A Bayesian approach to modelling spectrometer data chromaticity corrected using beam factors - II. Model priors and posterior odds

Peter H. Sims,[1,2,3]* Judd D. Bowman,[1] Steven G. Murray,[1,4] John P. Barrett,[5] Rigel C. Cappallo,[5] Colin J. Lonsdale,[5] Nivedita Mahesh,[6] Raul A. Monsalve,[1,7,8] Alan E. E. Rogers,[5] Titu Samson,[1] and Akshatha K. Vydula[1]

[1] *School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85287, USA*
[2] *Astrophysics Group, Cavendish Laboratory, J. J. Thomson Avenue, Cambridge CB3 0HE, UK*
[3] *Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, UK*
[4] *Scuola Normale Superiore (SNS), Piazza dei Cavalieri 7, I-56125 Pisa, PI, Italy*
[5] *MIT Haystack Observatory, Westford, MA 01886-1299, USA*
[6] *Cahill Center for Astronomy and Astrophysics, California Institute of Technology, Pasadena CA 91125, USA*
[7] *Space Sciences Laboratory, University of California Berkeley, Berkeley, CA 94720, USA*
[8] *Facultad de Ingeniería, Universidad Católica de la Santísima Concepción, Alonso de Ribera 2850, Concepción, Chile*

ABSTRACT

The reliable detection of the global 21-cm signal, a key tracer of Cosmic Dawn and the Epoch of Reionization, requires meticulous data modelling and robust statistical frameworks for model validation and comparison. In Paper I of this series, we presented the Beam-Factor-based Chromaticity Correction (BFCC) model for spectrometer data processed using BFCC to suppress instrumentally induced spectral structure. We demonstrated that the BFCC model, with complexity calibrated by Bayes factor-based model comparison (BFBMC), enables unbiased recovery of a 21-cm signal consistent with the one reported by EDGES from simulated data. Here, we extend the evaluation of the BFCC model to lower amplitude 21-cm signal scenarios where deriving reliable conclusions about a model's capacity to recover unbiased 21-cm signal estimates using BFBMC is more challenging. Using realistic simulations of chromaticity-corrected EDGES-low spectrometer data, we evaluate three signal amplitude regimes – null, moderate, and high. We then conduct a Bayesian comparison between the BFCC model and three alternative models previously applied to 21-cm signal estimation from EDGES data. To mitigate biases introduced by systematics in the 21-cm signal model fit, we incorporate the Bayesian Null-Test-Evidence-Ratio (BaNTER) validation framework and implement a Bayesian inference workflow based on posterior odds of the validated models. We demonstrate that, unlike BFBMC alone, this approach consistently recovers 21-cm signal estimates that align with the true signal across all amplitude regimes, advancing robust global 21-cm signal detection methodologies.

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

1 INTRODUCTION

Global 21-cm experiments operating in the frequency range $10 \lesssim \nu \lesssim 230$ MHz, corresponding to a redshift range[1] $150 \lesssim z \lesssim 5$, aim to provide conclusive and unbiased measurements of the sky-averaged redshifted 21-cm hyperfine line radiation emitted by neutral hydrogen in the high-redshift Universe. Observations of the 21-cm signal during the Universe's Dark Ages have the potential to provide precision cosmological constraints (e.g. Mondal & Barkana 2023; Mondal et al. 2024; Gessey-Jones et al. 2024; Naik et al. 2025), to

probe directly the initial stages of structure formation, and to characterise the properties of the first stars, proto-galaxies, and accreting black holes during Cosmic Dawn (CD) and the Epoch of Reionisation (e.g. Bevins et al. 2024; Pochinda et al. 2024; Cang et al. 2024; Gessey-Jones et al. 2025). However, to achieve this, the cosmological 21-cm signal must be extracted from spectrometer data containing astrophysical foreground emission, dominated by synchrotron radiation, that, depending on the frequency range and field observed, is 3–6 orders of magnitude brighter.

The foreground emission is intrinsically spectrally smooth and thus separable from the more spectrally structured global 21-cm signal. However, even low-level spectral structure in the instrumental transfer function mixes the foreground emission into the narrower spectral scales relevant for 21-cm signal detection complicating this separation in the measured data. For such structure to be treated as negligible requires that the instrument transfer function is smooth

---

* E-mail: psims3@asu.edu

[1] The 21-cm hyperfine line of neutral hydrogen has a rest-frame frequency of $\nu_{21} \approx 1420.4$ MHz. Due to the expansion of the Universe, the wavelength of radiation is stretched, which reduces its frequency and establishes a one-to-one mapping between the observation frequency, ν_{obs}, and the redshift, z, at which the 21-cm line is emitted: $\nu_{obs} = \nu_{21}/(1+z)$.

at the approximately[2] 10^5 : 1 foreground-to-noise ratio in the data, which is challenging to achieve in practice. Consequently, accurately accounting for spectral structure induced by instrumental chromaticity on scales relevant for 21-cm signal detection, remains one of the most significant challenges to achieving robust signal estimates. A primary contributor to this effect is the frequency-dependent weighting of the sky by the instrument beam. The Experiment to Detect the Global Epoch of Reionisation Signature (EDGES; Bowman et al. 2018a, hereafter B18) has pioneered a beam-factor-based Chromaticity Correction (BFCC) approach to mitigate this effect by dividing the calibrated spectrometer data by an estimated sky-weighted beam response.

If BFCC perfectly removed instrumental chromaticity, unbiased recovery of the global 21-cm signal would be possible by directly fitting an intrinsic astrophysical model to the corrected data. However, in Paper I of this series (Sims et al. 2023), it was demonstrated that while BFCC reduces the impact of instrumental chromaticity under realistic assumptions about the spectral structure of foregrounds, the correction is only partial, leaving residual spectral systematics that must be carefully modelled to avoid biases in signal estimation. To account for residual instrumental structure in BFCC data, we derived a flexible closed-form model for beam-factor chromaticity-corrected spectrometer data (hereafter the BFCC model) and demonstrated how to optimise the model's complexity for a given dataset using Bayes factor-based model comparison (BFBMC). Using realistic simulations of time-averaged EDGES data, we showed that when embedding a simulated 21-cm signal consistent with the deep ($A = 500^{+500}_{-200}$ mK) best fitting absorption trough reported in B18, fitting the data with an intrinsic sky model – one that describes the spectrum in the absence of chromatic effects (hereafter, the Intrinsic model) – yields biased estimates of the underlying 21-cm signal. In contrast, the BFCC model enables unbiased recovery of the simulated global 21-cm signal.

While the results of Paper I are encouraging, the 21-cm signal reported in B18 is deeper than expected at redshift 17 under standard cosmological assumptions (e.g. B18; Barkana 2018). Additionally, reanalyses of the data in B18, using alternate data models, have since been carried out that suggest either a lower amplitude 21-cm signal than reported in B18 or the absence of a detectable 21-cm signal altogether (e.g., Hills et al. 2018; Bradley et al. 2019; Singh & Subrahmanyan 2019; Sims & Pober 2020; Bevins et al. 2021; Cang et al. 2024). Accounting for this possibility, in this work, we extend our evaluation of the BFCC model to data sets with lower amplitude 21-cm signal and additional comparison models for which it is more challenging to draw reliable conclusions about the model's ability to recover unbiased estimates of the 21-cm signal using BFBMC alone. Specifically, we consider two additional scenarios for the amplitude of the 21-cm signal in the data:

(i) a null signal, which we use to validate the models within the Bayesian Null-Test-Evidence-Ratio (BaNTER) validation framework (Sims et al. 2025a; hereafter, S25),

(ii) a signal with a lower amplitude compared to that reported in B18, consistent with expectations for the 21-cm absorption trough associated with CD at redshift 17 under the standard cosmological assumption that the background brightness temperature at CD is dominated

by the Cosmic Microwave Background (CMB) and the minimum hydrogen gas temperature is determined by adiabatic cooling after decoupling from the radiation temperature.

Furthermore, we analyse the data using two additional comparison models for the non-21-cm component of the publicly available EDGES data. This nuisance component may, itself, be composed of multiple subcomponents describing, for example, astrophysical foregrounds, ionospheric effects and instrumental systematics. However, for brevity, here we refer to the non-21-cm component as the 'foreground' model with the understanding that they are intended to describe the full range of non-21-cm structure in the data. These models are the linearised physical model (hereafter, 'LinPhys' model) and the more general polynomial foreground model (hereafter, 'MultLin' model) used in the analysis carried out in B18 (see Section 4 for details).

When analysing simulated data, preferred models can be directly evaluated by comparing the input and inferred signal of interest. However, when analysing instrumental data where the detailed structure of the signal is a priori unknown, robust inference requires accounting for both the uncertainty in parameter estimates within models and the uncertainty in the model itself. Ignoring model uncertainty can lead to underestimated uncertainties in inferences and biased conclusions due to improperly weighted model-averaged parameter posteriors. Bayesian model comparison offers a unified and statistically consistent framework to address both sources of uncertainty (e.g., Jeffreys 1935, 1939 and Kass & Raftery 1995; hereafter, KR95).

In the context of 21-cm cosmology, Bayes-factor-based model comparison distinguishes between models that offer a compact explanation of the data (i.e., a good fit using relatively few effective degrees of freedom) and those that do not. However, it does not distinguish between:

(i) models in which the 21-cm signal component and the nuisance component (e.g., foregrounds, ionosphere, and instrument systematics) are each accurate, enabling unbiased recovery of the 21-cm signal, and

(ii) models in which the nuisance component is inaccurate, but its deficiencies are absorbed by the 21-cm signal model, still yielding a high Bayesian evidence fit – albeit with a biased signal estimate.

Here, we refer to the 21-cm signal component as the *model of interest*, and to the combination of all other components (e.g., astrophysical foregrounds, ionospheric effects, and residual instrumental systematics) as the *nuisance model*. While the nuisance model may itself have multiple subcomponents, we are concerned only with the possibility that inaccuracies in the nuisance model could correlate with the 21-cm signal model in a way that biases its inference.

Scenario (ii), above, can occur when a composite model has a nuisance model that fails to fully describe the nuisance signal component of the data, but is paired with a 21-cm model capable of fitting the sum of the true 21-cm signal *and* the residual systematics resulting from inaccuracies in the nuisance model. In this case, the composite model can provide a good fit to the data, but the 21-cm signal estimates it yields will be biased by the systematic residuals.

This situation introduces a form of *model-level degeneracy* (hereafter, *model degeneracy*), analogous to parameter degeneracy in conventional parameter estimation, but arising at the level of composite models. In such cases, different combinations of component models may provide similarly good fits to the data in aggregate, even though only some combinations yield accurate and unbiased recovery of the 21-cm signal.

To account for this phenomenon, two categories of composite

---

[2] The sky-averaged brightness temperature observed by the EDGES 2 instrument, when the Galaxy is low in the beam, is approximately 5000 K at 50 MHz (e.g. Bowman et al. 2018a). For a noise level of ∼ 20 mK – consistent with that reported in Bowman et al. (2018a) – this corresponds to a foreground-to-noise ratio of 2.5×10^5.

model comparison were defined in S25. Let the full set of models under consideration be denoted $\boldsymbol{M} = \{\boldsymbol{M}_{\text{inac}}, \boldsymbol{M}_{\text{ac}}\}$, where $\boldsymbol{M}_{\text{inac}}$ contains models that do not provide accurate or predictive fits to the data, and $\boldsymbol{M}_{\text{ac}}$ contains models that do. The key distinction between the two categories lies in the nature of the accurate models in $\boldsymbol{M}_{\text{ac}}$. In *category I* model comparison, all accurate models are only capable of fitting the data when their component sub-models are also accurate - no component can compensate for the inaccuracies of another. In this case, BFBMC is sufficient to distinguish between models that recover the true component signals and those that do not. Conversely, in *category II* model comparison, $\boldsymbol{M}_{\text{ac}}$ includes models that achieve high evidence fits by absorbing the inaccuracies of one component into another (e.g. a 21-cm signal component fitting residual foreground structure). In such cases, BFBMC alone leads to biased model-averaged inferences, and model validation becomes essential.

To address this issue, S25 introduced the BaNTER validation framework, which uses a Bayesian null test to derive model priors through comparison of a single component model and a composite model for single-component validation data set. This test is designed to identify composite models that, while fitting the data well over-all, yield biased signal inferences, enabling these poorly performing models to be downweighted or excluded a priori. Combining the BaNTER validation framework with Bayesian model comparison for observational data allows one to derive the model-validated posterior odds. This approach enables selection for models that are accurate and predictive of the data in aggregate and, crucially for unbiased 21-cm signal recovery, are composed of accurate and predictive component models.

In this work, we apply the BaNTER validation framework to derive model priors and use the model-validated posterior odds to compare the BFCC model with the Intrinsic, LinPhys, and MultLin models. We analyse realistic simulations of BFCC EDGES-low spectrometer data across null, moderate, and high amplitude 21-cm signals and demonstrate that this approach:

(i) favours the BFCC model over the alternatives, and

(ii) in contrast to using BFBMC-alone, reliably recovers 21-cm signal estimates consistent with the true signal in the data across all signal-amplitude regimes.

The remainder of the paper is organised as follows. In Section 2, we describe the Bayesian inference and BaNTER validation frameworks used to analyse the data and compare models. Section 3 summar-ises the simulations of BFCC EDGES-low data developed in Paper I and updates to the 21-cm signals included in the simulations in this work. In Section 4, we describe the models we fit to the data. Section 5 presents the analysis results of the simulated data sets and compares Bayesian model selection based on BFBMC-alone and the BaNTER-validated posterior-odds as methods for identify-ing preferred models for recovery of unbiased estimates of the 21-cm signal. In Section 6, we discuss the performance of the BFCC model relative to the Intrinsic, LinPhys, and MultLin models under differ-ent signal amplitude assumptions and emphasise the importance of model validation for deriving reliable inferences from the compar-ison of composite models for global 21-cm signal data. Finally, in Section 7, we summarise our findings and present our conclusions.

2 BAYESIAN INFERENCE AND MODEL VALIDATION FRAMEWORK

2.1 Bayesian inference

2.1.1 Bayes' theorem

Bayesian inference provides a consistent approach to estimate a set of parameters, $\boldsymbol{\Theta}$, from a model, $\boldsymbol{M}$, given a set of data, $\boldsymbol{D}$. Using Bayes' theorem we can write the posterior probability density of the parameters of the model as:

$$\mathcal{P}(\boldsymbol{\Theta}|\boldsymbol{D}, \boldsymbol{M}) = \frac{\mathcal{P}(\boldsymbol{D}|\boldsymbol{\Theta}, \boldsymbol{M}) \, \mathcal{P}(\boldsymbol{\Theta}|\boldsymbol{M})}{\mathcal{P}(\boldsymbol{D}|\boldsymbol{M})} \, . \tag{1}$$

Here, $\mathcal{P}(\boldsymbol{D}|\boldsymbol{\Theta}, \boldsymbol{M}) \equiv \mathcal{L}(\boldsymbol{\Theta})$ is the likelihood of the data, $\mathcal{P}(\boldsymbol{\Theta}|\boldsymbol{M}) \equiv \pi(\boldsymbol{\Theta})$ is the prior probability density of the parameters and $\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}) \equiv \mathcal{Z} = \int \mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})\mathrm{d}^n\boldsymbol{\Theta}$ is the Bayesian evidence, where n is the dimensionality of the parameter space.

2.1.2 Bayesian model comparison

Comparison of competing models in the light of observed data is a fundamental scientific goal. When one has a set of models for the data, $\boldsymbol{M} = \{\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_N\}$, preferred models can be determined from their marginal probabilities. Bayes' theorem for the marginal probability of a model gives:

$$\mathcal{P}(\boldsymbol{M}_i|\boldsymbol{D}, \boldsymbol{M}) = \frac{\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_i, \boldsymbol{M})\mathcal{P}(\boldsymbol{M}_i|\boldsymbol{M})}{\mathcal{P}(\boldsymbol{D}|\boldsymbol{M})} \, . \tag{2}$$

Here, $\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}) = \sum_{k=1}^{N} \mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_k, \boldsymbol{M})\mathcal{P}(\boldsymbol{M}_k|\boldsymbol{M})$ is the mar-ginal probability of the data over the models and their parameters, $\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_i, \boldsymbol{M})$ is the Bayesian evidence of $\boldsymbol{M}_i$ and $\mathcal{P}(\boldsymbol{M}_i|\boldsymbol{M})$ is the probability of $\boldsymbol{M}_i$ prior to analysing the data. For brevity, we leave the conditioning of the probability densities on $\boldsymbol{M}$ implicit going forward.

Bayesian methodology addresses model comparison between two possible models, $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$, for a data set, $\boldsymbol{D}$, via consideration of $\mathcal{R}_{ij}$, the posterior odds in favour of $\boldsymbol{M}_i$ over $\boldsymbol{M}_j$. Using Equation (2) we can write this as:

$$\mathcal{R}_{ij} = \frac{\mathcal{P}(\boldsymbol{M}_i|\boldsymbol{D})}{\mathcal{P}(\boldsymbol{M}_j|\boldsymbol{D})} = \frac{\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_i)\mathcal{P}(\boldsymbol{M}_i)}{\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_j)\mathcal{P}(\boldsymbol{M}_j)} \, . \tag{3}$$

Here, $\mathcal{P}(\boldsymbol{M}_i|\boldsymbol{D})$ is the posterior probability of model $\boldsymbol{M}_i$, $\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_i)/\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_j) \equiv \mathcal{B}_{ij}$ is the Bayes factor between the models, $\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_i) \equiv \mathcal{Z}_i$ and $\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_j) \equiv \mathcal{Z}_j$ are the Bayesian evidences of models $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$, respectively, and $\mathcal{P}(\boldsymbol{M}_i)/\mathcal{P}(\boldsymbol{M}_j)$ is the ratio of the prior probabilities of the two models before any conclusions have been drawn from the data.

As the model-prior-weighted average of the likelihood over the parameters priors, the marginal probability of a model is larger if the model is probable a priori and more of its parameter space is likely given the data; it is smaller for a model that is improbable a priori or if large areas of its parameter space have low likelihood values, even if the likelihood function is very highly peaked. It thus represents an updating of one's prior credence in the model, given the data, and automatically incorporates an 'Occam penalty' against a more complex theory with a broad parameter space. As such, in absence of an a priori reason to prefer it over a simpler alternative, it will be favoured only if it is significantly better at explaining the data.

In this work, we adopt the mapping of qualitative terms describ-ing the relative preference for one model over another defined in S25[3]. Specifically, we describe posterior odds in the range 1–3

---

[3] The mapping of qualitative terms defined in S25 generalises to $\mathcal{R}_{ij}$ the

$(0 \leq \ln(\mathcal{R}_{ij}) < 1)$ as a weak preference for M_i over M_j, posterior odds in the range 3–20 ($1 \leq \ln(\mathcal{R}_{ij}) < 3$) as a moderate preference, posterior odds in the range 20–150 ($3 \leq \ln(\mathcal{R}_{ij}) < 5$) as a strong preference, and posterior odds of greater than 150 ($5 \leq \ln(\mathcal{R}_{ij})$) as a decisive preference. When we have no information that lends additional credibility to one model over the other in advance of analysing the data, we set the prior odds to unity; therefore, the posterior odds are equal to the Bayes factor between the models ($\mathcal{R}_{ij} = \mathcal{B}_{ij}$). In this case, for the purpose of defining our qualitative descriptions, the Bayes factor between models takes the place of the posterior odds.

