Validating Posteriors Obtained by an Emulator When Jointly Fitting Mock Data of the Global 21 cm Signal and High-z Galaxy UV Luminosity Function

J. Dorigo Jones[1] , D. Rapetti[1,2,3] , J. Mirocha[4,5] , J. J. Hibbard[1] , J. O. Burns[1] , and N. Bassett[1]
[1] Center for Astrophysics and Space Astronomy, Department of Astrophysical and Planetary Sciences, University of Colorado Boulder, Boulder, CO 80309, USA
johnny.dorigojones@colorado.edu
[2] NASA Ames Research Center, Moffett Field, CA 94035, USA
[3] Research Institute for Advanced Computer Science, Universities Space Research Association, Washington, DC 20024, USA
[4] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[5] California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125, USA
Received 2023 August 5; revised 2023 September 22; accepted 2023 October 3; published 2023 December 5

Abstract

Although neural-network-based emulators enable efficient parameter estimation in 21 cm cosmology, the accuracy of such constraints is poorly understood. We employ nested sampling to fit mock data of the global 21 cm signal and high-z galaxy ultraviolet luminosity function (UVLF) and compare for the first time the emulated posteriors obtained using the global signal emulator globalemu to the "true" posteriors obtained using the full model on which the emulator is trained using ARES. Of the eight model parameters we employ, four control the star formation efficiency (SFE) and thus can be constrained by UVLF data, while the remaining four control UV and X-ray photon production and the minimum virial temperature of star-forming halos ($T_{\min}$) and thus are uniquely probed by reionization and 21 cm measurements. For noise levels of 50 and 250 mK in the 21 cm data being jointly fit, the emulated and "true" posteriors are consistent to within 1σ. However, at lower noise levels of 10 and 25 mK, globalemu overpredicts $T_{\min}$ and underpredicts γ_{lo}, an SFE parameter, by $\approx 3\sigma$–4σ, while the "true" ARES posteriors capture their fiducial values within 1σ. We find that jointly fitting the mock UVLF and 21 cm data significantly improves constraints on the SFE parameters by breaking degeneracies in the ARES parameter space. Our results demonstrate the astrophysical constraints that can be expected for global 21 cm experiments for a range of noise levels from pessimistic to optimistic, as well as the potential for probing redshift evolution of SFE parameters by including UVLF data.

Unified Astronomy Thesaurus concepts: Nested sampling (1894); Reionization (1383); Luminosity function (942); Neural networks (1933); Posterior distribution (1926); Radio astronomy (1338); Bayesian statistics (1900)

1. Introduction

A promising tool for probing the physics of the early Universe is the 21 cm cosmological signal arising from the neutral hydrogen gas that permeated the intergalactic medium (IGM) before, during, and after the formation of the first stars and galaxies (Madau et al. 1997; for reviews see Furlanetto et al. 2006; Bera et al. 2023). The spin-flip transition in neutral hydrogen emits low-frequency radiation at 1420.4 MHz ($\lambda \approx 21$ cm), which has been redshifted to low radio frequencies ($\nu \lesssim 200$ MHz, corresponding to redshifts $z \gtrsim 6$) owing to cosmic expansion and encodes the high-redshift evolution of the IGM. The 21 cm signal has both an anisotropic component (power spectrum) and an isotropic, sky-averaged component (global signal; Shaver et al. 1999), whose brightness temperature is measured as a differential temperature relative to the cosmic microwave background (CMB) radiation.

An unambiguous detection of the global 21 cm signal has the potential to reveal the true astrophysical and cosmological properties associated with the Dark Ages ($z > 30$–40), Cosmic Dawn (CD; $10 \lesssim z \lesssim 40$), and the Epoch of Reionization (EoR; ending by $z \approx 6$). However, the global 21 cm signal is particularly difficult to detect owing to the presence of significant foreground emission from the Milky Way that is

4–6 orders of magnitude brighter than the underlying signal, making a robust Bayesian forward modeling approach necessary to properly recover and exploit the global 21 cm signal (e.g., Bernardi et al. 2016; Liu & Shaw 2020; Shen et al. 2022).

Radio telescopes on Earth have provided some constraints on the 21 cm power spectrum (e.g., Paciga et al. 2011; Mertens et al. 2020; Trott et al. 2020; Garsden et al. 2021; The HERA Collaboration, et al 2022) and global 21 cm signal (e.g., Bowman et al. 2018; Singh et al. 2018, 2022). The claimed EDGES detection has been met with skepticism (see, e.g., Hills et al. 2018; Bradley et al. 2019; Sims & Pober 2020; Tauscher et al. 2020), particularly because of the systematics involved with measuring the global signal and recently because it has been found to be in tension with the nondetection published by SARAS 3 (Singh et al. 2022). To properly recover the underlying global 21 cm signal, the beam-weighted foreground (i.e., foreground emission convolved with the antenna beam) and instrumental systematics must be carefully fitted and removed (e.g., Hibbard et al. 2020, 2023; Rapetti et al. 2020; Tauscher et al. 2021; Murray et al. 2022; Pagano et al. 2022; Anstey et al. 2023; Saxena et al. 2023a). Radio frequency interference (RFI) is a large systematic due to artificial and ionospheric terrestrial contamination that can be avoided by measuring the 21 cm signal from the pristine radio environment of the far side of the Moon. Upcoming NASA Commercial Lunar Payload Services (CLPS) missions ROLSES (2023, at the lunar south pole; Burns et al. 2021b) and LuSEE-Night

(early 2026, on the far side; Bale et al. 2023) will lay the path for future lunar far side radio telescope arrays capable of measuring the 21 cm global signal and power spectrum (e.g., FARSIDE, Burns et al. 2021a; FarView, Polidan et al. 2022).

Physically motivated models for the global 21 cm signal have various astrophysical and cosmological parameters that affect the shape of the signal. Multiple studies have attempted to constrain such model parameters when fitting a measured global 21 cm signal via a Bayesian, likelihood-based approach (e.g., Monsalve et al. 2018; Mirocha & Furlanetto 2019; Monsalve et al. 2019; Qin et al. 2020; Bevins et al. 2022a, 2023). In this work, we perform a similar Bayesian parameter estimation analysis for eight astrophysical parameters using the publicly available model Accelerated Reionization Era Simulations (ARES;[6] Mirocha 2014; Mirocha et al. 2017) by fitting mock data of the global 21 cm signal and numerically sampling the full posterior distribution of these parameters via nested sampling. We examine the improvement in constraining power on these parameters when jointly fitting mock data of the high-z galaxy rest-frame ultraviolet (UV) luminosity function (LF) in addition to the global 21 cm signal. We present the first nested sampling constraints on ARES parameters when fitting a mock global 21 cm signal and UVLF that are calibrated to real UVLF data. In doing so, we forecast the level of astrophysical constraints that can be expected for different noise levels of global 21 cm experiments in combination with UVLF data.

The recent development of neural-network-based emulators for the global 21 cm signal, such as globalemu (Bevins et al. 2021; Bevins & Gessey-Jones 2023), 21CMVAE (Bye et al. 2022), and 21CMEMU (Breitman et al. 2023; which also emulates other quantities such as the 21 cm power spectrum and the UVLF), enables fast, efficient parameter estimations when fitting the global signal. To our knowledge, there is currently no study that shows a direct comparison of the parameter estimates obtained when using an emulator versus the corresponding full model of the global 21 cm signal in the likelihood. The accuracy of an emulator is determined by computing the rms error (RMSE) between model (i.e., simulated) and network (i.e., predicted) data realizations in a test set, while a fully Bayesian parameter inference and model comparison analysis is much more computationally demanding and yields a formal comparison of the posteriors (Trotta 2008).

Parameter estimation using a full model of the global signal in the likelihood is computationally expensive for most existing models. Most global 21 cm signal models are seminumerical and generate a realization of the signal on the order of minutes to hours (Thomas et al. 2009; Santos et al. 2010; Mesinger et al. 2011; Fialkov & Barkana 2014; Ghara et al. 2015, 2018; Murray et al. 2020; Hutter et al. 2023; Schaeffer et al. 2023; Schneider et al. 2023), which hinders the ability to perform an analysis that requires on the order of 10^5 likelihood evaluations. In contrast, the semianalytical code ARES generates a realization of the global 21 cm signal on the order of seconds, owing its speed primarily to the fact that it evolves the mean radiation background directly as opposed to averaging over large cosmological volumes. Therefore, we use ARES in a Bayesian nested sampling analysis to obtain the "true" posterior distributions and for the first time directly compare them to the emulated posteriors from globalemu.

We generate the mock global 21 cm signal and high-z UVLF using ARES with fiducial parameter values that are calibrated to the Bouwens et al. (2015) UVLF at $z = 5.9$ (Mirocha et al. 2017). We emphasize that the basic ARES UVLF model we employ accurately fits UVLFs at $z \approx 6$–10 obtained by either Hubble Space Telescope (HST) or JWST (Mirocha & Furlanetto 2023), and so our results would not change if we were to fit mock data calibrated to newer JWST UVLF measurements at these redshifts. However, given early indications of a departure from the predictions of HST-based models at $z \gtrsim 10$ (see, e.g., Naidu et al. 2022; Bouwens et al. 2023; Boylan-Kolchin 2023; Donnan et al. 2023; Finkelstein et al. 2023; Harikane et al. 2023; Lovell et al. 2023; Mason et al. 2023b), fitting JWST UVLFs at $z \gtrsim 10$ would require nontrivial changes to the UVLF model we employ (Mirocha & Furlanetto 2023). We defer such analysis to future work (see also Zhang et al. 2022).

To summarize, we pursue three main goals: (1) numerically sample the full posterior distribution of eight astrophysical parameters in ARES, which control the star formation efficiency (SFE) and UV and X-ray photon production per unit star formation in galaxies, when fitting mock global 21 cm signal data with varying noise levels; (2) validate and examine the accuracy of the posteriors obtained by our version of the publicly available neural network emulator globalemu that we trained with ARES; and (3) study the constraints from jointly fitting high-z galaxy UVLF mock data along with the simulated global 21 cm signal.

In Section 2, we describe our methods for obtaining marginalized posterior distributions via nested sampling when fitting mock data of the global 21 cm signal and UVLF. We also describe the training of the globalemu neural network and the generation of the mock data being fit. In Section 3, we present the results from nested sampling analyses, primarily comparing the posteriors obtained when using the emulator globalemu in the likelihood versus the full model ARES, as well as examining the effect on posteriors when jointly fitting with the high-z galaxy UVLF mock data. Finally, we summarize our results and conclusions in Section 4.

2. Analysis

In this section, we describe our analysis method for obtaining the posterior distributions for eight astrophysical parameters in ARES when fitting a mock global 21 cm signal plus statistical noise. The main steps to define our Bayesian analysis are (1) selecting a sampling method, (2) selecting a fiducial model for the global 21 cm signal, and (3) generating mock data by adding to the simulated global signal a noise realization at a statistical error level corresponding to a given integration time. We also train a neural network to emulate the ARES global signal model and study its accuracy versus the full ARES model in producing realizations of the signal.

Note that for this work we are not concerned with systematic uncertainties such as the beam-weighted foreground, RFI (from either terrestrial contamination or the instrument), and environmental horizon and surface conditions (for studies on such effects, see, e.g., Singh et al. 2018; Bassett et al. 2020, 2021; Hibbard et al. 2020, 2023; Kern et al. 2020; Leeney et al. 2022; Murray et al. 2022; Pagano et al. 2022; Anstey et al. 2023).

---

[6] https://github.com/mirochaj/ares; v0.9; git commit hash: fd77c4a869 82d25fdad790d717f8bf5eecff4eb8.

2.1. Likelihood

Bayesian inference allows us to estimate the posterior distribution $P(\theta|\boldsymbol{D}, m)$ of a set of parameters θ in a model m, given observed data $\boldsymbol{D}$ with priors π on the parameters (also written $P(\theta|m)$). This is achieved via the Bayes theorem:

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

where $\mathcal{L}$ is the likelihood function, or the probability of the data given the parameters of the model (also written $P(\boldsymbol{D}|\theta, m)$), and the normalizing factor Z is the Bayesian evidence, or marginal likelihood over the priors (also written $P(\boldsymbol{D}|m)$), which can be used for model comparison.

