Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **541,** 2262–2281 (2025) https://doi.org/10.1093/mnras/staf1109
Advance Access publication 2025 July 08

A general Bayesian model-validation framework based on null-test evidence ratios, with an example application to global 21-cm cosmology

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, University of Cambridge, J. J. Thomson Avenue, Cambridge CB3 0HE, UK*
[3]*Kavli Institute for Cosmology, University of Cambridge, 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]*Departamento de Ingenier´ıa Electrica, Universidad Católica de la Santísima Concepción, Alonso de Ribera 2850, Concepción, Chile*

Accepted 2025 June 26. Received 2025 June 26; in original form 2024 August 26

ABSTRACT

Comparing composite models for multicomponent observational data is a prevalent scientific challenge. When fitting composite models, there exists the potential for systematics from a poor fit of one model component to be absorbed by another, resulting in the composite model providing an accurate fit to the data in aggregate but yielding biased a posteriori estimates for individual components. We begin by defining a classification scheme for composite model comparison scenarios, identifying two categories: category I, where models with accurate and predictive components are separable through Bayesian comparison of the unvalidated composite models, and category II, where models with accurate and predictive components may not be separable due to interactions between components, leading to spurious detections or biased signal estimation. To address the limitations of category II model comparisons, we introduce the Bayesian Null Test Evidence Ratio-based (BaNTER) validation framework. Applying this classification scheme and BaNTER to a composite model comparison problem in 21-cm cosmology, where minor systematics from imperfect foreground modelling can bias global 21-cm signal recovery, we validate six composite models using mock data. We show that incorporating BaNTER alongside Bayes-factor-based comparison reliably ensures unbiased inferences of the signal of interest across both categories, positioning BaNTER as a valuable addition to Bayesian inference workflows with potential applications across diverse fields.

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

1 INTRODUCTION

Comparing competing models for observational data is a cornerstone of scientific research. In the absence of a definitive model, one must account for uncertainties in both the model parameters and the model's description of the data to draw robust inferences. Ignoring the latter can lead to biased[1] parameter estimates and underestimated uncertainties (e.g. Sims & Pober 2020; hereafter, S20).

Bayesian model comparison provides a statistically consistent means by which both model and parameter uncertainty can be accounted for (e.g. Jeffreys 1935; Kass & Raftery 1995). In its most general form, Bayesian model comparison calculates the posterior odds between models for the data. This is calculated from two factors: the prior odds of the models (our initial belief in the models) and the Bayes factor (the relative consistency of the models' fits to the data, weighted by parameter priors).

If there is no a priori information favouring any model, the prior odds are unity, and the Bayes factor and posterior odds are equivalent. In this limit, and assuming comparison of non-composite models of a data set that contains a single signal component, Bayes-factor-based model comparison (hereafter, BFBMC) provides a means of calculating the more predictive models for that signal.[2]

A more challenging problem arises when comparing models for data containing multiple signals, one of which is of primary interest,

★ E-mail: psims3@asu.edu

[1]We use 'bias' to refer to systematic errors in parameter estimates that result from model incompleteness or mis-specification, where the model does not adequately capture the true data-generating process.

[2]We use 'predictivity' to describe a model's statistical consistency with the data over its prior volume (equivalent to its Bayesian evidence).

while the rest represent nuisance structures. Examples in astrophysics include:

(i) modelling of data containing the cosmic microwave background (CMB) and nuisance foreground emission at microwave frequencies (e.g. Planck Collaboration VI 2020),
(ii) modelling of biomarker signatures in exoplanet transmission spectra in the presence of potential instrumental systematics (e.g. Madhusudhan et al. 2023; Moran et al. 2023), and
(iii) the modelling of data containing 21-cm radiation emitted by neutral hydrogen at Cosmic Dawn (e.g. Bowman et al. 2018) in the presence of nuisance radio frequency foreground emission (e.g. Spinelli, Bernardi & Santos 2019; Liu & Shaw 2020) and potential instrumental systematics (e.g. Hills et al. 2018; Bradley et al. 2019; Singh & Subrahmanyan 2019; S20; Bevins et al. 2021).

In such contexts, unlike with models of data sets measuring a single signal, there is a potential for the set of composite models under consideration to include models that are predictive of the data in aggregate but that can only obtain accurate fits to the data with biased component model fits. While BFBMC can determine the most predictive models, it cannot distinguish between models that provide accurate aggregate fits but biased component model fits. Thus, in this case, a model's inclusion in the preferred set, as determined by BFBMC, is necessary but not sufficient for ensuring accurate recovery of the signal components.

The above scenario can be made more concrete with the following example. Consider a data set $\boldsymbol{D} = \boldsymbol{S}_\mathrm{a} + \boldsymbol{S}_\mathrm{b} + \boldsymbol{n}$, where $\boldsymbol{S}_\mathrm{a}$ is the primary signal of interest (SOI), $\boldsymbol{S}_\mathrm{b}$ is a nuisance signal,[3] and $\boldsymbol{n}$ is noise. Furthermore, we have three competing composite models for the data $\boldsymbol{M}_{1\mathrm{c}} = \boldsymbol{M}_{1\mathrm{a}} + \boldsymbol{M}_{1\mathrm{b}}$, $\boldsymbol{M}_{2\mathrm{c}} = \boldsymbol{M}_{2\mathrm{a}} + \boldsymbol{M}_{2\mathrm{b}}$, and $\boldsymbol{M}_{3\mathrm{c}} = \boldsymbol{M}_{3\mathrm{a}} + \boldsymbol{M}_{3\mathrm{b}}$, such that each is the linear sum of two components that are intended to describe signals $\boldsymbol{S}_\mathrm{a}$ and $\boldsymbol{S}_\mathrm{b}$, respectively. Let us assume that $\boldsymbol{M}_{1\mathrm{c}}$ and $\boldsymbol{M}_{2\mathrm{c}}$ are comparably predictive models for the data (such that they fit the data equally well over their respective prior volumes; e.g. Trotta 2008) and both are significantly more predictive than $\boldsymbol{M}_{3\mathrm{c}}$. However, $\boldsymbol{M}_{1\mathrm{c}}$ accurately fits the data when its components accurately describe their signal counterparts. In contrast, accurate fits to the data with $\boldsymbol{M}_{2\mathrm{c}}$ are only[4] achievable through an inaccurate fit of $\boldsymbol{M}_{2\mathrm{b}}$ to $\boldsymbol{S}_\mathrm{b}$ with the residual structure in $\boldsymbol{S}_\mathrm{b}$ that is unmodelled by $\boldsymbol{M}_{2\mathrm{b}}$ being absorbed by $\boldsymbol{M}_{2\mathrm{a}}$.

Applying BFBMC to such a set of models and data, one will identify $\boldsymbol{M}_{1\mathrm{c}}$ and $\boldsymbol{M}_{2\mathrm{c}}$ as more predictive models of the data than $\boldsymbol{M}_{3\mathrm{c}}$; however, one will not be able to infer which of the models ($\boldsymbol{M}_{1\mathrm{c}}$ and $\boldsymbol{M}_{2\mathrm{c}}$) is more likely to yield unbiased inferences about the structure of $\boldsymbol{S}_\mathrm{a}$. Thus, using BFBMC to compare $\boldsymbol{M}_{1\mathrm{c}}$, $\boldsymbol{M}_{2\mathrm{c}}$, and $\boldsymbol{M}_{3\mathrm{c}}$ and weighting our models according to their relative Bayes factor will result in a spuriously high probability being assigned to biased parameter inferences from $\boldsymbol{M}_{2\mathrm{c}}$.

Breaking this degeneracy in the posterior model probabilities of $\boldsymbol{M}_{1\mathrm{c}}$ and $\boldsymbol{M}_{2\mathrm{c}}$ is possible if one has a priori knowledge of the model components that suggests that they are not equally credible[5] descriptions of their signal counterparts in the data. One way to obtain this prior information is through a null test comparing fits of the nuisance and composite models to SOI-free validation data, $\boldsymbol{D}_\mathrm{v}$. If the composite model is preferred over the nuisance model, it indicates the nuisance model is insufficiently predictive for unbiased SOI inference, and the null test is failed.

In this paper, we present BaNTER, a Bayesian Null Test Evidence Ratio-based validation framework. BaNTER calculates the evidence ratio (Bayes factor) between nuisance and composite models for SOI-free validation data. These validation results are used to make an a priori assessment of the probability that unbiased estimates of the SOI are obtainable with the composite model. We combine this with an analysis of the a posteriori probability of the data from BFBMC in a validated posterior-odds-based model comparison methodology.

We demonstrate an application of this approach to the problem of inferring accurate and precise estimates of the cosmological signal in global 21-cm cosmology mock data. In this test case, despite one of the models under consideration being the generative model (GM) for the mock data, using BFBMC alone does not reliably downweight biased alternatives. Consequently, it leads to biased 21-cm signal estimation. In contrast, we demonstrate that a BaNTER validated posterior-odds-based methodology enables unbiased estimates of the 21-cm SOI to be obtained.

In this work, we emphasize a detailed explanation of the BaNTER validation methodology. We use a test case with a compact set of models to demonstrate its efficacy and to highlight salient features. In Sims et al. (2025), we apply the BaNTER framework to a broader comparison of data models that have been applied to 21-cm signal estimation in the literature (Bowman et al. 2018; Hills et al. 2018; Sims et al. 2023).

The remainder of the paper is organized as follows. In Section 2, we introduce our composite model validation methodology. In Section 3, we apply it to a comparison of competing models for beam-factor-chromaticity-corrected global 21-cm signal data. We discuss the efficacy and limitations of the validation methodology, and potential improvements in Section 4. In Section 5, we provide a summary and discuss avenues for future work.

2 NULL-TEST-BASED BAYESIAN MODEL VALIDATION

The logical 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 one does not know whether models of this type are included in the set of models under consideration. When they are, treating these models as a priori equally probable will result in a degeneracy in the space of a posteriori model probabilities between these models and those that we wish to identify (those that are both predictive of the data in aggregate and have predictive component models). The purpose of the model validation framework introduced in this section is to identify when a set of competing composite models exhibits this degeneracy and to remove it by identifying and downweighting models that can only achieve accurate fits to the data using biased fits of its component models.

In Section 2.1, we begin by introducing the principles of Bayesian inference, upon which our model validation and posterior-odds-based model comparison method is based. In Section 2.2, we introduce

[3]More generally, the data may contain N nuisance signals that contribute structure to the data. For the purpose of forward modelling the data, these may be modelled individually; however, for this work we treat their sum as a single nuisance signal and the sum of the models for these signals as a single nuisance model.

[4]Here, we are referring to a model with components with which it is impossible to accurately describe their signal counterparts at the signal-to-noise level of the data but which in combination can describe the data in aggregate. It does not, for example, refer to a model for which the posteriors of the model parameters in a fit to the data are consistent with those parameters in nature, regardless of whether there are degeneracies between parameters.

[5]We use 'credibility' to refer to the degree of belief placed in a model.

a classification scheme of model comparison scenarios based on whether they exhibit the aforementioned degeneracy between preferred models for the data as judged by BFBMC. This also serves to introduce the model set notation used in the remainder of the paper. In Section 2.3, we present BaNTER validation that is designed to eliminate this degeneracy, enabling unbiased inferences of the SOI.

2.1 Bayesian inference

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.

When, rather than a single definitive model, one has a set of models for the data, $\boldsymbol{\mathcal{M}} = \{\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_N\}$, preferred models for the data can be determined from their marginal probabilities. Using Bayes' theorem at the model level, the marginal probability of a model $\boldsymbol{M}_i$, drawn from $\boldsymbol{\mathcal{M}}$, is given by:

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

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

For the purpose of comparing two specific models for the data, $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$, drawn from $\boldsymbol{\mathcal{M}}$, one can calculate their posterior odds:

$$\begin{aligned}\underbrace{\mathcal{R}_{ij}}_{\text{Posterior odds}} &= \frac{\mathcal{P}(\boldsymbol{M}_i|\boldsymbol{D})}{\mathcal{P}(\boldsymbol{M}_j|\boldsymbol{D})} \\ &= \frac{\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_i)\mathcal{P}(\boldsymbol{M}_i)}{\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_j)\mathcal{P}(\boldsymbol{M}_j)} \\ &= \underbrace{\mathcal{B}_{ij}}_{\text{Bayes factor}}\,\underbrace{\frac{\mathcal{P}(\boldsymbol{M}_i)}{\mathcal{P}(\boldsymbol{M}_j)}}_{\text{Prior odds}}\,.\end{aligned} \tag{3}$$

Here, $\mathcal{P}(\boldsymbol{M}_i|\boldsymbol{D})$ is given by equation (2), the Bayes Factor, $\mathcal{B}_{ij}$, is the posterior odds of the data given $\boldsymbol{M}_i$ and $\boldsymbol{M}_j$, $\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_i) \equiv \mathcal{Z}_i$ and $\mathcal{P}(\boldsymbol{D}|\boldsymbol{M}_j) \equiv \mathcal{Z}_j$ are the marginal likelihoods of the data in 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.

Table 1. Standards of a posteriori model preference for model $\boldsymbol{M}_i$ over $\boldsymbol{M}_j$ as a function of $\ln(\mathcal{R}_{ij})$, the log of the posterior odds between the two models. If one has no a priori information that makes any one of the models more probable than the others, the prior odds are unity and the Bayes factor and the posterior odds between models are equivalent. In this situation, we use the same qualitative descriptions for matching ranges of $\ln(\mathcal{B}_{ij})$.

$\ln(\mathcal{R}_{ij})$	Odds in favour of $\boldsymbol{M}_i$	Preference for $\boldsymbol{M}_i$
$0 \leq \ln(\mathcal{R}_{ij}) < 1$	1–3	Weak
$1 \leq \ln(\mathcal{R}_{ij}) < 3$	3–20	Moderate
$3 \leq \ln(\mathcal{R}_{ij}) < 5$	20–150	Strong
$\ln(\mathcal{R}_{ij}) \geq 5$	> 150	Decisive

In this work, we use the definitions in Table 1 to map between the posterior odds of two models and qualitative terms describing the relative preference for one over the other. 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. In this case, the posterior odds are equal to the Bayes factor between the models ($\mathcal{R}_{ij} = \mathcal{B}_{ij}$), and for the purpose of defining our qualitative descriptions the Bayes factor between models takes the place of the posterior odds. We have chosen boundaries between our qualitative descriptions such that in this limit they map to those given by Kass & Raftery (1995) for the preference of one model over another.

2.2 Classifying model comparison scenarios

2.2.1 *Motivation*

Suppose we wish to describe a data set of the form given in the Introduction:

$$\boldsymbol{D} = \boldsymbol{S}_\mathrm{a} + \boldsymbol{S}_\mathrm{b} + \boldsymbol{n}\,, \tag{4}$$

with $\boldsymbol{S}_\mathrm{a}$ the SOI. Consider two sets of models for describing $\boldsymbol{S}_\mathrm{a}$ and $\boldsymbol{S}_\mathrm{b}$:

$$\boldsymbol{\mathcal{M}}_\mathrm{a} = \{\boldsymbol{M}_{1\mathrm{a}}, \ldots, \boldsymbol{M}_{N_\mathrm{a}\mathrm{a}}\} = \{\boldsymbol{M}_{j\mathrm{a}} \mid j \in [1, N_\mathrm{a}]\} \tag{5}$$

and

$$\boldsymbol{\mathcal{M}}_\mathrm{b} = \{\boldsymbol{M}_{k\mathrm{b}} \mid k \in [1, N_\mathrm{b}]\}\,, \tag{6}$$

respectively. Additionally, let,

$$\boldsymbol{\mathcal{M}}_\mathrm{c} = \{\boldsymbol{M}_{i\mathrm{c}} \mid i \in [1, N_\mathrm{c}]\}\,, \tag{7}$$

denote the set of composite models obtained by summing pairs of models, one from $\boldsymbol{\mathcal{M}}_\mathrm{a}$ and one from $\boldsymbol{\mathcal{M}}_\mathrm{b}$[6]

Here, N_a, N_b, and N_c are the number of models in the three sets.

In the Introduction, we considered three competing composite models ($\boldsymbol{M}_{1\mathrm{c}}$, $\boldsymbol{M}_{2\mathrm{c}}$, and $\boldsymbol{M}_{3\mathrm{c}}$), with $\boldsymbol{M}_{1\mathrm{c}}$ and $\boldsymbol{M}_{2\mathrm{c}}$ being comparably predictive of the data and both being significantly more predictive

[6]More generally, one could consider data and models of the form $\boldsymbol{D} = f(\boldsymbol{\Theta}_\mathrm{a}, \boldsymbol{\Theta}_\mathrm{b}) + \boldsymbol{n}$ and $\boldsymbol{M}_{i\mathrm{c}} = f_{M_i}(\boldsymbol{\Theta}_{j\mathrm{a}}, \boldsymbol{\Theta}_{k\mathrm{b}})$, respectively. Here, $f(\cdot)$ denotes the function encoding the generative processes of the SOI and nuisance components, the propagation effects between the emission sources and the instrument, the instrument transfer function, and any corrections applied to the data prior to analysis. The function $f_{M_i}(\cdot)$ represents our model for this generative process. The parameters $\boldsymbol{\Theta}_\mathrm{a}$ and $\boldsymbol{\Theta}_\mathrm{b}$ describe the SOI and nuisance components, respectively, while $\boldsymbol{\Theta}_{j\mathrm{a}}$ and $\boldsymbol{\Theta}_{k\mathrm{b}}$ parametrize our model of these processes. The approach discussed in this work generalizes straightforwardly to this case (e.g. Sims et al. 2025).

than $\boldsymbol{M}_{3c}$. However, $\boldsymbol{M}_{1c}$ accurately fits the data when its components accurately describe their signal counterparts, while accurate fits to the data with $\boldsymbol{M}_{2c}$ are only achievable through an inaccurate fit of $\boldsymbol{M}_{2a}$ to $\boldsymbol{S}_a$, with the residual structure in $\boldsymbol{S}_a$ that is unmodelled by $\boldsymbol{M}_{2a}$ being absorbed by $\boldsymbol{M}_{2b}$. BFBMC will separate $\boldsymbol{M}_{1c}$, which we aim to identify, from the significantly lower predictivity composite model, $\boldsymbol{M}_{3c}$. However, it will not yield a preference for $\boldsymbol{M}_{1c}$ over $\boldsymbol{M}_{2c}$, due to their comparable predictivities for the data in aggregate. Thus, in this case, BFBMC alone is sufficient for identifying the most predictive composite models for the data, but not for uniquely identifying $\boldsymbol{M}_{1c}$ as the only composite models that additionally has component models that accurately describe their signal counterparts.

