Predicting the Neutral Hydrogen Distribution During Reionisation Using a GPR Emulator on N-Body Simulations

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Gaurav Pundir

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

May, 2025

Supervisor: Aseem Paranjape and Tirthankar Roy Choudhury

Certificate

This is to certify that this dissertation entitled *Predicting the Neutral Hydrogen Distribution During Reionisation Using a GPR Emulator on N-Body Simulations* towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents work carried out by Gaurav Pundir at the Indian Institute of Science Education and Research under the supervision of Prof. Aseem Paranjape and Prof. Tirthankar Roy Choudhury, Professors at the Inter-University Centre for Astronomy and Astrophysics (IUCAA), Pune and National Centre for Radio Astrophysics (NCRA), Pune respectively, during the academic year 2024-2025.

Asparanjape T. Roy Choudhury

Prof. Aseem Paranjape Prof. Tirthankar Roy Choudhury

Committee:

Prof. Aseem Paranjape

Prof. Tirthankar Roy Choudhury

Prof. Susmita Adhikari

Declaration

I hereby declare that the matter embodied in the report entitled *Predicting the Neutral Hydrogen Distribution During Reionisation Using a GPR Emulator on N-Body Simulations* are the results of the work carried out by me at the Indian Institute of Science Education and Research, Pune, as well as the Inter-University Centre for Astronomy and Astrophysics, Pune and National Centre for Radio Astrophysics, Pune, under the supervision of Prof. Aseem Paranjape and Prof. Tirthankar Roy Choudhury, and the same has not been submitted elsewhere for any other degree.

Gaurav

Gaurav Pundir
IISER Roll Number: 20201153
15th March 2025

This thesis is dedicated to my family

Acknowledgements

I am deeply grateful to my supervisors Prof. Aseem Paranjape and Prof. Tirthankar Roy Choudhury for their invaluable guidance and support over the last two years. Their trust in my ability to overcome various research challenges, combined with the enthusiasm they brought to our many discussions, has taught me great lessons and I owe a lot of my confidence, perseverance and articulation skills to them. Their mentorship has provided me with a well-rounded exposure to scientific research, better preparing me for the next step in my academic journey. I would also like to express my gratitude to Prof. Susmita Adhikari for interesting discussions, insights, and support in helping me navigate the challenges of my thesis year. I am thankful to Saee Dhawalikar for the valuable support that made this work possible — from helping me with code to engaging in discussions on various topics in cosmology.

I would also like to thank all my friends at IISER Pune for keeping things fun throughout my final year and, more importantly, for just being there for me. More broadly, I want to thank the student body at IISER Pune for organising various events and club activities, which provided many opportunities to make constructive use of my free time and, in turn, helped improve my productivity and mental well-being. Finally, words cannot quantify how grateful I am to my parents and my elder sister for believing in my ambitions and ability to pursue a career outside the perceived norms of our society. They never doubted me and always gave me the freedom I needed to understand myself as a person. Thank you for being my friends more than my guardians, and I hope I will make you proud one day.

Abstract

Building fast and accurate ways to model the distribution of neutral hydrogen during the Epoch of Reionisation (EoR) is essential for interpreting upcoming 21 cm observations. A key component of semi-numerical models of reionisation is the collapse fraction field $f_{\rm coll}(\mathbf{x})$, which represents the fraction of mass within dark matter haloes at each location. Using high-dynamic range N-body simulations to obtain this is computationally prohibitive and semi-analytical approaches, while being fast, end up compromising on accuracy.

In this work, we bridge the gap by developing a machine learning model that can generate $f_{\rm coll}$ maps by sampling from the full distribution of $f_{\rm coll}$ conditioned on the dark matter density contrast δ. The conditional distribution functions and the input density field to the model are taken from low-dynamic range N-body simulations that are more efficient to run. We evaluate the performance of our ML model by comparing its predictions to a high-dynamic range N-body simulation. Using these $f_{\rm coll}$ maps, we compute the HI and HII maps through a semi-numerical code for reionisation. We are able to recover the large-scale HI density field power spectra ($k \lesssim 1\ h\,{\rm Mpc}^{-1}$) at the $\lesssim 10\%$ level, while the HII density field is reproduced with errors well below 10% across all scales. Compared to existing semi-analytical prescriptions, our approach offers significantly improved accuracy in generating the collapse fraction field, providing a robust and efficient alternative for modelling reionisation.

Contents

List of Figures

Chapter 1

Introduction

1.1 Epoch of Reionisation and the 21 cm line

The *Epoch of Reionisation* (EoR) marks an important period in the history of the universe when the first luminous objects ionised the neutral hydrogen (HI) present in the intergalactic medium (IGM). The exact timing of this period is unknown, with most likely estimates for its start being around redshift of 20 – 30 with an end around 6. Studying this era is crucial for understanding many astrophysical processes, including the emergence of the first stars and galaxies and the growth of cosmic structure (for recent reviews, see [3, 4]). The observational signatures of EoR are extremely faint because of the large distances involved and are buried under much stronger astrophysical foregrounds. Luckily, there are multiple such observables on which reionisation leaves an imprint. The average flux in the Lyα absorption spectra from distant quasars can be used to study the end of reionisation. Furthermore, spatial fluctuations in the average flux can also be used as an indicator of the fluctuations in the distribution of neutral hydrogen, or HI [5–7]. The temperature of the IGM is also affected by the photoheating associated with reionisation and can be constrained from the Lyα absorption spectra [8–10].

Alternatively, one can also study the patchiness of reionisation either using the decrease in the intensity of Lyα emitters (LAEs) at redshifts around 6 [11], or using the kinetic Sunyaev-Zeldovich (kSZ) introduced in the CMB due to its interaction with the free electrons in ionised regions during the EoR [12, 13]. However, one of the most promising probes of

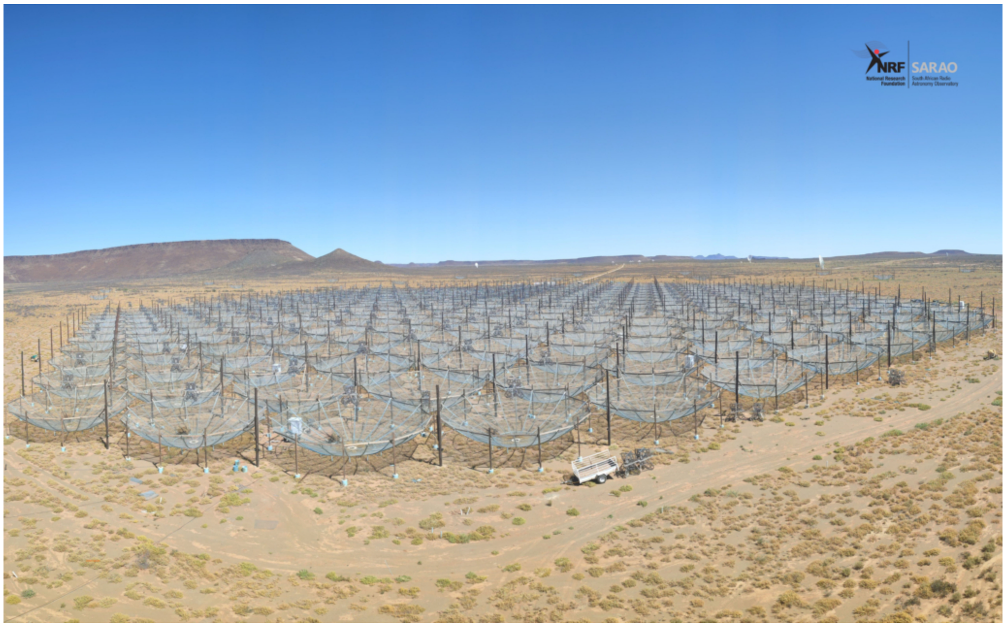


Figure 1.1: The Hydrogen Epoch of Reionisation (HERA) array situated at the Meerkat National Park in South Africa. *Image Credits*: South African Radio Astronomy Observatory - NRF/SARAO

the EoR is the *brightness temperature* fluctuations of the 21 cm line produced by neutral hydrogen [14–16]. The 21 cm signal refers to the transition between the two hyperfine levels of the hydrogen 1s state arising due to the interaction between the magnetic moments associated with the spin angular momentum of the electron and proton. Essentially, it is emitted when the electron and proton transition from a parallel to an anti-parallel spin state, with a frequency of around 1420 MHz. The brightness temperature T_b of a signal is defined as the effective temperature of a hypothetical blackbody with the same intensity at the same frequency as the signal. The sky-averaged brightness temperature excess over the CMB temperature is known as the 'global 21 cm signal' and can be measured using a single radiometer, which is the goal of experiments such as EDGES [17], SARAS [18, 19], and LEDA [20], among others. One can also aim to measure the *power spectrum* of the 21 cm brightness temperature fluctuations using radio interferometric arrays. Assuming an accurate subtraction of the contamination from astrophysical processes in the foreground, such a measurement would probe the density distribution of HI during the EoR. Therefore, given a model of cosmological structure formation describing the clustering of dark matter (to be traced by the baryonic matter such as HI and Helium) such as ΛCDM, along with a model describing the relevant astrophysical processes of reionisation, one can use these observations of the 21 cm signal fluctuations to constrain the Epoch of Reionisation. This goal has been pursued by various radio interferometers such as GMRT[1], MWA[2], PAPER[3], LOFAR[4] and will be continued by the upcoming HERA[5] Phase-II and SKA[6]. In a nutshell, we must be able to make efficient theoretical predictions for the 21 cm power spectrum in order to perform parameter inference from the upcoming observations and constrain models of the EoR.

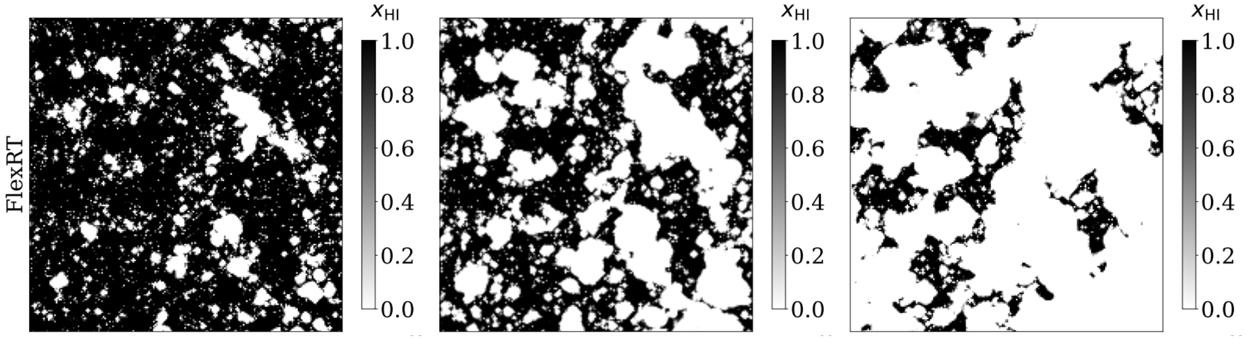


Figure 1.2: Evolution of the fraction of neutral hydrogen during reionisation as predicted by a cosmological radiative transfer simulation FlexRT [1]. The panels, from left to right, represent 20%, 50% and 80% volume ionised fractions with the white regions denoting ionised hydrogen.

1.2 Challenges in Modelling the EoR

In standard models of the EoR that assume galaxies to be the dominant contributors of ionising photons, reionisation proceeds via the formation of 'ionised bubbles' containing ionised hydrogen (HII). By modelling the distribution of these ionised bubbles, we can get the distribution of neutral hydrogen, which in turn provides information regarding fluctuations in the 21 cm signal. The most accurate way to achieve this is to run radiative transfer (RT) simulations that take into account the detailed physical interactions between matter and the photons emitted by the sources [21–30]. This includes various processes such as absorption, scattering, and photoionisation of the hydrogen atoms due to the high-energy photons. A set of slices showing the fraction of HI produced using an RT simulation are shown in Figure 1.2. However, these simulations must have a sufficiently large volume to achieve statistical

[1] https://www.gmrt.ncra.tifr.res.in/
[2] https://www.mwatelescope.org/
[3] http://eor.berkeley.edu/
[4] http://www.lofar.org/
[5] https://reionization.org/
[6] https://www.skao.int/en

convergence on the bubble distribution at large scales [31, 32]. Simultaneously, they need to resolve the smallest mass haloes capable of forming the first galaxies (typically down to $\sim 10^8 \ h^{-1} M_{\odot}$) due to their significant contribution to the ionising photon budget. This 'high-dynamic range' requirement adds significantly to the already high computational cost of modelling such a complex physical process, and makes these simulations highly inefficient in exploring the parameter space of EoR models.

