped sinusoid in the spectrum would disfavor a strong absorption feature (see eg. [Hills et al. 2018](#); [Singh & Subrahmanyan 2019](#), for prior studies that suggested a similar phenomenological systematic). Nevertheless, this statement is highly dependent on our prior knowledge; we know of no physical systematic that should arise as such a simple damped sinusoid. On a more nuanced view, it is possible that the combination of receiver gains, reflections and beam chromaticity would compound to yield a systematic with approximately sinusoidal structure; however, it then becomes important to know what the expected amplitude and period of such a sinusoid might be, and how likely it is (given the physical uncertainties of the parameters involved) that it would reach the strength and shape required to obviate the cosmological feature.

To properly address these questions, we require a full Bayesian forward-model. Such a model begins with the unknown physical parameters, for which we have reasonable estimates of uncertainty, and propagates those uncertainties self-consistently all the way through to the final signal estimation. This captures the full correlated, non-Gaussian probability distributions of the unknown parameters, allowing a more rigorous determination of their marginalised uncertainty. It also allows for comparing different models.

There is precedent for Bayesian models in global 21 cm experiments. Besides the use of Bayesian techniques to determine posteriors on the signal and foreground parameters (eg. [Monsalve et al. 2017b, 2018, 2019](#); [Singh & Subrahmanyan 2019](#); [Sims & Pober 2020](#); [Bevins et al. 2022](#)), there has been work on including various systematics in forward models, predominantly led by the REACH collaboration. This pioneering work encompasses foreground models and beam chromaticity ([Anstey et al. 2020](#)), antenna models ([Anstey et al. 2022](#)), generalized systematics ([Scheutwinkel et al. 2022a](#)) and non-Gaussian noise statistics ([Scheutwinkel et al. 2022b](#)). Perhaps most relevant for this work, [Roque et al. \(2020\)](#) considers receiver calibration under a Bayesian model-selection framework. In this work, the focus was on modeling the receiver gain posteriors, in order to determine a posterior on the calibrated temperature (and also to choose the number of polynomial terms required for calibration in a self-consistent way). To do this efficiently, [Roque et al. \(2020\)](#) use the method of conjugate priors, yielding an analytic solution to the posterior. In this paper, we implement a very similar Bayesian model for the receiver gains; however, we do not adopt the conjugate prior formalism, despite its efficiency. We do this because beyond the receiver gains, we are interested in simple models for the reflection coefficients. The required flexibility for these extended models makes using conjugate priors more difficult. Furthermore, we extend the forward model through to *joint analysis* of the receiver gain, foregrounds and cosmic signal.

This paper is the beginning of a larger project in which the entirety of the EDGES analysis chain is to be cast in a Bayesian forward-model. In this paper, we focus purely on receiver calibration. As this paper is primarily about the *technique*, we will apply the model to the data that constituted the result of [B18](#). As such, this paper is not intended to form a full ‘validation’ of the [B18](#) result; rather, it provides a necessary step in building confidence that certain systematics (in receiver gains) are unlikely to have caused the surprising results previously obtained. A more complete verification requires the modeling of all *known* systematics, and furthermore an expansion of the data (preferably to independent telescopes) to investigate potential *unknown* systematics.

The layout of the paper is as follows. §2 describes the data used throughout the paper. §3 introduces Bayesian inference in general, and derives a high-level likelihood for global experiments. Sec. 4

dives into the details of the probabilistic receiver gain calibration model adopted in this paper. §5 extends this probabilistic model to include sky data, before we analyse that data with our Bayesian model in Sec. 6. Finally, we summarize and conclude in Sec. 7.

All analysis in this work is open-source and available in Jupyter notebooks and Python scripts<sup>1</sup>.

## 2 DATA USED

All data used in this paper comes from [B18](#). Only part of the data required here was made publicly available by [B18](#), namely the time-averaged sky spectra. In addition to the sky spectrum itself, we require several calibration products in this paper, for which we use the exact data/settings used in [B18](#).

The sky spectrum consists of observations between day 250 of 2016 through to day 98 of 2017 (138 days after initial data quality cuts). Each integration from each night is filtered for RFI and other systematic outliers, including cuts on metadata such as local humidity and potential saturation of the analog-to-digital-converter (ADC). After filtering a day’s worth of data, all integrations within the 12 hours of LST corresponding to the galactic centre being below the horizon are averaged together. This averaged spectrum is then calibrated for the receiver gain, beam correction and path losses. Further filtering is performed on the calibrated, averaged spectrum from each night, checking for outliers. Finally, the 138 days are averaged together, and the spectra are binned in frequency bins of  $\sim 0.390$  MHz. This final spectrum is referred to as  $\hat{T}_{\text{sky},\text{bc}}$  in this paper (cf. Eq. 48), and we directly use the publicly available data<sup>2</sup> in this paper. We refer the interested reader to [B18](#) for details on the data analysis.

In this paper, we often need to ‘undo’ the calibration of the fiducial dataset described above, in order to recalibrate with different parameters. This process is defined in Eq. 48, and requires the nominal receiver calibration ( $\hat{T}_0^{\text{ant}}$  and  $\hat{T}_1^{\text{ant}}$ ), beam correction and path loss.

The path loss is, in general, a product of antenna, balun, connector and ground losses. The antenna loss is produced via simulation with FEKO ([Elsherbeni et al. 2014](#)). In the case of the public data from [B18](#), the ground loss is set to unity (i.e. ignored). The beam correction is produced via Eq. 39 (cf. [Mozdzen et al. 2019](#)), using the Haslam all-sky map ([Haslam et al. 1982](#)) with a spatially-invariant spectral index of -2.5, along with a beam model produced with FEKO ([Mahesh et al. 2021](#)). While these basic products are not publicly available, we show their final form in Fig. 8.

To obtain the receiver calibration we directly use outputs of the original C-code adopted in [B18](#)<sup>3</sup>. This includes frequency-dependent values of five receiver-calibration coefficients as well as the reflection coefficients of the receiver and antenna. These calibration parameters and reflection coefficients allow us to *de*-calibrate the public data (essentially taking it back to its raw form<sup>4</sup>). However, to *re*-calibrate the data requires the calibration measurements used to initially derive these calibration solutions, along with the various original settings

<sup>1</sup> Available at <https://github.com/edges-collab/bayesian-calibration-paper-code/releases/tag/submitted>

<sup>2</sup> Available at <https://www.nature.com/articles/nature25792/figures/1>

<sup>3</sup> This code, with scripts to run it with the same settings as [B18](#), is available at <https://github.com/edges-collab/alans-pipeline>

<sup>4</sup> As described further in the paper, we do this de-calibration, instead of starting directly from the raw data, because we wish to retain the exact averaging and flagging used on the public data, without re-performing this compute-heavy task.

used in the analysis. These calibration measurements were taken in the lab in September 2015. The observation includes the simultaneously measured spectra and temperatures from the four input ‘calibration’ sources (ambient, hot load, open and shorted long cable) plus an ‘antenna simulator’ designed to mimic the reflection coefficients of the antenna, as well as reflection coefficients for these input sources and the internal switch and receiver, and measurements of the resistance of the (SOL) calibration standards used to measure the reflection coefficients. These calibration measurements are here analysed with a new publicly-available calibration code, EDGES-CAL<sup>5</sup> to produce the five calibration parameters referenced previously. A detailed report of the use of the new code to produce the calibration parameters in this paper is available in Murray (2022b), which also demonstrates slight variations in the results between the codes – even with (nominally) the same input settings. The largest difference concerns the modelling of the various reflection parameters required (one for each calibration source, plus the internal switch, receiver and antenna). Since we are not concerned in this paper with re-modelling the reflection parameters, we simply take the direct output of the code used in B18 and input those values to our own calibration using EDGES-CAL<sup>6</sup>.

### 3 MATHEMATICAL AND BAYESIAN FRAMEWORK

#### 3.1 Notational Preliminaries

Throughout, bold upright quantities, eg.  $\mathbf{Q}$ , will refer to matrices, and bold italic, eg.  $\mathbf{q}$  will refer to vectors (where possible, vectors will also be lower case). Throughout, the symbol ‘ $\circ$ ’ will refer to Hadamard (i.e. element-wise) multiplication, and  $\oslash$  will refer to Hadamard division. The symbol  $\mathbf{I}_n$  will refer to the  $n \times n$  identity matrix. We will construct (row) vectors using square brackets surrounding elements separated by commas, eg.  $\mathbf{q}_{\text{cal}} = [q_1, q_2, \dots]$ , and assume that the transpose of a row vector,  $\mathbf{q}^T$ , is a column vector.

The ensemble average of a random variable will be denoted by angle brackets, eg.  $\langle q \rangle$ , while a sample mean will be denoted by an over-bar, eg.  $\bar{q}$ . An estimate of a quantity will be denoted by a hat, eg.  $\hat{q} = \bar{q}$ .

When denoting parameters that are statistics of a certain observable which itself is measured for multiple independent sources (eg. the variance,  $\sigma^2$  of the three-position-switch ratio,  $q$  for the open cable), we will denote the the observable as eg.  $q_{\text{open}}$ , but will ‘lift’ the source by one subscript level when denoting the statistics, i.e.  $\sigma_{q,\text{ant}}^2$  rather than  $\sigma_{q_{\text{ant}}}^2$ .

Throughout, we will use curly braces surrounding elements separated by commas to construct sets. Symbols denoting such sets will typically be in upper-case Latin calligraphic font, eg.  $\mathcal{T}_{\text{NW}} = \{\cos, \text{unc}, \sin\}$ , and usually these will denote sets of *labels* (eg. the three labels associated with ‘noise-wave’ temperatures). Sets of quantities associated with these labels may be represented in the short-hand notation  $T_{\mathcal{T}_{\text{cal}}} \equiv \{T_p \mid p \in \mathcal{T}_{\text{cal}}\}$ .

<sup>5</sup> Available at <https://github.com/edges-collab/edges-cal>

<sup>6</sup> In this work, we use the full suite of new EDGES pipeline codes, all open-source and available at <https://github.com/edges-collab>. Specifically, we use READ-ACQ v0.5.0, EDGES-IO v4.1.3, EDGES-CAL v6.2.3, EDGES-ANALYSIS v4.1.3 and EDGES-ESTIMATE v1.3.0

#### 3.2 Bayes’ Theorem

Bayesian approaches to parameter inference and model selection have become extremely popular in the astrophysics and cosmology literature. As such, we will only describe them briefly, referring the interested reader to more in-depth resources, such as Jaynes & Bretthorst (2003).

Bayesian statistics is fundamentally the update of the *credence* in a certain model given the acquisition of new data pertaining to the model. That is, it presupposes an existing credence (the “prior”) and some observations, and given a likelihood of obtaining those observations given the parameters of the model, it yields an updated credence. This process is described by Bayes’ formula:

$$P(\boldsymbol{\theta}|D, \mathcal{M}) = \frac{\mathcal{L}(D|\boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta})}{\mathcal{Z}(D|\mathcal{M})}. \quad (1)$$

The ‘model’,  $\mathcal{M}$ , is here parameterized by the set of parameters  $\boldsymbol{\theta}$ <sup>7</sup>. The LHS represents the ‘posterior’ credence of  $\boldsymbol{\theta}$  under model  $\mathcal{M}$  *after* observing data  $D$ , while the RHS takes the ‘prior’ credence,  $\pi(\boldsymbol{\theta})$ , and updates it with the ‘likelihood’ of the data,  $\mathcal{L}(D|\boldsymbol{\theta})$ , normalized by the ‘evidence’,  $\mathcal{Z}(D|\mathcal{M})$ .

Typically (although see Roque et al. (2020)) the evidence is impossible to write down analytically, but may be computed as the integral of the likelihood over the prior subspace. In this paper, we use the POLYCHORD sampler which is able to provide not only samples from the posterior, but an estimate of the evidence,  $\mathcal{Z}$ .

#### 3.3 The Gaussian Likelihood

In this paper, we will exclusively use a *Gaussian* likelihood. In this likelihood, we model the data as being sampled from a multivariate Gaussian distribution with mean vector  $\boldsymbol{\mu}(\boldsymbol{\theta}) \equiv \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{x})$  and covariance matrix  $\boldsymbol{\Sigma}(\boldsymbol{\theta}) \equiv \boldsymbol{\Sigma}(\boldsymbol{\theta}, \mathbf{x})$ . The mean vector is typically dependent both on the parameters of the model and a predicate variable,  $\mathbf{x}$ , which for this paper will be taken to be known with certainty (typically it will be frequency and/or input source). The data is taken to be sampled at particular values of this predicate variable,  $\mathbf{d} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$ .

The likelihood is thus given by

$$\mathcal{L}_g(\mathbf{d}|\boldsymbol{\theta}) \propto \sqrt{|\boldsymbol{\Sigma}^{-1}|} \exp \left\{ -\mathbf{r}^T \boldsymbol{\Sigma}^{-1} \mathbf{r} \right\}, \quad (2)$$

where the model residual is given by

$$\mathbf{r} = \mathbf{d} - \boldsymbol{\mu}(\boldsymbol{\theta}). \quad (3)$$

Note that the Gaussian likelihood is valid so long as the residuals,  $\mathbf{r}$  are Gaussian distributed. Given that some raw data  $\mathbf{d}$  is Gaussian distributed, the linearly transformed data  $\mathbf{d}' = \mathbf{A}\mathbf{d} + \mathbf{b}$  is also Gaussian distributed. Here the  $\mathbf{b}$  is simply absorbed into  $\boldsymbol{\mu}$ , but the scaling matrix  $\mathbf{A}$  results in a scaled covariance. Thus, in general with raw Gaussian-distributed data  $\mathbf{d}$  with covariance  $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$  we may write the likelihood as

$$\mathcal{L}_g(\mathbf{d}|\boldsymbol{\theta}) \propto \sqrt{|\boldsymbol{\Sigma}_A^{-1}|} \exp \left\{ -\mathbf{r}'^T \boldsymbol{\Sigma}_A^{-1} \mathbf{r}' \right\}, \quad (4)$$

with

$$\mathbf{r}' = \mathbf{A}_{\boldsymbol{\theta}} \mathbf{d} - \boldsymbol{\mu}'(\boldsymbol{\theta}) \quad (5)$$

$$\boldsymbol{\Sigma}_A = \mathbf{A}_{\boldsymbol{\theta}}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \mathbf{A}_{\boldsymbol{\theta}}. \quad (6)$$

<sup>7</sup> In principle, the total possible set of models may contain completely different parameterizations. In practice, we typically explore a single parameterization,  $\mathcal{M}$ , at a time, which has parameters  $\boldsymbol{\theta}$ .

While the two forms (2 and 4) are mathematically equivalent, it is sometimes convenient to use the scaled form in order to make certain properties of  $\mu$  clear, as we will now discuss.

### 3.4 Inference Method

The models we will encounter in this work generally have a large number of parameters. This is prohibitive for performing Bayesian inference via MCMC, due to the ‘curse of dimensionality’. Examples of inference methods that are applicable to high-dimensional data (under some conditions) are Gibbs sampling and Hamiltonian Monte Carlo (HMC). However, these techniques do not easily yield the Bayesian evidence, which is useful for comparing models, and especially for deciding on the relevant number of parameters to include in our smooth models.

Instead, we adopt a technique in which some of the parameters of the model are *pre-marginalized*. That is, we integrate the posterior distribution analytically for the *linear* parameters, reducing the effective dimensionality for the sampler, which must only deal with the remaining non-linear parameters. This technique has been previously described in (eg. [Lentati et al. 2017](#); [Monsalve et al. 2018](#); [Tauscher et al. 2021](#)), and we derive it for our purposes in App. B.

In short, the result is that in the context of a particular MCMC sample, we must sample only a set of *non-linear* parameters, for which we solve for the maximum-likelihood (ML) of the remaining linear parameters. Letting  $\hat{\mathbf{r}}$  be the residuals of the data to this conditional ML model, the posterior of the non-linear parameters is given by

$$p_{\text{NL}}(\theta_{\text{NL}} | \mathbf{d}) \propto \sqrt{|\Sigma^{-1}| |\Sigma_L|} \exp \left\{ -\frac{1}{2} \hat{\mathbf{r}}^T \Sigma^{-1} \hat{\mathbf{r}} \right\}, \quad (7)$$

where  $\Sigma_L$  is the covariance matrix of the linear sub-model. It is also possible to obtain samples from the posterior of the linear sub-model via sampling from a multivariate normal with mean  $\hat{\theta}_L$  and covariance  $\Sigma_L$  (cf. [Tauscher et al. 2021](#)).

### 3.5 Sampling Method

For all of our Bayesian sampling in this paper we use the POLYCHORD nested-sampling code ([Handley et al. 2015a,b](#)). For sampling, we use  $N_{\text{live}} \approx 100N_{\text{dim}}$ , where  $N_{\text{dim}}$  is the number of parameters sampled by the MCMC (i.e. not including linear parameters). Importantly, POLYCHORD is able to generate an estimate of the Bayesian evidence, which is useful for model comparison.

For all non-linear parameters in this work we employ uniform priors, i.e. any value of each parameter – within certain bounds – is equally likely *a priori*.

## 4 A PROBABILISTIC CALIBRATION MODEL

Measurements of EDGES’ antenna temperature, like all antennas, is accompanied by some multiplicative gain and additive noise<sup>8</sup>. While various external features, such as the angular response of the antenna, affect the voltage induced on the antenna itself, in this section we are not concerned with these effects, but rather the gain applied to this voltage on the signal path between the antenna and the analog-to-digital converter. That is, the gain applied by the receiver system itself before writing the measurements to disk.

<sup>8</sup> In principle, the gain may in fact be non-linear, but such a case is actively avoided and indications of such a state of affairs are flagged in our processing. Thus, we proceed with the assumption of linearity of the gains.

The primary – but not only – component that induces these gains is the low-noise amplifier (LNA), whose purpose is to amplify the incoming voltages from the antenna in order that additive noise in the rest of the system does not overwhelm the desired signal. Unfortunately, the value of this (complex-valued) receiver gain is not constant – either with frequency or time. It is dependent on the ambient temperature, humidity and other factors. To overcome this limitation, EDGES uses the well-known technique of Dicke-switching ([Rogers & Bowman 2012](#)) to perform gross calibration of the receiver gains. In this technique, measurements switch between the input (ostensibly from the antenna) to two different internal reference loads. In practice, this technique is not sufficient on its own for the high-precision required of 21 cm experiments; the signal path for the receiver input versus that of the internal references loads is slightly different (having an additional switch), and thus has slightly different reflection/propagation characteristics. These are accounted for by the noise-wave formalism ([Meys 1978](#)).

In this section, we present this technique of Dicke-switching along with the noise-wave formalism following [Monsalve et al. \(2017a\)](#) (hereafter M17). However, in doing so, we pay close attention to the probabilistic model, ultimately deriving a likelihood for the calibration parameters in a similar fashion to the recent work of [Roque et al. \(2020\)](#) for REACH. However, we do not follow [Roque et al. \(2020\)](#) in using conjugate priors to define our posterior distribution, using instead the linear marginalisation technique outlined in §3.4.

