Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **527,** 9833–9852 (2024) https://doi.org/10.1093/mnras/stad3849
Advance Access publication 2023 December 13

21CMEMU: an emulator of 21CMFAST summary observables

Daniela Breitman,[1★] Andrei Mesinger,[1,2] Steven G. Murray,[1,3] David Prelogović,[1] Yuxiang Qin[4,5] and Roberto Trotta[6,7,2]

[1] *Scuola Normale Superiore (SNS), Piazza dei Cavalieri 7, I-56125 Pisa, PI, Italy*
[2] *Centro Nazionale 'High Performance Computer, Big Data and Quantum Computing'*
[3] *School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85287-1404, USA*
[4] *School of Physics, University of Melbourne, Parkville, VIC 3010, Australia*
[5] *ARC Centre of Excellence for All-Sky Astrophysics in 3 Dimensions (ASTRO 3D)*
[6] *Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, I-34136 Trieste, Italy*
[7] *Imperial Centre for Inference and Cosmology (ICIC), Imperial College, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*

Accepted 2023 December 11. Received 2023 October 30; in original form 2023 September 12

ABSTRACT

Recent years have witnessed rapid progress in observations of the epoch of reionization (EoR). These have enabled high-dimensional inference of galaxy and intergalactic medium (IGM) properties during the first billion years of our Universe. However, even using efficient, seminumerical simulations, traditional inference approaches that compute 3D lightcones on-the-fly can take 10^5 core hours. Here we present 21CMEMU: an emulator of several summary observables from the popular 21CMFAST simulation code. 21CMEMU takes as input nine parameters characterizing EoR galaxies, and outputs the following summary statistics: (i) the IGM mean neutral fraction; (ii) the 21-cm power spectrum; (iii) the mean 21-cm spin temperature; (iv) the sky-averaged (global) 21-cm signal; (vi) the ultraviolet (UV) luminosity functions (LFs); and (vii) the Thomson scattering optical depth to the cosmic microwave background (CMB). All observables are predicted with sub- per cent median accuracy, with a reduction of the computational cost by a factor of over 10^4. After validating inference results, we showcase a few applications, including: (i) quantifying the relative constraining power of different observational data sets; (ii) seeing how recent claims of a late EoR impact previous inferences; and (iii) forecasting upcoming constraints from the sixth observing season of the *Hydrogen Epoch of Reionization Array* (*HERA*) telescope. 21CMEMU is publicly available, and is included as an alternative simulator in the public 21CMMC sampler.

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

1 INTRODUCTION

The cosmic dawn (CD) of the first luminous objects and eventual reionization of the Universe remain among the greatest mysteries in modern cosmology. Recent years have seen a dramatic increase in observations of the CD and epoch of reionization (EoR). These include: (i) the Ly α forest (e.g. Fan et al. 2006; Becker, Rauch & Sargent 2007; Becker et al. 2015; Bosman et al. 2018; D'Odorico et al. 2023); (ii) damping wings in quasar spectra (e.g. Bolton et al. 2011; Mortlock et al. 2011; Bañados et al. 2018; Wang et al. 2020; Yang et al. 2020); (iii) Ly α emission from galaxies (e.g. Ouchi et al. 2010; Clément et al. 2012; Konno et al. 2014; Drake et al. 2017; Hoag et al. 2019; Shibuya et al. 2019); (iv) large-scale polarization of the cosmic microwave background (CMB; e.g. Planck Collaboration 2020; de Belsunce et al. 2021; Heinrich & Hu 2021); (v) secondary kinetic Sunaev–Zeldovich (kSZ) CMB anisotropies (e.g. Das et al. 2014; George et al. 2015; Reichardt et al. 2021); (vi) upper limits on the cosmic 21-cm power spectrum (PS; e.g. Mertens et al. 2020; Trott et al. 2020; The HERA Collaboration 2022a, b). This is set to culminate with a 3D map of H I during the first billion years, expected with the upcoming Square Kilometer Array (SKA; e.g. Mellema et al. 2013; Koopmans et al. 2015; Mesinger 2019).

In step with these observational advances, Bayesian inference techniques have been developed that allow us to forward model the observations and constrain the parameters of reionization as well as the galaxies responsible (e.g. Choudhury & Ferrara 2005; Greig & Mesinger 2015; Mesinger et al. 2015; Mason et al. 2018; Greig, Mesinger & Bañados 2019; Ghara et al. 2020; Mondal et al. 2020; Choudhury, Paranjape & Bosman 2021; Qin et al. 2021; Abdurashidova et al. 2022; Maity & Choudhury 2022; Nikolić et al. 2023). These rely on efficient simulators, so-called seminumerical simulations (e.g. Mesinger & Furlanetto 2007; Thomas et al. 2009; Santos et al. 2010; Mesinger, Furlanetto & Cen 2011; Visbal et al. 2012; Ghara, Choudhury & Datta 2015; Murray et al. 2020; Maity & Choudhury 2022; Schneider et al. 2022; Trac et al. 2022), that typically approximate computationally expensive radiative transfer with approaches based on cheap fast fourier transforms. However, inference can be computationally expensive even with semi-numerical simulations. As an example, the recent, state-of-the-art inference

★ E-mail: daniela.breitman@sns.it

using nine galaxy parameters in Abdurashidova et al. (2022; hereafter HERA22) took $\sim 10^5$ core hours on an HPC center: roughly 10^5 likelihood evaluations each taking ~ 1 core hour to simulate the corresponding observables.

A popular alternative approach is to use emulators (e.g. Kern et al. 2017; Schmit & Pritchard 2017; Shimabukuro & Semelin 2017; Jennings et al. 2019; Ghara et al. 2020; Mondal et al. 2022; Bye, Portillo & Fialkov 2022; Lazare, Sarkar & Kovetz 2023). Once trained on a set of simulation outputs, an emulator can replace the expensive, on-the-fly simulation step in Bayesian inference: a single likelihood evaluation taking ~ 0.1 s instead of ~ 1 h. As such, the computational cost is *amortized*, requiring only the initial data base of simulations in order to perform subsequent, inexpensive inferences.[1] Of course, such amortized inference is restricted to the theoretical model that is used to train the emulator. Moreover, there is also the additional emulator error to account for, which can be non-negligible for high precision measurements and in corners of parameter space that are poorly sampled (e.g. Kern et al. 2017; appendix B of HERA22). Nevertheless, emulators allow us to rapidly perform many inferences of the same model, testing the impact on the posterior of different likelihood choices, priors, and new data. Moreover, the emulator error is sub-dominant compared with current, relatively low signal-to-noise ratio (S/N) observations, such as the 21-cm power spectrum upper limits.

Here we present 21CMEMU[2] – a public emulator of several summary outputs from the seminumerical code 21CMFAST.[3] These include (i) the volume-averaged hydrogen neutral fraction; (ii) the 21-cm power spectrum; (iii) the global 21cm brightness temperature; (iv) the neutral intergalactic medium (IGM) spin temperature; (v) the ultraviolet (UV) luminosity functions (LFs); (vi) the Thomson scattering optical depth of CMB photons. Our emulator was trained on summary observables from the *withHERA* inference in HERA22, which sampled nine astrophysical parameters that characterize galaxy properties. As a result, our work presents a few important improvements over previous emulators. The unprecedented number of summary outputs allows us to include complementary multiwavelength probes of high-z galaxies and the EoR when computing the likelihood. Moreover, our physically motivated galaxy parametrization (Park et al. 2019) allows us to easily motivate different choices of priors. We will periodically update 21CMEMU to include new summary outputs and astrophysical models.

We showcase our emulator by re-analysing the HERA power spectrum upper limits published in HERA22. We also perform inferences including various combinations of the data, illustrating the constraining power of each probe on the posterior. One call of 21CMEMU takes ~0.1 s (compared to ~1 h for 21CMFAST), with a typical inference finishing in a few hours.

This paper is organized as follows. In Section 2, we introduce the data used to train the emulator. In Section 3, we introduce the network and discuss its architecture, training procedure, and performance. In Section 4, we showcase applications of the emulator to EoR/CD inference problems. We conclude in Section 5. We assume a flat ΛCDM cosmology, with $(\Omega_\Lambda, \Omega_m, \Omega_b, h, \sigma_8, n_s) = (0.69, 0.31, 0.049, 0.68, 0.82, 0.97)$, consistent with results from (Planck Collaboration 2020). Unless stated otherwise, all lengths are in comoving units.

2 SIMULATED DATA SET

Our data sets for training and testing are taken from the *withHERA* inference from HERA22, using an increased number of livepoints (18k). This inference used the Multinest (Feroz, Hobson & Bridges 2009) sampler in 21CMMC[4] (Greig & Mesinger 2015, 2017, 2018) with a flat prior on all astrophysical parameters within the ranges shown in all of the corner plots (e.g. Fig. 6). The likelihood was determined by current observations of the EoR history, galaxy LFs and 21cm upper limits (discussed in detail in Section 4.1).

We use all of the MULTINEST outputs, including both accepted and rejected samples, resulting in 1.8M parameter samples. Of these, we randomly select 1.28M for training, 183k for validation and 330k for testing. The data base is standardized (subtract the mean and divide by the standard deviation of each summary statistic) before being passed into the network for training.

Our data sets are generated with the public 21CMFAST v3 code (Mesinger & Furlanetto 2007; Mesinger, Furlanetto & Cen 2011; Murray et al. 2020). 21CMFAST is a seminumerical simulation code that operates under the assumption that dark matter halos host galaxies which source inhomogeneous large-scale cosmic radiation fields. Matter density and velocity fields are generated using second-order Lagrangian perturbation theory (e.g. Scoccimarro 1998). Galaxy properties are assigned to dark matter halo fields using empirical scaling relations, following the parametrization in Park et al. (2019). The ionizing, X-ray, and soft UV cosmic radiation fields sourced by these galaxies are computed with a combination of excursion set and direct integration along the lightcone. The ionization and thermal state of the IGM gas are then tracked with a set of coupled differential equations, allowing us to compute the various observables discussed below. The HERA22 runs that form our data base assumed a simulation box length of 250 cMpc, with a 128^3 grid. For further details on the simulation code, the interested reader is directed to (Mesinger & Furlanetto 2007; Mesinger, Furlanetto & Cen 2011; Murray et al. 2020). Below we summarize the astrophysical parameters used as input to 21CMEMU, and the summary observables that are the corresponding output.

2.1 Galaxy model and astrophysical parameters

The input consists of nine parameters that characterize bulk galaxy properties. Two parameters $(f_{*,10}, \alpha_*)$ describe the stellar–to–halo mass relation (SHMR), which is a power law for the faint galaxies (hosted by $M_h \lesssim 10^{12} M_\odot$ halos) that dominate the cosmic radiation fields at $z > 5$ (e.g. Kuhlen & Faucher-Giguère 2012; Dayal et al. 2014; Behroozi & Silk 2015; Mutch et al. 2016; Sun & Furlanetto 2016; Yue, Ferrara & Xu 2016),

$$\frac{M_*}{M_h}(M_h) = f_{*,10}\left(\frac{M_h}{M_{10}}\right)^{\alpha_*}\left(\frac{\Omega_{\rm b}}{\Omega_{\rm m}}\right). \tag{1}$$

Here $\Omega_{\rm b}$ is the universal baryon energy density (as a fraction of the critical energy density), $\Omega_{\rm m}$ is the total matter (i.e. cold dark matter and baryon) energy density, and $f_* \equiv f_{*,10}\left(\frac{M_h}{M_{10}}\right)^{\alpha_*} \in [0, 1]$ is the stellar fraction, with $f_{*,10}$ corresponding to the fraction of galactic

[1] Another form of amortized inference is to train neural density estimators to fit the likelihood or likelihood/evidence ratio using simulated data (e.g. Alsing, Wandelt & Feeney 2018; Papamakarios, Sterratt & Murray 2018; Cole et al. 2022). This is referred to as simulation-based inference (SBI), and has the additional benefit of not having to specify an explicit functional form for the likelihood. SBI has recently been applied to mock 21cm observations (Zhao et al. 2022a; Zhao, Mao & Wandelt 2022b; Prelogović & Mesinger 2023; Saxena et al. 2023), with very promising results.

[2] https://github.com/21cmfast/21cmEMU

[3] https://github.com/21cmfast/21cmFAST

[4] https://github.com/21cmfast/21CMMC

gas in stars normalized to the amount in a halo of mass $M_{10} \equiv 10^{10}$ $\mathrm{M}_{\odot}$, and α_* the power-law index.

Star formation is assumed to occur on a time-scale that goes with the Hubble time, $H^{-1}(z)$ (or analogously the dynamical time, which also scales with the Hubble time during matter domination),

$$\dot{M}_* = \frac{M_*}{t_* H^{-1}(z)}. \tag{2}$$

The characteristic star formation time-scale, $t_* \in [0, 1]$, is another free parameter.

The typical ionizing escape fraction, $f_{\mathrm{esc}}(M_h) \in [0, 1]$ is similarly described by a power law (e.g. Paardekooper, Khochfar & Dalla Vecchia 2015; Kimm et al. 2017; Lewis et al. 2020),

$$f_{\mathrm{esc}}(M_h) = f_{\mathrm{esc},10}\left(\frac{M_h}{M_{10}}\right)^{\alpha_{\mathrm{esc}}}, \tag{3}$$

with two free parameters: the normalization, $f_{\mathrm{esc},\,10}$, and the power-law index, α_{esc}.

Star formation is suppressed in small mass halos due to inefficient gas cooling and/or feedback (e.g. Hui & Gnedin 1997; Springel & Hernquist 2003; Okamoto, Gao & Theuns 2008; Sobacchi & Mesinger 2013; Xu et al. 2016; Ma et al. 2020; Ocvirk et al. 2020). We account for this suppression by assuming only a fraction $\exp\left(-M_{\mathrm{turn}}/M_h\right)$ of halos host active star forming galaxies. The characteristic halo mass scale below which the abundance of galaxies is exponentially suppressed, M_{turn}, is another free parameter.

The specific X-ray luminosity escaping the galaxies is also taken to be a power law in energy (e.g. Das et al. 2017), $L_X \propto E^{-\alpha_X}$, with the index α_X left as a free parameter. The luminosity is normalized via the soft band X-ray luminosity per unit star formation rate (SFR), another free parameter,

$$L_{X<2\mathrm{keV}}/\mathrm{SFR} = \int_{E_0}^{2\mathrm{keV}} \mathrm{d}E\; L_X/\mathrm{SFR}, \tag{4}$$

where E_0, the last input parameter, is the minimum energy of X-ray photons capable of escaping their host galaxy.

In summary, the nine input parameters are:

(i) $\boldsymbol{f_{*,10}}$: normalization of the SHMR, defined at $M_h = M_{10}$.
(ii) $\boldsymbol{\alpha_*}$: power-law index of the SHMR.
(iii) $\boldsymbol{f_{\mathrm{esc},10}}$: normalization of the ionizing escape fraction to halo mass relation, defined at $M_h = M_{10}$.
(iv) $\boldsymbol{\alpha_{\mathrm{esc}}}$: power-law index of the ionizing UV escape fraction to halo mass relation.
(v) $\boldsymbol{t_*}$: characteristic star formation time-scale, defined as a fraction of the Hubble time.
(vi) $\boldsymbol{M_{\mathrm{turn}}/\mathrm{M}_{\odot}}$: characteristic mass below which halos become exponentially less likely to host an active star forming galaxy.
(vii) $\boldsymbol{\frac{L_{X<2\mathrm{keV}}/\mathrm{SFR}}{\mathrm{erg\ s^{-1}\ M_{\odot}^{-1}\ yr}}}$: soft-band X-ray luminosity per unit SFR escaping the galaxies.
(viii) $\boldsymbol{E_0/\mathrm{keV}}$: minimum X-ray energy that can escape the galaxies.
(ix) $\boldsymbol{\alpha_X}$: power-law index of the X-ray spectral energy distribution.

This simple parametrization is easy to interpret physically and is consistent with observations of the UV LFs as well as the scaling relations found in galaxy simulations and semi-analytic models.