One gets around this problem by resorting to the much faster but approximate semi-numerical models of reionisation. These aim to predict the 'ionisation field' – describing the fraction of hydrogen ionised at each location – by using the excursion-set approach [33] and a simple photon counting argument to define the barrier [34], thus bypassing the complicated radiative transfer physics [35–40]. These semi-numerical methods provide a reasonable match to RT simulations in terms of various statistics such as the neutral fraction, bubble size distribution, power spectrum of the ionisation field, and so on [41, 42]. When used along with semi-numerical galaxy formation codes, the input to these models can be the stellar mass [43] or the number of ionising photons entering the IGM at each cell [44], with the outpt being the ionisation fraction in the cell. When used in conjunction with dark-matter-only simulations, the required input is the 'collapse fraction field' denoted by $f_{\rm coll}(\mathbf{x})$, which is equal to the fraction of dark matter mass within haloes in the grid cell at $\mathbf{x}$. This can be computed by first filtering individual haloes using the excursion-set formalism from a dark matter density field evolved using Lagrangian perturbation theory [35, 45]. Alternatively, it can be prescribed semi-analytically without explicitly identifying sub-grid haloes from the conditional Press-Schechter (hereafter conditional PS) halo mass function [33, 46], conditioned on the dark matter density contrast $\delta(\mathbf{x})$ for each cell. One can also use the conditional Sheth-Tormen (hereafter conditional ST) mass function, which is based on the more general ellipsoidal collapse model [47, 48].

However, these analytical mass functions do not capture the full complexity of halo formation, are not universal and are only an approximate match to N-body simulation results [49–53]. In particular at the redshifts of our interest, the conditional PS mass function underestimates the abundance of high mass haloes ($M \gtrsim 10^{10} \ h^{-1} M_{\odot}$ at $z = 7$) and overpredicts the abundance at low masses ($M \lesssim 10^8 \ h^{-1} M_{\odot}$). While the ST mass function provides a better match, it still overpredicts the number of very massive haloes at high redshifts [49]. These considerations are important in studies of reionisation since correctly predicting the abundance of haloes is crucial for obtaining accurate ionised regions. Therefore, as the first step, one should transition away from the conditional PS and ST mass functions and use N-body simulations to calculate the conditional mass function empirically. However, these approaches only assign the mean $f_{\rm coll}$ conditioned on the density value of each cell $\langle f_{\rm coll}|\delta\rangle$, whereas in reality, the $f_{\rm coll}$ value can stochastically fluctuate across different cells with the same density value. Ignoring this 'scatter' or 'stochasticity' in the collapse fraction (which is primarily due to a dependence of $f_{\rm coll}$ on environmental variables other than the grid-scale δ) can lead to inaccurate recovery of the small-scale features in the HI and HII maps, as we show later in the paper. Hence, as the next step, one should use the conditional cumulative distribution function of $f_{\rm coll}$ conditioned on the density contrast, CDF$(f_{\rm coll}|\delta)$ to sample the $f_{\rm coll}$ field.

In either case, it is still important for the N-body simulations to have a high-dynamic range. This makes them computationally very expensive and thus one must explore alternatives to enable fast predictions of collapse fraction and subsequently the HI density fields. Attempts to resolve this issue have involved running low-resolution, large-volume simulations and using a high-resolution, small-volume simulation to populate the otherwise unresolved haloes. This has been implemented in [31, 54], although while not taking into account the scatter in the halo numbers for a given overdensity. Poisson fluctuations in the halo number count around the mean value predicted by the analytical conditional mass functions have been incorporated in certain studies [55–57], but this has the limitation of only being valid for large enough cell sizes [58, 59]. An alternative approach is to identify matching cells in the small-volume, high-resolution simulation and use haloes from these cells to populate the low-resolution box [60]. However, this method requires simultaneous access to both the large-volume and small-volume simulations during the construction of the effective high-dynamic-range box.

1.3 Goal and Outline of the Thesis

In this work, we aim to fully incorporate the effects of stochasticity in the collapse fraction values, by directly using the full CDF$(f_{\rm coll}|\delta)$ obtained from an N-body simulation for sampling the $f_{\rm coll}$ field. We still use a hybrid scheme of combining information from computationally inexpensive low-dynamic range boxes to mimic a high-dynamic range one, but do so using a machine learning algorithm based on Gaussian Process Regression (GPR). For the sake of comparison, we define this to be the *stochastic* case and also define the *deterministic* case in which $f_{\rm coll}$ predictions are made by simply assigning the conditional means $\langle f_{\rm coll}|\delta\rangle$. We use the $f_{\rm coll}$ fields from both the cases as inputs to a semi-numerical code for reionisation to obtain the HI and HII maps, and compare the results with those obtained from the $f_{\rm coll}$ field of a high-dynamic range simulation (ground truth). While we obtain the results for both the cases, the main focus of the thesis and the machine learning model is the stochastic case. Therefore, this work aims to establish an ML framework for efficiently modelling fields relevant to EoR by bypassing the need to run a high dynamic range N-body simulation, while improving upon the accuracy of semi-analytical prescriptions.

The outline of the thesis is as follows. We start by presenting the details of the methodology in chapter 2. We describe the various simulation boxes and their parameters in section 2.1, introduce Gaussian Process Regression and our implementation of it in section 2.2, the semi-analytical prescriptions of Sheth-Tormen and Press-Schechter for getting the $f_{\rm coll}$ fields in section 2.3, and some details regarding excursion set-based semi-numerical models of reionisation – in particular SCRIPT – in section 2.4. After this, we present the results obtained by combining these components for various different cases in chapter 3. We begin with a fiducial choice of parameters in section 3.1 and compare the fidelity of our GPR emulator in recovering statistics of the $f_{\rm coll}$ field and subsequently, the neutral hydrogen (HI) map with results from a high-dynamic range simulation box. In section 3.2, we first compare the performance of our emulator with the semi-analytical methods and later quantify the robustness of our method against a variation of the involved parameters. We interpret the results and discuss some interesting features of our machine learning model in light of small- and large-scale features of the fields in chapter 4. We conclude by summarising the work and addressing the future directions in chapter 5. Appendix A provides a convergence check of our results on the number of simulation boxes used to train our model, Appendix B highlights some more details regarding the optimisation of the binning to construct the training data, and finally Appendix C discusses the relation between the error in the power spectrum and in the global mean of mass-weighted $f_{\rm coll}$.

Chapter 2

Methodology

2.1 N-body Simulations

Here, we describe the various N-body simulation boxes that are used for training, sampling, and benchmarking the ML model. All of these were run using the GADGET-2[1] code [61], assuming a flat ΛCDM cosmology with $H_0 = 67.8~\mathrm{km\,s^{-1}Mpc^{-1}}$, $\Omega_m = 0.308$, $\Omega_b = 0.04$, $\sigma_8 = 0.829$, $n_s = 0.961$. On the simulation snapshots at the redshifts of interest, we compute the dark matter overdensity field $\delta(\mathbf{x})$ over a default grid size of $\Delta x = 0.5~h^{-1}\mathrm{Mpc}$, using a cloud-in-cell mass-assignment scheme. We then run the Friends-of-Friends (FoF) [62] halo finder on these snapshots (excluding the LB box) to get the discrete halo field. The collapse fraction field $f_{\mathrm{coll}}(\mathbf{x})$ is defined as

$$f_{\mathrm{coll}}(\mathbf{x}) = \frac{\sum_h m_h(\mathbf{x})}{M_{\mathrm{tot}}(\mathbf{x})}\,, \tag{2.1}$$

where the summation runs over the mass of all the haloes $m_h(\mathbf{x})$ contained in the cell at $\mathbf{x}$, and $M_{\mathrm{tot}}(\mathbf{x})$ is the total dark matter mass in the same cell. This field is then computed over the same grid as the density field $\delta(\mathbf{x})$. For the default case, we use 10 as the minimum number of particles for identifying a halo, which corresponds to a minimum halo mass of $4.08 \times 10^8~h^{-1}M_\odot$ for both the SB and RB as defined below, since they have the same particle mass resolution.

[1] https://wwwmpa.mpa-garching.mpg.de/gadget/

- **Small Boxes (SB)**: These have a volume of $V = (40\ h^{-1}\mathrm{Mpc})^3$ and contain $N = 512^3$ particles. 7 realisations of these are run with different seeds, and for each, the overdensity and collapse fraction fields are computed. These pairs of $(\delta, f_{\mathrm{coll}})$ found for each cell are then combined over all cells and over all 7 realisations to get a list of $80^3 \times 7 = 3584000$ $(\delta, f_{\mathrm{coll}})$ pairs, from which the training data is constructed (refer to 2.2.1). Each realisation of these simulations took ~ 210 CPU hours to run, consuming a maximum RAM of around 20 GB.

- **Reference Box (RB)**: This box has a volume of $V = (80\ h^{-1}\mathrm{Mpc})^3$ and number of particles $N = 1024^3$. With both the volume and the number of particles 8 times greater than the SBs, it has the same particle mass resolution $(M_{p,\ \mathrm{min}})$ as them (since $M_{p,\ \mathrm{min}} \propto \frac{V}{N}$), and consequently the same minimum halo mass as well. This box is our 'ground truth' – the goal of our emulator will be to recover the statistics of this high dynamic range box. This simulation took ~ 2900 CPU hours to run, consuming a maximum RAM of 160 GB.

- **Large Box (LB)**: This box has a volume of $V = (80\ h^{-1}\mathrm{Mpc})^3$ and number of particles $N = 512^3$. Therefore, it has a coarser particle resolution than the SBs, but the same volume as the RB. This box is used solely to provide the density values to be input into the emulator and make the f_{coll} predictions to be compared with the ground truth RB, and hence we do not run a halo finder on it. This simulation took ~ 220 CPU hours to run, consuming a maximum RAM of around 20 GB. Note that the combination of SB and LB requires significantly lesser RAM (20 GB) as compared to running the RB (160 GB).

2.2 Building the GPR Emulator

From the SB simulation boxes outlined in section 2.1, we obtain the $(\delta, f_{\mathrm{coll}})$ pairs for each cell. We can then bin the δ values from the SBs, and collect the f_{coll} values falling in each bin to either (a) compute their conditional mean $\langle f_{\mathrm{coll}} | \delta \rangle$ or (b) construct the conditional cumulative distribution function $\mathrm{CDF}(f_{\mathrm{coll}} | \delta)$. Using (a) and (b) to make the f_{coll} predictions precisely corresponds to the deterministic and stochastic cases as defined at the end of section 1, respectively. The GPR training as described in the next subsections is required only for the stochastic case.

2.2.1 Binning

The goal is to use the emulated CDF to directly sample an f_{coll} value, if a new δ value is given as the input. This amounts to the assumption that the spatial distribution of collapse fractions is primarily dictated by the local overdensity, and the cumulative effect of other environmental factors is modeled by random sampling from the conditional CDFs.

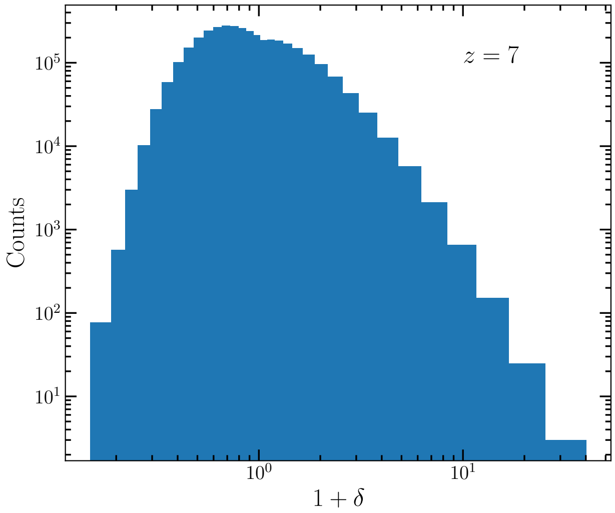


Figure 2.1: A histogram of the overdensity values in the 7 SB boxes combined at a redshift of 7. The logarithmic scale on the y axis indicates the rarity of high and low values. The binning used for the histogram is the same as the optimised binning used for constructing the training data, described in subsection 2.2.1.

The binning of the overdensity values is made trickier by their highly skewed distribution since extremely low and high values are quite rare, as shown in Figure 2.1. If a uniform binning scheme is adopted, to accurately capture the variation of the conditional CDF between two intermediate δ values, the bin width must be made sufficiently small. This causes too few f_{coll} values to be found in higher δ bins, leading to a very noisy CDF. Thus, to strike a balance between noise and systematic error, we adopt a variable binning scheme, where the bin width is set to a reference value at $\delta = 0$, and it increases along either direction. The bins are defined in $\log(1 + \delta)$, and usually have a reference value of around 0.03 dex at $\delta = 0$. The other parameter that we must decide in the training data is the number of bins in f_{coll} used to make the CDFs for a fixed δ bin. This, along with the δ bin widths at the two extremes are optimised for each case that we present separately. The optimal extreme bin widths are around ~ 0.05 dex and ~ 0.2 dex, while the optimal number of f_{coll} bins is either 500 or 900, depending on the case. We refer the reader to Appendix B for more details on the optimisation procedure, where we also study the effect of using a fixed binning scheme (optimised for the default $z = 7$ case) directly on the other cases.