All the quantities described in this section are frequency-dependent. However, to a good approximation, frequency channels are statistically uncorrelated in the observed spectra, and thus in this section we may consider each channel independently. Thus for notational clarity we omit frequency dependence throughout.

Since we introduce many variables throughout the next two sections, we provide a summary of the variable definitions in Tables 1 and 2. The first provides a summary of the different sets of labels used throughout the paper, and the second lists many of the important variables. Furthermore, we summarize the entire pipeline as a flowchart in Fig. 1.

### 4.1 The Noise-Wave Formalism

The Dicke switching technique in EDGES alternates between three switch positions: the input ‘source’ (src; typically the antenna), an internal ‘load’ (L) and an internal ‘load + noise-source’ (LNS). Given a ‘true’ temperature,  $T_{\text{switch}} \in \{T_{\text{src}}, T_L, T_{\text{LNS}}\}$ , for any of these switches at a particular frequency, the receiver imparts a time-dependent multiplicative gain, and adds its own noise, such that the output power is

$$P_{\text{switch}} = g T_{\text{switch}} + T_{\text{inst}} + n_{\text{switch}}, \quad (8)$$

where  $T_{\text{inst}}$  is the instrument’s thermal contribution, and  $n_{\text{switch}}$  is a zero-mean Gaussian random variable whose variance is proportional to  $g T_{\text{switch}}$ <sup>9</sup>.

We may thus form the power quotient

$$q_{\text{src}} \equiv \frac{P_{\text{src}} - P_L}{P_{\text{LNS}} - P_L}, \quad (9)$$

We note that the numerator and denominator are both Gaussian-distributed, and are correlated due to their mutual dependence on

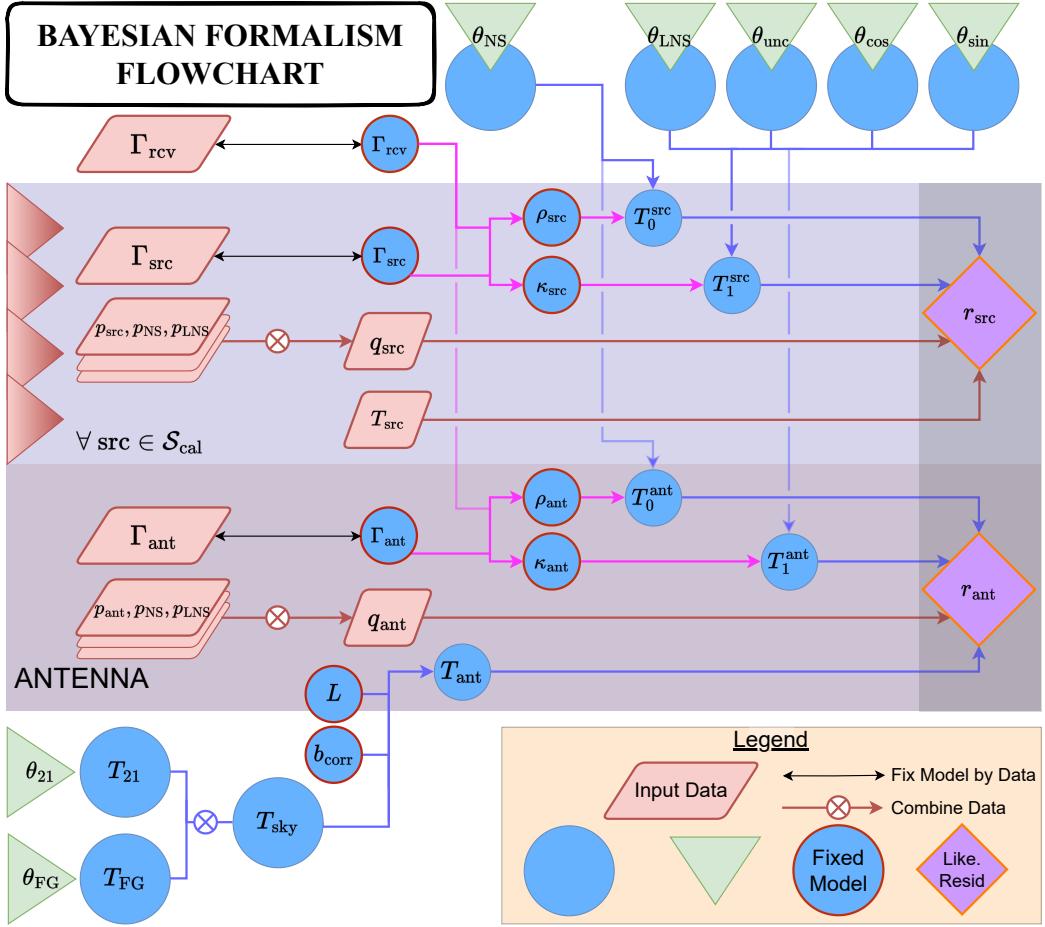
<sup>9</sup> We do not provide a definite form for  $n_{\text{switch}}$  here, as its true form is dependent on a number of subtle factors, such as the spectrometer and internal noise characteristics. For this paper, it is enough to assume it is zero-mean and Gaussian.

| SYMBOL        | ELEMENTS                     | SUBSETS   | VARS   | DESCRIPTION   | Eqs. |
|---------------|------------------------------|---|--------|---|------|
| $\mathcal{L}$ | {src, L, LNS}                |   | switch | Internal switch-position for the receiver: ‘src’ referring to receiver input port (may be substituted by label for the particular input source, see below), ‘L’ to internal load, and ‘LNS’ the internal ‘load plus noise-source’ | 8    |
| $\mathcal{S}$ | {amb, hot, open, short, ant} | $\mathcal{S}_{\text{cal}} = \mathcal{S} - \{\text{ant}\}$   | src    | Sources attached to the receiver input port   | 10   |
|               |                              | $\mathcal{T}_{\text{intload}} = \{\text{L, NS}\}$   |        |   |      |
| $\mathcal{T}$ | {unc, cos, sin, L, NS}       | $\mathcal{T}_{\text{NW}} = \mathcal{T} - \mathcal{T}_{\text{intload}}$<br>$\mathcal{T}_{\text{nl}} = \{\text{NS}\}$<br>$\mathcal{T}_{\text{lin}} = \mathcal{T} - \mathcal{T}_{\text{nl}}$ | $p$    | The five modeled temperature models (noise-waves and internal loads)  | 30   |

**Table 1.** Sets of labels used throughout this paper, with their defining symbols, and relevant subsets. The VARS column gives common variables used to stand in for elements of the set (usually as subscripts).

| SYMBOL                                      | DESCRIPTION   | DOMAIN   | Eqs.   |
|---|---|--|--------|
| $P_{\text{switch}}$                         | X Power from the receiver pointing to a given switch.   | $\mathbb{R}^+$                                     | 8      |
| $q_{\text{src}}$                            | X The ‘three-position-switch ratio’ which normalises input source power by measured internal powers   | $\mathbb{R}$                                       | 9, 17  |
| $\mathbf{T}_{\text{NW}}$                    | M Vector of noise-wave temperatures, $[T_{\text{unc}}, T_{\cos}, T_{\sin}]$ at one frequency, specific to receiver.   | $\mathbb{R}^3$                                     | 10     |
| $T_{\text{src}}$                            | X M Temperature of a source connected to the receiver input. Modeled for $\text{src}=\text{ant}$ , measured otherwise.  | $\mathbb{R}^+$                                     | 10     |
| $\Gamma_{\text{inst}}, \Gamma_{\text{src}}$ | F Reflection coefficients of the instrument and input sources respectively  | $\mathbb{C}, 0 \leq  \Gamma  \leq 1$               | 16     |
| $\mathbf{k}_{\text{src}}$                   | F 3-vector denoting the power transfer efficiency of an input source coupled to the receiver with respect to the noise-wave temperatures, $\{\text{unc, cos sin}\}$   | $\mathbb{R}^3$                                     | 15     |
| $c_{\text{src}}$                            | F Power transfer efficiency of input source coupled to receiver with respect to input temperature   | $\mathbb{R}, 0 \leq c_{\text{src}} \leq 1$         | 14     |
| $h$   | F Power transfer efficiency of the instrument   | $\mathbb{R}, 0 < h < 1$                            | 13     |
| $\sigma_{q, \text{src}}^2$                  | X Variance of $q_{\text{src}}$ . Estimated empirically using time-samples as independent realizations for $\text{src} \in \mathcal{S}_{\text{cal}}$ , and using residuals to high-order smooth polynomial fits over frequency for $\text{src}=\text{ant}$ | $\mathbb{R}^+$                                     | 20     |
| $T'_L, T'_{\text{NS}}$                      | M Effective ( $T_L, T_{\text{NS}}$ ) accounting for path differences between the source input and internal loads.   | $\mathbb{R}^+$                                     | 22     |
| $\mathbf{T}$                                | M The vector of four temperatures that compose the linear sub-model: $[T_{\text{unc}}, T_{\cos}, T_{\sin}, T'_L]^T$   | $\mathbb{R}^{+4}$                                  | 22     |
| $T_0^{\text{src}}, T_1^{\text{src}}$        | M Multiplicative and additive temperatures converting measured $q_{\text{src}}$ into source temperature, $T_{\text{src}}$   | $\mathbb{R}^+, \mathbb{R}$                         | 28, 29 |
| $\mathbf{T}_p$                              | M Length- $N_{\nu}$ model temperature spectrum of one of the five estimated temperature models, $p \in \mathcal{T}$   | $\mathbb{R}^{N_{\nu}}$                             | 30     |
| $\theta_p$                                  | P Length- $N_{\text{terms}}^P$ vector of polynomial parameters for $\mathbf{T}_p$   | $\mathbb{R}^{N_{\text{terms}}^P}$                  | 30     |
| $\Psi$                                      | C $N_{\nu} \times N_{\text{terms}}$ matrix of polynomial basis vectors, $\Psi_{ij} = (\nu_i / \nu_{\text{ref}})^j$  | $(\mathbb{R}^+)^{N_{\nu} \times N_{\text{terms}}}$ | 30     |
| $\theta_{\text{NW+L}}$                      | P The vector of all polynomial coefficients for $p \in \mathcal{T}_{\text{lin}}$ : $[\theta_{\text{unc}}, \theta_{\cos}, \theta_{\sin}, \theta_L]$  | $\mathbb{R}^{N_{\text{NW+L}}}$                     | C6     |
| $\mathbf{r}_{\text{src}}$                   | T Model residual vector for $\text{src} \in \mathcal{S}$  | $\mathbb{R}^{N_{\text{src}} N_{\nu}}$              | 31     |
| $\Sigma_{\text{src}}$                       | T The modeled diagonal covariance of $\mathbf{r}_{\text{src}}$ , equal to $T_{\text{NS}}^2 \sigma_{q, \text{src}}^2$  | $(\mathbb{R}^+)^{N_{\nu}}$                         | 32     |
| $c_{\text{terms}}, w_{\text{terms}}$        | C Short-hand for the number of terms, $N_{\text{terms}}^P$ used for $p \in \mathcal{T}_{\text{intload}}$ and $p \in \mathcal{T}_{\text{NW}}$ respectively   | $\mathbb{Z}^+$                                     |        |
| $T_{\text{sky}}$                            | M True radio temperature of the sky   | $\mathbb{R}^+$                                     | 36     |
| $T_{21}$                                    | M Temperature of the cosmic 21 cm radiation   | $\mathbb{R}$                                       | 40     |
| $\bar{\mathbf{T}}_{\text{BWFG}}$            | M LST-averaged beam-weighted foregrounds. Modeled as a linear sum of log-polynomials.   | $\mathbb{R}^+$                                     | 52     |
| $\mathbf{B}$                                | F Antenna beam as a function of line-of-sight   | $\mathbb{R}^+$                                     | 37     |
| $T_{\text{sky,beam}}$                       | M Sky temperature after attenuation by antenna beam   | $\mathbb{R}^+$                                     | 37     |
| $b_{\text{corr}}$                           | F Beam chromaticity correction  | $\mathbb{R}$                                       | 39     |
| $T_{\text{sky,bc}}$                         | M Sky temperature after beam attenuation, but correcting for chromatic beam structure via $b_{\text{corr}}$   | $\mathbb{R}^+$                                     | 38     |
| $L$   | F Fractional loss in the signal path (includes antenna, balun, connector and ground loss)   | $\mathbb{R}^+, 0 < L < 1$                          | 42, 43 |
| $p_{\text{sky,meas}}$                       | X The measured power from the deployed antenna integrated over 39 sec (uncalibrated)  | $\mathbb{R}^+$                                     | 44     |
| $\hat{T}_{\text{sky,bc}}$                   | T The estimated calibrated sky temperature, where calibration is derived from an iterative procedure, averaged over time. Equivalent to publicly available data.  | $\mathbb{R}^+$                                     | 48     |
| $\hat{T}'_{\text{sky,bc}}$                  | T An estimate of the sky temperature obtained from decalibrating $\hat{T}_{\text{sky,bc}}$ back to $q_{\text{ant}}$ then re-calibrating with an alternate calibration model   | $\mathbb{R}^+$                                     | E1     |

**Table 2.** Summary of symbols used in this work, with a description, appropriate domain and example of equations in which they are used. The colored symbol in the second column provides a ‘type’ for the quantity. **Key to symbols:** M: ‘Models’, P: ‘Model Parameters’, F: ‘Fixed Models’, C: ‘Model Choices’ (eg. number of parameters), X: ‘Measurements’, T: ‘Model-transformed Measurements’. Note that not all listed model parameters or models are independent: some are derived from others, or simply concatenations of others. Note also that the distinction between ‘model’ and ‘measurement’ is not always simple. Here, by ‘measurement’ we mean any symbol denoting a quantity that *can* be directly measured, without requiring use of a model parameter. For example,  $q_{\text{ant}}$  can be calculated directly from measured data without requiring use of a model parameter. Conversely, ‘models’ are here defined as quantities that, once model parameters are chosen, do not require any data to calculate and *cannot* be uniquely determined by measurements. Notably, ‘measurements’ as here defined may ‘modeled’ (which is the point of inference), but are nonetheless defined as measurements under the above definitions. In-text equations involving symbols here defined as ‘measurements’ may in fact be referring to either measurements or models, depending on context. ‘Fixed models’ are those for which we do not let the parameters vary *in this work*, but are also not direct measurements. ‘Model-transformed measurements’ are those quantities for which both the measured data and some choice of model parameter is required in order to calculate.



**Figure 1.** Flowchart for the Bayesian pipeline presented in this paper. The top shaded panel presents the pure calibration model (cf. §4). By using the point estimates of the calibration parameters (green triangles at the top) along with the network in the lower shaded panel, one is executing the *isolated* sky model fit (cf. §5), while if all parameters are estimated together, the *joint* inference is being performed. The top panel implicitly includes identical copies for each of the calibration sources,  $\mathcal{S}_{\text{cal}}$ . Dark red lines follow the flow of spectrum and thermistor measurements. Pink lines trace the flow of reflection parameter measurements/models. Blue lines trace the flow of noise-wave and internal load temperature models. The likelihood is computed as a zero-mean multivariate Gaussian distribution with diagonal covariance, evaluated at the concatenation of all purple diamonds with orange outlines (i.e.  $r_{\text{src}}$ ). Blue circles with red outlines are models which are considered ‘fixed’ by certain data in this analysis, though in principle in future analyses they should be left free. The only substantial difference between the calibration and antenna panels is that  $T_{\text{ant}}$  is a model instead of input data. Thus, the calibration parameters (top green triangles) are more effectively inferred from the calibration data, while the sky parameters (lower green triangles) require antenna data. In practice, the parameters co-vary and may be influenced indirectly through any measurement.

the realization of the load power,  $p_L$ . Here,  $q_{\text{src}}$  is a random value for a *single integration* (i.e. approximately 40 seconds worth of total measurement). We shall denote an average of  $N$  such integrations as  $\bar{q}_{\text{src}}$ . Note that to first-order, the receiver gain is cancelled in  $q_{\text{src}}$ , as it is present in each of the terms in both numerator and denominator<sup>10</sup>.

The three measured powers may be modeled using the noise-wave formalism (Meys 1978). Following M17, we write

$$\langle p_{\text{src}} \rangle = g [c_{\text{src}} T_{\text{src}} + \mathbf{k}_{\text{src}} \cdot \mathbf{T}_{\text{NW}}] + T_{\text{inst}}, \quad (10)$$

$$\langle p_L \rangle = g^* [h T_L] + T_{\text{inst}}, \quad (11)$$

$$\langle p_{\text{LNS}} \rangle = g^* [h(T_L + T_{\text{NS}})] + T_{\text{inst}}, \quad (12)$$

where

$$h = 1 - |\Gamma_{\text{inst}}|^2 \quad (13)$$

<sup>10</sup> This assumes, of course, that the receiver gain is stable over timescales of ~ 40 sec.

is the ‘power transfer efficiency’ of the instrument and  $\Gamma_{\text{inst}}$  is the complex-valued reflection coefficient of the receiver (measured in terms of the ‘ $S_{11}$ ’). Here,  $\mathbf{T}_{\text{NW}} = [T_{\text{unc}}, T_{\text{cos}}, T_{\text{sin}}]^T$  are the *noise-wave* temperatures, which quantify standing-wave contributions of the noise reflected from the receiver back to the antenna. Here ‘unc’ refers to the uncorrelated portion of the noise-wave, while ‘cos’ and ‘sin’ refer to the two correlated portions that in and out of phase.

Further note that the frequency-dependent gain is different for the input source and the internal loads. This is due to a small internal path difference on account of the switch. It is convenient to write  $g^* = g(1 + \delta_g)$ .

The coefficients  $\mathbf{k}_{\text{src}}$  and  $c_{\text{src}}$  are frequency-dependent functions of the reflections coefficients of the instrument,  $\Gamma_{\text{inst}}$  and input source,  $\Gamma_{\text{src}}$ . M17 provides the values of the coefficients as

$$c_{\text{src}} = (1 - |\Gamma_{\text{src}}|)^2 F_{\text{src}}^2 \quad (14)$$

$$\mathbf{k}_{\text{src}} = \left[ |\Gamma_{\text{src}}|^2 |F_{\text{src}}|^2, |\Gamma_{\text{src}}| |F_{\text{src}}| \cos \alpha_{\text{src}}, |\Gamma_{\text{src}}| |F_{\text{src}}| \sin \alpha_{\text{src}} \right] \quad (15)$$

where

$$F_{\text{src}} = \frac{\sqrt{h}}{1 - \Gamma_{\text{inst}} \Gamma_{\text{src}}}, \text{ and } \alpha_{\text{src}} = \arg(\Gamma_{\text{src}} F_{\text{src}}). \quad (16)$$

Both  $\Gamma_{\text{src}}$  and  $\Gamma_{\text{inst}}$  are independently measured in the lab (or, in the case of the antenna, repeatedly in the field), with their own thermal and systematic uncertainties. In general, our calibration likelihood should directly include the raw  $S_{11}$  measurements (from which  $\Gamma_{\text{inst}}$  and  $\Gamma_{\text{src}}$  are computed), with an estimate of their noise properties. However, our focus in this paper is not the estimation of  $\Gamma$ , and we ignore the thermal uncertainty in the measurements for now, instead using best-fit Fourier-series models to characterize  $\Gamma_{\text{src}}(\nu)$  and  $\Gamma_{\text{inst}}(\nu)$ .