2.2 BaNTER validation

When the set of competing models under consideration includes at least one model capable of providing accurate and predictive fits to the data using biased component fits, the Bayes factor between such models and those in the subset of interest – those containing models with accurate and predictive subcomponent models – is insufficient to distinguish them. Comparison of models for global 21-cm data can fall under this scenario (S25). In such a *category II* model comparison, BFBMC enables one to separate models that are predictive of the data in aggregate from those that are not, but informative prior odds on the models are necessary to separate predictive composite models that also have accurate and predictive component fits to the data from those that do not.

In this work, we use the BaNTER validation framework for composite models introduced in S25 to validate our composite models for the data and derive informative model priors. For a detailed description of general BaNTER validation, we refer the reader to S25. Here, we provide a brief overview of the method in the context of global 21-cm cosmology.

2.2.1 Global 21-cm data

Consider a global 21-cm signal data set of the form:

$$D = f(\mathbf{\Theta}_{21}, \mathbf{\Theta}_{\mathrm{Fg}}) + n \,, \tag{4}$$

where $f(\mathbf{\Theta}_{21}, \mathbf{\Theta}_{\mathrm{Fg}})$ is the sum of global 21-cm signal and foreground emission in the data, n is the noise and $\mathbf{\Theta}_{21}$ and $\mathbf{\Theta}_{\mathrm{Fg}}$ are parameters underlying the physical processes producing the 21-cm signal and foreground emission. The function $f(.)$ captures the generative processes of the signal components, the propagation effects between the source of emission and the instrument (such as ionospheric absorption and refraction), the instrument transfer function, and any corrections applied to the data, such as beam factor chromaticity correction (see Paper I).

In many practical cases, the function $f(\cdot)$ can be approximated as nearly linear over the relevant parameter ranges (see Section 3)[4]. Under this approximation, the data model in Equation (4) reduces to:

$$D \simeq S_{21} + S_{\mathrm{Fg}} + n \,, \tag{5}$$

where S_{21} and S_{Fg} represent the apparent global 21-cm signal and

foreground components in the data, respectively. This linearised form provides intuitive insight into the data structure, although it is not a prerequisite for the validity of the BaNTER framework.

2.2.2 Foreground, 21-cm signal, and composite models

Now, assume we have a definitive model for the signal component, denoted $M_{21}(\mathbf{\Theta}_{21})$, where $\mathbf{\Theta}_{21}$ represents the true values of the model parameters to be determined from the data.[5] In this case, the set of models for S_{21} is given by:

$$\mathcal{M}_{21} = \{M_{21}(\mathbf{\Theta}_{21})\} \,. \tag{6}$$

For the foreground component, we define a competing set of models as the union of different foreground model classes, including BFCC, Intrinsic, LinPhys, and MultLin models. Explicitly, the set of all foreground models is:

$$\mathcal{M}_{\mathrm{Fg}} = \{M_i \mid M_i \in (\mathcal{M}_{\mathrm{BFCC}} \cup \mathcal{M}_{\mathrm{Intrinsic}} \cup \mathcal{M}_{\mathrm{LinPhys}} \cup \mathcal{M}_{\mathrm{MultLin}})\} \,, \tag{7}$$

where $\mathcal{M}_{\mathrm{BFCC}}$, $\mathcal{M}_{\mathrm{Intrinsic}}$, $\mathcal{M}_{\mathrm{LinPhys}}$, and $\mathcal{M}_{\mathrm{MultLin}}$ are the sets of BFCC, Intrinsic, LinPhys, and MultLin models, respectively. These sets vary in size, ranging from single models in $\mathcal{M}_{\mathrm{Intrinsic}}$ and $\mathcal{M}_{\mathrm{LinPhys}}$ to multiple models with varying complexity in $\mathcal{M}_{\mathrm{BFCC}}$ and $\mathcal{M}_{\mathrm{MultLin}}$ (see Section 4). Each model in $\mathcal{M}_{\mathrm{Fg}}$ is of the form $M_{i\mathrm{Fg}}(\mathbf{\Theta}_{i\mathrm{Fg}})$, where $\mathbf{\Theta}_{i\mathrm{Fg}}$ are the parameters of the ith foreground model.

Finally, given $\mathcal{M}_{21}$ and $\mathcal{M}_{\mathrm{Fg}}$, we define a set of composite models as:

$$\mathcal{M}_{\mathrm{c}} = \{M_{i\mathrm{c}} \mid i = 1, \ldots, N_{\mathrm{c}}\} \,, \tag{8}$$

where each composite model is of the form:

$$M_{i\mathrm{c}} = g(\mathbf{\Theta}_{21}, \mathbf{\Theta}_{i\mathrm{Fg}}) \,. \tag{9}$$

Here, $g(\mathbf{\Theta}_{21}, \mathbf{\Theta}_{i\mathrm{Fg}})$ is a model for $f(\mathbf{\Theta}_{21}, \mathbf{\Theta}_{\mathrm{Fg}})$ and N_{c} represents the total number of composite models in the set. In the linearised case $M_{i\mathrm{c}}$ simplifies to:

$$M_{i\mathrm{c}} = M_{21} + M_{i\mathrm{Fg}} \,, \tag{10}$$

with $M_{21}(\mathbf{\Theta}_{21})$ being the model for the 21-cm signal component and $M_{i\mathrm{Fg}}(\mathbf{\Theta}_{i\mathrm{Fg}})$ being the ith model for the foreground component.

2.2.3 Bayesian null test

The BaNTER validation framework provides a means of separating models in $\mathcal{M}_{\mathrm{c}}$ that are able to accurately describe the data with accurate component models from those that provide accurate fits to the data but lead to biased inferences for the parameters of the 21-cm signal model, M_{21}, if present in the data. In the general (non-linear) case, this is achieved by comparing the Bayesian evidence of a composite model, $M_{\mathrm{c}}(\mathbf{\Theta}_{21}, \mathbf{\Theta}_{\mathrm{Fg}})$, against that of a foreground-only model, $M_{\mathrm{Fg}}(\mathbf{\Theta}_{\mathrm{Fg}})$, for a foreground-only validation data set of the form:

$$D_{\mathrm{v}} = S_{\mathrm{Fg}} + n \,. \tag{11}$$

Since the validation data used in this comparison is constructed to contain only a foreground component, any preference for the composite model over the foreground-only model indicates that the signal

---

mapping for $\mathcal{B}_{ij}$ introduced in KR95. When model priors are uninformative, it reduces to the model-odds thresholds established in KR95.

[4] Ionospheric absorption, as well as instrumental losses, act multiplicatively on both the astrophysical foregrounds and 21-cm signal. However, these effects are typically at the sub-10% level, and since their fits are dominated by foreground emission 4–6 orders of magnitude brighter than the 21-cm signal, the induced coupling is expected to be negligible in practice.

[5] See S25 for a discussion on how the BaNTER validation framework can be generalised to cases where both S_{21} and S_{Fg} are uncertain.

model is absorbing residual structure due to an inaccurate foreground model.

In this work, we use high-fidelity simulated foreground-only observations as the S_{Fg} component of our validation data (see Section 3.3).

Given D_{v} and a composite-foreground pair of models, $M_{i\mathrm{c}}$ and $M_{i\mathrm{Fg}}$, the BaNTER validation proceeds by fitting each model to the validation data and computing the null-test evidence ratio (Bayes factor):

$$\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{cFg}}) = \ln\left(\frac{\mathcal{Z}^{\mathrm{v}}_{\mathrm{c}}}{\mathcal{Z}^{\mathrm{v}}_{\mathrm{Fg}}}\right) , \qquad (12)$$

where $\mathcal{Z}^{\mathrm{v}}_{\mathrm{c}} = \mathcal{P}(D_{\mathrm{v}}|M_{i\mathrm{c}})$ and $\mathcal{Z}^{\mathrm{v}}_{\mathrm{Fg}} = \mathcal{P}(D_{\mathrm{v}}|M_{i\mathrm{Fg}})$ are the Bayesian evidences for the composite model $M_{i\mathrm{c}}$ and its foreground component $M_{i\mathrm{Fg}}$, respectively.

When $\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{cFg}}) \geq 0$, the composite model fits the validation data better than the foreground-only model. Since the validation data contains only foregrounds, this preference indicates that the 21-cm component is absorbing systematic residuals from an imperfect foreground model. For example, if a foreground model inadequately describes chromatic instrumental effects, the 21-cm component might fit both the true signal and these residual systematics, yielding a good overall fit to the data in aggregate but biased 21-cm signal estimates.

The validation metric $\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{cFg}})$ becomes large only when the foreground model is insufficiently accurate in describing the foreground component of the data *and* when a spurious fit of the 21-cm signal model absorbs residual structure in the validation data that the foreground model alone cannot fit.

The composite model is deemed to fail the null test if $\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{cFg}}) \geq \ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{threshold}})$, where $\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{threshold}})$ is a predefined threshold. We interpret different ranges of $\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{cFg}})$ as follows:

- < 0: Foreground-only model preferred. This is expected for foreground-only validation data.
- $0 - 3$: Moderate systematic contamination likely to bias signal estimates if a 21-cm signal is present.
- ≥ 3: Severe systematic contamination likely both to bias 21-cm signal recovery, if a 21-cm signal is present, or produce a false detection, if not.

Specifically, for $0 \leq \ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{cFg}}) < 3$, the composite model provides a better fit to the validation data than the foreground-only model. In the context of global 21-cm signal datasets, this suggests that inaccuracies in the foreground model are sufficient to bias estimates of the 21-cm signal, if present. However, under a Bayesian 21-cm detection criterion that requires strong evidence in favour of the composite model over the foreground-only model ($\ln(\mathcal{B}_{\mathrm{cFg}}) \geq 3$; see Section 2.4), these inaccuracies are too small to produce a false detection of the 21-cm signal, when it is absent.

For $\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{cFg}}) \geq 3$, the composite model provides a substantially better fit to the validation data than the foreground-only model, indicating that foreground model inaccuracies are severe enough to significantly bias 21-cm signal estimates, if present, or to lead to a false detection, if absent.

In this work, we follow S25 and adopt a conservative approach by setting $\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{threshold}}) = 0$. We treat the prior odds of failed composite models yielding unbiased estimates of the 21-cm signal in a global dataset as negligible when compared to models that successfully pass BaNTER validation.

2.2.4 Model comparison categorisation

The possibility exists for composite models that are predictive of the data in aggregate to obtain accurate fits with biased component models. However, in the absence of informative model priors, it is uncertain whether models of this type are included in the set of models under consideration. Following S25, we categorize the Bayesian comparison of composite models in $\mathcal{M}_{\mathrm{c}}$ as a *category I* model comparison problem if $\mathcal{M}_{\mathrm{c}}$ contains no such models, and a *category II* model comparison problem if such models are present.

For a *category I* model comparison problem, model validation is incidental to the recovery of unbiased 21-cm signal estimates through Bayesian analysis of the data. In contrast, for a *category II* model comparison problem, model validation becomes essential. Thus, understanding which of these categories applies to the Bayesian comparison of a given set of composite models is key to determining whether model validation is necessary for drawing robust inferences with them and, equivalently, for assessing the degree of confidence that can be placed in the conclusions drawn from the comparison of unvalidated models.

We use the BaNTER validation framework to determine whether the Bayesian comparison of the models considered here falls under *category I* or *category II* in Section 5.

2.3 Data likelihood

Let the data, vectorised over frequency, be denoted as D, and define a corresponding vectorised model, parameterised by Θ, as $M(\Theta)$.

We assume the noise in the data follows a zero-mean Gaussian distribution, uncorrelated between frequency channels. Consequently, we model the noise covariance matrix, $\mathbf{N}$, as diagonal, with elements given by:

$$N_{ij} = \left\langle n_i n_j^* \right\rangle = \delta_{ij}\sigma^2 , \qquad (13)$$

where $\langle \cdot \rangle$ denotes the expectation value, and σ is the root-mean-square (RMS) noise level in the data.

Defining the residuals between the data and model as $R = D - M(\Theta)$, the Gaussian likelihood function for R is given by:

$$\mathcal{P}(D|\Theta) = \frac{1}{\sqrt{(2\pi)^{N_{\mathrm{chan}}}\det(\mathbf{N})}} \exp\left[-\frac{1}{2}R(\Theta)^T \mathbf{N}^{-1} R(\Theta)\right] . \quad (14)$$

When fitting the data in Section 5, D represents vectorised, simulated beam factor chromaticity-corrected data (see Section 3.1). For a spectrum $X(\nu)$, we define the vectorisation operator $\mathrm{vec}(\cdot)$ such that:

$$\mathrm{vec}(X(\nu)) = [X_0, X_1, \ldots, X_{N_{\mathrm{chan}}}]^T , \qquad (15)$$

where X_i is the value of X at frequency channel i, and N_{chan} is the total number of channels in the dataset.

2.4 21-cm signal detection

We propose that a robust detection of the 21-cm signal should satisfy the following criteria:

(i) The subcomponents of the model must provide an accurate description of their respective signal components. This requirement prevents errors in one model component from being absorbed by another, which could result in an accurate fit to the data in aggregate but a biased recovery of the 21-cm signal.

(ii) The model incorporating the 21-cm signal must provide an accurate description of the data, such that the residuals are consistent with the expected noise level.

(iii) There must be strong Bayesian evidence favouring models that include a 21-cm signal component over those that do not.

Criteria (i) and (ii) can be assessed using BaNTER-validated posterior-odds-based model comparison (Sections 2.1 and 2.2.3) and an analysis of model residuals (see Appendix C), respectively. Given observational data D and a composite-foreground pair of models, $M_{i\mathrm{c}}$ and $M_{i\mathrm{Fg}}$, criterion (iii) is satisfied if the following threshold for 21-cm signal detection is met:

$$\ln(\mathcal{B}_{\mathrm{cFg}}) \geq 3 . \tag{16}$$

Here, $\ln(\mathcal{B}_{\mathrm{cFg}}) = \ln(\mathcal{Z}_{\mathrm{c}}/\mathcal{Z}_{\mathrm{Fg}})$ is the log Bayes factor in favour of the composite model, where $\mathcal{Z}_{\mathrm{c}} = \mathcal{P}(D|M_{i\mathrm{c}})$ and $\mathcal{Z}_{\mathrm{Fg}} = \mathcal{P}(D|M_{i\mathrm{Fg}})$ denote the Bayesian evidences for the composite model $M_{i\mathrm{c}}$ and its foreground-only counterpart $M_{i\mathrm{Fg}}$, respectively. This threshold corresponds to odds of at least 20:1 in favour of the composite model, providing strong evidence for the presence of a 21-cm signal in the data.

To test for the presence of a 21-cm signal in the data, we apply the criteria outlined above in Section 5 and define the full set of models under consideration as:

$$\mathcal{M} = \{M_i \mid M_i \in (\mathcal{M}_{\mathrm{c}} \cup \mathcal{M}_{\mathrm{Fg}})\} . \tag{17}$$

2.5 Computational techniques

2.5.1 Probability densities

In Section 5, when analysing the data, we estimate model evidences and sample from the posteriors on the model parameters, given the data, using nested sampling as implemented by the PolyChord algorithm (Handley et al. 2015b,a). Given samples from the posterior distribution of the parameters, $\mathcal{P}(\Theta|D, M)$, one can estimate $\mathcal{P}(y|\Theta, \nu, D, M)$, the posterior predictive density (posterior PD) of a function $y = f(\Theta, \nu)$ by calculating the corresponding set of samples from $\mathcal{P}(y|\Theta, \nu, D, M)$. We derive contour plots of prior and posterior PDs using the fgivenx software package (Handley 2018).

2.5.2 Summary statistics

Many of the aforementioned parameter posteriors will be characterised by non-Gaussian probability density functions (PDFs). Therefore, following Sims et al. (2025b), we use the highest probability density estimates (HPDEs) and highest probability density intervals (HPDIs; e.g. Hyndman 1996), $X_{\mathrm{HPD}}|_{-\sigma_-}^{+\sigma_+}$ as informative summary statistics of these distributions. Here, X_{HPD} is the HPDE value of the PDF of a parameter (or set of parameters), X, and $\sigma_\pm = |X_{\mathrm{HPDI}\pm} - X_{\mathrm{HPD}}|$ characterises its width, with $X_{\mathrm{HPDI}+}$ and $X_{\mathrm{HPDI}-}$ the upper and lower bound of the HPDI, respectively.

3 SIMULATED DATA

We construct realistic simulations of time-averaged BFCC data following the approach described in Paper I; for a detailed description we refer the reader to that work. In this section, we provide a summary of the approach including the updates to the 21-cm signal component in the simulations used here.

3.1 Spectrometer data in the snapshot limit

Working in the reference frame of the antenna, under the assumption that the integration time, Δt, is sufficiently short for the measurement to be accurately approximated as an instantaneous snapshot at the central time of the integration ($\Delta t \lesssim 10$ mins for EDGES 2), a calibrated autocorrelation spectrum derived from a zenith-pointing antenna, such as EDGES, can be written as (see Paper I):

$$T_{\mathrm{data}}(\nu, t) = \int_{\Omega^+} B(\nu, \Omega) T_{\mathrm{sky}}(\nu, \Omega, t)\, \mathrm{d}\Omega + n . \tag{18}$$

Here, $T_{\mathrm{sky}}(\nu, \Omega, t)$ represents the time-dependent sky brightness temperature distribution above the antenna, $d\Omega$ is a solid angle element, n denotes instrumental noise and Ω^+ is the skyward hemisphere centred on zenith. The term $B(\nu, \Omega) = \frac{1}{D_{\Omega^+}} D(\nu, \Omega)$ describes the frequency and direction-dependent antenna beam. It is normalised such that the beam pattern integrates to 1 over the skyward hemisphere with $D(\nu, \Omega)$ the antenna directivity pattern and $D_{\Omega^+} = \int_{\Omega^+} D(\Omega)\mathrm{d}\Omega$.

For an instrument incorporating a large ground plane below the antenna, such as EDGES, the region Ω^+ encompasses nearly the full integral antenna directivity. Specifically, for the H2 configuration of the EDGES 2 low-band instrument with a 30 m × 30 m sawtooth ground plane, detailed electromagnetic simulations of the antenna directivity (e.g. Mahesh et al. 2021) indicate that the fractional directivity towards the nadir-centred hemisphere at a fixed frequency is $1 - D_{\Omega^+}/D_{\mathrm{full}} \simeq 10^{-3}$. Here, $D_{\mathrm{full}} = \int_0^{4\pi} D(\Omega)\mathrm{d}\Omega$ is the integral antenna directivity over the full sphere. In this work, we assume that the fractional directivity towards the ground, on the order of 10^{-3}, has been accurately accounted for through ground-loss correction (e.g. Rogers & Bowman 2012; Monsalve et al. 2017b). Additionally, the data has been calibrated such that $\int_0^{\Omega^+} B(\nu, \Omega)\mathrm{d}\Omega = 1$ and $T_{\mathrm{data}}(\nu, t)$ is an absolute temperature measurement.

3.2 Beam factor chromaticity correction

For the purpose of global 21-cm signal data analysis, we can write the sky brightness temperature in the $10 \lesssim \nu \lesssim 230$ MHz frequency range as the sum of two components:

(i) A bright but spectrally smooth non-21-cm component comprised of synchrotron emission from the Galaxy and extragalactic sources, with a smaller contribution from Galactic free-free emission, and thermal emission from the Earth's ionosphere.

(ii) A redshifted 21-cm signal component with less smooth spectral structure determined in detail by the sky-averaged evolution with redshift of the ionization and temperature state of hydrogen and the relative coupling strength of the neutral hydrogen spin temperature to its kinetic temperature and the background radiation temperature.