2.2.2 Training Using Gaussian Process Regression

We employ the Gaussian Process Regression (GPR) technique to construct our interpolator function. A collection of random variables $\{Y(t)\}$ indexed by some label t constitute a *Gaussian process* if the joint distribution of $\{Y(t_1), Y(t_2), \ldots, Y(t_n)\}$ is a multivariate Gaussian for any finite subset of t denoted by $\{t_1, t_2, \ldots, t_n\}$. We shall be using Gaussian processes in the context of regression, and hence would like to use the notation f for the random variable and $\mathbf{x}$ for the label. GPR is a non-parametric method that approximates the collection of the target function values f as a Gaussian Process over the inputs $\mathbf{x}$. This is a standard regression algorithm described in Rasmussen & Williams [2] and has been implemented in Scikit-Learn[2]. For the sake of completeness, we briefly outline the basic principles here, following [2]. A Gaussian process is completely specified by a mean function, $\mu(\mathbf{x})$ and a covariance function, $k(\mathbf{x}, \mathbf{x}')$, where

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]\,; \tag{2.2}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]\,, \tag{2.3}$$

where $\mathbb{E}[X]$ denotes the expectation value of the random variable X. The mean function is usually taken to be $\mathbf{0}$ in the prior after appropriate normalisation of the data. The different choices for the covariance kernel are (using the notation $r = |\mathbf{x} - \mathbf{x}'|$) –

- *Radial Basis Function* (RBF) or *Squared Exponential* (SE):

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{r^2}{2l^2}\right) \tag{2.4}$$

- *Rational Quadratic* (RQ):

$$k(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{r^2}{2\alpha l^2}\right)^{-\alpha} \tag{2.5}$$

[2] https://scikit-learn.org/

- *Isotropic Matérn*:

$$k_\nu(\mathbf{x},\mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{\ell}\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}r}{\ell}\right), \tag{2.6}$$

$$k_{\nu=2.5}(\mathbf{x},\mathbf{x}') = \left(1+\frac{\sqrt{5}r}{\ell}+\frac{5r^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}r}{\ell}\right), \tag{2.7}$$

where K_ν is a modified Bessel function, Γ denotes the gamma function, and α, ν, l are all *hyperparameters* specifying the various kernels. The anisotropic variant of any of these kernels can be obtained by scaling the Euclidean distance between $\mathbf{x}$ and $\mathbf{x}'$ differently along each of the three directions, i.e. by setting

$$r^2(\mathbf{x},\mathbf{x}') \coloneqq \frac{(x_1-x_1')^2}{a^2}+\frac{(x_2-x_2')^2}{b^2}+\frac{(x_3-x_3')^2}{c^2}, \tag{2.8}$$

thus introducing three new hyperparameters a, b, and c. x_i is the component of $\mathbf{x}$ along the i^{th} Cartesian axis.

Given a choice of the covariance kernel, let us now briefly outline how GPR makes predictions. Suppose we are given a set of input points $\mathbf{x}_i$ and the true function values at these points $f_i \coloneqq f(\mathbf{x}_i)$, for $i = 1, 2, \ldots, n$. Let the input vector $\mathbf{x}$ be D-dimensional. The goal of regression using Gaussian Processes is to find a good approximation to the function value $f_{*j} \coloneqq f(\mathbf{x}_{*j})$ at any set of desired test inputs $\{\mathbf{x}_{*j} \mid j = 1, 2, \ldots, m\}$. In order to do this, one can utilise the fact that the collection $\{f_i, f_{*j} \mid \forall\, i, j\}$ constitutes an $n+m$ dimensional multivariate Gaussian. Let us denote the collection of *training* inputs $\{\mathbf{x}_i \mid i = 1, 2, \ldots, n\}$ as $\boldsymbol{x}$, and that of the *test* inputs $\{\mathbf{x}_{*j} \mid j = 1, 2, \ldots, m\}$ as $\boldsymbol{x}_*$. The corresponding set of training outputs is $\boldsymbol{f}$ and the set of test outputs, which constitutes what we wish to predict, is $\boldsymbol{f}_*$. We can now express the joint Gaussian distribution of $\boldsymbol{f}$ and $\boldsymbol{f}_*$ as

$$\begin{pmatrix}\boldsymbol{f}\\ \boldsymbol{f}_*\end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.9}$$

$$\implies \begin{pmatrix}\boldsymbol{f}\\ \boldsymbol{f}_*\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} k(\boldsymbol{x},\boldsymbol{x}) & k\left(\boldsymbol{x},\boldsymbol{x}_*\right)\\ k\left(\boldsymbol{x}_*,\boldsymbol{x}\right) & k\left(\boldsymbol{x}_*,\boldsymbol{x}_*\right)\end{pmatrix}\right), \tag{2.10}$$

where the covariance matrix $\boldsymbol{\Sigma}$ has been expanded to show its four sub-matrices which can be defined element-wise as follows

$$[k(\boldsymbol{x}, \boldsymbol{x})]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \tag{2.11}$$

$$[k(\boldsymbol{x}, \boldsymbol{x}_*)]_{ij} = k(\mathbf{x}_i, \mathbf{x}_{*j}) \tag{2.12}$$

$$[k(\boldsymbol{x}_*, \boldsymbol{x})]_{ij} = k(\mathbf{x}_{*i}, \mathbf{x}_j) \tag{2.13}$$

$$[k(\boldsymbol{x}_*, \boldsymbol{x}_*)]_{ij} = k(\mathbf{x}_{*i}, \mathbf{x}_{*j}) \,, \tag{2.14}$$

with the right-hand side of all equations coming from the usual covariance function definition in equation 2.3. If one samples from a multivariate Gaussian with a mean and covariance given by the RHS of equation 2.10, it is clear that the result is a set of 'function values' at the input points $\boldsymbol{x}$ and $\boldsymbol{x}_*$. Sampling repeatedly, we get a *family of functions*. This means that the specification of the mean and covariance gives us a distribution over functions, and since in equation 2.10, no information about the actually observed data outputs $\boldsymbol{f}$ has been incorporated, it can be thought of as a *prior* distribution over functions. The only information used so far is that the set of all function values is jointly Gaussian.

In order to get the *posterior* distribution of the allowed functions, we can use a straightforward property of jointly Gaussian random variables. Suppose $\boldsymbol{a}$ and $\boldsymbol{b}$ are two jointly Gaussian random *vectors* of lengths n_a and n_b, respectively. The full multivariate Gaussian distribution can be expressed as

$$\begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} A & C \\ C^T & B \end{pmatrix} \right) \,, \tag{2.15}$$

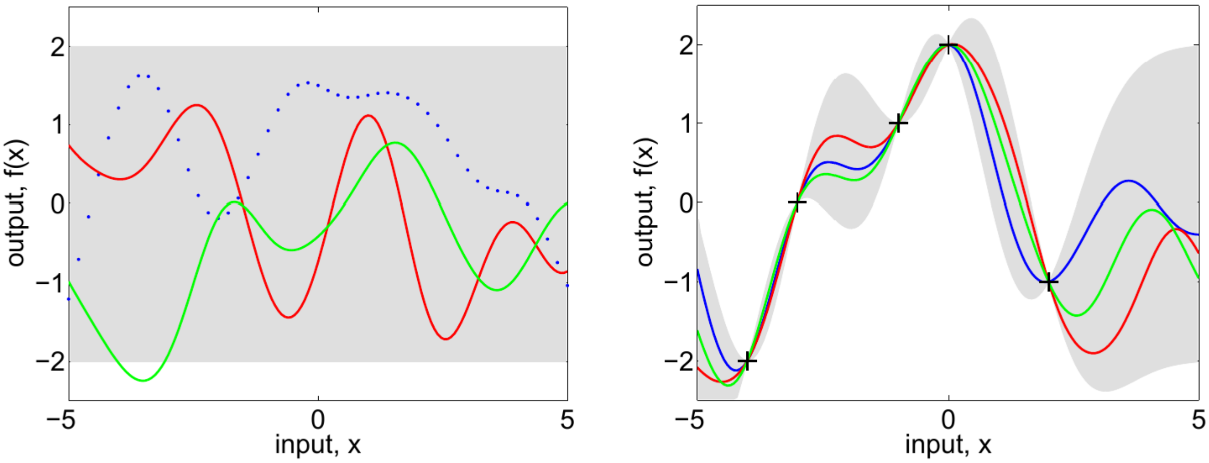
where A and B denote the covariance of the marginal distributions of $\boldsymbol{a}$ and $\boldsymbol{b}$, while C denotes the cross terms (analogous to equations 2.11–2.14). The mean vectors for $\boldsymbol{a}$ and $\boldsymbol{b}$ are denoted by $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$, respectively. Then, the *conditional* distribution of $\boldsymbol{b}$ given $\boldsymbol{a}$ is also a Gaussian with mean and variance given by

$$\boldsymbol{b}|\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}_b + C^T A^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_a), B - C^T A^{-1} C) \,. \tag{2.16}$$

We can directly apply this formula to get the posterior distribution of the function at the test inputs, $\boldsymbol{f}_*$ given all the inputs $\boldsymbol{x}$, $\boldsymbol{x}_*$ and the training outputs $\boldsymbol{f}$. Setting $\boldsymbol{\mu}_a = \boldsymbol{\mu}_b = \mathbf{0}$, $A = k(\boldsymbol{x}, \boldsymbol{x})$, $B = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$, $C = k(\boldsymbol{x}, \boldsymbol{x}_*)$, and $C^T = k(\boldsymbol{x}_*, \boldsymbol{x})$, we finally get

$$\boldsymbol{f}_*|\boldsymbol{x}_*, \boldsymbol{x}, \boldsymbol{f} \sim \mathcal{N}(k(\boldsymbol{x}_*, \boldsymbol{x})[k(\boldsymbol{x}, \boldsymbol{x})]^{-1}\boldsymbol{f}, k(\boldsymbol{x}_*, \boldsymbol{x}_*) - k(\boldsymbol{x}_*, \boldsymbol{x})[k(\boldsymbol{x}, \boldsymbol{x})]^{-1}k(\boldsymbol{x}, \boldsymbol{x}_*)) \,. \tag{2.17}$$

One can now generate functions from this posterior Gaussian to get the prediction of the GPR at the test inputs $\boldsymbol{x}_*$. These predictions are bound to be consistent with the already specified training data points. An example of three such functions randomly drawn from both the prior and the posterior distributions is shown in Figure 2.2. As can be observed, the posterior functions all pass through the data points denoted by '+' and can be rather unconstrained in regions with a lack of data points.



(a), prior

(b), posterior

Figure 2.2: Three functions drawn from (a) the GPR prior (equation 2.10) and (b) posterior (equation 2.17) distributions. In (a) the blue dots denote the function prediction at discrete test inputs, which has been connected by a continuous curve in the other cases. In both the panels, the grey shaded region represents the 2σ ($\sim$ 95%) confidence intervals. '+' is used to denote the training data points in (b). As we can see, the GPR prediction is quite well-constrained near $x = 0$, where there is a relatively high density of data points. *Image credits: [2]*

We only use the posterior mean value at each test input, which is also the maximum a posteriori (MAP) estimate, as the prediction of the GPR rather than drawing from the full Gaussian. For our specific case, we use the anisotropic Matérn kernel with $\nu = 2.5$ as the covariance function. Training the GPR model then entails learning the values of the *hyperparameters* associated with the Matérn kernel given in equation 2.7. While a simpler linear interpolation scheme can also provide similar results for the current work, we opt for GPR as our interpolation method due to its relatively better ability to generalise to the case of a greater number of conditioning variables beyond the dark matter density contrast δ alone. This constitutes future work and is addressed in chapter 5.

The hyperparameter optimisation is carried out using an anisotropic simulated annealing

(ASA) procedure, following [63]. As demonstrated in previous studies of reionisation [64, 65], this method works well for GPR hyperparameter training while preventing issues faced by some other Scikit-Learn methods of getting stuck at a local minimum of the cost function where the corresponding GPR is suboptimal in performance. We briefly outline the details of the algorithm for completeness. First, the data is divided into two parts – one for training and another for validation. Once the training data is specified, the ASA procedure involves evaluating the log marginal likelihood using algorithm 2.1 of [2] over a region with sparsely distributed values in hyperparameter space, which is then iteratively refined to zoom-in on the region of hyper-likelihood maximum (or equivalently, the cost function minimum).

The hyperparameter vector $\mathbf{h}$ that minimises the cost function is then used to make predictions on the validation data, again following algorithm 2.1 of [2] as implemented in Scikit-Learn. A convergence criterion is defined by requiring the magnitude of the 1st and 99th percentiles of $\hat{\alpha} - \alpha$ to be less than a threshold set by the user, called `cv_thresh` (here $\hat{\alpha}$ is the predicted value of the function and α is the true value at the same input). If the convergence criterion is not satisfied, the entire process repeats with a training data larger in size by 10%, and this cycle continues until the maximum number of iterations or $\geq 80\%$ of the full data is used for training. `cv_thresh` is usually taken to be around 0.015 in our case.