Inserting the models for the internal powers into Eq. 9, we find

$$q_{\text{src}} = \frac{c_{\text{src}} T_{\text{src}} + \mathbf{k}_{\text{src}} \cdot \mathbf{T}_{\text{NW}} - h T_{\text{L}} + [n_{\text{src}} - n_{\text{L}}] / g}{(1 + \delta_g) h T_{\text{NS}} + [n_{\text{LNS}} - n_{\text{L}}] / g}. \quad (17)$$

While the distribution of the noise in the numerator and denominator are both Gaussian, the distribution of  $q_{\text{src}}$  is not in detail – it is the distribution of a ratio of correlated Gaussian random variables. In general, with knowledge of the variance of  $(n_{\text{src}}, n_{\text{L}}, n_{\text{LNS}})$ , which themselves depend on the various temperatures involved, one can derive the distribution of  $q_{\text{src}}$ . In practice, doing so is rather complicated, and we defer this computation to future work. In this paper, we merely note that if  $T_{\text{NS}}$  is large compared to  $T_{\text{L}}$  (which is true for EDGES), the distribution of  $q_{\text{src}}$  is empirically close to Gaussian (cf. Fig. 2), with some evidence in our particular data for small non-Gaussianities in our shorted cable input (for models that account for non-Gaussianities, see Scheutwinkel et al. 2022a). We may also approximate the covariance as diagonal, as long as we average together  $\sim 16$  adjacent raw frequency channels (cf. Murray 2022a, for details). It can thus be approximately described simply by its expectation,  $\langle q_{\text{src}} \rangle$  and variance,  $\sigma_{q_{\text{src}}}^2$ . That is, we approximate

$$q_{\text{src}} \approx \langle q_{\text{src}} \rangle + n_{\text{src}} \quad (18)$$

$$n_{\text{src}} \sim \mathcal{N}(0, \sigma_{q_{\text{src}}}^2). \quad (19)$$

To estimate the variance, we assume the time axis to be statistically stationary<sup>11</sup> so that we compute

$$\sigma_{q_{\text{src}}}^2 = \langle (q_{\text{src}}(t) - \bar{q})(q_{\text{src}}(t) - \bar{q}) \rangle_t. \quad (20)$$

The expectation,  $\langle q_{\text{src}} \rangle$ , can be approximated by taking the second-order Taylor expansion of the expectation of a ratio, Eq. A1, applying it to the RHS of Eq. 9. We find that

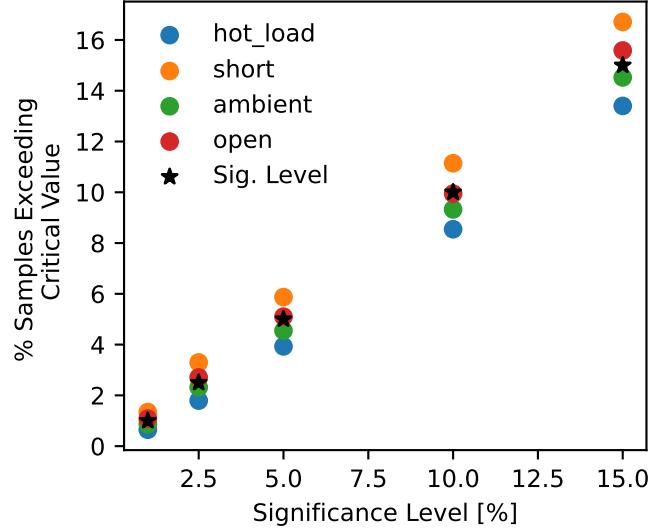
$$\langle q_{\text{src}} \rangle \approx \frac{c_{\text{src}} T_{\text{src}} + \mathbf{k}_{\text{src}} \cdot \mathbf{T}_{\text{NW}} - h T_{\text{L}}}{(1 + \delta_g) h T_{\text{NS}}} \left( 1 - \delta_0 + \delta_1^2 - \dots \right), \quad (21)$$

where the  $\delta_i$ 's are small dimensionless numbers dependent on the various source temperatures<sup>12</sup>. We ignore the terms involving  $\delta_i$  in this work, as they are very small (as long as  $T_{\text{NS}}$  is sampled with high signal-to-noise, and is large compared to  $T_{\text{L}}$ ).

We would like to solve for the noise-wave parameters and the internal temperatures of the receiver. Notice that with the exception

<sup>11</sup> This has been verified for calibration sources by using an augmented Dickey-Fuller test, which yields  $p$ -values of order  $10^{-28}$  or less for all sources. Note that this assumption is only made for *calibration sources*, not the in-field antenna, whose variance has sidereal dependence.

<sup>12</sup> In detail, this expansion makes  $q_{\text{src}}$  *not* linear in  $T_{\text{src}}$ , as it appears in the  $\delta_i$  terms in complicated ways. Nevertheless, this effect is small so long as  $T_{\text{NS}}$  is large.



**Figure 2.** Validation that the distribution of  $\mathbf{Q}$  is Gaussian. The Anderson-Darling (AD) metric was computed for all samples in a particular channel (where samples were taken over time from a single calibration input). This plot shows the percentage of channels where the AD metric was greater than the critical value for a particular significance level, i.e. the number of channels for which *rejection* of the hypothesis of Gaussianity is expected to be inappropriate to a certain level. For example, the right-most points show the percentage of channels for which Gaussianity can be rejected, while expecting 15% of the rejections to be incorrect. Colored points above each black star indicate that there is an excess of number of channels for which Gaussianity can be rejected, providing some evidence that the total spectrum has some non-Gaussianity. We find that only the shorted cable has some evidence of non-Gaussianity, and it is marginal (at most an excess of  $\sim 2\%$  of channels are considered non-Gaussian).

of  $T_{\text{NS}}$ , the equation is linear in its parameters. This is made more clear by re-writing our model as

$$T'_{\text{NS}} \langle q_{\text{src}} \rangle - \rho_{\text{src}} T_{\text{src}} \approx \frac{\mathbf{k}_{\text{src}}}{h} \cdot \mathbf{T}_{\text{NW}} - T'_{\text{L}} = \kappa_{\text{src}} \cdot \mathbf{T} \quad (22)$$

where

$$\rho_{\text{src}} = c_{\text{src}} / h, \quad (23)$$

$$T'_{\text{L}} = (1 + \delta_g) T_{\text{L}} \quad (24)$$

$$T'_{\text{NS}} = (1 + \delta_g) T_{\text{NS}}, \quad (25)$$

$$\kappa_{\text{src}} = [\mathbf{k}_{\text{src}} / h, -1] \text{ and} \quad (26)$$

$$\mathbf{T} = [\mathbf{T}_{\text{NW}}, T'_{\text{L}}]. \quad (27)$$

Since  $T'_{\text{L}}$  and  $T'_{\text{NS}}$  share the same essential properties as  $T_{\text{L}}$  and  $T_{\text{NS}}$  (i.e. they are smooth over frequency), it is just as reasonable to estimate them instead.

Inverting Eq. 22, we find that an estimate of the input source temperature may equivalently be written as a linear transformation of the measured  $q_{\text{src}}$ :

$$\hat{T}_{\text{src}} = T_0^{\text{src}} q_{\text{src}} + T_1^{\text{src}}, \quad (28)$$

where the sampling distribution of  $\hat{T}_{\text{src}}$  is Gaussian with variance  $T_0^{\text{src}} \sigma_{q_{\text{src}}}^2$ , and

$$T_0^{\text{src}} = T'_{\text{NS}} / \rho_{\text{src}}, \text{ and } T_1^{\text{src}} = -\kappa_{\text{src}} \cdot \mathbf{T} / \rho_{\text{src}}. \quad (29)$$

These two temperatures will be helpful in understanding the overall multiplicative and additive effects of the signal chain.

## 4.2 A Naive Calibration Likelihood

To infer the noise-wave parameters, the EDGES experiment takes the receiver to the lab, and replaces the antenna with four *known* input sources,  $\text{src} \in \mathcal{S}_{\text{cal}}$ , where  $\mathcal{S}_{\text{cal}}$  is the set {amb, hot, short, open}. Each source in  $\mathcal{S}_{\text{cal}}$  has different reflection characteristics as a function of frequency.

We measure three primary quantities for each source as a function of frequency: (i) spectra,  $\mathbf{q}_{\text{src}}$ , (ii) physical temperature,  $\mathbf{T}_{\text{src}}$  and (iii) reflection coefficient  $\mathbf{\Gamma}_{\text{src}}$ . Of these, in this paper we consider only the spectra to have non-negligible uncertainty.

We seek to generate posteriors on models for the noise-wave temperatures as well as the load and noise-source temperatures. We introduce some book-keeping notation for these sets of parameters; let  $\mathcal{T}_{\text{NW}}$  be the set of labels corresponding to the noise-wave terms:  $\mathcal{T}_{\text{NW}} = \{\text{unc, cos, sin}\}$ , and  $\mathcal{T}_{\text{intload}} = \{\text{L, NS}\}$  the labels corresponding to internal load temperature terms. Then the full set of modeled temperature terms is  $\mathcal{T} = \mathcal{T}_{\text{NW}} \cup \mathcal{T}_{\text{intload}}$ . An alternative useful partition of  $\mathcal{T}$  is into the terms that can be treated as linear (in the sense of App. B),  $\mathcal{T}_{\text{lin}} = \mathcal{T}_{\text{NW}} \cup \{\text{L}\}$  and those that must be considered non-linear,  $\mathcal{T}_{\text{nl}} = \{\text{NS}\}$ .

The modeled temperature noise-wave temperatures and the load and noise-source temperatures are not arbitrary; they are assumed to be smooth functions of frequency. We thus model each temperature as a low-order polynomial:

$$\mathbf{T}_p = \sum_i^{N_{\text{terms}}^p} \boldsymbol{\theta}_i^p \left( \frac{\nu}{\nu_0} \right)^i = \boldsymbol{\Psi} \boldsymbol{\theta}_p, \quad p \in \mathcal{T} \quad (30)$$

where  $\nu$  is the vector of observed frequencies (and the exponentiation is implicitly element-wise),  $\boldsymbol{\theta}_p$  are the unknown coefficients for temperature  $p$  (in temperature units) and  $\boldsymbol{\Psi}$  is the  $N_{\nu} \times N_{\text{terms}}^p$  matrix of polynomial basis vectors.

Let  $\mathbf{r}_{\text{src}}$  be the length- $N_{\nu}$  model-residual vector for a particular input source,  $\text{src} \in \mathcal{S}_{\text{cal}}$ :

$$\mathbf{r}_{\text{src}} = \mathbf{q}_{\text{src}} \circ \mathbf{T}_{\text{NS}} - \boldsymbol{\rho}_{\text{src}} \circ \mathbf{T}_{\text{src}} - \sum_{p \in \mathcal{T}_{\text{lin}}} \boldsymbol{\kappa}_{p,\text{src}} \mathbf{T}_p. \quad (31)$$

Under our assumptions of Gaussianity of  $\mathbf{q}$ , and independence between frequency channels, as justified in the previous subsection, we then have that the distribution of  $\mathbf{r}_{\text{src}}$  is a multivariate Gaussian with zero mean and diagonal covariance given by

$$\boldsymbol{\Sigma}_{\text{src}} = \mathbf{I}_{N_{\nu}} \boldsymbol{\sigma}_{q,\text{src}}^2 \mathbf{T}_{\text{NS}}^2. \quad (32)$$

Then, our final calibration likelihood is

$$\mathcal{L}_{\text{cal}}(\boldsymbol{\theta}_{\mathcal{T}}) = \prod_{\text{src} \in \mathcal{S}_{\text{cal}}} \mathcal{L}_{\text{src}}(\mathbf{q}_{\text{src}} | \boldsymbol{\theta}_{\mathcal{T}}), \quad (33)$$

$$\mathcal{L}_{\text{src}}(\mathbf{q}_{\text{src}} | \boldsymbol{\theta}_{\mathcal{T}}) \propto \sqrt{|\boldsymbol{\Sigma}_{\text{src}}^{-1}|} \exp \left\{ -\frac{1}{2} \mathbf{r}_{\text{src}}^T \boldsymbol{\Sigma}_{\text{src}}^{-1} \mathbf{r}_{\text{src}} \right\}, \quad (34)$$

This is conceptually the simplest representation of the likelihood, but for the purpose of exploiting the analytic marginalization of linear parameters (cf. §3.4), it is helpful to separate the linear parameters into a single term represented by a product of a matrix with the linear parameter vector. App. C details this process, showing that  $\boldsymbol{\theta}_p$  are linear for  $p \in \mathcal{T}_{\text{lin}}$ .

## 4.3 Comparison of likelihood to iterative approach

The fiducial calibration temperatures,  $\mathbf{T}_p$ , used in B18 were determined by an iterative process outlined in M17. This process has several differences with respect to the calibration likelihood presented

here, which result in somewhat differing calibration solutions. Our purpose in this paper is to understand the *posterior distribution* of the inferred cosmic signal, given uncertainties in the calibration parameters. To do this, we wish to keep the ‘point estimate’ of the calibration solutions rather similar to the results of B18, by choosing methods and other parameters that match as closely as possible. Thus, it is important to understand where differences in the point estimates of the calibration arise with respect to the previous methods, before moving on to propagating uncertainties forward to field data.

The primary differences between the solutions used in B18 and point estimates (nominally maximum-likelihood estimates) from our likelihood are as follows:

(i) B18 smoothed calibration data into 8-channel bins (with a Gaussian filter) before fitting calibration polynomials, whereas we bin the spectra into 32-channel bins with a top-hat filter. The purpose of this change is to ensure that the covariance is diagonal (cf. §4.1).

(ii) B18 inherently treated each frequency and source with the same weight (i.e. variance). Since doing so would bias our posterior distribution, we cannot use this assumption, and instead use the empirically-determined variance for each source and frequency.

(iii) The iterative method separates sources and model parameters. The amb and hot sources are essentially zero-length cables with extremely good impedance match to the receiver. This means that  $\Gamma_{\text{amb}}$  and  $\Gamma_{\text{hot}}$  are extremely small. Conversely, the cable measurements (open and short) are designed to have high reflections, which makes it possible to characterize the noise-wave temperatures. Since the solutions are very sensitive to the accuracy of the  $\Gamma_{\text{src}}$  measurements, the iterative solutions for  $\mathbf{T}_{\mathcal{T}}$  use the cable measurements *only* to directly fit the noise-wave terms,  $\mathcal{T}_{\text{NW}}$ , using amb and hot to fit the internal load temperatures,  $\mathbf{T}_{\mathcal{T}_{\text{intload}}}$ . This avoids leaking any potential biases from inaccurate  $\Gamma_{\text{open}/\text{short}}$  measurements to  $\mathbf{T}_{\mathcal{T}_{\text{intload}}}$ . This is not possible for the likelihood, which consistently accounts for all data.

Fig. 3 summarizes the differences in the calibration solutions for our likelihood compared to those used in B18, where all choices are kept as similar as possible with the exception of the three differences just mentioned. Note that B18 uses  $N_{\text{terms}}^p = c_{\text{terms}} = 6$  for  $p \in \mathcal{T}_{\text{intload}}$  and  $N_{\text{terms}}^p = w_{\text{terms}} = 5$  for  $p \in \mathcal{T}_{\text{NW}}$ . We plot each calibration solution individually in the left-hand panels, as a percentage difference from the solution used in B18. In the top right-hand panel we plot the induced absolute difference in the calibrated, beam-corrected sky temperature:

$$\Delta T_{21} = \hat{T}_{\text{sky, bc}}^{\text{B18}} - \hat{T}_{\text{sky, bc}}, \quad (35)$$

where  $\hat{T}_{\text{sky, bc}}^{\text{B18}}$  is the publicly-available calibrated sky temperature from B18 and  $\hat{T}_{\text{sky, bc}}$  is obtained by *de*-calibrating  $\hat{T}_{\text{sky, bc}}^{\text{B18}}$  with the exact B18 calibration solution, and *re*-calibrating with a new solution as specified in the legend (via Eq. E1). The lower right-hand panel shows the inferred cosmic signal from this new ‘recalibrated’ spectrum. Note that the inferred cosmic signals shown here are *not* jointly estimated along with the calibration, as we shall do in §5. Instead, after recalibrating the spectrum we fit a simple model,  $T_{\text{sky}} = T_{\text{FG}} + T_{21}$ , where the foreground term is given by a 5-term ‘linlog’ model (cf. Eq. 52) and  $T_{21}$  is a flattened Gaussian (cf. Eq. 40). The fit over the 21 cm parameters is performed via global minimization routine, where for each parameter-set choice, we find the MLE of the foregrounds using linear algebra. Thus, this lower-right panel gives a fast indication of how much the difference in calibration affects our target inference.

We first concentrate attention on the solid blue line, representing

the solutions using the same iterative procedure as used in B18 (but with updated Python code instead of the original C-code). While some small differences are apparent (especially in  $T_{\text{sin}}$ , which deviates by tens of percent at the lowest frequencies), their overall effect is extremely small, as evidenced by the excellent agreement of the inferred cosmic signal. To test difference (i) – i.e. increase frequency bin size – we plot in orange the result of binning in 32-channel bins, and still performing an iterative fit. While this results in some marked differences in the calibration parameters (up to 1% in  $T_{\text{cos}}$ ), the differences in the inferred cosmic signal are negligible. The likely reason for the differences, especially in the scaling,  $T'_{\text{NS}}$ , is the change from a Gaussian filter to a top-hat filter (blue to orange), rather than the change in bin size. We find that for a bin size of 32 channels (both for the top-hat and the Gaussian filter), the top-hat filter systematically produces higher average values for the hot load spectrum, at the level of  $\sim 0.05\%$ . A higher hot load temperature corresponds to a matching *decrease* in  $T'_{\text{NS}}$ , as witnessed. We choose the top-hat filter because it induces less correlation between neighbouring bins. Regardless of this difference, as already mentioned the estimated 21 cm signal is essentially unaffected. This is for two reasons: (i) the differences are smooth in frequency and largely ameliorated by the flexible foreground model, and (ii) the calibration parameters have correlated effects which partially cancel each other in the final calibrated spectrum.