Counterposing this example, the logical possibility also exists for the set of models under consideration to be composed exclusively of composite models that are predictive of the data and have components that accurately describe their signal counterparts (such as $\boldsymbol{M}_{1c}$), as well as composite models that are significantly less predictive of the data (such as $\boldsymbol{M}_{3c}$). In this simpler case, BFBMC alone is sufficient for identifying both the most predictive models for the data and the most predictive component models for the signals, rather than the former alone.

Let us label these two cases as category II (challenging) and category I (simple) composite model comparison scenarios, respectively. In the category I scenario, model validation is incidental, while in the category II scenario, it is essential. To determine under what conditions model validation is required, it is necessary to understand the properties of the composite and component models that lead to categories I and II model comparison scenarios. In the remainder of this section, we analyse the logical relations between the sets of models in $\mathcal{M}_a$, $\mathcal{M}_b$, and $\mathcal{M}_c$, and classify under what conditions these two scenarios arise.

2.2.2 *Categorizing models by their predictivity and accuracy*

To identify when one is dealing with a category I or II model comparison problem, and thus whether model validation is required, let us classify the models in $\mathcal{M}_a$, $\mathcal{M}_b$, and $\mathcal{M}_c$ according to whether they meet a given predictivity and accuracy[7] threshold. This will further serve to construct a notation for distinguishing the desired set of composite models from the remainder that are either not predictive of the data in aggregate, or predictive of the data in aggregate but composed of components that are not predictive of their signal counterparts.

We will use predictivity and accuracy criteria to classify models in each of $\mathcal{M}_a$, $\mathcal{M}_b$, and $\mathcal{M}_c$ into pairs of complementary subsets. Each pair will consist of one subset containing the preferred models that meet the predictivity and accuracy criteria for the signal component(s) they are intended to describe, and another containing the disfavoured models that do not meet these criteria.

We begin by defining the predictivity and accuracy criteria in the context of the composite models in $\mathcal{M}_c$, and will return to discuss their application to $\mathcal{M}_a$ and $\mathcal{M}_b$ afterwards.

For the composite models in $\mathcal{M}_c$, we define the subsets containing the preferred and disfavoured models as C and $\overline{C}$, respectively.

Given the data set $\boldsymbol{D}$ and the noise covariance matrix $\mathbf{N}$, we denote by $\mathcal{Z}_{max}$ the Bayesian evidence of the most predictive model for the data, $\boldsymbol{M}_{c,max}$, and by $\mathcal{B}_{imax}$ the Bayes factor between the ith model, $\boldsymbol{M}_{ic}$, and $\boldsymbol{M}_{c,max}$, such that

$$\mathcal{B}_{imax} = \frac{\mathcal{Z}_i}{\mathcal{Z}_{max}} . \tag{8}$$

We also define the median a posteriori likelihood of $\boldsymbol{M}_{ic}$ as $\ln(\overline{\mathcal{L}}_i)$. Writing the data likelihood $\mathcal{L}(\boldsymbol{r}_{ic}(\boldsymbol{\Theta}_{ic}))$ as a function of the residual vector, $\boldsymbol{r}_{ic}(\boldsymbol{\Theta}_{ic}) = [\boldsymbol{D} - \boldsymbol{M}_{ic}(\boldsymbol{\Theta}_{ic})]$ (see Section 3.1.7), the likelihood distribution for an ideal model (one that describes the data perfectly, excluding noise) can be sampled by substituting $\boldsymbol{r}_{ic}(\boldsymbol{\Theta}_{ic})$ in the likelihood expression with noise realizations drawn from the covariance matrix $\mathbf{N}$. We denote this ideal model-likelihood distribution as $\mathcal{L}_{noise}$. Finally, we define the model's accuracy parameter as the logarithm of the ratio of the median a posteriori likelihood of $\boldsymbol{M}_{ic}$ to the ideal model-likelihood distribution:

$$\lambda_i = \ln\left(\frac{\overline{\mathcal{L}}_i}{\mathcal{L}_{noise}}\right) . \tag{9}$$

When the distribution of λ_i is consistent with zero, this implies $\overline{\mathcal{L}}_i$ is comparable to typical values of $\mathcal{L}_{noise}$ and $\boldsymbol{M}_{ic}$ is accurate. In contrast, when most of the probability mass of λ_i is negative, $\boldsymbol{M}_{ic}$ is comparably inaccurate.

Qualitatively, we define a predictive and accurate composite model as one with a Bayesian evidence equal to or comparable to the most predictive model, and a fit likelihood that is credibly drawn from the ideal model-likelihood distribution. Quantitatively, we classify $\boldsymbol{M}_{ic}$ as an element of C if it satisfies the following predictivity and accuracy conditions:

(i) *Predictivity condition*.

$$\ln(\mathcal{B}_{imax}) \geq \ln(\mathcal{B}_{threshold}); \tag{10}$$

(ii) *Accuracy condition*.

$$Q_{q_{threshold}}(\lambda_i) \geq 0 . \tag{11}$$

Here, $Q(\cdot)$ is the quantile (or inverse cumulative distribution) function, defined such that for a random variable X, and $Q_q(X)$ is the value of x such that $P(X \leq x) = q$. The closer $q_{threshold}$ is to unity, the further $\overline{\mathcal{L}}_i$ can fall towards the lower end of the $\mathcal{L}_{noise}$ distribution while still being classified as an element of C.

In this work, we define $\ln(\mathcal{B}_{threshold}) = -3$ and $q_{threshold} = 0.999$. Condition (i) guarantees that the odds in favour of $\boldsymbol{M}_{c,max}$ over $\boldsymbol{M}_{ic}$ are at most 20-to-1 (see Table 1), meaning that $\boldsymbol{M}_{ic}$ is among the most predictive of the composite models under consideration. However, this condition alone does not address situations in which none of the models can accurately describe the data. Condition (ii) addresses this by ensuring that the residuals of the fitted model are consistent with the noise in the data. Specifically, $\boldsymbol{M}_{ic}$ will fail condition (ii) only if its median posterior likelihood is exceeded by 99.9 per cent of the probability mass of the $\mathcal{L}_{noise}$ distribution. If a composite model does not fulfill both of these conditions, it is classified as an element of $\overline{C}$.

For the purposes of model categorization in Sections 2.2.4 and 2.2.5, we assume that $\mathcal{M}_c$ contains at least one model that is an element of C. When this is known a priori, one can directly assess whether the models in $\mathcal{M}_c$ provide predictive and accurate descriptions of the data using condition (i). However, if this is uncertain, one can test whether condition (ii) holds for the models in $\mathcal{M}_c$, to validate that this is the case.

Given $\boldsymbol{D}$, $\mathbf{N}$, and the set of composite models for the data $\mathcal{M}_c$, there is no barrier to calculating the quantities required to assess whether a model fulfills conditions (i) and (ii), and thus to sort the models a posteriori into the sets C and $\overline{C}$. In contrast, direct

[7]We use 'accuracy' to refer to the statistical consistency of $\boldsymbol{M}(\overline{\boldsymbol{\Theta}})$ with the data, with $\overline{\boldsymbol{\Theta}}$ the median parameter sample for the posterior.

access to equivalent data sets containing only $\boldsymbol{S}_{\rm a}$ and $\boldsymbol{S}_{\rm b}$ may not be available, potentially necessitating more approximate, simulation-based methods to assess the predictivity of the models in $\boldsymbol{\mathcal{M}}_{\rm a}$ and $\boldsymbol{\mathcal{M}}_{\rm b}$. In either case, one can nevertheless define equivalent subsets of models in $\boldsymbol{\mathcal{M}}_{\rm a}$ and $\boldsymbol{\mathcal{M}}_{\rm b}$ to complete the conceptual framework.[8] We classify the models in these sets according to whether they are predictive and accurate descriptions of their corresponding signal components, using the same criteria described for models in $\boldsymbol{\mathcal{M}}_{\rm c}$, but with the update that the signal component being fitted is exclusively $\boldsymbol{S}_{\rm a}$ or $\boldsymbol{S}_{\rm b}$, respectively. With this update, we define the models in $\boldsymbol{\mathcal{M}}_{\rm a}$ and $\boldsymbol{\mathcal{M}}_{\rm b}$ that satisfy their corresponding predictivity and accuracy conditions as elements of the sets A and B, respectively, and their logical complements as elements of $\overline{A}$ and $\overline{B}$, respectively.

The predictivity and accuracy conditions given in equations (10) and (11) are chosen for their generality and because the quantities required to assess them are readily available, or can be easily computed, from the inputs and outputs of a Bayesian model comparison analysis using nested sampling (see Section 3.3).

2.2.3 *Component model permutations and the resulting composites*

As a stepping stone to categorizing the models in $\boldsymbol{\mathcal{M}}_{\rm c}$ based on the categorization of their component models, we consider the possible permutations of the sets A, $\overline{A}$, B, and $\overline{B}$ that can be used to construct composite models in C, and $\overline{C}$. Let us use arithmetic multiplication to represent the logical AND operation, such that AB describes a pair of models, $\boldsymbol{M}_{j\rm a}$ and $\boldsymbol{M}_{k\rm b}$, drawn from A and B, respectively. Composite models in C and $\overline{C}$ are formed by combining pairs of models, one from $\boldsymbol{\mathcal{M}}_{\rm a}$ and one from $\boldsymbol{\mathcal{M}}_{\rm b}$. These pairs can be drawn from four permutations of A, $\overline{A}$, B, and $\overline{B}$, which we can write as: AB, $A\overline{B}$, $\overline{A}B$, and $\overline{A}\,\overline{B}$.

An inaccurate or non-predictive composite model (i.e. an element of $\overline{C}$) can be constructed by combining models for $\boldsymbol{S}_{\rm a}$ and $\boldsymbol{S}_{\rm b}$ where at least one, and potentially both, models are inaccurate. This results in three permutations that can form composite models in $\overline{C}$, namely: $A\overline{B}\,\overline{C}$, $\overline{A}B\overline{C}$, and $\overline{A}\,\overline{B}\,\overline{C}$. We note that a composite model composed of accurate and predictive components (i.e. when both $\boldsymbol{S}_{\rm a}$ and $\boldsymbol{S}_{\rm b}$ are represented by accurate and predictive models) will necessarily be accurate and predictive of the data in aggregate, and thus an element of C. Therefore, the combination $AB\overline{C}$ is not a valid permutation, and the corresponding set is empty.

In contrast, a composite model that can accurately describe and is predictive of the data in aggregate (and thus belongs to the set C) can be constructed by combining models for $\boldsymbol{S}_{\rm a}$ and $\boldsymbol{S}_{\rm b}$ where both are accurate, or when either or both are inaccurate, subject to the condition that any inaccuracies in one model can be absorbed by the other, such that the composite model remains accurate overall. Therefore, all four permutations can be used to construct composite models in C, namely: ABC, $A\overline{B}C$, $\overline{A}BC$, and $\overline{A}\,\overline{B}C$.

To aid intuition for these set definitions their interpretation in the context of 21-cm cosmology is discussed in Appendix A.

2.2.4 *Category I model comparison*

Fig. 1 shows Euler diagrams[9] describing the relationships between the sets A, $\overline{A}$, B, $\overline{B}$, C, and $\overline{C}$ in three model comparison categories.

For clarity, and to distinguish them from $\boldsymbol{\mathcal{M}}_{\rm a}$, $\boldsymbol{\mathcal{M}}_{\rm b}$, $\boldsymbol{\mathcal{M}}_{\rm c}$, and other sets referred to in the remainder of the paper, we will refer to these sets, defined by the predictivity and accuracy conditions in equations (10) and (11), as 'Euler sets'.

The Euler diagram in Fig. 1(a) illustrates the logical relationships between these Euler sets in a scenario where the component models must be elements of A and B for their composite to be an element of C – that is, when accurate and predictive composite models for the data exclusively consist of accurate and predictive component models for their signal counterparts.

In this case, all elements of C are elements of ABC. Thus, the more predictive model will necessarily also have more accurate recovery of the signal components. Consequently, BFBMC provides a direct means of estimating the relative probability (weighted by model complexity and its ability to fit the data) that a model falls into the ABC set (or equivalently, that it will yield unbiased parameter estimates). As such, Fig. 1(b) illustrates the logical relations between Euler sets in a category I model comparison.

2.2.5 *Category II model comparison*

Figs 1(c) and (b) illustrate more complex scenarios in which composite models in C are drawn from multiple subsets. In Fig. 1(c), the composite models in C are drawn from four subsets: the set of accurate and predictive composite models with accurate and predictive component models for $\boldsymbol{S}_{\rm a}$ and $\boldsymbol{S}_{\rm b}$ (ABC), and the sets of accurate and predictive composite models with components for $\boldsymbol{S}_{\rm a}$, $\boldsymbol{S}_{\rm b}$, or both, that have low accurate and/or predictivity ($\overline{A}BC$, $A\overline{B}C$, and $\overline{A}\,\overline{B}C$, respectively).

Models in $\overline{A}BC$ have components $\boldsymbol{M}_{j\rm a}$ that are low-accuracy and/or predictivity models for $\boldsymbol{S}_{\rm a}$. However, in the composite model $\boldsymbol{M}_{ic}$, they are paired with components $\boldsymbol{M}_{k\rm b}$ that have sufficient flexibility to model both the underlying signal $\boldsymbol{S}_{\rm b}$ (if present) and the structure in $\boldsymbol{S}_{\rm a}$ that cannot be captured by $\boldsymbol{M}_{j\rm a}$. Similarly, $A\overline{B}C$ describes composite models with components $\boldsymbol{M}_{k\rm b}$ that are low-accuracy and/or predictivity models for $\boldsymbol{S}_{\rm b}$. In these composites, the $\boldsymbol{M}_{k\rm b}$ models are paired with components $\boldsymbol{M}_{j\rm a}$ that have sufficient flexibility to absorb any underlying signal $\boldsymbol{S}_{\rm a}$ (if present) and structure in $\boldsymbol{S}_{\rm b}$ that cannot be modelled by $\boldsymbol{M}_{k\rm b}$. Finally, $\overline{A}\,\overline{B}C$ describes composite models in which both components $\boldsymbol{M}_{j\rm a}$ and $\boldsymbol{M}_{k\rm b}$ are low-accuracy and/or predictivity models for $\boldsymbol{S}_{\rm a}$ and $\boldsymbol{S}_{\rm b}$, respectively. However, $\boldsymbol{M}_{j\rm a}$ has sufficient flexibility to absorb structure in $\boldsymbol{S}_{\rm b}$ that cannot be modelled by $\boldsymbol{M}_{k\rm b}$ and vice versa.

Fig. 1(b) illustrates a simplified model comparison scenario in which composite models in C are drawn from two subsets: ABC and $A\overline{B}C$.

Both Figs 1(c) and (b) illustrate category II model comparison problems. The logical relations between competing models for a data set simplify from Figs 1(c) to (b) when $\overline{A}\,BC$ and $\overline{A}\,\overline{B}C$ are empty sets (i.e. have no elements). This occurs when the models fulfill one of the following two conditions:

(i) *Condition 1*. There are no component models for $\boldsymbol{S}_{\rm b}$ that can absorb structure in $\boldsymbol{S}_{\rm a}$ that is impossible to model with $\boldsymbol{M}_{j\rm a}$. Thus, all composite models containing components from $\overline{A}$ are elements of $\overline{C}$.

(ii) *Condition 2*. There exists a definitive model, $\boldsymbol{M}_{\rm a}$, such that there are no component models for $\boldsymbol{S}_{\rm a}$ in $\overline{A}$.

In this paper, we derive a framework for unbiased inference of $\boldsymbol{S}_{\rm a}$ from a data set when either of the above conditions is fulfilled,

[8] We will discuss how the membership of these sets can be inferred from the combined results of model validation and BFBMC later in the text.

[9] An Euler diagram is a visual tool used to represent sets and their relationships. Unlike Venn diagrams, which show all possible set relations, Euler diagrams show only relevant relationships.