Once the training is complete, we have a properly trained interpolator function at our disposal, that can be used to sample the values of the function $\hat{\alpha}$ at any desired input. We use the GPR training to emulate the $\mathrm{CDF}(f_{\mathrm{coll}}|\delta)$ as obtained in the previous subsection, viewed as a function of f_{coll} and δ, thereby setting $\alpha =$ the CDF value. For most cases, the training ends within ~ 10 minutes on 4 CPU cores and uses around 10-15% of the full data for training.

2.2.3 Sampling

Our idea is to be able to recover the f_{coll} field of RB by a combination of information from SB and LB. We have used SB for obtaining the conditional means and the conditional CDFs to train the GPR, and we now use the δ values from the LB as the corresponding input to make the f_{coll} predictions.

For a given input δ_0 from the LB, we return (a) the conditional mean $\langle f_{\mathrm{coll}}|\delta_m \rangle$ for the deterministic case, where δ_m is the middle value of the bin that contains δ_0 and (b) an $f_{\rm coll}$ value randomly drawn from the emulated $\widehat{\rm CDF}(f_{\rm coll}|\delta_0)$ for the stochastic case. We use inverse transform sampling for this case where we draw a random number between 0 and 1 and return the smallest $f_{\rm coll}$ at which $\widehat{\rm CDF}(f_{\rm coll}|\delta_0)$ equals the random number. We can see that the latter method naturally accounts for scatter in the $f_{\rm coll}$ predictions for a fixed δ while the former does not. This sampling is done on a cell-by-cell basis, to produce a prediction of $f_{\rm coll}$ for each cell based on its δ value in LB.

2.3 Semi-Analytic Predictions for Collapse Fraction

An important problem in cosmology is to predict the abundance and spatial distribution of gravitationally bound dark matter haloes using the statistics of the initial dark matter density field. This is particularly important in the context of studying reionisation, since the sources of ionising photons sit within dark matter haloes and therefore the abundance and distribution of these objects as a function of their mass must be known. A widely used framework for this is the excursion set approach applied to the Press-Schechter (PS) formalism [33, 46]. This approach considers the dark matter overdensity field $\delta_R(\mathbf{x})$, smoothed on some scale R, and compares it with a collapse threshold δ_c motivated by the spherical collapse model to determine which regions will form haloes. This is done for successively smaller smoothing scales R and results in a random walk of δ_R, and the probability of crossing the constant barrier δ_c at some radius R can then be computed for a region with a prescribed overdensity δ_0 at a scale R_0. In this way, one can express the conditional collapse fraction of a region parametrised by an overdensity δ_0 and a radius R_0 as

$$f_{\rm coll}(> M_{\rm min}) = {\rm erfc}\left(\frac{\delta_c - \delta_0}{\sqrt{2(\sigma^2(M_{\rm min}) - \sigma^2(M_0))}}\right) , \tag{2.18}$$

where $M_{\rm min}$ is the minimum halo mass considered for computing the collapse fraction, and $\sigma^2(M)$ or $\sigma^2(R)$ denotes the variance in the overdensity field when smoothed over a scale enclosing mass M. For a spherical top-hat filter in real space, the relation between M and the smoothing scale R is simply $M = \dfrac{4}{3}\pi R^3 \bar{\rho}$, where $\bar{\rho}$ is the mean density of the universe. This quantity can be related to the linear matter power spectrum $P_{\rm lin}(k)$ and the Fourier transform of the smoothing kernel $W(k;R)$ as [66]

$$\sigma^2(R) = \int \frac{k^3}{2\pi^2} P_{\text{lin}}(k) |W(k;R)|^2 \, \mathrm{d}\ln k \,. \tag{2.19}$$

While the Press-Schechter formalism provided a decent match to N-body simulations initially, as their resolution improved, it was found that it underestimates the abundance of high-mass halos ($M \gtrsim 10^{10}\ h^{-1}M_\odot$ at $z = 7$) and overpredicts the abundance at low masses ($M \lesssim 10^{8}\ h^{-1}M_\odot$). Therefore, an improved model was suggested by Sheth and Tormen [47, 48] based on the physics of ellipsoidal collapse, which led to a *moving* barrier for the excursion sets unlike the fixed δ_c of Press-Schechter theory. This moving barrier represents a function of the variance $s_M \coloneqq \sigma^2(M)$ which is given at a redshift z by

$$B(s,z) = \sqrt{a}\delta_c(z)\left(1 + \frac{\beta}{(a\nu)^\alpha}\right) \,, \tag{2.20}$$

where $\delta_c(z) = \frac{1.686}{D(z)}$ is the critical overdensity at z, $\nu \coloneqq \frac{\delta_c^2(z)}{s}$, and a, α, β are parameters whose best-fit values are found by matching with N-body simulations. Given this, the conditional Sheth-Tormen (ST) collapse fraction can be computed as

$$f_{\text{coll}}(> M_{\text{min}}) = \int_{s_{\text{cell}}}^{s_{\text{min}}} f(s \mid s_{\text{cell}}, \delta_{\text{L},0}) \, \mathrm{d}s \,, \tag{2.21}$$

$$\text{where} \quad f(s \mid s_{\text{cell}}, \delta_{\text{L},0}) = \frac{1}{\sqrt{2\pi}} \frac{|T\left(s \mid s_{\text{cell}}\ \right)|}{\left(s - s_{\text{cell}}\ \right)^{3/2}} \exp\left[-\frac{\left[B(s) - \delta_{\text{L},0}\right]^2}{2\left(s - s_{\text{cell}}\ \right)}\right] \tag{2.22}$$

$$\text{and} \quad T\left(s \mid s_{\text{cell}}\ \right) = \sum_{n=0}^{5} \frac{\left(s_{\text{cell}}\ - s\right)^n}{n!} \frac{\partial^n \left[B(s) - \delta_{\text{L},0}\right]}{\partial s^n} \,. \tag{2.23}$$

Here $\delta_{\text{L},0}$ is the initial overdensity linearly extrapolated to the present day ($z = 0$). The values of the parameters appearing in equation 2.20 that provide the best fit to the global halo mass function obtained from an N-body simulation at high redshifts ranging between 6–15 (relevant to reionisation) are $a = 0.67, \alpha = 0.7, \beta = 0.4$.

We use SCRIPT's implementation based on equations 2.21–2.23 to obtain the conditional ST collapse fraction given the density field from the LB as an input (for more details on

SCRIPT, refer to the next section). The code requires a smoothing scale (Δx) to be provided, based on which it calculates the $f_{\rm coll}$ values for a set of non-linear overdensity values $\delta_{\rm NL}$. This involves using the spherical collapse approximation to compute the linear overdensities $\delta_{\rm L}$ from the $\delta_{\rm NL}$ values [67], since it is required by equations 2.21–2.23. In the next step, these ($\delta_{\rm NL}$, $f_{\rm coll}$) pairs are used to set up a spline interpolation, which shall then be used to compute the conditional ST $f_{\rm coll}$ for each of the $\delta_{\rm NL}$ values in the cells of LB. We used a similar implementation for obtaining the conditional PS $f_{\rm coll}$ as well, based on equation 2.18.

It turns out that for sufficiently low $\delta_{\rm NL}$ and Δx, a region ends up having its mass $M_0 < M_{\rm min}$ which implies $\sigma^2(M_{\rm min}) < \sigma^2(M_0)$, leading to a negative value inside the square root in the erfc function in equation 2.18. This breaks the spline interpolation and causes it to return only nan $f_{\rm coll}$ for all $\delta_{\rm NL}$. But physically, regions with $M_0 < M_{\rm min}$ should have an $f_{\rm coll} = 0$. If this condition is enforced, the interpolation works but still suffers from numerical errors which cause some $f_{\rm coll}$ values to become negative (with a very small magnitude) instead of 0. One can manually set these $f_{\rm coll}$ values to 0 in order to compare the performance of conditional PS at such fine resolutions. Similarly, one must properly account for such cases in the conditional ST code as well, where numerical errors during interpolation cause small negative $f_{\rm coll}$ values to appear, and the interpolation breaks if one does not set these to 0 at the appropriate place. For the set of $\delta_{\rm NL}$ values chosen for making the interpolator, this issue arises at resolutions finer than or equal to $\Delta x = 0.5\ h^{-1}$Mpc .

Apart from the issue of numerical errors due to interpolation, the non-linear to linear density mapping as predicted by the spherical collapse model becomes increasingly inaccurate as one goes to finer resolutions. In this regime, tidal effects play a significant role in influencing the collapse in high-density environments and spherical symmetry cannot be assumed. This is a limitation of the semi-analytical prescriptions which our GPR model is not subject to, and can thus provide a more accurate way of predicting the $f_{\rm coll}$ distribution and the subsequent HI maps.

2.4 Semi-Numerical Code for Reionisation: SCRIPT

Solving the radiative transfer equations for each cell in a simulation while simultaneously taking into account the coupling between different cells becomes computationally quite expensive, making full RT simulations a rather ineffective way to forward model observables related to the EoR. Therefore, *semi-numerical* models of reionisation have become a popular alternative and offer a much faster and resource-efficient way of generating the distribution of ionised bubbles. Widely employed among these models is the excursion-set approach for identifying ionised bubbles around collapsed dark matter haloes, similar to the one used by Bond et al. for formulating the halo mass function [33].

The simple physical argument at the core of all of these excursion-set based models of reionisation involves declaring a region as ionised if the number of ionising photons in it exceeds the number of atomic hydrogen. We define the *reionisation efficiency parameter* $\zeta(M)$ as the number of ionising photons available per hydrogen in a dark matter halo of mass M. Therefore, a halo of mass M produces $N_p(M)$ many ionising photons given by

$$\begin{aligned} N_p(M) &= \zeta(M) \cdot \text{(number of hydrogen atoms in halo)} && (2.24)\\ &= \zeta(M) \cdot \frac{\text{(mass of hydrogen in halo)}}{m_p} && (2.25)\\ &= \zeta(M) \cdot \frac{\text{(total mass in halo)}}{m_p} \cdot \frac{\Omega_b}{\Omega_m}(1-Y) && (2.26)\\ &= \zeta(M)\frac{M}{m_p}\frac{\Omega_b}{\Omega_m}(1-Y)\,. && (2.27) \end{aligned}$$

where Ω_b and Ω_m are the density parameters for baryons and all matter, respectively, m_p is the mass of the proton, and Y is the mass fraction of Helium. Now, consider an arbitrary region containing dark matter haloes of masses $M_1, M_2, \ldots, M_n$. The total number of ionising photons produced in this region can be obtained by summing over equation 2.27 for all the haloes,

$$\begin{aligned} \sum_{i \in \text{haloes}} N_p(M_i) &= \sum_{i \in \text{haloes}} \zeta(M)\frac{M}{m_p}\frac{\Omega_b}{\Omega_m}(1-Y) && (2.28)\\ &= \frac{\Omega_b}{\Omega_m}\frac{(1-Y)}{m_p}\sum_{i \in \text{haloes}} M\zeta(M) && (2.29)\\ &= \frac{\Omega_b}{\Omega_m}\frac{(1-Y)}{m_p}\zeta\ \Sigma_i M_i, && (2.30) \end{aligned}$$

where we have assumed mass-independence of the parameter ζ for simplicity. $\Sigma_i M_i$ denotes the mass contained in all the haloes of the region. Now let us write an expression for the number of hydrogen N_H in the region. If the mass of hydrogen in the region is M_H,

$$N_H = \frac{M_H}{m_p} \tag{2.31}$$

$$= \frac{M_{\rm tot}}{m_p}\frac{\Omega_b}{\Omega_m}(1-Y)\,, \tag{2.32}$$

where $M_{\rm tot}$ is the total mass contained in the region. Imposing the condition for the region being ionised and substituting the relevant expressions from 2.30 and 2.32,

$$\sum_{i\in \text{haloes}} N_p(M_i) \geq N_H \tag{2.33}$$

$$\frac{\Omega_b}{\Omega_m}\frac{(1-Y)}{m_p}\zeta\ \Sigma_i M_i \geq \frac{M_{\rm tot}}{m_p}\frac{\Omega_b}{\Omega_m}(1-Y) \tag{2.34}$$

$$\zeta\frac{\Sigma_i M_i}{M_{\rm tot}} \geq 1\,. \tag{2.35}$$

Defining the collapse fraction $f_{\rm coll}$ for a region as the ratio of its mass in haloes to its total mass, $f_{\rm coll} = \frac{\Sigma_i M_i}{M_{\rm tot}}$, we arrive at the condition for flagging a region as ionised,

$$\boxed{\zeta f_{\rm coll} \geq 1} \tag{2.36}$$

In practice, one computes the $f_{\rm coll}$ values not for arbitrary regions of space, but as a *field* over a grid with a specified resolution (Δx) over which the non-linear dark-matter density contrast field $\delta(\mathbf{x})$ is also defined. Therefore, in standard excursion set-based (ES) semi-numerical models, the above condition takes the following form

$$\zeta f_{\rm coll}(\mathbf{x}, R) \geq 1\,, \tag{2.37}$$

where $f_{\rm coll}(\mathbf{x}, R)$ refers to the $f_{\rm coll}$ field averaged over a spherical region of radius R, $f_{\rm coll}(\mathbf{x}, R) = \langle f_{\rm coll}(\mathbf{x})(1+\delta_{\rm NL}(\mathbf{x})\rangle_R$. This spherical region could be defined as a sphere of radius R in position space (known as a spherical top-hat filter) or in momentum space (known as a sharp-k filter). If the above condition is satisfied for any value of R centered around $\mathbf{x}$, the grid cell at that location is considered *ionised* with an ionisation fraction $x_{\rm HII}(\mathbf{x}) = 1$ assigned to that cell. Alternatively, some implementations flag the entire region of radius R as ionised if the condition is satisfied [35]. In case the condition is not satisfied, the *partially* ionised cell is assigned an $x_{\rm HII}(\mathbf{x}) = \zeta f_{\rm coll}(\mathbf{x})$, where the $f_{\rm coll}$ value used is the one evaluated at that cell only (without any averaging with others).