For all of the fits performed in this paper, we sample from a multivariate log-likelihood function assuming Gaussian-distributed noise:

$$\log \mathcal{L}(\theta) \propto [\boldsymbol{D} - m(\theta)]^T \boldsymbol{C}^{-1}[\boldsymbol{D} - m(\theta)], \quad (2)$$

where C is the noise covariance matrix of the data, which we assume to be diagonal. In this paper, we fit mock data realizations for the global 21 cm signal (D_{21}) and the UVLF (D_{UVLF}) instead of real data, although for the latter the mock data are calibrated to real measurements of the high-z galaxy UVLF (see Section 2.6). Hence, we know the input, or fiducial, values of the parameters whose posteriors we numerically sample and can evaluate the validity of the sampling methods and the accuracy of the ARES model and globalemu emulator based on the expectation of marginalized posterior distributions around the fiducial parameter values.

2.2. Combined Constraints

To better realize the constraints that are achievable from global 21 cm signal experiments, in addition to fitting only the mock global signal, we also perform joint fits that combine the model constraining powers from the global signal and high-z galaxy UVLF mock data. Using Equation (2), we construct separate log-likelihood functions for the global 21 cm signal and the UVLF. For the joint fits, we form a log-likelihood by adding both individual likelihoods (see, e.g., Chatterjee et al. 2021; Bevins et al. 2023):

$$\log \mathcal{L}_{\mathrm{joint}} = \log \mathcal{L}(\boldsymbol{D}_{21}|\theta) + \log \mathcal{L}(\boldsymbol{D}_{\mathrm{UVLF}}|\theta). \quad (3)$$

We evaluate the separate log-likelihood functions at the same set of parameters using the same priors to sample the full posterior distribution, as the models we employ for the 21 cm signal, $m_{21}(\theta)$, and for the UVLF, $m_{\mathrm{UVLF}}(\theta)$, are both generated using the ARES framework (see Section 2.4).[7] For the global 21 cm signal likelihood, the noise covariance matrix $\boldsymbol{C}_{21}$ is a diagonal array of constant values corresponding to the square of the estimated noise level σ_{21}. For the UVLF likelihood, the main diagonal elements of $\boldsymbol{C}_{\mathrm{UVLF}}$ are the same as the errors on the $z = 5.9$ UVLF data by Bouwens et al. (2015) (see Section 2.6).

2.3. Nested Sampling

We employ the Bayesian inference method of nested sampling (Skilling 2004; for reviews see Ashton et al. 2022; Buchner 2023). Conceptually, nested sampling algorithms converge on the best parameter estimates by iteratively removing regions of the prior volume with lower likelihood. Nested sampling computes both the evidence and posterior samples simultaneously (by recasting the multidimensional evidence integral into a 1D integral), whereas Markov Chain Monte Carlo (MCMC) samplers calculate only the posterior.

In general, Monte Carlo methods like nested sampling and MCMC are computationally expensive because they require many likelihood evaluations to sample the converged posterior distributions. We choose nested sampling instead of MCMC because the former is designed to better constrain complex parameter spaces with banana-shaped curved degeneracies and/or multimodal distributions (Buchner 2023). Another likelihood-based method that has been applied to parameter estimation of the global 21 cm signal is Fisher matrix analysis (Liu et al. 2013; Muñoz et al. 2020; Hibbard et al. 2022; Mason et al. 2023a), which assumes multivariate Gaussian posterior distributions and requires only $\mathcal{O}(N)$ likelihood evaluations for N parameters being sampled. Fisher analysis is efficient but provides an accurate description only when the posteriors are symmetric, Gaussian, and unimodal (e.g., Trotta 2008; Ryan et al. 2023); however, Fisher matrix generalizations exist (Heavens 2016), such as adding higher-order matrices (Sell-entin et al. 2014). There are also "likelihood-free" inference methods (also called simulation-based inference; see Cranmer et al. 2020), which have been shown to provide accurate posteriors at a relatively low computational cost (Prelogović & Mesinger 2023; Saxena et al. 2023b).

Two nested sampling algorithms in particular have been used in 21 cm cosmology and have been shown to efficiently sample posterior distributions: MultiNest (Feroz & Hobson 2008; Feroz et al. 2009, 2019) and PolyChord (Handley et al. 2015a, 2015b; Handley 2020). In both MultiNest and PolyChord, an initial number of "live" points, n_{live}, are generated in the prior volume, which are used to eventually converge on the best parameter estimates, but the two nested sampling algorithms differ in how they replace live points. For a more in-depth comparison of MultiNest and PolyChord see Section 2 of Lemos et al. (2023). Because of their different approaches for replacing live points, MultiNest and PolyChord are known to perform differently depending on the number of dimensions, or the number of parameters being constrained (see Figure 4 in Handley et al. 2015a). We primarily utilize MultiNest for our analyses, and we show for one joint fit that the two nested sampling algorithms converge on roughly the same result for the same n_{live} but with MultiNest being much more efficient than PolyChord for constraining eight astrophysical parameters in ARES (see Section 3.1).

2.4. Modeling the Global 21 cm Signal and UVLF

To simulate the global 21 cm signal (and high-z galaxy UVLF), we use the physically motivated, semianalytical code ARES, which is the union of a 1D radiative transfer code developed in Mirocha et al. (2012) and a uniform radiation background code described in Mirocha (2014). ARES outputs realizations of the global 21 cm signal and galaxy LF in just

---

[7] In Equation (3), we set the relative weights for the log-likelihoods equal to unity. It is also possible to explore weights by which to multiply the separate log-likelihoods when combining data sets from different experiments (e.g., Lahav et al. 2000; Hobson et al. 2002); we leave consideration of such analysis for future work.

seconds, which makes it computationally feasible to perform direct parameter estimation using the full model ARES rather than an emulator in the likelihood of a nested sampling analysis.[8] Although ARES contains cosmological parameters that affect the shape of the Dark Ages trough, in this work we focus on demonstrating the astrophysical constraints that are achievable when fitting the Cosmic Dawn and reionization redshift ranges.

For high-z galaxies, the observed LF probes the rest-frame UVLF, $\phi(M_{\rm UV})$, and so the UVLF model primarily depends on the star formation rate of massive, young stars. The ARES model is motivated by studies of the high-z galaxy LF based on abundance matching, and the fiducial model ignores dust extinction (which has a minor impact on the conversion between the observed and intrinsic LF at $z \gtrsim 6$) and suggested redshift evolution of the SFE. ARES assumes a multicolor disk spectrum for high-mass X-ray binaries (Mitsuda et al. 1984) and uses the BPASS version 1.0 single-star models for continuous star formation to derive the UV photon production efficiency (Eldridge & Stanway 2009).

For full descriptions of how ARES models the galaxy UVLF and the global 21 cm signal, see Section 2 of Mirocha et al. (2017). Here we will provide a brief description of the UVLF model to highlight the SFE parameterization. The two components required to calculate the UVLF are (1) the intrinsic luminosity L of galaxies as a function of dark matter (DM) halo mass M_h, and (2) the DM halo mass function (HMF; i.e., the number of DM halos per mass bin per comoving volume of the Universe). The HMF has been well studied (e.g., Press & Schechter 1974; Bond et al. 1991; Murray et al. 2013), and in ARES it is calculated a priori in lookup tables using an analytical construct that assumes that halos form by spherical collapse. The luminosity of each halo can be written in terms of the star formation rate, which is itself the product of the SFE, $f_\star$, and the baryon mass accretion rate (MAR). The MAR is derived directly from the HMF (see, e.g., Furlanetto et al. 2017; Mirocha et al. 2021), and so all that is needed to calculate the UVLF is a parameterization for the SFE. Here, as in Mirocha et al. (2017), we assume that the SFE is a double power law in M_h:

$$f_\star(M_h) = \frac{f_{\star,0}}{\left(\frac{M_h}{M_p}\right)^{\gamma_{\rm lo}} + \left(\frac{M_h}{M_p}\right)^{\gamma_{\rm hi}}}, \tag{4}$$

where $f_{\star,0}$ is the peak SFE at mass M_p, and $\gamma_{\rm lo}$ and $\gamma_{\rm hi}$ are the power-law indices at low and high masses, respectively.

We sample the full posterior distribution of eight parameters, including the four SFE parameters ($f_{\star,0}$, M_p, $\gamma_{\rm lo}$, and $\gamma_{\rm hi}$) and four other astrophysical parameters: $c_X, f_{\rm esc}, T_{\rm min}$, and $\log N_{\rm HI}$. The production and release of X-ray photons in galaxies are controlled by c_X and $\log N_{\rm HI}$, the escape of UV photons is controlled by $f_{\rm esc}$, and the minimum virial temperature that determines the number of collapsed star-forming halos is controlled by $T_{\rm min}$. In Table 1, we summarize these eight parameters and give the flat prior ranges used in the nested sampling analyses and also when training the globalemu network on ARES mock global 21 cm signals. The flexible ARES parameter space allows us to set wide, uninformative

priors over these free parameters that are still physically meaningful. In order to facilitate a complete exploration of the prior volume, for four parameters, c_X, $T_{\rm min}$, $f_{\star,0}$, and M_p, we sample from their prior ranges uniformly in log10 space, as shown in Table 1. Our prior ranges are centered on some empirically motivated values (see Section 2.6 for description of fiducial parameter values), but we give multiple orders of magnitude on either side of those values to accommodate potentially dramatic departures at high z and to capture the full resulting converged posterior distributions (see Section 3).

One of our main goals is to directly compare the posteriors when using an emulator for the global 21 cm signal versus when using the full model on which the emulator was trained. In the next subsection, we describe the construction of the training set for the emulator and directly assess the accuracy of the emulated signals compared to the "true," input ones.

2.5. Emulating ARES with globalemu

We employ the publicly available global 21 cm signal emulator globalemu (Bevins et al. 2021) for our analyses, though other emulators for the signal do exist, such as 21CMVAE (Bye et al. 2022), 21CMGEM (Cohen et al. 2020), and the recently released 21CMEMU (Breitman et al. 2023); we leave a comparison of the posteriors obtained from different global 21 cm signal emulators to future work. To obtain a trained globalemu neural network that accurately emulates ARES, we first create a large training set of simulated global 21 cm signals generated by ARES and then train globalemu on this training set. For the latter step, we test multiple network architectures (i.e., different numbers of nodes and hidden layers composing the network; see Bevins et al. 2021 for a detailed description of the network).

To create the training set, we generate global 21 cm signals from ARES by drawing random values[9] from the parameter ranges given in Table 1. Each signal spans the redshift range $z = 6$–55 with a redshift spacing of $\delta z = 0.1$, similar to Bevins et al. (2022a). The training set that is ultimately used to train the globalemu network used for analyses presented in this work contains 24,000 mock signals. A representative subset of this training set is shown in Figure 1. We also generated training sets of sizes 5000, 10,000, and 20,000, which all resulted in less accurate trained networks. The marginal improvement of 10% in the RMSE of the resulting trained network obtained when using a training set of size 24,000 compared to 20,000, however, indicates that increasing the number of global signals in the training set above 24,000 would not significantly affect our results. In addition, we also created a so-called test set of 2000 global signals using ARES and the same parameter ranges as used for the training set. Importantly, the test set is completely separate from the training set and is used to determine the accuracy of the trained globalemu network.

Using the 24,000-signal ARES training set, we train five globalemu networks each with a different network architecture. We test a similar, although less comprehensive, grid of

---

[8] Each fit using the full ARES model was performed on the Blanca compute cluster operated by University of Colorado Research Computing, employing three nodes with 30 CPU cores per node (parallelized via Intel/IMPI compiler) and utilizing ~ 100 GB of total memory.