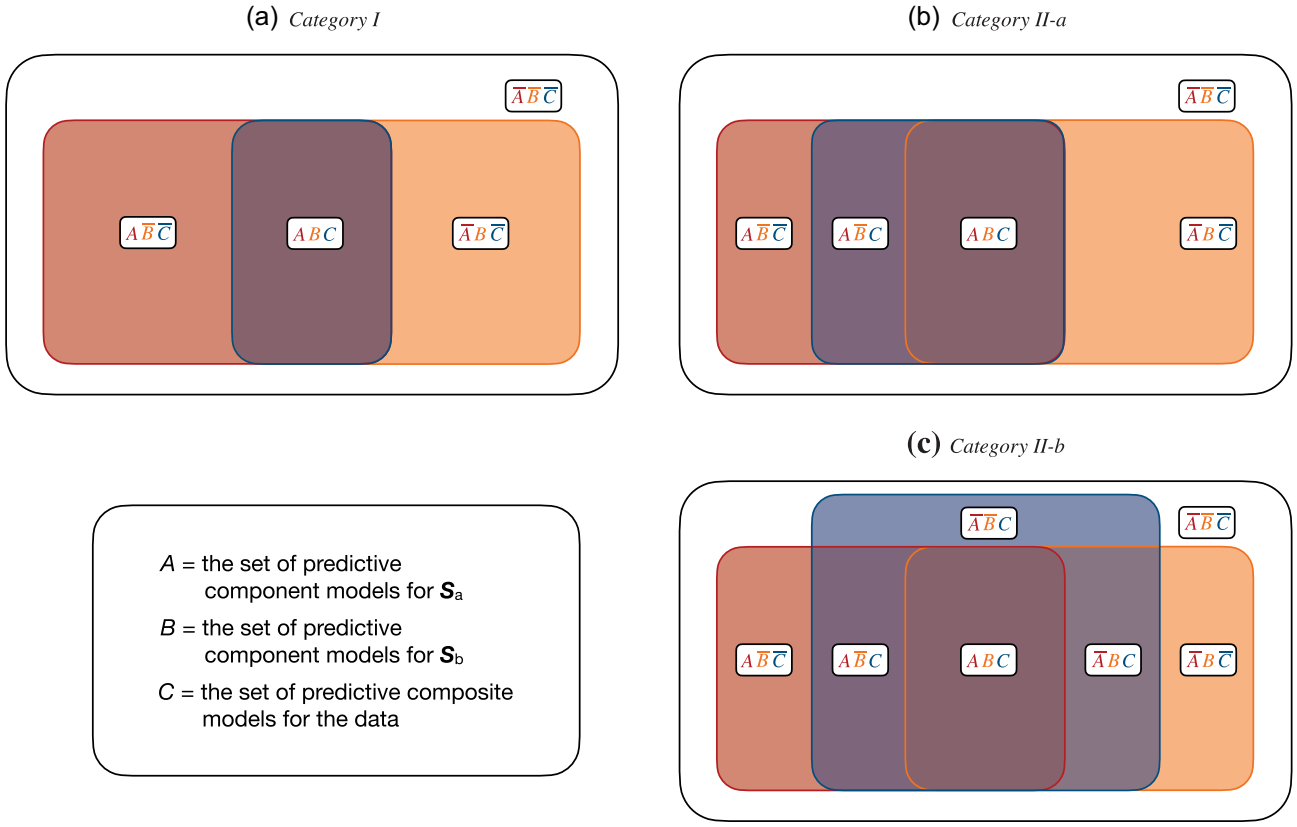

Figure 1. Euler diagrams representing the logical relations between a set of competing composite models for a data set $\boldsymbol{D} = \boldsymbol{S}_a + \boldsymbol{S}_b + \boldsymbol{n}$ in categories I and II model comparison scenarios. A, B, and C represent the Euler sets of predictive and accurate models for $\boldsymbol{S}_a$, $\boldsymbol{S}_b$, and the data in aggregate, respectively. In each plot, 'logical and' is represented by arithmetic multiplication (e.g. ABC denotes the Euler set of models for which A, B, and C are true.), and $\overline{A}$, $\overline{B}$, and $\overline{C}$ denote the logical complements to A, B, and C (e.g. the Euler set of component models for $\boldsymbol{S}_a$, $\boldsymbol{S}_b$ with relatively low predictivity and the Euler set of composite models with relatively low predictivity, respectively). Panel (a): Euler diagram representing a 'category I' scenario in which the Euler set of predictive composite models for the data derives exclusively from the intersection of composite models containing predictive component models for $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$. Panel (b): Euler diagram representing the category II-a scenario in which the Euler set of predictive composite models for the data contains two subsets: the Euler set of predictive composite models with predictive component models for $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ that we are interested in (ABC) and the Euler set of predictive composite models with component models for $\boldsymbol{S}_a$ (the SOI) that are predictive and for $\boldsymbol{S}_b$ (nuisance-structure in the data) that have low predictivity ($A\overline{B}C$). Panel (c): Euler diagram representing the category II-b scenario in which the Euler set of predictive composite models for the data contains four subsets: the Euler set of predictive composite models with predictive component models for $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ that we are interested in (ABC) and the Euler sets of predictive composite models with component models for $\boldsymbol{S}_a$ or $\boldsymbol{S}_b$ or both that have low predictivity ($\overline{A}BC$, $A\overline{B}C$, and $\overline{A}\,\overline{B}C$, respectively).

and the logical relations between the competing models for the data are as illustrated in Fig. 1(b) (category II-a). For concreteness, we assume that Condition 2 is fulfilled: we have a definitive model for $\boldsymbol{S}_a$, $\boldsymbol{M}_a$, which is an element of set A.[10] Thus, the set of models under consideration contains no elements in $\overline{A}$, and the composite models are elements of one of the following three sets: ABC, $A\overline{B}C$, or $A\overline{B}\,\overline{C}$.

In practice, if one does not have a definitive model for $\boldsymbol{S}_a$, the methodology described in this work can be used to derive a set of conditionally validated composite models. In this case, the validation of the composite models is contingent on the probability of $\boldsymbol{M}_a$. If there are several candidate models for $\boldsymbol{S}_a$, the validation methodology may be applied for each model, and the conclusions weighted by the relative probabilities assigned to each model of $\boldsymbol{S}_a$. We leave to future work a more detailed examination of the generalization to category II-b scenarios. Going forward, for brevity, we drop the 'a' label and describe analyses and model comparisons categorized by the Euler diagram in Fig. 1(b) as category II data analyses and model comparisons, respectively.

To aid intuition for the model comparison classification introduced in this and the preceding section, the interested reader can find a discussion of how it applies to a concrete example in the context of 21-cm cosmology in Appendix B.

2.2.6 Transforming a category II model comparison problem into a category I model comparison problem

A category II model comparison scenario is more likely when either or both of the component model sets contain flexible models with a priori poorly constrained parameters. In general, this increases the likelihood that $\boldsymbol{M}_a$ can absorb residual structures arising from imperfect modelling of $\boldsymbol{S}_b$ by $\boldsymbol{M}_{ib}$. Category I model comparison represents a special limiting case of category II model comparison, where $A\overline{B}C$ is the empty set. Approaching this limit becomes more likely when one can reduce the volume of the model space under

[10]Given a definitive model for $\boldsymbol{S}_a$, it follows that $\boldsymbol{\mathcal{M}}_a = \{\boldsymbol{M}_a\}$. Correspondingly, there is a 1-to-1 mapping between the composite and nuisance component models under consideration. To reflect this, we label the nuisance models with subscript i rather than k, with $\boldsymbol{M}_{ic} = \boldsymbol{M}_a + \boldsymbol{M}_{ib}$.

comparison using either implicit or explicit experimental[11] or theory-based priors.

In a category II scenario, BFBMC still allows the separation of models in set C from those in its complement, $\overline{C}$. This disfavours composite models with components that fail to provide a consistent model of their signal counterparts and cannot produce the observed data by any means. However, BFBMC does not enable one to distinguish between models of comparable complexity in the set ABC, for which one expects to recover unbiased estimates of the signal components, and those in the set $A\overline{B}C$, for which this is not the case. Therefore, using BFBMC alone limits one's ability to draw robust conclusions regarding which models will minimize bias in recovered parameter estimates.

However, if one can identify and exclude from the set of models under consideration those composite models in $A\overline{B}C$ a priori, then it is possible to transform a category II model comparison problem into a category I model comparison problem. We discuss a Bayesian-null-test-based methodology for achieving this transformation in the next section.

2.3 BaNTER

Let us define a validation data set in which the only signal component is $\boldsymbol{S}_\mathrm{b}$:

$$\boldsymbol{D}_\mathrm{v} = \boldsymbol{S}_\mathrm{b} + \boldsymbol{n}\,. \tag{12}$$

Ideally one would use $\boldsymbol{S}_\mathrm{b}$-only observations for $\boldsymbol{D}_\mathrm{v}$; however, if obtaining such a data set is not experimentally feasible, realistic high-fidelity simulated $\boldsymbol{S}_\mathrm{b}$-only observations can provide a good substitute.

Two characteristic signatures of models in the set $A\overline{B}C$ are:

(i) the $\boldsymbol{M}_{i\mathrm{b}}$ component of $\boldsymbol{M}_{i\mathrm{c}}$ is a non-predictive model of the $\boldsymbol{S}_\mathrm{b}$ component of the data,

(ii) the sum of statistically significant systematics resulting from inaccuracies of $\boldsymbol{M}_{i\mathrm{b}}$ as a model for $\boldsymbol{S}_\mathrm{b}$ can be absorbed by $\boldsymbol{M}_\mathrm{a}$ such that $\boldsymbol{M}_{i\mathrm{c}} = \boldsymbol{M}_\mathrm{a} + \boldsymbol{M}_{i\mathrm{b}}$ yields a predictive fit to the data in aggregate.

Fundamentally, we are interested in assessing whether unbiased estimates of the SOI will be obtained in a fit of $\boldsymbol{M}_{i\mathrm{c}} = \boldsymbol{M}_\mathrm{a} + \boldsymbol{M}_{i\mathrm{b}}$ to $\boldsymbol{D}$. Thus, to validate a model $\boldsymbol{M}_{i\mathrm{b}}$ in $\mathcal{M}_\mathrm{b}$, one must answer the question: is $\boldsymbol{M}_{i\mathrm{b}}$ sufficiently predictive of $\boldsymbol{S}_\mathrm{b}$ for a fit of $\boldsymbol{M}_{i\mathrm{c}} = \boldsymbol{M}_\mathrm{a} + \boldsymbol{M}_{i\mathrm{b}}$ to $\boldsymbol{D}$ to yield unbiased inferences of $\boldsymbol{S}_\mathrm{a}$?

If one can identify and exclude models in $\overline{B}$ a priori, the corresponding composite models in $A\overline{B}C$ will also be eliminated from the set of models under consideration for $\boldsymbol{D}$. Thus, initially one may assume that characteristic (i), alone, is sufficient for defining a model validation metric. However, in practice, this approach has limitations.

Equation (2) can be used to calculate the posterior probabilities of models in $\mathcal{M}_\mathrm{b}$ for $\boldsymbol{D}_\mathrm{v}$ or the posterior odds of the models can be determined using BFBMC (see Section 2.1). The former provides an absolute measure of the model predictivities and the latter a measure of their relative predictivities; however, neither directly answers whether $\boldsymbol{M}_{i\mathrm{b}}$ is *sufficiently* accurate for our science case. Ultimately, this question is dependent on the models in $\mathcal{M}_\mathrm{a}$ as well as those in $\mathcal{M}_\mathrm{b}$, with larger imperfections in $\boldsymbol{M}_{i\mathrm{b}}$ acceptable if they are not describable with $\boldsymbol{M}_\mathrm{a}$, since in this case they cannot bias the SOI parameter inference. Thus, in this work, we take an alternate validation approach that addresses this question directly.

[11]For example, through the analysis of alternate datasets that depend on the same underlying model.

2.3.1 *Identifying and excluding models in $A\overline{B}C$*

Combining characteristics (i) and (ii), it follows that models in $\mathcal{M}_\mathrm{c}$ that are elements of the set $A\overline{B}C$ are identifiable with a null test between $\boldsymbol{M}_{i\mathrm{b}}$ and $\boldsymbol{M}_{i\mathrm{c}} = \boldsymbol{M}_\mathrm{a} + \boldsymbol{M}_{i\mathrm{b}}$ as models for $\boldsymbol{D}_\mathrm{v}$. Here, a preference for $\boldsymbol{M}_{i\mathrm{c}}$ over $\boldsymbol{M}_{i\mathrm{b}}$ as a model for $\boldsymbol{D}_\mathrm{v}$ suggests that $\boldsymbol{M}_{i\mathrm{b}}$ provides an incomplete description of the structure in $\boldsymbol{S}_\mathrm{b}$, *and* that a spurious fit of $\boldsymbol{M}_\mathrm{a}$ absorbs residual structure in the validation data that $\boldsymbol{M}_{i\mathrm{b}}$ alone cannot fit. In a fit of $\boldsymbol{M}_{i\mathrm{c}}$ to $\boldsymbol{D}$, the same residual structure modelled by $\boldsymbol{M}_\mathrm{a}$ in the validation data will bias the SOI parameter inference. This signifies that $\boldsymbol{M}_{i\mathrm{b}}$ is not sufficiently predictive for our science case.

Since $\boldsymbol{M}_{i\mathrm{b}}$ is a component of $\boldsymbol{M}_{i\mathrm{c}}$, the latter will always provide an equally good or better maximum-likelihood fit to the data. Thus, we judge the relative preference for $\boldsymbol{M}_{i\mathrm{c}}$ over $\boldsymbol{M}_{i\mathrm{b}}$ by the Bayes factor between the models.[12] Let us define a log-Bayes-factor between a composite model $\boldsymbol{M}_{i\mathrm{c}} = \boldsymbol{M}_{i\mathrm{b}} + \boldsymbol{M}_\mathrm{a}$ and the component model for the nuisance structure, $\boldsymbol{M}_{i\mathrm{b}}$, as

$$\ln(\mathcal{B}^\mathrm{v}_\mathrm{cb}) = \ln(\mathcal{Z}^\mathrm{v}_\mathrm{c}/\mathcal{Z}^\mathrm{v}_\mathrm{b})\,, \tag{13}$$

where $\mathcal{Z}^\mathrm{v}_\mathrm{c} = \mathcal{P}(\boldsymbol{D}_\mathrm{v}|\boldsymbol{M}_{i\mathrm{c}})$ and $\mathcal{Z}^\mathrm{v}_\mathrm{b} = \mathcal{P}(\boldsymbol{D}_\mathrm{v}|\boldsymbol{M}_{i\mathrm{b}})$ are the Bayesian evidence of $\boldsymbol{M}_{i\mathrm{c}}$ and $\boldsymbol{M}_{i\mathrm{b}}$, respectively.[13] When $\ln(\mathcal{B}^\mathrm{v}_\mathrm{cb}) < 0$, $\boldsymbol{M}_{i\mathrm{b}}$ is preferred over $\boldsymbol{M}_{i\mathrm{c}}$ for $\boldsymbol{D}_\mathrm{v}$, as expected when fitting $\boldsymbol{S}_\mathrm{b}$-only validation data.

Model classes with validation Bayes factors in the range $0 < \ln(\mathcal{B}^\mathrm{v}_\mathrm{cb}) \leq 3$ have a weak-to-moderate preference for $\boldsymbol{M}_{i\mathrm{c}}$ over $\boldsymbol{M}_{i\mathrm{b}}$ (odds in favour of $\boldsymbol{M}_{i\mathrm{c}}$ over $\boldsymbol{M}_{i\mathrm{b}}$ of 20-to-1 or less; see Table 1). This suggests that inaccuracies in the nuisance model are sufficient to bias estimates of SOI if present. However, under a Bayesian 21-cm detection criterion that requires strong evidence in favour of the composite model over the model for nuisance structure ($\ln(\mathcal{B}_\mathrm{cb}) \geq 3$; see Section 3.2), this is below the level at which we would consider $\boldsymbol{S}_\mathrm{a}$ to have been spuriously detected.

$\ln(\mathcal{B}^\mathrm{v}_\mathrm{cb}) \geq 3$ corresponds to a strong preference for $\boldsymbol{M}_{i\mathrm{c}}$ over $\boldsymbol{M}_{i\mathrm{b}}$ in the validation data, with odds in favour of $\boldsymbol{M}_{i\mathrm{c}}$ over $\boldsymbol{M}_{i\mathrm{b}}$ of 20-to-1 or better (see Table 1). This indicates that nuisance model inaccuracies are severe enough to significantly bias 21-cm signal estimates if present, or to lead to a spurious detection of the SOI if it is absent.

Here, we set a threshold above which we judge a nuisance model to have failed the null test of: $\ln(\mathcal{B}^\mathrm{v}_\mathrm{cb}) \geq \ln(\mathcal{B}^\mathrm{v}_\mathrm{threshold})$, with $\ln(\mathcal{B}^\mathrm{v}_\mathrm{threshold}) = 0$. In general, higher values of $\ln(\mathcal{B}^\mathrm{v}_\mathrm{threshold})$ correspond to less stringent validation criteria and vice versa. Ultimately, the value used for $\ln(\mathcal{B}^\mathrm{v}_\mathrm{threshold})$ should be informed by the quality of the validation data. We come back to this in Section 4.

[12]Here, we use the Bayes factor as our validation metric due to its general applicability and simple interpretability. However, if simpler to calculate for the problem at hand, one might achieve a similar goal with alternate validation metrics including information criteria such as the Bayesian, Akaike or Deviance Information Criteria (e.g. Trotta 2008) or comparison of Kolmogorov–Smirnov test p-values deriving from the residuals of fits of $\boldsymbol{M}_{i\mathrm{b}}$ and $\boldsymbol{M}_{i\mathrm{c}}$ to $\boldsymbol{D}_\mathrm{v}$.

[13]$\mathcal{Z}^\mathrm{v}_\mathrm{b}$ and $\mathcal{Z}^\mathrm{v}_\mathrm{c}$ can be estimated using nested sampling (see Section 3.3). Additionally, for composite models where $\boldsymbol{M}_{i\mathrm{c}}$ reduces to $\boldsymbol{M}_{i\mathrm{b}}$ for a specific set of parameters of $\boldsymbol{M}_\mathrm{a}$ one can use the computationally efficient Savage–Dicke method (e.g. Wagenmakers et al. 2010) to calculate $\ln(\mathcal{B}^\mathrm{v}_\mathrm{cb})$ directly from the prior and posterior distributions of $\boldsymbol{M}_{i\mathrm{c}}$. In this case, one negates the computational expense associated with deriving $\mathcal{Z}^\mathrm{v}_\mathrm{b}$ and $\mathcal{Z}^\mathrm{v}_\mathrm{c}$.

2.3.2 SOI model dependence

While here we have assumed that we have a definitive model for $\boldsymbol{S}_{\rm a}$ that is an element of set A (see Section 2.2.5), more generally, since our proposed null test will generically identify the subset of models where the systematics resulting from an inaccurate fit of $\boldsymbol{M}_{i\rm b}$ to $\boldsymbol{S}_{\rm b}$ can be accurately fit with $\boldsymbol{M}_{\rm a}$, we should expect its results to be dependent on the specific choice of model for $\boldsymbol{M}_{\rm a}$. The more flexible the $\boldsymbol{M}_{\rm a}$ component model, the more likely it is that it will be able to accurately describe systematics resulting from an inaccurate $\boldsymbol{M}_{i\rm b}$. However, the use of the Bayes factor between models for the validation data in BaNTER validation balances this against the larger Occam penalty associated with the greater complexity of $\boldsymbol{M}_{\rm a}$ (see Section 2.1). Thus, $\boldsymbol{M}_{i\rm b}$ will only fail the null test if $\boldsymbol{M}_{i\rm c}$ provides a sufficiently better fit to the data to justify its greater complexity.

2.3.3 Null test results as model priors

If the BaNTER results indicate a statistically significant preference for $\boldsymbol{M}_{i\rm c}$ over $\boldsymbol{M}_{i\rm b}$, $\boldsymbol{M}_{i\rm b}$, and $\boldsymbol{M}_{i\rm c}$ are judged to have failed the null test. Given this information, one can update their assessment of the probability of these models. Here, we consider models that fail the validation null test to have negligible probability as models with which unbiased inferences of the SOI can be recovered in subsequent analysis of the data.

2.3.4 Identifying and excluding models in $A\overline{B}\,\overline{C}$

In addition to its primary purpose of identifying models in $A\overline{B}C$, the BaNTER results will additionally identify models that produce a spurious detection of $\boldsymbol{S}_{\rm a}$ in the set $A\overline{B}\,\overline{C}$.[14] Unlike models in $A\overline{B}C$, which we need validation data to identify, we expect models that are elements of $A\overline{B}\,\overline{C}$ to independently be identified and disfavoured through BFBMC applied to the observational data. However, by disfavouring or eliminating all models that fail the null test, the additional disfavouring of models in $A\overline{B}\,\overline{C}$ a priori can be treated as an ancillary benefit of the validation analysis.