2.2 Observational summaries

For a given set of cosmological and astrophysical parameters, 21CMFAST calculates the corresponding 3D lightcones of IGM properties. When performing inference, these lightcones are generally compressed into summary statistics that are compared directly with observations. Here we do not attempt to directly emulate the 3D lightcones of the various cosmological quantities, and instead only emulate the following summary observables (motivated by existing EoR/CD observations discussed in Section 4.1):

(i) $\boldsymbol{\overline{x}_{\mathrm{HI}}(z)}$ – the volume-averaged neutral fraction of hydrogen and helium as a function of redshift (also commonly referred to as the EoR history).

(ii) $\boldsymbol{\overline{T}_{\mathrm{b}}(z)}$ – the volume-averaged (global) 21-cm brightness temperature (e.g. Madau, Meiksin & Rees 1997; Furlanetto 2006; Pritchard & Loeb 2012):

$$\begin{aligned} T_{\mathrm{b}}(\boldsymbol{x}, z) &= \frac{T_{\mathrm{S}} - T_{\mathrm{R}}}{1+z}(1 - e^{\tau_{21}}) \\ &\approx 27\, x_{\mathrm{HI}}(1+\delta_b)\left(\frac{\Omega_b h^2}{0.023}\right)\left(\frac{0.15}{\Omega_m h^2}\frac{1+z}{10}\right)^{1/2}\ \mathrm{mK} \\ &\quad \times \left(\frac{T_{\mathrm{S}} - T_{\mathrm{R}}}{T_{\mathrm{S}}}\right)\left[\frac{\partial_r v_r}{(1+z)H(z)}\right], \end{aligned} \tag{5}$$

where τ_{21} is the 21cm optical depth of the intervening gas, $\delta_b \equiv \rho/\bar{\rho} - 1$ is the baryon overdensity, with ρ being the baryon density, and T_{S} and T_{R} are the spin and background temperatures, respectively. We assume throughout that the radio background is provided by the CMB, $T_{\mathrm{R}} = T_{\mathrm{CMB}}$ is the temperature of the CMB. We note that 21CMFAST computes the brightness temperature at each cell location, $\mathbf{x}$, using the exact expression in the first line of the equation above; the second line is a Taylor expansion in the limit of $\tau_{21} \ll 1$ that provides physical intuition.

(iii) $\boldsymbol{\overline{T}_{\mathrm{S}}(z)}$ – the mean spin temperature of the neutral IGM as a function of redshift. The IGM spin temperature is only defined for neutral hydrogen that is outside of the cosmic H II regions that surround galaxies. Specifically, the volume average is performed over those cells in the simulation box with $\overline{x}_{\mathrm{HI}} \geq 95$ per cent.

(iv) $\boldsymbol{\Delta^2_{21}(k, z)}$ – spherically averaged 21-cm PS: $\Delta^2_{21}(k, z)\left[\mathrm{mK}^2\right] \equiv k^3/(2\pi^2)\langle \tilde{T}_b \tilde{T}_b^* \rangle$, where $k = |\boldsymbol{k}|$, and $\tilde{T}_b(\boldsymbol{k}, z)$ is the Fourier dual of the brightness temperature from equation (5).

(v) $\boldsymbol{\phi(M_{1500}, z)}$ – the non-ionizing UV LF, defined as the number density of galaxies per UV magnitude, M_{1500}, as a function of redshift. The ~1500 Å rest frame luminosity is calculated from the SFR: $\dot{M}_*(M_h, z) = \mathcal{K}_{\mathrm{UV}} \times L_{\mathrm{UV}}$, where $\mathcal{K}_{\mathrm{UV}} = 1.15 \cdot 10^{-28} \mathrm{M}_{\odot}\mathrm{yr}^{-1}\ \mathrm{Hz\ s\ erg}^{-1}$ assumes a Salpeter initial mass function (e.g. Madau & Dickinson 2014; Sun & Furlanetto 2016). The UV luminosity is related to the AB magnitude using (Oke & Gunn 1983): $\log\left(\frac{L_{\mathrm{UV}}}{\mathrm{erg\ s^{-1}\ Hz^{-1}}}\right) = 0.4 \times (51.63 - M_{\mathrm{UV}})$.

(vi) $\boldsymbol{\tau_{\mathrm{e}}}$ – the Thompson optical depth to the last scattering surface: $\tau_e = \sigma_T \int_0^{z_{\mathrm{LSS}}} dz \left|\frac{cdt}{dz}\right| n_e$, where σ_T is the Thompson scattering cross section and n_e is the electron number density calculated assuming hydrogen and helium are singly ionized at a fraction $(1 - \overline{x}_{\mathrm{HI}})$ and that helium is doubly ionized at $z < 3$.

Although the last two quantities are computed analytically by 21CMFAST and are therefore reasonably fast, we still emulate them for two reasons. The first is to provide users of 21CMEMU with a standalone package. The second is that the analytic calculation is still slower than the emulator prediction time: emulation reduces the runtime from ~1 s to ≲ 50 ms for a single parameter combination (≲ 1 ms per parameter set if in a large (≳ 100) batch), with a relatively low emulation error (see Section 3).

We use 84 redshift bins in the range $z \sim 5$–35 for all summaries except the 21-cm PS. For the 21-cm PS we exclude high redshift bins that generally have a very weak signal, keeping 60 redshift bins spanning $z \sim 6$–21, and 12 k bins spanning $k \sim 0.04$–1 Mpc^{-1}. We

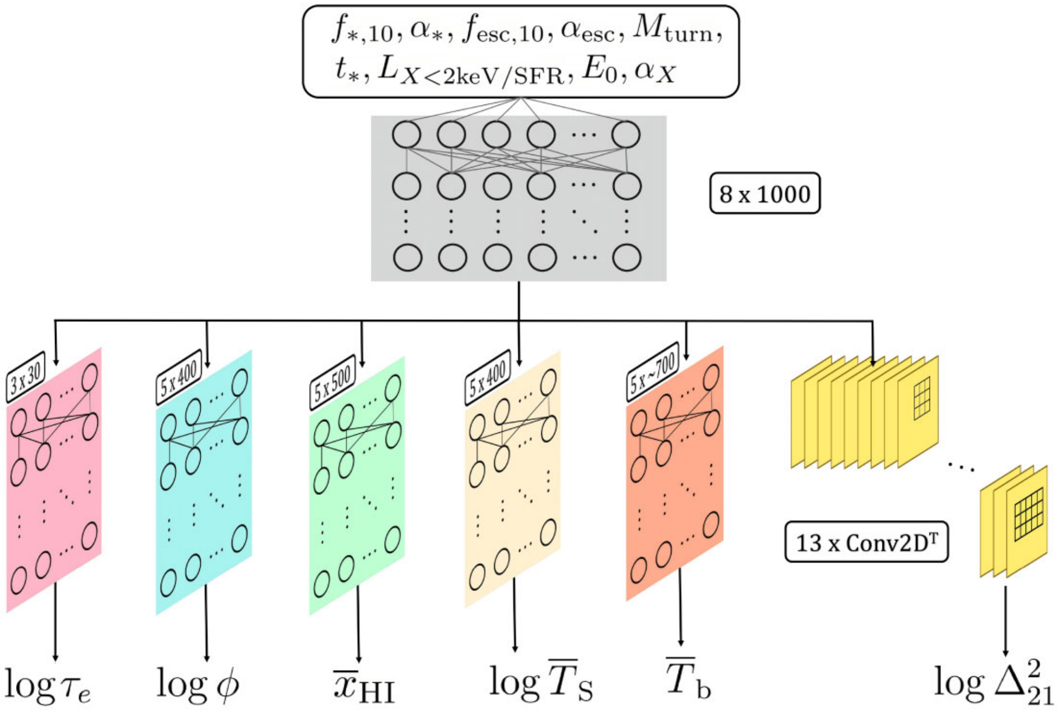
$f_{*,10}, \alpha_*, f_{\mathrm{esc},10}, \alpha_{\mathrm{esc}}, M_{\mathrm{turn}}, t_*, L_{X<2\mathrm{keV}/\mathrm{SFR}}, E_0, \alpha_X$
8 x 1000
3 x 30
5 x 400
5 x 500
5 x 400
5 x ~700
13 x Conv2D$^{\mathrm{T}}$
$\log \tau_e$
$\log \phi$
$\overline{x}_{\mathrm{HI}}$
$\log \overline{T}_{\mathrm{S}}$
$\overline{T}_{\mathrm{b}}$
$\log \Delta^2_{21}$

Figure 1. Schematic of the 21CMEMU architecture. Astrophysical parameters (*top*; c.f. Section 2.1 are inputted through a large block of fully connected layers. The output from this shared block is then passed on into five blocks (much smaller than the shared block). The first four fully-connected branches, from left to right, output the Thomson scattering optical depth, UV LFs, mean hydrogen neutral fraction, spin temperature, and global signal, respectively. The output from the shared block is also reshaped into an image and is passed into a 2D convolutional neutral network which outputs the 21-cm power spectrum (rightmost branch). The convolutions gradually build the PS image. The window size varies among the layers. The number of filters (stacked squares) decreases toward the end of the convolutional neural network.

also floor the PS values to 0.1 mK^2, in order to reduce the dynamic range of the data and improve training. We note that the value of the floor is an order of magnitude smaller than the accuracy of the 21CMFAST simulator itself (e.g. Mesinger, Furlanetto & Cen 2011; Zahn et al. 2011), and thus has no effective impact on the accuracy of our emulator.

3 EMULATOR ARCHITECTURE AND PERFORMANCE

21CMEMU is implemented using `Tensorflow` (Abadi et al. 2015), with an architecture consisting of (see diagram in Fig. 1):

(i) one large block (eight layers with 1k nodes each) of fully connected (dense) layers whose output is fed into all of the branches.
(ii) one branch per summary observable.

Since the 21-cm power spectrum is a smooth function of wave-mode and redshift (e.g. Fig. 3), it can be interpreted as a 2D image. Therefore we use a convolutional neural network (CNN) in the 21-cm PS branch and fully connected layers in the other branches. Note that the branches are not connected with one another. The only nodes they have in common are those from the main block which each branch receives as input.

The network trains on all of the summaries at once (i.e. multitask learning), using a weighted sum of root mean squared error (RMSE) losses with one loss term per branch, where each branch loss has a different weight. We assign the largest weight to the 21-cm PS branch as it is higher dimensional with the largest dynamic range, and thus more difficult to learn. The final set of weights chosen is obtained from a trial of about 50 different weight combinations with the goal of choosing the best weights such that the 21-cm power spectrum, brightness temperature, and neutral fraction are learned best. The performance of the other summaries is not significantly affected by the choice of weights. These trials are done ad hoc since the training is computationally expensive.

We perform a few tests to motivate the importance of the block of fully connected layers. First, we train a network equivalent to the brightness temperature branch alone i.e. whose input is the astrophysical parameters that are straight away passed into the brightness temperature branch of fully connected layers. We find that the median brightness temperature fractional error over the test set in this network is ~ 45 per cent larger than the one in the final network.[5] This means that, on average,[6] our final architecture performs better than just having individual networks for each summary. The final architecture can contribute to improving the performance in two ways: (i) combining the losses of the summaries allows the network to learn from the correlations between the summaries; and (ii) simply making the network larger and deeper. To test the relative importance of (i) and (ii), we train a network without the shared block but with

[5] We also test slightly changing the brightness temperature branch itself: adding an additional layer and increasing the number of nodes increases the median fractional error (see below for specific definition) by about 50 per cent, while increasing the number of layer nodes slightly and adding one additional layer increases it by about 20 per cent.

[6] We did not perform this test for all of the other summaries. We did perform it for the 21-cm power spectrum and found that the performance of the final network is a few per cent better than that of the CNN branch alone.

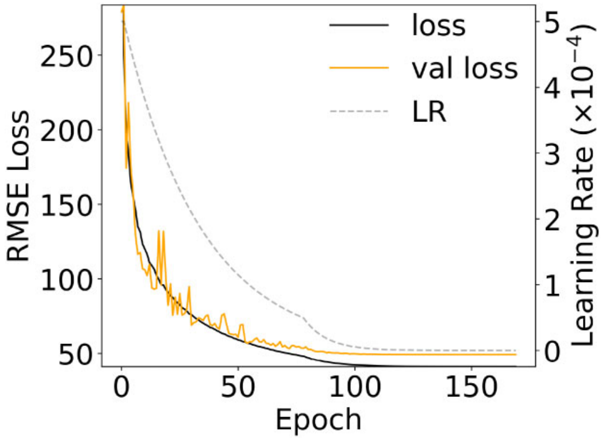

Figure 2. Training loss (black line) and validation loss (orange line) as a function of training epoch. The learning rate curve is also shown with the dashed grey line and the corresponding right axis.

the rest of the architecture the same. This significantly reduces the number of trainable parameters in the network (by about 50 per cent), but still allows different summaries to influence each other through the shared loss. We do see an increase of up to a few per cent in the median and 68 per cent CL of the fractional error for the smaller network as expected. Most notably, for the brightness temperature we see an increase of ∼ 1 per cent, ∼13 per cent, and ∼27 per cent for the median, 68 per cent CL, and 95 per cent CL of the fractional error, respectively. We conclude that combining the losses of all the summaries is the main cause of performance improvement, while the large shared block is needed to get the best performance for the most challenging summaries: the brightness temperature and 21-cm power spectrum.

In Fig. 2, we show the total training and validation losses as a function of epoch in black and orange, respectively. We also show the learning rate schedule used during training with a dashed grey line. We see a smooth decline of the validation loss up to ∼100 epochs. Our final network is taken at the minimum of the validation loss, at epoch 150. The training takes about eleven GPU hours (∼3.5 min per epoch) with the full data base (1.8M samples).

Below, we discuss the branch architecture and performance for each summary observable in turn, summarizing the results in Table 1. Throughout, we illustrate the emulator performance using examples from the test set, as well as the distributions of absolute differences (Abs Diff) and fractional errors (FE) over the entire test set. The latter two are defined for each observational summary, y, as

$$\text{Abs Diff} \equiv |\text{y}_{\text{true}} - \text{y}_{\text{pred}}| \tag{6}$$

$$\text{FE}(\%) \equiv \frac{\text{Abs Diff}}{\max\left(|\text{y}_{\text{true}}|, \text{y}_{\text{floor}}\right)}, \tag{7}$$

where y_{true} refers to the 21CMFAST direct simulation output and y_{pred} is the corresponding 21CMEMU prediction. We compute the above averaged over different bins in y and/or different models in the test set, as described below. One drawback of the FE metric is that it can diverge to infinity as the denominator goes to zero. To avoid this, we use floors for the values of the denominator: $\log(\Delta^2_{21,\text{floor}}) = 0.1$; $\bar{T}_{b,\text{floor}} = 5$ mK, and $\bar{x}_{\text{HI,floor}} = 10^{-4}$. The specific values of these floors was chosen relatively arbitrarily; however, they are lower than the expected accuracy achievable by any near term experiment.[7] The other summaries, τ_{e}, $\overline{T}_{\text{S}}$, and UV LFs do not have a floor value.

Table 1. Performance of the 21CMEMU network when trained on the full data base, half of the data base and 1 per cent of the data base.