However, we find a very strong difference when we use our default maximum-likelihood model (pink dotted line). In particular, we note how the calibration solutions incur extra spectral structure, and the inferred cosmic signal is significantly more shallow. We first ask whether this difference may be due to difference (ii), i.e. the fact that our likelihood accounts for frequency- and source- dependent noise. The dashed yellow curves shows the results of a likelihood in which the variance is constant (since this figure show a maximum-likelihood estimate, the magnitude of the variance is inconsequential) over frequency and source. While the agreement between this curve and B18 is somewhat better, this does not seem to be the dominant difference.

Difference (iii) concerns weighting different input sources for different parameters. The justification for the iterative method (largely) ignoring the contribution of the cable for estimating  $T'_{\text{L}}$  and  $T'_{\text{NS}}$  is that there are potential biases in the measurements of  $\Gamma_{\text{open}}$  and  $\Gamma_{\text{short}}$  that skew the estimates. Fig. 4 demonstrates that for this calibration dataset, this is indeed the case. It shows significant ( $>20\sigma$ ) ‘wiggles’ in the *open* and *short* residuals for the iterative approach (and all approaches). The default likelihood (solid pink) is able to achieve far better precision on the *short* measurement, but sacrifices accuracy on all other measurements to do so. Since the structure being fit in the *short* measurement is *a priori* expected to be due to biases in  $\Gamma_{\text{short}}$ , the increased residuals in the other measurements can be considered ‘leakage’ from the bias in *short*. In other words, we would like the estimation of scale and offset parameters to be dominated by the very accurate *ambient* and *hot\_load* measurements, restricting any potential biases in the cable measurements to affect the noise-wave parameters. To check whether this difference is dominating our discrepancy with B18, we fit a likelihood in which the cable measurements are significantly down-weighted, by artificially increasing their variance<sup>13</sup>. The result in both Figs. 3 and 4 is shown in grey.

<sup>13</sup> We note that doing so is *not self-consistent*. While we may achieve reasonable maximum a-posteriori estimates based on intuitive expectations with this approach, the posteriors will have an incorrect spread, and the Bayesian evidence will be wrong.

Importantly, the scale and offset parameters exhibit far less spectral structure than the default likelihood (pink dotted), and the inferred cosmic signal is in much better agreement with B18. In Fig. 4, we notice that while the ‘down-weighted cable’ model performs similarly to the default likelihood in calibrating the cable measurements, its errors there do not leak into the loads. The remaining differences between the down-weighted cable likelihood and B18 are a combination of the frequency- and source-dependent weighting, and differences in the exact strength with which the cable measurements are down-weighted (the iterative method does not afford an obvious translation of its effective weighting of the measurements).

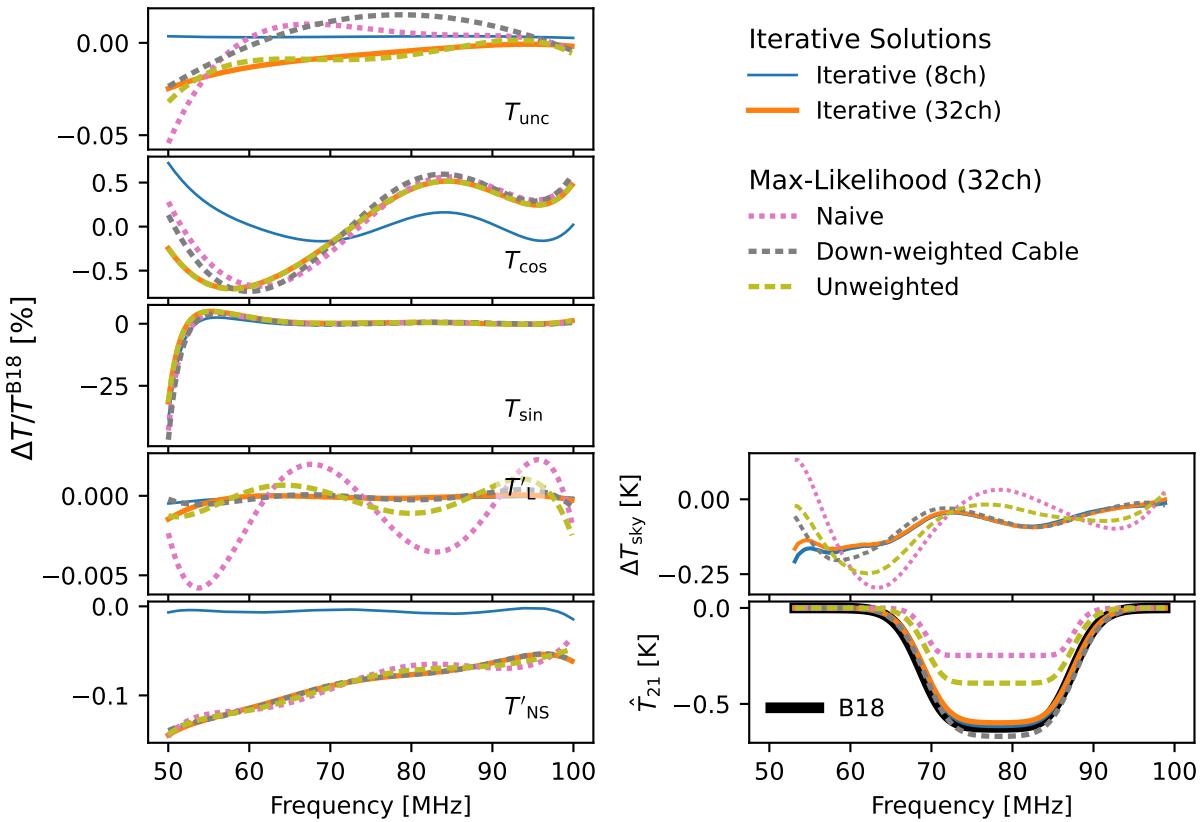
#### 4.4 A De-biased Calibration Likelihood

The previous section showed that there is a significant systematic in the cable measurements used to derive our calibration, and that the effect of this systematic can be largely avoided by minimizing the impact of the cable measurements on the estimation of the load and noise-source temperatures. In this subsection, we outline a modified likelihood that takes advantage of this knowledge so as to decrease its bias. We note that the ‘down-weighted cable’ likelihood used in §4.3 is not appropriate, since its noise model is known to be incorrect, and therefore it will yield incorrect posterior distributions and Bayesian evidence.

The first question is whether the iterative solution really does avoid being biased by the cable systematic. The proper way to answer this would be to construct a model for the cable measurements that included a flexible  $\Gamma_{\text{src}}$  systematic component, then determine the model with the highest Bayesian evidence. However, choosing a flexible form for the (complex-valued)  $\Gamma_{\text{src}}$  that is able to capture the systematic is a rather involved task, and we defer it to future work<sup>14</sup>. In the meantime, we can gain some confidence by noting Fig. 5, which shows that different numbers of  $w_{\text{terms}}$  yield largely consistent inferred cosmic signals (bottom panel). Indeed, also shown in this figure is the residuals to an antenna simulator – a known input source with  $\Gamma_{\text{src}}$  designed to approximate the antenna itself. While this source is not used to fit the calibration, it can be used to check the results. Fig. 5 indicates that  $w_{\text{terms}} = 5$  (orange) – the choice used in B18 – minimizes the its residuals. Higher  $w_{\text{terms}}$  decrease the performance of the antenna simulator, indicating that they are fitting systematics in the cable measurements themselves. This is not a perfect test. It is possible that the  $w_{\text{terms}} = 5$  fit is partially biased by cable systematics, or that the antenna simulator itself has independent systematics in its  $\Gamma_{\text{src}}$  measurement. However, without performing a full investigation into the source and nature of the cable systematic, we can be reasonably confident that  $w_{\text{terms}} = 5$  is providing a good, stable calibration.

The next question is how to define a likelihood that is able to restrict the cable measurements’ impact on the noise-wave terms. In principle, this is impossible with a self-consistent likelihood. In this paper, we take the following approach: we first perform an iterative fit, and then, given the measured temperature of the cable inputs, we ‘decalibrate’ to determine  $\hat{q}_{\text{src}}$ . To this, we add simulated Gaussian noise with variance determined empirically from the measurements. We then substitute these simulated cable ‘measurements’ for the observed data. In this way, to within correlations between the scale and offset and noise-wave parameters, we are guaranteed to obtain point-

<sup>14</sup> We can report that a simple polynomial scaling and delay is not a good model.



**Figure 3.** Summary of differences in calibration solutions with respect to those used in B18. The left column shows the five calibration parameters as a function of frequency, each displayed as a percentage deviation from B18. The right panels show the effect of the deviated solutions on the sky data, with the upper plot showing the absolute difference in (re)calibrated sky temperature (cf. Eq. E1), and the lower plot showing the ‘best-fit’ cosmic signal assuming a 5-term LinLog foreground model (Eq. 52). Solid blue curves represent a calibration performed with the traditional iterative algorithm of M17, and these match B18 very well. Smoothing over 32 channels (solid orange) does not affect the inferred signal significantly. All other (dashed) curves represent maximum-likelihood calibrations using Eq. 33, with varying assumptions. Pink is our naive likelihood in which all sources and frequencies have variance empirically estimated. This results in large differences with respect to B18 for the inferred signal. Assuming a constant variance with respect to frequency and sources (yellow) increases correspondence, but is still quite discrepant. The discrepancy can be largely removed by down-weighting the cable measurements (grey), minimizing their contribution to the estimation of the offset and scale parameters ( $T'_L$  and  $T'_{\text{NS}}$ ). See text for details.

estimates of the noise-waves consistent with the iterative approach, with a posterior distribution consistent with the observed noise<sup>15</sup>.

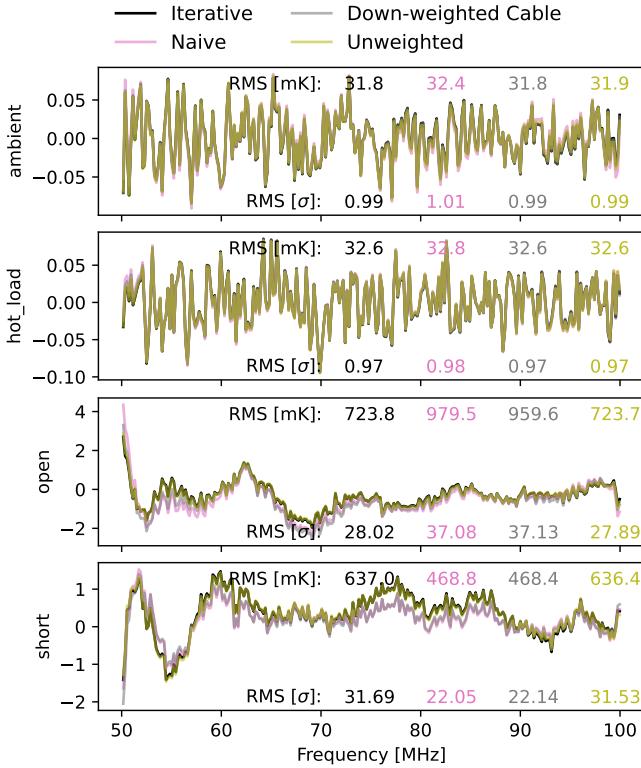
Table 3 shows the Bayesian evidence computed by sampling models and data computed in this way, where the initial iterative calibration (to set the simulated cable data) was performed with the default  $c_{\text{terms}} = 6$  and  $w_{\text{terms}} = 5$  as used in B18. The highest Bayesian evidence is obtained for the fiducial number of terms. For  $w_{\text{terms}}$  this is merely taken to be a consistency check of the code, as the highest evidence must be obtained for the  $w_{\text{terms}}$  used in the simulation. However, this reasoning does not apply as strongly to  $c_{\text{terms}}$ , since this is predominantly set by the `ambient` and `hot_load` sources, which are not simulated. Thus, the fact that we obtain the highest

evidence for  $c_{\text{terms}} = 6$  is an indication that we truly require 6 terms for  $T'_L$  and  $T'_{\text{NS}}$ <sup>16</sup>. This is further emphasised by the fourth column of Table 3, which shows the Bayesian evidence for models fit to data in which the simulated cable measurements were constructed based on fits with  $(c, w) = (8, 5)$ . Even in this case, the strongest evidence is obtained for a model with  $c = 6$ , which is a strong justification for our choice to use this number of terms for the rest of this paper.

As a further check, we show the calibration residuals of the three highest-evidence models from Table 3 in Fig. 6. As expected, each can perfectly reproduce the cable measurements, while differences in the loads and antenna simulator are very small. There are noticeable differences in the inferred cosmic signal, however even these are stable to within the expected posterior reported in B18.

<sup>15</sup> In this work we use one specific noise realization for this simulated cable data. Given that we are in a high SNR regime, we do not expect results to be sensitive to the realization itself.

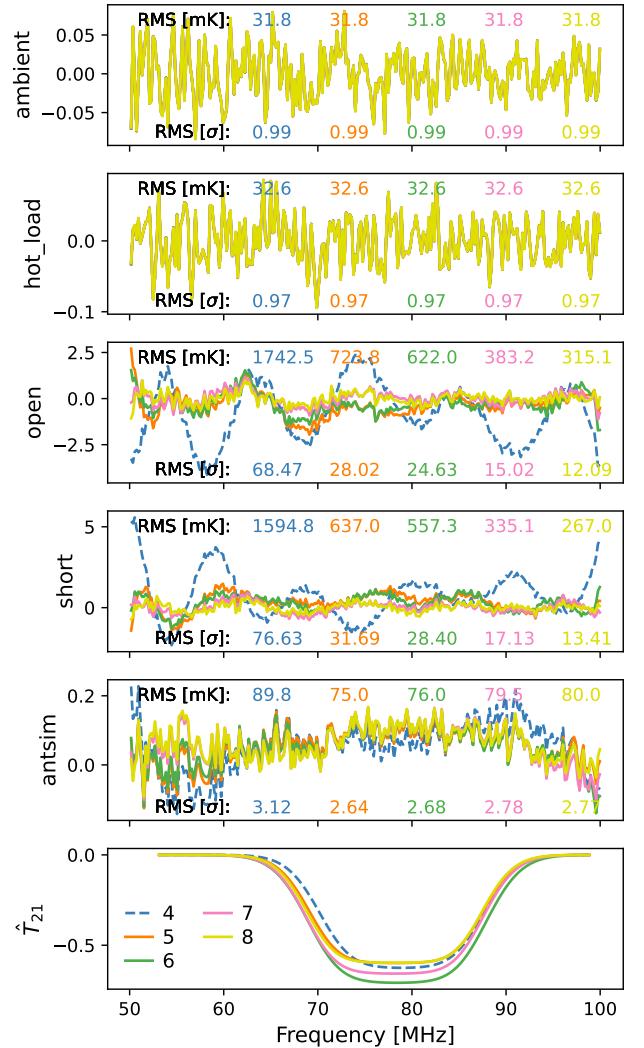
<sup>16</sup> We note that using 6 terms also provides the best RMS on the antenna simulator, which is why it was chosen to be used in B18.



**Figure 4.** Residuals to the known temperature for each input source, for a range of likelihood models (see caption of Fig. 3 for a key to the models). Also listed for each model and source are the RMS of the residuals, in units of mK and also numbers of standard deviation (according to the empirical noise model). Overall, differences between models are small. However, the cable measurements (bottom two rows) have larger weighted RMS (in terms of numbers of standard deviations) when their contribution is down-weighted. Conversely, when down-weighting the cable measurements, the fit to the ambient and hot load are significantly improved.

| $c_{\text{terms}}$ | $w_{\text{terms}}$ | $\ln \mathcal{Z} [6,5]$ | $\ln \mathcal{Z} [8,5]$ |
|--------------------|--------------------|-------------------------|-------------------------|
| 6                  | 4                  | -10065.5                |                         |
| <b>6</b>           | <b>5</b>           | <b>3083.7</b>           |                         |
| 6                  | 6                  | 3075.0                  |                         |
| 3                  | 5                  | -1787.6                 |                         |
| 4                  | 5                  | 2944.7                  |                         |
| 5                  | 5                  | 3029.1                  | 3014.7                  |
| <b>6</b>           | <b>5</b>           | <b>3083.7</b>           | <b>3078.7</b>           |
| 7                  | 5                  | 3075.4                  | 3072.8                  |
| 8                  | 5                  | 3068.2                  | 3068.3                  |
| 9                  | 5                  | 3062.2                  |                         |

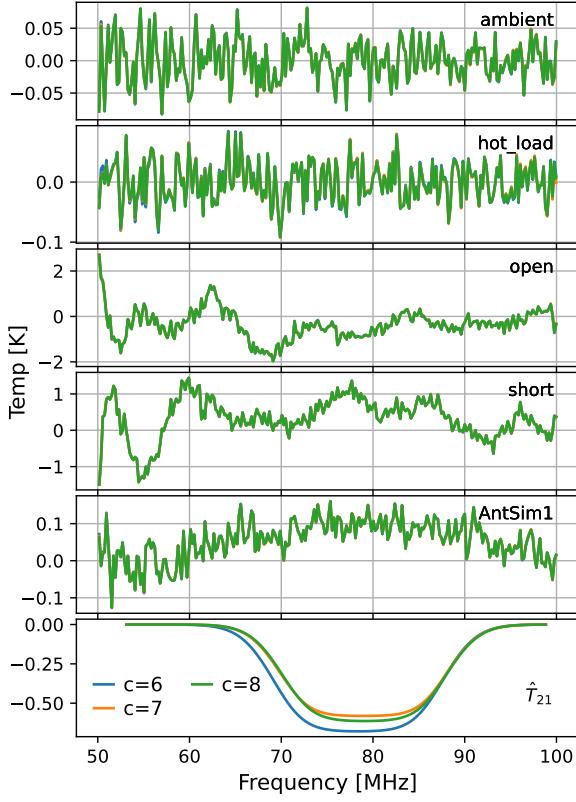
**Table 3.** Bayesian evidence for pure calibration models (i.e. without field data) where the cable measurements are replaced by simulated data constructed from less-biased iterative solutions. The iterative solution uses  $c = 6$  and  $w = 5$ . Since only the cable measurements are substituted for simulations, only the noise-wave parameters (set by  $w_{\text{terms}}$ ) are significantly affected by this choice. Thus the top portion of the table is a check that indeed the simulations yield a maximum evidence at the notional number of terms. The lower section modifies  $c_{\text{terms}}$  and thus provides justification for choosing  $c_{\text{terms}} = 6$  to describe the calibration. We find that we obtain the maximum evidence for  $c_{\text{terms}} = 6$  whether we use simulated cable data from a fit with  $c'_{\text{terms}} = 6$  (third column) or 8 (fourth column).



**Figure 5.** Calibration source residuals to models with different number of  $w_{\text{terms}}$ . As  $w_{\text{terms}}$  increases, the load residuals (top two panels) stay the same, as they are not sensitive to  $w_{\text{terms}}$ . The cable measurements (third and fourth panels) show ever-decreasing residuals, but with a mean  $\chi^2 > 10$  even for the highest terms. We affirm that the structure being fit in the cable measurements by the extra terms is largely systematics, by noting that an antenna simulator (fifth panel) is fit more poorly for higher  $w_{\text{terms}}$ , peaking at  $w_{\text{terms}} = 5$ . Regardless, for all these models, the inferred cosmic signal is reasonably consistent.