2.3.5 Using null test results and model accuracy to determine model comparison classification

A composite model's failure of the validation null test identifies it as an element of either $A\overline{B}\,\overline{C}$ or $A\overline{B}C$. One can obtain unbiased inferences regarding the SOI by downweighting models that fail the validation null test irrespective of which of these two sets they are elements. Thus, determining which of the two sets a failed model is an element of is not an essential component of our proposed model-validated Bayesian inference workflow discussed in the next section. Nevertheless, confirmation that models can be correctly identified as elements of $A\overline{B}\,\overline{C}$ and $A\overline{B}C$ using the validation analysis provides an additional cross-check of the methodology, and thus we briefly consider how it can be achieved.

[14]This is not guaranteed to identify all models in $A\overline{B}\,\overline{C}$, since the logical possibility exists for systematics resulting from inaccuracies of $\boldsymbol{M}_{i\rm b}$ as models for $\boldsymbol{S}_{\rm b}$ to be of a form not modellable with $\boldsymbol{M}_{\rm a}$. Nevertheless, as such models will be elements of $\overline{C}$ they will still be identified by BFBMC and any inferences from them will be appropriately downweighted according to the degree to which they are disfavoured as models for the observational data, $\boldsymbol{D}$.

One can predict whether a composite model that fails BaNTER validation is an element of $A\overline{B}\,\overline{C}$ or $A\overline{B}C$ based on its ability to fit the validation data. If, in addition to failing the null test, the composite model does not meet the accuracy condition given in equation (11), this implies it is likely to be an element of $A\overline{B}\,\overline{C}$. In contrast, if it fulfills the accuracy condition, albeit using a spurious fit of the SOI model, it is more likely to be an element of $A\overline{B}C$.

2.4 Bayesian inference workflow

Fig. 2 shows a flowchart of our BaNTER model validation framework and the way it interfaces with a typical Bayesian inference workflow. We divide the resulting model-validated Bayesian inference workflow into three main sections: model specification (orange), model validation (green), and data analysis (blue). The steps joined by dashed-red lines describe a typical Bayesian workflow with uninformative model priors. The steps joined by solid black lines describe the model-validated Bayesian inference workflow proposed in this work. We describe each below.

2.4.1 Bayesian workflow with uninformative model priors

In the model specification component of a typical Bayesian framework, one begins by specifying a set of models, $\boldsymbol{\mathcal{M}}$, that are considered as candidate models for the data. In Fig. 2, we label this as the 'unvalidated model space'. For each model in this set one specifies the parameter priors, $\mathcal{P}(\boldsymbol{\Theta}_i|\boldsymbol{M}_i)$ and the model parametrization $\boldsymbol{M}_i(\boldsymbol{\Theta}_i)$ that defines the mapping between $\boldsymbol{\Theta}_i$ and the modelled observables that will be compared with the data.

In the data analysis component, one collects the observational data, $\boldsymbol{D}$, and specifies a data likelihood, $\mathcal{P}(\boldsymbol{\Theta}|\boldsymbol{D}, \boldsymbol{M})$ that reflects the noise covariance structure of the data. One then uses the unvalidated candidate models to fit the data according to the assumed likelihood. Conceptually, the analysis can then be divided into two tasks:[15] (i) parameter estimation, in which one draws samples from the parameter priors, calculates the corresponding data likelihood and posterior probability and uses the posterior samples to estimate the posterior density function of the model parameters; and (ii) model comparison, in which one calculates the model evidence, $\mathcal{Z}_i$, and the Bayes factor between models, $\mathcal{B}_{ij}$ (see Section 2.1 for details).

In the unvalidated model case, the Bayes factors define the relative probabilities of the models. If the weight of evidence in favour of one model over the others is sufficiently large, one may perform model selection. Alternatively and more generally, the model evidences can be used as weights for the parameter posteriors when deriving Bayesian model-averaged parameter estimates and posterior predictive distributions for the signal components. In Fig. 2, we label these as unvalidated signal inference. If the unvalidated model space is well approximated by category I model comparison, unbiased inferences can be obtained at this stage. If instead it represents a category II model comparison problem, this procedure will yield biased inferences resulting from undue probability being assigned to biased inferences from models in $A\overline{B}C$.

[15]In practice, if one uses a nested sampling algorithm to draw samples from the prior, as done in this work (see Section 3.3), model evidence and parameter estimation are carried out simultaneously, with the latter being a by-product of the former.

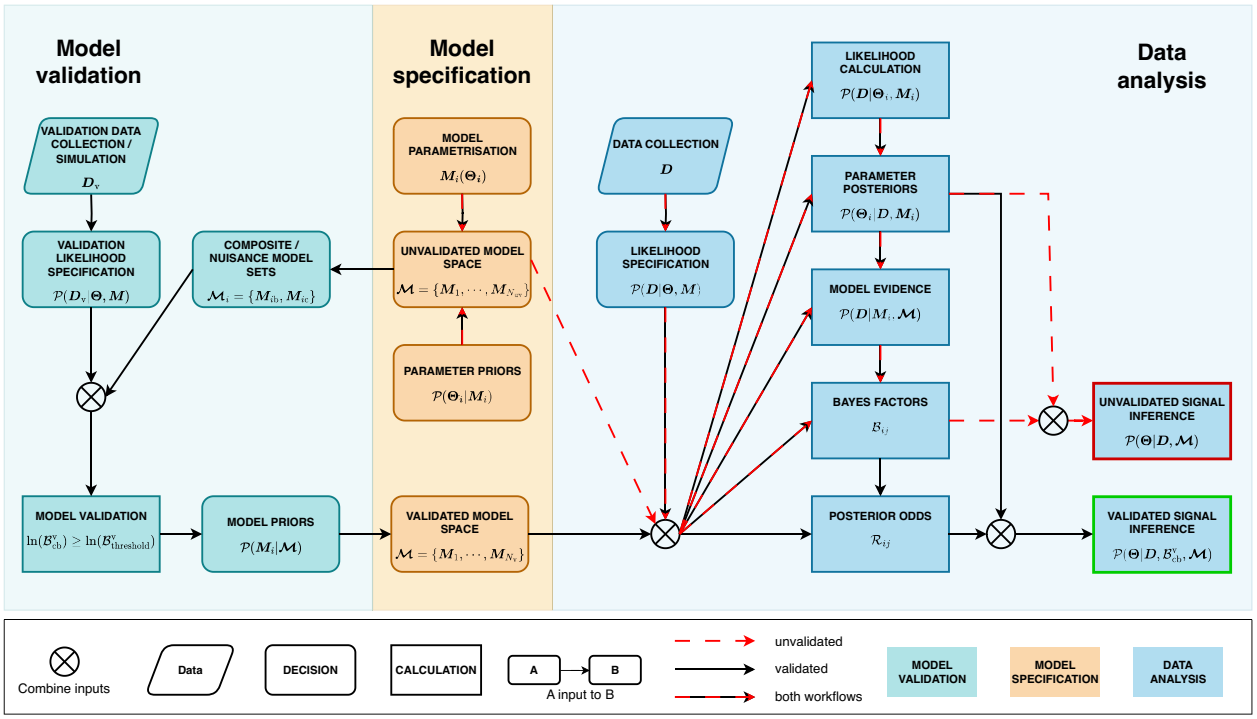

Figure 2. Flowchart of a model-validated Bayesian inference workflow incorporating the BaNTER validation methodology introduced in this work. The workflow is divided into three main sections: model specification (orange), model validation (green), and data analysis (blue). Steps joined by dashed-red lines describe a Bayesian workflow with uninformative model priors (see Section 2.4.1). The model-validated Bayesian inference framework we propose in this work augments this workflow with BaNTER model validation. Steps joined by solid black lines describe our model-validated Bayesian inference workflow (see Section 2.4.2). Steps joined by solid black overlaid by dashed-red lines are used in both workflows.

2.4.2 Model-validated Bayesian inference workflow

In the model-validated Bayesian inference workflow we begin by defining the unvalidated model space, as before. The BaNTER model validation component of the workflow is then responsible for identifying composite models in the unvalidated model space that are elements of the set $A\overline{B}C$ versus $A\overline{B}\,\overline{C}$ or $A\overline{B}\,\overline{C}$, enabling one to separate the former from the latter and update the probability one assigns to the models in advance of analysing the observational data $\boldsymbol{D}$.[16] In this process, the models that fail the validation analysis are either,

(i) discarded from the set of models for the observational data under consideration, or,
(ii) assigned a lower prior model probability than those which pass.

In this work, we treat the composite models that fail the validation null test as having negligible probability as models with which one may extract unbiased inferences of the SOI. As such, the above two options are equivalent. We use the second approach because it enables additional cross-checks regarding the Euler set that models which fail model validation derive from and, correspondingly, the category of the model comparison problem that is being undertaken.

However, more generally, by eliminating the computational expense associated with deriving the Bayesian analysis data products from fits of the models that fail the validation analysis to the observational data, the first approach benefits from improved computational efficiency. Therefore, in future work, when applying the model-validated Bayesian inference workflow rather than demonstrating features of the model validation framework, the first approach may be preferable. We refer to the subset of models that pass the validation null test, which we consider credible models for deriving unbiased inferences regarding the SOI, as the 'validated model space'.

The data analysis component of the model-validated Bayesian inference workflow proceeds in the same manner as the typical workflow with the following update. At the model comparison stage, one uses the parameter posteriors and posterior odds (rather than Bayes factors[17]) to derive Bayesian model averaged parameter estimates and posterior predictive distributions for the signal components via equation (2). This validated signal inference will yield unbiased estimates of the SOI in both categories I and II model comparison problems.[18]

[16]Alternatively, one could achieve validated signal inference by applying BaNTER validation to the preferred set of models identified via BFBMC in the model comparison stage of the Bayesian workflow with uninformative model priors described in Section 2.4.1. In this case, the initial model comparison is used to separate models in ABC and $A\overline{B}C$ from those in $A\overline{B}\,\overline{C}$ and BaNTER is used to distinguish between models in ABC and $A\overline{B}C$.

[17]The posterior odds are used implicitly if one has discarded models that fail the validation analysis from the set of models for the observational data under consideration. However, the practical procedure is to weight the model posteriors according to their Bayes factors (in the same manner as the typical workflow, but here applied to the validated model subset). In contrast, if models that fail the validation analysis are downweighted but not discarded, when calculating the posterior odds the reduced probability assigned to those models explicitly enters equation (2), at this stage, via the model prior terms.

[18]A caveat is that one should assign a probability to the results of the validation analysis in proportion to the quality of the validation data. Thus, if one lacks

2.4.3 Model accuracy condition

In both the unvalidated and validated Bayesian inference workflows discussed in Sections 2.4.1 and 2.4.2, respectively, we assumed that $\mathcal{M}_c$ contains at least one model that is an element of C. If this is uncertain a priori, before proceeding with Bayesian model-averaged parameter estimation we can additionally validate that the preferred models identified by model comparison meet the model accuracy condition given in equation (11). Assuming they do, having also passed the BaNTER null test, they can be identified as elements of ABC. In contrast, if they do not, they are judged to have failed model validation and considered insufficiently accurate for reliable inferences of the SOI. We assign negligible probability to these model, the same approach as with those that fail the Bayesian null test.

3 APPLICATION TO GLOBAL 21-CM COSMOLOGY

Determining whether robust inferences from a set of models requires category I or II model comparison holds significant importance in the field of global 21-cm cosmology. The signal measured in a radiometric global 21-cm experiment can be written as the sum of two components: (a) a faint sky-averaged 21-cm cosmology signal (that we are interested in) and (b) intense foreground emission (a nuisance signal), both of which propagate through an instrumental transfer function. One would expect category II model comparison to be necessary in global 21-cm cosmology if a subset of the foreground model components of a competing set of models cannot fully describe the foreground component of the data, but one or more of the composite models under consideration includes a 21-cm model that is capable of fitting the sum of systematics resulting from the inaccuracy of the foreground model and the 21-cm signal in the data (if present).

In this section, we apply the BaNTER model validation methodology, described in Section 2, to an example model comparison problem motivated by global 21-cm cosmology data modelling. We consider three sets of data models ($\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$) that together illustrate salient features of the model validation methodology. We analyse the data using both the Bayesian workflow with uninformative model priors and the model-validated Bayesian inference workflow described in Section 2.4. This enables us to demonstrate that only with the latter workflow can one:

(i) distinguish between sets of models requiring categories I and II model comparison,
(ii) identify models in $A\overline{B}C$ if present,
(iii) ascertain that the composite models in the aforementioned three sets are elements of the ABC, $A\overline{B}C$, and $A\overline{B}\,\overline{C}$, respectively,
(iv) recover unbiased inferences of the cosmological 21-cm signal in the category II model comparison problem.

3.1 Data models and mock data

We consider a set of three competing composite 1D spectral models for a mock global 21-cm signal data set $\boldsymbol{D} = \boldsymbol{S}_a + \boldsymbol{S}_b + \boldsymbol{n}$, where $\boldsymbol{S}_a$ is the 21-cm signal and $\boldsymbol{S}_b$ is foreground emission:[19]

(i) *Model 1.* $\boldsymbol{M}_{1c} = \boldsymbol{M}_a + \boldsymbol{M}_{1b}$.
(ii) *Model 2.* $\boldsymbol{M}_{2c} = \boldsymbol{M}_a + \boldsymbol{M}_{2b}$.
(iii) *Model 3.* $\boldsymbol{M}_{3c} = \boldsymbol{M}_a + \boldsymbol{M}_{3b}$.

Each composite model has a 21-cm signal component with a functional form matching the true GM for the 21-cm signal in the mock data ($\boldsymbol{M}_a$) and one of three foreground model components ($\boldsymbol{M}_{1b}$, $\boldsymbol{M}_{2b}$, and $\boldsymbol{M}_{3b}$). Going forward, we use $\mathcal{M}_i = \{\boldsymbol{M}_{ic}, \boldsymbol{M}_{ib}\}$ to denote the ith composite/nuisance model set. When analysing the data, we assume that $\boldsymbol{S}_b$ is known to be present in the data, a priori; however, the presence, or not, of $\boldsymbol{S}_a$ is unknown. The full set of models under consideration for the data is given by $\mathcal{M} = \{\boldsymbol{M}_{1c}, \boldsymbol{M}_{1b}, \boldsymbol{M}_{2c}, \boldsymbol{M}_{2b}, \boldsymbol{M}_{3c}, \boldsymbol{M}_{3b}\}$. Prior to model validation or analysis of the data we treat each of the models in $\mathcal{M}$ as equally probable.

3.1.1 21-cm signal model

We model the 21-cm signal as a flattened-Gaussian absorption trough with a functional form matching the model parametrization used in Bowman et al. (2018, hereafter B18):

$$\boldsymbol{M}_a = \boldsymbol{T}_{21} = -A\left(\frac{1-\mathrm{e}^{-\tau \mathrm{e}^{\boldsymbol{B}_{21}}}}{1-\mathrm{e}^{-\tau}}\right), \tag{14}$$

with,

$$\boldsymbol{B}_{21} = \frac{4(\boldsymbol{\nu}-\nu_0)^2}{w^2}\ln\left[-\frac{1}{\tau}\ln\left(\frac{1+\mathrm{e}^{-\tau}}{2}\right)\right]. \tag{15}$$

Here, A, ν_0, w, and τ describe the amplitude, central frequency, width, and flattening of the absorption trough, respectively, and $\boldsymbol{\nu}$ is a vector of central frequencies of the channels over the observing band of the data.

3.1.2 Foreground models

As the basis of our foreground description we use the class of models defined by the flexible closed-form parametrization derived for beam-factor chromaticity corrected radio spectrometer data [20] in S23:

$$\begin{aligned}\boldsymbol{T}^{\mathrm{Fg}}_{\mathrm{BFCC}}(\boldsymbol{\Theta}_{\mathrm{BFCC}}) = \Bigg[\bar{T}_{\mathrm{m}_0}\left(\frac{\boldsymbol{\nu}}{\nu_c}\right)^{-\beta_0}\left(1+\sum_{\alpha=1}^{N_{\mathrm{pert}}} p_\alpha \ln\left(\frac{\boldsymbol{\nu}}{\nu_c}\right)^{\alpha}\right) \\ + \frac{\left(1-\left(\frac{\boldsymbol{\nu}}{\nu_c}\right)^{-\beta_0}\right)T_\gamma}{\bar{\boldsymbol{B}}_{\mathrm{factor}}(\boldsymbol{\nu})}\Bigg]\mathrm{e}^{-\tau_{\mathrm{ion}}(\boldsymbol{\nu})} \\ + \frac{T_e}{\bar{\boldsymbol{B}}_{\mathrm{factor}}(\boldsymbol{\nu})}\left(1-\mathrm{e}^{-\tau_{\mathrm{ion}}(\boldsymbol{\nu})}\right).\end{aligned} \tag{16}$$

Here, all operations on vector quantities are elementwise, and we parametrize the effective ionospheric optical depth as,

$$\tau_{\mathrm{ion}} = \tau_0(\nu/\nu_c)^{-2}, \tag{17}$$

high-quality validation data this reduces the ability to rule out models that fail model validation with high confidence. We consider this issue further, and ways in which its impact can be mitigated, in Section 4.

[19] In practice, the instrument and ionosphere impart multiplicative effects on the astrophysical signals such that the data are more accurately represented as $\boldsymbol{D} = f(\boldsymbol{S}_a + \boldsymbol{S}_b) + \boldsymbol{n}$. However, these effects are small under ideal conditions regarding ionospheric conditions and instrumental calibration and here, for simplicity, we neglect them and treat the data as the sum of independent components.

[20] In Sims et al. (2025), we test a set of competing models drawn from several model classes that have been applied to 21-cm signal estimation in the literature. We find that, assuming high-fidelity modelling of the instrument, models of the form given in equation (16) are most predictive of beam-factor chromaticity corrected radio spectrometer data.

Table 2. A summary of the structural properties of the models in $\mathcal{M} = \{\boldsymbol{M}_{1c}, \boldsymbol{M}_{1b}, \boldsymbol{M}_{2c}, \boldsymbol{M}_{2b}, \boldsymbol{M}_{3c}, \boldsymbol{M}_{3b}\}$.