Training size	Summary	Median FE (per cent)	68 per cent CL (per cent)
1.3M Full	$\log\Delta^2_{21}$	0.55	2.4
	$\overline{T}_{\text{b}}$	0.34	1.2
	$\log\overline{T}_{\text{S}}$	0.032	0.13
	$\overline{x}_{\text{HI}}$	0.0073	0.10
	τ_{e}	0.11	0.26
	$\log\phi$	0.50	2.1
640k Random	$\log\Delta^2_{21}$	0.71	3.0
	$\overline{T}_{\text{b}}$	0.43	1.51
	$\log\overline{T}_{\text{S}}$	0.047	0.17
	$\overline{x}_{\text{HI}}$	0.0086	0.12
	τ_{e}	0.15	0.35
	$\log\phi$	0.57	2.5
13k Random	$\log\Delta^2_{21}$	3.2	13.0
	$\overline{T}_{\text{b}}$	4.8	16.6
	$\log\overline{T}_{\text{S}}$	0.40	1.2
	$\overline{x}_{\text{HI}}$	0.035	0.57
	τ_{e}	0.45	1.0
	$\log\phi$	2.5	10.0

3.1 The 21-cm power spectrum

The power spectrum branch consists of 13 2D convolution layers with wide (up to seven redshift bins × 3k bins) kernels and two upsampling layers that gradually build the (k, z) PS image based on the output of the shared block, as seen in Fig. 1. We use a pixel-based RMSE loss, weighted by the inverse of the estimated thermal noise corresponding to a 1000h SKA1-low observation (taken from Prelogović et al. 2022; for more details see section 2.2.1 in that work). Weighting by the inverse of the noise forces the CNN to be more accurate in (k, z) bins that are easier to observe: generally corresponding to lower redshifts and larger scales.

In Fig. 3, we compare the emulator prediction for the 21-cm power spectrum with its corresponding target from 21CMFAST. We show a single sample from the test set, with the 21CMEMU prediction on the left and the 21CMFAST target in the middle panel. This sample was chosen as it has the closest median fractional error to that of the entire test set; thus it can be considered representative of the typical emulator performance. It is difficult to see a difference between the two PS with the naked eye. We the FE of this single sample in the rightmost panel. The FE is generally sub-per cent, rising to ∼ per cent in regions of low power.

In these 2D images we clearly see the well-known trend of three peaks in the redshift evolution of the large-scale 21-cm PS and two

[7] For the 21-cm power spectrum for example, the expected mean noise level from thermal noise and sample variance for a 1000hr observation with the SKA1-low instrument is ≳0.1 mK^2 (e.g. see Fig. 2, bottom left panel in Kaur, Gillet & Mesinger 2020). Similarly, global signal experiments have measurement noise that is orders of magnitude larger than the floor value we chose (e.g. Murray et al. 2022; Singh et al. 2022), and are instead limited mostly by foregrounds and instrument systematics. For the mean neutral fraction, estimates have typical uncertainties of order 0.1 (see e.g. Greig et al. 2022 and references therein), orders of magnitude larger than the floor value we use.

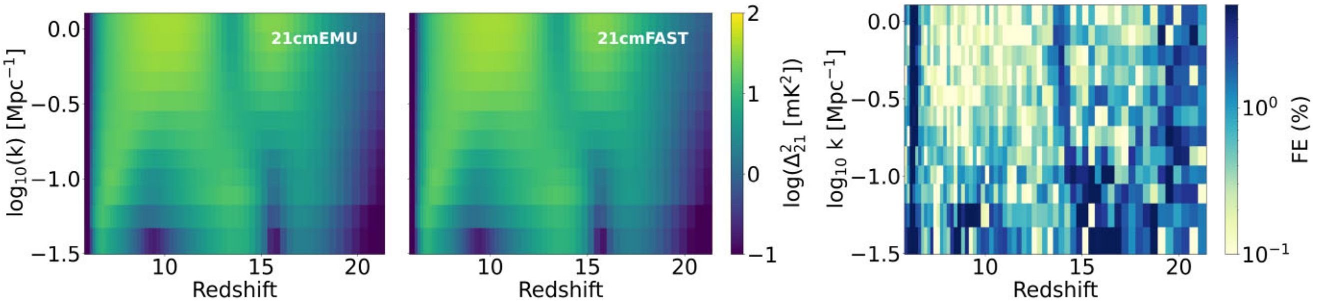

Figure 3. The spherically averaged 21-cm power spectrum as a function of wavemode and redshift for a sample in the test set. The 21CMEMU prediction is shown on the left while the 21CMFAST result is on the right. This sample has a 21-cm PS fractional errors (FE) that is roughly comparable to the median value of the whole the test set, and can thus be considered representative of the emulator performance. The rightmost panel shows the fractional error for this single sample.

peaks in the small-scale evolution (e.g. Pritchard & Furlanetto 2007). In general, the features evolve smoothly over (k, z), showcasing why we use a CNN in the 21-cm PS branch of 21CMEMU.

We quantify the 21-cm PS prediction error in the top left panel of Fig. 4. In the top sub-panel, we plot the redshift evolution of the PS amplitude at $k = 0.1$ Mpc^{-1}, with 21CMEMU predictions shown via dash–dotted lines and the corresponding 21CMFAST targets shown with solid lines. We chose to plot $k = 0.1$ Mpc^{-1} because the strongest constraints by current interferometers are around these scales; smaller scales are dominated by thermal noise and larger scales by foregrounds (e.g. Mertens et al. 2020; Trott et al. 2020; The HERA Collaboration 2022a, b). The 10 models plotted here were chosen at random from the test set. We again see that the differences between the emulator and 'truth' are difficult to spot with the naked eye.

In the bottom sub-panel we show the Abs Diff between each pair of curves in the top sub-panel, as well as the median Abs Diff (dashed black line) and the 68 per cent/|95 per cent confidence limits (CL; dark/light grey) computed over the entire test set. We see that the median (68 per cent) 21CMEMU absolute error at $k \sim 0.1$ Mpc^{-1} is $\left|\log\left(\Delta^2_{21,\mathrm{true}}/\mathrm{mK}^2\right)\text{ - }\log\left(\Delta^2_{21,\mathrm{pred}}/\mathrm{mK}^2\right)\right| \leq 0.01$ (~ 0.02). This translates to a median (68 per cent) fractional error of 0.70 per cent (1.0 per cent)[8] at this wavemode and 0.55 per cent (2.4 per cent) when averaged over all wavemodes. This is far below observational uncertainties in the near-term, thus justifying the use of an emulator. The error rises slightly at lower redshifts, owing to the broader distributions of possible PS, including very small values post reionization. In Appendix A, we show the evolution of the 21-cm power spectrum fractional error as a function of the input 9D astrophysical parameters.

3.2 The 21-cm global signal

The 21-cm global signal branch consists of seven fully connected layers with 600–1000 nodes each. We quantify the performance of 21CMEMU on the global signal in the top right panel of Fig. 4. We show the redshift evolution of the global signal (*top*) and Abs Diff (*bottom*) for the same 10 random samples from the test set.

As for the 21-cm power spectra, the difference between the 21CMFAST calculation and 21CMEMU prediction is difficult to see with the naked eye and is generally $\lesssim 1$ mK. We see from the bottom sub-panel that the 95 per cent CL of the errors in the test set is also $\lesssim 1$ mK. This translates to a median (68 per cent) FE of 0.34 per cent (1.2 per cent).

We see from both the global signal and the PS that our training set spans a wide range of heating and ionization histories. This is due to the fact that we include both accepted and rejected livepoints of the HERA22 inference in the training set, in order to have the largest data set possible. Extending beyond the ranges of the most likely models allows 21CMEMU to generalize beyond the HERA22 posterior distribution, accurately predicting even unlikely models that, e.g. have not reionized by $z = 5$.

3.3 The 21-cm spin temperature in the neutral IGM

The $\overline{T}_\mathrm{S}$ branch consists of five fully connected layers with 400 nodes each. We quantify the network performance on the mean 21-cm spin temperature in the right panel of the middle row of Fig. 4. In the top sub-panel, we show 10 examples of the emulated spin temperature curve (dash–dotted line) and the corresponding true curves from the test set (solid line). In the bottom sub-panel of the plot, we show the absolute error for each of the 10 examples, the median for the entire test set with the black dashed line, and the 68 per cent/95 per cent CL regions in shaded in dark/pale grey as a function of redshift. We can see that the Abs Diff is $\left|\log\left(\mathrm{T}_{\mathrm{S,true}}/\mathrm{K}\right) - \log\left(\mathrm{T}_{\mathrm{S,pred}}/\mathrm{K}\right)\right| < 0.01$ at 95 per cent CL over most of the redshift range. The FE of the log of the mean spin temperature over the entire test set is 0.032 per cent and the 68 per cent CL is 0.13 per cent.

We recall that the spin temperature is calculated by taking the global average over all cells in the simulation box that have $\bar{x}_\mathrm{HI} \geq 95$ per cent. When there are no cells satisfying this condition, the spin temperature becomes undefined. We account for this by having the emulator predict the redshift at which the spin temperature becomes undefined.[9] The emulator correctly predicts the exact redshift bin below which $\overline{T}_\mathrm{S}$ becomes undefined for 95.1 per cent of the models

[8] Note that these errors are calculated on the emulator PS output which is in log space. Computing the corresponding error distributions in linear space, we obtain a median (68 per cent) FE of 1.53 per cent (1.94 per cent) at $k \sim 0.1$ Mpc^{-1}, and 1.39 per cent (3.76 per cent) over the entire test set. Note that since we return to linear space, we do not need to apply a floor on the power spectrum in this FE calculation.

[9] In principle, one could use the EoR history emulator prediction to find the redshift at which the volume averaged neutral fraction drops below 0.05. However, this is not identical to our definition for $\overline{T}_\mathrm{S}$, since our simulations account for partially neutral and self-shielded clumps inside the reionized cells. Therefore we include a separate output for the redshift at which there are no cells with $\bar{x}_\mathrm{HI} \geq 95$ per cent. We note that 21CMFAST also includes partially ionized cells, both by UV and X-rays. Partial ionization by UV

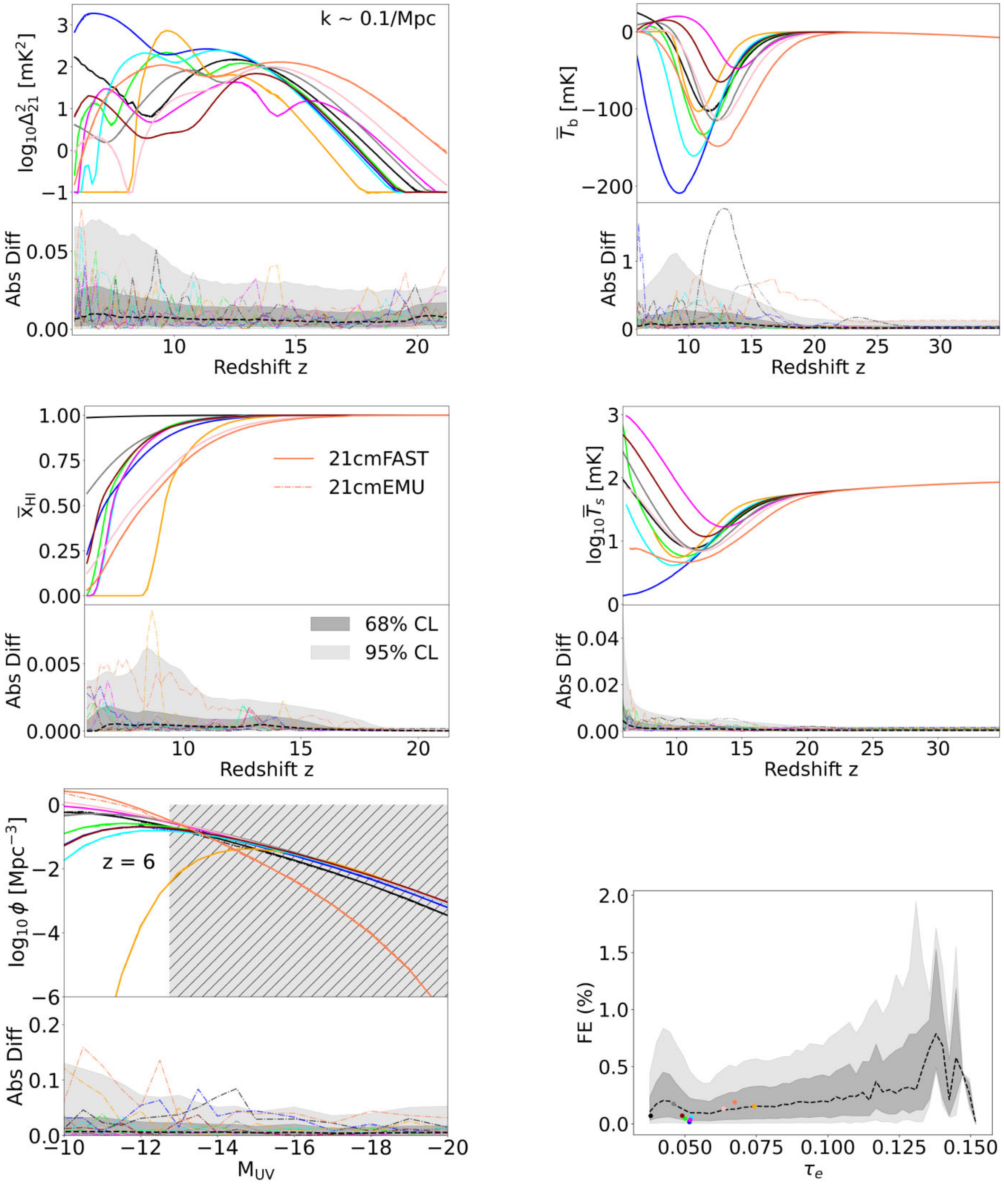

Figure 4. A subset of summary outputs from 21CMEMU for 10 random samples from the test set. Panels show: Redshift evolution of the $k = 0.1$ Mpc^{-1} 21-cm PS amplitude, redshift evolution of the mean 21-cm brightness temperature, redshift evolution of the mean spin temperature in the neutral IGM, the CMB optical depth, UV LFs at $z = 6$, the EoR history (*clockwise from upper left*). Colours denote the astrophysical parameter sample with solid (dashed) lines corresponding to outputs from 21CMFAST (21CMEMU). In the bottom sub-panels, we show the absolute differences (Abs Diff) between the predicted and true quantities shown in the top sub-panels. Abs Diff of the 10 random samples are shown with the corresponding colours, while the median Abs Diff (FE in the case of τ_e) computed over the entire test set are shown with dashed black curves. Dark (light) shaded regions enclose 68 per cent (95 per cent) CL.

in our test set, and is only one bin off ($\Delta z \sim 0.1$) for 4.89 per cent of the models.

3.4 The global history of reionization

The EoR history branch, like the spin temperature branch, consists of five fully connected hidden layers with 500 nodes each. In the left panel of the middle row in Fig. 4, we show the EoR histories of our 10 parameter samples (*top sub-panel*), and the corresponding prediction error (*bottom sub-panel*). We see that the Abs Diff are $\lesssim$0.005 for 95 per cent of the models in the test set. The FE is 0.0075 per cent for the median and 0.095 per cent at the 68 per cent CL.

3.5 The CMB Thomson scattering optical depth

The Thomson scattering optical depth branch consists of three layers of 30 nodes each as it outputs only one number. We show the FE of the $\tau_{\rm e}$ prediction in the lower right panel of Fig. 4. The 10 parameter samples are denoted with different colour dots. Over the entire test set, we see a median fractional error of 0.1 per cent and a 0.25 per cent FE at 68 per cent CL. There is a notable increase in the prediction error as well as its bin-to-bin variance toward higher values of $\tau_{\rm e}$. This is due to a small number of samples in this unlikely corner of parameter space: fewer than 1 per cent of the models in the test set have $\tau_{\rm e} > 0.11$.

3.6 Galaxy UV LFs

The LFs branch consists of five layers of 400 nodes each. The network outputs the LFs at four redshifts (z = 6, 7, 8, 10) and magnitude bins ranging from −20 to −10. In the lower left panel of Fig. 4, we show the emulated and simulated LFs at $z = 6$ (the performance at the other redshifts is comparable). The hatched region denotes the range spanned by LF observations used in the inference in the following section.

We can see that the emulator is very accurate in the flat range spanned by the existing observations, while it is less accurate around the faint-end turnover. At all of the redshift bins, we have that the Abs Diff$\left|\log\left(\phi_{\rm true}/{\rm Mpc}^{-3}\right) - \log\left(\phi_{\rm pred}/{\rm Mpc}^{-3}\right)\right| < 0.1$ over the majority of the magnitude range.