Additionally, the Earth's ionosphere refracts and absorbs both of these components in a frequency-dependent manner (e.g. Vedantham et al. 2014; Shen et al. 2021).

Despite a dynamic range of several orders of magnitude between these two components, in the absence of instrumental effects and barring a significant level of Faraday rotated polarised foreground emission (e.g. Spinelli et al. 2019), the effective foreground after passing through the ionosphere is expected to be spectrally separable from the 21-cm signal. However, instrumental chromaticity, if unaccounted for and in excess of the dynamic range between the foregrounds and 21-cm signal, will eliminate this separation of characteristic spectral scales and will introduce foreground systematics

greater than or equal in amplitude to the 21-cm signal of interest, biasing its recovery by spectral means.

The impact of instrumental chromaticity on the separation of the 21-cm signal from the non-21-cm component of the data can be significantly mitigated (although not entirely removed) by dividing the calibrated autocorrelation spectrum by a beam chromaticity correction factor, B_{factor}, that describes the average spectral structure of the beam weighted by the brightness temperature distribution of the sky at a given reference frequency (see Paper I for details). In the short integration snapshot limit, B_{factor} is given by (e.g. Mozdzen et al. 2017, 2019),

$$B_{\text{factor}}(\nu, t) = \frac{\int_{\Omega^+} B^{\text{m}}(\nu, \Omega) T_{\text{fg}}^{\text{m}}(\nu_{\text{c}}, \Omega, t) \text{d}\Omega}{\int_{\Omega^+} B^{\text{m}}(\nu_{\text{c}}, \Omega) T_{\text{fg}}^{\text{m}}(\nu_{\text{c}}, \Omega, t) \text{d}\Omega}, \quad (19)$$

and the BFCC data has the form,

$$T_{\text{corrected}}(\nu, t) = T_{\text{data}}(\nu, t)/B_{\text{factor}}(\nu, t). \quad (20)$$

Time-averaged BFCC data, $T_{\text{corrected}}(\nu)$, such as that analysed in B18 and also the subject of the analysis here, is formed by averaging $T_{\text{corrected}}(\nu, t)$ over t.

Here, as in Paper I, we focus on the effectiveness of BFCC when one has an accurate model for B^{m} and $T_{\text{fg}}^{\text{m}}(\nu_{\text{c}}, \Omega, t)$. In upcoming work we will explore how 21-cm signal recovery with the BFCC model derived in Paper I is impacted by realistic deviations from the assumption of an error-free model for $T_{\text{fg}}^{\text{m}}(\nu_{\text{c}}, \Omega, t)$ and B^{m}.

3.3 Simulations

To construct $T_{\text{corrected}}(\nu)$, we first construct simulated time-dependent EDGES-low spectrometer data, $T_{\text{data}}(\nu, t)$, following Equation (18), at 120 times, spaced by 6 minute intervals, in the LST range $0 \leq LST < 12$ h, selected to match the LST window of the publicly available EDGES-low data, when the Galactic plane is relatively low in the beam. We simulate data over a $50 - 100$ MHz spectral band, assuming a 1 MHz channel width and observation of a sky model composed of:

- foregrounds with realistic spatially dependent spectral structure,
- spectrally-dependent absorption by the ionosphere,
- ionospheric emission, and
- a flattened Gaussian redshifted 21-cm signal profile.

We simulate our sky model including the aforementioned components as,

$$T_{\text{sky}}(\nu, \Omega, t) = \left[(T_{\text{fg}}(\nu, \Omega, t) + T_{21}) e^{-\tau_{\text{ion}}(\nu)} \right.$$
$$\left. + T_{\text{e}}(1 - e^{-\tau_{\text{ion}}(\nu)}) \right), \quad (21)$$

where T_{e} and τ_{ion} are the temperature of electrons and opacity of the ionosphere, respectively. Here, following Paper I, we use, $T_{\text{e}} = 450$ K and $\tau_{\text{ion}} = \tau_0 (\nu/\nu_{\text{c}})^{-2}$, with $\tau_0 = 0.014$ at reference frequency $\nu_{\text{c}} = 75$ MHz (e.g. Rogers et al. 2015). The foreground brightness temperature distribution is given by,

$$T_{\text{fg}}(\nu, \Omega, t) = T_{\text{fg}}(\nu_{\text{c}}, \Omega, t) \left(\frac{\nu}{\nu_{\text{c}}} \right)^{-\beta_{\Omega,t}} + T_\gamma. \quad (22)$$

Here, $T_{\text{fg}}(\nu_{\text{c}}, \Omega, t)$ is the spectral power law component of the foreground brightness temperature at reference frequency ν_{c}. $\beta_{\Omega,t}$ is the spatially dependent spectral index distribution characterising the power law structure of that emission. $T_\gamma = 2.725$ K is the CMB temperature and T_{21} is the 21-cm signal in the data.

Following Paper I, we derive $T_{\text{fg}}(\nu_{\text{c}}, \Omega, t)$ as a spectral extrapolation from the Haslam 408 MHz all-sky map (Haslam et al. 1981, 1982) reprocessed by Remazeilles et al. (2015) and $\beta_{\Omega,t}$ as a spatially dependent spectral index distribution derived from the global sky model (GSM; Zheng et al. 2017). Furthermore, we model the global 21-cm signal as a flattened-Gaussian absorption trough, matching the model parametrisation used in B18:

$$T_{21}(\nu) = -A \left(\frac{1 - e^{-\tau e^{B_{21}}}}{1 - e^{-\tau}} \right), \quad (23)$$

where,

$$B_{21} = \frac{4(\nu - \nu_0)^2}{w^2} \log \left[-\frac{1}{\tau} \log \left(\frac{1 + e^{-\tau}}{2} \right) \right], \quad (24)$$

and A, ν_0, w and τ describe the amplitude, central frequency, width and flattening of the absorption trough, respectively.

In the simulated data sets analysed in this work, we incorporate absorption profiles with position and shape parameters: $\nu_0 = 78$ MHz, $w = 19$ MHz and $\tau = 8$ matching the 21-cm signal shape parameters considered in Paper 1, for ease of comparison[6]. For the amplitude of the signal, we consider three cases:

(i) *A null-amplitude 21-cm signal* ($A = 0$ mK). We use this as a null-test to identify models that lead to spurious signal detection through joint estimation of a 21-cm signal with an insufficiently accurate foreground and ionosphere model.

(ii) *A moderate-amplitude 21-cm signal* ($A = 150$ mK), with a signal amplitude consistent with expectations under standard cosmological assumptions regarding cooling of the hydrogen gas during the Dark Ages and a background radiation temperature during CD dominated by the CMB.

(iii) *A high-amplitude 21-cm signal* ($A = 500$ mK), consistent with the best-fit recovered in B18 and explainable in a physically motivated manner with additional cooling of the hydrogen gas beyond that due to adiabatic expansion and/or an additional radio background raising the total radio background temperature in excess of the CMB.

We construct our time-dependent beam factor model, $B_{\text{factor}}(\nu, t)$, and BFCC data, $T_{\text{corrected}}(\nu, t)$, in the manner outlined in Section 3.2. For our beam model, $B(\nu, \Omega)$, we use the FEKO EM simulation of the EDGES-low blade dipole antenna with a 30 m × 30 m sawtooth ground plane from Mahesh et al. (2021). We calculate our time-averaged BFCC data, $T_{\text{corrected}}(\nu)$, by averaging $T_{\text{corrected}}(\nu, t) = T_{\text{data}}(\nu, t)/B_{\text{factor}}(\nu, t)$ over the simulated snapshot spectra. We add noise to the data at a level such that the resultant noise in the BFCC data, after time-averaging, is Gaussian and white, with an RMS amplitude of 20 mK that is comparable to estimates of the noise in the publicly available EDGES-low data (e.g. Singh & Subrahmanyan 2019). In all of our simulations, we assume the receiver calibration of the data is unbiased and uncertainty free (see Murray et al. 2022 for a discussion of the impact of uncertainty and bias in the receiver calibration parameter estimation).

Figure 1 illustrates the key astrophysical components of our simulated data sets. Our intrinsic foreground brightness temperature distribution model, evaluated at the centre of our simulated spectral band, $T_{\text{fg}}(75 \text{ MHz}, l, b)$, is shown in Figure 1a. Our model for the foreground spectral index distribution $\beta(l, b)$ is shown in Figure 1b.

---

[6] Given that there are potentially correlated effects that would result from varying both amplitudes and shape simultaneously, by taking this approach we are able to explicitly isolate and test the effect of underlying 21-cm amplitude on signal recovery.

Parameter	Prior
A	$U(0, 1)$ K
ν_0	$U(55, 95)$ MHz
w	$U(5, 30)$ MHz
τ	$U(0, 20)$

Figure 1c shows the output simulated time-averaged, beam factor chromaticity corrected spectrum and Figure 1d illustrates the injected 21-cm signals in the data in the three signal amplitude regimes we consider.

4 DATA MODELS & LIKELIHOOD

We consider four classes of composite models, each assuming the data is composed of 21-cm signal and non-21-cm signal components and instrumental noise (Equation (4)). The non-21-cm signal component models (i) astrophysical foreground emission following propagation through the ionosphere, where it undergoes chromatic absorption, (ii) ionospheric emission, and (iii) any residual instrumental effects in the data not perfectly correct by beam-factor-based chromaticity correction. For brevity, we use 'foreground component' as a shorthand for the non-21-cm signal component going forward.

4.1 Signal model

In each case, following B18, we model the 21-cm signal component of the data as a flattened Gaussian model (Equation (23)). We assume priors on the amplitude, central frequency, width and flattening of the absorption trough as listed in Table 1.

When fitting the data, we jointly estimated the 21-cm signal component model with one of four foreground components. The first is the BFCC model, which is the model we developed in Paper I of this series. The second is the Intrinsic model, which is a physically motivated parametrisation of the foreground component of the sky signal after propagation through the ionosphere. The third is the LinPhys model, which is a linear approximation to the Intrinsic model with uninformative priors on the parameters of the model. The fourth is the MultLin model, which is a more general polynomial foreground model also used as the foreground model in some recovered 21-cm signal estimates in B18, again assuming uninformative priors on the parameters of the model. We describe the four models in more detail below.

4.2 BFCC foreground model

In Paper I of this series we derived the BFCC model: a flexible closed-form model for BFCC spectrometer data. The BFCC model explicitly accounts for and models:

- the effect of realistic spatially dependent spectral structure of foreground emission,
- frequency dependent absorption of the foreground and 21-cm emission while propagating through the ionosphere,
- emission by high-temperature electrons in the ionosphere, and
- re-weighting of all components of the data, including the 21-cm signal and noise during BFCC.

In Paper I, the BFCC model is shown to enable unbiased recovery of a high-amplitude simulated global 21-cm signal. For a detailed derivation of the model, we refer the reader to that work. Here, we quote the final form of the model:

$$
\begin{aligned}
T_{\text{BFCC,fg}}^{\text{model}}(\nu) = &\left[\bar{T}_{\text{m}_0} \left(\frac{\nu}{\nu_c} \right)^{-\beta_0} \left(1 + \sum_{\alpha=1}^{N_{\text{pert}}} p_\alpha \ln \left(\frac{\nu}{\nu_c} \right)^\alpha \right) \right. \\
&+ \frac{\left(1 - \left(\frac{\nu}{\nu_c} \right)^{-\beta_0} \right) T_\gamma}{\bar{B}_{\text{factor}}(\nu)} + \frac{T_{21}}{\bar{B}_{\text{factor}}(\nu)} \left. \right] e^{-\tau_{\text{ion}}(\nu)} \\
&+ \frac{T_e}{\bar{B}_{\text{factor}}(\nu)} \left(1 - e^{-\tau_{\text{ion}}(\nu)} \right) . \quad (25)
\end{aligned}
$$

The first and second terms in Equation (25) describe the spatially-isotropic and -anisotropic subcomponents of the power-law component of the foreground emission, respectively. The third term accounts for the beam-factor weighted (following BFCC) CMB temperature, along with the spatially isotropic subcomponent of the foreground emission where the beam-factor does not cancel during BFCC. The fourth term represents the beam-factor weighted global 21-cm signal temperature. The common product of the terms in square brackets, $e^{-\tau_{\text{ion}}(\nu)}$, models ionospheric absorption, with the effective ionospheric optical depth modelled as $\tau_{\text{ion}} = \tau_0 (\nu/\nu_c)^{-2}$. Finally, the fifth term models the beam-factor weighted net emission from hot electrons in the ionosphere.

Equation (25) has $N_{\text{pert}} + 2$ foreground parameters, 2 ionospheric parameters and N_{21} 21-cm model parameters. Of the foreground model parameters, $\bar{T}_{\text{m}_0}$ describes the time- and sky-averaged non-21-cm-signal component of the sky brightness temperature at reference frequency ν_c. β_0 describes the mean temperature spectral index of the power law component of the foreground emission. p_α describes the fractional amplitude of the αth log-polynomial model vector for describing spectral fluctuations about the sky-averaged spectrum of the foreground brightness temperature field, normalised to the fractional amplitude of the perturbation relative to the mean brightness temperature at the reference frequency ν_c. We use Bayesian model comparison to determine the preferred number of log-polynomial model vectors to describe the data, N_{pert}. The two free parameters of the ionospheric model, T_e and τ_0 describe the temperature of ionospheric electrons and the effective ionospheric optical depth at ν_c, respectively. For the flattened Gaussian 21-cm absorption trough considered in this work, $N_{21} = 4$ and the parameters of the model are the amplitude, A, central frequency, ν_0, width, w and flattening, τ, of the absorption trough (see Equation (23)).

Of the above parameters, one can define physical priors for $\bar{T}_{\text{m}_0}, \beta_0$, T_e and τ_{ion} based on existing observations (see Paper I and references therein for details). The p_α parameters correspond to the temperatures of individual perturbation spectral model vectors at reference frequency $\nu_c = 75$ MHz. The fraction of the antenna temperature described by these terms is expected to be small relative to $\bar{T}_{\text{m}_0}$. In Paper I, it was found that limiting individual perturbation model vectors to 10% absolute fractional perturbations provided sufficient flexibility to accurately model simulated foreground-only BFCC data, without adding a significant degree of superfluous flexibility. We adopt the same range here.

Following Paper I, we incorporate this information when fitting Equation (25) to the simulated data in Section 5, in a conservative manner, using broad physical priors on the parameters of the model as listed in Table 2.
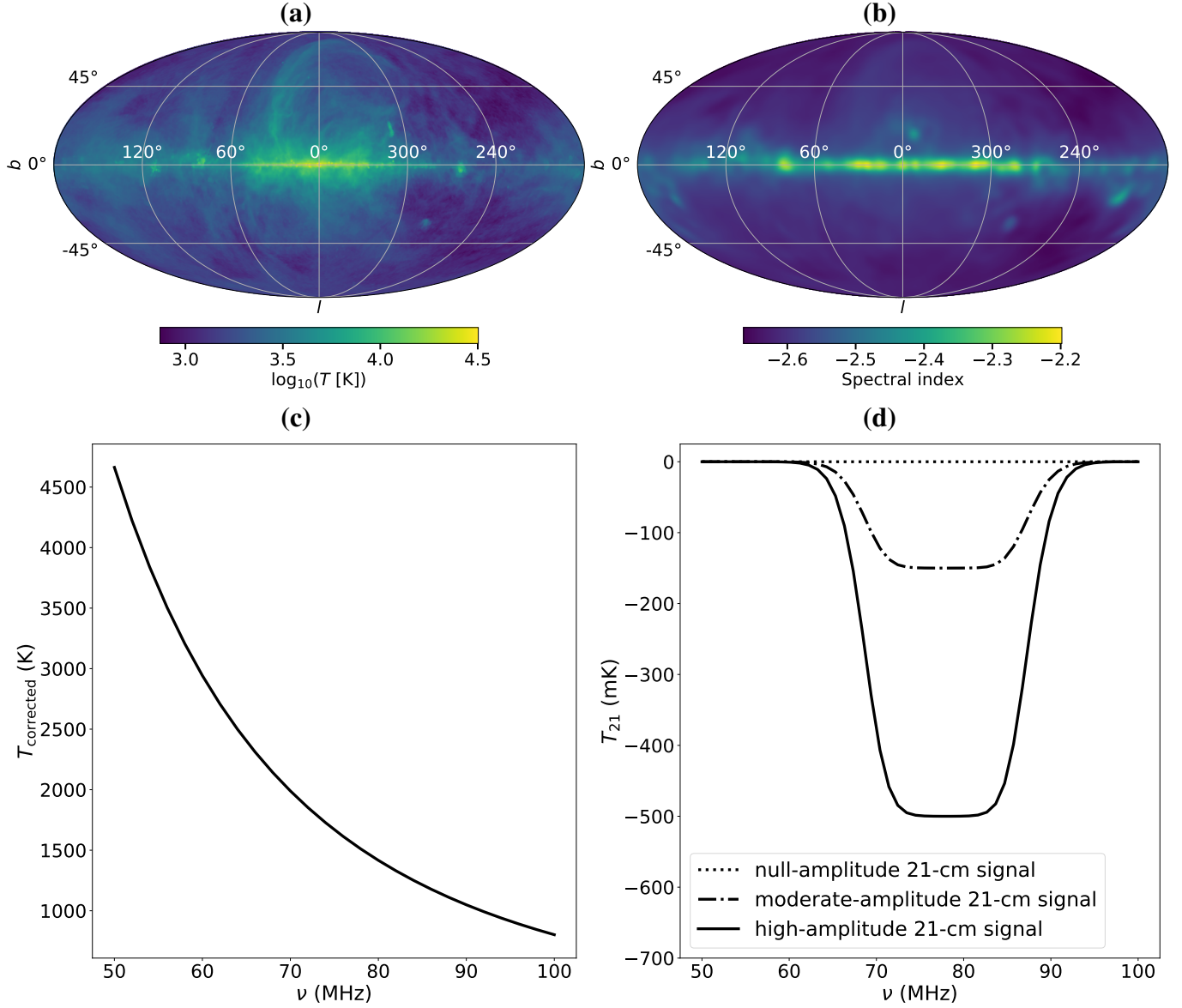
Figure 1. Astrophysical components of our simulated data sets. Figure 1a: Our intrinsic foreground brightness temperature distribution model, evaluated at the centre of our simulated spectral band, $T_{\rm fg}(75\,{\rm MHz},l,b)$. Figure 1b: The spatially-dependent foreground spectral index distribution $\beta(l,b)$ used when constructing simulated observational data. Figure 1c: Simulated time-averaged, beam factor chromaticity corrected spectrum resulting from time-averaging simulated BFCC EDGES low-band data over 120 simulated snapshot spectra derived at 6 minute intervals in the LST range $0 \le LST < 12$ h, matching the LST window of the publicly available EDGES low-band data. Figure 1d: Input 21-cm signal, in the simulated BFCC data, in the three signal amplitude regimes analysed in Section 5.

4.3 Intrinsic foreground model

A detailed description and derivation of the Intrinsic sky model is given in Paper I. Here, we quote the final form of the model:

$$T_{\rm Intrinsic,fg}^{\rm model}(\nu) = b_0 \left(\frac{\nu}{\nu_{\rm c}}\right)^{-2.5+b_1+b_2\log\left(\frac{\nu}{\nu_{\rm c}}\right)} {\rm e}^{-b_3\left(\frac{\nu}{\nu_{\rm c}}\right)^{-2}} + b_4 \left(\frac{\nu}{\nu_{\rm c}}\right)^{-2}.$$

(26)

In this construction, b_i with $i \in [0, \cdots, 4]$ are foreground and ionospheric parameters to be determined in the fit of the model to the data and they acquire direct interpretations in terms of physical properties of the foreground sky and the ionosphere in Equation (25) as follows (see Paper I, Appendix E for details).