[9] At this time, we do not impose constraints on the CMB optical depth, τ_e, or the neutral hydrogen fraction, $x_{\rm H\,I}$, when creating the training sets. This results in a number of unphysical signals with $\tau_e > 1\sigma$ from the value obtained by Planck Collaboration et al. (2020) and/or $x_{\rm H\,I} \gtrsim 5\%$ at $z = 5.3$, despite the signals being generated from physically reasonable parameter ranges. The existence of unphysical signals in our training set is thus a contaminating factor toward constraining actual data, which is beyond our scope of testing the accuracy of the globalemu emulation of ARES.

Table 1
Astrophysical Parameters in ARES to Be Fit with Mock Global 21 cm Signal and High-z UVLF Data

Parameter	Description	Prior Range (with Units)	Fiducial Value, θ_0
c_X	normalization of X-ray luminosity–SFR relation	Log unif. $[10^{36}, 10^{41}]$ erg s^{-1} $(M_\odot$ yr$^{-1})^{-1}$	2.6×10^{39}
f_{esc}	escape fraction of UV photons	Uniform $[0, 1]$	0.2
T_{min}	minimum virial temperature of star-forming halos	Log unif. $[3 \times 10^2, 5 \times 10^5]$ K	10^4
$\log N_{HI}$	neutral hydrogen column density in galaxies	Uniform $[18, 23]$	21
$f_{\star,0}$	peak star formation efficiency	Log unif. $[10^{-5}, 10^0]$	0.05
M_P	dark matter halo mass at $f_{\star,0}$	Log unif. $[10^8, 10^{15}]M_\odot$	2.8×10^{11}
γ_{lo}	low-mass slope of $f_\star(M_h)$	Uniform $[0, 2]$	0.49
γ_{hi}	high-mass slope of $f_\star(M_h)$	Uniform $[-4, 0]$	−0.61
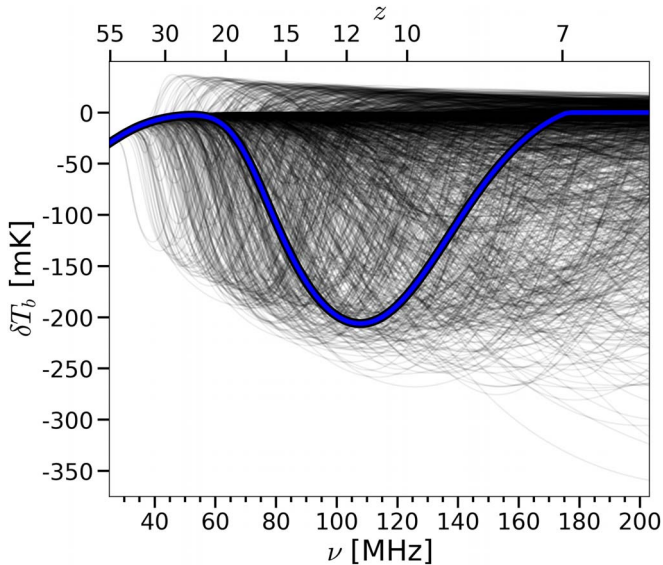
Figure 1. Representative subset of the training set (10% out of 24,000 total) containing mock global 21 cm signals generated by ARES when varying eight astrophysical parameters. The full training set was used to train globalemu (see Table 1 for the parameter ranges). Shown in thick blue is the fiducial global 21 cm signal to which we add Gaussian-distributed noise at different levels to form the mock 21 cm data sets that we fit. The mock UVLF data that we add in our joint fits are also generated by ARES using the same fiducial parameter values (see Table 1) that were obtained via calibration to the Bouwens et al. (2015) $z = 5.9$ UVLF by Mirocha et al. (2017), as described in Section 2.6.

architectures to those tested in Bevins et al. (2021, see their Figure 8): [8, 8, 8], [64, 64], [16, 16, 16, 16], [16, 16, 16], [32, 32, 32], where the values of each component in a given bracket are the numbers of nodes in each hidden layer, and the number of components in each bracket is the number of layers. The network stops learning once the loss function does not improve by 10^{-5} within the last 20 epochs of training, which ensures that the trained network is as accurate as possible for the chosen network architecture.[10] For the data preprocessing step that is required before training the network (see Section 4 of Bevins et al. 2021), we turn off the astrophysics-free baseline subtraction and resampling options because we find that they

---

[10] Training a network until the loss function on the training set stops decreasing can sometimes result in overfitting, where the network learns the training set specifically rather than the underlying patterns. We test for overfitting in our network trained with an architecture of three hidden layers of 32 nodes by comparing the distribution of loss values (i.e., RMSE normalized by the maximum $|\delta T_b|$) across the training and test data sets (see also Bevins et al. 2021). We find that the loss distributions for the test and training sets, when emulated with the trained neural network, are nearly identical, and so we can consequently conclude that the network is not overfitting the training data.

have a slightly negative impact on the accuracy of the resulting trained network. The lack of benefit from the preprocessing steps may be due to the fact that the "astrophysics-free" Dark Ages composes a small portion of our simulated signals.

We determine the accuracy of each trained network by evaluating them at the parameter values of the 2000 ARES signals in the test set and comparing the resulting emulated signals to their corresponding "true" signals. The top panel of Figure 2 shows a subset of the test set (in black) plotted along with the corresponding emulations (in red) generated by the globalemu network used for analyses presented in this work. The bottom panel of Figure 2 shows the residuals between the emulated and "true" signals in the test set. We find that the network architecture of [32, 32, 32] gives the lowest mean RMSE of 1.25 mK (with a maximum RMSE of 18.5 mK) between the 2000 emulated and "true" signals (see the horizontal dotted red line in the bottom panel of Figure 2), while the other network architectures gave mean RMSEs ranging from 1.8 to 4.5 mK. Network training for the architecture [32, 32, 32] took 10 hr, as performed on a 2018 MacBook Pro with a six-core i9 processor and 32 GB of memory. The mean RMSE of 1.25 mK is comparable to or better than those achieved in other studies that trained globalemu on large training sets (e.g., Bevins et al. 2022a, 2022b, 2023), and Bevins et al. (2021) also found [32, 32, 32] to give the lowest mean RMSE of the trained network.

The efforts described above to optimize the accuracy of the ARES-trained globalemu network provide robustness to the accuracy limits determined in Section 3. Even so, the small RMSE of the trained network should contribute to bias on the resulting emulated parameter constraints. We briefly investigate this by determining whether or not there is a correlation in the test set between the depth of a signal's Cosmic Dawn (CD) trough and the accuracy of its corresponding emulation. We find no statistically significant correlation between CD trough depth and the mean, median, or maximum emulation residual, obtaining Kendall rank and Pearson correlation coefficients between −0.1 and −0.6 with p-values all $<10^{-3}$. Therefore, we infer that emulated posterior biases are not correlated to emulation residuals in the CD trough depth, although we defer to future work a detailed investigation of the relationship between RMSE network uncertainties and the accuracy of the emulator model constraints.

2.6. Mock Data

For all analyses, we fit the same mock data realization for the global 21 cm signal and galaxy UVLF at $z = 5.9$ generated by ARES using a fiducial set of parameter values, θ_0 (see Figure 1
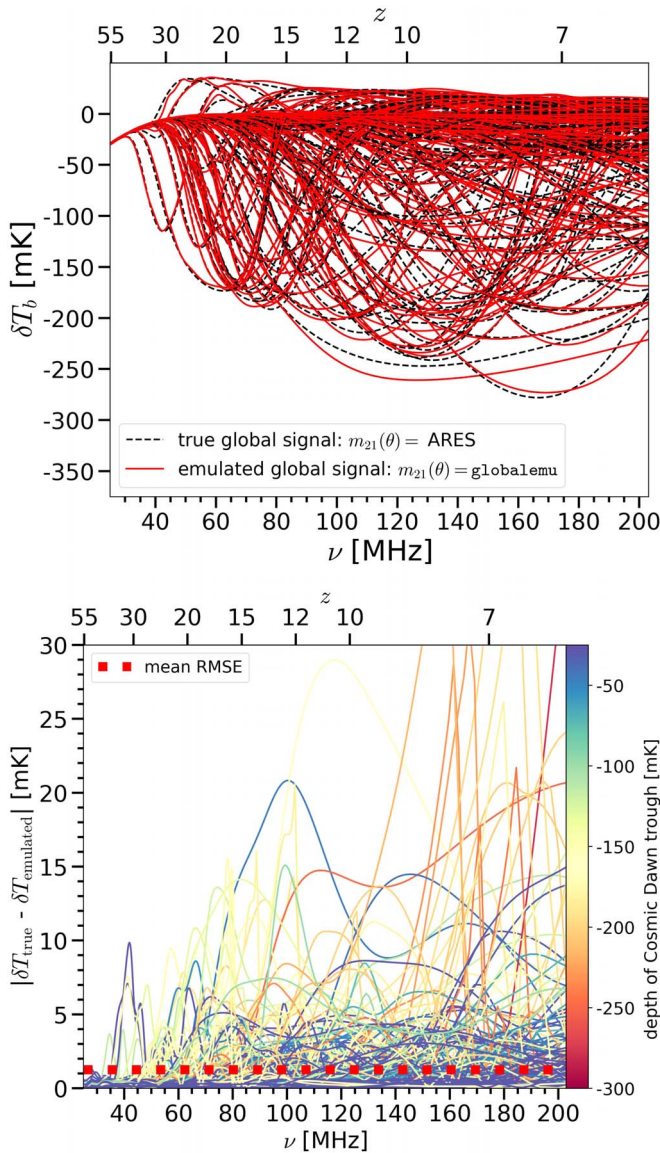
Figure 2. Top: representative subset of the test set (200 out of 2000) generated by ARES ("true" global signals; black dashed curves) and the corresponding subset of emulations from the globalemu network (solid red curves) trained on the ARES training set using the architecture [32, 32, 32]. Bottom: differences between the emulated and "true" signals in the top panel (i.e., emulation residuals), with color depicting the depth of the Cosmic Dawn (CD) trough of the respective signal. The horizontal dotted red line indicates the mean RMSE of 1.25 mK between the emulated and "true" signals in the full test set (see Section 2.5).

and Table 1). The fiducial values used for the four parameters that the UVLF is sensitive to (i.e., the four SFE parameters: $f_{\star,0}$, M_p, γ_{lo}, and γ_{hi}; see Equation 4) were determined empirically via calibration to the $z = 5.9$ UVLF measured by Bouwens et al. (2015; see Mirocha et al. 2017 for details on this calibration). For the other four astrophysical parameters that we constrain (i.e., the four "non-SFE" parameters: c_X, f_{esc}, T_{min}, and $\log N_{HI}$) we use typical, physically motivated fiducial values based on observations or simulations.

Because the non-SFE parameters have no effect on the ARES UVLF model, their values are not constrained by the UVLF calibration procedure. In particular, the fiducial value for c_X is motivated by studies of low-z star-forming galaxies (e.g., Mineo et al. 2012), and the fiducial value for $\log N_{HI}$ is

motivated by simulations (e.g., Das et al. 2017). The difference in our fiducial values for f_{esc} and $\log N_{HI}$ compared to those used in Mirocha et al. (2017) results in our fiducial mock global 21 cm signal (see the blue curve in Figure 1) having a Cosmic Dawn trough that is located at the same frequency but is ≈ 50 mK deeper.

The fiducial mock global 21 cm signal is created in the same manner as the training set (i.e., $z = 6$–55 with step $\delta z = 0.1$), and the mock galaxy UVLF is created at the same 10 magnitudes as the UVLF at $z = 5.9$ measured by Bouwens et al. (2015). Therefore, the mock UVLF that we fit is a collection of 10 data points that resembles the actual $z = 5.9$ UVLF measured by Bouwens et al. (2015), but with small vertical offsets from the real data points due to the UVLF calibration procedure that allows us to identify the input model parameters (see the left panel of Figure 2 in Mirocha et al. 2017 for a comparison of the fiducial ARES UVLF model and the Bouwens et al. 2015 UVLF).