In summary, while investigation into the source of the systematics in the cable measurements is a high priority for future work, the results of this section indicate that using an iterative approach to solve for the noise-wave parameters and  $T'_L$  and  $T'_{\text{NS}}$  produces results that maintain consistency between their respective inferred cosmic signals. We thus adopt this approach, wherein we use the iterative solutions to produce simulated cable data for our likelihood, for the remainder of this paper. Hereafter, we use this model with  $c_{\text{terms}} = 6$  and  $w_{\text{terms}} = 5$ , which maximizes the Bayesian evidence.

Finally, we show the posterior distributions of the calibration parameters and calibrated antenna simulator and field data in Fig. 7. We note that all calibration parameters have posteriors with width  $\ll 1\%$ , and are consistent with the iterative solutions, with the slight



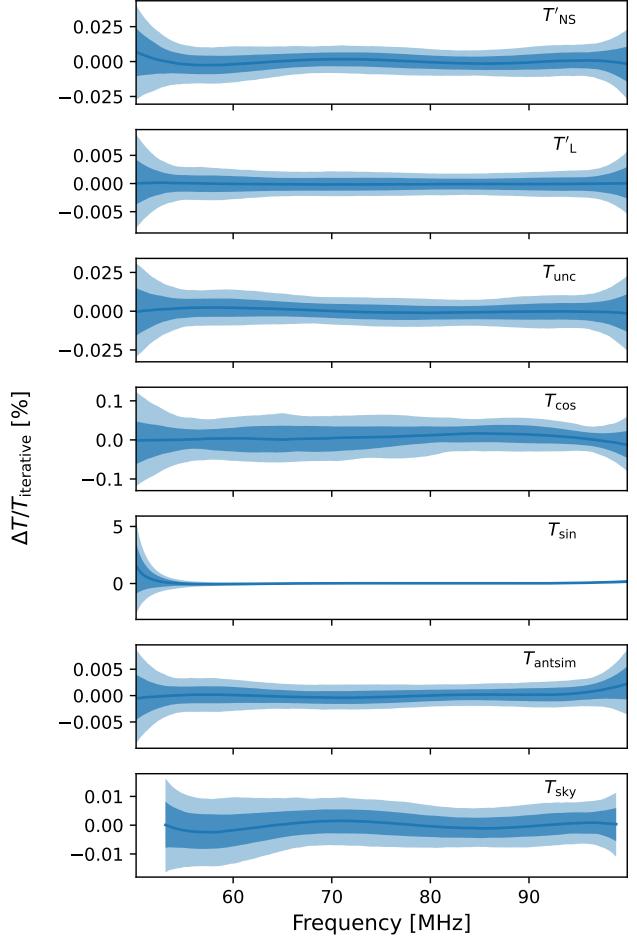
**Figure 6.** Calibration residuals for differing number of  $c_{\text{terms}}$ , using the likelihood and data outline in §4.4 – i.e. in which simulated cable measurements are injected. The models shown correspond to the three with highest evidence in Table 3. All choices of  $c_{\text{terms}}$  produce similar estimates of the cosmic signal.

exception of  $T_{\text{sin}}$ , which has a 5% width at frequencies  $< 53$  MHz. The induced extra uncertainty on the sky data is of order 0.01%. All curves exhibit higher uncertainty at the band edges, which is expected for the flexible polynomials we fit.

## 5 A PROBABILISTIC SKY DATA MODEL

We now turn to derive a probabilistic model for data measured with the EDGES antenna in the field, which will include the previously-derived calibration likelihood as a subset.

As in the previous section, throughout this section, almost all of the quantities are frequency-dependent, and thus represented by a length- $N_{\nu}$  vector. However, since frequencies are not coupled in any of the modelling steps, we will omit their frequency-dependence in this section, and write each as a scalar (to be interpreted as a single



**Figure 7.** Posterior 1- and 2- $\sigma$  regions for the calibration parameters using the pure calibration likelihood described in §4.4. Each is shown as a fractional difference with respect to the iterative solution. Also shown is the posterior re-calibrated sky temperature, whose 1- $\sigma$  posterior fully agrees with the iterative solution, and is a hundredth of a percent in width.

element of a length- $N_{\nu}$  vector<sup>17</sup>). We will explicitly note the rare quantities that are (assumed to be) independent of frequency.

The true average sky temperature is assumed to be (in absence of ionospheric distortions) merely a sum of foreground and cosmic signal:

$$T_{\text{sky}}(t) = T_{21} + \int \frac{d\Omega}{2\pi} T_{\text{FG}}(t, \Omega), \quad (36)$$

where  $t$  is the time of observation, and  $\Omega$  is solid angle in the reference frame of the antenna (e.g. azimuth and altitude) with the integral extending over the upper hemisphere. We have made the assumption that the cosmic signal is isotropic and has negligible fluctuations on the scale of the horizon. This equation adopts the formalism in which the sky drifts across the frame of reference (i.e.  $T_{\text{FG}}$  changes with time with respect to  $\Omega$ ).

The EDGES antenna does not perform an unweighted integral over the sky; it is more sensitive to regions close to zenith, and this

<sup>17</sup> This is in contrast to a continuous function of frequency, as it encodes not only the frequency dependence but also information about the frequency bin width.

sensitivity pattern is defined by its primary beam,  $B$ . Thus, EDGES in principle measures

$$T_{\text{sky,beam}}(t) = \int \frac{d\Omega}{2\pi} B(\Omega) [T_{21} + T_{\text{FG}}(t, \Omega)]. \quad (37)$$

The beam is normalized such that its integral is  $2\pi$  over the upper hemisphere. Notice that beyond an overall modification to the amplitude, the beam introduces distortions to the *spectrum* of  $T_{\text{sky,beam}}$  compared to  $T_{\text{sky}}$ . This is true even if the beam is achromatic itself, as it couples spatial structure in the sky into spectral structure. This effect is commonly known as beam chromaticity.

Nevertheless, while detailed modelling of the beam-weighted foregrounds is an important line of inquiry (Tauscher et al. 2020a,b; Mawesh et al. 2021), when averaged over a wide range of LSTs, the beam chromaticity tends to average out, as demonstrated in B18 by verifying consistency of cosmic signal estimate with and without beam correction. Thus, in this paper, given that the FG temperature is an *a priori* unknown smooth function of time and frequency, we simply replace the entire beam-weighted foreground term with a similar smooth function—allowing the higher-order terms of the unknown FG model to absorb any remaining structure from the beam—and correct for the beam chromaticity explicitly:

$$T_{\text{sky,bc}}(t) \equiv \frac{T_{\text{sky,beam}}(t)}{b_{\text{corr}}(t)} \approx T_{21} + T_{\text{BWFG}}(t), \quad (38)$$

where the ‘beam chromaticity correction’ is given by (Mozdzen et al. 2019):

$$b_{\text{corr}}(\nu) = \frac{\int d\Omega B(\Omega, \nu) T_{\text{haslam}}(\Omega, \nu_{\text{ref}})}{\int d\Omega B(\Omega, \nu_{\text{ref}}) T_{\text{haslam}}(\Omega, \nu_{\text{ref}})}, \quad (39)$$

with  $\nu_{\text{ref}} = 75$  MHz. Note that this is an approximation; it accounts for the first-order frequency-dependent effects of the beam under the assumption that the cosmic signal is isotropic on the angular scales over which the beam modulates. For an achromatic beam and accurate sky model, this correction accounts for all chromatic structure leaked from angular scales to frequency. For realistic chromatic beams, there is unavoidably residual chromatic structure (after beam correction). This is why we denote the foreground term as “ $T_{\text{BWFG}}$ ”, which indicates that the foregrounds we finally estimate must themselves account for this structure.

Throughout this work we use a phenomenological model for the 21 cm signal during Cosmic Dawn as used in B18:

$$T_{21,i} = -A \left( \frac{1 - e^{-\tau e^{\psi}}}{1 - e^{-\tau}} \right), \quad (40)$$

where

$$\psi = \frac{4(\nu - \nu_0)^2}{w^2} \log \left[ -\frac{1}{\tau} \log \left( \frac{1 + e^{-\tau}}{2} \right) \right] \quad (41)$$

and the 21 cm parameters are  $\theta_{21} = [A, \nu_0, w, \tau]$ .

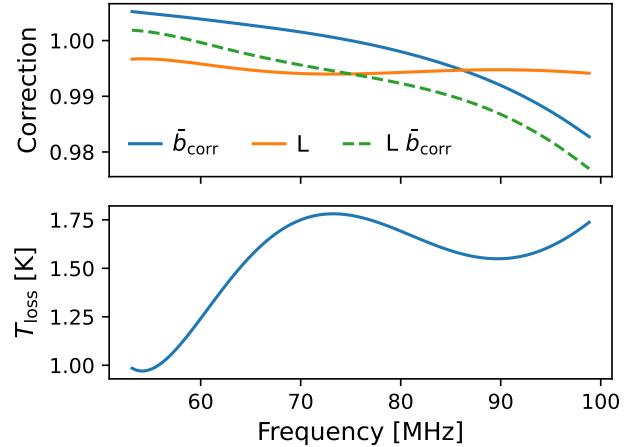
The antenna imposes additional frequency- and time-dependent gains on the incoming signal after the beam-convolution:

$$T_{\text{sky,loss}}(t) = L T_{\text{sky,beam}}(t) + (1 - L) T_{\text{amb}}(t), \quad (42)$$

where  $T_{\text{amb}}$  quantifies the *frequency-independent* ambient temperature at the antenna-site at any time, and  $L$  is considered here to be time-independent and encodes the product of losses incurred by the antenna, balun, connectors and ground:

$$L = L_{\text{ant}} L_{\text{balun}} L_{\text{conn}} L_{\text{ground}}. \quad (43)$$

In this work, we consider uncertainties in the modelling of the loss to be negligible. Its value is very close to unity for all frequencies.



**Figure 8.** Various correction factors applied to the field data, as expressed in Eq. 48. In particular, the total loss (from antenna and ground) as well as the ‘beam correction’ (Eq. 39).

Finally, the signal passes through the receiver, at which point a multiplicative gain and additive noise are imposed (cf. §4), and we recognize that the entire signal chain (including the sky) has been stochastic:

$$p_{\text{sky,meas}}(t) = g(t) T_{\text{sky,loss}}(t) + T_{\text{inst}}(t) + n_{\text{sky,meas}}, \quad (44)$$

where  $n_{\text{sky,meas}}$  is a zero-mean Gaussian random variable.

Applying the Dicke-switching and noise-wave formalism presented in §4, we find the final measured three-position-switch power ratio is given by

$$q_{\text{ant}}(t) = \frac{T_{\text{sky,loss}}(t) - T_1^{\text{ant}}}{T_0^{\text{ant}}} + n_{\text{ant}}(t), \quad (45)$$

with  $T_0^{\text{ant}}, T_1^{\text{ant}}$  given by Eq. 29, and where we assume that  $n_{\text{ant}}$  is drawn from a zero-mean Gaussian distribution with variance  $\sigma_{q,\text{ant}}^2(t)$ <sup>18</sup>.

## 5.1 Data Processing

In B18 the spectra from all times are averaged together. This clearly *loses* information, and makes it more difficult to verify the truly global nature of the cosmological background (Tauscher et al. 2020a; Liu et al. 2014), however without an accurate model of the low-frequency sky, it is necessary in order to average down systematics that decorrelate as the sky evolves.

Data-flagging and averaging was applied to an estimate of the *sky temperature* rather than the raw measured quotients. That is, the data

$$\widehat{T}_{\text{sky,bc}}(t) = \frac{1}{L b_{\text{corr}}(t)} \left[ \widehat{T}_0^{\text{ant}} q_{\text{ant}}(t) + \widehat{T}_1^{\text{ant}} - (1 - L) T_{\text{amb}}(t) \right] \quad (46)$$

are used to evaluate flags and perform averaging. Here, the estimated calibration functions were computed based on the iterative scheme outlined in §4.3, with  $(c_{\text{terms}}, w_{\text{terms}}) = (6, 5)$ .

<sup>18</sup> Note that  $n_{\text{ant}}$  is different than  $n_{\text{sky,meas}}$  and while the latter is normally distributed to a very good approximation, the former is only Gaussian under the approximations outlined in §4.1.

The final averaged spectrum is

$$\widehat{\bar{T}}_{\text{sky,bc}} = \frac{1}{w} \sum_j \xi(t_j) \widehat{T}_{\text{sky,bc}}(t_j), \quad (47)$$

$$\approx \frac{1}{L\bar{b}_{\text{corr}}} \left[ \widehat{T}_0^{\text{ant}} \bar{q}_{\text{ant}} + \widehat{T}_1^{\text{ant}} - \bar{T}_{\text{loss}} \right], \quad (48)$$

where  $\bar{T}_{\text{loss}} \equiv (1 - L)\bar{T}_{\text{amb}}$  is the mean loss temperature,  $\xi \in \{0, 1\}$  are the per-frequency flags at each time-stamp, and  $w = \sum_j \xi(t_j)$  is the number of (unflagged) samples per-frequency (over all measured time samples), and the effective quotient, ambient temperature and beam correction are given by weighted average over time samples,

$$\bar{X} = \frac{1}{w} \sum_j \xi(t_j) X(t_j). \quad (49)$$

Note the the second relation (Eq. 48) is an approximation, due to the fact that the beam correction term is time-dependent and the mean of a product is the not the product of means. Nevertheless, we expect this effect to be small, and delay its proper treatment to future work.

Note that  $\widehat{\bar{T}}_{\text{sky,bc}}$  is precisely the publicly-available spectrum shown in Fig. 1 of B18.

Our likelihood does not use  $\widehat{\bar{T}}_{\text{sky,bc}}$  directly, but instead uses the more basic quantity  $\bar{q}_{\text{ant}}$ . This has the benefit of being only slightly dependent on the estimates  $\widehat{T}_0^{\text{ant}}, \widehat{T}_1^{\text{ant}}$ . This slight dependence arises through the fact that the flags,  $\xi$ , are computed based on the calibrated data. This dependence is extremely insensitive to small changes in the calibration temperatures, because the flags are empirically assigned based on a non-parametric estimate of whether the datum is an outlier (over either the frequency or LST axis). Changes in the calibration temperatures introduce very smooth changes in the data as a function of frequency/LST, and therefore are highly unlikely to change the flags. We obtain  $\bar{q}_{\text{ant}}$  simply by inverting Eq. 48 using the publicly-available data  $\widehat{\bar{T}}_{\text{sky,bc}}$  and values for  $L, \bar{b}_{\text{corr}}, \bar{T}_{\text{loss}}$  and  $\widehat{T}_{0,1}^{\text{ant}}$  obtained directly from the B18 analysis code.

## 5.2 Data Model

The antenna-sourced power quotient,  $q_{\text{ant}}$  can be treated on the same footing as the input calibration sources in the calibration likelihood, Eq. 33, i.e. by expanding the set of sources summed over to  $S_{\text{joint}} = S_{\text{cal}} \cup \{\text{ant}\}$ . Just like the other input sources, its distribution is assumed to be well-approximated by an uncorrelated multivariate Gaussian, with mean  $\langle \bar{q}_{\text{ant}} \rangle$  and variance  $\sigma_{q_{\text{ant}}}^2$ .

In contrast to the other sources, however, we do not have a low-noise measurement of the true input temperature of the antenna,  $T_{\text{src}} \equiv T_{\text{ant}}$ . Instead, we have a *model* for the true temperature:

$$T_{\text{ant}} = L\bar{b}_{\text{corr}} \left[ \bar{T}_{\text{BWFG}} + T_{21} \right] + \bar{T}_{\text{loss}} \quad (50)$$

$$\equiv T_0^{\text{ant}} \langle \bar{q}_{\text{ant}} \rangle + T_1^{\text{ant}}. \quad (51)$$

where  $\bar{T}_{\text{BWFG}}$  models the time-averaged beam-weighted foregrounds, and should be a smooth function of frequency.

The variance is in principle an unknown function that should be modelled and inferred. However, in this work we simply estimate the variance by analysis of the residuals of the data to high-order smooth models. Note that this variance model is different to that use in B18, who assumed a frequency-independent variance<sup>19</sup>.

<sup>19</sup> The magnitude of this variance in B18 was not important for the main results as they were merely maximum likelihood estimates.

We assume a very spectrally smooth, but otherwise flexible, model for the time-averaged beam-weighted foregrounds, which allows it to absorb potential calibration and beam chromaticity errors so long as they are spectrally distinct from the expected 21 cm signal. The model we employ here is colloquially termed the LINLOG model:

$$\bar{T}_{\text{BWFG}} = \left( \frac{\nu}{\nu_0} \right)^{-2.5} \sum_i^{N_{\text{FG}}} \theta_{\text{FG},i} \ln \left( \frac{\nu}{\nu_0} \right)^i = \Phi \theta_{\text{FG}}, \quad (52)$$

where  $\Phi$  is the  $N_{\nu} \times N_{\text{FG}}$  matrix of LINLOG basis functions. Note that this is equivalent to assuming a LINLOG model of foregrounds of the same order at each time  $t$ , where  $\theta_{\text{FG},i} = \int \theta_{\text{FG},i}(t) dt$ .

With these models, we have a joint calibration and sky model likelihood that is a simple extension of the pure-calibration likelihood (cf. Eq. 33):

$$\mathcal{L}_{\text{joint}}(\theta_{\text{mdl}}, \theta_{21}, \theta_{\text{FG}}) = \mathcal{L}_{\text{cal}}(\theta_{\text{mdl}}) \cdot \mathcal{L}_{\text{ant}}(q_{\text{ant}} | \theta_{\text{mdl}}, \theta_{21}, \theta_{\text{FG}}), \quad (53)$$

where  $\mathcal{L}_{\text{ant}}$  is exactly Eq. 34 with  $\text{src}=\text{ant}$ .

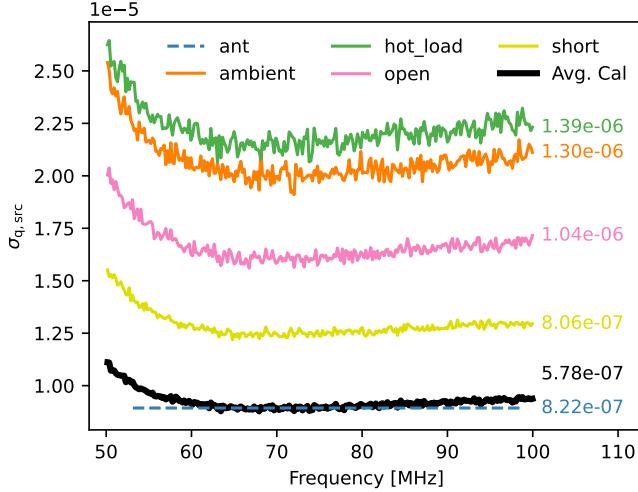
Similar to the pure-calibration likelihood, we can also represent the joint likelihood in a way that highlights the linear parameters, so that we can use the AMLP method described in §3.4. We give this representation in App. D. Briefly, the LINLOG foreground parameters are linear (along with the  $T_{\text{lin}}$ ), while  $\theta_{\text{NS}}$  and  $\theta_{21}$  are non-linear.