We provide an alternative setting in 21CMEMU that allows the user to skip the emulation and directly calculate the CMB optical depth and UV LFs using 21CMFAST. This improved accuracy however comes at the cost of a slower runtime: ∼ 700 ms per call compared with $\lesssim$ 50 ms using emulation.

3.7 Summary of 21CMEMU performance and context with other emulators

In Table 1, we summarize the performance of 21CMEMU for each summary in the first row, using the fiducial training set of 1.3M samples. In general, the median (68 per cent) emulation fractional error is at the level of $\lesssim$ 0.5 per cent (1 per cent). The most accurate prediction is achieved with the EoR history, most likely due to the fact that it is a monotonic and smooth function, making it easier to learn. The least accurate summary is the power spectrum, which is understandable as it is two dimensional with the largest dynamic range.

It is difficult to directly compare the performance of 21CMEMU with other emulators of EoR/CD observables, due to their different astrophysical parametrizations and training set sizes. Nevertheless, at face value 21CMEMU's accuracy is better than achievable with state-of-the-art emulators (e.g. Mondal et al. 2020; Bevins et al. 2021; Bye, Portillo & Fialkov 2022; Yoshiura, Minoda & Takahashi 2023). For example, comparing with the recent, bespoke 21-cm global signal emulator 21CMVAE (Bye, Portillo & Fialkov 2022), we obtain a factor of 2.2 (1.5) lower median (95th percentile) RMS error (see their equation 1). Our median 21-cm PS FE is a factor of ∼10–100 lower than that of the bespoke PS emulators in Kern et al. (2017) and Ghara et al. (2020), when compared over the same redshift/wavemode ranges.

This improvement in 21CMEMU over previous works could be attributed to several factors. First, we have a training set of unprecedented size: 1.3M samples. This is orders of magnitude larger than used in previous works (generally ranging from thousands to tens of thousands). We quantify how 21CMEMU's accuracy changes with the training set size in the following section.

Secondly, the improvement in power spectrum emulation could be attributed in part to our novel CNN architecture. Previous 21-cm PS emulators used only fully connected layers which are not as efficient in processing 2D images such as the PS.

Finally, the fact that 21CMEMU emulates many different observables allows the prediction of any one of these to be helped by the others. Indeed, we verified explicitly that the 21-cm PS emulation is improved when the other summary outputs are included in the loss (i.e. when all branches are trained together). In addition to improving performance, including multiple EoR/CD observables is extremely important in the current era where 21-cm observations are not strongly constraining. As we show in Section 4.2, complementary galaxy and EoR observations are needed to obtain a likelihood-dominated (as opposed to prior-dominated) posterior (see also HERA22).

3.8 Varying the size of the training set

Since 21CMEMU was trained on an uncharacteristically large training set, it is useful to see how it performs with smaller training sets. To do so, we remove some models at random, retrain 21CMEMU on the reduced training set, and test its performance on the same test set.

In Fig. 5, we plot the median FE in each summary as a function of the training set size. We normalize the FE so that unity corresponds to the fiducial, 1.3 M training set. We also explicitly list the performance using half of the data base (640k samples), and 1 per cent of the data base (13k samples) in the middle and bottom rows of Table 1.

We see that there is a sharp increase in emulator accuracy with training set size, up to a size of ∼ 100k. Doubling the size of the training set roughly doubles the emulator accuracy. This relationship flattens beyond sizes of $\gtrsim$ 100k, such that a ten-fold increase in the training set from ∼100k → 1.3 M only improves the FE by a factor of ∼two.

4 APPLICATION TO INFERENCE

In this section, we apply 21CMEMU to inference problems. We use the 21CMMC driver (Greig & Mesinger 2015), which now includes the option to use either 21CMFAST or 21CMEMU as the simulator. 21CMMC incorporates three highly parallel samplers: EMCEE (Foreman-Mackey et al. 2013), Multinest (Feroz, Hobson & Bridges 2009), and Ultranest (Buchner 2016, 2019; Buchner 2021); in this work we use the latter two as discussed further below.

is assumed to correspond to unresolved H II regions surrounding nascent galaxies (see the discussion in Zahn et al. 2011).

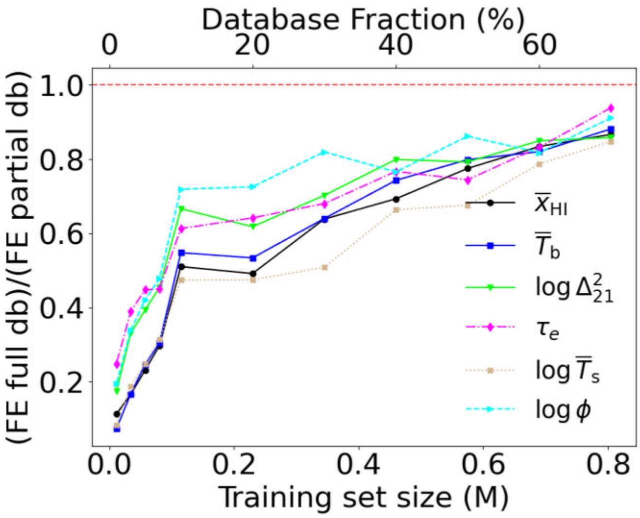

Figure 5. Median fractional error of each summary as a function of the training set size. The FE is normalized so that unity corresponds to the fiducial, 1.3 M training set.

First, we run the same inference as was previously run in HERA22 using 21CMFAST in order to see how emulator error affects the posterior. After this validation, we showcase the potential of 21CMEMU by performing several new inferences demonstrating: (i) how different observations are complimentary; (ii) the approximate impact of new, late-ending EoR constraints; (iii) the potential impact of upcoming H6C HERA observations. Each 21CMEMU inference took roughly a day on a GPU, compared with a few weeks on a cluster had we used 21CMFAST directly.

4.1 Comparison with direct simulation

We run the same inference as in HERA22 (10k livepoints) with 21CMEMU. Doing this allows us to directly compare the inference results between the two methods.

The likelihood in HERA22 incorporates four data sets:

(i) *Thomson scattering CMB optical depth* – this term compares the Thomson scattering CMB optical depth from the proposed model with the one from the analysis of Planck Collaboration (2020) by Qin et al. (2020), whose posterior is characterized by median and 68 per cent credible interval (CI): $\tau_e = 0.0569^{+0.0081}_{-0.0086}$. The likelihood function is a two-sided Gaussian.

(ii) *The Lyman forest dark fraction* – this term compares the mean neutral fraction at $z = 5.9$ with the upper limit of $\overline{x}_{\rm HI} < 0.06 \pm 0.05$ at 68 per cent CI obtained from QSO dark fraction (McGreer, Mesinger & D'Odorico 2015). The likelihood function is unity at $\overline{x}_{\rm HI}(z = 5.9) < 0.06$, decreasing as a one-sided Gaussian for higher neutral fraction values.

(iii) *UV LFs* – this term compares the model with $z = 6, 7, 8, 10$ UV LFs observed with *Hubble* (Bouwens et al. 2015, 2016; Oesch et al. 2018) in the magnitude range $M_{\rm UV} \in [-20, -10]$. This likelihood term is also a two-sided Gaussian.

(iv) *21-cm power spectrum upper limits* – this term accounts for HERA H1C 94 night observations at $z = 8$ and $z = 10$, presented in The HERA Collaboration (2022b). The likelihood is the upper limit likelihood discussed in HERA22.

These individual likelihood terms are multiplied to obtain the total likelihood. When using 21CMEMU for inference, we add the median emulator error in quadrature to the measurement uncertainties for each corresponding likelihood term.

In Figs 6 and 7, we compare posteriors obtained using 21CMFAST (*cyan*) to that using 21CMEMU (*orange*). Both were run using the `MultiNest` sampler with the same number of livepoints (10k, yielding ~60k posterior samples). In the lower left of Fig. 6 we plot the 1D and 2D marginal PDFs for our astrophysical parameters, while in the top right we plot 95 per cent CI of some of the summary observations (see caption for details). In Fig. 7 we plot the corresponding spin temperature PDFs in the two HERA bands, which was one of the main results of the HERA22 paper. We note that the 21CMFAST and 21CMEMU posteriors are nearly identical, testifying that the emulation error is fairly negligible when performing inference using current data sets. The only notable difference is in the t_* PDF, which is slightly broader when 21CMEMU is used as a simulator compared with 21CMFAST. We find no notable trends of the emulator error with this parameter, concluding the small difference could be due to stochasticity in sampling and/or a higher dimensional covariance of the emulator error.

In Fig. 6 we also include a run using 21CMEMU and the same HERA22 likelihood, but with the `UltraNest` sampler (*purple curves*; 5k livepoints, yielding ~70k posterior samples). The resulting posterior is consistent with the previous two. Interestingly, the choice of sampler (purple versus orange) results in a larger difference than the choice of simulator (orange versus cyan) for some marginal PDFs. In particular, the `UltraNest` posterior is more accurate towards the edges of the prior range, resulting in flatter posteriors at the edges: this behavior is also recovered using the EMCEE sampler as shown in Lazare, Sarkar & Kovetz (2023). Moreover, `UltraNest`'s vectorization makes it ~10× faster when using an efficient simulator like 21CMEMU. Therefore, in subsequent sections we only show posteriors generated with `UltraNest`.

We remind the reader that the emulator was trained on the HERA22 nested sampling output. This inference took ~400k core hours. Once trained however, the emulator performs amortized posterior estimation in only 225 core hours using `Multinest` or in 30 core hours using `Ultranest`.

4.2 Impact of different observations on the posterior

Having tested the emulator in the previous section, we now use it to perform multiple inferences that would be too costly with direct simulation. We begin by quantifying how the individual terms from the HERA22 likelihood discussed in the previous section affect the posterior. We do this by removing the terms one by one, and comparing the resulting posteriors in Fig. 8.

In orange we show the full HERA22 posterior from the previous section, including all likelihood terms. In green, we remove the HERA power spectrum upper limit constraint. We see that the only consequence is that the L_X/SFR parameter becomes unconstrained. In the panel on the right, we can also see the 95 per cent CI of the power spectrum and 21-cm global signal becoming wider around z ~6–10. As discussed in HERA22, the 21-cm power spectrum limits is the only measurement sensitive to the IGM temperature during the CD.

Next, if we remove constraints on the EoR history (here corresponding to the dark fraction and τ_e likelihood terms), using only the UV LFs in the likelihood, we obtain the posterior shown in blue. We see that EoR history measurements allow us to set (lose) constraints on the ionizing escape fraction (here parametrized via $f_{\rm esc,\,10}$ and $\alpha_{\rm esc}$), which disappear completely when their corresponding terms are not

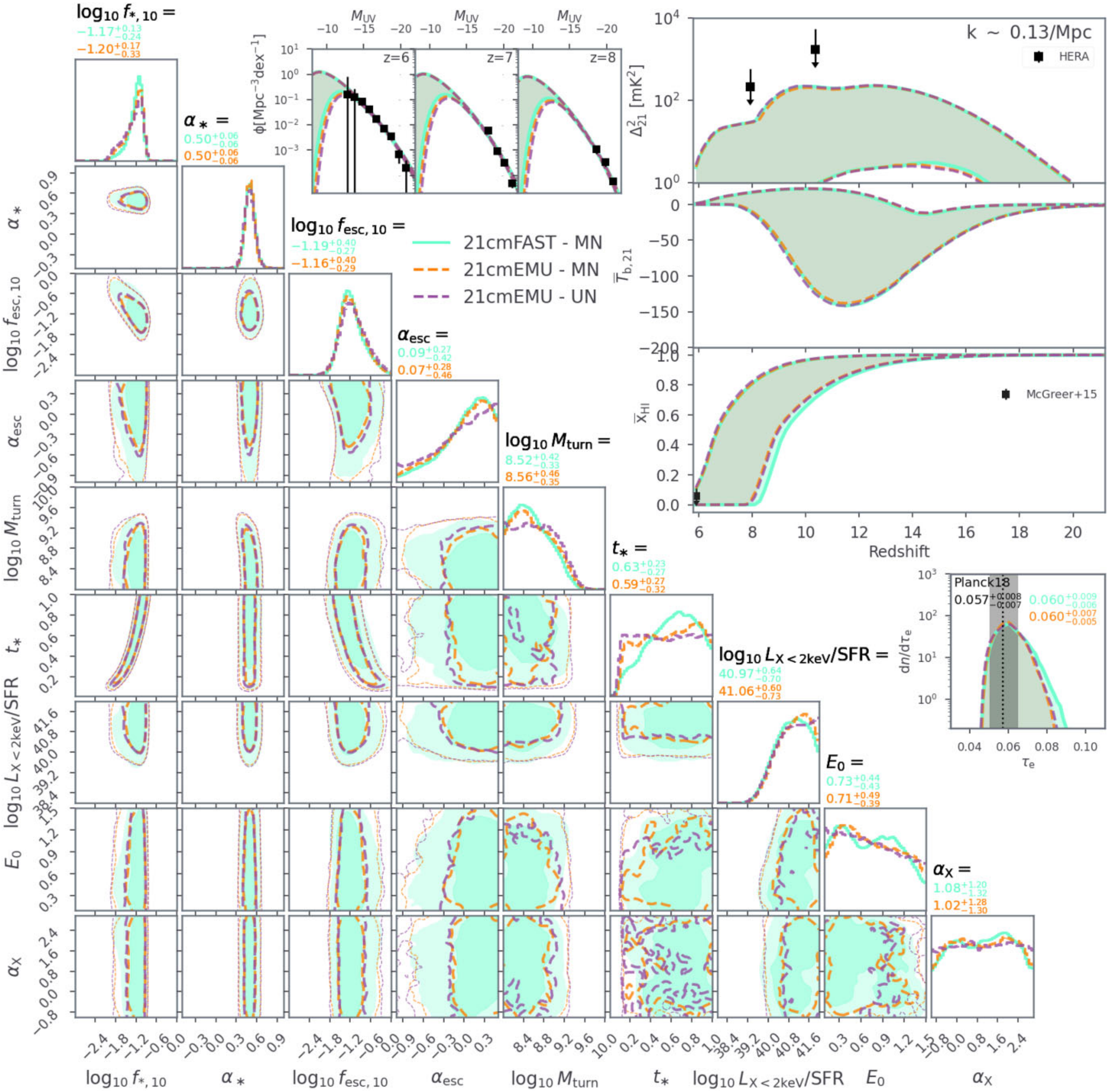

Figure 6. Comparison of posteriors obtained using 21CMFAST and 21CMEMU after performing an inference with the same HERA22 likelihood. The darker/thick dashed regions represent 68 per cent credible intervals (CIs), while pale/thin dashed regions represent 95 per cent CIs. The orange and purple posterior distributions are obtained using the MultiNest sampler (10k livepoints, ∼60k posterior samples), while the cyan posterior distribution is obtained using the UltraNest sampler (5k livepoints, ∼ 70k posterior samples). The median value and the 68 per cent CIs of the 1D marginal PDFs are written above each column of the corner plot. In the panels on the top right, all highlighted regions correspond to 95 per cent CIs. In the top middle panel, we plot the LFs for redshifts 6, 7, and 8. For the LF likelihood, we use the data shown in black squares (Bouwens et al. 2015, 2016; Oesch et al. 2018). In the top right, we show a panel with three summary statistics, namely the redshift evolution of the 21-cm power spectrum at $k = 0.13$ cMpc^{-1}, the 21-cm global signal and mean neutral fraction, from top to bottom. The black squares in the power spectrum plot correspond to the two deepest limits for each HERA redshift band ($k = 0.13$ cMpc^{-1} at z ∼8 and $k = 0.17$ cMpc^{-1} at z ∼10). In the bottom plot, the black square denotes the upper limit on the average neutral hydrogen fraction obtained from the QSO dark fraction (McGreer, Mesinger & D'Odorico 2015). In the bottom right, we show the PDFs of the Thomson optical depth together with the *Planck* result used in the likelihood. The astrophysical parameter ranges shown in the corner plot correspond to the extent of the flat priors assumed for the inferences.

included in the likelihood. Including only the UV LFs does disfavor very early reionization, $z > 11$, because the redshift evolution of the SFR density implied by UV LF observations is too steep to allow arbitrarily early EoR, even with escape fractions close to unity.