The noise that we add to the fiducial mock 21 cm signal is Gaussian distributed with a standard deviation noise estimate σ_{21}. For our analyses, we test five different 21 cm noise levels (including the optimistic, fiducial, and pessimistic scenarios used for the REACH radiometer in de Lera Acedo et al. 2022): $\sigma_{21} = 5$ mK or 10 mK (referred to as "optimistic"), $\sigma_{21} = 25$ mK or 50 mK (referred to as "standard"), and $\sigma_{21} = 250$ mK (referred to as "pessimistic"). We also note that the noise added to D_{21} is constant in frequency space, whereas in practice the noise on the measured global 21 cm signal is expected to decrease with increasing frequency according to the radiometer equation. It has been suggested that such frequency dependence has little impact on the derived parameter constraints (Bevins et al. 2022b), but full treatment is left for future work. For the UVLF, we use the error reported for the $z = 5.9$ UVLF data from Bouwens et al. (2015).

3. Results

In this section, we present the results of fitting mock global 21 cm signal data, with and without mock high-z galaxy UVLF data, using various noise levels to be expected from 21 cm experiments. For the astrophysical modeling, we employ either an ARES-trained globalemu network or the full ARES model.

We first discuss the posteriors obtained when jointly fitting the mock 21 cm and UVLF data (Section 3.1, Figures 3 to 5, followed by those obtained when separately fitting the individual data sets (Section 3.2, Figure 6), and lastly we discuss the concept of posterior consistency in our results (Section 3.3). Because the joint fits produce unimodal posteriors with well-behaved means, we focus primarily on the posteriors from joint fits when comparing globalemu and ARES.

We determine the accuracy of the ARES-trained globalemu model by comparing the mean (see top panel of Figure 5) or the shape (see Appendix A) of the emulated posterior distributions with those of the "true," full ARES posteriors. Note that this comparison is driven by the global 21 cm signal since globalemu does not emulate the UVLF, which we continue to model with ARES. To our knowledge, the recently released 21CMEMU (Breitman et al. 2023) is the only publicly available emulator that includes the UVLF; see, however, Kern et al. (2017) for a more general emulator.
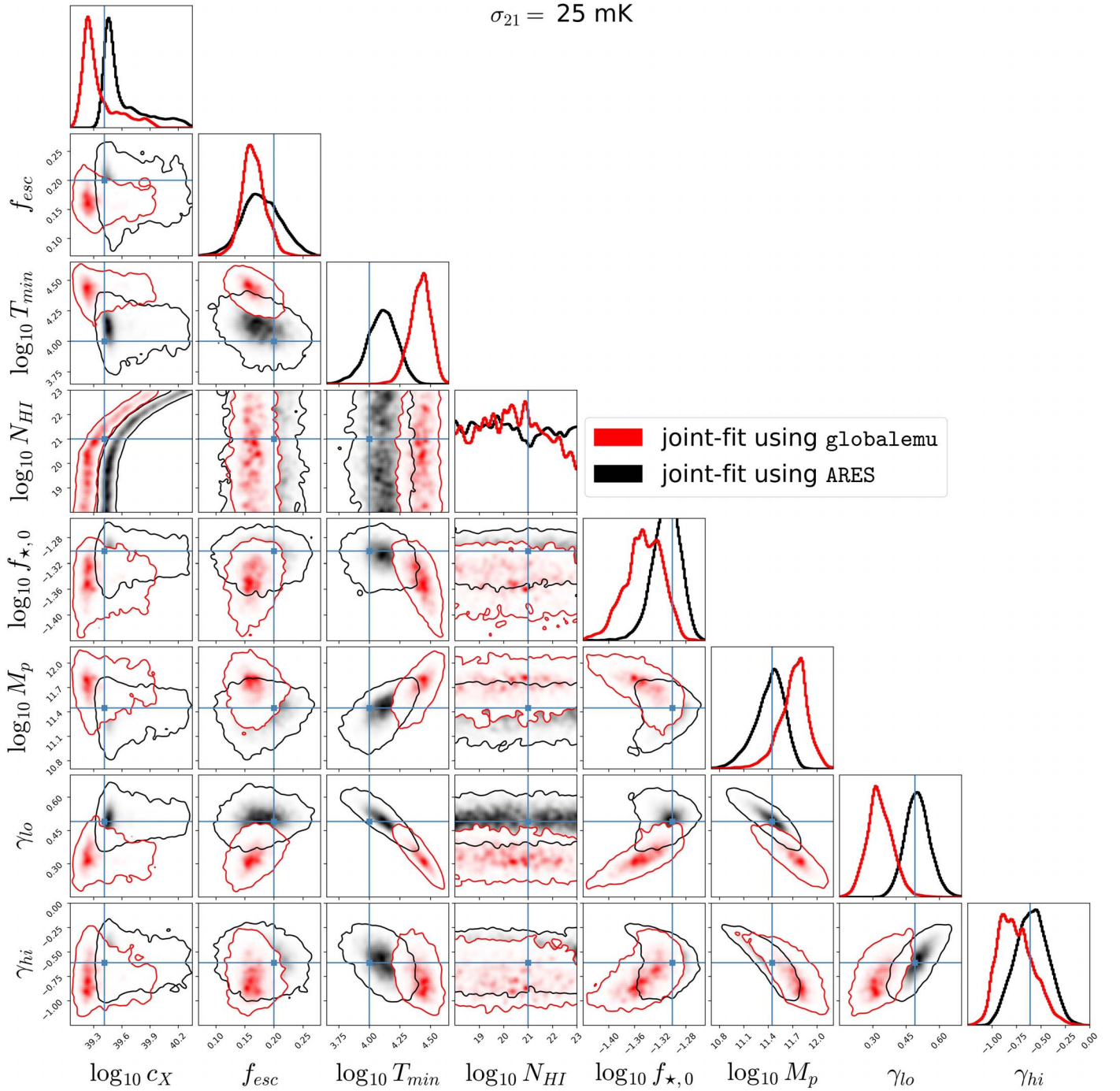
Figure 3. Marginalized 1D and 2D posterior distributions for eight astrophysical parameters in ARES when jointly fitting mock global 21 cm signal and UVLF data. These eight parameters control the SFE and the UV and X-ray photon production in galaxies (see Table 1). The red posterior is obtained using the ARES-trained globalemu network model, and the black posterior is obtained using the full ARES model. Blue vertical and horizontal lines indicate the input, or fiducial, parameter values used to generate the mock data being fit (see Table 1), which are calibrated to real observations of the UVLF (see Section 2.6). The statistical noise in the 21 cm data being fit is $\sigma_{21} = 25$ mK, which among the five tested we find gives the most accurate "true" ARES posteriors with respect to the fiducial parameter values and also highlights for which parameters globalemu obtains biased constraints (see also Figure 5). The UVLF data noise is the same as the error on the $z = 5.9$ UVLF measurements from Bouwens et al. (2015). Contour lines in the 2D histograms represent the 95% confidence levels, and density color maps are shown. Axis ranges are zoomed in with respect to the full prior ranges given in Table 1. See Table 2 for further details on each fit.

For most fits, we find it necessary to use more than the default number of initial live points in MultiNest of $n_{\text{live}} = 400$ in order to fully sample the posterior and obtain convergence (see Table 2 for details on fits performed). For the choice of sampling efficiency (i.e., the ratio of points accepted to points sampled), we use the recommended value for parameter estimation in MultiNest, $e = 0.8$, and for the evidence tolerance, the recommended, default value of tol = 0.5.[11] All of the triangle plots shown in this paper were generated using the Python module corner.py (Foreman-Mackey 2016; Foreman-Mackey et al. 2016) with 100 bins and

---

[11] Note that when testing the decrease of this value to 0.1 it only extended the length of the run without significantly altering the final results.

Table 2
Summary of Key Nested Sampling Analyses

Type of Mock Data Being Fit	Model Used in Likelihood	σ_{21} (mK)	σ_{UVLF} (mag^{-1}cMpc^{-3})	n_{live}	$\log Z$	$n_{\mathrm{evaluations}}$	f_{accept}	sec./eval. (s)
Both UVLF and global 21 cm signal	`globalemu`	25	B+15	600	-280.6 ± 0.2	94,843	0.147	24.60
		50	B+15	600	-278.5 ± 0.2	91,614	0.163	24.35
		250	B+15	600	-274.4 ± 0.1	44,628	0.244	24.20
	`ARES`	5	B+15	800	-290.1 ± 0.3	445,729	0.052	34.20
		10	2xB+15	400	-282.2 ± 0.2	128,745	0.087	35.64
		25	B+15	800	-280.0 ± 0.1	129,367	0.154	36.04
		50	B+15	800	-275.8 ± 0.1	104,890	0.185	34.62
		250	B+15	800	-273.2 ± 0.1	57,448	0.254	37.56
	`ARES` using `PolyChord`	10	2xB+15	400	-284.3 ± 0.2	3,650,406	0.004	39.92
Only global 21 cm signal	`globalemu`	25	⋯	1200	-268.9 ± 0.1	142,805	0.145	0.01
		50	⋯	1200	-266.5 ± 0.1	86,308	0.202	0.01
	`ARES`	25	⋯	1200	-268.8 ± 0.1	166,331	0.128	17.23
		50	⋯	1200	-266.2 ± 0.1	103,231	0.176	18.11
Only UVLF	`ARES`	⋯	B+15	400	-11.6 ± 0.2	15,578	0.348	16.37

Note. The pieces of information provided for each fit are the noise level of the mock 21 cm signal (σ_{21}) and/or UVLF (σ_{UVLF}) being fit, the number of initial live points used (n_{live}), and the final output metrics, including the evidence ($\log Z$), the total number of likelihood evaluations ($n_{\mathrm{evaluations}}$), the acceptance rate (f_{accept}), and average CPU-time required per evaluation (sec./eval.). "B+15" denotes that the UVLF error used is the same as that of the $z = 5.9$ UVLF data by Bouwens et al. (2015) (see Section 2.6). All fits shown were performed using `MultiNest`, except for one joint fit for which we used `PolyChord`, the result of which is consistent with the equivalent `MultiNest` fit (see Figure 4). The fit using `PolyChord` required over an order of magnitude more computational time to converge compared to the equivalent `MultiNest` fit, and so we used twice the "B+15" UVLF error to aid convergence in a reasonable amount of time without significantly affecting the results (see Section 3.1). The results from each fit included here are presented in Section 3 and Appendix B (see Figures 3 to 6 and Figures B1 and B2 in Appendix B), except for the $\sigma_{21} = 25$ mK only global signal fits.

a Gaussian smoothing kernel of 2σ. For case examples, we tested that increasing the number of bins did not affect the essence of the results presented. The resulting posteriors were plotted using the samples and weights output by the converged nested sampling runs.

3.1. Jointly Fitting 21 cm and UVLF Mock Data

In Figure 3, we present the posteriors obtained from a joint fit using either `globalemu` or `ARES`, for a "standard" noise level of $\sigma_{21} = 25$ mK in the 21 cm data being fit. We present this as the main joint-fit result because we find that $\sigma_{21} = 25$ mK gives the least biased mean parameter values for `ARES` with respect to the fiducial ones and therefore provides the best representation of the accuracy limits of the `globalemu` model with respect to `ARES`. In Figure 4, we compare the 1D posteriors obtained from joint fits using `ARES` for three characteristic 21 cm noise levels. In Figure 5, we summarize the biases between the emulated and "true" posteriors, as well as between the "true" posteriors and the fiducial values, for the five tested 21 cm noise levels. We present the full posterior distributions obtained from joint fits for $\sigma_{21} = 50$ and 250 mK in Appendix B.