## 6 JOINT CALIBRATION AND SKY MODEL RESULTS

We first answer the question of the relative thermal uncertainty between the calibration data and field data. Fig 9 shows the thermal noise,  $\sigma_{q,\text{src}}$ , on each input source as a function of frequency. While each individual calibration source has a higher uncertainty per frequency channel, the combination of higher frequency resolution and multiple calibration sources results in a slightly lower overall calibration uncertainty when averaged over all frequency channels and sources (black curve and number). Thus, we naively expect the calibration data to be slightly more constraining than the field data.

Next, we turn to the posteriors of the joint calibration and sky data likelihood. Fig. 11 shows the resulting Bayesian posteriors when the likelihood defined in Eq. 53 is applied to the full set of available data (i.e. both lab calibration data as discussed in §4.4 and the averaged sky spectrum as given in Eq. 48). In fact, the figure shows these posteriors as the colored regions, while also showing posteriors from an ‘isolated’ sky model fit as grey hashed regions. The ‘isolated’ fit is obtained simply by using the sky data alone, where the data has been pre-calibrated using the *maximum a posteriori* point from our calibration model. Thus, this plot reveals the impact of performing a joint ‘calibration and sky model’ fit on the cosmic inference, as compared to the traditional process of choosing a calibration and then performing the sky model fit in isolation.

We look for two things: bias and posterior spread differences. Note that the regions shown are the 68% and 95% (i.e. 1- and 2- $\sigma$ ) quantiles, while the solid/dashed lines are the median value. Considering biases between the two approaches, we note that for low  $N_{\text{FG}} \leq 6$ , the posterior regions are discrepant to varying degrees: extremely so for  $N_{\text{FG}} = 4$ , with slightly better agreement ( $\sim 1\sigma$ ) for  $N = 5$ , and  $2\sigma$  for  $N_{\text{FG}} = 6$ . Conversely, for  $N_{\text{FG}} > 6$ , we find extremely good agreement between the approaches, with almost perfect overlap of their distributions. The reason for this is likely that the sky-averaged data contains non-cosmic structure that is unable to be fit by a LINLOG model with fewer than 7 terms. For the isolated approach, this simply causes the cosmic inference to be biased, as it is correlated with the

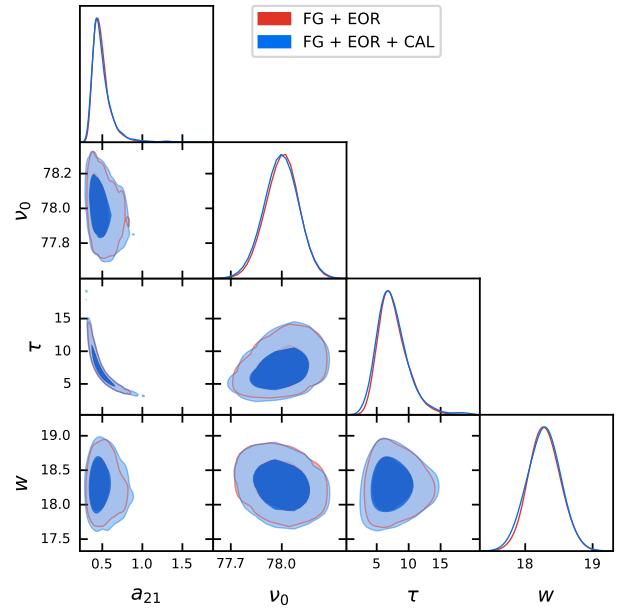


**Figure 9.** Empirical thermal uncertainty of the different input sources,  $\sigma_{q,\text{src}}$ , including the antenna measurements from the field (blue). The thicker black line is the mean uncertainty over all calibration sources, i.e.  $\sqrt{\sum_{\text{src}} \sigma_{q,\text{src}}^2 / N_{\text{src}}}$ . Numbers to the right of each curve are the mean uncertainty over frequency for the given source. While the field antenna measurements individually have the lowest per-channel uncertainty (at the displayed frequency resolution), the combination of all calibration data has slightly lower total mean uncertainty ( $5.78 \times 10^{-7}$  vs.  $8.22 \times 10^{-7}$ ). Note that both calibration and field data have been averaged over a different amount of time and number of frequency channels before being displayed, making the numbers to the right the only ‘comparable’ figure.

foreground model. For the joint model, the effect is partially ameliorated by the calibration itself absorbing some of the extra structure.

This is made clear in Fig. 12, which shows the posteriors on the polynomial coefficients of the calibration scaling parameter,  $T'_{\text{NS}}$ . In that figure, we show only  $N_{\text{FG}} = (4, 6, 8, 10)$  for visual clarity. The grey-dashed cross-hairs show the results of the traditional iterative solution, on which the calibration likelihood is based. As speculated based on Fig. 11, the low- $N_{\text{FG}}$  posteriors are discrepant with the iterative solutions, while the high- $N_{\text{FG}}$  posteriors are consistent with them. On its own, this merely indicates that the calibration solutions are biased when they need to fit out structure in the sky data that the sky model itself is too inflexible to deal with. However, comparing this to Fig. 11 reveals that the *manner* in which the calibration solutions is biased is precisely to absorb the residual smooth FG structure, allowing the cosmic signal to remain at its preferred location consistent with B18. This is further clarified in Fig. 13, which shows the posteriors for  $T_0^{\text{ant}}$  and  $T_1^{\text{ant}}$  as a comparison to the iterative solution, for  $N_{\text{FG}} \in (4, 10)$ . While the additive temperature is almost unaffected by the change in foreground model complexity, the multiplicative temperature is altered in the  $N_{\text{FG}} = 4$  case, with a noticeable dip of magnitude 0.02% around 75 MHz. This lends weight to the proposition that the cosmic feature is truly preferred by the data, as it appears stable with respect to  $N_{\text{FG}}$  and is strong enough to influence the calibration to remain stable.

In terms of the posterior spread, we note the general trend in both approaches for the spread to increase as  $N_{\text{FG}}$  increases. This is to be expected, since the extra flexibility leaves more room for the cosmic signal to vary. This trend is broken only by  $N_{\text{FG}} = 6$ , which is an outlier for both its median estimate and its posterior spread. It is unclear what the precise source of this anomaly is, though it is not of much importance, since  $N_{\text{FG}} = 6$  is not the best model (from the



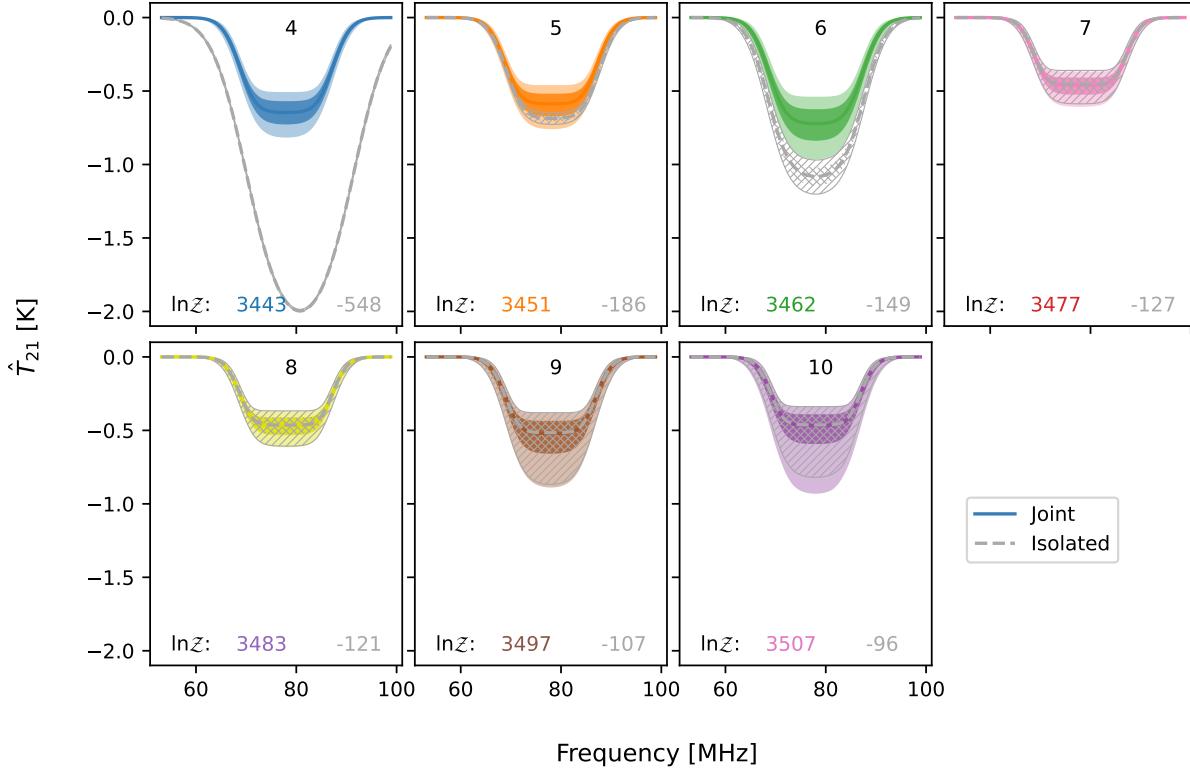
**Figure 10.** Posteriors on the 21 cm flattened-gaussian absorption model, Eq. 40, from both isolated and joint fits for  $N_{\text{FG}} = 10$ . Here, the foreground model is complex enough to describe the non-cosmological components of the data, and the calibration model is not required to account for the beam-weighted foregrounds. This results in very similar predictions for the absorption parameters (cf. Fig. 11). We also see very similar posterior spread in the parameters, due to the fact that the calibration uncertainty is extremely low (cf. Fig. 7).

Bayesian evidence, see below). Finally, we note that for  $N_{\text{FG}} > 6$ , the spread of both approaches is extremely similar. This is to be expected, since Fig. 7 shows such a small (0.01%) spread in the calibrated sky data posterior from the calibration alone. That is, as long as the FG model is sufficiently flexible so that the calibration is not drawn away from the lab data, the uncertainty on the calibration itself does not add significant uncertainty to the cosmic inference. This is reinforced by Fig. 10, which shows the posteriors on  $\theta_{21}$  for both the isolated and joint likelihoods, for  $N_{\text{FG}} = 10$ . The posteriors are almost identical, implying that the uncertainty on the calibration parameters is extremely small.

A final point to note is that Fig. 11 also lists the Bayesian Evidence,  $\ln \mathcal{Z}$ , for each of the models. For both the joint and isolated fits, the evidence increases indefinitely with  $N_{\text{FG}}$ . This strongly indicates that there is structure in the sky data that the flattened-Gaussian signal and LINLOG FG model are insufficient to account for. While the evidence *may* peak for even higher  $N_{\text{FG}}$ , it is unclear that these would be preferred, given our prior of spectrally smooth foregrounds. It is much more likely that the extra residual structure is due to residual beam chromaticity or inaccuracies in the measurement of the antenna’s reflection characteristics. Pursuing the constraint of such systematics is planned for future work.

## 7 CONCLUSIONS

In this paper, we have developed a Bayesian likelihood for the joint estimation of receiver calibration parameters, foregrounds and 21 cm signal for the EDGES global experiment. Our approach is similar to that of (Roque et al. 2020), except that, with an eye towards including



**Figure 11.** The inferred posteriors of the 21 cm signal using different numbers of FG terms (number at top of each panel). Colored regions show 1- and 2- $\sigma$  quantiles from a joint calibration and sky model inference, while the grey hashes show the same for an ‘isolated’ fit of the sky model to pre-calibrated sky data. For  $N_{\text{FG}} \leq 6$  the joint and isolated fits are discrepant, due to the FG model being insufficiently flexible to fit the data. For  $N_{\text{FG}} \geq 7$ , the posteriors are remarkably similar, indicating that the calibration uncertainty is minimal.

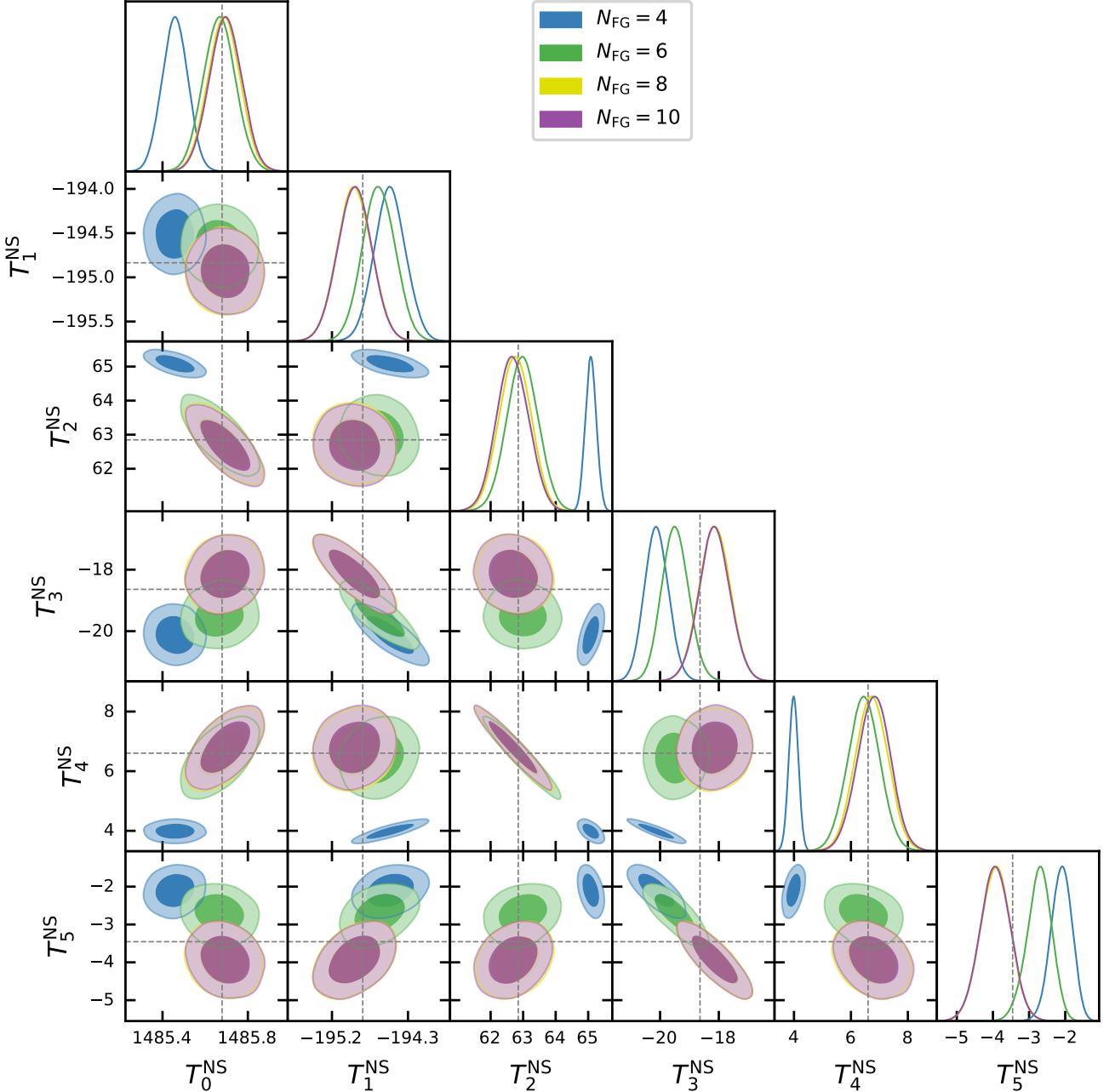
more sophisticated systematics in the future, we do not utilize conjugate priors, but instead improve efficiency by marginalizing over our many linear parameters (Monsalve et al. 2018; Tauscher et al. 2020a). We applied this joint fit to data from the first reported evidence of a detection of the global 21 cm signal in Bowman et al. (2018).

Our first investigation was for data taken purely in the lab, meant to inform the receiver calibration alone. This data consists of spectra and reflection parameters from four sources with known temperature. Two of the sources are designed to have very low reflections, while the other two – the open and shorted cable – have large reflections, designed to probe the noise-wave temperatures. We found that a significant systematic exists in the two cable measurements, resulting in  $\sim 20 - 30\sigma$  residuals in those measurements after calibration. Such systematics were entirely absent from the other two measurements, whose reflections were small (i.e.  $|\Gamma_{\text{src}}| \ll 1$ ). We found that the these systematics significantly bias resulting cosmic signal estimates when calibration is performed using a likelihood that treats all input sources on the same footing. However, the established iterative solution technique (Monsalve et al. 2017a) is able to largely avoid this bias by restricting the influence of the cable measurements. Despite the cable systematics, applying the iterative calibration solutions to an ‘antenna simulator’ designed to mimic the reflection characteristics of the EDGES antenna yields reasonable residuals ( $\sim 2\sigma$ ), and the effect of added calibration flexibility on cosmic inference is minimal. Thus, to avoid this bias in our Bayesian

likelihood, we adopt an approximate method in which we use the iterative solutions to simulate cable measurements without systematic biases, adding Gaussian noise consistent with empirical estimates. Using this method, we find that using 6 terms for the scale and offset temperatures (i.e.  $c_{\text{terms}} = 6$ ) maximizes the Bayesian evidence, irrespective of the number of  $c_{\text{terms}}$  used in the iterative solution. This confirms the choice of B18, where this number of terms was chosen based on residuals of the antenna simulator.

We then performed a joint fit of our calibration model along with a sky model consisting of LINLOG foregrounds and a flattened-Gaussian cosmic signal. We were careful to include all other losses and corrections applied in B18, including beam correction, ground loss and balun loss. We found that our joint model infers a cosmic signal consistent with B18, for  $N_{\text{FG}} = 4 - 10$ . This is in contrast to an ‘isolated’ inference in which the ‘best-fit’ calibration is applied to the sky data and the sky model is fit alone – in this approach only high- $N_{\text{FG}}$  fits are consistent in their predictions. This, along with the rising Bayesian evidence with  $N_{\text{FG}}$ , indicates that there is structure in the sky that requires  $N_{\text{FG}} > 6$  to begin to capture. Nevertheless, the inferred feature is strong enough that in the joint fit, the calibration tends to absorb the extra structure that the foregrounds are unable to fit, keeping the cosmic inference consistent between different numbers of foreground terms.