Model	Number of parameters	Instrument–foreground coupling terms ($N_{\rm pert}$)	Includes ionospheric model	Includes 21-cm signal model
$\boldsymbol{M}_{1c}$	11	3	Y	Y
$\boldsymbol{M}_{2c}$	9	1	Y	Y
$\boldsymbol{M}_{3c}$	8	2	N	Y
$\boldsymbol{M}_{1b}$	7	3	Y	N
$\boldsymbol{M}_{2b}$	5	1	Y	N
$\boldsymbol{M}_{3b}$	4	2	N	N

Table 3. Priors on the parameters of the 21-cm signal and foreground model components. The relation between τ_0 and $\tau_{\rm ion}$ is given in equation (17).

Model component	Model	Parameter	Prior
21-cm signal	$\boldsymbol{M}_{a}$	A	$U(0, 1)$ K
	$\boldsymbol{M}_{a}$	ν_0	$U(55, 95)$ MHz
	$\boldsymbol{M}_{a}$	w	$U(5, 30)$ MHz
	$\boldsymbol{M}_{a}$	τ	$U(0, 20)$
Foreground	$\boldsymbol{M}_{1b}$, $\boldsymbol{M}_{2b}$, $\boldsymbol{M}_{3b}$	$\bar{T}_{\rm m_0}$	$U(1000, 6000)$ K
	$\boldsymbol{M}_{1b}$, $\boldsymbol{M}_{2b}$, $\boldsymbol{M}_{3b}$	β_0	$U(2.0, 3.0)$
	$\boldsymbol{M}_{1b}$, $\boldsymbol{M}_{2b}$	$T_{\rm e}$	$U(100, 800)$ K
	$\boldsymbol{M}_{1b}$, $\boldsymbol{M}_{2b}$	τ_0	$U(0.005, 0.025)$
	$\boldsymbol{M}_{1b}$, $\boldsymbol{M}_{2b}$, $\boldsymbol{M}_{3b}$	p_0	$U(-0.1, 0.1)$
	$\boldsymbol{M}_{1b}$, $\boldsymbol{M}_{3b}$	p_1	$U(-0.1, 0.1)$
	$\boldsymbol{M}_{1b}$	p_2	$U(-0.1, 0.1)$

where $\nu_c = 75$ MHz is a reference frequency. The free parameters in equation (16) are $\boldsymbol{\Theta}_{\rm BFCC} = [\bar{T}_{\rm m_0}, \beta_0, T_{\rm e}, \tau_0, p_1, \cdots, p_{N_{\rm pert}}]$. We describe these parameters and provide an overview of the individual terms in the equation in Appendix C. For a detailed description of the physical and instrumental effects described by equation (16), we refer the reader to S23. This, however, is not a prerequisite for understanding the features of Bayesian model validation and comparison that are the focus of this paper.

We define our three foreground component models using the following variants of the equation (16) model class:

(i) $\boldsymbol{M}_{1b}$ has $N_{\rm f} = 7$ fitted foreground parameters, including: the amplitude and spectral index of the foreground emission ($\bar{T}_{\rm m_0}$ and β_0), the temperature of ionospheric electrons and the effective optical depth of the ionosphere (at reference frequency ν_c) through which the foreground emission travels ($T_{\rm e}$ and τ_0, respectively) and $N_{\rm pert} = 3$ terms for describing the amplitude of instrumentally coupled spectral fluctuations about the sky-averaged spectrum of the foreground brightness temperature field. Hereafter, for brevity, we refer to the $N_{\rm pert}$ model components as instrument-foreground coupling terms. As will be discussed in Section 3.1.4, we use $\boldsymbol{M}_{1b}$ as a GM for $\boldsymbol{S}_a$ when constructing the mock data.

(ii) $\boldsymbol{M}_{2b}$ has $N_{\rm f} = 5$ free parameters, matching those in model $\boldsymbol{M}_{1b}$, but with only $N_{\rm pert} = 1$ instrument–foreground coupling term. Thus, this model assumes that structure in the data from higher than first-order instrument-foreground coupling is negligible. If this assumption does not hold, meaning $\boldsymbol{M}_{2b}$ provides a poor description of $\boldsymbol{S}_b$, but the model $\boldsymbol{M}_{2c}$ can fit the data accurately by biasing the foreground and the 21-cm signal parameters, the models in $\boldsymbol{\mathcal{M}}_2$ will fail the null test.

(iii) $\boldsymbol{M}_{3b}$ has $N_{\rm f} = 4$ free parameters, including the amplitude and spectral index of the foreground emission and $N_{\rm pert} = 2$ instrument–foreground coupling terms. This model assumes that structures in the data described by the third-order instrument-foreground coupling term and ionospheric effects are negligible. As with $\boldsymbol{M}_{2b}$, if these assumptions do not hold, meaning $\boldsymbol{M}_{3b}$ provides a poor description of $\boldsymbol{S}_b$, but $\boldsymbol{M}_{3c}$ can better fit the data by biasing the other foreground parameters and the 21-cm signal parameters, then the models in $\boldsymbol{\mathcal{M}}_3$ will fail the null test.

We note that for continuity, foreshadowing the results, the model labelling has been chosen such that the results with the composite models incorporating these components, $\boldsymbol{M}_{1c}$, $\boldsymbol{M}_{2c}$, and $\boldsymbol{M}_{3c}$, respectively, align qualitatively with the properties of the corresponding numbered models discussed in the Introduction.

3.1.3 Model summary and parameter priors

The structural properties of the full set of models under consideration for the data are summarized in Table 2. When fitting these models to the mock data we impose priors on the parameters of the component models that are summarized in Table 3. These priors have a combination of physical and experimental motivations (see S23 for details).

3.1.4 Mock data

We generate mock data, $\boldsymbol{D}$, of the form given in equation (4), using $\boldsymbol{M}_a$ and $\boldsymbol{M}_{1b}$ as GMs for $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ such that $\boldsymbol{M}_{1c}$ (for a specific set of model parameters) is the true GM of the data.

We generate signals (A) and (B) as $\boldsymbol{S}_a = \boldsymbol{M}_a(\boldsymbol{\Theta}_a)$ and $\boldsymbol{S}_b = \boldsymbol{M}_{1b}(\boldsymbol{\Theta}_b)$, respectively. The values of the 21-cm parameter vector, $\boldsymbol{\Theta}_a = [A, \nu_0, w, \tau]$, are listed in the top panel of Table 4. They are selected to be consistent with the 21-cm signal estimates in B18, except for the amplitude of the signal, which we set to a value of 150 mK. The values of the foreground parameter vector, $\boldsymbol{\Theta}_b = [\bar{T}_{\rm m_0}, \beta_0, p_1, p_2, p_3, T_{\rm e}, \tau_0]$ are listed in the middle panel of Table 4. They are derived from the maximum-likelihood fit of $\boldsymbol{M}_{1b}$ to simulations of foreground emission as observed by the Experiment to Detect the Global Epoch of Reionisation Signature (EDGES; e.g. B18) experiment (see S23). $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$ are illustrated in Figs 3(a) and (b), respectively.[21]

To generate the mock data, we add to the signal components zero-mean uncorrelated Gaussian random noise drawn from the distribution $N(\boldsymbol{\Theta}_n)$. The values of the noise parameter vector, $\boldsymbol{\Theta}_n = [\mu_n, \sigma_n]$ are listed in the middle panel of Table 4, and are chosen to produce a noise realization with properties comparable to those estimated for the publicly available EDGES-low data from B18 (e.g. Singh & Subrahmanyan 2019).

[21] We note that this definition of $\boldsymbol{S}_a$ makes equation (14) a definitive model of the SOI for the purposes of fitting the mock data in this work (see Section 2.2.5, condition 2).

Table 4. True values of the 21-cm signal parameters used to generate $\boldsymbol{S}_{\rm a}$, the foreground parameters used to generate $\boldsymbol{S}_{\rm b}$, and noise parameters used to generate $\boldsymbol{n}$. The GM of the signal components is noted in the component descriptions.

Data component	Component description	Parameter	Value	Parameter description
$\boldsymbol{S}_{\rm a}$	21-cm signal	A	150 mK	Absorption depth
	[GM: equation (14)]	ν_0	78 MHz	Central frequency
		w	19 MHz	Width
		τ	8	Flattening parameter
$\boldsymbol{S}_{\rm b}$	Foreground emission	$\bar{T}_{\rm m_0}$	1627.46 K	Foreground brightness temperature at $\nu = \nu_{\rm c}$
	[GM: equation (16)]	β_0	2.488	Mean temperature spectral index
		p_1	−0.080	First-order spectral fluctuations fractional amplitude
		p_2	−0.002	Second-order spectral fluctuations fractional amplitude
		p_3	0.013	Third-order spectral fluctuations fractional amplitude
		$T_{\rm e}$	117.31 K	Ionospheric electron temperature
		τ_0	0.009	Ionospheric optical depth at $\nu = \nu_{\rm c}$
$\boldsymbol{n}$	Noise	$\mu_{\rm n}$	0 mK	Noise mean
	[GM: $N(\mu_{\rm n}, \sigma_{\rm n})$]	$\sigma_{\rm n}$	20 mK	Noise standard deviation

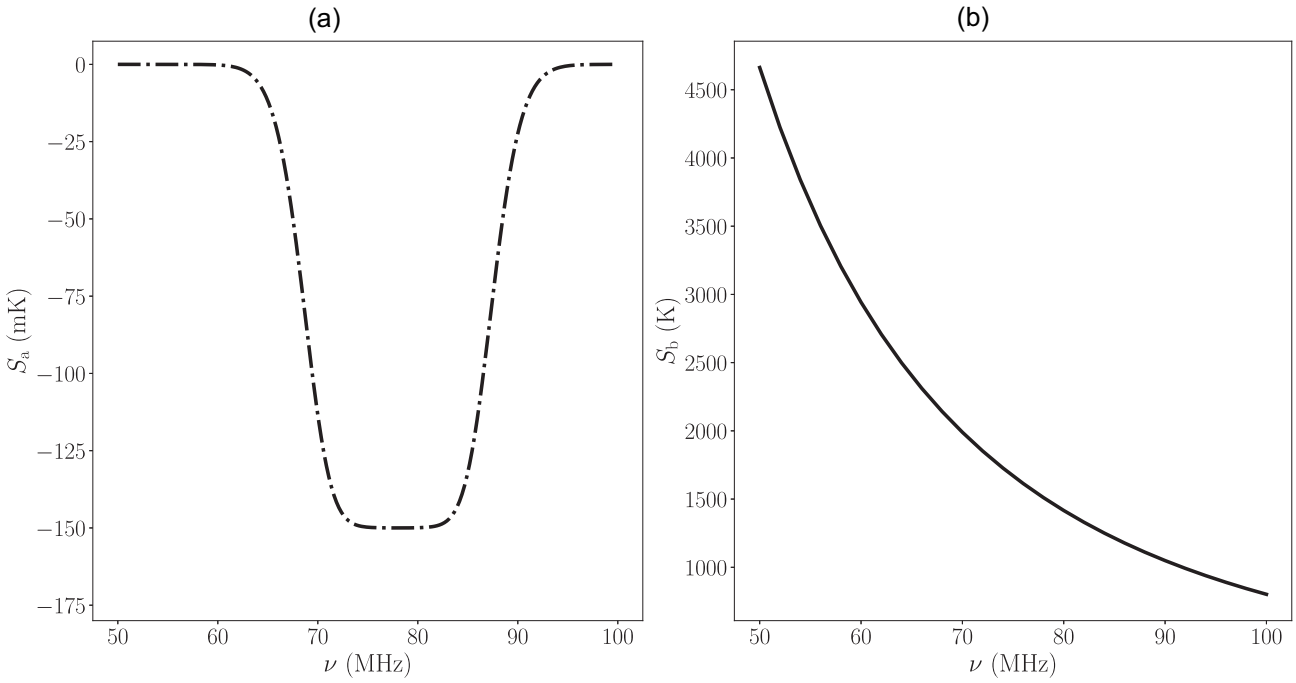

Figure 3. The signal components of the mock data. Panel (a): signal component $\boldsymbol{S}_{\rm a}$ – a flattened-Gaussian absorption trough with an amplitude of 150 mK and central frequency of 78 MHz. Panel (b): signal component $\boldsymbol{S}_{\rm b}$ – radio foreground emission in beam-factor chromaticity corrected spectrometer data. Note the difference in units on the vertical axes of the two panels. The amplitude of $\boldsymbol{S}_{\rm b}$ at 78 MHz is $\sim$ 1500 K, which is 4 orders of magnitude brighter than the peak amplitude of $\boldsymbol{S}_{\rm a}$ at this frequency. As a result, spectral structure in $\boldsymbol{S}_{\rm b}$ unmodelled at the level of 1 part in 10^4 by $\boldsymbol{M}_{i\rm b}$ is sufficient to leave systematic structure comparable in amplitude to $\boldsymbol{S}_{\rm a}$.

3.1.5 Including the true generative model in the set of models under consideration

In practice, all statistical models that are fit to observational data are approximate (e.g. Box 1976); thus, having one of the models under consideration being the true GM is an idealization. However, this test case has a number of benefits over a comparison involving only approximate models:

(i) it illustrates that category II model comparison scenarios can arise in this limit and thus that they are not a by-product of comparing competing composite models for the data in which all models in the set under consideration are approximate,

(ii) with this test-case, the preferred model for the data that one aims to identify with an ideal model comparison framework is evident a priori, simplifying the evaluation of the efficacy of the model validation framework and subsequent posterior-odds-based comparison of the models under consideration.

Furthermore, a situation in which all models in the set under consideration are approximate can be expected to converge on that considered here, in the limit that at least a subset models for the data accurately describe it at a given signal-to-noise level. The test case is thus also illustrative of this common situation.[22]

[22]In Sims et al. (2025), we demonstrate that this expectation is borne out, in the context of fitting models to simulated global 21-cm signal data, when all models of the data under consideration are approximate.

3.1.6 Validation data

We construct validation data, $\boldsymbol{D}_{\rm v}$, of the form given in equation (12) that contains only the nuisance signal $\boldsymbol{S}_{\rm b}$. This is generated in the same manner as the mock data except, in this case, omitting $\boldsymbol{S}_{\rm a}$ from the data vector.

3.1.7 Data likelihood

We include uncorrelated Gaussian white noise in our simulated data. We correspondingly model its covariance matrix, $\mathbf{N}$, as diagonal, with elements given by:

$$N_{ij} = \left\langle n_i n_j^* \right\rangle \delta_{ij} \sigma^2 \, . \tag{18}$$

Here, $\langle .. \rangle$ represents the expectation value and $\sigma = 20$ mK.

Given this noise covariance model and defining a vectorized model for the signal, $\boldsymbol{m}$, constructed from a vector of parameters $\boldsymbol{\Theta}$ and a residuals vector between the data and model given by $\boldsymbol{r} = \boldsymbol{D} - \boldsymbol{m}(\boldsymbol{\Theta})$ we can specify a Gaussian likelihood for the simulated data as,

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{\sqrt{(2\pi)^{N_{\rm chan}} \det(\mathbf{N})}} \exp\left[-\frac{1}{2}\boldsymbol{r}(\boldsymbol{\Theta})^T \mathbf{N}^{-1} \boldsymbol{r}(\boldsymbol{\Theta})\right] \, . \tag{19}$$

We make use of this likelihood when fitting the validation and mock data in Sections 3.4 and 3.5, respectively.

3.2 Signal detection

Bayesian model comparison provides a means by which the detectability of $\boldsymbol{S}_{\rm a}$ can be assessed. Let us assume that we have a predictive composite model for the data of the form $\boldsymbol{M}_{i\rm c} = \boldsymbol{M}_{\rm a} + \boldsymbol{M}_{i\rm b}$, where $\boldsymbol{M}_{i\rm c}$ is an element of the set ABC. For a data set known to contain $\boldsymbol{S}_{\rm b}$, we can ascertain the evidence in favour of a detection of $\boldsymbol{S}_{\rm a}$ using BFBMC of $\boldsymbol{M}_{i\rm c}$ and $\boldsymbol{M}_{i\rm b}$ as models for the data. Given that $\boldsymbol{M}_{i\rm b}$ is a subcomponent of $\boldsymbol{M}_{i\rm c}$, a fit of $\boldsymbol{M}_{i\rm c}$ to the data will necessarily have a likelihood equal or higher than that of $\boldsymbol{M}_{i\rm b}$. However, by using the difference in Bayesian evidence of the models as the comparison metric, BFBMC of $\boldsymbol{M}_{i\rm c}$ and $\boldsymbol{M}_{i\rm b}$ will yield a preference for $\boldsymbol{M}_{i\rm c}$ over $\boldsymbol{M}_{i\rm b}$ only when $\boldsymbol{M}_{i\rm c}$ provides a fit that is improved sufficiently to counterbalance the Occam penalty associated with its additional complexity relative to $\boldsymbol{M}_{i\rm b}$.

We define the threshold for a detection of $\boldsymbol{S}_{\rm a}$, as $\ln(\mathcal{B}_{\rm cb}) \geq 3$, where $\ln(\mathcal{B}_{\rm cb})$ is as defined in equation (13) except, in this case, for the observational rather than validation data. This corresponds to odds in favour of $\boldsymbol{M}_{i\rm c}$ over $\boldsymbol{M}_{i\rm b}$ of 20-to-1 or better (see Table 1) and is consistent with guidelines established by Kass & Raftery (1995) for strong evidence in favour of one model relative to another.[23]

In the general case where it is unknown whether $\boldsymbol{M}i\rm c$ is an element of ABC, a preference for $\boldsymbol{M}i\rm c$ over $\boldsymbol{M}i\rm b$ indicates that $\boldsymbol{S}\rm a$ is either detected in the data or that a spurious detection has occurred due to the presence of systematics.[24] Model validation is crucial for distinguishing between these two possibilities.

[23] In principle, one could also adopt a simulation-based inference approach for this task. For example, one could assess signal detection using a neural binary classifier (e.g. Saxena et al. 2024), trained on simulated examples of detections and non-detections. This approach offers flexibility and the ability to optimize the detection metric for signals of the sort expected in the data; however, it also requires additional computational resources and careful validation to ensure robustness. We leave this as a direction for future exploration.

[24] We use the term 'systematics' in a general sense to refer to any residual structure in the data resulting from imperfect modelling of $\boldsymbol{S}\rm b$ by $\boldsymbol{M}i\rm b$. This can include unmodelled instrumental, environmental, or astrophysical effects, or a combination of these factors.