Finally we show the prior distribution in the space of UF LFs, 21-cm PS, 21-cm global signal, and EoR history in grey. We see that all of the posteriors in these spaces are significantly broader than the priors, and are thus likelihood dominated (i.e. are not sensitive to the

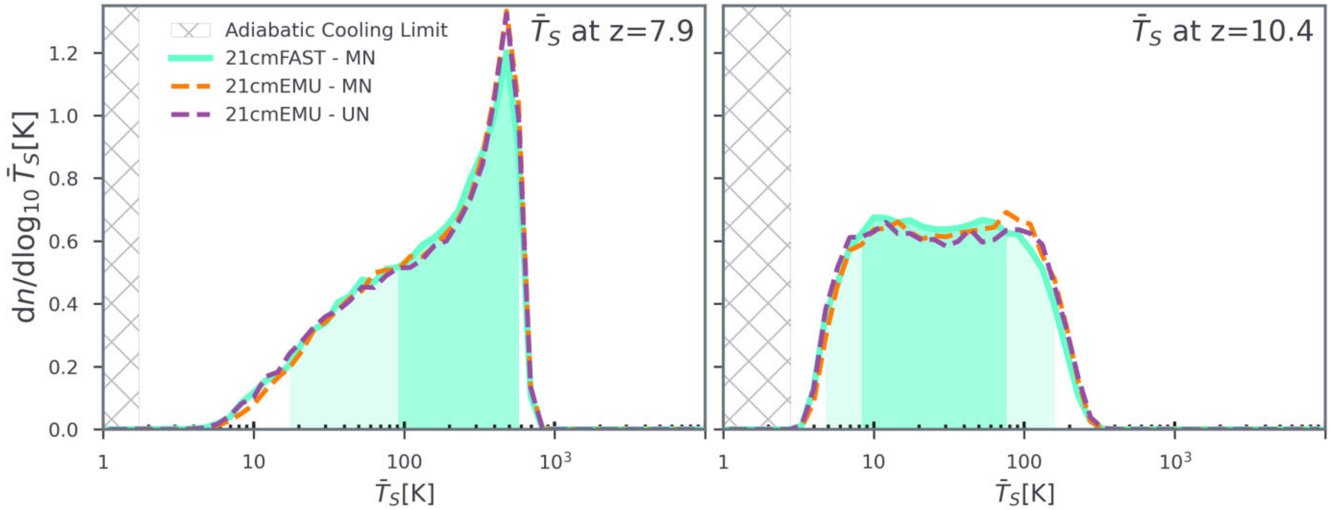

Figure 7. Comparison of the mean spin temperature distribution from 21CMFAST and 21CMEMU for each of the two HERA bands after performing an inference with the exact same likelihood. The CIs have been calculated using the highest posterior density method. The dark (light) cyan shaded region shows the 68 per cent (95 per cent) CI. The solid cyan line shows the distribution for 21CMFAST with 10k livepoints using `MultiNest`. The dashed orange line shows the same but for 21CMEMU. The dashed purple line shows the distribution for 21CMEMU but using the `UltraNest` sampler with 5k livepoints.

prior choices). Moreover, each likelihood term adds complimentary information, highlighting the importance of combining observational data sets when interpreting the high-redshift Universe.

4.3 Impact of late reionization

Recent observations of the large-scale opacity fluctuations in the Ly α forest (e.g. Becker et al. 2015; Bosman et al. 2018, 2022) imply a late end to reionization $z < 5.6$ (Choudhury, Paranjape & Bosman 2021; Qin et al. 2021). In this section, we explore how such new EoR history constraints would impact the previously shown posterior. Unfortunately, the current version of 21CMEMU does not emulate the Ly α forest, and so we cannot compute a likelihood directly in the observed space of Ly α opacity fluctuations. Instead we take a more approximate approach, computing the likelihood in the space of EoR histories, i.e. $\bar{x}_{\rm HI}(z)$. To construct a likelihood in this space, we use the EoR history posterior from Qin et al. (2021), who did in fact compute a likelihood from forward-modeled Ly α opacities in addition to the dark fraction and $\tau_{\rm e}$ observations. Specifically, we compute a new *Late EoR* posterior by replacing the dark fraction and $\tau_{\rm e}$ likelihood terms with a Gaussian likelihood evaluated at three redshifts, $\bar{x}_{\rm HI}(z = 6) = 0.25 \pm 0.07$, $\bar{x}_{\rm HI}(z = 7) = 0.58 \pm 0.1$, and $\bar{x}_{\rm HI}(z = 8) = 0.79 \pm 0.09$, ignoring any covariance between redshifts. Although this is obviously an approximation to computing the likelihood directly in the space of the observations, it suffices to qualitatively show the impact of new EoR history constraints.

In Fig. 9, we show the previous (*Fiducial*) posterior in purple (71k samples) together with the new (*Late EoR*) posterior in orange (18k samples). Understandably, the corresponding recovered EoR history in orange is narrowly centered around the three points at $z = 6$, 7, 8 used to define the likelihood. As a consequence, the posterior of the Thomson optical depth also becomes more narrow, shifting toward lower values while still being within the range allowed by Planck observations. The resulting PDF of $f_{\rm esc,\,10}$ for *Late EoR* is also narrower, and shifted towards smaller values. Even the power-law scaling of the escape fraction with halo mass, $\alpha_{\rm esc}$, is constrained to within ± 0.3 (68 per cent C.I.) for *Late EoR*, whereas the *Fiducial* posterior only sets a lower limit for this parameter. The remaining parameters are unaffected by the change to the Late EoR likelihood.

We also see that the recovered 21-cm large-scale PS for *Late EoR* is narrower at $z < 8$. The large-scale 21-cm PS during the EoR peaks around its midpoint (e.g. Lidz et al. 2007; Pritchard & Furlanetto 2007), which occurs at $z \sim$7–8. The HERA22 upper limits disfavor higher values of the 21-cm PS at $z \sim 8$, but the tail towards small PS values seen in the *Fiducial* posterior (corresponding to small $\bar{x}_{\rm HI}$), shrinks when moving to the *Late EoR* posterior.

4.4 Forecasts for HERA Phase II sixth-season observations

We now forecast parameter constraints that could be achievable with the sixth season of HERA observations, taken in 2022–2023 (Berkhout et al., in preparation). This season of observing used Phase II of the HERA instrument, spanning 50–230 MHz (omitting the FM band, 90–110 MHz), expanding coverage to CD and late reionization with respect to Phase I (which was used for HERA22). While analysis of this season's data is ongoing, its broad characteristics are known (Dillon & Murray 2021): approximately 1300 h of unflagged data over ~150 nights, with an average of ~148 un-flagged antennas per night. Although further flagging will certainly occur during processing, this data set will be HERA's most sensitive data release to date, by a significant factor.

We create a mock observation corresponding to this upcoming data set. For the 'true' cosmic signal, we use the Evolution of Structure (EOS) 2021 release (Muñoz et al. 2022). EOS2021 is a large simulation (1.5 cGpc per side with 1000^3 cells) made with 21CMFAST, with the goal of being our current 'best guess' for the true cosmic signal. Although it used the same parametrization for galaxy scaling relations as is used here (see Section 2.1), the physical model of EOS2021 has a few notable differences. Instead of leaving $M_{\rm turn}$ as a free parameter, EOS2021 explicitly calculated a local $M_{\rm turn}(\mathbf{x}, z)$ based on feedback from the local ionizing and photo-disassociating backgrounds, as well as the relative velocities of baryons and dark matter. Furthermore, EOS2021 explicitly accounted for putative PopIII star formation in the first, H_2-cooled galaxies (e.g. Tegmark et al. 1997; Abel, Bryan & Norman 2002; Bromm & Larson 2004;

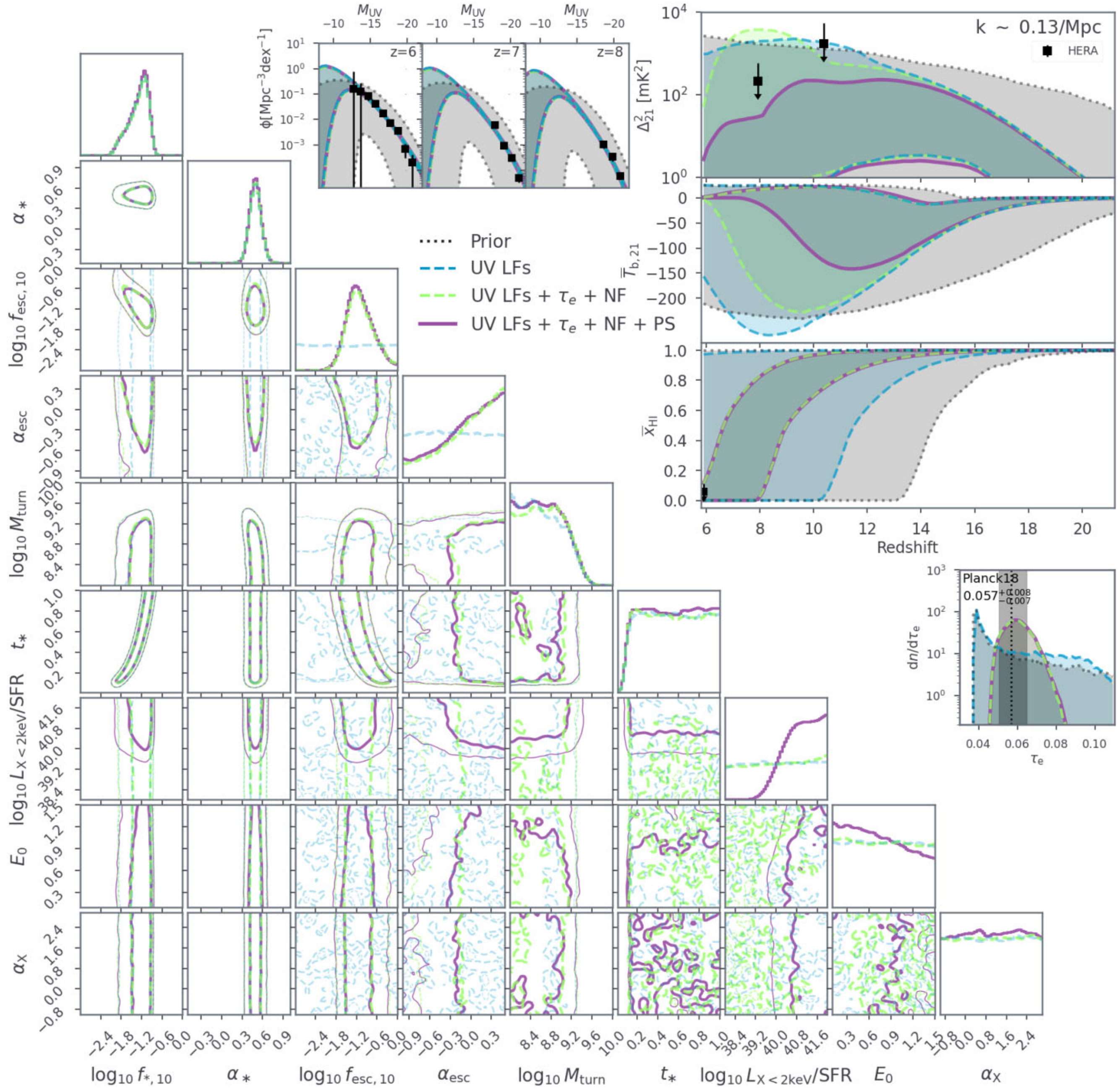

Figure 8. Contribution of various likelihood terms to the total posterior. The corner plot on the left shows the 95 per cent CI of three inferences, all run with 21CMEMU and UltraNest. The full posterior with all four probes is plotted in purple (exactly the same as the purple in Fig. 6). In green, we show the posterior without the HERA power spectrum upper limits term. In blue, we additionally remove the neutral fraction and Thomson optical depth terms, leaving only the UV LFs terms. On the top right half of the plot, we show the 95 per cent CI of the same three posteriors but in the space of summary statistics: first the UV LFs, and then a panel with the 21-cm power spectrum, 21-cm global signal, and EoR history, top to bottom, and finally a panel with the Thomson optical depth. In grey, we plot the summary statistic 95 per cent CI assuming a flat distribution across all nine astrophysical parameters which is what was used for the prior for the 21CMFAST inference.

Haiman & Bryan 2006), which dominated the background radiation fields at $z > 16$, for their fiducial parameter choices. As a result of the models being different, 21CMEMU could result in a biased recovery of the EOS2021 signal; we quantify this below.

We use 21CMSENSE[10] (Pober et al. 2013, 2014) to obtain thermal and sample variance estimates of the HERA sixth season data, and describe our methodology and assumptions in Appendix B. We consider our sensitivity estimate to be realistic, with a few important caveats, for example the potential over-estimation of sensitivity when treating ‘similar’ baselines as identical (Zhang, Liu & Parsons 2018). The largest unpredictable caveat is of course the presence of instrumental systematics, for which we describe our approach in more detail below.

Radio telescopes, including HERA, impose their own signature on observations – dependent on the primary beam attenuation, antenna

[10] https://github.com/rasg-affiliates/21cmSense

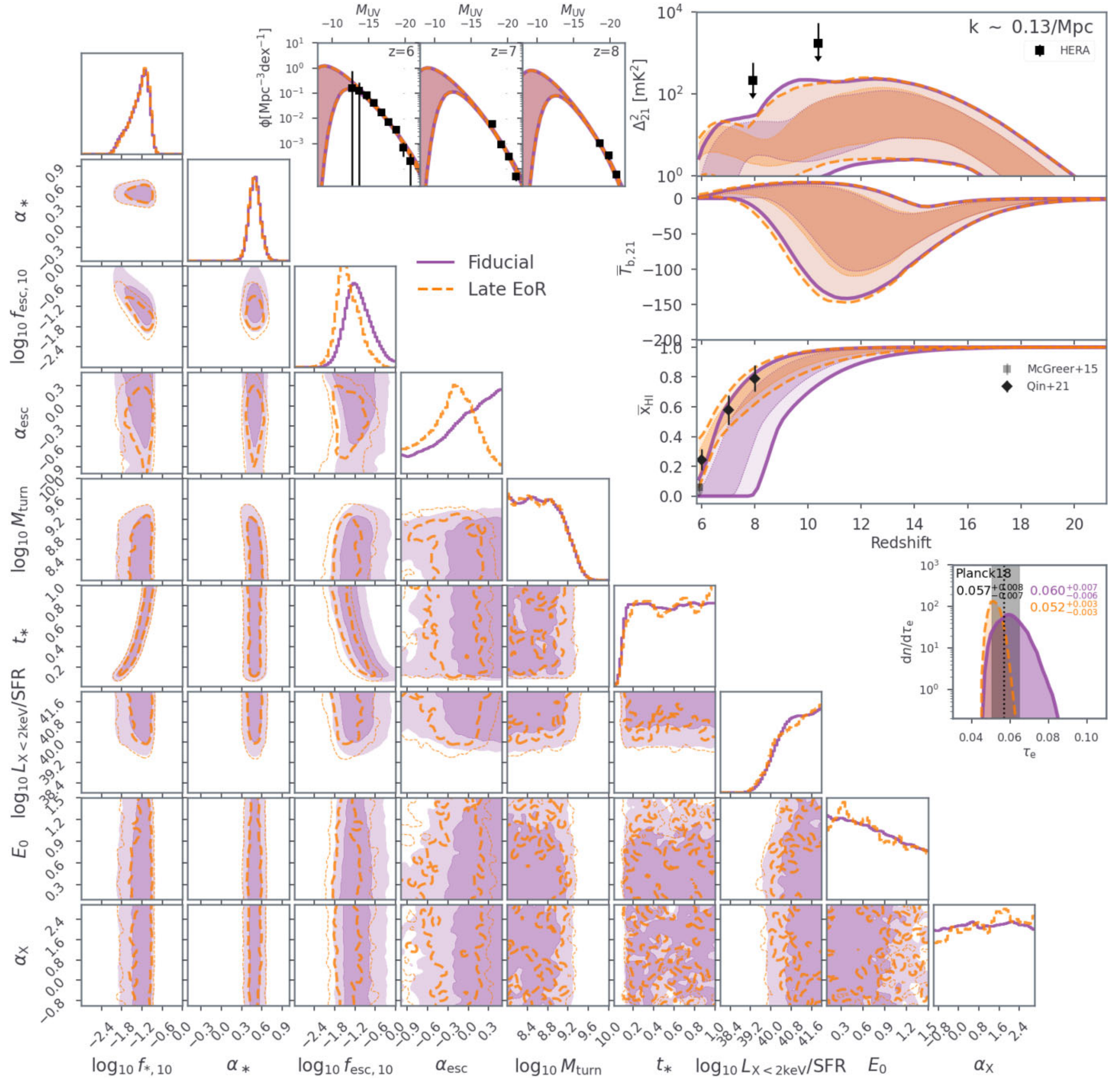

Figure 9. Same as Fig. 6, but comparing the fiducial posterior (*solid purple*) with one obtained by replacing the QSO dark fraction and τ_e likelihood terms with a 'Late EoR' likelihood denoted by the three points with error bars in the middle right panel (*dashed orange*). The 'Late EoR' likelihood is based on the inference results in Qin et al. 2021, which included recent measurements of opacity fluctuations in the Ly α forest. In the top right sub-panels, we show both the 68 per cent (darker) and 95 per cent (paler) C.I.