As to be expected, because the four SFE parameters ($f_{\star,0}$, M_p, γ_{lo}, and γ_{hi}) directly determine the UVLF model (Section 2.4), their posteriors are well constrained when adding the UVLF to the 21 cm data. For joint fits, the four SFE posteriors are unimodal and centered on the fiducial value, which is not the case when fitting only the 21 cm data as we discuss in Section 3.2. Interestingly, the bimodalities in the 1D posteriors for M_p and γ_{lo} when fitting only the 21 cm data disappear when adding the UVLF data in the joint fit, showing that the combination of both data sets can break degeneracies in the `ARES` parameter space and reduce biases.

Comparing the emulated distributions (in red) and "true" distributions (in black) in Figure 3, we see that the `globalemu` model produces similar posteriors as the `ARES` model, in both shape (see Appendix A) and mean (top panel of Figure 5), except for a few exceptions discussed below. In Figure 5, we summarize the two different types of parameter biases discussed, for the different 21 cm noise levels tested: emulation bias (Equation 5) and true bias (Equation 6). Emulation bias refers to the accuracy of the emulated posterior parameter means, $\mu_{\mathrm{globalemu}}$, with respect to the "true" posterior parameter means, μ_{ARES}, and true bias refers to the accuracy of μ_{ARES} with respect to the fiducial parameter value, θ_0. We note that these two biases provide all of the information necessary to evaluate the accuracy of `globalemu` and `ARES` and that defining a third bias between $\mu_{\mathrm{globalemu}}$ and θ_0 does not further aid our results.

We therefore define and compute an emulation bias as the difference in the emulated and "true" posterior parameter means divided by the standard deviation of the "true" posterior, σ_{ARES}:

$$\text{emulation bias} \equiv \frac{|\mu_{\mathrm{globalemu}} - \mu_{\mathrm{ARES}}|}{\sigma_{\mathrm{ARES}}}. \quad (5)$$

In the same manner, we define a true bias between an `ARES` posterior parameter mean and its fiducial value:

$$\text{true bias} \equiv \frac{|\mu_{\mathrm{ARES}} - \theta_0|}{\sigma_{\mathrm{ARES}}}. \quad (6)$$

We find that, in general, the emulation bias decreases as the 21 cm noise level increases. For $\sigma_{21} = 50$ and 250 mK, all parameters' emulation biases are $\leqslant 1$ (marked with a black horizontal line in Figure 5), while at lower noise levels ($\sigma_{21} = 5$, 10, and 25 mK) the emulation bias increases above 1 for
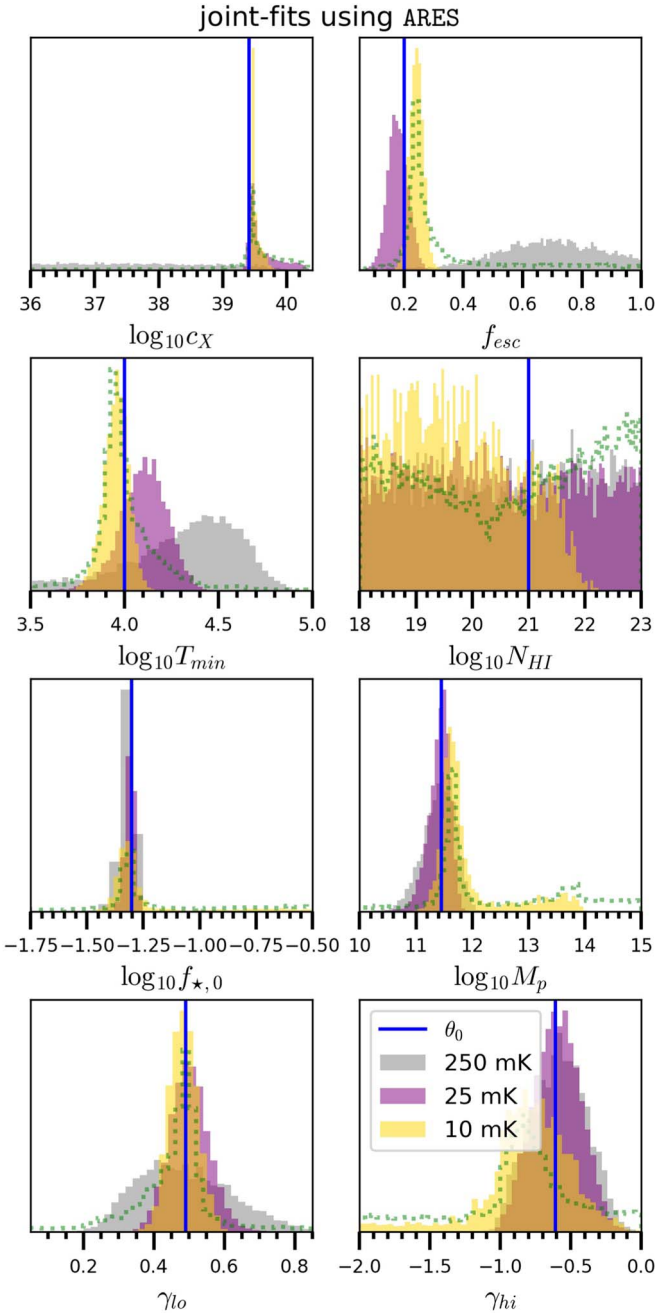
Figure 4. Marginalized 1D posterior distributions when jointly fitting mock global 21 cm signal and UVLF data using the full ARES model, for three different 21 cm noise levels: 10 mK (optimistic), 25 mK (standard), and 250 mK (pessimistic). These eight parameters control the SFE and the UV and X-ray photon production in galaxies (see Table 1). Blue vertical lines indicate the input, or fiducial, parameter values used to generate the mock data (see Section 2.6). The dotted green histograms result from using PolyChord with $\sigma_{21} = 10$ mK and match well the corresponding distributions obtained by using MultiNest. The noise on the mock UVLF being fit is the same as the error on the $z=5.9$ UVLF measurements from Bouwens et al. (2015), except for the posteriors for 10 mK shown here, for which we used twice the UVLF error to allow for a reasonable convergence time of the PolyChord run (see Section 3.1). The posteriors for 25 and 250 mK are the same as those in Figure 3 and B2 in Appendix B, respectively. Axis ranges are zoomed in from the full prior ranges given in Table 1.

certain parameters. For $\sigma_{21} = 10$ and 25 mK, $T_{\min}$ and γ_{lo} have emulation biases of 3–4, and for $\sigma_{21} = 5$ mK, $T_{\min}$ and f_{esc} have even higher emulation biases of $\approx$6–10, while the emulation bias of γ_{lo} drops below 1.
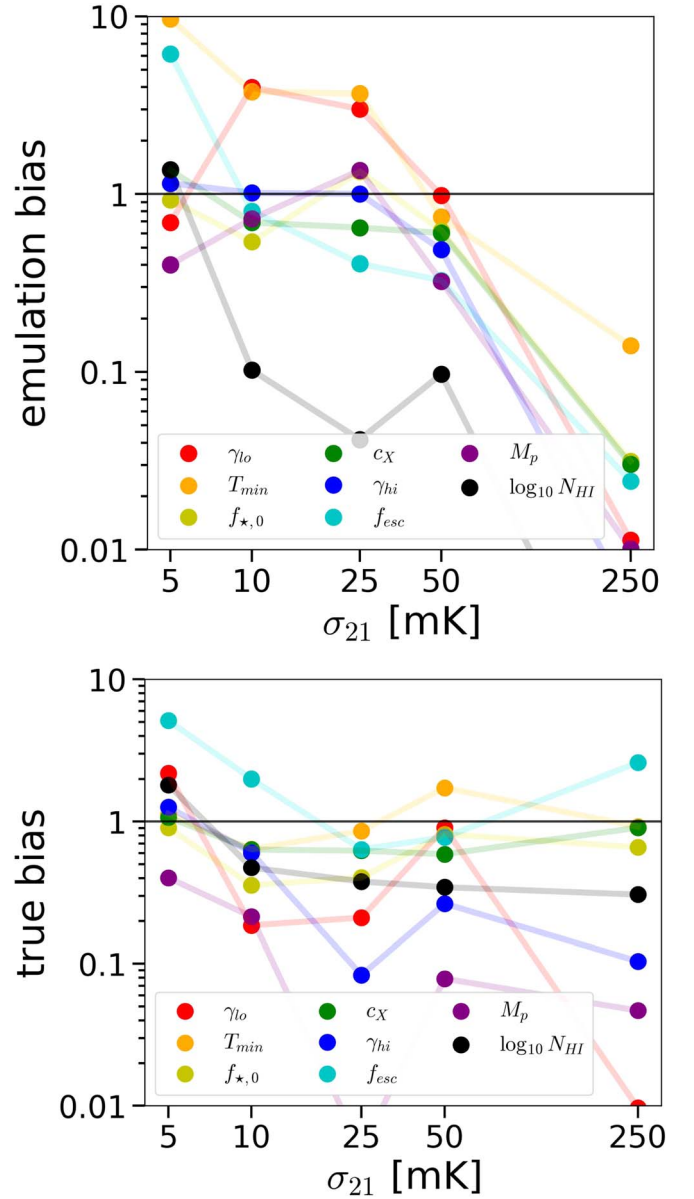
Figure 5. Top: emulation bias (number of standard deviations; see Equation (5)) between globalemu and ARES for different noise levels of the mock 21 cm data being jointly fit with the mock UVLF data. Generally, the emulation bias decreases as the 21 cm noise level increases. For $\sigma_{21} = 50$ and 250 mK, the emulation biases are <1 for all eight parameters, as indicated by the horizontal black line. The emulation biases for γ_{lo}, $T_{\min}$, and f_{esc} can be significantly higher than the rest for certain lower 21 cm noise levels. Bottom: true bias (Equation 6) between ARES and the fiducial parameter values, for the same joint fits. True bias is lowest at 25 mK (<1 for all parameters) and increases at high and low 21 cm noise levels owing to increased uncertainty and difficulty in sampling, respectively (see Section 3.1). As also discussed in the text, note that the high emulation bias on f_{esc} at 5 mK is dominated by its high true bias.

The relatively high emulation biases on $T_{\min}$ and γ_{lo} are due only to the globalemu posteriors being less accurate, given that the true biases on $T_{\min}$ and γ_{lo} are low (see the bottom panel of Figure 5). In contrast, the high emulation bias on f_{esc} at 5 mK is influenced by the high true bias on the ARES posterior for f_{esc}. We find that true biases $\geqslant 1$ at low 21 cm noise levels for f_{esc} and $T_{\min}$ also exist when using other samplers such as PolyChord (see Figure 4) and emcee (Foreman-Mackey et al. 2013), and so we infer that these biases could be due to

accuracy limitations of the sampling algorithms to produce unbiased constraints at very low noise levels. Future work could further explore sampling biases at such low noise levels by using other algorithms such as dynesty (Speagle 2020) and in particular UltraNest (Buchner 2016, 2019, 2021), which was created for the purpose of mitigating bias in complex posteriors.

We can also compare the final evidences output from the nested sampling analyses and compute the Bayes factor (i.e., the ratio of evidences, or difference of log evidences) to select the favored model given the data and priors (Trotta 2008). For $\sigma_{21} = 25$ mK, the Bayes factor between globalemu and ARES is 0.6; for 50 mK, it is 2.7; and for 250 mK, it is 1.2. The natural logarithm of these Bayes factors being <1 indicates that there is no preference for one model over the other in fitting the mock data (see, e.g., Kass & Raftery 1995; Jeffreys 1998; Trotta 2008).

In Figure 4, we see that as σ_{21} increases, the non-SFE posteriors become less constrained around the fiducial value, except for $\log N_{HI}$, which is unconstrained at all noise levels. At high/pessimistic σ_{21}, the 21 cm data provide much less constraining power, which causes degeneracies in the eight-dimensional ARES parameter space to grow larger (i.e., the space becomes flatter). This subsequently widens the posterior distributions for those parameters that are most sensitive to the 21 cm data (see also Section 3.3). For the 21 cm noise level of 250 mK, the true biases are ≈ 1 for c_X and T_{min} and ≈ 2.5 for f_{esc}. In contrast, for $\sigma_{21} = 10$, 25, and 50 mK, there is no true bias $\geqslant 1$, except for T_{min} at 50 mK and f_{esc} at 10 mK, which each have true bias of ≈ 2 (see bottom panel of Figure 5).