Table 5. Validation null test results. Bayes factors between $\boldsymbol{M}_{i\rm c}$ and $\boldsymbol{M}_{i\rm b}$, as models for the validation data, where i runs over the three model classes defined in Section 3.1. A positive $\ln(\mathcal{B}^{\rm v}_{\rm cb})$ means that $\boldsymbol{M}_{i\rm c}$ is preferred over $\boldsymbol{M}_{i\rm b}$ as a model of $\boldsymbol{D}_{\rm v}$. The reverse is true when $\ln(\mathcal{B}^{\rm v}_{\rm cb})$ is negative.

Model class	$\ln(\mathcal{B}^{\rm v}_{\rm cb})$	Comment
$\mathcal{M}_1$	−3.2	Null test passed
$\mathcal{M}_2$	184.6	Spurious detection – null test failed
$\mathcal{M}_3$	2997.2	Spurious detection – null test failed

3.3 Computational techniques

In Sections 3.4 and 3.5, we estimate model evidences and sample from the posteriors of model parameters, given the data, using nested sampling as implemented by the PolyChord algorithm (Handley, Hobson & Lasenby 2015b, a). Given samples from the posterior distribution of the parameters, $\mathcal{P}(\boldsymbol{\Theta}|\boldsymbol{D}, \boldsymbol{M})$, we can estimate the posterior predictive density (posterior PD), $\mathcal{P}(y|\boldsymbol{\Theta}, \nu, \boldsymbol{D}, \boldsymbol{M})$, for a function $y = f(\boldsymbol{\Theta}, \nu)$. This is achieved by calculating the set of samples that correspond to $\mathcal{P}(y|\boldsymbol{\Theta}, \nu, \boldsymbol{D}, \boldsymbol{M})$. We generate contour plots of prior and posterior PDs using the fgivenx software package (Handley 2018).

In our analysis, the foreground and composite models are nested. The Savage–Dickey ratio provides a computationally efficient method for calculating the Bayes factor between nested models using the prior and posterior distributions of the most complex model, $\boldsymbol{M}_{1\rm c}$ (e.g. Wagenmakers et al. 2010). In this approach, one calculates the ratio of the posterior and prior densities evaluated at zero of the components that the complex model has in addition to the nested model. However, for the purposes of confirming model accuracy and BaNTER model classification predictions, we require not only the Bayes factors but also the parameter posteriors of the individual models. Using nested sampling, we obtain both the parameter posteriors and Bayesian evidences; consequently, we do not explicitly employ the Savage–Dickey ratio approach in this work. Nevertheless, we note that it can be an efficient method for performing BaNTER validation and BFBMC when one's sole objective is to use these tools for validated Bayesian model comparison, without the additional objective of demonstrating the efficacy of the framework.

3.4 BaNTER model validation results

3.4.1 Null test

Table 5 lists values of $\ln(\mathcal{B}^{\rm v}_{\rm cb})$ – the Bayes factors between $\boldsymbol{M}_{i\rm c}$ and $\boldsymbol{M}_{i\rm b}$ as models for the validation data, $\boldsymbol{D}_{\rm v}$ – calculated for the three model classes defined in Section 3.1. For positive $\ln(\mathcal{B}^{\rm v}_{\rm cb})$, $\boldsymbol{M}_{i\rm c}$ is preferred over $\boldsymbol{M}_{i\rm b}$. Under the assumption that $\boldsymbol{M}_{i\rm c}$ is an accurate model for $\boldsymbol{D}_{\rm v}$ this would constitute evidence for a detection of $\boldsymbol{S}_{\rm a}$; however, since $\boldsymbol{S}_{\rm b}$ is the only signal component that $\boldsymbol{D}_{\rm v}$ contains, any detection of $\boldsymbol{S}_{\rm a}$ in the validation data is necessarily spurious.

For $\mathcal{M}_1$, $\ln(\mathcal{B}^{\rm v}_{\rm cb}) = -3.2$, which corresponds to a strong preference (see Table 1) for $\boldsymbol{M}_{1\rm b}$, relative to $\boldsymbol{M}_{1\rm c}$, as a model for the validation data. As such, $\mathcal{M}_1$ passes the model validation test. In contrast, $\boldsymbol{M}_{i\rm c}$ is decisively preferred over $\boldsymbol{M}_{i\rm b}$ for both $\mathcal{M}_2$ and $\mathcal{M}_3$ ($\ln(\mathcal{B}^{\rm v}_{\rm cb}) = 184.6$ and 2997.2, respectively). Correspondingly, $\mathcal{M}_2$ and $\mathcal{M}_3$ fail the validation null test, and we can identify $\boldsymbol{M}_{2\rm c}$ and $\boldsymbol{M}_{3\rm c}$ as elements of either $A\overline{B}C$ or $A\overline{B}\,\overline{C}$.

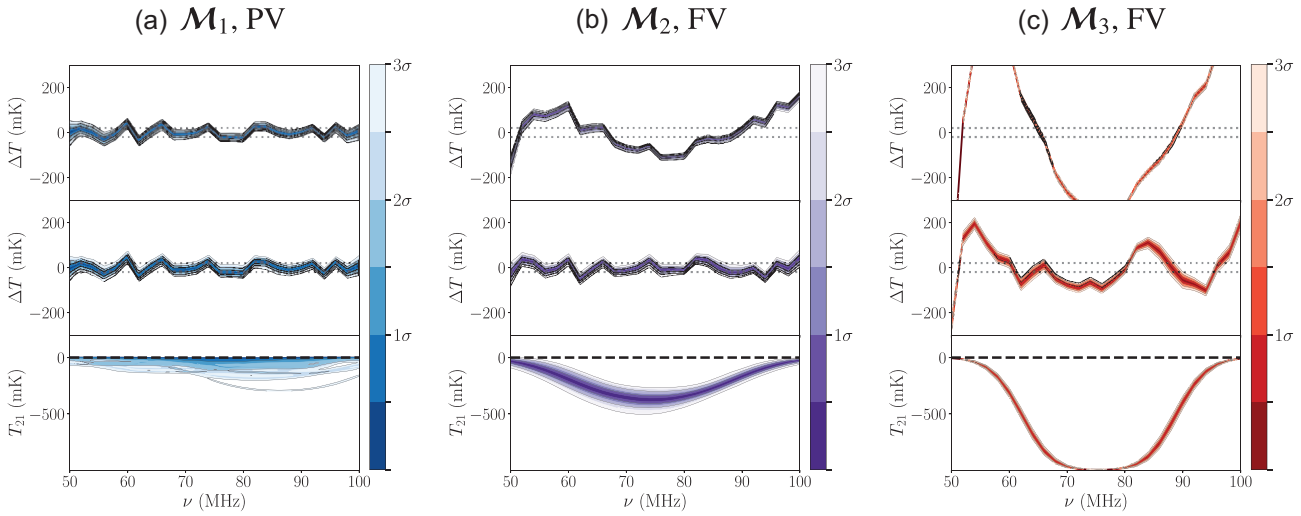

Figure 4. Validation test results. Posterior predictive densities of the foreground fit residuals $\boldsymbol{r}_{i\mathrm{b}} = [\boldsymbol{D}_\mathrm{v} - \boldsymbol{M}_{i\mathrm{b}}(\boldsymbol{\Theta}_{i\mathrm{b}})]$ (top panels), of the composite model's fit residuals $\boldsymbol{r}_{i\mathrm{c}} = [\boldsymbol{D}_\mathrm{v} - \boldsymbol{M}_{i\mathrm{c}}(\boldsymbol{\Theta}_{i\mathrm{c}})]$ (middle panels), and of the model of signal (A) recovered with the composite model ($\boldsymbol{M}_\mathrm{a}(\boldsymbol{\Theta}_{i\mathrm{a}})$; bottom panels) for model classes $i = 1$, 2, and 3 (subfigures a, b, and c, respectively). The dotted lines in the top and middle panels denote the noise level in the simulated data. The dashed black line shows the input 21-cm signal, $\boldsymbol{S}_\mathrm{a}$, in the mock data. The subfigure captions indicate whether the model class passed model validation (PV) or failed (FV).

The null test results can be understood intuitively by examining Fig. 4, which shows the posterior PDs of the foreground fit residuals to the validation data $\boldsymbol{r}_{i\mathrm{b}} = [\boldsymbol{D}_\mathrm{v} - \boldsymbol{M}_{i\mathrm{b}}(\boldsymbol{\Theta}_{i\mathrm{b}})]$ (top panels), of the composite model's fit residuals $\boldsymbol{r}_{i\mathrm{c}} = [\boldsymbol{D}_\mathrm{v} - \boldsymbol{M}_{i\mathrm{c}}(\boldsymbol{\Theta}_{i\mathrm{c}})]$ (middle panels), and of $\boldsymbol{M}_\mathrm{a}(\boldsymbol{\Theta}_{i\mathrm{a}})$, the model of signal (A) recovered with the composite model (bottom panels) for composite/nuisance model sets $i = 1$, 2, and 3 (Figs 4a–c, respectively). Here, $\boldsymbol{\Theta}_{i\mathrm{b}}$ and $\boldsymbol{\Theta}_{i\mathrm{c}}$ are the posterior distributions of the parameters of the ith foreground [signal (B)] model and composite 21-cm + foreground [signal (A) + signal (B)] model, respectively, derived from fits of those models to the data. $\boldsymbol{\Theta}_{i\mathrm{a}}$ is the marginal posterior distribution of the parameters of the ith 21-cm model derived in the fit of the ith composite model to the data.

Across the simulated spectral band, the posterior PDs of the residuals of both the $\boldsymbol{\mathcal{M}}_1$ foreground and the composite model fits to $\boldsymbol{D}_\mathrm{v}$ are consistent at 95 per cent credibility with the 20 mK expected noise level in the data and in both the root mean square (RMS) value of the median residual is 17 mK. The posterior PD of the 21-cm signal associated with the fit of $\boldsymbol{M}_{1\mathrm{c}}$ has an amplitude consistent with zero across the band, implying its signal component is adding superfluous complexity in the context of this data set. The $\ln(\mathcal{B}^\mathrm{v}_{\mathrm{cb}}) = -3.2$ log-Bayes-factor found disfavouring $\boldsymbol{M}_{1\mathrm{c}}$ relative to $\boldsymbol{M}_{1\mathrm{b}}$ is qualitatively consistent with these results.

The posterior PDs of the foreground fit residuals of $\boldsymbol{M}_{2\mathrm{b}}$ to $\boldsymbol{D}_\mathrm{v}$ have a median RMS value of $\mathrm{RMS}_\mathrm{med}(\boldsymbol{r}_{2\mathrm{b}}) = 74$ mK and thus are inconsistent with the expected noise level. In contrast, the posterior PDs of the fit residuals of $\boldsymbol{M}_{2\mathrm{c}}$ to $\boldsymbol{D}_\mathrm{v}$ are consistent at 95 per cent credibility with the expected noise level in the data across the band and have a median RMS value of $\mathrm{RMS}_\mathrm{med}(\boldsymbol{r}_{2\mathrm{c}}) = 18$ mK. The posterior PD of the 21-cm signal associated with the fit of $\boldsymbol{M}_{2\mathrm{c}}$ to the validation data has a mean amplitude $A = 381 \pm 36$ mK and central frequency at $\nu_0 = 74 \pm 1$ MHz, corresponding to a high significance spurious detection of the signal in the validation data. Relative to $\boldsymbol{M}_{2\mathrm{b}}$, the significantly improved fit of $\boldsymbol{M}_{2\mathrm{c}}$ to $\boldsymbol{D}_\mathrm{v}$ and the fact that the amplitude of the signal component of the model is highly inconsistent with zero ($\sim 10\sigma$) are qualitatively consistent with the $\ln(\mathcal{B}^\mathrm{v}_{\mathrm{cb}}) = 184.6$ log-Bayes-factor found in favour of $\boldsymbol{M}_{2\mathrm{c}}$ over $\boldsymbol{M}_{2\mathrm{b}}$.

The posterior PDs of both the $\boldsymbol{\mathcal{M}}_3$ foreground and composite model fit residuals to the validation data have median RMS values of 294 and 85 mK, respectively; thus both are inconsistent with the expected noise level in the data. The posterior PD of the 21-cm signal associated with the fit of $\boldsymbol{r}_{3\mathrm{c}}$ to $\boldsymbol{D}_\mathrm{v}$ has a mean amplitude $A = 998 \pm 2$ mK and central frequency at $\nu_0 = 75.5 \pm 0.1$ MHz, corresponding to a very high significance spurious detection of the signal in the validation data. Despite its poor fit in absolute terms, the improvement in the fit of $\boldsymbol{M}_{2\mathrm{c}}$ to $\boldsymbol{D}_\mathrm{v}$, relative to $\boldsymbol{M}_{2\mathrm{b}}$, and the fact that the amplitude of the signal component of the model is highly inconsistent with zero ($\sim 500\sigma$) are qualitatively consistent with the $\ln(\mathcal{B}^\mathrm{v}_{\mathrm{cb}}) = 2997.2$ log-Bayes-factor in favour of $\boldsymbol{M}_{3\mathrm{c}}$ over $\boldsymbol{M}_{3\mathrm{b}}$.

3.4.2 *Euler set predictions*

To predict the Euler sets of the validated composite models, we evaluate the predictivity and accuracy conditions given in equations (10) and (11) for the set of foreground models and the accuracy condition[25] for the composite models for $\boldsymbol{D}_\mathrm{v}$.

Assuming perfect validation data, foreground models that pass both conditions are elements of Euler set B. Assuming our signal model is an element of A, composite models that pass the accuracy condition are likely but not guaranteed to be elements of Euler set C. The lack of guarantee derives from the fact that while composite models with components drawn from A and B will necessarily be elements of ABC those with components drawn from A and $\overline{B}$ may be elements of $A\overline{B}C$ with respect to $\boldsymbol{D}_\mathrm{v}$ while being elements

[25]We do not use the relative predictivities of the composite models for the foreground-only validation data as a useful metric for Euler set prediction since they suffer from the same limitation that necessitates the introduction of the BaNTER null test. Namely, BFBMC imposes an Occam penalty against a composite model composed of elements of A and B if the component A represents unnecessary model complexity (as would be the case when fitting $\boldsymbol{D}_\mathrm{v}$). Thus, there is no guarantee that the aforementioned model will be more predictive than a less complex composite model composed of elements of A and $\overline{B}$ that fits $\boldsymbol{D}_\mathrm{v}$ using spurious fits of its components.

Table 6. Validation predictivity and accuracy condition results for the foreground models. $\mathcal{B}_{j\max} = \mathcal{Z}_j/\mathcal{Z}_{\rm b,max}$, with $\mathcal{Z}_j$ the Bayesian evidence of the jth foreground model for the validation data and $\mathcal{Z}_{\rm b,max}$ the highest Bayesian evidence foreground model for the validation data. $Q_{0.999}(\lambda_{j\rm b})$ is the accuracy condition (see Section 2.4.3) evaluated for the fit of the jth foreground model to the validation data.

Foreground model	$\ln(\mathcal{B}_{j\max})$	$Q_{0.999}(\lambda_{j\rm b})$
$\boldsymbol{M}_{1\rm b}$	0	10
$\boldsymbol{M}_{2\rm b}$	−194	−188
$\boldsymbol{M}_{3\rm b}$	−3331	−3320

of $A\overline{B}\ \overline{C}$ with respect to true data. Specifically, this difference in model classification with respect to the validation and true data will occur if structure in the validation data that is unmodellable by the foreground model can be absorbed by the signal model but the sum of this structure and the true signal in the data cannot.

The results of the predictivity and accuracy conditions for the foreground and composite models are summarized in Table 6. We find that $\boldsymbol{M}_{1\rm b}$ is an element of B and $\boldsymbol{M}_{2\rm b}$ and $\boldsymbol{M}_{3\rm b}$ are elements $\overline{B}$. Calculating the corresponding accuracy metrics for $\boldsymbol{M}_{1\rm c}$, $\boldsymbol{M}_{2\rm c}$, and $\boldsymbol{M}_{3\rm c}$, we find that $\boldsymbol{M}_{1\rm c}$ and $\boldsymbol{M}_{2\rm c}$ pass the accuracy condition with $Q_{0.999}(\lambda_{1\rm c}) = 10$ and $Q_{0.999}(\lambda_{2\rm c}) = 4$, while $\boldsymbol{M}_{3\rm c}$ fails with $Q_{0.999}(\lambda_{3\rm c}) = -309$. Combining these results yields predictions of the Euler sets of the composite models for data of ABC, $A\overline{B}C$, and $A\overline{B}\ \overline{C}$ for $\boldsymbol{M}_{1\rm c}$, $\boldsymbol{M}_{2\rm c}$, and $\boldsymbol{M}_{3\rm c}$, respectively.

3.4.3 BaNTER validation results as model priors

Given BaNTER null test results and the aforementioned classification predictions for $\boldsymbol{M}_{1\rm c}$, $\boldsymbol{M}_{2\rm c}$, and $\boldsymbol{M}_{3\rm c}$, we should anticipate the comparison of $\boldsymbol{M}_{1\rm c}$ and $\boldsymbol{M}_{3\rm c}$ being a category I model comparison problem and comparison of $\boldsymbol{M}_{1\rm c}$ and $\boldsymbol{M}_{2\rm c}$ being a category II model comparison problem.

In principle, given that both $\boldsymbol{\mathcal{M}}_2$ and $\boldsymbol{\mathcal{M}}_3$ have failed the validation null test, we would be justified in judging them not to be credible models with which one could recover unbiased estimates of the 21-cm signal in the mock data. This would leave $\boldsymbol{\mathcal{M}}_1$ as a decisively preferred set of models for the data. However, for the purpose of confirming the above predictions regarding the Euler set in which each model is an element and their model comparison categorizations, in the next section, we perform a full Bayes-factor-based comparison of $\boldsymbol{M}_{1\rm b}$, $\boldsymbol{M}_{2\rm b}$, $\boldsymbol{M}_{3\rm b}$, $\boldsymbol{M}_{1\rm c}$, $\boldsymbol{M}_{2\rm c}$, and $\boldsymbol{M}_{3\rm c}$ as models for the mock data.

3.5 Mock data results

3.5.1 Bayes factors, predictivity, and accuracy

Table 7 lists values of $\ln(\mathcal{B}_{\rm cb})$ – the Bayes factors between $\boldsymbol{M}_{i\rm c}$ and $\boldsymbol{M}_{i\rm b}$ as models for the mock data, $\boldsymbol{D}$ – and $\ln(\mathcal{B}_{i\max})$ – the Bayes factor between the highest Bayesian evidence composite model for $\boldsymbol{D}$ and the remaining models.