layout, channelization and other instrumental characteristics. The effect of this instrumental signature on the observed power spectrum is such that neighbouring Fourier modes are linearly'mixed' via a 'window function' matrix (e.g. Liu & Tegmark 2011; Gorce et al. 2023). We calculated this window function using the `hera_pspec`[11] package. We did not use the exact HERA beam as in Gorce et al. (2023). Instead, we used the Gaussian beam approximation which we deemed sufficient for this forecast (see fig. 7 in Gorce et al. 2023 for a comparison). Once we obtain the HERA window function, we matrix multiply it with the emulated model to properly compare with the forecast.

We perform inference using the EOS2021 cosmological signal with the sensitivity estimates from 21CMSENSE as the mock observation (see Fig. B2 in Appendix B). This inference takes about 30 GPU h to run to convergence with UltraNest. In Fig. 10 we show the resulting posterior (HERA sixth season in orange) together with the previous result (*Fiducial* in purple). In the top right panel we show the mock PS at k ∼ 0.16 Mpc^{-1} as orange points with associated error bars. We see that based purely on the available S/N, the HERA sixth season data have the potential to detect the cosmic PS during the EoR (6 < z

[11] https://github.com/HERA-Team/hera_pspec

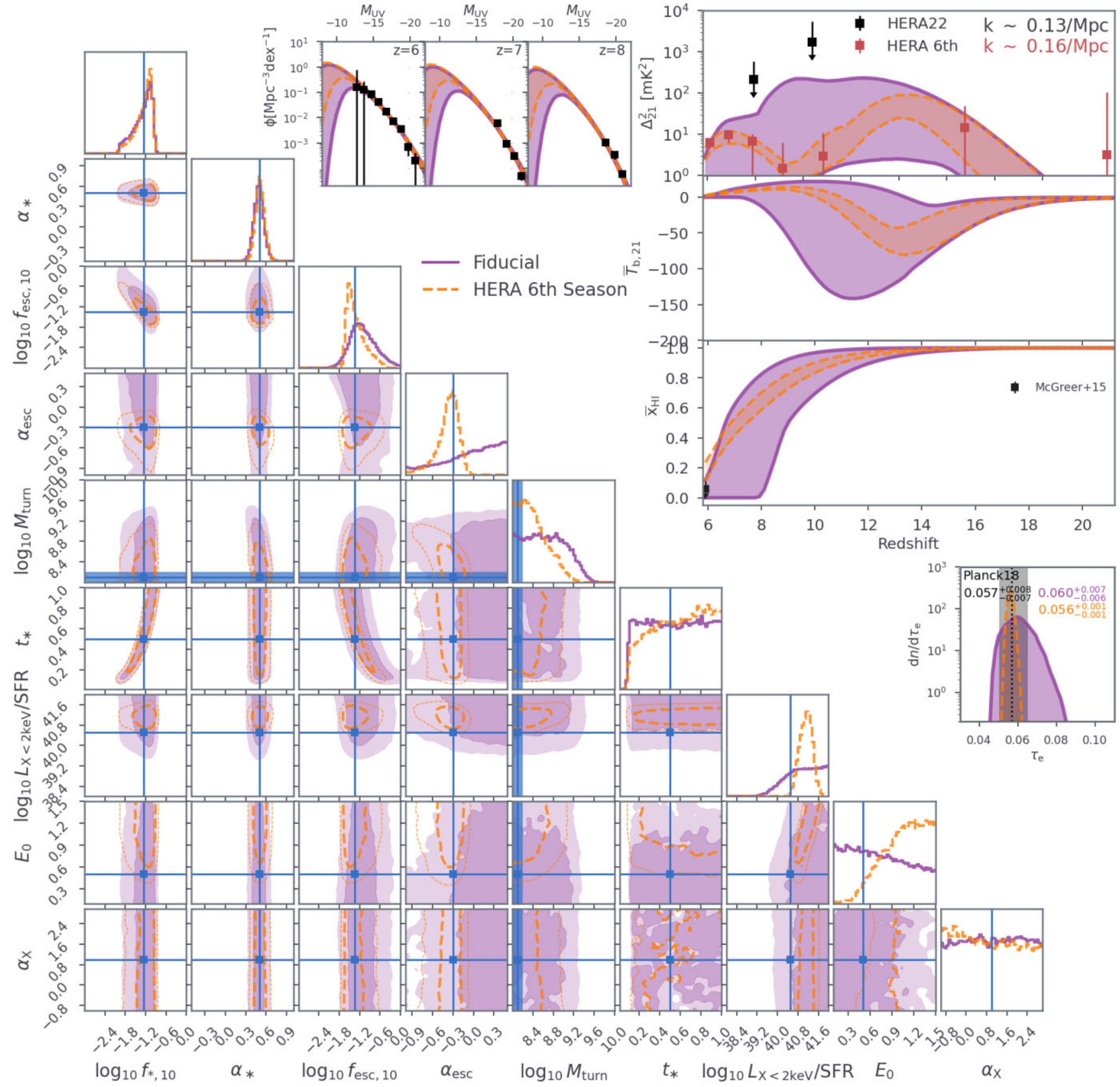

Figure 10. Forecasted constraints from mock HERA Phase II season six observations (see text for details) are shown orange (dashed). The mock PS amplitudes at $k \sim 0.16$ Mpc^{-1} are shown as red squares with error bars in the top right panel, together with current upper limits from HERA22 as black squares. The parameters of the cosmological simulation used for the mock observation, EOS2021, are denoted with blue lines and squares in the corner plot. We caution that the theoretical model used in EOS2021 and that used in 21CMEMU are somewhat different, as discussed in the text. As $M_{\rm turn}$ in EOS2021 evolves with redshift (see fig. 5 in Muñoz et al. 2022), here we demarcate its range during the EoR (i.e. $6 < z < 8$ where the mock observations imply a detection). For more details about the panels, see the legend in Fig. 6.

< 8). The 95 per cent CI of the inferred PS (orange) go tightly around the data points. This unbiased recovery is reassuring, given the above-mentioned differences in the theoretical models used for the mock and forward-modeled data. Indeed, most of the ‘true’ astrophysical parameters from EOS2021 (denoted with blue lines in the corner plot) are consistent with the recovered orange posteriors. Parameters governing X-ray heating, $L_{\rm X<2keV}$/SFR and E_0, are recovered with the lowest accuracy, with the true values residing outside of the 68 per cent CI of their 2D PDF. This is understandable, because the 21CMEMU forward models do not include the additional radiation from H$_2$-cooling galaxies, which dominate the X-ray heating at z > 16.

Comparing to current constraints (*Fiducial* posterior in purple), we see that that HERA sixth season data have the potential to drastically improve our knowledge of the EoR. The *HERA sixth season* EoR history $\overline{x}_{\rm HI}(z)$ is constrained to within ±0.06 (95 per cent C.I.): a factor of $\gtrsim 7$ improvement over current limits. As a result, we can place strong constraints on the characteristic ionizing escape fraction,

$f_{\rm esc,\,10}$, and its dependence on galaxy mass, $\alpha_{\rm esc}$, which are almost completely unknown currently.

It is important to note that these two posteriors use a different form for the likelihood. For the *HERA sixth season* forecast, we assume that there are *no residual systematics* after processing of the HERA data. This is in contrast to the previous likelihoods, which assume that each k-mode has a positive systematic whose prior amplitude is uniform and unbounded (cf. The HERA Collaboration 2022b). In practice, assuming no residual systematics results in a two-sided Gaussian likelihood, corresponding to a 'detection', whereas the traditional likelihood has been a one-sided error-function corresponding to an 'upper-limit'. We make this choice as it is not straightforward to sample from the unbounded uniform prior for systematics when creating the mock data for the forecast. The resulting tighter parameter posteriors for the new data are therefore the result of an admixture of the new more sensitive data *and* the (effectively) more constrained priors on systematics.

5 CONCLUSION

Here we introduced 21CMEMU: a publicly available emulator of several summary observables from 21CMFAST. We trained the emulator on 1.3M pseudo-posterior samples from the inference in HERA22. The input consists of a nine parameter model characterizing the UV and X-ray outputs of high redshift galaxies. The output consists of: (i) the 21-cm power spectrum as a function of redshift and wavemode; (ii) the IGM mean neutral fraction as a function of redshift; (iii) the UV LF at four redshifts 6, 7, 8, and 10; (iv) the Thompson scattering optical depth to the CMB; (v) the mean spin temperature as a function of redshift; and (vi) the 21-cm global signal as a function of redshift. The emulator predicts all of these quantities with under ~ 2.4 per cent error at 68 per cent CL, and a computational cost that is lower by a factor of $\sim$10 000 compared to 21CMFAST.

We varied the size of the training set, finding only a modest decrease in performance (a factor of $\sim$2 decrease in the FE) as the number of samples was reduced from 1.3M to $\sim$100k. Below $\sim$100k samples, we saw a sharp drop in performance, with the fractional error increasing roughly as the inverse of the size of the training set.

We validated the emulator's performance in inference by comparing the posteriors obtained with 21CMEMU versus 21CMFAST using the same likelihood (taken from HERA22). We found a very modest difference between these two posteriors, further illustrating that the emulator error is negligible when performing inference using current data.

Next, we profited from the speed of our trained emulator to perform multiple inferences that would otherwise be very costly using direct simulation. First, we studied the constraining power of each term in our fiducial likelihood. We found that current observations are very complementary, with UV LFs constraining the SHMRs, EoR history probes constraining the ionizing escape fraction, and the addition of 21-cm PS upper limits constraining the X-ray luminosity to SFR relation.

We also explored the impact of new EoR history constraints, driven by opacity fluctuations in the Ly α forest. These recent observations imply much tighter constraints on the EoR history, finishing at $z < 5.6$ (e.g. Choudhury, Paranjape & Bosman 2021; Qin et al. 2021). The inclusion of these new limits tightened the recovered constraints on the ionizing escape fraction and its scaling with halo mass. The impact on other parameters was modest.

Finally, we presented forecasts of parameter constraints achievable with ongoing sixth season phase II observations with the HERA telescope. Optimistically, we could expect a detection of the 21-cm PS at $z \sim$6–7. This would result in a dramatic improvement in the recovered timing of the EoR, allowing us to infer $\overline{x}_{\rm H\,I}(z)$ to within ± 0.06 (95 per cent C.I.): a factor of $\gtrsim 7$ improvement over current limits. As a result, we could place strong constraints on the characteristic ionizing escape fraction and its dependence on galaxy mass, which are almost completely unknown currently. We cautioned however that this forecast is optimistic, mainly because it assumed there are no residual systematics in the processed data (see Appendix B for more details).

21CMEMU was trained on a data base of summary observables where only one seed i.e. random set of initial conditions is available per combination of astrophysical parameters. In the future, we hope train the emulator on a data base that samples many different seeds in order to emulate a full likelihood function rather than only approximate the mean as we do right now. This is important since Prelogović & Mesinger (2023) showed that using a single random seed when forward modeling can bias the inference results.

We make 21CMEMU publicly available at https://github.com/21cmfast/21cmEMU, and include it as an alternative simulator in the public 21CMMC[12] sampler. We will periodically release updated versions, trained on the latest galaxy models and expanding the choice of summary outputs.

ACKNOWLEDGEMENTS

We thank A Liu and J Dillon for useful comments on a draft version of this paper. We also thank A Gorce for helpful discussion concerning the HERA window function calculation. We gratefully acknowledge computational resources of the Center for High Performance Computing (CHPC) at SNS. AM and RT acknowledge support from the Italian Ministry of Universities and Research (MUR) through the PRO3 project 'Data Science methods for Multi-Messenger Astrophysics and Cosmology', as well as partial support from the Fondazione ICSC, Spoke 3 'Astrophysics and Cosmos Observations', Piano Nazionale di Ripresa e Resilienza Project ID CN00000013 'Italian Research Center on High-Performance Computing, Big Data and Quantum Computing' funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di 'campioni nazionali di R&S (M4C2-19)'—Next Generation EU (NGEU). SM has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101067043. YQ is supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project # CE170100013. RT acknowledges co-funding from Next Generation EU, in the context of the National Recovery and Resilience Plan, Investment PE1 – Project FAIR 'Future Artificial Intelligence Research'. This resource was co-financed by the Next Generation EU [DM 1555 del 11.10.22].

DATA AVAILABILITY

The trained emulator is on a publicly accessible github repository, as well as available for install as a PYTHON package using `pip`.