As briefly mentioned, we performed one joint fit using the PolyChord nested sampling algorithm to compare the result to an equivalent joint fit using MultiNest. In Figure 4, the posteriors from PolyChord are shown as dotted green histograms, and the equivalent posteriors from MultiNest are shown in solid yellow. For the 21 cm data being fit we assume the optimistic noise level $\sigma_{21} = 10$ mK, and for the UVLF we assume twice the error on the $z = 5.9$ UVLF measurements from Bouwens et al. (2015) (i.e., "2xB+15"). We use "2xB+15" UVLF error instead of "B+15" because this allows the PolyChord run to converge in a more reasonable amount of time. In addition, we find that doing so has no effect on the non-SFE posteriors and only slightly increases the width of the SFE posteriors. We find close agreement between the posterior distributions and final evidences (see Table 2) obtained when using PolyChord versus those when using MultiNest. Comparing the two runs, we find that PolyChord required 28 times more likelihood evaluations to reach roughly the same result (with an acceptance rate of 0.38% vs. 8.7% for MultiNest; see Table 2). PolyChord, however, is expected to become more efficient than MultiNest for a larger number of parameters (Handley et al. 2015a) and could thus be a better choice for 21 cm analyses including additional free parameters to account for systematics such as the beam-weighted foreground, RFI, subsurface conditions, etc.

3.2. Fitting Individual Mock Data Sets

In Figure 6, we present the posterior distributions when separately fitting our individual mock data sets. When fitting only the 21 cm data, using either the full ARES model (in black) or the ARES-trained globalemu model (in red) for $\sigma_{21} = 50$ mK, the posterior presents large degeneracies and in general larger true biases than the corresponding joint fit at the

same σ_{21} (shown in Figure B1 in Appendix B). In particular, for the SFE parameters, bimodalities and degeneracies exist when fitting only the global signal that are removed when jointly fitting the UVLF (see Section 3.1). Among the four SFE parameters, M_p and γ_{hi} are the least constrained when fitting only the 21 cm data. This is expected because these two parameters control the brightest sources, which contribute relatively little to the global photon budget, making the global signal rather insensitive to these parameters and motivating the inclusion of the UVLF data to aid these constraints. In addition, even though the posteriors of the non-SFE parameters, c_X, f_{esc}, T_{min}, and $\log N_{HI}$, remain largely the same after adding the UVLF data, the joint fit does significantly reduce the presence of long tails in these parameters, in particular for f_{esc} and T_{min}.

When only fitting the UVLF data (green posterior in Figure 6), we find as expected strong constraints on the SFE parameters and a lack of constraints on the rest. This is because the ARES UVLF model only depends on the four SFE parameters and is independent of the other four. The green together with the black or red posteriors in Figure 6 illustrate how jointly fitting the UVLF with the 21 cm data is expected to break significant degeneracies in this parameter space, to obtain the tight constraints shown in Figure 3.

Comparing the red and black constraints from the 21 cm data in Figure 6, we find that using the ARES-trained globalemu model produces rather similar 1D and 2D posterior distributions to those from the full ARES model, with all emulation biases <1, except for f_{esc}, which has an emulation bias of ≈ 1. As stated in Table 2, the runs using globalemu and ARES reach nearly the same final evidence, further demonstrating the agreement between the two results. This close agreement shows that globalemu is able to represent the ARES parameter space more easily when the constraints are significantly weaker with respect to those from the joint fit with the UVLF data.

3.3. Posterior Consistency

Bayesian consistency of a posterior distribution is the concept that as the number of data observations grows, the posterior distribution converges on the truth (Schwartz 1965).[12] A posterior is considered consistent if it eventually concentrates on the true parameter value as the number of degrees of freedom in the data vector increases to infinity. As shown in Figure 4, we observe posterior consistency when comparing the 1D posteriors obtained for decreasing levels of the 21 cm noise: larger integration times result in posteriors generally becoming more peaked around the input, fiducial values (marked by blue lines). As briefly mentioned in Section 3.1, for lower integration times (i.e., higher σ_{21}), the 21 cm data provide relatively little constraining power, which grows the covariance in the multidimensional parameter space, producing probability density biases.[13] As expected from Bayesian consistency, we thus find that the posteriors are more biased from their fiducial values at increasing noise levels.

Posterior consistency is most apparent for these four parameters: c_X, f_{esc}, T_{min}, and γ_{lo}. Their pessimistic noise level posteriors ($\sigma_{21} = 250$ mK; gray in Figure 4) are clearly not centered on their fiducial values, presenting a relatively

---

[12] We also refer the reader to Prof. Surya Tapas Tokdar's notes on Bayesian consistency: http://www2.stat.duke.edu/~st118/sta941/Asymp.pdf.

[13] Additionally, this should produce further departures from the standard assumptions taken in building the likelihood in Equation (2) (see, e.g., Prelogović & Mesinger 2023) and likely contribute to bias in the posterior.
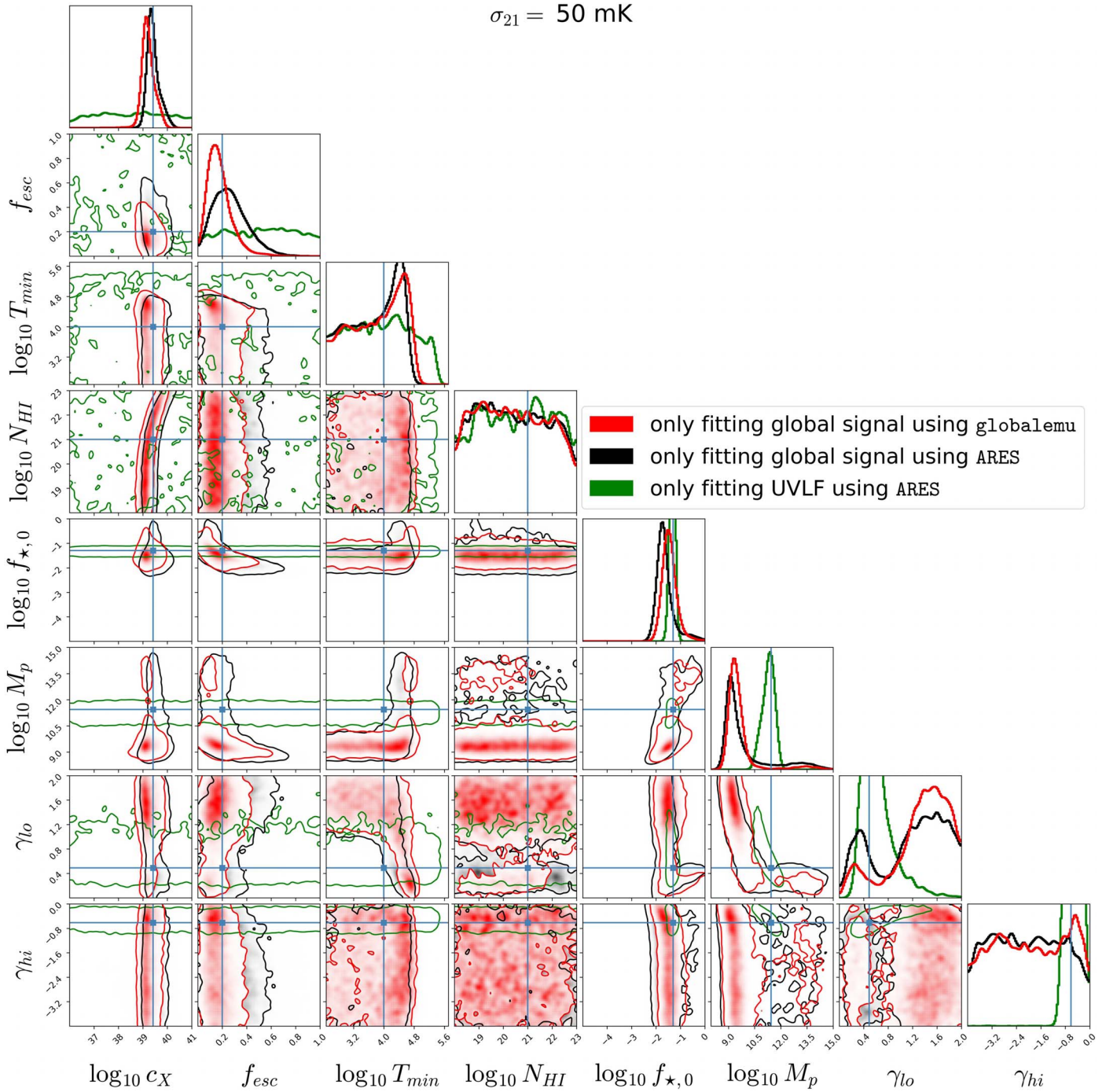
Figure 6. Marginalized 1D and 2D posterior distributions obtained when fitting either mock global 21 cm signal data (red and black) or mock UVLF data (green). All is the same as in Figure 3, except that the statistical noise in the 21 cm data being fit is $\sigma_{21} = 50$ mK, and the axis ranges are the full prior ranges given in Table 1. See Table 2 for further details on each fit.

slow "rate of convergence," while the three SFE parameters $f_{\star,0}$, M_p, and γ_{hi} have faster rates of convergence and thus require less integration time to concentrate on their input, fiducial values. As also shown in the triangle plots above, $\log N_{HI}$ remains largely unconstrained for all the noise levels, though `globalemu` still accurately emulates its posterior.

4. Conclusions

In this paper, we present the 1D and 2D posterior distributions for eight astrophysical parameters in `ARES` obtained when fitting mock data of the global 21 cm signal and/or the high-z galaxy UVLF via nested sampling. We compare for the first time the posteriors obtained from a global 21 cm signal emulator to those obtained using the full model on which it is trained, at various 21 cm noise levels. Use of an emulator such as `globalemu` is desirable, as it speeds up model evaluations by several orders of magnitude, but the accuracy of such constraints is poorly understood. The eight parameters employed control in `ARES` the SFE and the efficiency of UV and X-ray photon production per unit star formation in galaxies (see Table 1).

We assess the accuracy of the parameter constraints obtained by an ARES-trained globalemu network and determine for which parameters and 21 cm noise levels globalemu is biased compared to ARES. We test optimistic, standard, and pessimistic 21 cm noise levels ranging between $\sigma_{21} = 5$ and 250 mK to show the astrophysical constraints that can be expected for non-systematics-limited 21 cm experiments. We optimize the accuracy of the trained globalemu network by testing multiple network architectures and training set sizes, obtaining a mean RMSE between the emulated and true ARES signals in the test set of 1.25 mK.

We find that adding the UVLF to the 21 cm data provides significant improvements to the constraints on the four SFE parameters, and it has little to no effect on the constraints on the non-SFE parameters. These results imply that combining 21 cm observations with HST and JWST measurements of the UVLF at different redshifts may provide key insights into the suggested redshift evolution of the SFE and the degree of stochasticity.

The ARES-trained globalemu model produces relatively accurate posteriors with respect to the "true" ARES model at the tested 21 cm noise levels, in both shape and mean, except for the following. In particular, T_{min} and γ_{lo} present significant emulation biases at $\sigma_{21} = 25$ mK or lower, for which globalemu overpredicts T_{min} and underpredicts γ_{lo} by $\approx 3\sigma - 4\sigma$ (see the top panel of Figure 5, and Figure 3 for the full posterior distributions), except for at $\sigma_{21} = 5$ mK, where γ_{lo} has a negligible bias. For noise levels of $\sigma_{21} = 50$ and 250 mK, the globalemu emulator reproduces the posterior means found by ARES at the 68% confidence level for all eight parameters (see the top panel of Figure 5, and Appendix B for the full posterior distributions).