We find a decisive preference ($\ln(\mathcal{B}_{\rm cb}) \geq 5$) for $\boldsymbol{M}_{i\rm c}$ over $\boldsymbol{M}_{i\rm b}$ in all three composite/nuisance model sets. In the absence of validation results, this implies decisive detections of either $\boldsymbol{S}_{\rm a}$ or systematic structure fit by the 21-cm model.[26]

[26]Models $\boldsymbol{M}_{2\rm c}$ and $\boldsymbol{M}_{3\rm c}$ failed validation, indicating these models will yield biased 21-cm signal inferences. As such, the detections of $\boldsymbol{S}_{\rm a}$ with these models cannot be treated as definitive. Only $\boldsymbol{M}_{1\rm c}$, which passes validation, provides robust evidence for $\boldsymbol{S}_{\rm a}$ in the data.

Table 7. Mock data results. For each model class we list the Bayes factors in favour of a detection of $\boldsymbol{S}_{\rm a}$ in the mock data, $\ln(\mathcal{B}_{\rm cb})$, and between the highest Bayesian evidence composite model for the mock data and the remaining models, $(\ln(\mathcal{B}_{i\max}))$. We also note the corresponding validation test results for each model class (see Table 5) and the Euler set in which each composite model in each model set is an element, as inferred from combining the validation test results and $\ln(\mathcal{B}_{i\max})$. We note that the true Euler sets of the composite models confirm the predicted Euler sets in Section 3.4 so we do not list them separately.

Model class	$\ln(\mathcal{B}_{\rm cb})$	$\ln(\mathcal{B}_{i\max})$	Validation test results	True/predicted model classifications
$\boldsymbol{\mathcal{M}}_1$	18.5	−1.6	Passed	ABC
$\boldsymbol{\mathcal{M}}_2$	523.5	0	Failed	$A\overline{B}C$
$\boldsymbol{\mathcal{M}}_3$	3910.2	−431.3	Failed	$A\overline{B}\ \overline{C}$

Comparing models with BFBMC alone yields $\boldsymbol{M}_{2\rm c}$ as the highest Bayesian evidence model for the data. However, the Bayes factor $\ln(\mathcal{B}_{12}) = -1.6$ implies only a moderate preference for $\boldsymbol{M}_{2\rm c}$ over $\boldsymbol{M}_{1\rm c}$, while $\boldsymbol{M}_{3\rm c}$ is decisively disfavoured relative to both.

Calculating the accuracy condition given by equation (11) for $\boldsymbol{M}_{1\rm c}$, $\boldsymbol{M}_{2\rm c}$, and $\boldsymbol{M}_{3\rm c}$ as models for the mock data, we find $\boldsymbol{M}_{1\rm c}$ and $\boldsymbol{M}_{2\rm c}$ pass with $Q_{0.999}(\lambda_{1\rm c}) = 9$ and $Q_{0.999}(\lambda_{2\rm c}) = 7$ and $\boldsymbol{M}_{3\rm c}$ fails with $Q_{0.999}(\lambda_{3\rm c}) = -412$. Thus, $\boldsymbol{M}_{3\rm c}$ is an element of $\overline{C}$ and, consistent with its low evidence, is not a credible model with which unbiased inferences of the 21-cm signal can be made. In contrast, $\boldsymbol{M}_{1\rm c}$ and $\boldsymbol{M}_{2\rm c}$ pass the predictivity and accuracy conditions given by equations (10) and (11). In the absence of the BaNTER validation analysis, this would limit them to being elements either of ABC or $A\overline{B}C$, and, thus, it does not guarantee that unbiased inferences of the 21-cm signal are recoverable using them.

3.5.2 Signal recovery using Bayesian workflow with uninformative model priors

Fig. 5 shows posterior PDs derived from fits of the models to the mock data. Analysis of these results shows the following:

(i) $\boldsymbol{M}_{1\rm c}$. Accurately fits the mock data (Fig. 5a, middle panel) and yields unbiased inferences of the 21-cm signal (Fig. 5a, bottom panel).

(ii) $\boldsymbol{M}_{2\rm c}$. Accurately fits the mock data (Fig. 5b, middle panel) but results in biased inferences of the 21-cm signal (Fig. 5b, bottom panel).

(iii) $\boldsymbol{M}_{3\rm c}$. Fails to fit the mock data (Fig. 5c, middle panel) and produces biased inferences of the 21-cm signal (Fig. 5c, bottom panel).

The posterior odds of $\boldsymbol{M}_{1\rm c}$, $\boldsymbol{M}_{2\rm c}$, and $\boldsymbol{M}_{3\rm c}$ are determined by the product of prior odds and Bayes factors between them. Under the Bayesian workflow with uninformative model priors (Fig. 2), equal prior odds are assumed, and BFBMC alone is used to calculate model weights based on observational data. This assigns approximately 83 per cent weight to $\boldsymbol{M}_{2\rm c}$ and 17 per cent to $\boldsymbol{M}_{1\rm c}$, with $\boldsymbol{M}_{3\rm c}$ contributing negligibly (see equation 2 and Table 1). The high weighting of $\boldsymbol{M}_{2\rm c}$ results in significantly biased Bayesian model-averaged 21-cm signal inferences (with the averaged 21-cm signal being similar to Fig. 5b, bottom panel).

The limitation of BFBMC-alone arises because it compares model predictivity for the data in aggregate. This does not distinguish

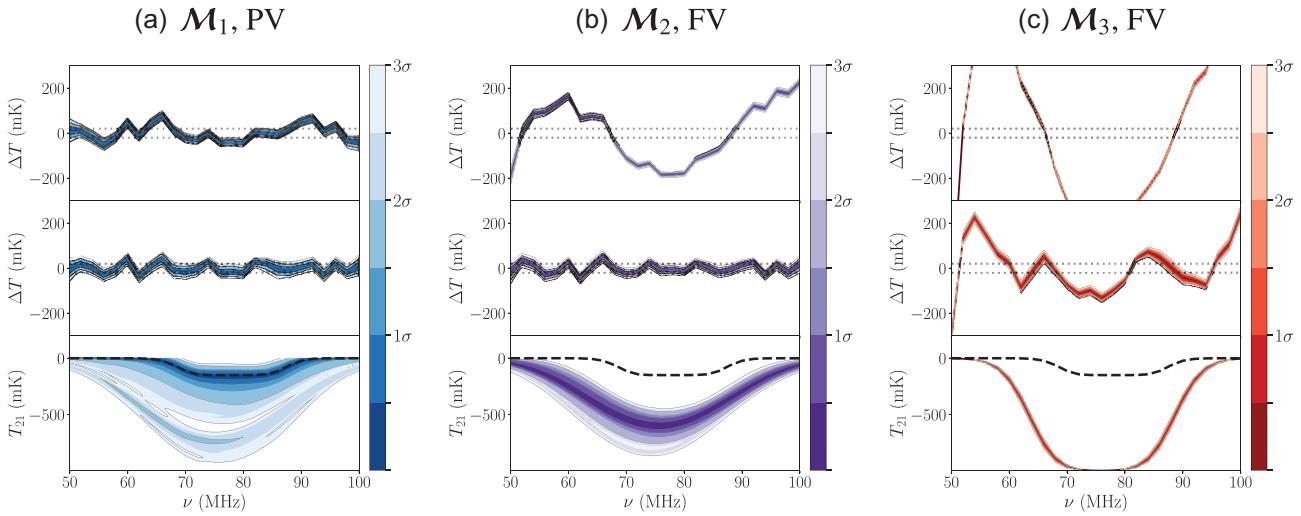

Figure 5. Mock data results. Posterior predictive densities of the foreground fit residuals $\boldsymbol{r}_{i\mathrm{b}} = [\boldsymbol{D} - \boldsymbol{M}_{i\mathrm{b}}(\boldsymbol{\Theta}_{i\mathrm{b}})]$ (top panels), composite models fit residuals $\boldsymbol{r}_{i\mathrm{c}} = [\boldsymbol{D} - \boldsymbol{M}_{i\mathrm{c}}(\boldsymbol{\Theta}_{i\mathrm{c}})]$ (middle panels) and of the model of signal (A) recovered with the composite model ($\boldsymbol{M}_{\mathrm{a}}(\boldsymbol{\Theta}_{i\mathrm{a}})$; bottom panels) for model classes $i = 1$, 2, and 3 (subfigures a, b and c, respectively). The dotted lines in the top and middle panels denote the noise level in the simulated data. The dashed black line shows the input 21-cm signal, $\boldsymbol{S}_{\mathrm{a}}$, in the mock data. The subfigure captions indicate whether the model class passed model validation (PV) or failed (FV).

between a model in ABC, with a high Bayesian evidence due to accurate fits of the model components, as with $\boldsymbol{M}_{1\mathrm{c}}$, and a model in $A\overline{B}C$, with biased fits of the model components that still produce good overall predictions, as with $\boldsymbol{M}_{2\mathrm{c}}$.

3.5.3 Signal recovery using BaNTER-validated Bayesian inference workflow

Following the model-validated Bayesian inference workflow (Fig. 2), we update the prior odds we assign to the models based on the results of the BaNTER validation. Both $\boldsymbol{M}_{2\mathrm{c}}$ and $\boldsymbol{M}_{3\mathrm{c}}$ fail the BaNTER null test, indicating that their inferred 21-cm signals are likely to be biased even if they achieve a high Bayesian evidence and/or accurate fit to the data in aggregate (as is the case with $\boldsymbol{M}_{2\mathrm{c}}$). As can be seen from Figs 5(b) and (c), the results of the mock data analysis confirm this to be the case.

In contrast, $\boldsymbol{M}_{1\mathrm{c}}$ passes the accuracy and predictivity conditions and BaNTER validation, indicating that unbiased estimates of the 21-cm signal are recoverable with this model. The results of the mock data analysis in Fig. 5(a) confirm the predicted result, recovering a posterior PD of the 21-cm signal consistent with the true 21-cm signal in the data.

Since $\boldsymbol{M}_{2\mathrm{c}}$ and $\boldsymbol{M}_{3\mathrm{c}}$ fail BaNTER validation, their prior odds can be approximated as negligible, leaving $\boldsymbol{M}_{1\mathrm{c}}$ as the only credible model for unbiased signal estimation. Thus, the ultimate result of the model-validated Bayesian inference workflow is the recovery of unbiased 21-cm signal inferences with $\boldsymbol{M}_{1\mathrm{c}}$ (Fig. 5a, bottom panel).

3.5.4 Comparison of mock data results and validation analysis predictions

From the results in Sections 3.5.1–3.5.3, one can conclude the following:

(i) $\boldsymbol{M}_{2\mathrm{b}}$ and $\boldsymbol{M}_{3\mathrm{b}}$ are insufficiently accurate models of $\boldsymbol{S}_{\mathrm{b}}$ to yield unbiased estimates of the 21-cm signal when they are jointly fit to the data with $\boldsymbol{M}_{\mathrm{a}}$,

(ii) in contrast, $\boldsymbol{M}_{1\mathrm{b}}$ is a sufficiently accurate model of $\boldsymbol{S}_{\mathrm{b}}$ to yield unbiased estimates of the 21-cm signal when jointly fit to the data with $\boldsymbol{M}_{\mathrm{a}}$,

(iii) the true model classifications of $\boldsymbol{M}_{1\mathrm{c}}$, $\boldsymbol{M}_{2\mathrm{c}}$, and $\boldsymbol{M}_{3\mathrm{c}}$ are ABC, $A\overline{B}C$, and $A\overline{B}\,\overline{C}$, respectively,

(iv) comparison of $\boldsymbol{M}_{1\mathrm{c}}$ and $\boldsymbol{M}_{3\mathrm{c}}$ is a category I model comparison problem that is successfully solved by the Bayesian workflow with uninformative model priors (see Fig. 2),

(v) comparison of $\boldsymbol{M}_{1\mathrm{c}}$ and $\boldsymbol{M}_{2\mathrm{c}}$ is a category II model comparison problem. In this case, the Bayesian workflow with uninformative model priors yields biased inferences of $\boldsymbol{S}_{\mathrm{a}}$, whereas unbiased inferences of $\boldsymbol{S}_{\mathrm{a}}$ are recovered by using the BaNTER validation results in the model-validated Bayesian inference workflow.

These observations are fully consistent with the expected outcomes based on the results of the validation analysis (see Section 3.4).

4 DISCUSSION

4.1 Nested models

The flexible nature and utility of nested or partially nested models has led to their widespread usage in various fields spanning psychology and finance (e.g. Ratcliff & McKoon 2008; Fama & French 1993) to ecology and astronomy (e.g. Elith & Leathwick 2009; Sims et al. 2016; Sims & Pober 2019; Sims, Pober & Sievers 2022a, b; Burba, Sims & Pober 2023). Their ability to account for the a priori unknown complexity of foreground emission and uncertain presence of systematics has also led to their use in global 21-cm cosmology (e.g. B18, H18, Tauscher et al. 2018; Hibbard et al. 2020; Rapetti et al. 2020; Sims & Pober 2020; Anstey, de Lera Acedo & Handley 2021; Shen et al. 2022; S23, Saxena et al. 2023; Pagano et al. 2024; Pattison, Anstey & de Lera Acedo 2024).

The set of models, $\boldsymbol{\mathcal{M}} = \{\boldsymbol{M}_{1\mathrm{c}}, \boldsymbol{M}_{1\mathrm{b}}, \boldsymbol{M}_{2\mathrm{c}}, \boldsymbol{M}_{2\mathrm{b}}, \boldsymbol{M}_{3\mathrm{c}}, \boldsymbol{M}_{3\mathrm{b}}\}$, considered in this work are nested. Specifically, the parameters of $\boldsymbol{M}_{1\mathrm{b}}$, $\boldsymbol{M}_{2\mathrm{c}}$, $\boldsymbol{M}_{2\mathrm{b}}$, $\boldsymbol{M}_{3\mathrm{c}}$, and $\boldsymbol{M}_{3\mathrm{b}}$ are a subset of the parameters of $\boldsymbol{M}_{1\mathrm{c}}$ and the parameters of $\boldsymbol{M}_{2\mathrm{b}}$ and $\boldsymbol{M}_{3\mathrm{b}}$ are a subset of the parameters of $\boldsymbol{M}_{1\mathrm{b}}$. Thus, given that we use $\boldsymbol{M}_{1\mathrm{c}}$ as the GM for the data, the nested structure of the other two nuisance models mean that $\boldsymbol{M}_{2\mathrm{b}}$

and $\boldsymbol{M}_{3\mathrm{b}}$ will, in general, provide incomplete descriptions of the validation data. Thus, in some respects, one may find it unsurprising that $\mathcal{M}_2$ and $\mathcal{M}_3$ fail BaNTER validation. However, we note that $\mathcal{M}_2$ and $\mathcal{M}_3$ will fail BaNTER validation only if the errors due to the imperfect modelling of the validation data by $\boldsymbol{M}_{2\mathrm{b}}$ and $\boldsymbol{M}_{3\mathrm{b}}$ are accurately modellable by $\boldsymbol{M}_{\mathrm{a}}$. Thus, their failure is not a foregone conclusion; rather, it derives from the properties of the $\boldsymbol{M}_{1\mathrm{b}}$ and $\boldsymbol{M}_{1\mathrm{c}}$ model incompletenesses and the structure of $\boldsymbol{M}_{\mathrm{a}}$, in combination.

Additionally, we note that while nested models were selected to construct illustrative categories I and II model comparison problems in this work, category II model comparison problems will arise whenever the set of models under consideration include elements of both ABC and $A\overline{B}C$ (see Section 2.2.5), regardless of whether the nuisance components of the models are nested. Category II model comparison problems with non-nested nuisance structure models are considered further in Sims et al. (2025).

4.2 Other modelling scenarios – overly complex models that pass model validation

In this work, we used $\mathcal{M}_1$ – the most complex of the three model classes under consideration – as our GM for the mock data, with $\mathcal{M}_2$ and $\mathcal{M}_3$ containing a subset of the structural components present in $\mathcal{M}_1$. A question of interest, not answered by the analysis considered thus far is: how would models that have a superset of the components of $\mathcal{M}_1$ behave in the validation and data analysis?

Intuitively, one would expect such models to pass model validation but to be disfavoured at the data analysis stage. This disfavouring occurs due to the Occam penalty associated with the superfluous complexity and reduced predictivity of said models not being compensated by an improved fit to the data. To test this scenario, we consider an $\mathcal{M}_4$ model class containing models defined identically to those in $\mathcal{M}_1$ (see Section 3.1) except that the foreground component $\boldsymbol{M}_{4\mathrm{b}}$ has an additional instrument–foreground coupling term ($N_{\mathrm{pert}} = 4$) relative to the GM. Results derived from adding this model to the validation and data model comparison analyses conducted in Sections 3.4 and 3.5 confirm the above prediction. Specifically, we find the following.

(i) $\mathcal{M}_4$ passes the model validation test with $\ln(\mathcal{B}^{\mathrm{v}}_{\mathrm{cb}}) = -3.5$, which corresponds to a strong preference (see Table 1) for $\boldsymbol{M}_{4\mathrm{b}}$, relative to $\boldsymbol{M}_{4\mathrm{c}}$, as a model for the validation data.

(ii) A fit of $\boldsymbol{M}_{4\mathrm{c}}$ to $\boldsymbol{D}$ yields a comparable recovery of the 21-cm signal to that obtained with $\boldsymbol{M}_{1\mathrm{c}}$. However, BFBMC of the models for the data yields $\ln(\mathcal{B}_{4\mathrm{max}}) = -3.7$ and $\ln(\mathcal{B}_{41}) = -2.1$. Thus, $\boldsymbol{M}_{4\mathrm{c}}$ is strongly and moderately disfavoured relative to $\boldsymbol{M}_{2\mathrm{c}}$ and $\boldsymbol{M}_{1\mathrm{c}}$, respectively, with the latter statistic representing the Occam penalty associated with the superfluous complexity of the nuisance structure component of the model.

4.3 Model validation challenges and mitigation approaches

4.3.1 Validation data fidelity

The efficacy of the model validation methodology introduced in this work is dependent on the fidelity of the model validation data. As such, if the nuisance component of the data is, for example, time-dependent,[27] care must be taken to ensure acquisition of the $\boldsymbol{S}_{\mathrm{b}}$-only observations required for $\boldsymbol{D}_{\mathrm{v}}$ and the observational data, $\boldsymbol{D}$, in a concomitant manner. Similarly, if it is not experimentally feasible to obtain $\boldsymbol{S}_{\mathrm{b}}$-only observations for $\boldsymbol{D}_{\mathrm{v}}$ and simulated $\boldsymbol{S}_{\mathrm{b}}$-only observations are used instead, the efficacy of the validation analysis will be a function of the accuracy of the simulated observations.