One would expect that in the absence of recombinations, the number of ionising photons and the number of ionised hydrogen match. This can be quantified by comparing the mass-averaged global ionised fraction $Q^M_{\rm HII} = \langle x_{\rm HII}(\mathbf{x})(1 + \delta(\mathbf{x}))\rangle$ with the global average $\langle \zeta f_{\rm coll}(\mathbf{x})(1 + \delta(\mathbf{x}))\rangle$. It turns out that the prescriptions of assigning ionised fractions to cells as discussed above end up violating this condition of *photon number conservation* [68]. This leads to a more drastic problem of the large-scale power spectrum of the HI density field becoming dependent on the resolution of the grid Δx used to generate the $f_{\rm coll}$ and density fields. This in turn poses an issue for accurately modelling the 21 cm power spectrum at the level of precision expected from upcoming observational projects such as the SKA [40].

The **S**emi-numerical **C**ode for **ReI**onisation with **P**ho**T**on-conservation (SCRIPT)[3] [40] offers a solution to this problem by evaluating the distribution of ionised regions through an explicitly photon-conserving algorithm in two rounds. Firstly, the number of ionising photons generated by sources within a cell at $\mathbf{x}$ is computed as

$$N_\gamma(\mathbf{x}) = \zeta f_{\rm coll}(\mathbf{x})[1 + \delta(\mathbf{x})]\bar{n}_H \,, \tag{2.38}$$

where $\bar{n}_H$ is the mean hydrogen number density. If $\zeta f_{\rm coll}(\mathbf{x})$ is greater than 1, then there is an excess of ionising photons in this region which gets isotropically distributed to the nearest cells, with the cell at $\mathbf{x}'$ consuming $[1 + \delta(\mathbf{x}')]\bar{n}_H$ of these photons. If $N_\gamma(\mathbf{x})$ is sufficient to ionise all of these cells, they are given an ionised fraction of 1 and the process repeats until there are insufficient photons to fully ionise a set of nearest cells. In this case, the photons are equally distributed among all such cells and they acquire an ionised fraction less than 1 at the end of the first round. This process is done independently centered on all cells, and may lead to *overionised* cells with $x_{\rm HII} > 1$.

In the second round, these overionised cells are treated as secondary sources and their excess photons given by $N^{(2)}_\gamma(\mathbf{x}) = [x_{\rm HII}(\mathbf{x}) - 1][1 + \delta(\mathbf{x})]\bar{n}_H$ are distributed to the nearest underionised cells with $\zeta f_{\rm coll}(\mathbf{x}) < 1$ in a similar isotropic manner as described before. The nearest overionised cells from the first round are left unchanged. This process is similarly repeated for all overionised cells until none remain. This algorithm is manifestly photon conserving since it always uses all the photons produced by all the sources to ionise the hydrogen. A small difference between other ES-based models and SCRIPT is that a partially

[3] https://bitbucket.org/rctirthankar/script

ionised cell at $\mathbf{x}$ in the former will always have a value of ionisation fraction equal to $\zeta f_{\rm coll}(\mathbf{x})$, whereas in the latter it can be exceeded due to extra contribution from secondary ionising sources (the overionised cells).

As a result, SCRIPT can be used to produce HI maps during reionisation that have a large-scale power spectrum that is properly converged with respect to the grid resolution of the $f_{\rm coll}$ and density fields [40]. In this work, we will use SCRIPT as our model of reionisation to obtain the HI and HII maps from the $f_{\rm coll}$ field supplied using various methods (semi-analytical or simulation-based). We assume a constant ζ value throughout, which corresponds to an ionising emissivity for dark matter haloes that is proportional to their mass.

Chapter 3

Results

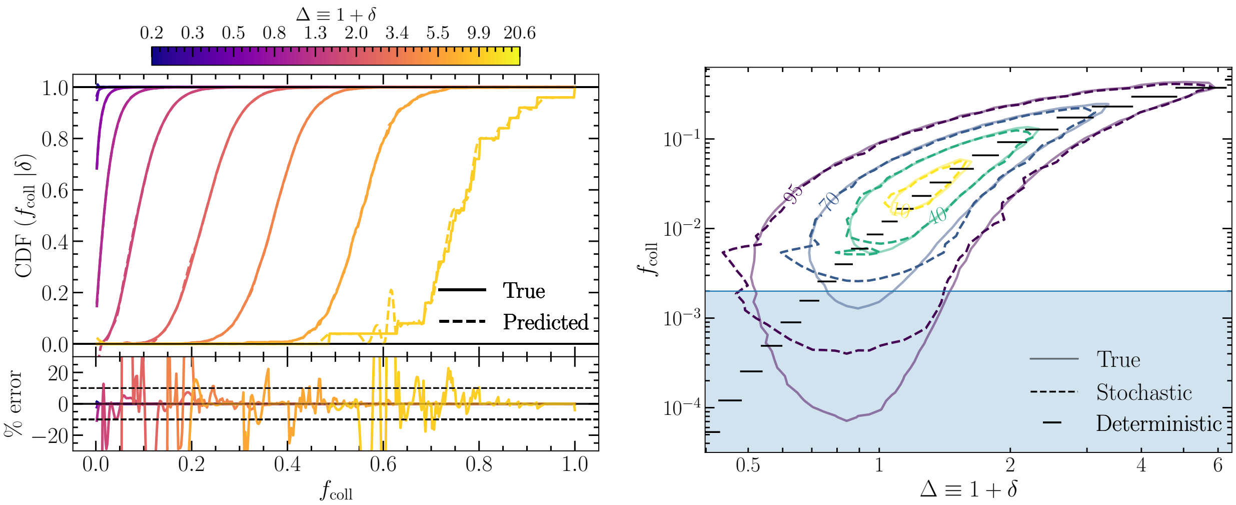
In this chapter, we benchmark the various f_{coll} predictions against the ground truth taken from the RB. Our primary interest lies in modelling the neutral hydrogen density field during the Epoch of Reionisation (EoR), and we obtain this from the collapse fraction field by using **S**emi-numerical **C**ode for **R**e**I**onisation with **P**ho**T**on-conservation (SCRIPT)[1] [40]. The section is divided into two parts - Fiducial, where we discuss the f_{coll} and SCRIPT results corresponding to the fiducial choice of parameters and Variation, where we compare the SCRIPT results for the semi-analytical methods and extend them to variations in the parameters.

3.1 Fiducial

3.1.1 Collapse Fraction

We consider our fiducial case to have redshift $z = 7$, grid size $\Delta x = 0.5\ h^{-1}$Mpc , and minimum halo mass $M_{h,\text{min}} = 4.08 \times 10^8\ h^{-1}M_\odot$ (corresponding to 10 particles per halo for the SB and RB). Henceforth, we shall use the notation $\Delta \equiv 1 + \delta$. We first look at the results of GPR training, by comparing the emulated and training CDFs as a function of f_{coll} conditioned on 10 different Δ values, as shown in Figure 3.1a. It can be seen that both the

[1] https://bitbucket.org/rctirthankar/script



(a)

(b)

Figure 3.1: (a) Comparison between the true CDF from the training data and the interpolator's prediction, shown at 10 different Δ values. The relative error occasionally blows up due to the small values of the CDFs, (b) Comparison of the joint distribution of $f_{\rm coll}$ and Δ. The 10, 40, 70 and 95 percentile contours are shown and the blue region demarcates the $f_{\rm coll}$ less than the first bin edge defined during the $f_{\rm coll}$ binning. The conditional means calculated in the deterministic case for each delta bin are also shown using black horizontal lines. This shows that over most of the $f_{\rm coll}$ and Δ range, our interpolator recovers the true distribution to a high accuracy.

training and prediction CDF become noisy at very high Δ, due to a smaller number of $f_{\rm coll}$ values.

The recovery of the joint distribution of non-zero $f_{\rm coll}$ values and their corresponding Δ is shown in Figure 3.1b. While the contours are very similar between truth and prediction at intermediate to high $f_{\rm coll}$ and Δ, the very low $f_{\rm coll}$ values are not recovered as well. We understand this to be a limitation of the way we set up the training data for the GPR, where the smallest value of the training CDF that is fed into the GPR is CDF($f_{\rm coll} = 0.002$). The region below $f_{\rm coll} = 0.002$ is shaded in blue. The interpolator ends up overestimating the CDF at $f_{\rm coll}$ below this threshold and that leads to an oversampling of $f_{\rm coll} = 0$ values, and consequently an undersampling of very low $f_{\rm coll} \lesssim 10^{-3}$. Attempting to fix this problem by incorporating CDF($f_{\rm coll} = 0$) during the training does not provide any significant improvement over our current choice for the joint distribution or the rest of our results. In Figure 3.1b, we have also shown the $\langle f_{\rm coll}|\delta\rangle$ values for various δ bins using short horizontal black lines. The variable length of the horizontal line reflects the variable bin widths in δ. We distinguish between the collapse fraction $f_{\rm coll}$ computed from equation 2.1 (constrained to be between 0 and 1) and the mass-averaged collapse fraction $f_{\rm coll}^M(\mathbf{x}) \equiv f_{\rm coll}(\mathbf{x})(1+\delta(\mathbf{x}))$, where as usual, for the predicted (true) $f_{\rm coll}^M$ the δ is taken to be from the LB (RB). We use the following expressions to compute the auto power spectrum of a field $g(\mathbf{x})$, denoted by $P_g(k)$, and its cross power spectrum with another field $h(\mathbf{x})$, denoted by $P_{gh}(k)$:

$$\frac{\langle g(\mathbf{k}) g^*(\mathbf{k}')\rangle}{\bar{g}^2} = (2\pi)^3 P_g(k)\delta_D(\mathbf{k}-\mathbf{k}')\,, \tag{3.1}$$

$$\frac{\langle g(\mathbf{k}) h^*(\mathbf{k}')\rangle}{\bar{g}\bar{h}} = (2\pi)^3 P_{gh}(k)\delta_D(\mathbf{k}-\mathbf{k}')\,, \tag{3.2}$$

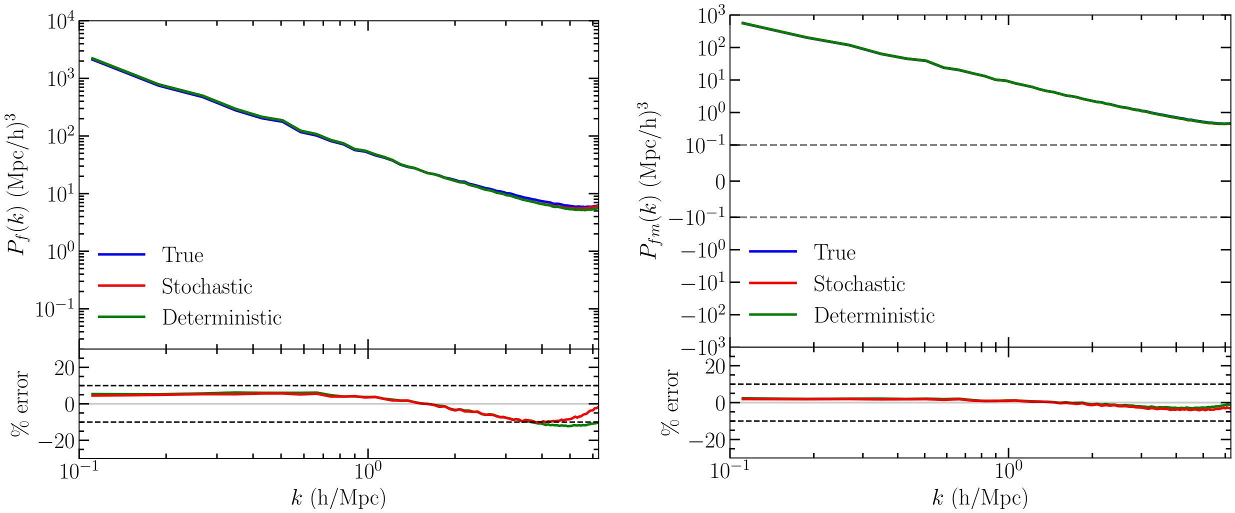
where $g(\mathbf{k})$ and $\bar{g}$ are, respectively, the Fourier conjugate and mean of $g(\mathbf{x})$, δ_D denotes the Dirac delta function, an asterisk denotes complex conjugation and the angular brackets represent an average over Fourier space such that $|\mathbf{k}| = k$.