One question that naturally arises in the context of this work is whether a full joint model is necessary. We have shown that, un-

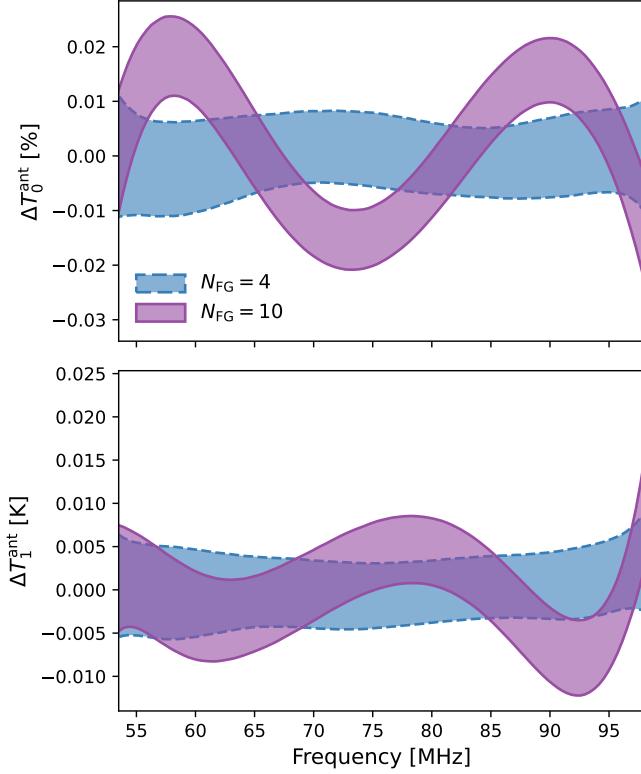


**Figure 12.** Posteriors for the calibration function  $T'_{\text{NS}}$  after running the joint calibration and sky model inference for different numbers of FG terms. Grey dashed cross-hairs are the iterative solutions for the fiducial calibration model. Using a flexible FG model lets the calibration remain close to the best fit from purely lab-calibration data. The calibration absorbs some of the sky data structure when the sky model itself is not flexible enough. See Fig. 13 for a projection of this posterior to frequency-space.

der the calibration assumptions made in this work, a joint model is unnecessary *if the foreground model is sufficiently complex to describe the sky data*. The uncertainty on the calibration parameters is small enough that the extra uncertainty propagated to the cosmic signal parameters is negligible (cf. Fig 9). However, if the foreground model is too inflexible to account for the sky data, a joint model is more robust. Nevertheless, it would seem more appropriate to set the foreground model to be sufficiently flexible, and use an isolated fit,

rather than a joint fit. This may not remain true as further calibration uncertainties are included.

A further question is whether the joint model presented here may lend itself to more bias than an isolated model. Such a conclusion suggests itself as a possibility upon consideration of the pure calibration likelihood developed in this work. In that likelihood, we necessarily combined data from all calibration sources, weighted according to their thermal uncertainty. However, since our model for



**Figure 13.** Posteriors of linear calibration parameters,  $T_0^{\text{ant}}$  and  $T_1^{\text{ant}}$  for the joint calibration and data likelihood. Shown are the posteriors for  $N_{\text{FG}} = 4$  and  $N_{\text{FG}} = 10$ , which are represented by the same colors in Figs. 11 and 12. The curves shown are in comparison to the iterative solution with calibration data only. The multiplicative gain,  $T_0$  is shown as a fractional difference, while the additive temperature,  $T_1$ , is shown as an absolute difference.

some of those sources was incomplete (i.e. the cable measurements), this leaked bias into the models for all sources. In this case, ‘isolating’ the measurements and models reduces the overall bias. This reasoning may also be the case for the sky model and data. By letting the calibration models “see” the sky data in the joint model, they are able to be influenced by it. This is not a problem, and indeed is the correct thing to do, if our sky model is accurate. However, if it is not, the calibration solutions will be pushed away from the solutions they would obtain purely from calibration data. Whether this is really a problem hinges on one’s prior credence on the sky model’s accuracy. If one is very confident in the sky model, then it is perfectly appropriate for the calibration model to be moved away from the lab data by the sky. If not, then it is appropriate to conclude that such movement is systematic bias. In this work we established that while the foreground and calibration models are correlated, the sharp features of our deep flattened-gaussian model are sufficiently uncorrelated with either such that its estimate remains constant against their changing complexity. In principle, model selection will play the crucial role of deciding which model is most appropriate.

We note that while the full joint model presented here is potentially unnecessary, in the sense that the posteriors for the parameters of interest are not significantly affected by including the calibration parameters, the Bayesian formalism for the calibration itself has proven to be highly useful. The bias in the calibration solutions for the cable measurements, noted and discussed in §4.3, is only able to be properly diagnosed under the Bayesian framework presented

here. Comparing Bayesian Evidence between different systematics models should be an effective solution to understanding where the bias comes from – a problem we leave to future work.

## 7.1 Future Work

The results in this paper represent the foundation of a broader program that will be required to verify the results of B18. The foundation is the Bayesian statistical framework, in which various systematic biases and uncertainties can be added and jointly inferred along with the cosmic signal. In this paper, we have merely focused on the ‘easiest’ of these uncertainties – the receiver calibration – as an initial exploration. Five additional instrumental systematics are candidates for future modelling:

- (i) The calibration cable measurements  $\Gamma_{\text{short/open}}$ , which are behind the bias seen in Fig. 4.
- (ii)  $\Gamma_{\text{inst}}$  and  $\Gamma_{\text{ant}}$ , for which we have multiple measurements taken over multiple years (in-situ for  $\Gamma_{\text{ant}}$ ).
- (iii) The beam chromaticity, for which the beam model itself could be better characterized, as well as the formalism in which the correction is applied.
- (iv) The various losses involved: antenna, ground, balun etc.

In particular, the cable measurement bias will be an important systematic to characterize, in order to enable a more self-consistent likelihood.

## ACKNOWLEDGEMENTS

This work was supported by the NSF through research awards for EDGES (AST-1609450, AST-1813850, and AST-1908933). N.M. was supported by the Future Investigators in NASA Earth and Space Science and Technology (FINESST) cooperative agreement 80NSSC19K1413. PHS was supported in part by a McGill Space Institute fellowship and funding from the Canada 150 Research Chairs Program. EDGES is located at the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamatji people as the traditional owners of the Observatory site. We thank CSIRO for providing site infrastructure and support. *Software:* This paper has made use of many excellent software packages, including NUMPY (Harris et al. 2020), SCIPY (Virtanen et al. 2020), POLYCHORD (Handley et al. 2015a,b), ASTROPY (Robitaille et al. 2013; Astropy Collaboration et al. 2018), GETDIST (Lewis 2019), MATPLOTLIB (Hunter 2007), H5PY (Collette 2013) and YABF<sup>20</sup>.

## DATA AVAILABILITY

All publicly-available data used in this work, as well as analysis software in the form of Jupyter notebooks, can be accessed at <https://github.com/edges-collab/bayesian-calibration-paper-code>. Raw calibration data is available on reasonable request via email to the corresponding author. Software used to perform work with this data is publicly available at <https://github.com/edges-collab>. Output data products, such as MCMC chains, resulting from this work are also available upon reasonable request to the corresponding author.

<sup>20</sup> <https://github.com/steven-murray/yabf>

## REFERENCES

- Anstey D., de Lera Acedo E., Handley W., 2020, arXiv e-prints, 2010, arXiv:2010.09644
- Anstey D., Cumner J., de Lera Acedo E., Handley W., 2022, *Monthly Notices of the Royal Astronomical Society*, 509, 4679
- Astropy Collaboration et al., 2018, *The Astronomical Journal*, 156, 123
- Bernardi G., et al., 2016, *Monthly Notices of the Royal Astronomical Society*, 461, 2847
- Bevins H. T. J., Handley W. J., Fialkov A., Acedo E. d. L., Javid K., 2021, arXiv:2104.04336 [astro-ph]
- Bevins H. T. J., de Lera Acedo E., Fialkov A., Handley W. J., Singh S., Subrahmanyam R., Barkana R., 2022, *Monthly Notices of the Royal Astronomical Society*, 513, 4507
- Bowman J. D., Rogers A. E. E., Hewitt J. N., 2008, *The Astrophysical Journal*, 676, 1
- Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018, *Nature*, 555, 67
- Collette A., 2013, Python and HDF5. O'Reilly
- Elsherbeni A. Z., Nayeri P., Reddy C. J., 2014, Antenna Analysis and Design Using FEKO Electromagnetic Simulation Software. IET Digital Library, doi:10.1049/SBEW521E, <https://digital-library.theiet.org/content/books/ew/sbew521e>
- Furlanetto S. R., 2016, arXiv:1511.01131 [astro-ph] 10.1007/978-3-319-21957-8\_9, 423, 247
- Furlanetto S. R., Peng Oh S., Briggs F. H., 2006, *Physics Reports*, 433, 181
- Girish B. S., et al., 2020, *Journal of Astronomical Instrumentation*, 09, 2050006
- Handley W. J., Hobson M. P., Lasenby A. N., 2015a, *Monthly Notices of the Royal Astronomical Society*, 450, L61
- Handley W. J., Hobson M. P., Lasenby A. N., 2015b, *Monthly Notices of the Royal Astronomical Society*, 453, 4384
- Harris C. R., et al., 2020, *Nature*, 585, 357
- Haslam C. G. T., Salter C. J., Stoffel H., Wilson W. E., 1982, *Astronomy and Astrophysics, Suppl. Ser.*, Vol. 47, p. 1-143 (1982), 47, 1
- Hills R., Kulkarni G., Meerburg P. D., Puchwein E., 2018, arXiv:1805.01421 [astro-ph, physics:hep-ph]
- Hunter J. D., 2007, *Computing in Science and Engineering*, 9, 90
- Jaynes E. T., Bretthorst G. L., 2003, Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, doi:10.1017/CBO9780511790423, <https://www.cambridge.org/core/books/probability-theory/9CA08E224FF30123304E6D8935CF1A99>
- Jelić V., Zaroubi S., Labropoulos P., Bernardi G., de Bruyn A. G., Koopmans L. V. E., 2010, *Monthly Notices of the Royal Astronomical Society*, 409, 1647
- Lentati L., Sims P. H., Carilli C., Hobson M. P., Alexander P., Sutter P., 2017, preprint, p. arXiv:1701.03384
- Lewis A., 2019, GetDist: A Python Package for Analysing Monte Carlo Samples (arXiv:1910.13970), doi:10.48550/arXiv.1910.13970, <http://arxiv.org/abs/1910.13970>
- Liu A., Parsons A. R., Trott C. M., 2014, *Physical Review D*, 90, 023019
- Mahesh N., Bowman J. D., Mozdzen T. J., Rogers A. E. E., Monsalve R. A., Murray S. G., Lewis D., 2021, *The Astronomical Journal*, 162, 38
- McKinley B., Trott C. M., Sokolowski M., Wayth R. B., Sutinjo A., Patra N., Nambissan T. J., Ung D. C. X., 2020, *Monthly Notices of the Royal Astronomical Society*, 499, 52
- Meys R., 1978, *IEEE Transactions on Microwave Theory and Techniques*, 26, 34
- Monsalve R. A., Rogers A. E. E., Bowman J. D., Mozdzen T. J., 2017a, *The Astrophysical Journal*, 835, 49
- Monsalve R. A., Rogers A. E. E., Bowman J. D., Mozdzen T. J., 2017b, *The Astrophysical Journal*, 847, 64
- Monsalve R. A., Greig B., Bowman J. D., Mesinger A., Rogers A. E. E., Mozdzen T. J., Kern N. S., Mahesh N., 2018, *The Astrophysical Journal*, 863, 11
- Monsalve R. A., Fialkov A., Bowman J. D., Rogers A. E. E., Mozdzen T. J., Cohen A., Barkana R., Mahesh N., 2019, *The Astrophysical Journal*, 875, 67
- Mozdzen T. J., Mahesh N., Monsalve R. A., Rogers A. E. E., Bowman J. D., 2019, *Monthly Notices of the Royal Astronomical Society*, 483, 4411
- Murray S. G., 2022a, Technical Report 196, Quantifying Correlations Between Frequency Bins, [http://loco.lab.asu.edu/loco-memos/edges\\_reports/EDGES\\_Memo\\_196\\_v0.pdf](http://loco.lab.asu.edu/loco-memos/edges_reports/EDGES_Memo_196_v0.pdf). Arizona State University
- Murray S. G., 2022b, Technical Report 199, Correspondence of Edges-Cal to Alans-Pipeline, [http://loco.lab.asu.edu/loco-memos/edges\\_reports/EDGES\\_Memo\\_199\\_v0.pdf](http://loco.lab.asu.edu/loco-memos/edges_reports/EDGES_Memo_199_v0.pdf). Arizona State University
- Nambissan T. J., et al., 2021, SARAS 3 CD/EoR Radiometer: Design and Performance of the Receiver, <https://ui.adsabs.harvard.edu/abs/2021arXiv210401756N>
- Pritchard J. R., Loeb A., 2012, *Reports on Progress in Physics*, 75, 086901
- Robitaille T. P., et al., 2013, *Astronomy & Astrophysics*, 558, A33
- Rogers A. E. E., Bowman J. D., 2012, *Radio Science*, 47
- Roque I. L. V., Handley W. J., Razavi-Ghods N., 2020, arXiv e-prints, 2011, arXiv:2011.14052
- Scheutwinkel K. H., de Lera Acedo E., Handley W., 2022b, Bayesian Evidence-Driven Diagnosis of Instrumental Systematics for Sky-Averaged 21-Cm Cosmology Experiments, <https://ui.adsabs.harvard.edu/abs/2022arXiv220404445S>
- Scheutwinkel K. H., Handley W., de Lera Acedo E., 2022a, Bayesian Evidence-Driven Likelihood Selection for Sky-Averaged 21-Cm Signal Extraction, <https://ui.adsabs.harvard.edu/abs/2022arXiv220404491S>
- Shaver P. A., Windhorst R. A., Madau P., de Bruyn A. G., 1999, *Astronomy & Astrophysics*, 345, 380
- Sims P. H., Poher J. C., 2020, *Monthly Notices of the Royal Astronomical Society*, 492, 22
- Singh S., Subrahmanyam R., 2019, *The Astrophysical Journal*, 880, 26
- Singh S., et al., 2022, *Nature Astronomy*, 6, 607
- Sokolowski M., et al., 2015, *Publications of the Astronomical Society of Australia*, 32, e004
- Tauscher K., Rapetti D., Burns J. O., 2020a, *The Astrophysical Journal*, 897, 132
- Tauscher K., Rapetti D., Burns J. O., 2020b, *The Astrophysical Journal*, 897, 175
- Tauscher K., et al., 2021, *The Astrophysical Journal*, 915, 66
- Virtanen P., et al., 2020, *Nature Methods*, 17, 261

## APPENDIX A: EXPECTATION OF A RATIO OF GAUSSIAN VARIABLES

For random variables  $X, Y$ , where the density of  $Y$  at zero is negligible, the expectation of the ratio can be approximated via Taylor series to second-order as

$$E[X/Y] \approx \frac{E[X]}{E[Y]} \left( 1 - \frac{\text{Cov}(X, Y)}{E[X]E[Y]} + \frac{\text{Var}(Y)}{E[Y]^2} \right). \quad (\text{A1})$$

## APPENDIX B: ANALYTIC MARGINALIZATION OF LINEAR PARAMETERS

MCMC techniques are typically inefficient when dealing with large numbers of parameters, due to the curse of dimensionality. In this appendix, we review the derivation of a technique used to reduce the effective dimensionality of the model to be explored, via analytical marginalization over some of the parameters. We note that this technique is not new, even in the context of global experiments (eg. Lentati et al. 2017; Monsalve et al. 2018; Tauscher et al. 2021). However, at first glance, these papers suggest differing results (i.e. they conclude with different formulae that are not obviously identical). In this appendix, we derive the same results, and show that they are not in disagreement.

The technique is applicable when the following conditions hold (cf. Tauscher et al. (2021)):

(i) The likelihood of the data,  $\mathbf{d}$ , is Gaussian, i.e.

$$\mathcal{L}(\mathbf{d}|\boldsymbol{\theta}) \propto |\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})| \exp \left\{ -\frac{1}{2} \mathbf{r}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \mathbf{r} \right\}, \quad (\text{B1})$$

where the  $\mathbf{r} = (\mathbf{d} - \mathbf{m}(\boldsymbol{\theta}))$  is the residual of the data to a model,  $\mathbf{m}$ , dependent on the parameters  $\boldsymbol{\theta}$  and evaluated at the same coordinates as the data, and  $\boldsymbol{\Sigma}$  is a model for the covariance of the data, potentially dependent on the parameters as well.

(ii) After specification of the values of a *subset* of the parameters, to be called the non-linear parameters  $\boldsymbol{\theta}_{\text{NL}}$ , the model is linear in the remaining parameters,  $\boldsymbol{\theta}_{\text{L}}$ . That is, the parameters can be split into two groups,  $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\text{L}}, \boldsymbol{\theta}_{\text{NL}}\}$  such that when the model is conditioned on  $\boldsymbol{\theta}_{\text{NL}}$ , the gradient of the model with  $\boldsymbol{\theta}_{\text{L}}$  is independent of  $\boldsymbol{\theta}_{\text{NL}}$ .

(iii) The covariance depends only on the non-linear parameters.

(iv) The priors on the two sets of parameters are independent, i.e.  $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}_{\text{L}})\pi(\boldsymbol{\theta}_{\text{NL}})$

(v) The linear prior is either Gaussian or improper uniform.

We note that in this work we consider both the linear and non-linear priors to be improper uniform for simplicity, and thus they drop out of our derivations. We furthermore note that our formulation in which  $\mathbf{r}$  is the residual of raw data to a model, while being sufficiently general, is not the only way – nor always the most practical – to formulate the residuals. Indeed, the ‘data’  $\mathbf{d}$  may be taken to be some function of the raw data,  $\mathbf{d} = f(\mathbf{d}_{\text{raw}}, \boldsymbol{\theta}_{\text{NL}})$ , so long as the resulting data has a Gaussian distribution. In practice, the only realistic non-trivial function that preserves Gaussianity is a scaling, i.e.  $\mathbf{d} = f(\boldsymbol{\theta}_{\text{NL}})\mathbf{d}_{\text{raw}}$ . In this case, the parameters used in  $f$  must be considered ‘non-linear’ as they affect the covariance of  $\mathbf{d}$ . Given the constraints, this is mathematically equivalent to keeping  $\mathbf{d} = \mathbf{d}_{\text{raw}}$  and dividing the model  $\mathbf{m}$  by  $f(\boldsymbol{\theta}_{\text{NL}})$ , in which case the covariance is constant (but the parameters must still be non-linear as they are divisors in the model). While mathematically equivalent, the two are not algorithmically equivalent, and which is computed more efficiently depends on the way a particular code is written.