When examining the 1D posteriors obtained from joint fits at various noise levels in Figure 4, we find that as the noise in the 21 cm data decreases, the 1D posteriors become more concentrated around their input, fiducial values, as expected for "posterior consistency." For standard noise levels of $\sigma_{21} = 25$ and 50 mK, the true biases for all parameters are <1, except for at $\sigma_{21} = 50$ mK, where T_{min} has a true bias of ≈ 1.5. For the pessimistic noise level of $\sigma_{21} = 250$ mK, three parameters (c_X, T_{min}, and f_{esc}) have "true" ARES posterior means that are $\approx 1\sigma - 3\sigma$ away from their fiducial value (i.e., have "true biases" $\approx 1-3$; see the bottom panel of Figure 5). This indicates a slow rate of convergence for these parameter fits and the need for a longer integration time to achieve posteriors centered around the true value.

In summary, this work provides insights on the statistical constraints that are achievable from global 21 cm measurements in combination with high-z UVLF data when using an emulator. We obtain strong constraints on eight ARES parameters when jointly fitting such data using either the full ARES model or an ARES-trained globalemu model. The most accurate ARES constraints are achieved for a 21 cm noise level of 25 mK, where all eight ARES parameter means are within 1σ of their fiducial values. At this noise level, however, globalemu overpredicts T_{min} and underpredicts γ_{lo}. For larger noise levels of 50 and 250 mK, while in general the true biases increase, the emulated and true posteriors match more closely such that their parameter means are within 1σ of each other.

Acknowledgments

Appendix A
K-S Tests Comparing Emulated and "True" Posterior Distributions

To further evaluate the accuracy of the ARES-trained globalemu model, we also perform a two-sample Kolmogorov–Smirnov (K-S) test on each pair of emulated and "true" 1D posteriors obtained from joint fits, shown in Figure 3, B1, and B2. We compute the cumulative distribution functions (CDFs) for each 1D marginalized posterior probability distribution function (48 total: 16 from $\sigma_{21} = 25$ mK, 16 from $\sigma_{21} = 50$ mK, and 16 from $\sigma_{21} = 250$ mK) and employ scipy.stats to calculate the K-S statistics and associated p-values for each of the 24 pairs of CDFs.

We find that the p-values are all >0.05, except for at $\sigma_{21} = 250$ mK, where two parameters (M_p and γ_{hi}) have $p < 0.05$. For six of the eight parameters, the p-values are highest at $\sigma_{21} = 50$ mK, which is the same 21 cm noise level that we found gives the most similar means of the emulated and "true" posteriors (Figure 5). We also computed the K-S statistics between the emulated and true 1D posteriors obtained when fitting only the global signal (Figure 6) and found that $p > 0.05$ for all eight parameters. Therefore, based on the K-S tests, we conclude that the null hypothesis that the emulated and "true" posteriors originate from the same parent distribution is not rejected. We note that even though globalemu is deemed to be a good representation of ARES based on K-S tests and the Bayes factor (see Section 3.1), significant biases exist on the emulated posteriors for T_{min} and γ_{lo}, in particular at lower 21 cm noise levels of $\sigma_{21} \lesssim 25$ mK (see Figure 5).

Appendix B
Posterior Distributions from Joint Fits for 21 cm Noise Levels of 50 and 250 mK

In Figures B1 and B2, we present the full posterior distributions obtained from mock 21 cm and UVLF data joint fits for 21 cm noise levels of $\sigma_{21} = 50$ and 250 mK, respectively. As discussed in Section 3.1, we found that these noise levels give the best match (i.e., most similar parameter means) between the emulated and "true" posteriors (see the top panel of Figure 5). The 1D posteriors for $\sigma_{21} = 250$ mK are also shown in Figure 4.
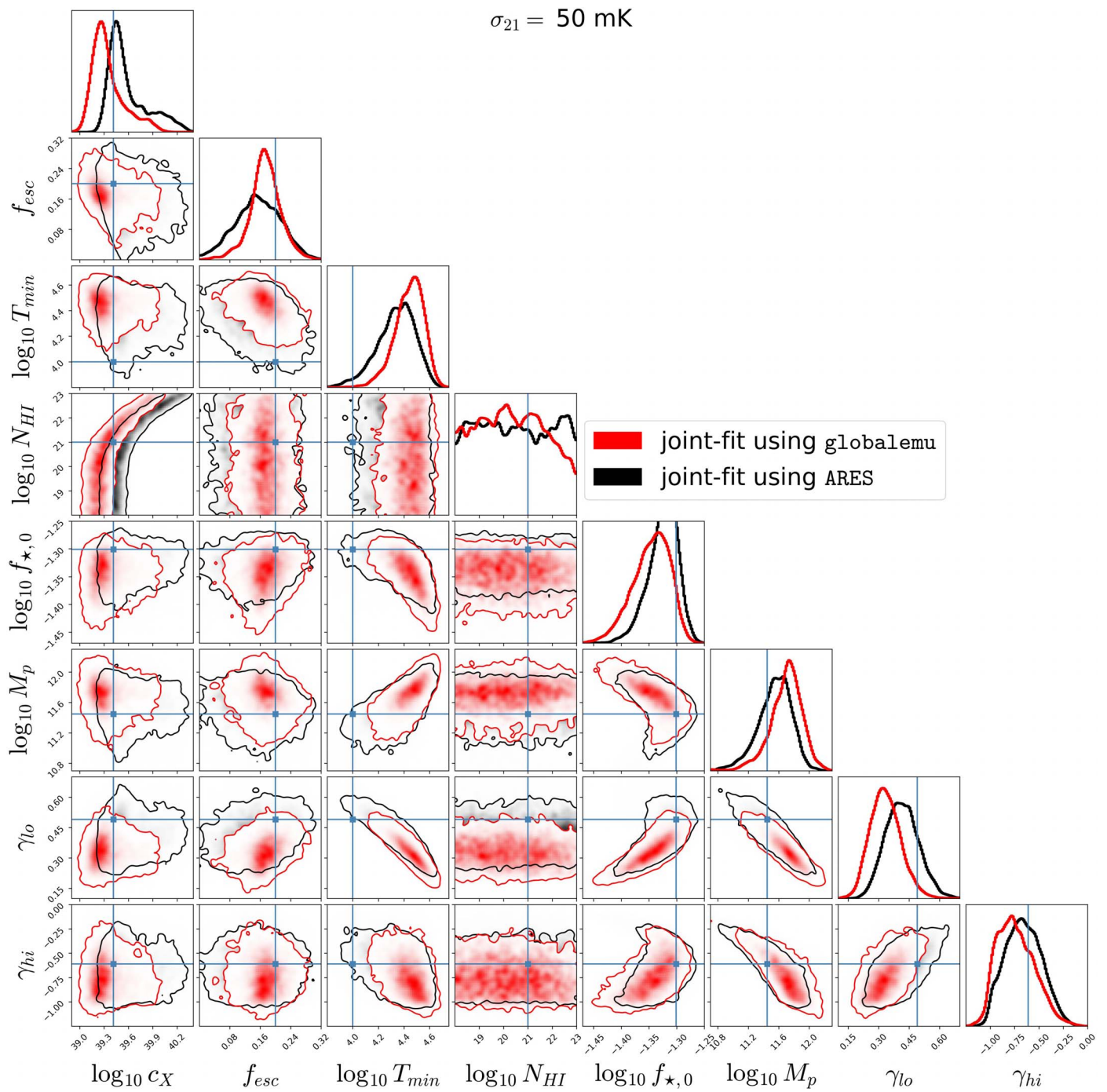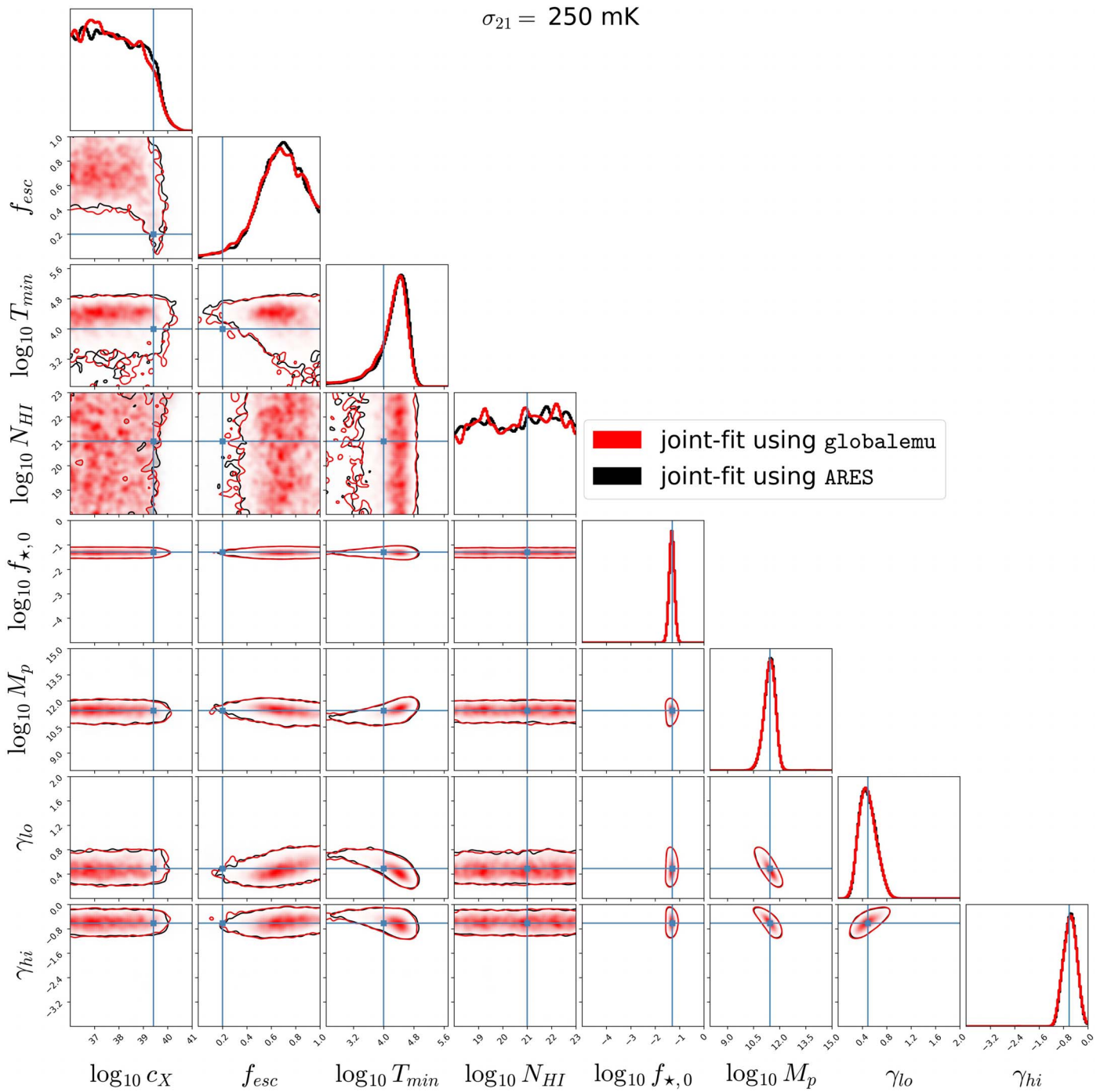
Figure B1. Marginalized 1D and 2D posterior distributions for eight astrophysical parameters in ARES when jointly fitting mock global 21 cm signal and UVLF data. All is the same as in Figure 3, except that the noise in the 21 cm data being fit is $\sigma_{21} = 50$ mK.

Figure B2. Marginalized 1D and 2D posterior distributions for eight astrophysical parameters in ARES when jointly fitting mock global 21 cm signal and UVLF data. All is the same as in Figures 3 and B1, except that the noise in the 21 cm data being fit is $\sigma_{21} = 250$ mK, and the axis ranges are zoomed out to show the full prior ranges.