REFERENCES

Abadi M. et al., 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/

[12] https://github.com/21cmfast/21CMMC

Abdurashidova Z. et al., 2022, ApJ, 924, 51

Abel T., Bryan G. L., Norman M. L., 2002, Science, 295, 93

Alsing J., Wandelt B., Feeney S., 2018, MNRAS, 477, 2874

Bañados E. et al., 2018, Nature, 553, 473

Becker G. D., Rauch M., Sargent W. L. W., 2007, ApJ, 662, 72

Becker G. D., Bolton J. S., Madau P., Pettini M., Ryan-Weber E. V., Venemans B. P., 2015, MNRAS, 447, 3402

Behroozi P. S., Silk J., 2015, ApJ, 799, 32

Bevins H. T. J., Handley W. J., Fialkov A., de Lera Acedo E., Javid K., 2021, MNRAS, 508, 2923

Bolton J. S., Haehnelt M. G., Warren S. J., Hewett P. C., Mortlock D. J., Venemans B. P., McMahon R. G., Simpson C., 2011, MNRAS, 416, L70

Bosman S. E. I., Fan X., Jiang L., Reed S., Matsuoka Y., Becker G., Haehnelt M., 2018, MNRAS, 479, 1055

Bosman S. E. I. et al., 2022, MNRAS, 514, 55

Bouwens R. J. et al., 2015, ApJ, 803, 34

Bouwens R. J. et al., 2016, ApJ, 830, 67

Bromm V., Larson R. B., 2004, ARA&A, 42, 79

Buchner J., 2016, Stat. Comput., 26, 383

Buchner J., 2019, PASP, 131, 108005

Buchner J., 2021, JOSS, 6, 3001

Bye C. H., Portillo S. K. N., Fialkov A., 2022, ApJ, 930, 79

Choudhury T. R., Ferrara A., 2005, MNRAS, 361, 577

Choudhury T. R., Paranjape A., Bosman S. E. I., 2021, MNRAS, 501, 5782

Clément B. et al., 2012, A&A, 538, A66

Cole A., Miller B. K., Witte S. J., Cai M. X., Grootes M. W., Nattino F., Weniger C., 2022, J. Cosmol. Astropart. Phys., 2022, 004

D'Odorico V. et al., 2023, MNRAS, 523, 1399

Das S. et al., 2014, J. Cosmol. Astropart. Phys., 2014, 014

Das A., Mesinger A., Pallottini A., Ferrara A., Wise J. H., 2017, MNRAS, 469, 1166

Datta K. K., Mellema G., Mao Y., Iliev I. T., Shapiro P. R., Ahn K., 2012, MNRAS, 424, 1877

Dayal P., Ferrara A., Dunlop J. S., Pacucci F., 2014, MNRAS, 445, 2545

de Belsunce R., Gratton S., Coulton W., Efstathiou G., 2021, MNRAS, 507, 1072

de Oliveira-Costa A., Tegmark M., Gaensler B. M., Jonas J., Landecker T. L., Reich P., 2008, MNRAS, 388, 247

Dillon J. S., Murray S., 2021, Technical Report, HERA Memo #122: H6C (2023 season) Summary of Season Flags

Drake A. B. et al., 2017, MNRAS, 471, 267

Fagnoni N. et al., 2021, MNRAS, 500, 1232

Fan X. et al., 2006, AJ, 132, 117

Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Furlanetto S. R., 2006, MNRAS, 371, 867

George E. M. et al., 2015, ApJ, 799, 177

Ghara R., Choudhury T. R., Datta K. K., 2015, MNRAS, 447, 1806

Ghara R. et al., 2020, MNRAS, 493, 4728

Gorce A. et al., 2023, MNRAS, 520, 375

Greig B., Mesinger A., 2015, MNRAS, 449, 4246

Greig B., Mesinger A., 2017, MNRAS, 472, 2651

Greig B., Mesinger A., 2018, MNRAS, 477, 3217

Greig B., Mesinger A., Bañados E., 2019, MNRAS, 484, 5094

Greig B., Mesinger A., Davies F. B., Wang F., Yang J., Hennawi J. F., 2022, MNRAS, 512, 5390

Haiman Z., Bryan G. L., 2006, ApJ, 650, 7

Heinrich C., Hu W., 2021, Phys. Rev. D, 104, 063505

Hoag A. et al., 2019, ApJ, 878, 12

Hui L., Gnedin N. Y., 1997, MNRAS, 292, 27

Jennings W. D., Watkinson C. A., Abdalla F. B., McEwen J. D., 2019, MNRAS, 483, 2907

Kaur H. D., Gillet N., Mesinger A., 2020, MNRAS, 495, 2354

Kern N. S., Liu A., Parsons A. R., Mesinger A., Greig B., 2017, ApJ, 848, 23

Kimm T., Katz H., Haehnelt M., Rosdahl J., Devriendt J., Slyz A., 2017, MNRAS, 466, 4826

Konno A. et al., 2014, ApJ, 797, 16

Koopmans L. et al., 2015, in Advancing Astrophysics with the Square Kilometre Array (AAemu4). p. 1, preprint (arXiv:1505.07568)

Kuhlen M., Faucher-Giguère C.-A., 2012, MNRAS, 423, 862

Lazare H., Sarkar D., Kovetz E. D., 2023, preprint (arXiv:2307.15577)

Lewis J. S. W. et al., 2020, MNRAS, 496, 4342

Lidz A., Zahn O., McQuinn M., Zaldarriaga M., Dutta S., Hernquist L., 2007, ApJ, 659, 865

Liu A., Shaw J. R., 2020, PASP, 132, 62001

Liu A., Tegmark M., 2011, Phys. Rev. D, 83, 103006

Liu A., Parsons A. R., Trott C. M., 2014a, Phys. Rev. D, 90, 23018

Liu A., Parsons A. R., Trott C. M., 2014b, Phys. Rev. D, 90, 23019

Ma X., Quataert E., Wetzel A., Hopkins P. F., Faucher-Giguère C.-A., Kereš D., 2020, MNRAS, 498, 2001

McGreer I. D., Mesinger A., D'Odorico V., 2015, MNRAS, 447, 499

Madau P., Dickinson M., 2014, ARA&A, 52, 415

Madau P., Meiksin A., Rees M. J., 1997, ApJ, 475, 429

Maity B., Choudhury T. R., 2022, MNRAS, 515, 617

Mason C. A., Treu T., Dijkstra M., Mesinger A., Trenti M., Pentericci L., de Barros S., Vanzella E., 2018, ApJ, 856, 2

Mellema G. et al., 2013, Exp. Astron., 36, 235

Mertens F. G. et al., 2020, MNRAS, 493, 1662

Mesinger A., 2019, The Cosmic 21-cm Revolution; Charting the First Billion Years of Our Universe

Mesinger A., Furlanetto S., 2007, ApJ, 669, 663

Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955

Mesinger A., Aykutalp A., Vanzella E., Pentericci L., Ferrara A., Dijkstra M., 2015, MNRAS, 446, 566

Mondal R. et al., 2020, MNRAS, 498, 4178

Mondal R., Mellema G., Murray S. G., Greig B., 2022, MNRAS, 514, L31

Mortlock D. J. et al., 2011, Nature, 474, 616

Muñoz J. B., Qin Y., Mesinger A., Murray S. G., Greig B., Mason C., 2022, MNRAS, 511, 3657

Murray S., Greig B., Mesinger A., Muñoz J., Qin Y., Park J., Watkinson C., 2020, J. Open Source Softw., 5, 2582

Murray S. G., Bowman J. D., Sims P. H., Mahesh N., Rogers A. E. E., Monsalve R. A., Samson T., Vydula A. K., 2022, MNRAS, 517, 2264

Mutch S. J., Geil P. M., Poole G. B., Angel P. W., Duffy A. R., Mesinger A., Wyithe J. S. B., 2016, MNRAS, 462, 250

Nikolić I., Mesinger A., Qin Y., Gorce A., 2023, MNRAS, 526, 3170

Ocvirk P. et al., 2020, MNRAS, 496, 4087

Oesch P. A., Bouwens R. J., Illingworth G. D., Labbé I., Stefanon M., 2018, ApJ, 855, 105

Okamoto T., Gao L., Theuns T., 2008, MNRAS, 390, 920

Oke J. B., Gunn J. E., 1983, ApJ, 266, 713

Ouchi M. et al., 2010, ApJ, 723, 869

Paardekooper J.-P., Khochfar S., Dalla Vecchia C., 2015, MNRAS, 451, 2544

Papamakarios G., Sterratt D. C., Murray I., 2018, preprint (arXiv:1805.07226)

Park J., Mesinger A., Greig B., Gillet N., 2019, MNRAS, 484, 933

Planck Collaboration, 2020, A&A, 641, A6

Pober J. C. et al., 2013, AJ, 145, 65

Pober J. C. et al., 2014, ApJ, 782, 66

Prelogović D., Mesinger A., 2023, MNRAS, 524, 4239

Prelogović D., Mesinger A., Murray S., Fiameni G., Gillet N., 2022, MNRAS, 509, 3852

Pritchard J. R., Furlanetto S. R., 2007, MNRAS, 376, 1680

Pritchard J. R., Loeb A., 2012, Rep. Prog. Phys., 75, 86901

Qin Y., Poulin V., Mesinger A., Greig B., Murray S., Park J., 2020, MNRAS, 499, 550

Qin Y., Mesinger A., Bosman S. E. I., Viel M., 2021, MNRAS, 506, 2390

Reichardt C. L. et al., 2021, ApJ, 908, 199

Santos M. G., Ferramacho L., Silva M. B., Amblard A., Cooray A., 2010, MNRAS, 406, 2421

Saxena A., Cole A., Gazagnes S., Meerburg P. D., Weniger C., Witte S. J., 2023, MNRAS, 525, 6097

Schmit C. J., Pritchard J. R., 2017, MNRAS, 475, 1213

Schneider A., Giri S. K., Amodeo S., Refregier A., 2022, MNRAS, 514, 3802

Scoccimarro R., 1998, MNRAS, 299, 1097

Shibuya T., Ouchi M., Harikane Y., Nakajima K., 2019, ApJ, 871, 164
Shimabukuro H., Semelin B., 2017, MNRAS, 468, 3869
Singh S. et al., 2022, Nat. Astron., 6, 607
Sobacchi E., Mesinger A., 2013, MNRAS, 432, L51
Springel V., Hernquist L., 2003, MNRAS, 339, 312
Sun G., Furlanetto S. R., 2016, MNRAS, 460, 417
Tegmark M., Silk J., Rees M. J., Blanchard A., Abel T., Palla F., 1997, ApJ, 474, 1
The HERA Collaboration, 2022a, Improved Constraints on the 21 cm EoR Power Spectrum and the X-Ray Heating of the IGM with HERA Phase I Observations
The HERA Collaboration, 2022b, ApJ, 925, 221
Thomas R. M. et al., 2009, MNRAS, 393, 32
Trac H., Chen N., Holst I., Alvarez M. A., Cen R., 2022, ApJ, 927, 186
Trott C. M., 2016, MNRAS, 461, 126
Trott C. M. et al., 2020, MNRAS, 493, 4711
Visbal E., Barkana R., Fialkov A., Tseliakhovich D., Hirata C. M., 2012, Nature, 487, 70
Wang F. et al., 2020, ApJ , 896, 23
Xu H., Wise J. H., Norman M. L., Ahn K., O'Shea B. W., 2016, ApJ, 833, 84
Yang J. et al., 2020, ApJ, 897, L14
Yoshiura S., Minoda T., Takahashi T., 2023, preprint (arXiv:2305.11441)
Yue B., Ferrara A., Xu Y., 2016, MNRAS, 463, 1968
Zahn O., Mesinger A., McQuinn M., Trac H., Cen R., Hernquist L. E., 2011, MNRAS, 414, 727
Zhang Y. G., Liu A., Parsons A. R., 2018, ApJ, 852, 110
Zhao X., Mao Y., Cheng C., Wandelt B. D., 2022a, ApJ, 926, 151
Zhao X., Mao Y., Wandelt B. D., 2022b, ApJ, 933, 236

APPENDIX A: PARAMETER SPACE DEPENDENCE OF THE 21-CM PS EMULATION ERROR

In Appendix A, we look at how the emulation error is distributed over the 9D input parameter space. In Fig. A1, we show the 21-cm power spectrum test set fractional error as a 2D histogram as a function of each pair of input astrophysical parameters. On the diagonal, we show the histogram (probability density) of each astrophysical parameter in the test set.

As expected, the emulation error peaks at the edges of parameter space where the density of samples is the lowest [see also fig. 9 in Kern et al. (2017) and top plot in fig. 18 in Abdurashidova et al. (2022)]. However, the inclusion of the rejected livepoints in the

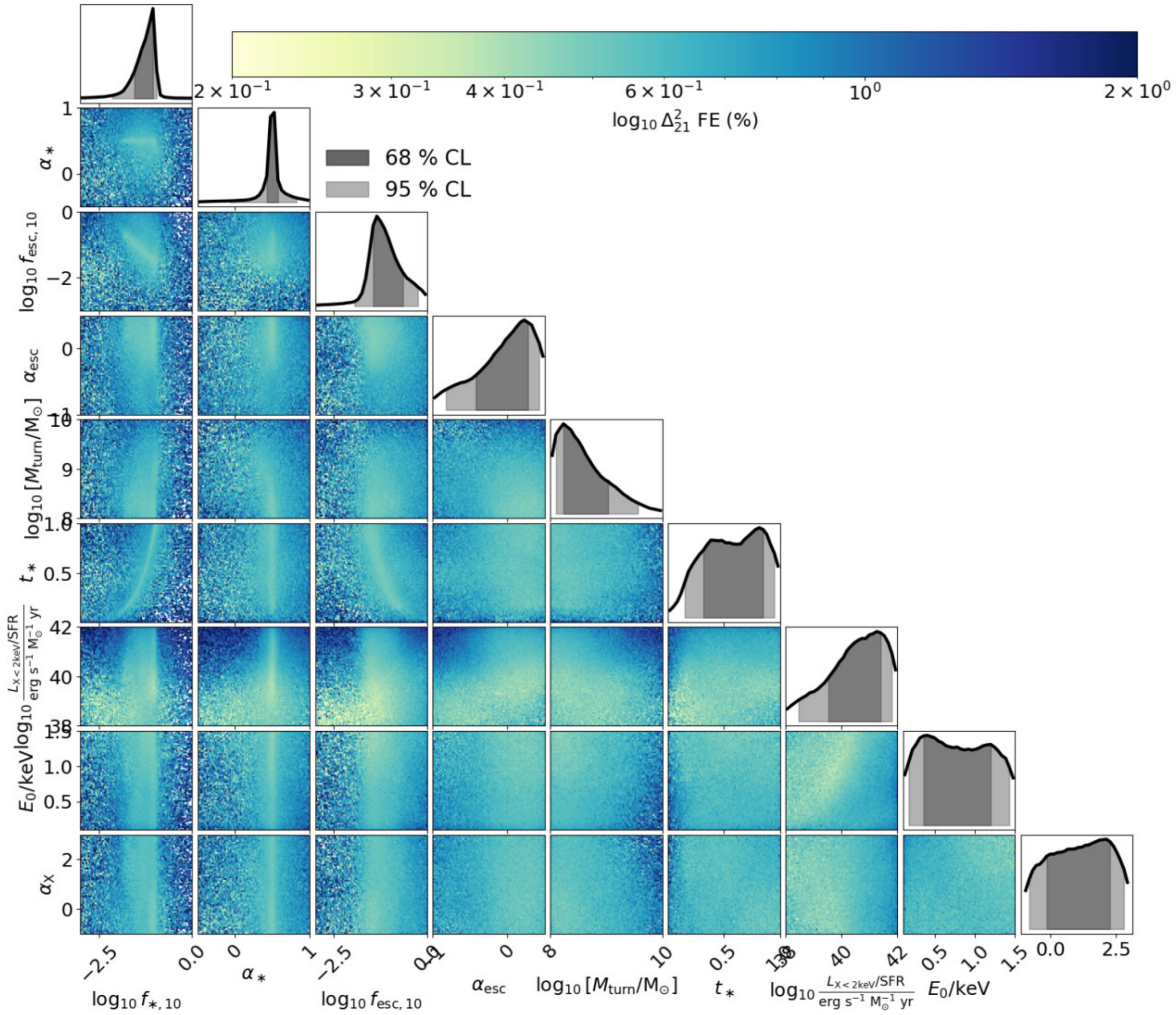

Figure A1. Distribution of the mean fractional error of the emulated log Δ^2_{21}. The colour of each bin in the 2D histogram is a weighted mean of the fractional error of the samples therein. On the diagonal, we show the 1D marginal density distribution of each astrophysical parameter in the test set. Note that the range of astrophysical parameters in the corner plot corresponds to the ranges taken for the flat prior of the inference used to generate the data base.

training allowed our emulator to generalize well beyond the peak of the posterior (c.f. Fig. 6). Importantly, the mean FE remains modest ($\lesssim 2$ per cent) throughout the prior volume.

APPENDIX B: 21CMSENSE SENSITIVITY ESTIMATES FOR HERA'S SIXTH SEASON

To obtain mock error estimates for the forecasted sixth season of HERA observations, we used the updated open-source 21CMSENSE[13] tool. The general algorithm of 21CMSENSE can be found in Pober et al. (2014) and in the extensive documentation and tutorials of the updated codebase[14] (see also Liu & Shaw (2020) for a review including a similar argument). A brief outline of the calculations is as follows: 21CMSENSE estimates thermal noise on any 3D $\vec{k}$-mode as

$$P_N(\vec{k}_\perp, k_{||}) \propto \frac{T_{\rm sys}^2}{N_{k_\perp}\Delta\nu\tau_{\rm int}}\xi(\vec{k}_\perp, k_{||}), \tag{B1}$$

where $T_{\rm sys}$ is frequency-dependent system temperature

$$T_{\rm sys} = T_{\rm sky}(\nu) + T_{\rm rcv}(\nu), \tag{B2}$$

$\Delta\nu = 122.07$ kHz is the channel width of the observation and $\tau_{\rm int} = 300$ s is the coherently averaged local sidereal time (LST)-bin size used in the analysis.[15] Furthermore, ξ is a 'flag' function that takes the value 0 or 1 depending on the location of the 3D mode with respect to the foreground wedge (see below).