4.3.2 Simulated validation data

When validation simulations imperfectly represent the nuisance structure in the data, two scenarios must be considered: either there may be additional structure in the data not present in the simulations, or the simulations may contain additional structure not present in the data.

If the simulation has additional structure not present in the data models that are sufficient to describe the true nuisance structure in the data may fail the validation test. This could result in less stringent constraints on the SOI due to the nuisance structure model that pass validation being unnecessarily complex. However, these constraints will still be unbiased since models that pass the validation analysis with additional nuisance structure can also be expected to pass the simpler validation null test that would be associated with a more accurate validation data set.

Conversely, if the data contain additional nuisance structure not modelled in the validation simulations, a model that passes validation with the simulated data but fails to describe the additional structure might cause biased final inferences despite passing the validation test. The Bayesian-averaged posterior inference resulting from the model-validated Bayesian inference workflow will, nevertheless, be improved relative to those from the Bayesian workflow with uninformative model priors by the downweighting or removal of the models that fail the validation null test.

4.3.3 Accounting for uncertainties in validation data simulations

Most often, simulations are a simplified version of nature; thus, the second case considered in Section 4.3.2 is more likely than the first. If the primary source of errors in the simulation derives from uncertainties in the correct specification of its input parameters (for example, due to the finite precision with which those parameters can be experimentally constrained), the validation methodology proposed in this work can be updated as follows. To account for the range of structure potentially present in the nuisance component of the data, one can replace the single validation data set considered thus far with a testing set of validation simulations spanning the range of possible nuisance structure consistent with the uncertainties on its input parameters. In this regime, the unperturbed simulation represents a fiducial validation data set that describes the most probable representation of the nuisance structure in the data. However, now, rather than using the fiducial validation data alone, for each model the validation analysis is iterated over all validation data sets. In so doing, the updated validation analysis accounts for other possible realizations of the nuisance structure in the data.

In this updated framework, one might conservatively discard any model that fails validation for any validation data set. Alternatively, if the values of input parameters for the validation data are drawn from a prior distribution one may calculate a prior probability associated with a given validation data set. This prior can then be used to either,

[27] In 21-cm cosmology time-dependence of the astrophysical foregrounds, ionospheric conditions, ambient conditions at the observation site and the instrumental calibration may all contribute to the time dependence of the data.

(i) inform the probability one places in a model that passes or fails the validation analysis with that simulated data, or,

(ii) update the validation threshold, $\ln(\mathcal{B}^{\rm v}_{\rm threshold})$, above which one considers the model to have failed the validation test.[28]

5 SUMMARY AND FUTURE WORK

In this paper, we introduced the concept of categories I and II model comparison problems. A category I scenario arises when all predictive composite models for the data also have predictive component models. In contrast, a category II scenario involves a subset of models that can accurately fit the data but only with biased component models.

BFBMC is a common Bayesian approach to comparing competing models for a data set. The comparison of competing models using BFBMC enables one to identify preferred models for the data in a category I model comparison scenario. However, we demonstrated that in a category II model comparison scenario, it will favour predictive models for the data, regardless of whether the components of those models are accurate.

To address the challenges of deriving unbiased inferences in category II scenarios, we presented the BaNTER validation framework. BaNTER calculates the Bayes factor between composite and single-component models for single-signal validation data. This allows one to identify, a priori, models in which a statistically significant improvement in the accuracy of their fit to the data is liable to result from biased fits of their model components. Single-component models that are favoured over their composite variants are judged to have passed. Those for which the composite model is preferred, implying evidence in favour of a spurious detection of the absent signal component, fail.

We demonstrated the efficacy of this methodology for deriving unbiased inferences from a mock data set in a category II model comparison scenario in a global 21-cm cosmology context. Comparing three competing composite models, $\boldsymbol{M}_{1c}$, $\boldsymbol{M}_{2c}$, and $\boldsymbol{M}_{3c}$, for a data set where $\boldsymbol{M}_{1c}$ is the GM for the data, $\boldsymbol{M}_{2c}$ accurately fits the data in aggregate but only with biased estimates of the foregrounds and 21-cm signal, and $\boldsymbol{M}_{3c}$ is a less predictive model of the data than $\boldsymbol{M}_{1c}$ and $\boldsymbol{M}_{2c}$, we demonstrated the following.

(i) BFBMC successfully resolves the category I problem by decisively favouring $\boldsymbol{M}_{1c}$ over $\boldsymbol{M}_{3c}$.

(ii) The Bayesian evidence of $\boldsymbol{M}_{1c}$ and $\boldsymbol{M}_{2c}$ are comparable because each provides a comparable fit to the data over their respective prior volumes. Consequently, a Bayesian workflow with uninformative model priors in which BFBMC alone is used to derive the posterior odds of the models, fails to solve the category II model comparison problem of identifying $\boldsymbol{M}_{1c}$ as superior to $\boldsymbol{M}_{2c}$. Consequently, the Bayesian-model-averaged 21-cm signal inferences derived from this analysis are biased.

(iii) Applying BaNTER validation to the three models classes we find that $\boldsymbol{M}_{1c}$ passes. In contrast, $\boldsymbol{M}_{2c}$ and $\boldsymbol{M}_{3c}$ fail the null test due to spurious detections of a 21-cm signal in the foreground-only validation data.

(iv) Incorporating the BaNTER validation results to derive informative model priors used in a model-validated Bayesian inference workflow enables the derivation of unbiased inferences of the 21-cm signal in the data.

In Sims et al. (2025), we use BaNTER and the model-validated Bayesian inference framework, developed in this work, as part of a broader comparison of a set of data models that have been applied to 21-cm signal estimation in the literature (Bowman et al. 2018; Hills et al. 2018; Sims et al. 2023). There it will be shown that selecting the best models for unbiased 21-cm signal inference from time-averaged spectrometer data constitute a category II model comparison problem, underscoring the necessity of model validation in this context. Application of BaNTER validation to other scientific problems involving the comparison of competing composite models for a data set provides a rich avenue for future work.

[28]For example, if the perturbations in the simulation of the validation data are relatively improbable one may use a larger threshold $\ln(\mathcal{B}^{\rm v}_{\rm threshold})$ for the validation null- test or, for a fixed threshold, assign a model that narrowly fails this iteration of the updated validation analysis a reduced but non-zero probability assuming it passes other iterations.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation through research awards for EDGES (AST-1813850, AST-1908933, and AST-2206766). PHS thanks Irina Stefan for valuable discussions and helpful comments on a draft of this manuscript. SGM has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 101067043. EDGES is located at the Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamatji people as the traditional owners of the Observatory site. We thank CSIRO for providing site infrastructure and support.

DATA AVAILABILITY

The data from this study will be shared on reasonable request to the corresponding author.

REFERENCES

Anstey D., de Lera Acedo E., Handley W., 2021, MNRAS, 506, 2041
Bevins H. T. J., Handley W. J., Fialkov A., de Lera Acedo E., Greenhill L. J., Price D. C., 2021, MNRAS, 502, 4405
Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018, Nature, 555, 67
Box G. E. P., 1976, J. Am. Stat. Assoc., 71, 791
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
Elith J., Leathwick J. R., 2009, Annu. Rev. Ecol. Evol. Syst., 40, 677
Fama E. F., French K. R., 1993, J. Financial Econ., 33, 3
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
Handley W., 2018, J. Open Source Softw., 3, 849
Hibbard J. J., Tauscher K., Rapetti D., Burns J. O., 2020, ApJ, 905, 113
Hills R., Kulkarni G., Meerburg P. D., Puchwein E., 2018, Nature, 564, E32
Jeffreys H., 1935, Proc. Camb. Phil. Soc., 31, 203
Kass R. E., Raftery A. E., 1995, J. Am. Stat. Assoc., 90, 773
Liu A., Shaw J. R., 2020, PASP, 132, 062001
Madhusudhan N., Sarkar S., Constantinou S., Holmberg M., Piette A. A. A., Moses J. I., 2023, ApJ, 956, L13
Moran S. E. et al., 2023, ApJ, 948, L11
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
Planck Collaboration VI, 2020, A&A, 641, A6
Rapetti D., Tauscher K., Mirocha J., Burns J. O., 2020, ApJ, 897, 174
Ratcliff R., McKoon G., 2008, Neural Comput., 20, 873
Saxena A., Meerburg P. D., de Lera Acedo E., Handley W., Koopmans L. V. E., 2023, MNRAS, 522, 1022
Saxena A., Meerburg P. D., Weniger C., de Lera Acedo E., Handley W., 2024, RAS Tech. Instrum., 3, 724
Shen E., Anstey D., de Lera Acedo E., Fialkov A., 2022, MNRAS, 515, 4565
Sims P. H. et al., 2023, MNRAS, 521, 3273
Sims P. H. et al., 2025, preprint (arXiv:2506.20042)
Sims P. H., Lentati L., Alexander P., Carilli C. L., 2016, MNRAS, 462, 3069
Sims P. H., Pober J. C., 2019, MNRAS, 488, 2904
Sims P. H., Pober J. C., 2020, MNRAS, 492, 22
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
Singh S., Subrahmanyan R., 2019, ApJ, 880, 26
Spinelli M., Bernardi G., Santos M. G., 2019, MNRAS, 489, 4007
Tauscher K., Rapetti D., Burns J. O., Switzer E., 2018, ApJ, 853, 187
Trotta R., 2008, Contemp. Phys. 49, 71
Wagenmakers E.-J., Lodewyckx T., Kuriyal H., Grasman R., 2010, Cogn. Psychol., 60, 158

APPENDIX A: EULER SET NOTATION IN THE CONTEXT OF 21-CM COSMOLOGY

To provide intuition for the set notation introduced in Section 2.2.2, in this appendix we consider how it applies to an example drawn from 21-cm cosmology, in which one aims to model data containing 21-cm radiation emitted by neutral hydrogen at Cosmic Dawn (the SOI; $\boldsymbol{S}_{\rm a}$) as well as nuisance radio frequency foreground emission and potential instrumental systematics ($\boldsymbol{S}_{\rm b}$; hereafter, foregrounds for brevity).

In the global 21-cm cosmology context, the seven possible permutations of composite models described in Section 2.2.3 are characterized by the attributes summarized in Table A1.

In detail, composite models in the Euler set ABC are ideal models for the data with which unbiased estimates of the SOI can be obtained. These models are constructed by combining models for the 21-cm signal and foregrounds that are both accurate and predictive and the goal of composite model comparison is to identify models in this Euler set. Composite models in $A\overline{B}C$, $\overline{A}BC$, and $\overline{A}\,\overline{B}C$, in contrast, are more problematic. These models are capable of accurately fitting the data in aggregate, but only via inaccuracies in the 21-cm model being absorbed by the foreground model or inaccuracies in the foreground model being absorbed by the 21-cm model or both of these effects occurring in unison, in the three sets, respectively. In any of these cases, one will recover biased estimates of the SOI despite the composite model being predictive of the data in aggregate. Thus, when models of this sort are in the set of models for the data under consideration, model comparison using BFBMC alone will result in a degeneracy, in the space of a posteriori model probability, between these models and those in the set ABC. To eliminate this degeneracy and, with it, the bias in one's inferences of the SOI, identifying and removing models in the sets $A\overline{B}C$, $\overline{A}BC$, and $\overline{A}\,\overline{B}C$ is essential. The BaNTER model validation framework introduced in Section 2.3 is designed for this purpose.

Models in $\overline{A}B\overline{C}$, $A\overline{B}\,\overline{C}$, and $\overline{A}\,\overline{B}\,\overline{C}$ are inaccurate and/or not predictive in aggregate (see equations 10 and 11) and are constructed by combining models for the 21-cm signal and foregrounds where the former, latter or both are inaccurate and/or not predictive. Unlike models in $A\overline{B}C$, $\overline{A}BC$, and $\overline{A}\,\overline{B}C$, these models are identifiable and separable from those of interest, in ABC, using BFBMC of the composite models for the data alone. Thus, if the set of models under consideration derives exclusively from ABC, $\overline{A}B\overline{C}$, $A\overline{B}\,\overline{C}$, and $\overline{A}\,\overline{B}\,\overline{C}$, BaNTER validation is incidental to recovery of unbiased inferences of the SOI.

In practice, the Euler set classifications of the set of models under consideration may be uncertain. In this case, BaNTER validation can be used to identify models that are likely to be biased when fit to the data, if present, and in so doing, facilitates unbiased inferences of the SOI following model comparison, regardless of the Euler set classifications of the models.

Table A1. List of Euler sets permutations that composite models may be elements of, in the context of modelling a 21-cm cosmology data set containing two components: 21-cm radiation emitted by neutral hydrogen at Cosmic Dawn (the SOI; $\boldsymbol{S}_{\rm a}$) and nuisance foregrounds ($\boldsymbol{S}_{\rm b}$). We list for each permutation: whether the 21-cm and foreground component models associated with that Euler set are predictive and accurate of the 21-cm signal and foreground components of the data in isolation, whether the composite model is a predictive and accurate description of the data in aggregate, whether unbiased parameter inferences are recoverable when fitting a composite model in this Euler set to the data and whether model validation is necessary to distinguish between the composite model in the given Euler set and a model in the Euler set of interest, ABC.

Euler set permutation	Predictive and accurate 21-cm signal model	Predictive and accurate foreground model	Predictive and accurate composite model	Unbiased parameter inferences	Model validation necessary
ABC	✓	✓	✓	✓	–
$\overline{A}BC$	✗	✓	✓	✗	✓
$A\overline{B}C$	✓	✗	✓	✗	✓
$\overline{A}\,\overline{B}C$	✗	✗	✓	✗	✓
$A\overline{B}\,\overline{C}$	✓	✗	✗	✗	✗
$\overline{A}B\overline{C}$	✗	✓	✗	✗	✗
$\overline{A}\,\overline{B}\,\overline{C}$	✗	✗	✗	✗	✗

APPENDIX B: CATEGORY I AND II MODEL COMPARISON IN THE CONTEXT OF 21-CM COSMOLOGY

To provide intuition for the model comparison categories introduced in Sections 2.2.4 and 2.2.5, in this appendix we describe how they apply to the 21-cm cosmology example described in Appendix A.

Suppose we have a global 21-cm signal data set, of the form described in equation (4), composed of two signal components – 21-cm radiation emitted by neutral hydrogen at Cosmic Dawn (the SOI; $\boldsymbol{S}_{\rm a}$) and nuisance foregrounds ($\boldsymbol{S}_{\rm b}$) – and noise.

Furthermore, we have a set of composite models for describing the data, $\boldsymbol{\mathcal{M}}_{\rm c}$. Each composite model is obtained by summing a model for the global 21-cm signal, $\boldsymbol{M}_{j\rm a}$ and a model for the foregrounds, $\boldsymbol{M}_{k\rm b}$, where these models are elements of the sets of models under consideration for these components: $\boldsymbol{\mathcal{M}}_{\rm a}$ and $\boldsymbol{\mathcal{M}}_{\rm b}$, respectively.

Categories I and II global 21-cm composite model comparison problems, in the context of global 21-cm cosmology, are distinguished by whether the set of composite models for the data is composed exclusively of models for which unvalidated BFBMC is sufficient for identifying models that are both accurate and predictive models for the data and accurate predictive component models for the 21-cm signal and foreground (those in Euler set ABC; see Appendix A).

A category I global 21-cm composite model comparison problem exists when unvalidated BFBMC is sufficient. This occurs when the elements of $\boldsymbol{\mathcal{M}}_{\rm c}$ are elements of the Euler sets ABC, $A\overline{B}\,\overline{C}$, $\overline{A}B\overline{C}$, or $\overline{A}\,\overline{B}\,\overline{C}$. Here, composite models in the Euler set ABC are composed of a 21-cm signal and foregrounds that are both accurate and predictive and fits of these models to the data enable unbiased estimates of the SOI to be obtained. In contrast, models in $\overline{A}B\overline{C}$, $A\overline{B}\,\overline{C}$, and $\overline{A}\,\overline{B}\,\overline{C}$ are inaccurate and/or not predictive in aggregate and are constructed by combining models for the 21-cm signal and foregrounds where the former, latter or both are inaccurate and/or not predictive.

A category II global 21-cm composite model comparison problem exists when unvalidated BFBMC is insufficient for uniquely identifying models in the Euler set ABC. This occurs when the elements of $\boldsymbol{\mathcal{M}}_{\rm c}$ are elements of the Euler sets ABC, and at least one of $\overline{A}BC$, $A\overline{B}C$, or $\overline{A}\,\overline{B}\,\overline{C}$. Composite models from the latter three Euler sets are capable of accurately fitting the data in aggregate, but only via inaccuracies in the 21-cm model being absorbed by the foreground model or inaccuracies in the foreground model being absorbed by the 21-cm model or both of these effect occurring in unison. These models will yield biased estimates of the 21-cm signal when fit to the data but will be degenerate with models in ABC in the space of model probability as judged by unvalidated BFBMC. In this case, identifying and removing models in these three Euler sets is essential and can be achieved using the BaNTER model validation framework introduced in Section 2.3.

APPENDIX C: FOREGROUND MODEL DESCRIPTION

The individual terms in equation (16) are defined as follows. $\bar{\boldsymbol{B}}_{\rm factor}(\boldsymbol{\nu})$ is an instrumental beam factor associated with the spectral structure of the spectrometer's directivity pattern. The first and second terms describe the spatially isotropic and -anisotropic subcomponents of a power-law component of the foreground emission, respectively. The third term describes the beam-factor-weighted CMB temperature and the spatially isotropic subcomponent of the power-law component of the foreground emission, for which the beam factor is not cancelled during a beam factor chromaticity correction (see S23). The common product of the terms in square brackets, $\mathrm{e}^{-\tau_{\rm ion}(\nu)}$, models ionospheric absorption. The fifth term models the beam-factor-weighted net emission by hot electrons in the ionosphere.

Equation (16) has $N_{\rm pert}+2$ astrophysical foreground parameters and two ionospheric foreground parameters. The free parameters of the astrophysical foreground model are defined as follows. $\bar{T}_{\rm m_0}$ describes the time- and sky-averaged foreground brightness temperature at reference frequency $\nu_{\rm c}$. 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, normalized to the fractional amplitude of the perturbation relative to the mean brightness temperature at the reference frequency $\nu=\nu_{\rm c}$. The two free parameters of the ionospheric foreground model, $T_{\rm e}$ and τ_0 describe the temperature of ionospheric electrons and the effective ionospheric optical depth at $\nu=\nu_{\rm c}$, respectively.

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