ORCID iDs

J. Dorigo Jones https://orcid.org/0000-0002-3292-9784
D. Rapetti https://orcid.org/0000-0003-2196-6675
J. Mirocha https://orcid.org/0000-0002-8802-5581
J. J. Hibbard https://orcid.org/0000-0002-9377-5133
J. O. Burns https://orcid.org/0000-0002-4468-2117
N. Bassett https://orcid.org/0000-0001-7051-6385

References

Anstey, D., de Lera Acedo, E., & Handley, W. 2023, MNRAS, 520, 850
Ashton, G., Bernstein, N., Buchner, J., et al. 2022, NRvMP, 2, 39
Bale, S. D., Bassett, N., Burns, J. O., et al. 2023, arXiv:2301.10345
Bassett, N., Rapetti, D., Burns, J. O., Tauscher, K., & MacDowall, R. 2020, AdSpR, 66, 1265
Bassett, N., Rapetti, D., Tauscher, K., et al. 2021, APJ, 923, 33
Bera, A., Ghara, R., Chatterjee, A., Datta, K. K., & Samui, S. 2023, JApA, 44, 10
Bernardi, G., Zwart, J. T. L., Price, D., et al. 2016, MNRAS, 461, 2847
Bevins, H., & Gessey-Jones, T. 2023, htjb/globalemu: globalemu v1.8.0, Zenodo, doi:10.5281/zenodo.8178850
Bevins, H. T. J., de Lera Acedo, E., Fialkov, A., et al. 2022a, MNRAS, 513, 4507
Bevins, H. T. J., Fialkov, A., de Lera Acedo, E., et al. 2022b, NatAs, 6, 1473
Bevins, H. T. J., Handley, W. J., Fialkov, A., de Lera Acedo, E., & Javid, K. 2021, MNRAS, 508, 2923
Bevins, H. T. J., Heimersheim, S., Abril-Cabezas, I., et al. 2023, arXiv:2301.03298
Bond, J. R., Cole, S., Efstathiou, G., & Kaiser, N. 1991, ApJ, 379, 440
Bouwens, R., Illingworth, G., Oesch, P., et al. 2023, MNRAS, 523, 1009
Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2015, ApJ, 803, 34
Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J., & Mahesh, N. 2018, Natur, 555, 67
Boylan-Kolchin, M. 2023, NatAs, 7, 731
Bradley, R. F., Tauscher, K., Rapetti, D., & Burns, J. O. 2019, ApJ, 874, 153
Breitman, D., Mesinger, A., Murray, S., et al. 2023, arXiv:2309.05697
Buchner, J. 2016, S&C, 26, 383
Buchner, J. 2019, PASP, 131, 108005
Buchner, J. 2021, JOSS, 6, 3001
Buchner, J. 2023, Stat. Surv., 17, 169
Burns, J., Hallinan, G., Chang, T. C., et al. 2021a, arXiv:2103.08623
Burns, J. O., MacDowall, R., Bale, S., et al. 2021b, PSJ, 2, 44
Bye, C. H., Portillo, S. K. N., & Fialkov, A. 2022, ApJ, 930, 79
Chatterjee, A., Choudhury, T. R., & Mitra, S. 2021, MNRAS, 507, 2405
Cohen, A., Fialkov, A., Barkana, R., & Monsalve, R. A. 2020, MNRAS, 495, 4845
Cranmer, K., Brehmer, J., & Louppe, G. 2020, PNAS, 117, 30055
Das, A., Mesinger, A., Pallottini, A., Ferrara, A., & Wise, J. H. 2017, MNRAS, 469, 1166
de Lera Acedo, E., de Villiers, D. I. L., Razavi-Ghods, N., et al. 2022, NatAs, 6, 984
Donnan, C. T., McLeod, D. J., Dunlop, J. S., et al. 2023, MNRAS, 518, 6011
Eldridge, J. J., & Stanway, E. R. 2009, MNRAS, 400, 1019
Feroz, F., & Hobson, M. P. 2008, MNRAS, 384, 449
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2019, OJAp, 2, 10
Fialkov, A., & Barkana, R. 2014, MNRAS, 445, 213
Finkelstein, S. L., Bagley, M. B., Ferguson, H. C., et al. 2023, ApJL, 946, L13
Foreman-Mackey, D. 2016, JOSS, 1, 24
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Foreman-Mackey, D., Vousden, W., Price-Whelan, A., et al. 2016, corner.py: corner.py v2.0.0, Zenodo, doi:10.5281/zenodo.53155
Furlanetto, S. R., Mirocha, J., Mebane, R. H., & Sun, G. 2017, MNRAS, 472, 1576
Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, PhR, 433, 181
Garsden, H., Greenhill, L., Bernardi, G., et al. 2021, MNRAS, 506, 5802
Ghara, R., Choudhury, T. R., & Datta, K. K. 2015, MNRAS, 447, 1806
Ghara, R., Mellema, G., Giri, S. K., et al. 2018, MNRAS, 476, 1741
Handley, W., 2020 PolyChord: Next Generation Nested Sampling,, Zenodo, doi:10.5281/zenodo.3598030
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
Harikane, Y., Ouchi, M., Oguri, M., et al. 2023, ApJS, 265, 5

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Heavens, A. 2016, Entrp, 18, 236
The HERA Collaboration, Abdurashidova, Z., Adams, T., et al. 2022, arXiv:2210.04912
Hibbard, J. J., Mirocha, J., Rapetti, D., et al. 2022, ApJ, 929, 151
Hibbard, J. J., Rapetti, D., Burns, J. O., Mahesh, N., & Bassett, N. 2023, arXiv:2304.09959
Hibbard, J. J., Tauscher, K., Rapetti, D., & Burns, J. O. 2020, ApJ, 905, 113
Hills, R., Kulkarni, G., Meerburg, P. D., & Puchwein, E. 2018, Natur, 564, E32
Hobson, M. P., Bridle, S. L., & Lahav, O. 2002, MNRAS, 335, 377
Hunter, J. D. 2007, CSE, 9, 90
Hutter, A., Trebitsch, M., Dayal, P., et al. 2023, MNRAS, 524, 6124
Jeffreys, H. 1998, The Theory of Probability (Oxford: Oxford Univ. Press)
Kass, R. E., & Raftery, A. E. 1995, J. Am. Stat. Assoc., 90, 773
Kern, N. S., Liu, A., Parsons, A. R., Mesinger, A., & Greig, B. 2017, ApJ, 848, 23
Kern, N. S., Parsons, A. R., Dillon, J. S., et al. 2020, ApJ, 888, 70
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, ed. F. Loizides & B. Scmidt (Netherlands: IOS Press)
Lahav, O., Bridle, S. L., Hobson, M. P., Lasenby, A. N., & Sodré, L. 2000, MNRAS, 315, L45
Leeney, S. A. K., Handley, W. J., & de Lera Acedo, E. 2022, arXiv:2211.15448
Lemos, P., Weaverdyck, N., Rollins, R. P., et al. 2023, MNRAS, 521, 1184
Liu, A., Pritchard, J. R., Tegmark, M., & Loeb, A. 2013, PhRvD, 87, 043002
Liu, A., & Shaw, J. R. 2020, PASP, 132, 062001
Lovell, C. C., Harrison, I., Harikane, Y., Tacchella, S., & Wilkins, S. M. 2023, MNRAS, 518, 2511
Madau, P., Meiksin, A., & Rees, M. J. 1997, ApJ, 475, 429
Mason, C. A., Muñoz, J. B., Greig, B., Mesinger, A., & Park, J. 2023a, MNRAS, 524, 4711
Mason, C. A., Trenti, M., & Treu, T. 2023b, MNRAS, 521, 497
Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, MNRAS, 493, 1662
Mesinger, A., Furlanetto, S., & Cen, R. 2011, MNRAS, 411, 955
Mineo, S., Gilfanov, M., & Sunyaev, R. 2012, MNRAS, 419, 2095
Mirocha, J. 2014, MNRAS, 443, 1211
Mirocha, J., & Furlanetto, S. R. 2019, MNRAS, 483, 1980
Mirocha, J., & Furlanetto, S. R. 2023, MNRAS, 519, 843
Mirocha, J., Furlanetto, S. R., & Sun, G. 2017, MNRAS, 464, 1365
Mirocha, J., La Plante, P., & Liu, A. 2021, MNRAS, 507, 3872
Mirocha, J., Skory, S., Burns, J. O., & Wise, J. H. 2012, ApJ, 756, 94
Mitsuda, K., Inoue, H., Koyama, K., et al. 1984, PASJ, 36, 741
Monsalve, R. A., Fialkov, A., Bowman, J. D., et al. 2019, ApJ, 875, 67
Monsalve, R. A., Greig, B., Bowman, J. D., et al. 2018, ApJ, 863, 11
Muñoz, J. B., Dvorkin, C., & Cyr-Racine, F. Y. 2020, PhRvD, 101, 063526
Murray, S., Greig, B., Mesinger, A., et al. 2020, JOSS, 5, 2582
Murray, S. G., Bowman, J. D., Sims, P. H., et al. 2022, MNRAS, 517, 2264
Murray, S. G., Power, C., & Robotham, A. S. G. 2013, A&C, 3, 23
Naidu, R. P., Oesch, P. A., van Dokkum, P., et al. 2022, ApJL, 940, L14
Paciga, G., Chang, T. C., Gupta, Y., et al. 2011, MNRAS, 413, 1174
Pagano, M., Sims, P., Liu, A., et al. 2022, arXiv:2211.10448
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6
Polidan, R., Lopez, J., & Burns, J. 2022, AAS Meeting, 54, 206.10
Prelogović, D., & Mesinger, A. 2023, MNRAS, 524, 4239
Press, W. H., & Schechter, P. 1974, ApJ, 187, 425
Qin, Y., Mesinger, A., Park, J., Greig, B., & Muñoz, J. B. 2020, MNRAS, 495, 123
Rapetti, D., Tauscher, K., Mirocha, J., & Burns, J. O. 2020, ApJ, 897, 174
Ryan, B., Stevenson, B., Trendafilova, C., & Meyers, J. 2023, PhRvD, 107, 103506
Santos, M. G., Ferramacho, L., Silva, M. B., Amblard, A., & Cooray, A. 2010, MNRAS, 406, 2421
Saxena, A., Meerburg, P. D., de Lera Acedo, E., et al. 2023a, MNRAS, 522, 1022
Saxena, A., Cole, A., Gazagnes, S., et al. 2023b, MNRAS, 525, 6097
Schaeffer, T., Giri, S. K., & Schneider, A. 2023, MNRAS, 526, 2942
Schneider, A., Schaeffer, T., & Giri, S. K. 2023, PhRvD, 108, 043030
Schwartz, L. 1965, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 4, 10
Sellentin, E., Quartin, M., & Amendola, L. 2014, MNRAS, 441, 1831
Shaver, P. A., Windhorst, R. A., Madau, P., & de Bruyn, A. G. 1999, A&A, 345, 380
Shen, E., Anstey, D., de Lera Acedo, E., & Fialkov, A. 2022, MNRAS, 515, 4565

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

Singh, S., Jishnu, N. T., Subrahmanyan, R., et al. 2022, NatAs, 6, 607

Singh, S., Subrahmanyan, R., Udaya Shankar, N., et al. 2018, ApJ, 858, 54

Skilling, J. 2004, in AIP Conf. Ser. 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th Int. Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, ed. R. Fischer, R. Preuss, & U. V. Toussaint (Melville, NY: AIP), 395

Speagle, J. S. 2020, MNRAS, 493, 3132

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

Tauscher, K., Rapetti, D., Nhan, B. D., et al. 2021, ApJ, 915, 66

Thomas, R. M., Zaroubi, S., Ciardi, B., et al. 2009, MNRAS, 393, 32

Trott, C. M., Jordan, C. H., Midgley, S., et al. 2020, MNRAS, 493, 4711

Trotta, R. 2008, ConPh, 49, 71

Van Rossum, G., & Drake, F. L., Jr 1995, Python Reference Manual (Amsterdam: Centrum voor Wiskunde en Informatica)

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261

Zhang, Z., Shan, H., Gu, J., et al. 2022, MNRAS, 516, 1573