We note that these conditions hold for our likelihood, Eq. 53 (cf. Eqs. C6 and D7).

We now integrate the posterior over the linear parameters:

$$p_{\text{NL}}(\boldsymbol{\theta}_{\text{NL}}|\mathbf{d}) = \int p(\boldsymbol{\theta}_{\text{NL}}, \boldsymbol{\theta}_{\text{L}}) d\boldsymbol{\theta}_{\text{L}} \quad (\text{B2})$$

$$\propto \pi_{\text{NL}}(\boldsymbol{\theta}_{\text{NL}}) \int \pi_{\text{L}}(\boldsymbol{\theta}_{\text{L}}) \mathcal{L}(\mathbf{d}|\boldsymbol{\theta}_{\text{NL}}, \boldsymbol{\theta}_{\text{L}}) d\boldsymbol{\theta}_{\text{L}} \quad (\text{B3})$$

$$= \int \mathcal{L}(\mathbf{d}|\boldsymbol{\theta}_{\text{NL}}, \boldsymbol{\theta}_{\text{L}}) d\boldsymbol{\theta}_{\text{L}} \quad (\text{B4})$$

$$\equiv \mathcal{L}_{\text{eff}}(\mathbf{d}|\boldsymbol{\theta}_{\text{NL}}). \quad (\text{B5})$$

Here the second last equality makes the assumption that the priors on both linear and non-linear parameters are improper uniform distributions and the last equality defines an “effective” likelihood.

Let  $\mathbf{m}'$  be the model conditioned on the non-linear parameters. We can then write  $\mathbf{m}' = \mathbf{A}\boldsymbol{\theta}_{\text{L}}$ , i.e. the remaining model is a linear model with design matrix  $\mathbf{A}$  (note that  $\mathbf{A}$  may be dependent on the non-linear parameters). Further, let the data covariance (after any transformation by non-linear parameters) be  $\boldsymbol{\Sigma}$  and the maximum-likelihood estimate of the linear parameters (conditional on the chosen non-linear parameters) be  $\hat{\boldsymbol{\theta}}_{\text{L}}$ , that is,

$$\hat{\boldsymbol{\theta}}_{\text{L}} = (\mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{d} = \boldsymbol{\Sigma}_{\text{L}} \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{d}, \quad (\text{B6})$$

with  $\boldsymbol{\Sigma}_{\text{L}}$  the covariance matrix of the linear parameters. Now, express

the residuals as  $\mathbf{r} = \mathbf{d} - \mathbf{A}(\hat{\boldsymbol{\theta}}_{\text{L}} + \boldsymbol{\delta}_{\text{L}}) \equiv \hat{\mathbf{r}} - \mathbf{A}\boldsymbol{\delta}_{\text{L}}$ , i.e. the sum of the maximum-likelihood residuals and a small model component.

Following the derivation in Monsalve et al. (2018), we express the exponent of the conditional likelihood (not the effective likelihood) as

$$-\frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{r}} - \frac{1}{2} \boldsymbol{\delta}_{\text{L}}^T \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} \boldsymbol{\delta}_{\text{L}} + \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} \boldsymbol{\delta}_{\text{L}}. \quad (\text{B7})$$

We may then integrate over the linear parameters:

$$\mathcal{L}_{\text{eff}}(\mathbf{d}|\boldsymbol{\theta}_{\text{NL}}) \propto \frac{\exp \left\{ -\frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{r}} \right\}}{\sqrt{|\boldsymbol{\Sigma}|}} \int \exp \left\{ -\frac{1}{2} \chi_{\text{L}}^2 \right\} d\boldsymbol{\delta}_{\text{L}}, \quad (\text{B8})$$

where

$$-\frac{1}{2} \chi_{\text{L}}^2 = -\frac{1}{2} \boldsymbol{\delta}_{\text{L}}^T \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} \boldsymbol{\delta}_{\text{L}} + \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} \boldsymbol{\delta}_{\text{L}}. \quad (\text{B9})$$

This may be solved using the identity Eq. 13 in Monsalve et al. (2018) to give

$$\int \exp \left\{ -\frac{1}{2} \chi_{\text{L}}^2 \right\} d\boldsymbol{\delta}_{\text{L}} = \sqrt{(2\pi)^{N_{\text{L}}} |\boldsymbol{\Sigma}_{\text{L}}|} \exp \left\{ \frac{1}{2} \mathbf{b}^T \boldsymbol{\Sigma}_{\text{L}} \mathbf{b} \right\}, \quad (\text{B10})$$

with  $\mathbf{b} = \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{r}}$ . Substituting this result back into the effective likelihood we have

$$\mathcal{L}_{\text{eff}}(\mathbf{d}|\boldsymbol{\theta}_{\text{NL}}) \propto \frac{\exp \left\{ -\frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{r}} + \frac{1}{2} \mathbf{b}^T \boldsymbol{\Sigma}_{\text{L}} \mathbf{b} \right\}}{\sqrt{|\boldsymbol{\Sigma}| |\boldsymbol{\Sigma}_{\text{L}}^{-1}|}} \quad (\text{B11})$$

$$= \frac{\exp \left\{ -\frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{r}} + \frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} \mathbf{A} \boldsymbol{\Sigma}_{\text{L}} \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{r}} \right\}}{\sqrt{|\boldsymbol{\Sigma}| |\boldsymbol{\Sigma}_{\text{L}}^{-1}|}} \quad (\text{B12})$$

$$= \frac{\exp \left\{ -\frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} [\hat{\mathbf{r}} - \mathbf{A} \boldsymbol{\Sigma}_{\text{L}} \mathbf{A}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{r}}] \right\}}{\sqrt{|\boldsymbol{\Sigma}| |\boldsymbol{\Sigma}_{\text{L}}^{-1}|}}. \quad (\text{B13})$$

Now, departing from the derivation of Monsalve et al. (2018), we note that the last term in the exponential contains the standard “hat matrix”,

$$\mathbf{H} = \mathbf{A} \boldsymbol{\Sigma}_{\text{L}} \mathbf{A}^T \boldsymbol{\Sigma}^{-1}, \quad (\text{B14})$$

which “puts a hat” on the data model, i.e.  $\mathbf{H}\mathbf{d} = \mathbf{A}\hat{\boldsymbol{\theta}}_{\text{L}}$  and is idempotent. We thus have

$$\mathcal{L}_{\text{eff}}(\mathbf{d}|\boldsymbol{\theta}_{\text{NL}}) \propto \frac{\exp \left\{ -\frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} [\hat{\mathbf{r}} - \mathbf{H}(\mathbf{d} - \mathbf{A}\hat{\boldsymbol{\theta}}_{\text{L}})] \right\}}{\sqrt{|\boldsymbol{\Sigma}| |\boldsymbol{\Sigma}_{\text{L}}^{-1}|}} \quad (\text{B15})$$

$$= \frac{\exp \left\{ -\frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} [\hat{\mathbf{r}} - (\mathbf{H}\mathbf{d} - \mathbf{H}\mathbf{H}\mathbf{d})] \right\}}{\sqrt{|\boldsymbol{\Sigma}| |\boldsymbol{\Sigma}_{\text{L}}^{-1}|}} \quad (\text{B16})$$

$$= \frac{\exp \left\{ -\frac{1}{2} \hat{\mathbf{r}}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{r}} \right\}}{\sqrt{|\boldsymbol{\Sigma}| |\boldsymbol{\Sigma}_{\text{L}}^{-1}|}}, \quad (\text{B17})$$

where the last equality follows due to idempotency of  $\mathbf{H}$ .

This final equation, Eq. (B17), is the same result as given in Tauscher et al. (2021).

## APPENDIX C: LINEAR REPRESENTATION OF CALIBRATION LIKELIHOOD

In §4.2 we presented the likelihood of the calibration data in a clear conceptually-oriented notation. In practice, to use the analytical

marginalization over the linear parameters, as outlined in §3.4 and App. B, it is useful to represent the data model as a combination of linear and non-linear parameters, in which the the linear parameters enter exclusively through a single term,  $\mathbf{A}\theta_L$ . Here, we derive this representation.

Let the linear temperature terms,  $T_{\mathcal{T}_{\text{lin}}}$ , be formed into a vector

$$\boldsymbol{\theta}_{\text{NW+L}} = [\boldsymbol{\theta}_{\text{unc}}^T, \boldsymbol{\theta}_{\text{cos}}^T, \boldsymbol{\theta}_{\text{sin}}^T, \boldsymbol{\theta}_L^T]^T. \quad (\text{C1})$$

Also, for a particular input source,  $\text{src} \in \mathcal{S}_{\text{cal}}$ , and term,  $p \in \mathcal{T}_{\text{lin}}$ , define an  $N_V \times N_{\text{terms}}^p$  sub-design matrix

$$\mathbf{V}_{ij}^{\text{src},p} = \boldsymbol{\kappa}_p^{\text{src}}(\nu_i) \boldsymbol{\Psi}_{ij}^p. \quad (\text{C2})$$

Then define a design matrix:

$$\mathbf{K}_{\text{cal}} = \begin{pmatrix} \mathbf{V}_{\text{amb,unc}} & \mathbf{V}_{\text{amb,cos}} & \mathbf{V}_{\text{amb,sin}} & \mathbf{V}_{\text{amb,L}} \\ \mathbf{V}_{\text{hot,unc}} & \mathbf{V}_{\text{hot,cos}} & \mathbf{V}_{\text{hot,sin}} & \mathbf{V}_{\text{hot,L}} \\ \mathbf{V}_{\text{short,unc}} & \mathbf{V}_{\text{short,cos}} & \mathbf{V}_{\text{short,sin}} & \mathbf{V}_{\text{short,L}} \\ \mathbf{V}_{\text{open,unc}} & \mathbf{V}_{\text{open,cos}} & \mathbf{V}_{\text{open,sin}} & \mathbf{V}_{\text{open,L}} \end{pmatrix} \quad (\text{C3})$$

Furthermore, let the LHS of Eq. 22 for a particular input source be written as a  $N_V$ -vector  $\boldsymbol{d}_{\text{src}}$ <sup>21</sup>:

$$\boldsymbol{d}_{\text{src}} = \boldsymbol{q}_{\text{src}} \circ \boldsymbol{\Psi} \boldsymbol{\theta}_{\text{NS}} - \boldsymbol{\rho}_{\text{src}} \circ \boldsymbol{T}_{\text{src}}, \quad (\text{C4})$$

and form the  $N_V |\mathcal{T}_{\text{cal}}|$  vector concatenated over sources:

$$\boldsymbol{d}_{\text{cal}} = [\boldsymbol{d}_{\text{amb}}^T, \boldsymbol{d}_{\text{hot}}^T, \boldsymbol{d}_{\text{open}}^T, \boldsymbol{d}_{\text{short}}^T]^T. \quad (\text{C5})$$

Under the approximations of Gaussianity and independence of frequency channels, we can write the model residual vector  $\boldsymbol{r}_{\text{cal}}$  in the following form (equivalent to the concatenation of  $\boldsymbol{r}_{\text{src}}$  vectors, cf. Eq. 31):

$$\boldsymbol{r}_{\text{cal}} = (\boldsymbol{d}_{\text{cal}} - \mathbf{K}_{\text{cal}} \boldsymbol{\theta}_{\text{NW+L}}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\text{cal}}), \quad (\text{C6})$$

where  $\boldsymbol{\Sigma}_{\text{cal}}$  is the diagonal scaled covariance matrix given by

$$\boldsymbol{\Sigma}_{\text{cal}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\text{amb}} & 0 & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{\text{hot}} & 0 & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_{\text{short}} & 0 \\ 0 & 0 & 0 & \boldsymbol{\Sigma}_{\text{open}} \end{pmatrix}. \quad (\text{C7})$$

Eq. C6 represents a Gaussian likelihood for the (transformed) data  $\boldsymbol{d}_{\text{cal}}$  given a purely linear model  $\mathbf{K}_{\text{cal}} \boldsymbol{\theta}_{\text{NW+L}}$  in which  $\mathbf{K}_{\text{cal}}$  in general depends on non-linear parameters (but we do not consider any such parameters in this paper). This is precisely the form we need to apply the effective likelihood, Eq. B17, with  $\mathbf{A} = \mathbf{K}_{\text{cal}}$ . Note that we have adopted the scaled Gaussian likelihood (Eq. 4) in which both  $\boldsymbol{d}_{\text{cal}}$  and  $\boldsymbol{\Sigma}_{\text{cal}}$  are scaled by the non-linear parameters  $\boldsymbol{\theta}_{\text{NS}}$ .

## APPENDIX D: LINEAR REPRESENTATION OF JOINT LIKELIHOOD

Here, we apply the same process to the joint likelihood, Eq. 53, as App. C applied to the calibration likelihood, Eq. 33. Our aim is to represent the likelihood in the same form as C6.

We treat the antenna simply as another (i.e. fifth) source. Thus we have

$$\boldsymbol{d}_{\text{ant}} = \bar{\boldsymbol{q}}_{\text{ant}} \circ \boldsymbol{\Psi} \boldsymbol{\theta}_{\text{NS}} - \boldsymbol{\rho}_{\text{ant}} \circ \boldsymbol{T}_{21, \text{meas}}, \quad (\text{D1})$$

<sup>21</sup> We choose the notation  $\boldsymbol{d}$  here as this quantity represents our ‘data’, though in truth it is a linear transformation of the data, in which the transformation itself is being modelled.

where

$$\boldsymbol{T}_{21, \text{meas}} = \mathbf{L} \circ \bar{\boldsymbol{b}}_{\text{corr}} \circ \boldsymbol{T}_{21}(\boldsymbol{\theta}_{21}) + \bar{\boldsymbol{T}}_{\text{loss}} \quad (\text{D2})$$

and

$$\boldsymbol{d}_{\text{sml}} = [\boldsymbol{d}_{\text{amb}}^T, \boldsymbol{d}_{\text{hot}}^T, \boldsymbol{d}_{\text{open}}^T, \boldsymbol{d}_{\text{short}}^T, \boldsymbol{d}_{\text{ant}}^T]^T. \quad (\text{D3})$$

Note that this has removed the contribution of the foregrounds to the expected antenna temperature, as these are linear and will be added to the linear term, rather than  $\boldsymbol{d}$ . To do this, we modify the  $\mathbf{K}$  matrix to be

$$\mathbf{K}_{\text{sml}} = \begin{pmatrix} \mathbf{V}_{\text{amb,unc}} & \mathbf{V}_{\text{amb,cos}} & \mathbf{V}_{\text{amb,sin}} & \mathbf{V}_{\text{amb,L}} & 0 \\ \mathbf{V}_{\text{hot,unc}} & \mathbf{V}_{\text{hot,cos}} & \mathbf{V}_{\text{hot,sin}} & \mathbf{V}_{\text{hot,L}} & 0 \\ \mathbf{V}_{\text{short,unc}} & \mathbf{V}_{\text{short,cos}} & \mathbf{V}_{\text{short,sin}} & \mathbf{V}_{\text{short,L}} & 0 \\ \mathbf{V}_{\text{open,unc}} & \mathbf{V}_{\text{open,cos}} & \mathbf{V}_{\text{open,sin}} & \mathbf{V}_{\text{open,L}} & 0 \\ \mathbf{V}_{\text{ant,unc}} & \mathbf{V}_{\text{ant,cos}} & \mathbf{V}_{\text{ant,sin}} & \mathbf{V}_{\text{ant,L}} & \mathbf{V}_{\text{ant,FG}} \end{pmatrix}, \quad (\text{D4})$$

where

$$\mathbf{V}_{ij}^{\text{ant,FG}} = \rho_{i,\text{ant}} \mathbf{L}_i \bar{\boldsymbol{b}}_{\text{corr}} \boldsymbol{\Phi}_{ij}, \quad (\text{D5})$$

and write  $\boldsymbol{\theta}_{\text{lin}}$  as

$$\boldsymbol{\theta}_{\text{lin}} = [\boldsymbol{\theta}_{\text{unc}}^T, \boldsymbol{\theta}_{\text{cos}}^T, \boldsymbol{\theta}_{\text{sin}}^T, \boldsymbol{\theta}_L^T, \boldsymbol{\theta}_{\text{FG}}^T]^T. \quad (\text{D6})$$

Under the same assumptions and arguments employed in App. C, this yields

$$\boldsymbol{r}_{\text{sml}} = \boldsymbol{d}_{\text{full}}(\boldsymbol{\theta}_{21}, \boldsymbol{\theta}_{\text{NS}}) - \mathbf{K}_{\text{sml}} \boldsymbol{\theta}_{\text{lin}} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\text{sml}}), \quad (\text{D7})$$

with  $\boldsymbol{\Sigma}_{\text{sml}}$  is simply

$$\boldsymbol{\Sigma}_{\text{cal}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\text{amb}} & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{\text{hot}} & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_{\text{short}} & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\Sigma}_{\text{open}} & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{\Sigma}_{\text{ant}} \end{pmatrix}. \quad (\text{D8})$$

## APPENDIX E: OBTAINING ‘RECALIBRATED’ SKY TEMPERATURE

Given calibrated sky temperature data,  $\widehat{\boldsymbol{T}}_{\text{sky,bc}}$  that was calibrated with the multiplicative and additive temperatures  $\widehat{T}_0^{\text{ant}}$  and  $\widehat{T}_1^{\text{ant}}$  respectively, the data may be re-calibrated using new estimates of the calibration temperatures,  $\widehat{T}_0^{\text{ant'}}$  and  $\widehat{T}_1^{\text{ant'}}$  as follows:

$$\begin{aligned} \widehat{\boldsymbol{T}}'_{\text{sky,bc}} &= \frac{1}{L \bar{\boldsymbol{b}}_{\text{corr}}} \left[ x_0 \left( L \bar{\boldsymbol{b}}_{\text{corr}} \widehat{\boldsymbol{T}}_{\text{sky,bc}} + \bar{\boldsymbol{T}}_{\text{loss}} - \widehat{T}_1^{\text{ant}} \right) + \widehat{T}_1^{\text{ant'}} - \bar{\boldsymbol{T}}_{\text{loss}} \right] \\ &= x_0 \widehat{\boldsymbol{T}}_{\text{sky,bc}} + \frac{1}{L \bar{\boldsymbol{b}}_{\text{corr}}} \left[ \bar{\boldsymbol{T}}_{\text{loss}} (x_0 - 1) - x_0 \widehat{T}_1^{\text{ant}} + \widehat{T}_1^{\text{ant'}} \right] \end{aligned} \quad (\text{E1})$$

where  $x_0 = \widehat{T}_0^{\text{ant'}} / \widehat{T}_0^{\text{ant}}$  is the ratio of the new to old scaling temperatures. Here,  $L$ ,  $\bar{\boldsymbol{b}}_{\text{corr}}$  and  $\bar{\boldsymbol{T}}_{\text{loss}}$  are defined in §5. This re-calibrated temperature is used in Fig. 3.

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