- $b_0 = \bar{T}_{\rm m_0} e^{-\tau_{\rm ion}(\nu)}$ is the (attenuated) mean amplitude of the foreground power-law emission at the reference frequency ν_c.

- $b_1 = 2.5 - \beta_0$, with $\beta_0 \approx 2.5$ the power-law spectral index of the radio foreground emission, in the $50 \lesssim \nu \lesssim 190$ MHz band, when the Galactic Centre is in the sky (e.g. Mozdzen et al. 2017, 2019), and b_1 represents a deviation from this value.

- $b_2 = \sigma_\beta^2/2$, where σ_β^2 is the variance of the spectral index across the sky. This term encodes the amount of foreground spectral curvature generated by averaging over the spatially dependent spectral index distribution visible to the instrument.

- $b_3 = \tau_0$ gives the ionospheric opacity at ν_c (with frequency scaling as $[\nu/\nu_c]^{-2}$).

- $b_4 = T_e\tau_0$, with T_e as the electron temperature.

Table 2. Priors on the parameters of the foreground and ionospheric models defined in Section 4 and fit in Section 5.

.

Model	Parameter	Model component	Prior
BFCC	$\bar{T}_{\mathrm{m}_0}$	foreground	$U(1000, 6000)$ K
	β_0	foreground	$U(2.0, 3.0)$
	p_α	foreground	$U(-0.1, 0.1)$
	T_e	ionosphere	$U(100, 800)$ K
	τ_0	ionosphere	$U(0.005, 0.025)$
Intrinsic	b_0	foreground	$U(1000, 6000)$ K
	b_1	foreground	$U(-0.5, 0.5)$
	b_2	foreground	$U(0, 0.2)$
	b_3	ionosphere	$U(0.005, 0.025)$
	b_4	ionosphere	$U(0.5, 20.0)$ K
LinPhys	a_0	foreground + ionosphere	$U(1000, 6000)$ K
	$a_{1\cdots4}$	foreground + ionosphere	$U(-10^4, 10^4)$
MultLin	c_0	foreground + ionosphere	$U(1000, 6000)$ K
	c_i	foreground + ionosphere	$U(-10^4, 10^4)$

The priors we use when fitting Equation (26) are listed in Table 2. They are set to be equivalent to the priors on $\bar{T}_{\mathrm{m}_0}$, β_0, T_e and τ_ion in the BFCC model.

4.4 LinPhys foreground model

Assuming $b_i \ll 1$ with $i \in [1, 2, 3]$, the linearisation of Equation (26), over these parameters, will accurately approximate the full non-linear model. Performing this linearisation yields the polynomial foreground model used for recovery of the 21-cm signal in B18 (their Equation 1),

$$T_{\mathrm{LinPhys}}^{\mathrm{model}}(\nu) = a_0 \left(\frac{\nu}{\nu_\mathrm{c}}\right)^{-2.5} + a_1 \left(\frac{\nu}{\nu_\mathrm{c}}\right)^{-2.5} \log\left(\frac{\nu}{\nu_\mathrm{c}}\right) +$$
$$a_2 \left(\frac{\nu}{\nu_\mathrm{c}}\right)^{-2.5} \left[\log\left(\frac{\nu}{\nu_\mathrm{c}}\right)\right]^2 + a_3 \left(\frac{\nu}{\nu_\mathrm{c}}\right)^{-4.5} + a_4 \left(\frac{\nu}{\nu_\mathrm{c}}\right)^{-2} . \quad (27)$$

A more detailed discussion of the linearisation of Equation (26) can be found in Hills et al. (2018, hereafter H18). In brief, this linearisation is performed by Taylor expanding Equation (26) about the point $b_i = 0$ with $i \in [1, 2, 3]$ and retaining terms up to second order in these parameters. The resulting linearised model is a polynomial in ν with coefficients a_i, which are related to the coefficients of the non-linear Intrinsic model, b_i, and thus the physical parameters of the BFCC model, Equation (25), as follows:

$$a_0 = b_0 = \bar{T}e^{-\tau_\mathrm{ion}(\nu)} , \quad (28)$$
$$a_1 = b_0 b_1 = (2.5 - \beta_0)\bar{T}e^{-\tau_\mathrm{ion}(\nu)} ,$$
$$a_2 = b_0(b_1^2/2 + b_2) = \frac{\bar{T}e^{-\tau_\mathrm{ion}(\nu)}[(2.5 - \beta_0)^2 + \sigma_\beta^2]}{2} ,$$
$$a_3 = -b_0 b_3 = -\tau_0 \bar{T}e^{-\tau_\mathrm{ion}(\nu)} ,$$
$$a_4 = b_4 = T_\mathrm{e}\tau_0 .$$

From Equation (28), it can be seen that for Equation (27) to provide a physical model for the emission components it describes, it is necessary that a_i with $i \in [0, 2, 4]$ are strictly positive and a_3 is strictly negative. These constraints were not imposed when fitting the EDGES-low data in B18. As such, the component of the data fit using Equation (27) was not limited to astrophysical and ionospheric effects. The increased flexibility of Equation (27) in the absence of

physical priors increases the level of correlation between the nominal-foreground and 21-cm components of the model; however, it has the benefit that the LinPhys has some additional flexibility to model systematics such as those expected to arise from imperfect correction of the data for antenna chromaticity (Paper I)[7].

The fact that the maximum likelihood parameters of this nominal-foreground component of the sky model, recovered when jointly fitting it with a flattened Gaussian 21-cm model in B18, do not respect the physicality constraints given above (see e.g. Hills et al. 2018) indicates that these parameters are indeed being used, in part, to model systematic structures in the data in the fits presented in B18. However, given that the LinPhys model is not explicitly tailored to fitting the non-intrinsic sky structure remaining after BFCC, it is unclear, a priori, whether this additional flexibility is sufficient to model such systematics. This will be tested in Section 5, where we will fit realistic simulated EDGES low-band data with Equation (27) using the broad priors listed in Table 2.

4.5 MultLin foreground model

For a number of 21-cm signal recovery tests, in place of Equation (27), B18 use a more general polynomial model (their Equation 2). This model has the form,

$$T_{\mathrm{MultLin}}^{\mathrm{model}}(\nu) = \sum_{n=0}^{N-1} c_n \nu^{n-2.5} . \quad (29)$$

Here, the exponent -2.5 is chosen for the same reason as in the Intrinsic model: to enable more accurate modelling of the dominant synchrotron component of the foreground emission. Additional terms aim to model higher order spectral structure in the foreground emission and can also partially capture some instrumental effects, such as additional spectral structure from chromatic beams or small errors in calibration (B18).

---

[7] Receiver calibration error has also been identified as a possible source of systematic structure (e.g. Bowman et al. 2018b; Murray et al. 2022); however, we leave more detailed investigation of this possibility to future work.
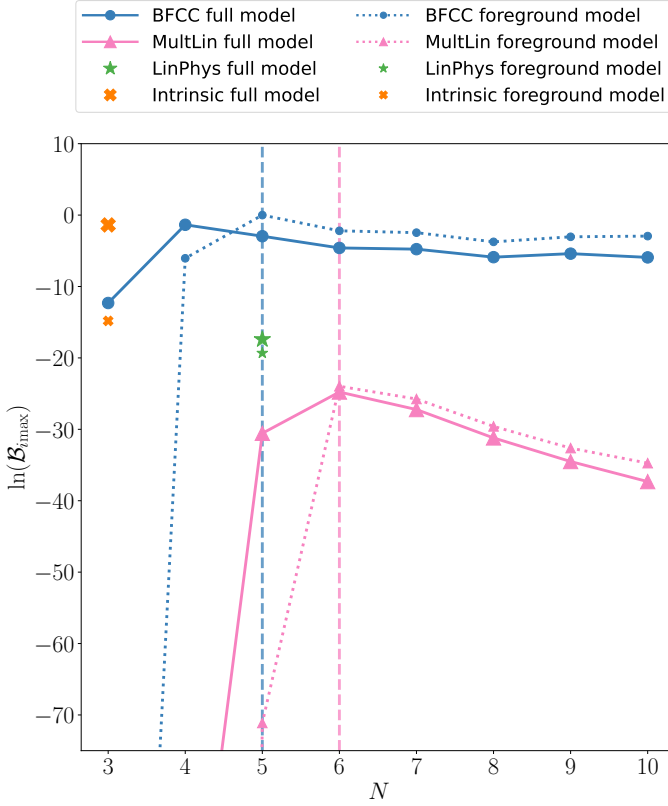
Figure 2. Bayes factors ($\mathcal{B}_{i\max}$) of model $\boldsymbol{M}_i$ relative to the highest-evidence model $\boldsymbol{M}_{\max}$ for the foreground-only validation data set $\boldsymbol{D}_v$. The parameter count N denotes the number of terms in the model component with a priori unknown complexity (see main text). In the flexible-complexity BFCC and MultLin parametrisations, models that include a 21-cm signal are connected by solid lines, and those without a 21-cm signal are connected by dotted lines. In both these and the fixed-complexity Intrinsic and LinPhys models, the presence or absence of a 21-cm signal is also indicated by large and small symbols, respectively (see legend). Models with maximum total evidence in the BFCC and MultLin classes are marked by vertical dashed lines in blue and pink, respectively.

Table 3. BaNTER model validation results. Bayes factors between $\boldsymbol{M}_{ic}$ and $\boldsymbol{M}_{i\mathrm{Fg}}$, as models for the validation data, where i runs over the models defined in Section 4. A positive $\ln(\mathcal{B}^v_{c\mathrm{Fg}})$ indicates that the composite model, $\boldsymbol{M}_{ic}$, is preferred over the foreground model, $\boldsymbol{M}_{i\mathrm{Fg}}$, for the foreground-only validation data, $\boldsymbol{D}_v$, and the model has failed the BaNTER null test. The reverse is true when $\ln(\mathcal{B}^v_{c\mathrm{Fg}})$ is negative.

Model	$\ln(\mathcal{B}^v_{c\mathrm{Fg}})$	Pass/Fail	Comment
BFCC ($N = 3$)	203.3	Fail	Spurious signal detection
BFCC ($N = 4$)	4.7	Fail	Spurious signal detection
BFCC ($N = 5$)	-3.0	Pass	
BFCC ($N = 6$)	-2.4	Pass	
BFCC ($N = 7$)	-2.3	Pass	
BFCC ($N = 8$)	-2.1	Pass	
BFCC ($N = 9$)	-2.4	Pass	
BFCC ($N = 10$)	-3.0	Pass	
MultLin ($N = 3$)	8396.4	Fail	Spurious signal detection
MultLin ($N = 4$)	769.3	Fail	Spurious signal detection
MultLin ($N = 5$)	40.6	Fail	Spurious signal detection
MultLin ($N = 6$)	-0.8	Pass	
MultLin ($N = 7$)	-1.5	Pass	
MultLin ($N = 8$)	-1.6	Pass	
MultLin ($N = 9$)	-1.9	Pass	
MultLin ($N = 10$)	-2.6	Pass	
LinPhys ($N = 5$)	1.9	Fail	Moderate preference for $\boldsymbol{M}_{ic}$ over $\boldsymbol{M}_{i\mathrm{Fg}}$, but below the spurious signal detection threshold
Intrinsic ($N = 3$)	13.4	Fail	Spurious signal detection

5 RESULTS

Table A1 of Appendix A summarises the 21-cm signal detection and parameter inference for the four models – BFCC, Intrinsic, LinPhys, and MultLin – across three simulated 21-cm signal amplitude scenarios (null test, moderate amplitude, and high amplitude). We describe the results in detail below.

5.1 BaNTER validation results

5.1.1 Null test

Figure 2 shows the Bayes factors ($\mathcal{B}_{i\max}$) between model $\boldsymbol{M}_i$ and $\boldsymbol{M}_{\max}$ for the validation data generated as described in Section 3.3. Here, i runs over all models in $\mathcal{M}$, we perform BFBMC over models including (solid lines and/or large symbols) and excluding (dotted lines and/or small symbols) a 21-cm signal component, and the validation data contains simulated observations of foreground emission and noise, corresponding to a scenario where no observable 21-cm signal exists within the observation band of interest. Such a situation

could arise, for instance, if the first stars had not yet formed during the redshift interval corresponding to the frequency range of the data[8].

Following Paper I, we plot the Bayes factor as a function of N, the number of parameters associated with the component of the model whose complexity is a priori unknown. In models that incorporate a foreground component designed to describe a combination of effects that cannot be physically separated (the LinPhys and MultLin models), N corresponds to the complexity of this component. In models where the foreground can be explicitly decomposed into physically motivated ionospheric and astrophysical subcomponents (the Intrinsic and BFCC models), with only the complexity of the astrophysical foreground subcomponent being a priori unknown, N refers to the complexity of the latter subcomponent. $\boldsymbol{M}_{\max}$ is the model with the highest Bayesian evidence for the data, which we find to be the BFCC model with $N = 5$ terms and no 21-cm signal component.

Table 3 lists the corresponding values of $\ln(\mathcal{B}^v_{c\mathrm{Fg}})$, the Bayes factors between $\boldsymbol{M}_{ic}$ and $\boldsymbol{M}_{i\mathrm{Fg}}$ as models for the validation data, $\boldsymbol{D}_v$. For positive $\ln(\mathcal{B}^v_{c\mathrm{Fg}})$, $\boldsymbol{M}_{ic}$ is preferred over $\boldsymbol{M}_{i\mathrm{Fg}}$. Since $\boldsymbol{S}_{\mathrm{Fg}}$ is the only signal component present in $\boldsymbol{D}_v$, a preference for $\boldsymbol{M}_{ic}$ indicates inaccuracy of $\boldsymbol{M}_{i\mathrm{Fg}}$ and any detection of $\boldsymbol{S}_{21}$ in the validation data is necessarily spurious.

5.1.2 Spurious signal detection and bias predictions with unvalidated models

Using the composite model validation criteria defined in Section 2.2.3, we find that the Intrinsic model, LinPhys model, and variants of the BFCC with $N < 5$, as well as variants of the MultLin model with $N < 6$ foreground terms, exhibit evidence in favour of

---

[8] $z = \frac{\nu_{21}}{\nu_{\mathrm{obs}}} - 1$, with $\nu_{21} \simeq 1420.4$ MHz.

including a spurious 21-cm component when fitting the non-21-cm simulated validation data. These models thus fail the null test.

It follows from this result that, if these models were used to analyse an equivalent data set containing a significant 21-cm signal, they would fit a combination of the true 21-cm signal and the systematics that caused them to fail BaNTER validation. As a result, these models would produce biased estimates of the 21-cm signal. The extent of this bias depends on the degree to which the sum of the true 21-cm signal and the systematics are fittable with the 21-cm model. This, in turn, depends on the level of systematics in the data, the amplitude and shape of the true 21-cm signal, and the flexibility of the 21-cm model.

In Table 3, we note that most models failing the BaNTER null test do so with sufficiently large $\ln(\mathcal{B}_{\mathrm{cFg}}^{\mathrm{v}})$ values to yield spurious signal detections in a foreground-only data set. The exception is the LinPhys ($N = 5$) model, which fails the BaNTER null test with only a moderate preference for $\boldsymbol{M}_{i\mathrm{c}}$ over $\boldsymbol{M}_{i\mathrm{Fg}}$, but below the spurious signal detection threshold ($\ln(\mathcal{B}_{\mathrm{cFg}}^{\mathrm{v}}) = 3$) defined in Section 2.2.3. This suggests that we should expect this model to yield biased inferences of the 21-cm signal, but with the level of that bias being lower than for the other models that fail the BaNTER null test.

5.1.3 BaNTER validation results as binary model priors

Based on the BaNTER null test results, we judge the Intrinsic and LinPhys models, as well as variants of the BFCC with $N < 5$ and variants of the MultLin model with $N < 6$ foreground terms as inadequate for reliably recovering unbiased estimates of the 21-cm signal. To confirm this conclusion, we perform two model comparison analyses in Sections 5.2 and 5.3:

(i) *Bayesian model comparison with uninformative model priors:* An unvalidated model comparison analysis is conducted, where the null test is not applied, and all models are treated as equally likely a priori.

(ii) *BaNTER-validated posterior-odds-based Bayesian model comparison:* In this case, models that fail the BaNTER null test are assigned negligible prior odds of yielding unbiased 21-cm signal estimates and are excluded from $\boldsymbol{M}$. We denote the resulting validated subset as $\boldsymbol{M}_{\mathrm{v}}$. Models in $\boldsymbol{M}_{\mathrm{v}}$ are treated as equally likely a priori and are weighted by their Bayesian evidence as models for the observational data a posteriori.

5.2 Moderate amplitude 21-cm signal

5.2.1 Bayesian model comparison with uninformative model priors

In Figure 3 (left), we present the Bayes factors ($\ln(\mathcal{B}_{i\mathrm{max}})$) comparing models $\boldsymbol{M}_i$ to the highest-evidence model, $\boldsymbol{M}_{\mathrm{max}}$, for the data set $\boldsymbol{T}_{\mathrm{corrected}}$. Here, $\boldsymbol{T}_{\mathrm{corrected}}$ corresponds to the moderate-amplitude 21-cm signal scenario with $A = 150$ mK, i runs over all models in the set $\boldsymbol{M}$ (i.e. including both the models that pass and those that fail BaNTER validation), and $\boldsymbol{M}_{\mathrm{max}}$ is the BFCC composite model with $N = 4$.

We find that a subset of models – including the Intrinsic composite model and BFCC composite models with $N = 3$ and $N = 5$–10 – describe the data comparably well to the highest-evidence model ($N = 4$), with no strong Bayesian evidence favouring one over the others. Relative to this subset, the remaining models are decisively disfavoured ($\ln(\mathcal{B}) > 5$ when comparing any model in the high-evidence subset to models outside it; see Table A1).

Among the high-evidence models, the Intrinsic model and BFCC models with $N = 3$ and 4 are the highest-evidence candidates, with

the Intrinsic model and BFCC model with $N = 3$ being only weakly disfavoured relative to $\boldsymbol{M}_{\mathrm{max}}$. All three models strongly support the inclusion of a 21-cm component ($\ln(\mathcal{B}_{\mathrm{cFg}}) > 3.0$).

The remaining models in the preferred subset are moderately disfavoured relative to $\boldsymbol{M}_{\mathrm{max}}$. Two of them provide strong evidence for a 21-cm detection, while the others show no detection.

The LinPhys and MultLin models are decisively disfavoured by the Bayesian evidence. The only model in the LinPhys class and the highest-evidence models in the MultLin class (MultLin $N = 5$ and 6) yield detections of the 21-cm signal. The remaining MultLin models show a mix of detections and non-detections.

In general, weaker detections or non-detections are more probable in higher-complexity foreground models, as their greater flexibility allows them to fit both the foregrounds and a significant fraction of the 21-cm signal simultaneously. Consequently, the difference in Bayesian evidence between models with and without a 21-cm signal decreases, eventually falling below the detection threshold.