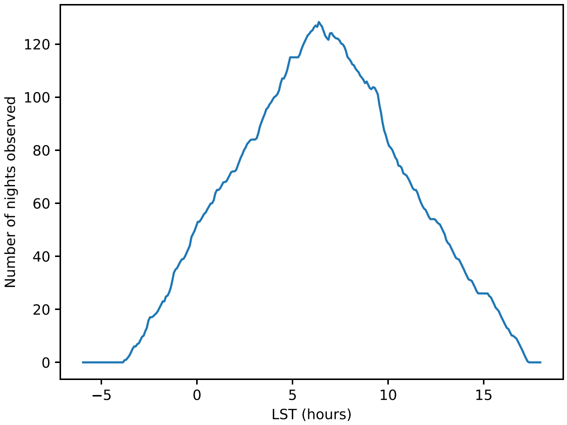

Figure B1. The number of times each 300-s LST bin was observed and unflagged in HERA's sixth season, used for sensitivity estimates. Note that this accounts only for flags arising from strong effects that affect large swathes of the observed antennas and/or channels (e.g. lightning storms, correlator outages), and further flags are applied in the downstream analysis.

In this equation, $N_{\rm k_\perp}$ represents the number of samples of this angular scale observed *coherently* throughout the observing season (i.e. observations that are averaged together as visibilities). In 21CMSENSE, this is estimated by creating a grid on the $\vec{k}_\perp$ plane, whose cells are the size of the primary beam of the instrument in Fourier-space (for HERA, this is 7λ at 150 MHz), and binning the the baseline coordinates into this grid.[16] In addition to the number of samples in a bin coming from different (redundant) baselines, we also have samples from the same baseline at different *times*. Here, samples at the same LST on different nights are averaged coherently, but samples at different LSTs are averaged *incoherently* (i.e. after forming power spectra). Currently, 21CMSENSE only has support for specifying the number of nights observed and the number of hours observed each night (thereby specifying the number of LST bins in conjunction with the LST bin duration). However, in realistic observational programmes, the same LST bins are not observed each night (whether due to the evolution of the sky throughout the season, or through flagging/data quality concerns). To partially account for this, we define a function $n_{\rm obs}({\rm LST})$ which counts the number of unflagged days observed over the season for any given 300-s-long LST bin (note that this accounts for flags of the entire observation, due to things like poor weather or correlator malfunctions, but not antenna- or channel-specific flags). To map this non-constant function of LST bin onto the schema available in 21CMSENSE, which assumes the same LST bins are observed each night, we set

$$n_{\rm days,\,eff} = \sqrt{\frac{\sum_{\rm LST}^{n_{\rm LST}} n_{\rm obs}^2({\rm LST})}{n_{\rm LST}}} \tag{B3}$$

and $t_{\rm day} = n_{\rm LST} \times 300$ s. This achieves the same resulting thermal noise level, under the assumption that the sky temperature is constant over the LST bins. We use actual sixth-season HERA measurements for $n_{\rm obs}$, as shown in Fig. B1. We calculate $n_{\rm days,\,eff} = 67.4$ for coherent averaging and $t_{\rm day} = 21$ h for incoherent averaging (i.e. the thermal noise from our observing pattern is equivalent to observing 253 300-s LST bins each for 67.4 d). Finally, we apply a further factor of $\epsilon = 0.9$ to $n_{\rm days,\,eff}$ to broadly account for finer-scale flags applied during analysis that are unaccounted in the LST-bin observing pattern of Fig. B1. In summary, we have

$$N_{k_\perp} = \epsilon N_{{\rm bl},k_\perp} N_{\rm days,eff}\sqrt{n_{\rm LST}}. \tag{B4}$$

The line-of-sight modes observed depend on the channel width, as already defined, and also the bandwidth of the observation. While HERA Phase II observes 200 MHz of bandwidth from 50–250 MHz, power spectra are estimated in smaller 'spectral windows' whose size is determined by a number of factors. Chiefly, the windows are as wide as possible, so as to include the largest scales where the signal is strongest, but are constrained by lightcone evolution (Datta et al. 2012; Trott 2016; Greig & Mesinger 2018) to be effectively smaller than ~ 10 MHz. In practice, spectral windows are chosen to lie between strongly flagged channels (e.g. FM band and Orbcomm), which means their width is not constant. Here, we use constant 20 MHz spectral windows, where we assume a Blackman tapering function is applied to each window to reduce the effective bandwidth to ~ 10 MHz (and an appropriate scaling factor of 1.737 is applied to the final noise level). We calculate noise estimates for all spectral windows between 50 and 250 MHz, excluding the FM band between 90 and 110 MHz.

We use a model for $T_{\rm sky}$ that is a power law in frequency, with amplitude and spectral-index obtained from simulated auto-correlations of the diffuse sky, using the GSM (de Oliveira-Costa et al. 2008) and the HERA Phase II primary beam (Fagnoni et al.

[13] https://github.com/rasg-affiliates/21cmSense.

[14] e.g. https://21cmsense.readthedocs.io/en/latest/tutorials/understanding_21cmsense.html

[15] In general, 21CMSENSE uses the more fundamental snapshot integration time of the instrument, and re-phases observations over a longer 'coherent observation duration', however HERA is a drift-scan telescope that performs no re-phasing, and all observations within an LST bin are considered coherent without re-phasing.

[16] This is probably the greatest departure from the actual HERA analysis, which coherently averages only *redundant* baselines, i.e. those that are equivalent to within several centimeters.

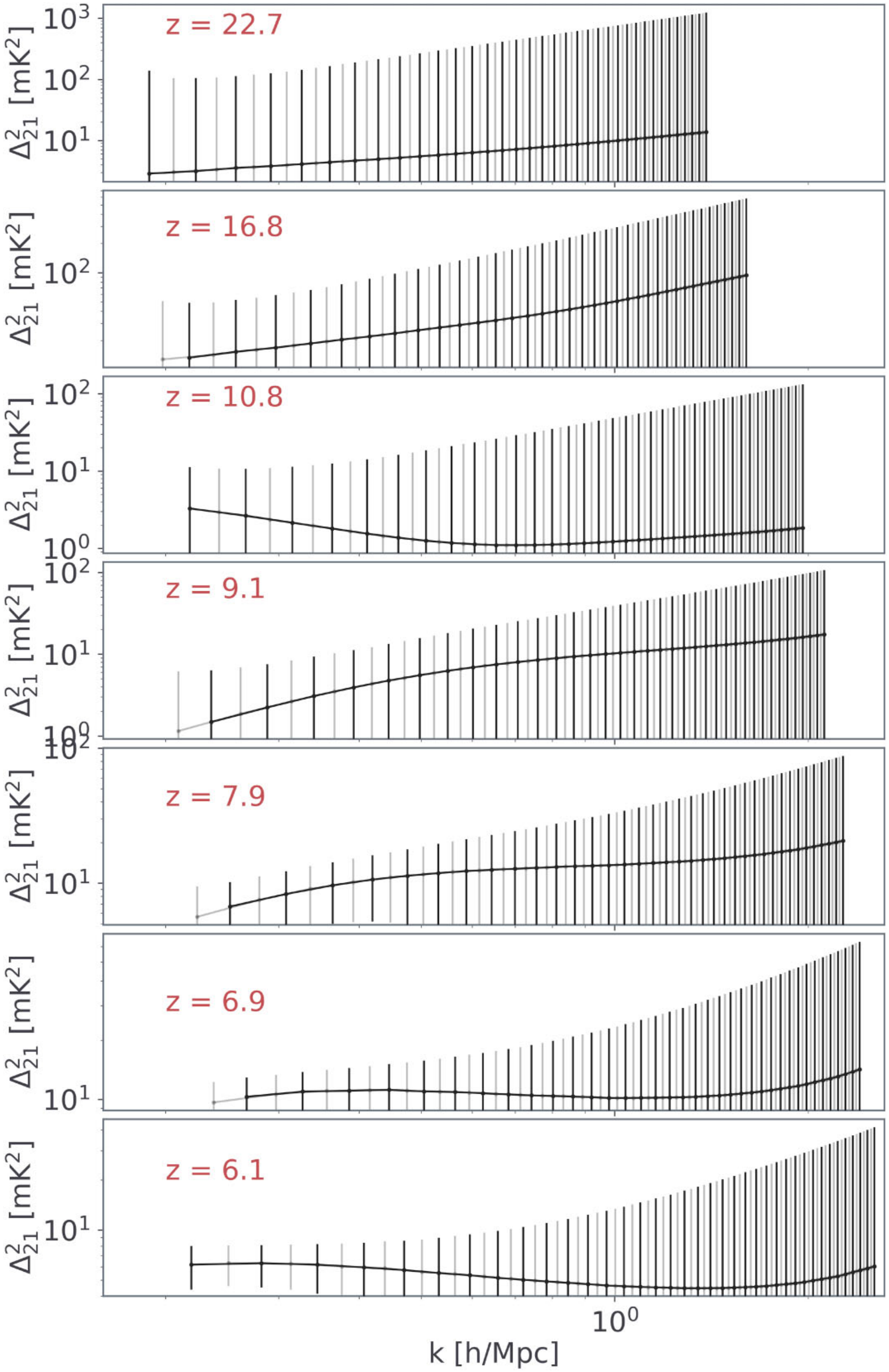

Figure B2. HERA phase II sixth season sensitivity forecast obtained using 21CMSENSE with the parameters specified in Table B1. Note that in practice, HERA decimates the k-bins to avoid requiring non-diagonal covariance (e.g. Abdurashidova et al. 2022). Here we have approximated this practice by using only half of the above k-bins (those highlighted in black) when computing the likelihood for our inference.

Table B1. Parameters used within 21CMSENSE to obtain sensitivity estimates for the sixth season of HERA observations.

Parameter	Description	Value
$N_{\rm ants}$	Number of antennas within the 209 available actually observed.	75, 100, 120, 148, 209
$T_{\rm sky}$	Sky temperature model	$150{\rm K}(\nu/150\ {\rm MHz})^{-2.5}$
$T_{\rm rcv}$	Receiver Temperature	Empirical, 600 K at 50 MHz, 60 K above 200 MHz
$\Delta\nu$	Channel width	122.07 kHz
$\tau_{\rm int}$	Coherent integration time (LST bin width)	300 sec
Δ u	UV-grid size for coherent baseline averaging	7 λ
$N_{\rm days,\ eff}$	Effective number of days observed coherently	67.4†
$t_{\rm day}$	Effective observed hours per day	21 h (253 LST bins)
ϵ	Efficiency factor for frequency-dependent flags	0.9
B	Spectral window bandwidth	20 MHz
$B_{\rm eff}$	Effective spectral window bandwidth after Blackman taper	11.51 MHz
FG wedge level	Line-of-sight scale below which modes are filtered	0.15 h/Mpc + horizon
Theory model	Cosmological power spectrum from which to calculate cosmic variance	Muñoz et al. 2022

See Appendix B for details on the algorithm. †Note that $N_{\rm days,\ eff}$ and $t_{\rm int}$ are effectively equivalent to the actual LST footprint of the season in terms of thermal noise, under the assumption that the sky temperature is constant with LST.

2021) at LST = 7 h,

$$T_{\rm sky} = 150\,{\rm K} \times \left(\frac{\nu}{150\,{\rm MHz}}\right)^{-2.5}. \tag{B5}$$

Currently, 21CMSENSE is not able to use different sky models for different LST-bins, so this choice represents the temperature for the most-observed LST bin. For $T_{\rm rcv}$, we use a frequency-dependent model based on electromagnetic simulations performed in Fagnoni et al. (2021), interpolated by a cubic spline. This model is close to a power law at low frequencies, with an amplitude of ∼600 K at 50 MHz and asymptoting to a const ∼ 60 K by 200 MHz.

We construct several estimates of the noise based on different effective observing arrays. The sixth season of HERA data observed with a maximum of 209 antennas simultaneously in any given night (of the total 350 antennas available). The bulk of these antennas observed consistently throughout the season, though a fraction of them were swapped in and out. In our estimates here, we assume that the same antennas observe consistently throughout the season, which is a reasonable approximation. Nevertheless, in practice, even though 209 antennas are being correlated at any given moment, some fraction of them are flagged over all channels (e.g. due to swapped polarizations, non-redundancies from physical effects such as feed displacement, or X-engine failures that affect a subset of antennas, etc.). The average number of antennas actually observing per-night throughout the season is as-yet unknown, though initial estimates place it at ∼150 antennas (Dillon & Murray 2021). Here we use $N_{\rm ants} = 148$, where the antennas are drawn randomly from the set of 209 antennas that actually observed throughout the season. In all cases, we use only baselines whose East-West length is greater than 15 m (i.e. we exclude North–South baselines, as their systematics are more difficult to filter out), and also only include baselines shorter than 150 m, similar to analyses of previous HERA seasons.

After obtaining the 3D sensitivity grid, we incoherently average into 1D spherical $|k|$-shells with bins of width $\Delta k_{||}$. In this process, we flag out $(|k_\perp|, k_{||})$-bins within the foreground 'wedge' (Liu, Parsons & Trott 2014a, b), defined by

$$k_{||}^{\rm wedge} = 0.15h{\rm Mpc}^{-1} + \frac{{\rm d}k_{||}}{{\rm d}\eta}(\nu)\frac{|b|}{c}, \tag{B6}$$

with $|b|$ the baseline length (in meters) corresponding to a given $k_\perp$, and $dk_{||}/d\eta$ a redshift-dependent cosmological factor converting bandwidth into cosmic distance. This corresponds to the 'horizon' limit of foregrounds in delay-space, plus a conservative buffer of 0.1 h/Mpc (corresponding to the buffer used in first-season HERA analyses).

In addition to the thermal variance, cosmic- (or sample-) variance is added, proportional to a fiducial cosmological power spectrum, $P^2_{\rm theory}$ divided by the number of LST bins and $k_\perp$-modes in a spherical shell. We note that using the number of LST-bins is inspired by the idea that LST-bins should capture the entire duration of 'coherence', equal to roughly the beam-crossing time for an antenna. However, HERA is conservative in using shorter coherence times, resulting in many more LST-bins. This reduces the thermal sensitivity, but artificially reduces the cosmic variance estimated by 21CMSENSE. Nevertheless, since cosmic variance is generally a sub-dominant contribution to the total variance, this should not have a large effect on the results presented here. For the fiducial theoretical model, we here use the model from Muñoz et al. (2022).

We summarize the parameters used in Table B1 and show the full HERA phase II sixth season sensitivity forecast in Fig. B2.

There are a few caveats to these estimates. Most importantly, baselines found within 7λ UV-bins together are considered redundant, while in the HERA analysis only baselines within 10 cm are considered redundant. This will artificially increase thermal sensitivity estimates. Secondly, the sky temperature is considered constant over the LST bins. To minimize the effect of this limitation, we use a sky model that is based at the most-observed LST (7 h). Thirdly, cosmic variance is reduced as the square root of the number of LST bins, instead of the number of independent 'fields' observed. This artificially increases the sensitivity from cosmic variance, though this should not have a large effect since this is the sub-dominant contribution on most scales and redshifts. Finally, in this forecast we did not decimate the k-bins as was done in previous analyses. This results in some unaccounted covariance between k-bins that would tend to over-estimate the sensitivity. We do not expect this to significantly affect the qualitative conclusions derived from the forecast.

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