We compute the auto and cross power spectra by setting $g = f_{\rm coll}^M$ and $h = \Delta$ respectively in the above, for both the deterministic and stochastic cases, and compare them with the truth in Figure 3.2. The agreement between the auto power spectra is within 5% for $k \lesssim 2$ $h\,{\rm Mpc}^{-1}$, and at the smallest scales stays within 10% for the stochastic case whereas for the deterministic case it worsens to slightly below -10%. This implies that there is some extra small-scale power in the stochastic $f_{\rm coll}$ field, and this difference will become more stark once we go to the HI density field in 3.1.2. The cross-power spectrum is recovered better as expected, with sub-2% errors for most of the k range and only becoming $\sim 5\%$ at the smallest scales. We can see that the level of agreement is very similar between the stochastic and deterministic cases.

If we take a closer look at Figure 3.2a, the error in the large-scale power is mostly constant for $k \leq 0.7$ $h\,{\rm Mpc}^{-1}$. Moreover, this error arises predominantly due to the error in the mean of $f_{\rm coll}^M$ between the truth and predictions. This implies a good agreement ($< 1\%$) at large-scales between the un-normalised power spectra, computed by dropping the $\bar{g}^2$ in the auto power as defined in equation 3.1. We address this issue in Appendix C.

3.1.2 Hydrogen Density Fields

As mentioned earlier, we use SCRIPT to model the HI and HII fields relevant to EoR. SCRIPT constructs ionised bubbles around sources by allowing regions to receive photons from mul-



(a)

(b)

Figure 3.2: Comparison of (a) $f_{\rm coll}^M$-$f_{\rm coll}^M$ auto and (b) $f_{\rm coll}^M$-$\mathbf{\Delta}$ cross power spectra, between truth and predictions using the deterministic and stochastic cases. The two cases behave similarly for most of the k range, with the stochastic case having slightly more power at the smallest scales, and this effect gets amplified in the corresponding HI maps (see Figure 3.4).

tiple sources and possibly get 'overionised' with an ionisation fraction greater than 1. These excess photons are then distributed in the nearby cells causing them to become ionised, until the ionisation levels of all overionised cells have been properly adjusted. In this manner, SCRIPT incorporates photon conservation explicitly and hence achieves a large-scale HI power spectrum that is converged with respect to the spatial resolution of the $f_{\rm coll}$ and density fields.

The code requires $f_{\rm coll}(\mathbf{x})$ at the desired redshift, along with the *reionisation efficiency parameter* ζ as the primary inputs. ζ gives the number of ionising photons in the intergalactic medium per hydrogen in dark matter haloes. The output is an ionisation (HII) fraction field $x_{\rm HII}(\mathbf{x})$, which can then be used to get an HI fraction field, $x_{\rm HI}(\mathbf{x}) = 1 - x_{\rm HII}(\mathbf{x})$. Upon mass-averaging these, we get the HII and HI density fields upto normalisation:

$$x_{\rm HI}^M(\mathbf{x}) = x_{\rm HI}(\mathbf{x})(1 + \delta(\mathbf{x})) \propto \rho_{\rm HI}(\mathbf{x})\,; \tag{3.3}$$

$$x_{\rm HII}^M(\mathbf{x}) = x_{\rm HII}(\mathbf{x})(1 + \delta(\mathbf{x})) \propto \rho_{\rm HII}(\mathbf{x})\,. \tag{3.4}$$

We use the stochastic, deterministic and true collapse fraction fields as the input to SCRIPT and generate the HI and HII maps. We assume a constant ionising efficiency ζ

and calibrate it for all the three cases separately such that the global ionisation fraction, $Q^M_{\rm HII} \equiv \langle x^M_{\rm HII}(\mathbf{x}) \rangle$ is 0.5 (this is our fiducial setting). A comparison of the HI density field $x^M_{\rm HI}(\mathbf{x})$, at a slice through $z = 50\ h^{-1}$Mpc is then shown in Figure 3.3. We also compute statistics such as the auto and cross (with Δ) power spectra of the HII and HI density fields, computed as given in equation 3.1. The comparison between the deterministic, stochastic and true cases for the fiducial $Q^M_{\rm HII} = 0.5$ can be seen in Figure 3.4.

For the HI auto power spectrum, it is clear that the error in the recovery of large-scale power ($k \lesssim 1\ h\,{\rm Mpc}^{-1}$) is similar between the deterministic and stochastic cases, with both being around 10% in magnitude. Interestingly, at the smallest scales, the deterministic case underestimates the power with a large error of around $35 - 40\%$ whereas the stochastic case has a better agreement of around $20 - 25\%$. For the HII auto power, the recovery is more consistent between the two cases, being well within 10% for the entire k range. This highlights the crucial role played by stochasticity in correctly predicting specifically the HI map during reionisation, and we discuss this further in section 4.

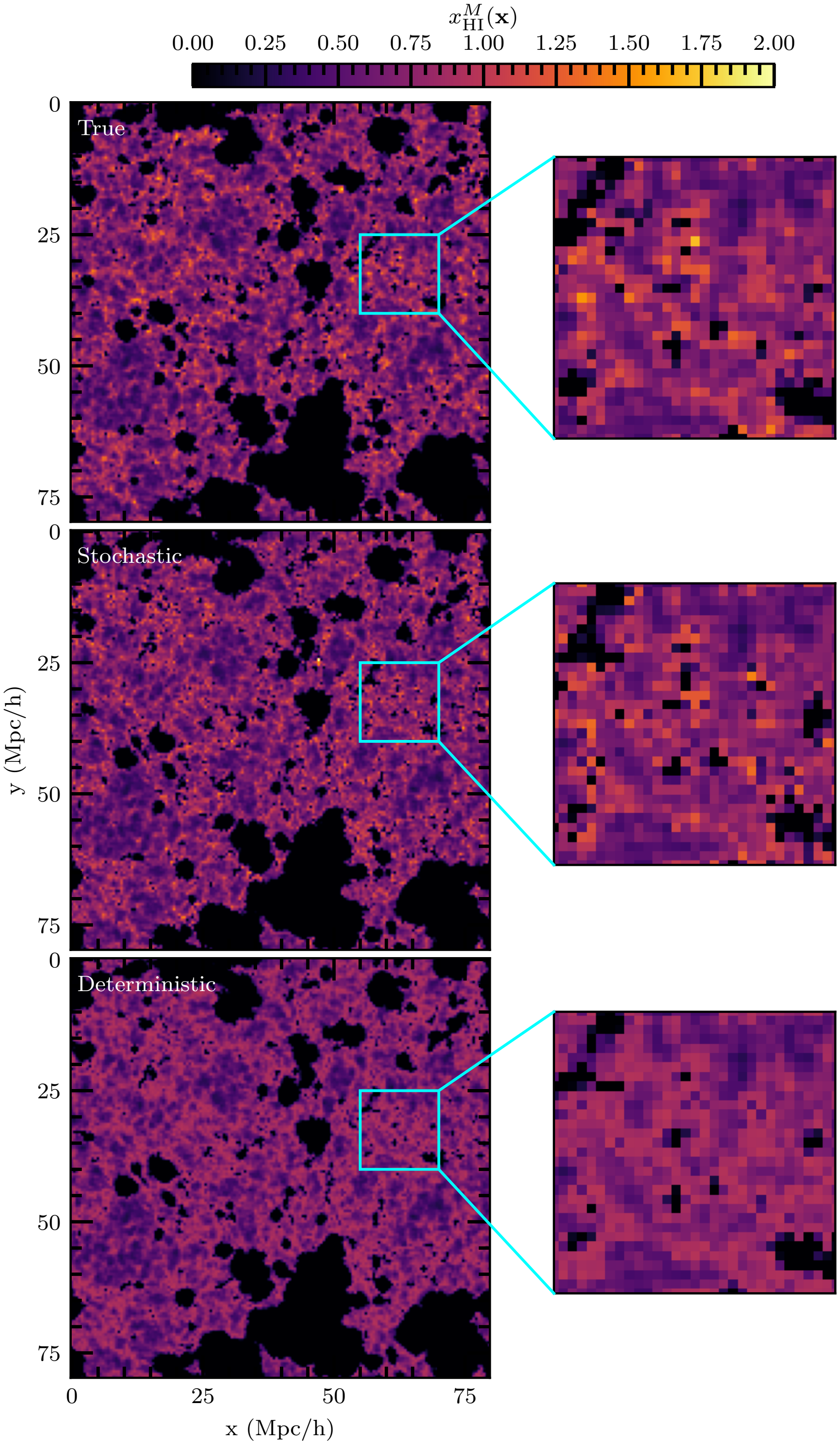
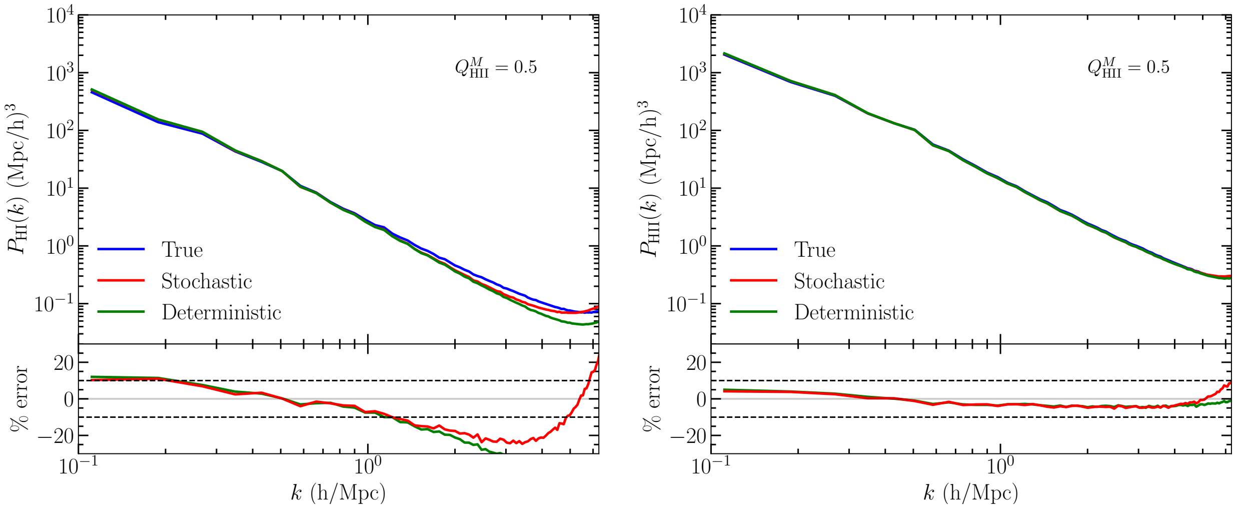


Figure 3.3: The neutral HI density field at $Q^M_{\rm HII} = 0.5$ in the ground truth *(top panel)*, as recovered by our ML interpolator (stochastic, *middle panel*), and as recovered using the conditional means (deterministic, *bottom panel*) at a slice through $z = 50\ h^{-1}$Mpc . The black regions are the ionised bubbles. The deterministic case washes out small-scale HI fluctuations to a large extent. The stochastic case recovers them better, but with some inaccuracy in the small-scale HI density correlations (refer to chapter 4 for a detailed discussion).



(a)

(b)

Figure 3.4: (a) HI-HI, (b) HII-HII power spectra for truth and predictions using the stochastic and deterministic cases ($Q^M_{\mathrm{HII}} = 0.5$). In (a), while the large-scale power is recovered similarly across both cases, the stochastic case has a better accuracy at small scales. This highlights the importance of scatter in f_{coll} around the mean in contributing to the HI density fluctuations at small scales.