5.2.2 BaNTER-validated posterior-odds-based model comparison

In Figure 3 (right), we present the posterior odds ($\ln(\mathcal{R}_{i\mathrm{max}})$; c.f. Section 5.1.3 for our binary prior odds) comparing models $\boldsymbol{M}_i$ to the highest posterior odds model, $\boldsymbol{M}_{\mathrm{max,v}}$, for the moderate amplitude 21-cm signal data set. The highest posterior odds model for this data set is the BFCC composite model with $N = 5$ foreground terms. We treat the prior odds of models that failed the BaNTER null test as negligible, excluding them from the analysis. Consequently, selecting the highest posterior odds model in $\boldsymbol{M}$ corresponds to selecting the highest evidence model in the validated subset, $\boldsymbol{M}_{\mathrm{v}}$.

The validated posterior-odds-based model comparison shows that the BFCC composite models with $N = 6$–10 describe the data comparably well to $\boldsymbol{M}_{\mathrm{max,v}}$, with no strong preference for one model over the others ($\mathcal{R}_{i\mathrm{max}} < 3$). The remaining models in $\boldsymbol{M}_{\mathrm{v}}$, outside this subset, are decisively disfavoured.

Among the highest posterior odds models, the BFCC composite models with $N = 5$ and 6 have the greatest evidence as models for the data. Both yield detections of the 21-cm signal ($\ln(\mathcal{B}_{\mathrm{cFg}}) > 3.0$). The higher complexity BFCC models are weakly to moderately disfavoured relative to $\boldsymbol{M}_{\mathrm{max,v}}$ and do not detect the 21-cm signal.

MultLin is the other model class that passes BaNTER validation; however, these models are decisively disfavoured by BFBMC relative to the BFCC models. None of the validated MultLin models detect the 21-cm signal in the moderate amplitude 21-cm signal data.

When comparing the conclusions drawn from the BaNTER-validated Bayesian model comparison and the Bayesian model comparison with uninformative priors, we observe the following differences in preferred models:

(i) *Unvalidated workflow:* The Intrinsic model and BFCC models with $N = 3$ and 4 are the highest-evidence models. All three of these models provide strong support for the inclusion of a 21-cm component ($\ln(\mathcal{B}_{\mathrm{cFg}}) > 3.0$).

(ii) *BaNTER-validated workflow:* The three highest-evidence models from the unvalidated workflow fail the BaNTER null test and are thus excluded from the validated model set. The BFCC models with $N = 5$ and 6 are the highest posterior odds models for the moderate amplitude 21-cm signal data. Both of these models provide strong support for the detection of the 21-cm signal ($\ln(\mathcal{B}_{\mathrm{cFg}}) > 3.0$).
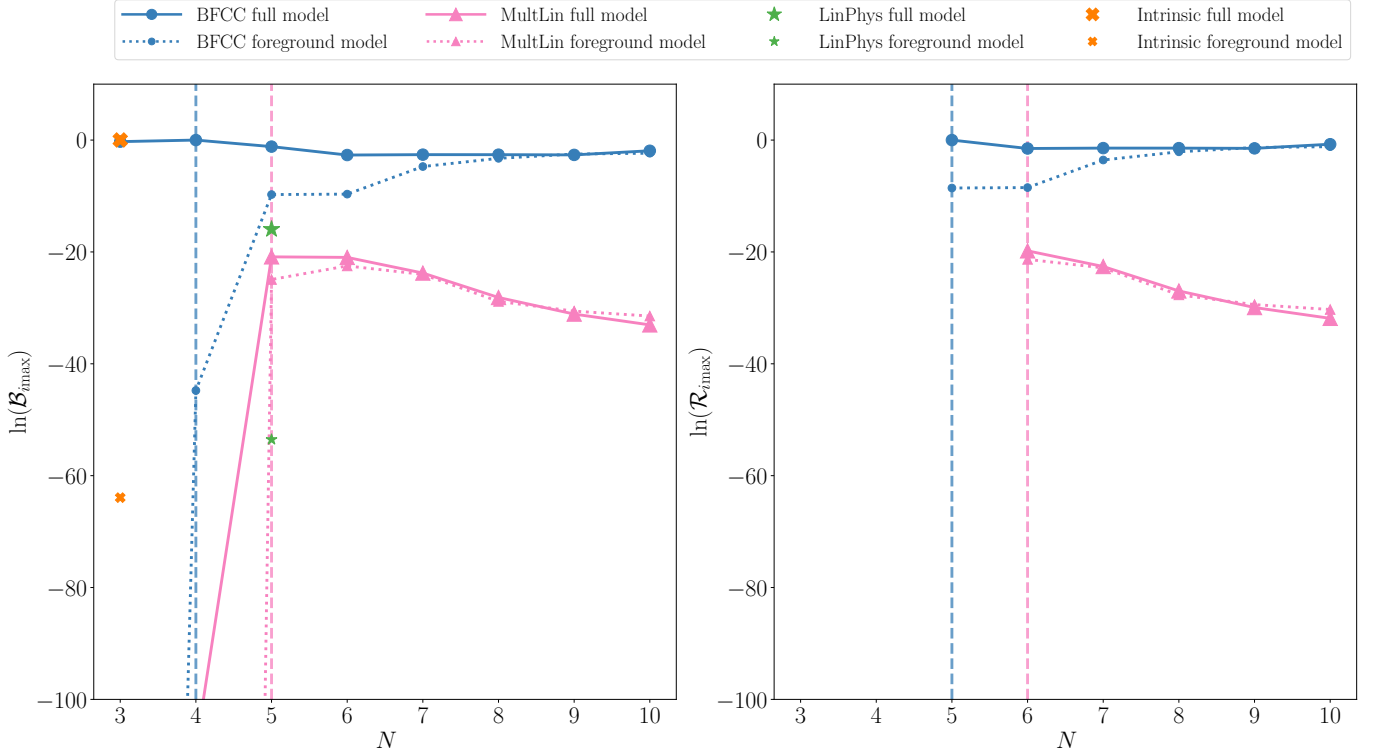
Figure 3. Results of the Bayesian comparison of models for simulated data incorporating a moderate amplitude 21-cm signal ($A = 150$ mK). **Left**: Bayes factors ($\mathcal{B}_{i\max}$) comparing model M_i to the maximum evidence model, $M_{\max}$. Here, i runs over all models in the set $\mathcal{M}$, which includes both the models that pass and those that fail BaNTER validation). **Right**: Posterior odds ($\mathcal{R}_{i\max}$) of model M_i over the validated model $M_{\max,v}$. Here, $M_{\max}$ and $M_{\max,v}$ represent the models with the highest Bayesian evidence and posterior odds, respectively. Symbols and solid and dashed lines have the same meanings as in Figure 2. The subset of models that are present in the left panel but are absent in the right represent the set of models that failed the BaNTER null test. The number of foreground terms with the highest evidence (left) and highest posterior odds (right) for these models are indicated by blue and pink vertical, dashed lines, respectively.

5.2.3 21-cm signal estimates

Having identified the preferred model using Bayes factors and posterior odds, we now examine whether these preferences align with the ground truth, as determined by the consistency of the recovered parameter posteriors with the true input parameters of the 21-cm signal in the data. Figure 4 illustrates the fit results for all composite models in $\mathcal{M}$ that exhibit strong evidence for a 21-cm signal detection. For each model, the figure shows the posterior PDs of: (i) the residuals obtained by fitting the data using only the foreground component of the model, (ii) the residuals obtained by fitting the data using the full model, and (iii) the recovered 21-cm signal derived from fitting the data with the full model.

The first seven subplots (Figures 4a to 4d and 4g to 4i) show results for models that failed the BaNTER validation. While the majority of these models achieve reasonable fits to the data (evidenced by the relative consistency between the full model residuals and the expected noise level in the data; see Figures 4c to 4d and 4g to 4i, middle panels), they nevertheless yield substantial (Figures 4d, 4h and 4i) to very substantial (Figures 4a to 4c and 4g) biases in their recovered 21-cm signals (amplitudes, location or shape parameters inconsistent with the underlying parameters of the 21-cm signal in the data at 95% credibility; see Table A1).

Additionally, models that most severely failed the BaNTER null test – the MultLin model with $N = 3$ and 4, and the BFCC model with $N = 3$ (see Table 3) – also exhibit the most biased 21-cm signal recovery. Furthermore, the MultLin model with $N = 3$ and 4 provides poor

fits to the data, even with biased 21-cm signal modelling absorbing some structure due to foreground systematics.

In contrast, barring biased recovery of the flatness parameter, τ, the 21-cm signal recovered with the LinPhys model, which failed the BaNTER null test by the smallest margin, shows 21-cm parameter estimates that are consistent with the underlying 21-cm signal in the data at 95% credibility.

By comparison, models that passed the BaNTER validation and detected the 21-cm signal yield fully unbiased recovery of the 21-cm signal in the data (Figures 4e and 4f and Table A1), with all 95% credibility HPDI parameter estimates consistent with the true signal parameters (Table A1).

Comparing the consistency of the recovered 21-cm signals (represented by the contours in the bottom panels of each subfigure) with the true 21-cm signal data (indicated by the dashed black lines) in Figure 4 yields the ground truth efficacy of our models (also see Table A1). By comparing this true efficacy to the expected efficacy based on the BaNTER-validated Bayesian model comparison and the Bayesian model comparison with uninformative priors, we draw the following conclusions regarding these two model comparison methodologies:

(i) *Unvalidated workflow:* The three highest-evidence models (the Intrinsic model and BFCC models with $N = 3$ and 4), as determined by the BFBMC analysis, yield biased estimates of the underlying 21-cm signal in the data. This bias demonstrates that comparison of the full set of models considered in the analysis, as judged by BFBMC, is insufficient to identify models that yield unbiased recovery of the
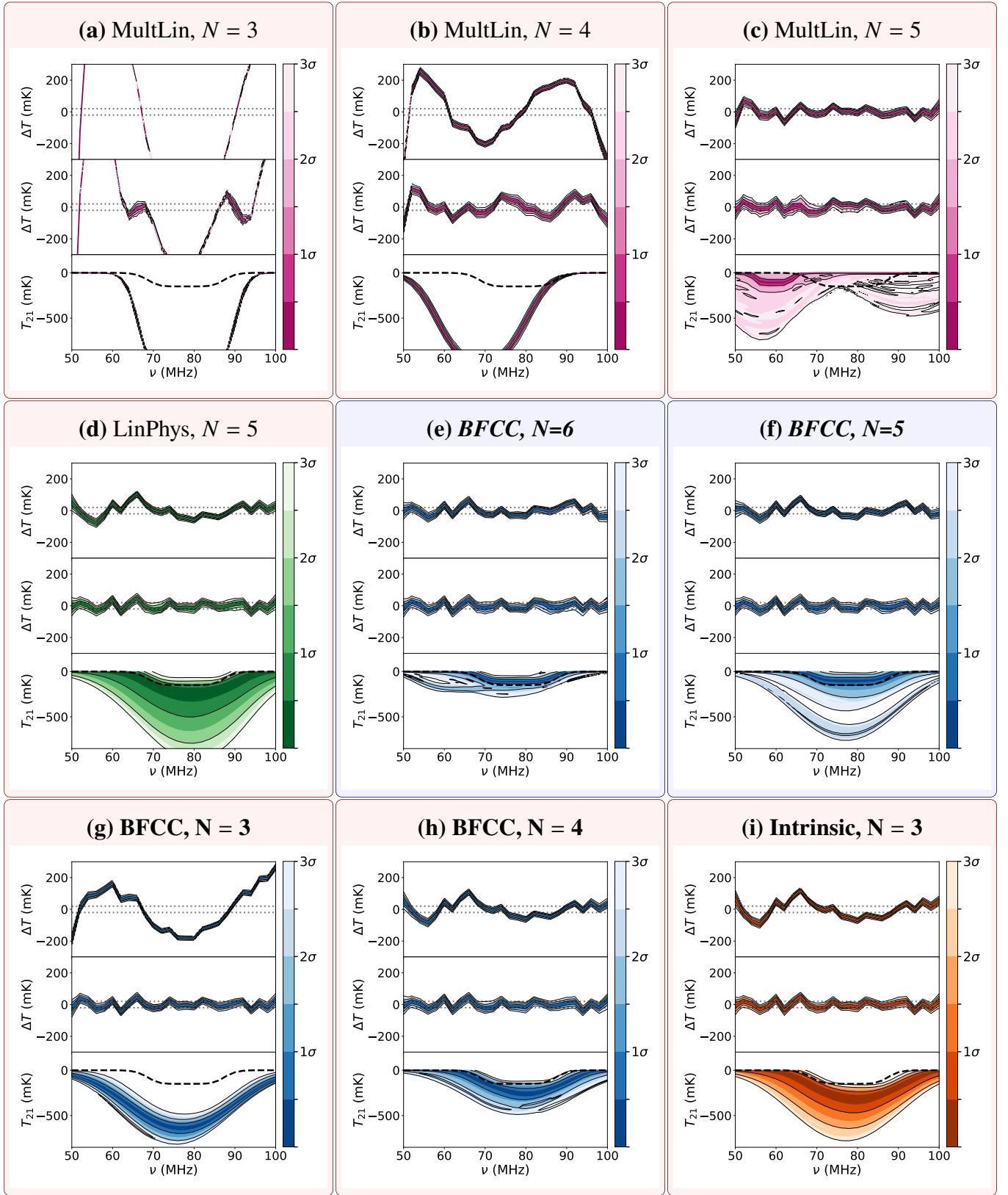
Figure 4. Signal recovery plots for models detecting a 21-cm signal in simulated data containing a moderate amplitude signal ($A = 150$ mK). Each subplot shows posterior probability densities of foreground-only residuals (top), full-model residuals (middle), and recovered 21-cm signal (bottom). Dotted lines in the top and middle panels indicate the noise level; the dashed black line shows the true input signal. Models are arranged by increasing Bayesian evidence (lowest evidence in top-left, highest in bottom-right). Background colours distinguish BaNTER validation results: red backgrounds indicate failed validation and blue backgrounds indicate passed validation. We highlight models using **bold** for highest evidence models (those with $\ln(\mathcal{B}_{i\max}) \geq -3$) and *italic* for highest BaNTER-validated posterior odds models (those with $\ln(\mathcal{R}_{i\max}) \geq -3$). Models meeting both criteria appear in ***bold italic***. Subfigure captions indicate the model type and foreground complexity (N).

21-cm signal. This conclusion is particularly driven by the inclusion of the Intrinsic model and the BFCC models with $N = 3$ or 4 in the set of models under consideration (see Appendix B1 for details). This result implies that the comparison of models in the unvalidated model set for the moderate amplitude 21-cm signal data constitutes a *category II* model comparison problem for which the conclusions drawn from BFBMC alone are not robust.

(ii) *BaNTER-validated workflow:* All models shown to yield biased estimates of the 21-cm signal in Figure 4 were correctly identified and excluded from the validated model set by the BaNTER null test. The highest posterior odds models with 21-cm signal detections in the BaNTER-validated posterior-odds-based analysis are correctly identified as the BFCC models with $N = 5$ and 6. These models are found to be the only ones that yield unbiased estimates of the 21-cm signal in the data in Figure 4.

The excellent agreement between the expected performance of the models, based on the results of BaNTER validation in Section 5.1, and the validity of the recovered 21-cm signal estimates demonstrates the necessity and efficacy of the BaNTER-validated posterior-odds-based analysis in the context of the moderate amplitude 21-cm signal scenario.

5.3 High amplitude 21-cm signal

5.3.1 Bayesian model comparison with uninformative model priors

In Figure 5 (left), we show the Bayes factors for the high-amplitude 21-cm signal scenario ($A = 500$ mK), comparing each model M_i to the highest-evidence model, M_{max}, for $T_{corrected}$. As in the unvalidated moderate-amplitude 21-cm signal analysis, i runs over all models in the set $\mathcal{M}$. Here, we find that M_{max} is the BFCC composite model with $N = 5$.

In contrast to the moderate-amplitude 21-cm signal scenario, all composite models in $\mathcal{M}$ show strong support for the inclusion of a 21-cm signal component in the high-amplitude case ($\ln(\mathcal{B}_{cFg}) \geq 3.0$; see Table A1). Additionally, the subset of highest-evidence models for the high-amplitude data set is smaller. Specifically, similar to the moderate-amplitude regime, the BFCC composite models with $N = 4$ and 5 remain comparably probable, with no strong Bayesian evidence favouring one over the other ($\ln(\mathcal{B}_{imax}) < 3$). However, several models that were in the highest-evidence subset for the moderate-amplitude case are now absent. These include the Intrinsic composite model and the BFCC composite models with $N = 3$ and 6 to 10.

The decrease in $\ln(\mathcal{B}_{imax})$ for the BFCC composite models with $N = 6$ to 10 can be attributed to the Occam penalty associated with their increased complexity (see Section 2.1.2), combined with the minimal improvement in their ability to fit the high-amplitude data relative to M_{max} (see Figures 6k, 6m to 6p and 6r, middle panels).

In contrast, the decrease in $\ln(\mathcal{B}_{imax})$ of the BFCC composite model with $N = 3$ and of the Intrinsic model is driven by their poorer performance in fitting the high-amplitude 21-cm signal data set. The most significant difference in Bayes factor between the two data sets is observed for the BFCC composite model with $N = 3$, which transitions from being in the highest-evidence subset for the moderate-amplitude 21-cm signal data set to being decisively disfavoured for the high-amplitude 21-cm signal data. Specifically, we find that for the high-amplitude signal data, $\ln(\mathcal{B}_{imax})$ decreases by ~ 17 (see Table A1), corresponding to odds in favour this model relative to M_{max} of worse than $1 : 10^7$.

To understand this difference, recall that the BFCC composite model with $N = 3$ failed the BaNTER null test in Section 5.1. This

failure indicates that when this model is used to fit a data set containing a non-zero 21-cm signal, the 21-cm component necessarily fits the combined contribution of both the 21-cm signal and the foreground systematics. In the moderate-amplitude regime, while the foreground model imperfectly describes the foregrounds, it also captures a significant fraction of the 21-cm signal. This behaviour allows a biased fit of the 21-cm model component to absorb residual systematics, leading to a relatively better fit to the data in aggregate. In contrast, in the high-amplitude regime, a smaller fraction of the 21-cm signal is described by the foreground model. As a result, the parameters of the 21-cm model are more strongly constrained by the remaining 21-cm component, reducing its flexibility to absorb residual foreground systematics.

It follows from this explanation that one should expect the reduction in Bayes factor to be accompanied by a decrease in the accuracy of the BFCC composite model with $N = 3$ in the high-amplitude 21-cm signal regime relative to the moderate-amplitude case. To quantify this decrease, we use the model accuracy statistic introduced in S25 (see Appendix C). Indeed, we find that this expectation holds: specifically, in the moderate-amplitude regime, we obtain $Q_{0.999}(\lambda) \simeq 2$, which satisfies the S25 accuracy condition ($Q_{0.999}(\lambda) > 0$). However, in the high-amplitude case, this value drops to $Q_{0.999}(\lambda) \simeq -17$, implying that the fit residuals are inconsistent with the expected noise distribution in the data and are thus contaminated by residual systematics.

The same reasoning applies to the Intrinsic model, which also failed the BaNTER null test in Section 5.1. Here, we observe a similar but less pronounced decrease in model accuracy when transitioning from the moderate- to high-amplitude 21-cm signal data sets.

5.3.2 BaNTER-validated posterior-odds-based model comparison

In Figure 5 (right), we show the posterior odds ($\ln(\mathcal{R}_{imax})$), comparing models M_i to the highest posterior odds model, $M_{max,v}$, for the high-amplitude 21-cm signal data set. The highest posterior odds model in this case is the BFCC composite model with $N = 5$ foreground terms, which coincides with the highest Bayesian evidence model identified in Section 5.3.1. As in the moderate-amplitude regime, models that failed the BaNTER null test are assigned negligible prior odds and thus excluded from the analysis. The validated subset, $\mathcal{M}_v$, contains only the remaining models.

The remaining BFCC composite models in $\mathcal{M}_v$ are strongly disfavoured, while the remaining MultLin models are decisively disfavoured. Within the MultLin class, the $N = 6$ and 7 foreground complexity composite models have the highest posterior odds, with the remaining MultLin models strongly to decisively disfavoured relative to these two models.

Comparing the conclusions drawn from the BaNTER-validated posterior odds and the Bayesian model comparison with uninformative model priors, in the high amplitude 21-cm signal regime one finds the following similarities and differences in preferred models:

(i) *Unvalidated workflow:* The BFCC composite models with $N = 4$ and 5 are the highest-evidence models.

(ii) *BaNTER-validated workflow:* BFCC composite models with $N = 4$ fails the BaNTER null test and thus is excluded from the validated model set. The set of the highest posterior odds models for the high amplitude 21-cm signal data contains only the BFCC models with $N = 5$.
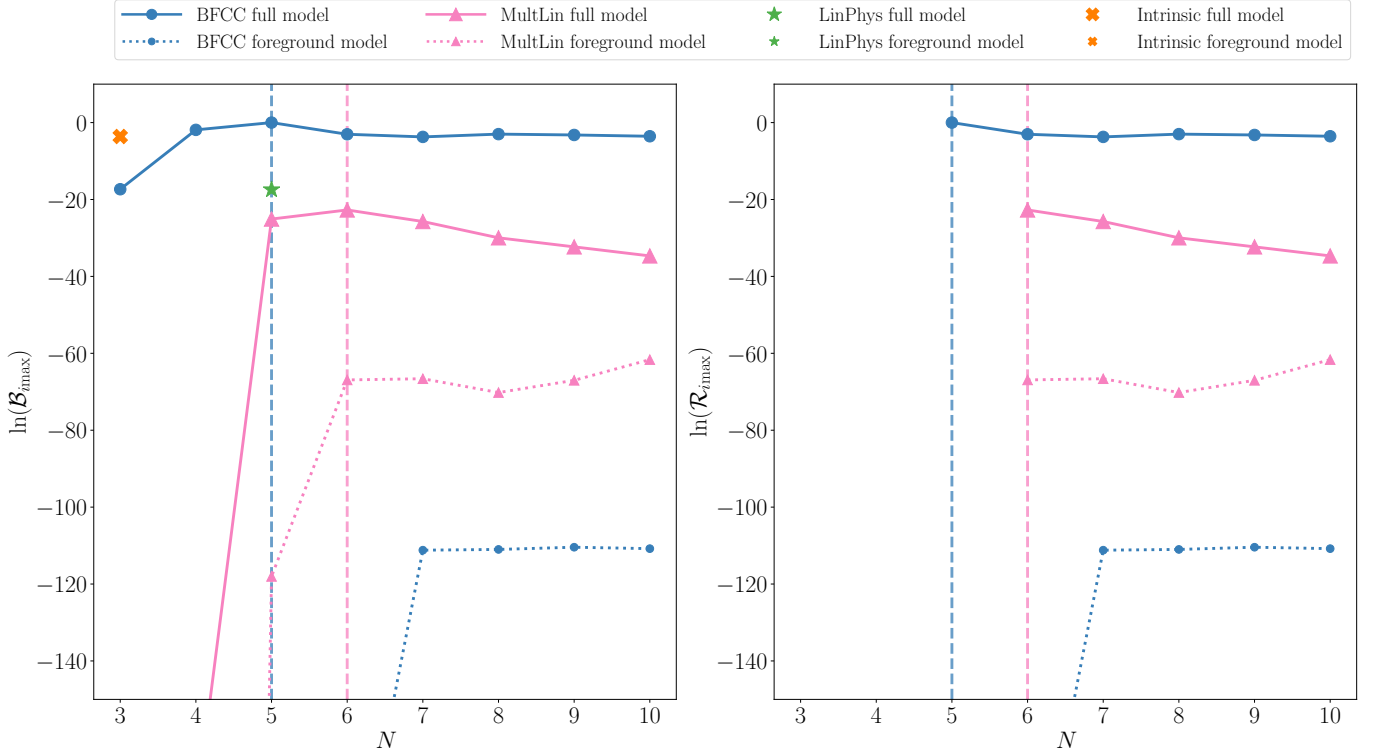
Figure 5. As in Figure 3 but for the Bayesian comparison of models for simulated data incorporating a high amplitude 21-cm signal model ($A = 500$ mK).

5.3.3 21-cm signal estimates

Having established the preferred models for the data set containing the high-amplitude 21-cm signal using Bayes factors and posterior odds, we now examine whether the recovered parameter posteriors align with these conclusions.

Figure 6 illustrates the fit results for all composite models in $\mathcal{M}$ that exhibit strong evidence for a 21-cm signal detection. For the high-amplitude 21-cm signal data set, $\mathcal{M}$ includes all models considered. For each model, the figure shows the posterior PDs of the foreground (top panels) and of the composite model (middle panels) fit residuals, and the 21-cm signal component of the composite model fit (bottom panels).

Figures 6a, 6b, 6g, 6i, 6j, 6l and 6q show results for models that failed BaNTER validation. As in the moderate-amplitude 21-cm signal scenario, we find that while all but two of the failed models achieve good fits to the full data – evidenced by the consistency between full model residuals and the expected noise level in Figures 6g, 6j, 6l and 6q – they all result in biased recovery of the amplitude, location or shape parameters of the underlying 21-cm signal at 95% credibility (see Table A1 for details).

Additionally, mirroring our findings in the moderate-amplitude scenario, the models that most severely failed the BaNTER null test (MultLin models with $N = 3$ and 4 and the BFCC model with $N = 3$; c.f. Table 3) also yield the most biased 21-cm signal recovery. Conversely, the LinPhys model, which failed the BaNTER null test by the smallest margin, also yields the least biased 21-cm signal parameter estimates of the failed models in the high-amplitude 21-cm signal scenario, with only the central frequency of the recovered signal being inconsistent with the underlying 21-cm signal in the data at 95% credibility.

By comparison, all models that passed BaNTER validation yield

unbiased recovery of the 21-cm signal in the data (see Figures 6c to 6f, 6h, 6k, 6m to 6p and 6r and Table A1), with all 95% credibility HPDI parameter estimates consistent with the true signal parameters (Table A1).

As in Section 5.3.3, by comparing the consistency of the recovered 21-cm signals in Figure 6 with the true 21-cm signal data, one can derive the ground truth efficacy of our models for the high amplitude data set (also see Table A1). By comparing this true efficacy to the expected efficacy based on the BaNTER-validated Bayesian model comparison and the Bayesian model comparison with uninformative priors, we draw the following conclusions regarding these two model comparison methodologies in this regime:

(i) *Unvalidated workflow:* The subset of highest-evidence models ($\ln(\mathcal{B}_{i\max}) < 3$) for the high-amplitude 21-cm signal data set contains the BFCC composite models with $N = 4$ and 5. Within this subset, only the BFCC composite model with $N = 5$ yields unbiased estimates of the underlying 21-cm signal, while the BFCC composite model with $N = 4$ results in biased recovery. Thus, Bayesian model comparison with uninformative priors (BFBMC alone) remains insufficient to uniquely identify models that yield unbiased recovery of the 21-cm signal. This result implies that, as in the moderate-amplitude scenario, model comparison in the unvalidated workflow corresponds to a *category II* model comparison problem (see Appendix B2). However, the degree of bias in the recovered 21-cm signal estimates is reduced relative to the moderate-amplitude scenario: whereas three-quarters of models yielded biased recovery in the moderate case, this fraction decreases to one-half in the high-amplitude regime.

(ii) *BaNTER-validated workflow:* All models that yielded biased estimates of the 21-cm signal in Figure 6 were correctly identified a priori and excluded from the validated model set by the BaNTER

null test. The highest posterior odds model in the BaNTER-validated posterior-odds-based analysis is the BFCC composite model with $N = 5$, which yields unbiased estimates of the 21-cm signal. Additionally, all remaining models that passed BaNTER validation are found to yield unbiased recovery of the 21-cm signal, albeit generally with lower precision.

The excellent agreement between the expected performance of models, based on the null-test-based Bayesian validation analysis in Section 5.1, and the validity of the recovered 21-cm signal estimates further demonstrates the value of BaNTER-validated posterior-odds-based analysis, even in the high-amplitude 21-cm signal scenario. However, the more moderate differences in preferred models between the BaNTER-validated and unvalidated BFBMC workflows indicate that, while the high-amplitude data modelling problem remains a *category II* model comparison problem, BFBMC provides more reliable inferences in this regime than in the moderate-amplitude scenario.

6 DISCUSSION

6.1 Model efficacy

The primary goal of global 21-cm signal experiments is to obtain unbiased inferences about the redshifted 21-cm signal in the data. This can be subdivided into two distinct but related sub-goals: (i) the recovery of unbiased estimates of the 21-cm signal when it is present in the data, and (ii) the avoidance of spuriously detecting a 21-cm signal if none is present.

Our results in Section 5 demonstrate broad agreement in the models that best achieve sub-goal (i) in the moderate- and high-amplitude 21-cm signal regimes. Specifically, we find that signal detections using the BFCC composite models with $N \geq 5$ and the MultLin composite models with $N \geq 6$ yield unbiased parameter inferences, with 95% credibility HPDIs consistent with the parameters of the underlying 21-cm signals in the data. In contrast, the Intrinsic and LinPhys composite models, as well as lower-complexity BFCC and MultLin models, produce biased parameter inferences.

A similar distinction holds for models that best achieve sub-goal (ii) in our null-21-cm signal validation dataset. The only difference is that while the LinPhys composite model is preferred over the foreground model for foreground-only validation data, it is not sufficiently preferred for us to consider it a spurious detection of the 21-cm signal in the data ($\ln(\mathcal{B}_{cFg}) \leq 3$). As such, it satisfies sub-goal (ii) despite failing sub-goal (i).

In a lower-noise observation than considered here, the LinPhys composite model may yield spurious 21-cm signal detection in the null-amplitude regime, failing both sub-goals (i) and (ii). Nevertheless, the fact that LinPhys only fails sub-goal (i) here indicates that the LinPhys foreground model provides only a mildly insufficient description at the 20 mK RMS noise level considered in this work. Therefore, it should be expected to cause only moderate bias in 21-cm signal recovery with that model. This expectation was confirmed by the results in Sections 5.2 and 5.3.

Comparing the BFCC composite models with $N \geq 5$ to the MultLin composite models with $N \geq 6$, we find that while both satisfy sub-goals (i) and (ii), the BFCC models provide more precise parameter inferences. This result suggests the following hierarchy of model efficacy, from most to least effective:

(i) *BFCC composite models with $N \geq 5$:* These models avoid spurious detection of a 21-cm signal in the null-amplitude regime and yield unbiased and relatively precise estimates of the 21-cm signal in both the moderate- and high-amplitude regimes. These are the only models considered in this work that meet both the model accuracy and constraining power requirements for unbiased recovery of the 21-cm signal in the moderate-amplitude regime.

(ii) *MultLin composite models with $N \geq 6$:* These models avoid spurious detection of a 21-cm signal when none is present. In the high-amplitude regime, they have sufficient constraining power to recover unbiased 21-cm signal estimates. However, they yield unbiased yet less precise estimates compared to the BFCC composite models. Nevertheless, they remain viable for global 21-cm cosmology and serve as a consistency check for the more stringent results obtained with BFCC composite models.

(iii) *The LinPhys composite model:* This model avoids spurious detection of a 21-cm signal when none is present, but yields biased estimates of the 21-cm signal in the moderate- and high-amplitude 21-cm signal regimes considered here. As such, this model is not recommended.

(iv) *The Intrinsic composite model, BFCC models with $N \leq 4$, and MultLin models with $N \leq 5$:* These models result in the spurious detection of a 21-cm signal when none is present and produce biased estimates of the 21-cm signal when one is present. They are therefore unsuitable for global 21-cm signal analysis.

6.2 Validated Bayesian model comparison

Our comparison in Section 5 of the ground truth results to the conclusions drawn from BaNTER-validated Bayesian model comparison and Bayesian model comparison with uninformative model priors demonstrates the necessity of the former and insufficiency of the latter for deriving reliable inferences regarding the 21-cm signal in the null-, moderate-, and high-amplitude regimes.

Specifically, we find that Bayesian model comparison with uninformative model priors fails to uniquely identify the models that yield unbiased estimates of the 21-cm signal in the data across all amplitude regimes. In the null-amplitude regime, two-thirds of the highest-evidence models yield spurious detections of the 21-cm signal. In the moderate-amplitude regime, three-quarters of the highest-evidence models produce biased 21-cm signal recovery. This fraction improves in the high-amplitude regime, dropping to one-half of the highest-evidence models yielding biased recovery.

In contrast, the BaNTER null test effectively identifies and eliminates models that produce poor foreground model fits and/or spurious detections in the null-amplitude regime and biased estimates in both moderate- and high-amplitude regimes. Furthermore, the BaNTER-validated Bayesian model comparison framework assigns the highest posterior odds to the BFCC composite models that yield both accurate and precise estimates of the underlying 21-cm signal in the data. Finally, in both signal amplitude regimes, the remaining validated models used for 21-cm signal detection are also found to yield unbiased recovery of the underlying signal parameters, albeit generally with reduced precision compared to the highest posterior odds models.

6.2.1 Accounting for imperfectly simulated data

The results of the BaNTER-validated Bayesian model comparison framework applied to the simulated EDGES low-band datasets considered here are highly promising. However, their effectiveness when applied to observational data depends on the accuracy of the simulated data used in model validation. If systematic effects present in real observations are absent from simulations and have amplitudes
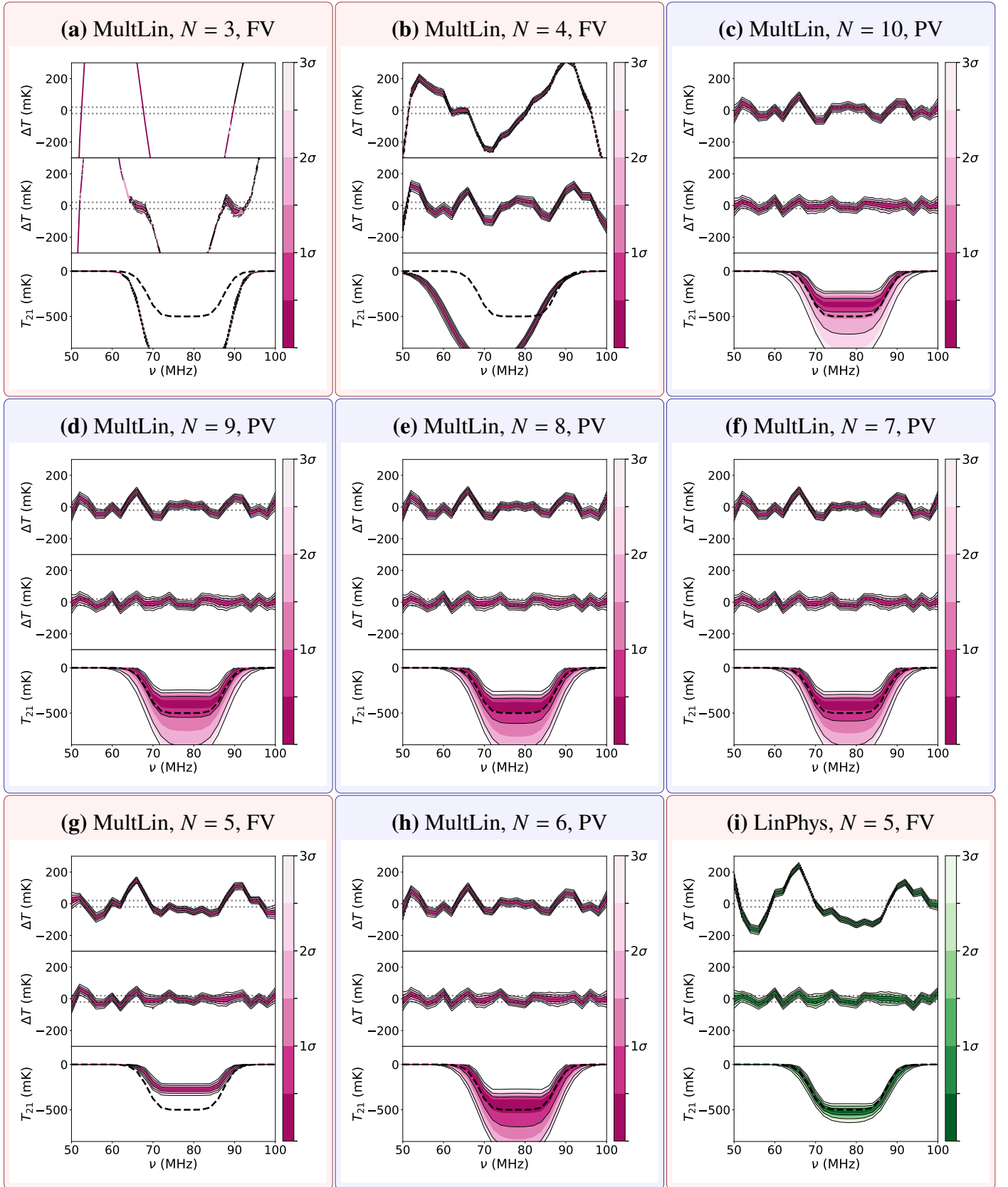
Figure 6. As in Figure 4 but for simulated data, $T_{\text{corrected}}$, containing a high amplitude 21-cm signal ($A = 500$ mK).
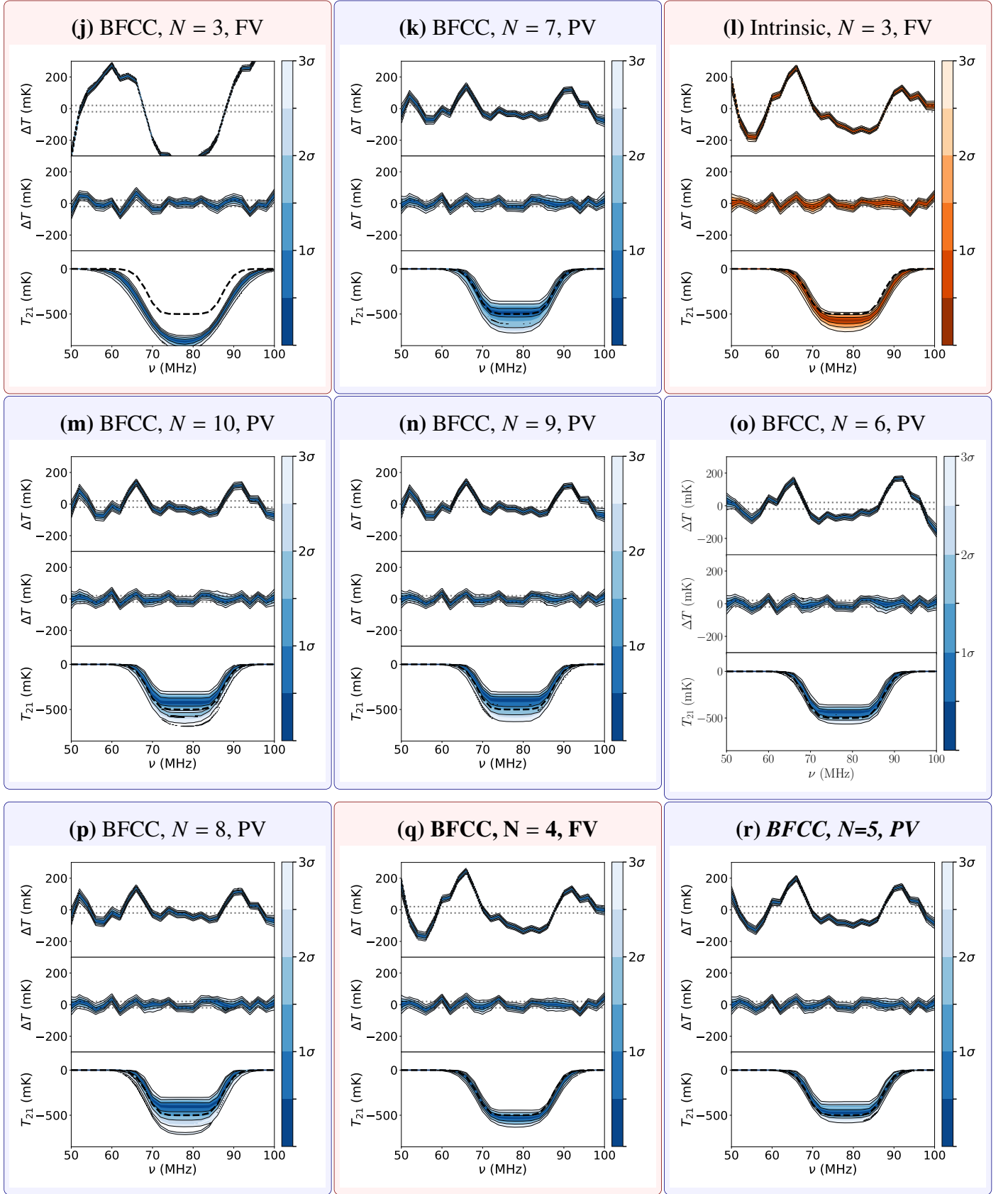
Figure 6 – *continued*

significant relative to the noise level, certain models that should fail BaNTER validation in an ideal case may instead pass (see Appendix D for a discussion of possible sources). Such models could then lead to spurious inferences in subsequent data analysis if they also provide an accurate description of the observational data while containing biased component models.