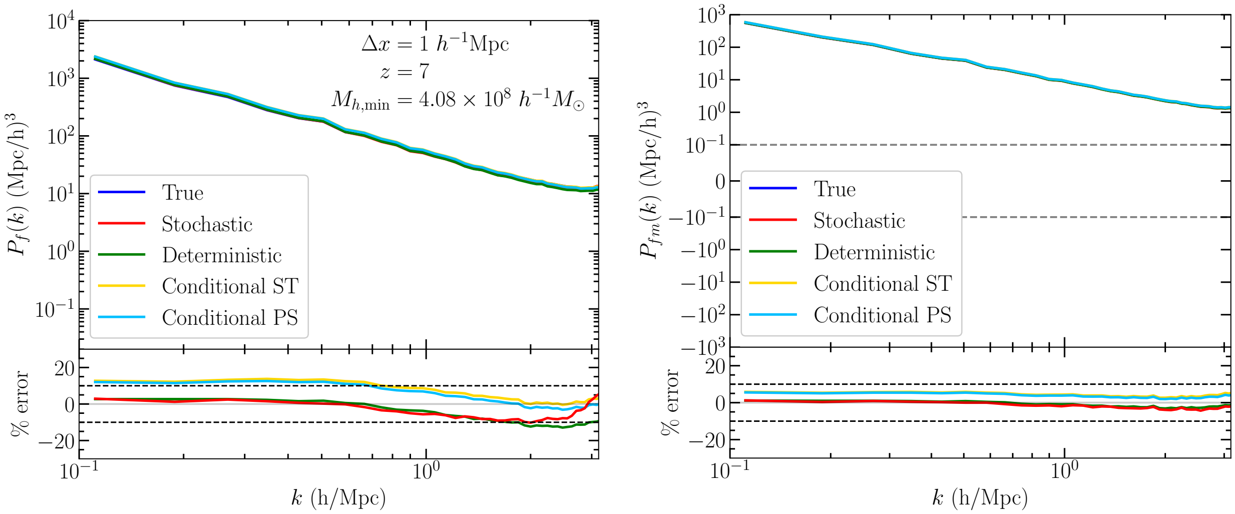
3.2 Variation

3.2.1 Semi-Analytical Methods

The extended Press-Schechter formalism [33, 46] calculates the conditional mass function of dark matter haloes by using an excursion set approach for identifying gravitationally collapsed regions, with a constant barrier given by the threshold linear density for spherical collapse δ_c. A better match to N-body simulations is provided by the conditional mass function calculated by Sheth & Tormen [47, 48] by accounting for ellipsoidal collapse. This leads to a barrier definition that depends on the variance of the smoothed density field at the scale under consideration. Both of these semi-analytical methods have been commonly used in semi-numerical models of reionisation [35, 38, 69] to prescribe the $f_{\mathrm{coll}}(\mathbf{x})$ field, given the underlying density field. It is also important to note that the resultant $f_{\mathrm{coll}}(\mathbf{x}|R, \delta_0)$ is obtained using a deterministic formula in the smoothing scale R and the overdensity of the region δ_0. This leads to an f_{coll} field prediction analogous to our deterministic case where the scatter around the f_{coll} values due to varying environmental features beyond δ is neglected.

We aim to use the density field of the LB as an input to obtain the conditional PS and ST collapse fraction fields and compare the results with our stochastic and deterministic cases. It turns out that when the conditional ST and PS $f_{\rm coll}$ fields are evaluated at $\Delta x = 0.5\ h^{-1}$Mpc, some $f_{\rm coll}$ values that should be 0 instead take on small negative values due to numerical errors in the interpolation between the input δ and the semi-analytical $f_{\rm coll}$. This problem does not arise in the $\Delta x = 1\ h^{-1}$Mpc case, and hence we use that as our primary comparison between the semi-analytical, stochastic and deterministic cases.

Therefore, keeping all the other parameters ($Q^M_{\rm HII}$, z, $M_{h,{\rm min}}$) fixed at the fiducial, we vary Δx to 1 h^{-1}Mpc and rerun all the cases. To get the conditional PS and ST results, we use the density field from the LB as the input. The resulting $f^M_{\rm coll}$ power spectra comparison is shown in Figure 3.5.



(a)

(b)

Figure 3.5: Comparison of (a) $f^M_{\rm coll}$-$f^M_{\rm coll}$ auto and (b) $f^M_{\rm coll}$-Δ cross power spectra, between truth, stochastic, deterministic and the semi-analytical predictions. The semi-analytical cases show a greater error in the large-scale $f^M_{\rm coll}$ power as well as in the global $f^M_{\rm coll}$ mean (see Appendix C).

We can clearly observe that both the methods that use the simulations to generate $f_{\rm coll}$ values (stochastic and deterministic) perform better in recovering the power than the semi-analytical prescriptions, except at very small scales. Proceeding to the HI and HII density fields and computing their power spectra, we compare the results in Figure 3.6. At least for the HI density field, the power at the largest and the smallest scales has a significantly greater error as compared to the stochastic case. Even upon making these comparisons for the $\Delta x = 0.5\ h^{-1}$Mpc case by setting the small negative $f_{\rm coll}$ values to 0, we find a similar improvement in accuracy over the semi-analytical methods. This demonstrates the effectiveness of our method at such small scales where tidal effects become important, and the spherical collapse model used by the semi-analytical prescriptions to get the mapping between non-linear and linear density fields is inaccurate.

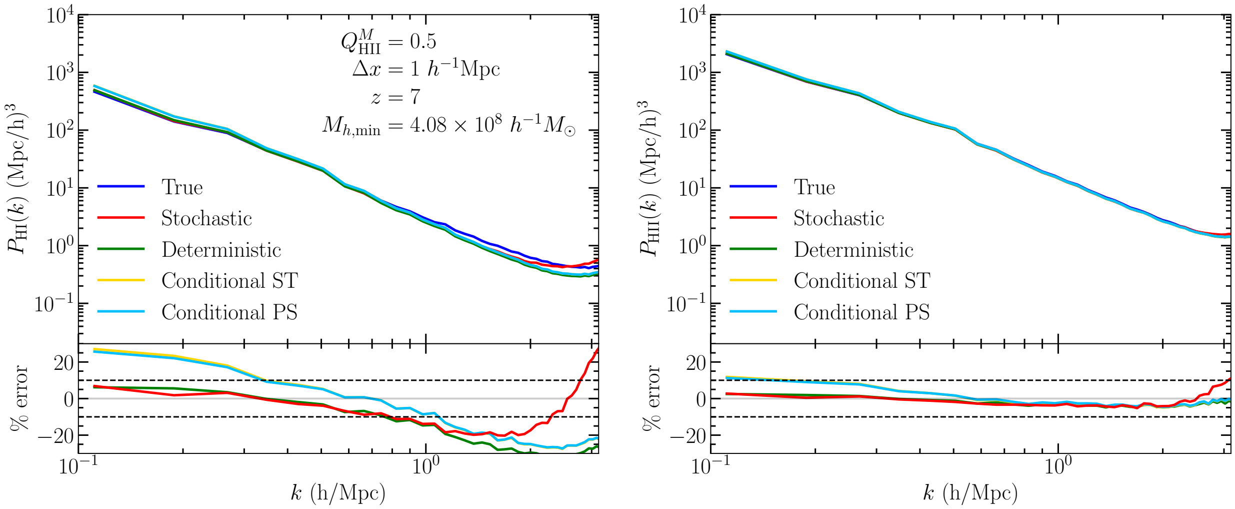


Figure 3.6: Comparison of HI-HI *(left)*, HII-HII *(right)* power spectra between truth, stochastic, deterministic and the semi-analytical predictions at $Q^M_{\rm HII} = 0.5$. The stochastic and deterministic methods present a huge improvement over the semi-analytical cases in recovering the large-scale HI power, adding to their pre-existing advantage of being more viable at scales finer than $\Delta x = 1\ h^{-1}$Mpc.

We now settle on the stochastic case and investigate the robustness of the method, in particular the SCRIPT results, against a variation of the involved parameters. Hereafter, the 'Predicted' label on the plots refers to the stochastic case. We perform convergence tests with respect to simulation box parameters such as grid size on the one hand, and physical parameters such as ionised fraction, redshift and minimum halo mass on the other.

3.2.2 Ionisation Fraction

The fraction of ionised hydrogen $Q^M_{\rm HII}$ in the whole box is controlled by the ionising efficiency ζ. This is a crucial quantity that directly controls the size of the ionised bubbles and, hence, the evolution of the IGM during reionisation. It is commonly treated as a free parameter in studies of reionisation, and therefore, it is important for our method to work well for a range of ζ values. So far, we have used a ζ for the stochastic case which gives $Q^M_{\rm HII} = 0.5$, and this value comes out to be ~ 10.8. We now change $Q^M_{\rm HII}$ to 0.25 and 0.75, keeping everything else the same ($M_{h,\,\rm min} = 4.08 \times 10^8\ h^{-1}M_\odot, \Delta x = 0.5\ h^{-1}{\rm Mpc}, z = 7$). Figure 3.7 shows the results for the auto and cross power spectra of the HI and HII density fields.

As seen before, the large-scale HI auto power is recovered at the $\sim 10\%$ level for the $Q^M_{\rm HII} = 0.5$ case, down to $k \sim 1.5\ h\,{\rm Mpc}^{-1}$. The $Q^M_{\rm HII} = 0.75$ case is even better, with a 5% error over a similar k range. The HI cross power is also similar, with sub-5% errors initially that increase to around 10% by $k \sim 1.5\ h\,{\rm Mpc}^{-1}$. The ionisation field auto and cross are recovered much better, with at least a fidelity of $\sim 5\%$ down to $k \sim 2\ h\,{\rm Mpc}^{-1}$, regardless of the $Q^M_{\rm HII}$ value. We see relatively larger errors in the $Q^M_{\rm HII} = 0.25$ case and at small scales ($k \gtrsim 2\ h\,{\rm Mpc}^{-1}$) even for other ionised fractions, at least in the HI results. The relatively greater disagreement for $Q^M_{\rm HII} = 0.25$ at large-scales is related to the behaviour of the large-scale HI bias (defined below) in the truth at ionisation fractions close to 0.25, and is described in chapter 4.

We then proceed to calculate the HI (HII) bias denoted by $b_{\rm HI}$ ($b_{\rm HII}$) and given by the expressions

$$b^2_{\rm HI}(k) = \frac{P_{\rm HI}(k)}{P_m(k)}\,; \qquad\qquad b^2_{\rm HII}(k) = \frac{P_{\rm HII}(k)}{P_m(k)}, \tag{3.5}$$

where $P_{\rm HI}(k)$ ($P_{\rm HII}(k)$) is the HI (HII) auto power spectrum and $P_m(k)$ is the matter power spectrum, both computed using equation 3.1. We compute the HI and HII bias only at three different low k values, and study their variation as a function of the global ionised fraction $Q^M_{\rm HII}$ in Figure 3.8. In the HII bias plot, we have also plotted the $f_{\rm coll}$ bias (which is independent of $Q^M_{\rm HII}$ by construction). One can observe the HII bias to be clearly approaching the $f_{\rm coll}$ bias, at sufficiently low $Q^M_{\rm HII}$ [40]. We can also see that for higher k, the deviation from the $f_{\rm coll}$ bias happens for a lower value of $Q^M_{\rm HII}$.

3.2.3 Redshift

The redshift z directly affects structure formation, with an increasing number of collapsed haloes forming at lower redshifts. This changes the distribution and abundance of the sources of ionising photons, leading to different ionised bubble topologies for a given $Q^M_{\rm HII}$. Therefore, if we wish to use our emulator for studying the redshift evolution of the HI density field, we must ensure that it is accurate for multiple redshifts. We vary the redshift to two other values $z = 5$ and $z = 9$, while keeping a fixed $Q^{M}_{\rm HII} = 0.5$ and $\Delta x = 0.5h^{-1}$Mpc. The effects on the HI and HII power spectra are captured in Figure 3.9. For the HI field, the agreement remains within around 10%, at least upto $k \sim 1\ h\,\text{Mpc}^{-1}$. The overall trend is similar across redshifts, with small scales around $k \sim 3\ h\,\text{Mpc}^{-1}$ showing a significant dip in the power. The HII results are a lot better with the errors not exceeding 5% for almost the entire k range.

3.2.4 Minimum Halo Mass

The calculation of $f_{\rm coll}$ at each cell assumes a halo mass cutoff $M_{h,\,\rm min}$, and this corresponds to the lowest mass haloes that are capable of producing ionising photons through the formation of luminous objects such as stars or galaxies. For haloes where cooling of the infalling baryonic matter is governed by radiation through atomic transition lines, $M_{h,\,\rm min}$ is around $10^8\ h^{-1}M_{\odot}$ [4]. Changing $M_{h,\,\rm min}$ can significantly alter the ionising photon budget and hence the distribution of ionised bubbles, since the low-mass haloes are more abundant than very high-mass ones.

Since different reionisation models may assume different values of $M_{h,\,\rm min}$, we now vary it by varying the minimum number of particles used by the FoF group finder for identifying a halo from the default 10 to 40 and 80, while fixing $z = 7$, $Q^{M}_{\rm HII} = 0.5$ and $\Delta x = 0.5h^{-1}$Mpc. This changes $M_{h,\rm min}$ from $4.08 \times 10^8\ h^{-1}M_{\odot}$ to $1.63 \times 10^9\ h^{-1}M_{\odot}$ and $3.26 \times 10^9\ h^{-1}M_{\odot}$ respectively, and the corresponding results are shown in Figure 3.10. We see $\sim 12\%$ error at the largest scales in the HI results, that falls and stays within 10% till $k \sim 1\ h\,\text{Mpc}^{-1}$, and the HII power spectra remain within 7-8% for almost the entire k range.