In such a scenario, by eliminating a subset of models that, while accurately describing the observational data, contain biased components, BaNTER validation still enhances the robustness of conclusions compared to those drawn from BFBMC alone. However, ultimately, the goal of validation is to ensure that the simulation is sufficiently accurate to identify *all* models incapable of unbiased 21-cm signal recovery. To strengthen confidence in this criterion, BaNTER *model validation* can be supplemented with additional *simulation validation* tests to evaluate the accuracy of the validation dataset. S25 propose assessing consistency between the modelling complexity required to describe simulated observations and that needed for observational data as a potential approach. We plan to explore this approach further in future work.

6.2.2 *Extending BaNTER validation to other 21-cm cosmology models*

In this paper, we have demonstrated the value of BaNTER validation in facilitating robust identification of models that can both accurately describe realistic simulated BFCC EDGES-low spectrometer data and recover unbiased estimates of the 21-cm signal from them. However, we anticipate the framework to be similarly useful for comparing alternative composite models across other datasets.

Analysing time-averaged spectrometer data with alternative models provides an additional potential use case in global 21-cm cosmology. For example, Anstey et al. (2021) present a forward modelling analysis of spectrometer data, showing that the model enables reliable 21-cm signal detection in data from a relatively smooth conical log spiral antenna but not in data from a more chromatic conical sinuous antenna. By identifying models prone to biased recovery a priori, incorporating BaNTER validation into such analyses could yield benefits similar to those demonstrated for the BFCC and MultLin models in this work.

Fitting time-dependent or multi-instrument data enables one to leverage both angular and spectral information to distinguish between the foregrounds and the 21-cm signal (e.g. Liu et al. 2013; Tauscher et al. 2020b,a; Hibbard et al. 2023; Saxena et al. 2023; Anstey et al. 2023), reducing the correlation between anisotropic foregrounds and isotropic 21-cm model components. However, accurately describing the data typically necessitates increased model complexity. Additionally, for a single instrument, the extent of correlation reduction between the foreground and 21-cm signal model components depends on the LST range and time-binning of the data[9].

In cases where time-dependent data modelling weakens but does not entirely eliminate the correlation between the 21-cm signal and other model components, BaNTER validation can be anticipated to similarly improve model selection, thereby facilitating unbiased 21-cm signal recovery.

Finally, BaNTER validation is not limited to composite spectrometer models. For example, the methodology demonstrated here could also be applied to validating Bayesian forward modelling approaches for interferometric 21-cm datasets, which incorporate foregrounds, 21-cm signal fluctuations about the mean, and potential systematics (e.g. Furlanetto et al. 2006). Potential applications include validating whether forward models possess sufficient accuracy for unbiased signal recovery amid foregrounds and instrumental systematics, particularly in interferometric calibration (e.g. Byrne et al. 2021; Sims et al. 2022a,b), instrumental modelling (e.g. Wilensky et al. 2024), and 21-cm signal recovery (e.g. Sims et al. 2016, 2019; Sims & Pober 2019; Burba et al. 2023, 2024). Additionally, joint analyses combining spectrometer, interferometric, and other data types could further benefit from BaNTER validation, enhancing model selection and ensuring unbiased 21-cm signal recovery across diverse observational scenarios.

7 SUMMARY & CONCLUSIONS

In Paper I of this series, we derived the physically motivated flexible-complexity BFCC model for spectrometer data post-processed to suppress instrumentally induced spectral structure using beam-factor-based chromaticity correction (e.g. B18). We demonstrated that the BFCC model, with complexity calibrated using BFBMC, enables unbiased recovery of a flattened Gaussian 21-cm signal consistent with the one reported by B18 from simulated data.

In this work, we applied the BFCC model to the analysis of realistic simulations of chromaticity-corrected EDGES-low spectrometer datasets, considering a broader range of scenarios regarding the 21-cm signal in the data. We analysed data containing 21-cm signals in three amplitude regimes: null ($A = 0$ mK), moderate ($A = 150$ mK), and high ($A = 500$ mK). Additionally, we extended the Bayesian comparison of the BFCC model to three competing model classes previously considered in the literature: the Intrinsic model used in Hills et al. (2018), as well as the LinPhys model and an extended set of MultLin models applied to 21-cm signal estimation from EDGES data in B18.

By comparing 21-cm parameter posteriors recovered with competing models to the true 21-cm signal parameters in the data, we identify a broad agreement in models that enable unbiased parameter inferences. Our analysis reveals that only BFCC composite models with $N \geq 5$ and MultLin composite models with $N \geq 6$ avoid spurious detections and yield unbiased 21-cm signal estimates, with BFCC models providing superior precision. The complete model efficacy hierarchy is presented in Section 6.1.

Additionally, we investigated the extent to which Bayesian model comparison can identify the models that yield unbiased 21-cm signal estimates in the data. To address challenges arising from systematics that bias the 21-cm signal model fit while still maintaining an accurate fit to the data in aggregate, we employed the BaNTER validation framework introduced in S25. This framework uses a Bayesian null test to identify composite models that are likely to yield biased 21-cm signal estimates. We used BaNTER validation results to derive model priors and conduct a posterior-odds-based Bayesian comparison of the models.

By comparing models that enable unbiased inferences of the underlying 21-cm signal to conclusions drawn from BaNTER-validated posterior-odds-based model comparison and BFBMC alone, we found that the latter fails to reliably identify models yielding unbiased estimates across all amplitude regimes. Using BFBMC alone, we found that 2/3, 3/4, and 1/2 of the highest-evidence models led to spurious 21-cm signal detections or biased estimates in the null, moderate, and high amplitude regimes, respectively.

---

[9] Correlation between the foreground and 21-cm signal model components is more likely when jointly modelling data over a shorter LST interval or using coarser time-binning.

In contrast, BaNTER validation successfully identified and eliminated models that yield spurious detections in the null-amplitude regime and biased estimates in the moderate- and high-amplitude regimes. Furthermore, the BaNTER-validated posterior-odds-based model comparison framework assigns the highest posterior odds to BFCC composite models that simultaneously provide accurate and precise estimates of the underlying 21-cm signal in the simulated data. Finally, in both signal amplitude regimes, the remaining validated models used to detect the 21-cm signal also yield unbiased recovery of the underlying signal parameters (though generally with reduced precision compared to the highest posterior odds models).

We conclude that the BFCC model holds excellent promise for unbiased inference of the global 21-cm signal from spectrometer data, and we plan to test it on EDGES observations in future work. Moreover, Bayesian validation and model comparison methods, such as those discussed here, provide a powerful framework for identifying optimal models for global 21-cm data sets, ensuring robust signal recovery, and, ultimately, enabling detailed astrophysical insights into the radiative background and structure formation at Cosmic Dawn.

DATA AVAILABILITY

The data from this study will be shared on reasonable request to the corresponding author. Software used in this work to generate simulated data and beam-factors, given an electromagnetic simulation of the beam, is publicly available at `https://github.com/edges-collab`.

References

Agrawal Y., Kavitha K., Singh S., 2024, ApJ, 974, 137
Anstey D., de Lera Acedo E., Handley W., 2021, MNRAS, 506, 2041
Anstey D., de Lera Acedo E., Handley W., 2023, MNRAS, 520, 850
Barkana R., 2018, Nature, 555, 71
Bassett N., Rapetti D., Tauscher K., Nhan B. D., Bordenave D. D., Hibbard J. J., Burns J. O., 2021, ApJ, 923, 33
Bevins H. T. J., Handley W. J., Fialkov A., de Lera Acedo E., Greenhill L. J., Price D. C., 2021, MNRAS, 502, 4405
Bevins H. T. J., Heimersheim S., Abril-Cabezas I., Fialkov A., de Lera Acedo E., Handley W., Singh S., Barkana R., 2024, MNRAS, 527, 813
Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018a, Nature, 555, 67
Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018b, Nature, 564, E35
Bradley R. F., Tauscher K., Rapetti D., Burns J. O., 2019, ApJ, 874, 153
Burba J., Sims P. H., Pober J. C., 2023, MNRAS, 520, 4443
Burba J., Bull P., Wilensky M. J., Kennedy F., Garsden H., Glasscock K. A., 2024, MNRAS, 535, 793
Byrne R., Morales M. F., Hazelton B. J., Wilensky M., 2021, MNRAS, 503, 2457
Cang J., Mesinger A., Murray S. G., Breitman D., Qin Y., Trotta R., 2024, arXiv e-prints, p. arXiv:2411.08134
Datta A., Bradley R., Burns J. O., Harker G., Komjathy A., Lazio T. J. W., 2016, ApJ, 831, 6
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Gessey-Jones T., Pochinda S., Bevins H. T. J., Fialkov A., Handley W. J., de Lera Acedo E., Singh S., Barkana R., 2024, MNRAS, 529, 519
Gessey-Jones T., et al., 2025, arXiv e-prints, p. arXiv:2502.18098
Handley W., 2018, The Journal of Open Source Software, 3, 849
Handley W. J., Hobson M. P., Lasenby A. N., 2015a, MNRAS, 450, L61
Handley W. J., Hobson M. P., Lasenby A. N., 2015b, MNRAS, 453, 4384
Harris C. R., et al., 2020, Nature, 585, 357
Haslam C. G. T., Klein U., Salter C. J., Stoffel H., Wilson W. E., Cleary M. N., Cooke D. J., Thomasson P., 1981, A&A, 100, 209
Haslam C. G. T., Salter C. J., Stoffel H., Wilson W. E., 1982, A&AS, 47, 1
Hibbard J. J., Tauscher K., Rapetti D., Burns J. O., 2020, ApJ, 905, 113
Hibbard J. J., Rapetti D., Burns J. O., Mahesh N., Bassett N., 2023, ApJ, 959, 103
Hills R., Kulkarni G., Meerburg P. D., Puchwein E., 2018, Nature, 564, E32
Hunter J. D., 2007, Computing in Science & Engineering, 9, 90
Hyndman R. J., 1996, The American Statistician, 50, 120
Jeffreys H., 1935, Proceedings of the Cambridge Philosophical Society, 31, 203
Jeffreys H., 1939, Theory of Probability
Kass R. E., Raftery A. E., 1995, Journal of the American Statistical Association, 90, 773
Kirkham C. J., et al., 2024, arXiv e-prints, p. arXiv:2412.14023
Liu A., Pritchard J. R., Tegmark M., Loeb A., 2013, Phys. Rev. D, 87, 043002
Mahesh N., Bowman J. D., Mozdzen T. J., Rogers A. E. E., Monsalve R. A., Murray S. G., Lewis D., 2021, AJ, 162, 38
Mondal R., Barkana R., 2023, Nature Astronomy, 7, 1025
Mondal R., Barkana R., Fialkov A., 2024, MNRAS, 527, 1461
Monsalve R. A., Rogers A. E. E., Bowman J. D., Mozdzen T. J., 2017a, ApJ, 835, 49
Monsalve R. A., Rogers A. E. E., Bowman J. D., Mozdzen T. J., 2017b, ApJ, 847, 64
Monsalve R. A., et al., 2024, ApJ, 961, 56
Mozdzen T. J., Bowman J. D., Monsalve R. A., Rogers A. E. E., 2017, MNRAS, 464, 4995
Mozdzen T. J., Mahesh N., Monsalve R. A., Rogers A. E. E., Bowman J. D., 2019, MNRAS, 483, 4411
Murray S. G., Bowman J. D., Sims P. H., Mahesh N., Rogers A. E. E., Monsalve R. A., Samson T., Vydula A. K., 2022, MNRAS, 517, 2264
Naik S. S., Chingangbam P., Singh S., Mesinger A., Furuuchi K., 2025, arXiv e-prints, p. arXiv:2501.02538
Pagano M., Sims P., Liu A., Anstey D., Handley W., de Lera Acedo E., 2024, MNRAS, 527, 5649
Pattison J. H. N., Anstey D. J., de Lera Acedo E., 2024, MNRAS, 527, 2413
Pochinda S., et al., 2024, MNRAS, 531, 1113
Rapetti D., Tauscher K., Mirocha J., Burns J. O., 2020, ApJ, 897, 174
Remazeilles M., Dickinson C., Banday A. J., Bigot-Sazy M. A., Ghosh T., 2015, MNRAS, 451, 4311
Rogers A. E. E., Bowman J. D., 2012, Radio Science, 47, RS0K06
Rogers A. E. E., Bowman J. D., Vierinen J., Monsalve R., Mozdzen T., 2015, Radio Science, 50, 130
Rogers A. E. E., et al., 2022, Radio Science, 57, e2022RS007558
Roque I. L. V., Handley W. J., Razavi-Ghods N., 2021, MNRAS, 505, 2638
Roque I. L. V., et al., 2025, Experimental Astronomy, 59, 7
Saxena A., Meerburg P. D., de Lera Acedo E., Handley W., Koopmans L. V. E., 2023, MNRAS, 522, 1022
Shen E., Anstey D., de Lera Acedo E., Fialkov A., Handley W., 2021, MNRAS, 503, 344
Sims P. H., Pober J. C., 2019, MNRAS, 488, 2904

Sims P. H., Pober J. C., 2020, MNRAS, 492, 22

Sims P. H., Lentati L., Alexander P., Carilli C. L., 2016, MNRAS, 462, 3069

Sims P. H., Lentati L., Pober J. C., Carilli C., Hobson M. P., Alexander P.,
    Sutter P. M., 2019, MNRAS, 484, 4152

Sims P. H., Pober J. C., Sievers J. L., 2022a, MNRAS, 517, 910

Sims P. H., Pober J. C., Sievers J. L., 2022b, MNRAS, 517, 935

Sims P. H., et al., 2023, MNRAS, 521, 3273

Sims P. H., et al., 2025a, arXiv e-prints, p. arXiv:2502.14029

Sims P. H., et al., 2025b, arXiv e-prints, p. arXiv:2504.09725

Singh S., Subrahmanyan R., 2019, ApJ, 880, 26

Spinelli M., Bernardi G., Santos M. G., 2019, MNRAS, 489, 4007

Spinelli M., Kyriakou G., Bernardi G., Bolli P., Greenhill L. J., Fialkov A.,
    Garsden H., 2022, MNRAS, 515, 1580

Tauscher K., Rapetti D., Burns J. O., Switzer E., 2018, ApJ, 853, 187

Tauscher K., Rapetti D., Burns J. O., 2020a, ApJ, 897, 132

Tauscher K., Rapetti D., Burns J. O., 2020b, ApJ, 897, 175

Tauscher K., et al., 2021, ApJ, 915, 66

Vedantham H. K., Koopmans L. V. E., de Bruyn A. G., Wijnholds S. J., Ciardi
    B., Brentjens M. A., 2014, MNRAS, 437, 1056

Virtanen P., et al., 2020, Nature Methods, 17, 261

Wilensky M. J., et al., 2024, RAS Techniques and Instruments, 3, 400

Zheng H., et al., 2017, MNRAS, 464, 3486

APPENDIX A: 21-CM SIGNAL HPD PARAMETER ESTIMATES SUMMARY

We summarise in Table A1 our 21-cm signal detection and parameter inference results described in Section 5.

APPENDIX B: MODEL COMPARISON CATEGORISATION

In Section 5.1, we found that several models failed the BaNTER null test. These include the Intrinsic model, the LinPhys model, BFCC variants with $N = 3$ or 4, and MultLin variants with $N = 3$, 4, or 5 foreground terms.

Composite models that fail BaNTER validation cannot be considered credible for 21-cm cosmology, as they are prone to yielding spurious detections or biased estimates of the 21-cm signal (if present). However, the extent to which their inclusion in the set of considered models biases conclusions from BFBMC-alone depends on their Bayesian evidence relative to accurate and predictive composite models with accurate and predictive components (see S25 for details).

When the Bayesian evidence of failed models is substantially lower than that of the highest-evidence models, BFBMC naturally downweights these erroneous models. Consequently, their inclusion does not meaningfully influence Bayesian model-averaged conclusions or the selection of the most probable model. This corresponds to the *category I* model comparison scenario defined in S25.

In contrast, the inclusion of models that fail the BaNTER null test yet have Bayesian evidence comparable to that of the highest-evidence models is more problematic. Because these models pass Bayesian selection criteria despite failing BaNTER validation, they introduce systematic biases that cannot be corrected by BFBMC alone. In the absence of model validation, their inclusion in the model set risks significantly biasing conclusions. This corresponds to the *category II* model comparison scenario defined in S25.

B1 Moderate amplitude 21-cm signal

In the moderate amplitude analysis in Section 5.2, we determined that, among the models that detect the 21-cm signal, only the BFCC composite models with $N = 5$ and 6 yield unbiased recovery of the 21-cm signal parameters (see Figure 4). From Figure 3 (or Table 3), it can be seen that, among the models that failed BaNTER validation, only the LinPhys model and BFCC models with $N = 3$ and 4 have Bayesian evidence comparable to that of the highest posterior odds model, the BFCC model with $N = 5$.

Thus, including the LinPhys model and MultLin models with $N = 3$, 4, or 5 – all of which are decisively disfavoured relative to the BFCC model with $N = 5$ – does not significantly bias BFBMC conclusions. If only this subset of models is included alongside accurate and predictive composite models with accurate and predictive components in $\mathcal{M}$, Bayesian comparison of these models would constitute a *category I* model comparison, for which BFBMC is sufficient.

In contrast, the inclusion of the Intrinsic model or the BFCC model with $N = 4$ or 5 in the set of models under consideration means that, in the absence of model validation, BFBMC applied to the moderate amplitude data will yield biased 21-cm inferences. This represents a *category II* model comparison problem, for which BaNTER validation is essential for unbiased recovery of the 21-cm signal.

B2 High amplitude 21-cm signal

In the high amplitude analysis in Section 5.3, we determined that the BFCC composite models with $N \geq 5$ and MultLin models with $N \geq 6$ yield unbiased recovery of the 21-cm signal parameters (see Figure 6). From Figure 5, it can be seen that, among the models that failed BaNTER validation (see Table 3), only the BFCC model with $N = 4$ has Bayesian evidence comparable to that of the highest posterior odds model, the BFCC model with $N = 5$.

Thus, including the BFCC model with $N = 3$, the Intrinsic model, the LinPhys model, and MultLin models with $N = 3$, 4, or 5 – all of which are strongly or decisively disfavoured relative to the BFCC model with $N = 5$ – does not significantly bias BFBMC conclusions. If only this subset of models were included alongside accurately predictive composite models in $\mathcal{M}$, Bayesian comparison of these models would constitute a *category I* model comparison, for which BFBMC is sufficient.

In contrast, the inclusion of the BFCC model with $N = 4$ in the set of models under consideration means that, in the absence of model validation, BFBMC applied to the high amplitude data will yield biased 21-cm inferences. This represents a *category II* model comparison problem, for which BaNTER validation is essential for unbiased recovery of the 21-cm signal.

APPENDIX C: S25 ACCURACY CONDITION

Following the model-validated Bayesian inference workflow introduced in S25 we apply the Bayesian null-test described in Section 2.2.3 a priori and use the results in combination with the relative evidences of the models to derive the posterior odds we ascribe to the models using Equation (2). This approach yields a validated set of models that, of the set of models under consideration, are a posteriori most probable for recovering unbiased estimates of the global 21-cm signal in the data.

However, edge-case possibility remains that despite being the most probable models of those under consideration these models nevertheless provide insufficiently accurate descriptions of the data to be credible models for unbiased recovery of the component signals. To

Table A1. Summary of 21-cm signal detection and parameter inference for the four models (BFCC, Intrinsic, LinPhys and MultLin) and three simulated 21-cm signal amplitude scenarios (21-cm signal null test, moderate amplitude 21-cm signal and high amplitude 21-cm signal). Models in which a 21-cm signal is detected are labelled with a checkmark. Detection of a signal in the 21-cm signal null test scenario corresponds to a failure of the validation null test (see Section 5.1.1 for details). In both other scenarios, detection of a 21-cm signal is positive, while its non-detection is associated with significant correlation between the 21-cm signal and the non-21-cm component of the data model. For each data set, we list the Bayes factor ($\ln(\mathcal{B}_{i\mathrm{max}})$) between model $\boldsymbol{M}_i$ and $\boldsymbol{M}_{\mathrm{max}}$. Here, i runs over the models in $\boldsymbol{M}$ and $\boldsymbol{M}_{\mathrm{max}}$ is the highest Bayesian evidence model for the data. Additionally, we list the posterior odds ($\mathcal{R}_{i\mathrm{max}}$) between models $\boldsymbol{M}_j$ and $\boldsymbol{M}_{\mathrm{max,v}}$, where j runs over the models in $\boldsymbol{M}_{\mathrm{v}}$ and $\boldsymbol{M}_{\mathrm{max,v}}$ is a posteriori most probable BaNTER validated model. We treat models which fail model validation as having negligible probability of facilitating unbiased estimates of the 21-cm signal a priori; these models have log-posterior-odds marked with a '-'. For those models with detected signals ($\ln(\mathcal{B}_{\mathrm{cFg}}) > 3.0$), HPD parameter estimates and uncertainties corresponding to the 95% HPDI of the posterior distributions are quoted. The input parameters of the flattened Gaussian absorption troughs in the moderate- and high-amplitude signal models are $\nu_0 = 78$ MHz, $w = 19$ MHz and $\tau = 8$, in both cases, and $A = 0.15$ and 0.5 K in the moderate- and high-amplitude signal cases, respectively. The names of models that simultaneously detect the 21-cm signal and recover signal parameters consistent with the underlying signal in the data are highlighted in italic. These are found to exclusively be elements of the BaNTER validated model set $\boldsymbol{M}_{\mathrm{v}}$ (identifiable by their finite $\mathcal{R}_{i\mathrm{max}}$ values).

Scenario	Model	21-cm signal detection	$\ln(\mathcal{B}_{i\mathrm{max}})$	$\ln(\mathcal{R}_{i\mathrm{max}})$	A (K)	ν_0 (MHz)	w (MHz)	τ	Consistent
Foreground-only validation data					-	-	-	-	
	BFCC ($N = 3$)	✓	-10.9	-	$0.36^{+0.06}_{-0.05}$	$76.90^{+1.53}_{-1.70}$	$30.00^{+0.00}_{-1.39}$	$2.13^{+1.64}_{-1.17}$	✗
	BFCC ($N = 4$)	✓	0.0	-	$0.08^{+0.08}_{-0.05}$	$85.25^{+4.44}_{-4.03}$	$24.95^{+5.05}_{-6.31}$	$3.03^{+14.34}_{-3.03}$	✗
	BFCC ($N = 5$)		-1.6	0.0					
	BFCC ($N = 6$)		-3.2	-1.7					
	BFCC ($N = 7$)		-3.4	-1.8					
	BFCC ($N = 8$)		-4.5	-2.9					
	BFCC ($N = 9$)		-4.0	-2.4					
	BFCC ($N = 10$)		-4.6	-3.0					
	Intrinsic ($N = 3$)	✓	-0.0	-	$0.11^{+0.12}_{-0.04}$	$85.19^{+4.32}_{-3.54}$	$26.76^{+3.24}_{-6.23}$	$2.42^{+14.14}_{-2.42}$	✗
	LinPhys ($N = 5$)		-16.0	-					
	MultLin ($N = 3$)	✓	-4342.7	-	$1.00^{+0.00}_{-0.00}$	$78.09^{+0.13}_{-0.15}$	$25.39^{+0.25}_{-0.28}$	$18.44^{+1.56}_{-2.11}$	✗
	MultLin ($N = 4$)	✓	-113.7	-	$1.00^{+0.00}_{-0.04}$	$67.92^{+0.42}_{-0.40}$	$21.10^{+0.55}_{-0.67}$	$1.81^{+0.50}_{-0.56}$	✗
	MultLin ($N = 5$)	✓	-29.2	-	$1.00^{+0.00}_{-0.43}$	$95.00^{+0.00}_{-39.99}$	$25.36^{+2.14}_{-5.88}$	$0.00^{+0.62}_{-0.00}$	✗
	MultLin ($N = 6$)		-23.4	-21.8					
	MultLin ($N = 7$)		-25.9	-24.3					
	MultLin ($N = 8$)		-29.8	-28.2					
	MultLin ($N = 9$)		-33.1	-31.6					
	MultLin ($N = 10$)		-36.0	-34.4					
Moderate amplitude 21-cm signal					0.15	78.0	19.0	8.0	
	BFCC ($N = 3$)	✓	-0.3	-	$0.65^{+0.12}_{-0.10}$	$76.74^{+0.93}_{-1.13}$	$30.00^{+0.00}_{-2.95}$	$0.72^{+0.59}_{-0.72}$	✗
	BFCC ($N = 4$)	✓	-0.0	-	$0.27^{+0.14}_{-0.12}$	$79.30^{+1.35}_{-1.01}$	$20.12^{+6.35}_{-2.12}$	$0.81^{+5.84}_{-0.81}$	✗
	BFCC ($N = 5$)	✓	-1.2	0.0	$0.10^{+0.19}_{-0.06}$	$78.25^{+1.35}_{-1.01}$	$17.85^{+10.67}_{-4.61}$	$0.61^{+14.95}_{-0.61}$	✓
	BFCC ($N = 6$)	✓	-2.7	-1.5	$0.11^{+0.10}_{-0.04}$	$77.75^{+1.99}_{-10.37}$	$18.65^{+11.34}_{-3.03}$	$9.90^{+10.10}_{-9.69}$	✓
	BFCC ($N = 7$)		-2.6	-1.4					
	BFCC ($N = 8$)		-2.6	-1.4					
	BFCC ($N = 9$)		-2.7	-1.5					
	BFCC ($N = 10$)		-1.9	-0.8					

Table A1 – *continued*

Scenario	Model	21-cm signal detection	$\ln(\mathcal{B}_{i\max})$	$\ln(\mathcal{R}_{i\max})$	A (K)	ν_0 (MHz)	w (MHz)	τ	Consistent
Moderate amplitude 21-cm signal					0.15	78.0	19.0	8.0	
	Intrinsic ($N = 3$)	✓	0.0	-	$0.31^{+0.32}_{-0.11}$	$79.41^{+1.10}_{-2.10}$	$22.46^{+7.54}_{-2.06}$	$0.00^{+3.01}_{-0.00}$	✗
	LinPhys ($N = 5$)	✓	-16.0	-	$0.23^{+0.54}_{-0.11}$	$79.25^{+1.32}_{-1.65}$	$21.30^{+8.70}_{-1.96}$	$0.00^{+7.65}_{-0.00}$	✗
	MultLin ($N = 3$)	✓	-5120.2	-	$1.00^{+0.00}_{-0.00}$	$78.14^{+0.12}_{-0.14}$	$24.37^{+0.25}_{-0.26}$	$19.72^{+0.28}_{-2.78}$	✗
	MultLin ($N = 4$)	✓	-108.7	-	$1.00^{+0.00}_{-0.03}$	$69.21^{+0.45}_{-0.48}$	$22.39^{+0.52}_{-0.65}$	$2.19^{+0.50}_{-0.47}$	✗
	MultLin ($N = 5$)	✓	-20.9	-	$0.11^{+0.23}_{-0.08}$	$59.45^{+2.02}_{-2.83}$	$9.56^{+12.11}_{-3.03}$	$16.77^{+3.23}_{-16.77}$	✗
	MultLin ($N = 6$)		-21.0	-19.8					
	MultLin ($N = 7$)		-23.8	-22.6					
	MultLin ($N = 8$)		-28.2	-27.0					
	MultLin ($N = 9$)		-31.1	-29.9					
	MultLin ($N = 10$)		-33.1	-31.9					
High amplitude 21-cm signal					0.5	78.0	19.0	8.0	
	BFCC ($N = 3$)	✓	-17.3	-	$0.82^{+0.06}_{-0.04}$	$77.86^{+0.26}_{-0.27}$	$21.11^{+0.72}_{-0.65}$	$2.44^{+0.99}_{-0.88}$	✗
	BFCC ($N = 4$)	✓	-1.9	-	$0.54^{+0.06}_{-0.06}$	$78.40^{+0.26}_{-0.29}$	$19.35^{+0.69}_{-0.62}$	$6.22^{+3.96}_{-2.34}$	✗
	BFCC ($N = 5$)	✓	0.0	0.0	$0.48^{+0.05}_{-0.09}$	$78.23^{+0.26}_{-0.24}$	$19.06^{+0.61}_{-0.77}$	$8.14^{+9.18}_{-3.71}$	✓
	BFCC ($N = 6$)	✓	-3.0	-3.0	$0.45^{+0.08}_{-0.05}$	$78.21^{+0.25}_{-0.50}$	$19.03^{+0.62}_{-0.85}$	$9.73^{+8.35}_{-4.18}$	✓
	BFCC ($N = 7$)	✓	-3.7	-3.7	$0.41^{+0.22}_{-0.04}$	$78.15^{+0.34}_{-0.64}$	$18.95^{+0.62}_{-0.54}$	$13.83^{+1.98}_{-9.88}$	✓
	BFCC ($N = 8$)	✓	-3.0	-3.0	$0.40^{+0.16}_{-0.05}$	$77.97^{+0.42}_{-0.34}$	$18.75^{+0.85}_{-0.34}$	$13.27^{+5.13}_{-8.31}$	✓
	BFCC ($N = 9$)	✓	-3.2	-3.2	$0.40^{+0.16}_{-0.05}$	$78.03^{+0.39}_{-0.32}$	$19.03^{+0.47}_{-0.77}$	$7.59^{+9.16}_{-2.52}$	✓
	BFCC ($N = 10$)	✓	-3.6	-3.6	$0.43^{+0.14}_{-0.09}$	$77.94^{+0.43}_{-0.30}$	$18.90^{+0.68}_{-0.72}$	$11.84^{+3.44}_{-6.71}$	✓
	Intrinsic ($N = 3$)	✓	-3.6	-	$0.58^{+0.06}_{-0.07}$	$78.50^{+0.26}_{-0.27}$	$19.49^{+0.70}_{-0.63}$	$5.14^{+3.11}_{-1.83}$	✗
	LinPhys ($N = 5$)	✓	-17.4	-	$0.52^{+0.07}_{-0.06}$	$78.34^{+0.27}_{-0.27}$	$19.32^{+0.62}_{-0.66}$	$5.96^{+5.43}_{-2.17}$	✗
	MultLin ($N = 3$)	✓	-7418.5	-	$1.00^{+0.00}_{-0.00}$	$78.20^{+0.12}_{-0.10}$	$22.57^{+0.22}_{-0.23}$	$20.00^{+0.00}_{-0.88}$	✗
	MultLin ($N = 4$)	✓	-179.0	-	$1.00^{+0.00}_{-0.01}$	$71.98^{+0.42}_{-0.47}$	$22.70^{+0.58}_{-0.65}$	$2.43^{+0.47}_{-0.41}$	✗
	MultLin ($N = 5$)	✓	-25.1	-	$0.27^{+0.04}_{-0.04}$	$78.25^{+0.44}_{-0.47}$	$18.85^{+0.89}_{-0.80}$	$20.00^{+0.00}_{-8.64}$	✗
	MultLin ($N = 6$)	✓	-22.7	-22.7	$0.43^{+0.50}_{-0.11}$	$78.25^{+0.37}_{-0.35}$	$19.10^{+0.74}_{-0.74}$	$4.33^{+13.05}_{-2.05}$	✓
	MultLin ($N = 7$)	✓	-25.7	-25.7	$0.41^{+0.44}_{-0.11}$	$78.10^{+0.42}_{-0.44}$	$19.02^{+0.72}_{-0.72}$	$4.83^{+15.17}_{-2.01}$	✓
	MultLin ($N = 8$)	✓	-30.0	-30.0	$0.42^{+0.46}_{-0.12}$	$78.08^{+0.46}_{-0.46}$	$18.92^{+0.82}_{-0.66}$	$4.59^{+13.60}_{-1.99}$	✓
	MultLin ($N = 9$)	✓	-32.3	-32.3	$0.40^{+0.46}_{-0.12}$	$78.09^{+0.47}_{-0.44}$	$19.11^{+0.67}_{-0.78}$	$4.74^{+15.26}_{-1.65}$	✓
	MultLin ($N = 10$)	✓	-34.7	-34.7	$0.37^{+0.30}_{-0.12}$	$78.12^{+0.47}_{-0.45}$	$19.37^{+0.77}_{-0.86}$	$18.38^{+1.62}_{-13.34}$	✓

ensure that this is not the case, we test each of these models using the absolute accuracy condition introduced in S25.

Given the dataset D, the noise covariance matrix $\mathbf{N}$, and composite model $M_{i\mathrm{c}}$, we define the median a posteriori likelihood of $M_{i\mathrm{c}}$ as $\ln(\overline{\mathcal{L}_i})$. Writing the data likelihood $\mathcal{L}(r_{i\mathrm{c}}(\Theta_{i\mathrm{c}}))$ as a function of the residual vector, $r_{i\mathrm{c}}(\Theta_{i\mathrm{c}}) = [D - M_{i\mathrm{c}}(\Theta_{i\mathrm{c}})]$ (see Section 2.3), the likelihood distribution for an ideal model (one that describes the data perfectly, excluding noise) can be sampled by substituting $r_{i\mathrm{c}}(\Theta_{i\mathrm{c}})$ in the likelihood expression with noise realizations drawn

from the covariance matrix $\mathbf{N}$. We denote this ideal model likelihood distribution as $\mathcal{L}_{\mathrm{noise}}$. We define the model's accuracy parameter as the logarithm of the ratio of the median a posteriori likelihood of $M_{i\mathrm{c}}$ to the ideal model likelihood distribution:

$$\lambda_i = \ln\left(\frac{\overline{\mathcal{L}_i}}{\mathcal{L}_{\mathrm{noise}}}\right). \tag{C1}$$

When the distribution of λ_i is consistent with zero, this implies $\overline{\mathcal{L}_i}$

is comparable to typical values of $\mathcal{L}_{\text{noise}}$ and $\boldsymbol{M}_{i\text{c}}$ is accurate. In contrast, when most of the probability mass of λ_i is negative, $\boldsymbol{M}_{i\text{c}}$ is comparably inaccurate.

Qualitatively, we define an accurate composite model as one with a fit likelihood that is credibly drawn from the ideal model likelihood distribution. Quantitatively, we classify $\boldsymbol{M}_{i\text{c}}$ as accurate if it satisfies the following generalised accuracy condition:

$$Q_{q_{\text{threshold}}}(\lambda_i) \geq 0 \ . \tag{C2}$$

Here, $Q(.)$ is the quantile (or inverse cumulative distribution) function, defined such that for a random variable X, $Q_q(X)$ is the value of x such that $P(X \leq x) = q$. The closer $q_{\text{threshold}}$ is to unity, the further $\overline{\mathcal{L}_i}$ can fall towards the lower end of the $\mathcal{L}_{\text{noise}}$ distribution while still being classified as an element of C. In this work, we use $q_{\text{threshold}} = 0.999$ meaning that $\boldsymbol{M}_{i\text{c}}$ will fail the accuracy condition if its median posterior likelihood is exceeded by 99.9% of the probability mass of the $\mathcal{L}_{\text{noise}}$ distribution.

APPENDIX D: SOURCES OF POSSIBLE UNMODELLED SYSTEMATICS

Examples of potential sources of systematic effects that in this work are approximated as being sufficiently small to be neglected without impacting 21-cm signal inference include:

- uncertainties in the antenna beam and foreground models (e.g. Liu et al. 2013; Tauscher et al. 2018, 2020b; Rapetti et al. 2020; Hibbard et al. 2020; Anstey et al. 2021; Bassett et al. 2021; Shen et al. 2021; Mahesh et al. 2021; Rogers et al. 2022; Spinelli et al. 2022; Hibbard et al. 2023; Pagano et al. 2024; Pattison et al. 2024; Monsalve et al. 2024; Agrawal et al. 2024), which may impact the effectiveness of instrumental chromaticity correction;
- receiver calibration uncertainties (e.g. Monsalve et al. 2017a; Roque et al. 2021; Tauscher et al. 2021; Murray et al. 2022; Kirkham et al. 2024; Roque et al. 2025);
- antenna[10] and/or ground[11] loss correction uncertainties (e.g. Monsalve et al. 2017b, 2024);
- spectral chromaticity induced by ionospheric effects beyond those captured by a static, isotropic, time-averaged ionospheric model (e.g. Vedantham et al. 2014; Datta et al. 2016; Shen et al. 2021);
- polarised foreground emission (e.g. Spinelli et al. 2019).

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

---

[10] We include here resistive losses in antenna panels, the balun, and connectors.

[11] Resulting from partial absorption by the ground of radio emission visible to the antenna due to its non-zero beam directivity below the horizon.