3.2.5 Grid Resolution

The grid size used for CIC-smoothing the density and $f_{\rm coll}$ fields, Δx, determines the resolution at which the sources are identified and impacts the distribution of ionised regions. Rather than a physically interpretable parameter, this is a choice that one must make in order to construct the density field from the simulation box and make $f_{\rm coll}$ predictions. Therefore, we check our model's sensitivity to it by varying it from the fiducial value of $\Delta x = 0.5$

h^{-1}Mpc to two other values, $\Delta x = 0.25\ h^{-1}$Mpc and $\Delta x = 1\ h^{-1}$Mpc while keeping a fixed $z = 7$ and $Q^{M}_{\rm HII} = 0.5$, with the results shown in Figure 3.11.

The $\Delta x = 0.25$ case suffers larger errors for the HI power spectra, but the other two cases have similar $\lesssim 10\%$ agreements at large scales below $k = 1\ h\,{\rm Mpc}^{-1}$. However, achieving results corresponding to $\Delta x = 0.25\ h^{-1}$Mpc is simply not possible using the conditional PS and ST prescriptions without resorting to ad hoc assumptions, such as setting negative $f_{\rm coll}$ values to zero. Our interpolator enables this and represents a significant improvement over the current state of the art. The HII results are again a lot more robust, always performing better than 10% across all k.

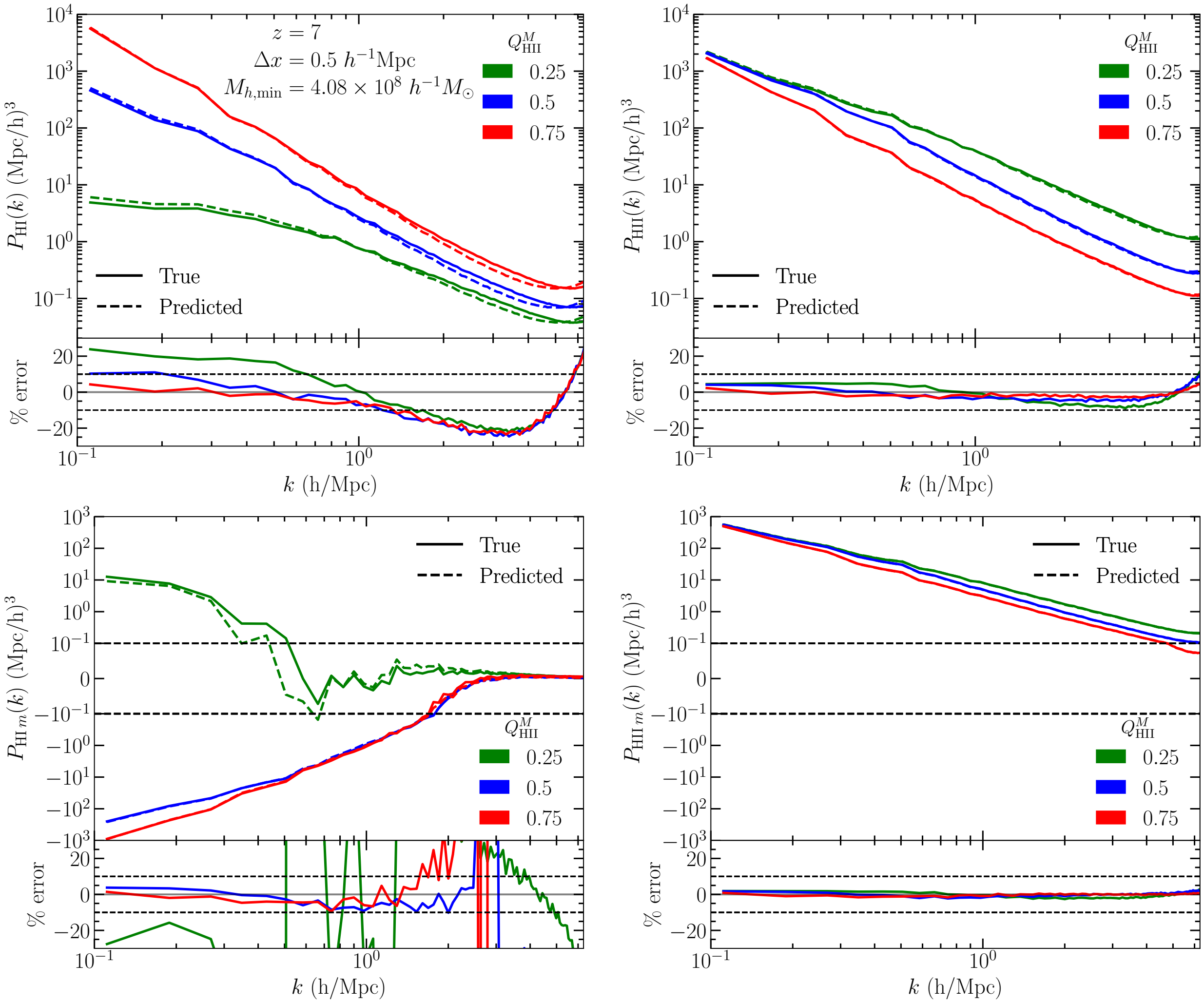
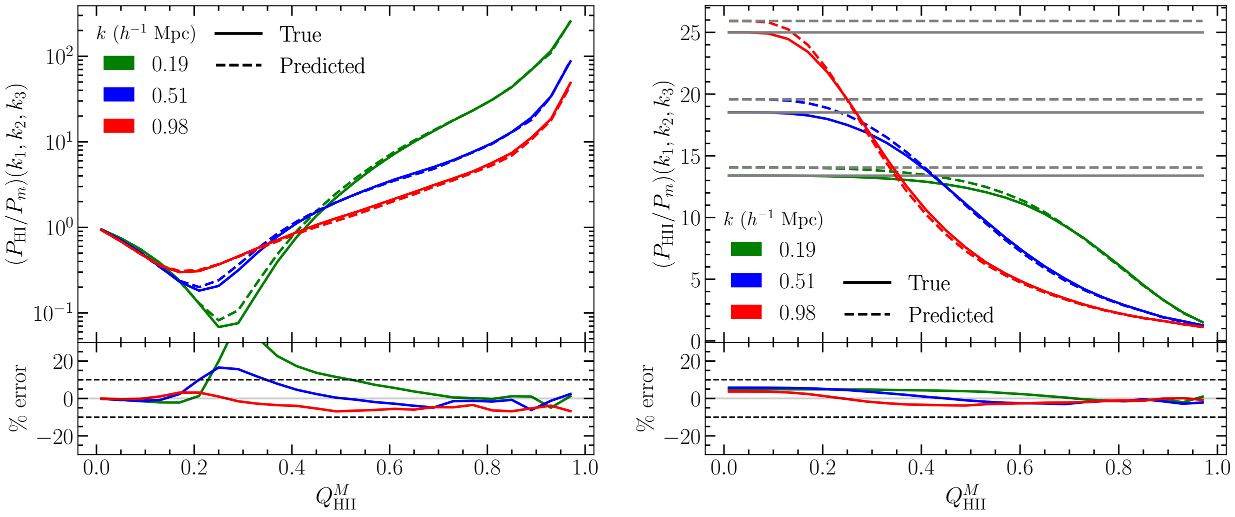


Figure 3.7: HI-HI *(top left)*, HII-HII *(top right)*, HI-$(1+\delta)$ *(bottom left)*, HII-$(1+\delta)$ *(bottom right)* power spectra for truth and prediction (always stochastic hereafter), for different values of the global ionisation fraction $Q^M_{\mathrm{HII}} = 0.25, 0.5, 0.75$. This demonstrates the validity of our method across different ionised fractions. The relatively large errors in the case of $Q^M_{\mathrm{HII}} = 0.25$ are discussed in chapter 4.



(a)

(b)

Figure 3.8: Comparison between truth and prediction of (a) HI bias, (b) HII bias evaluated for three different low k values as a function of the ionised fraction. In the right panel, the three sets of grey horizontal lines represent the $f_{\rm coll}$ bias for each of the three k values. The HII bias approaches the $f_{\rm coll}$ bias at large scales during sufficiently early stages of reionisation, and this can be used to study the large-scale HI bias error at $Q^M_{\rm HII} = 0.25$ (see chapter 4).

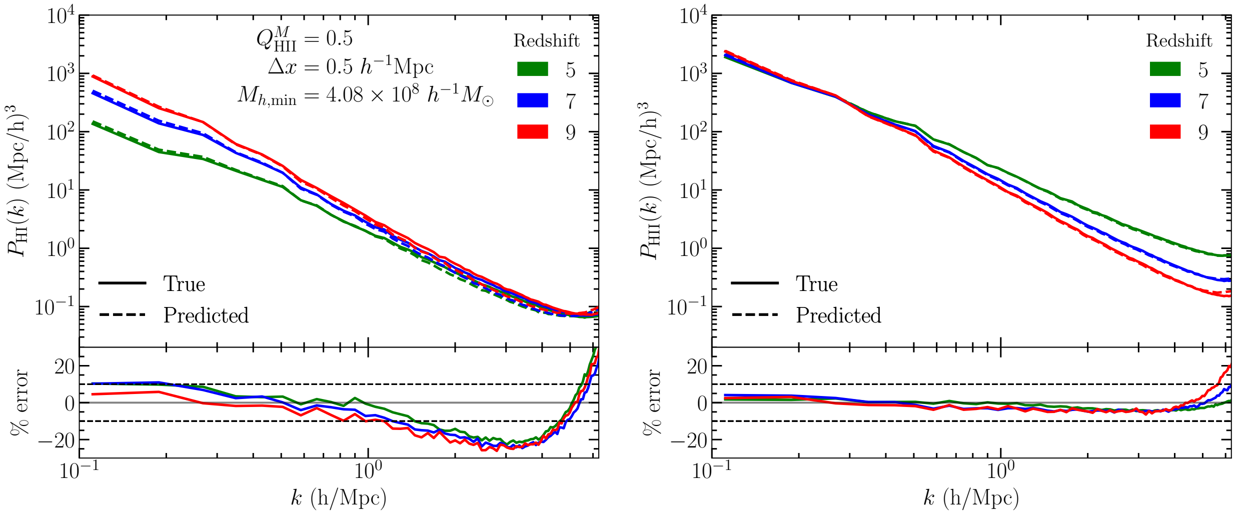


Figure 3.9: HI-HI *(left panel)*, HII-HII *(right panel)* power spectra for truth and prediction, for three different values of redshift at fixed ionisation fraction $Q^M_{\rm HII} = 0.5$. This shows the robustness of our method for variations in redshift, which can be important in using it to study the redshift evolution of reionisation.

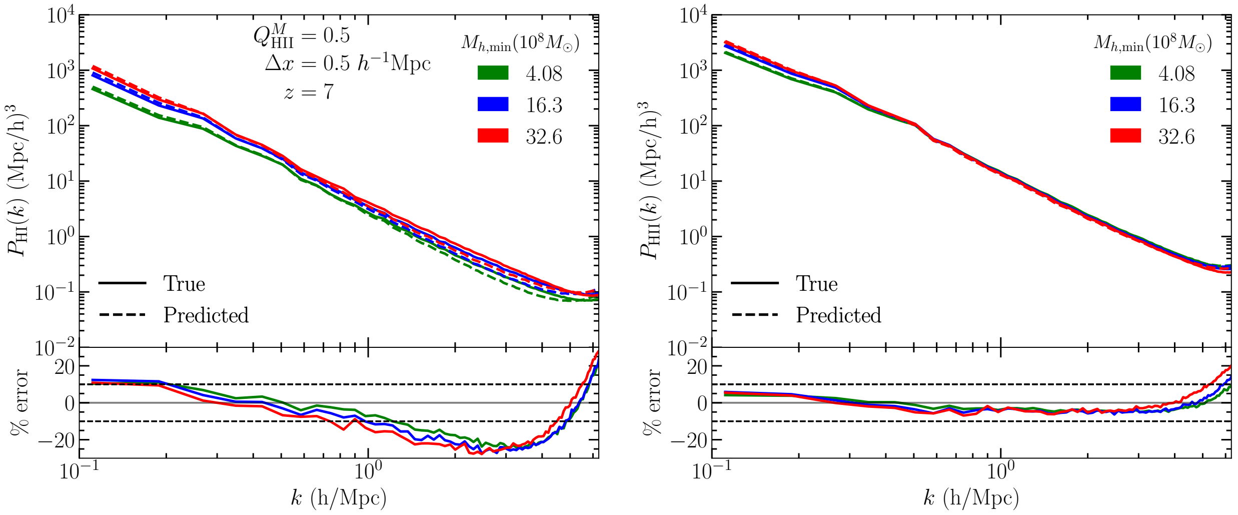


Figure 3.10: HI-HI *(left panel)* and HII-HII *(right panel)* power spectra for truth and prediction, for three different values of minimum halo mass $M_{h,\rm min}$ at fixed ionisation fraction $Q^M_{\rm HII} = 0.5$ and grid size $\Delta x = 0.5\ h^{-1}$Mpc . This shows that our method works with a similar fidelity for a range of $M_{h,\rm min}$, and hence for different choices of the minimum number of particles used for identifying the FoF haloes.

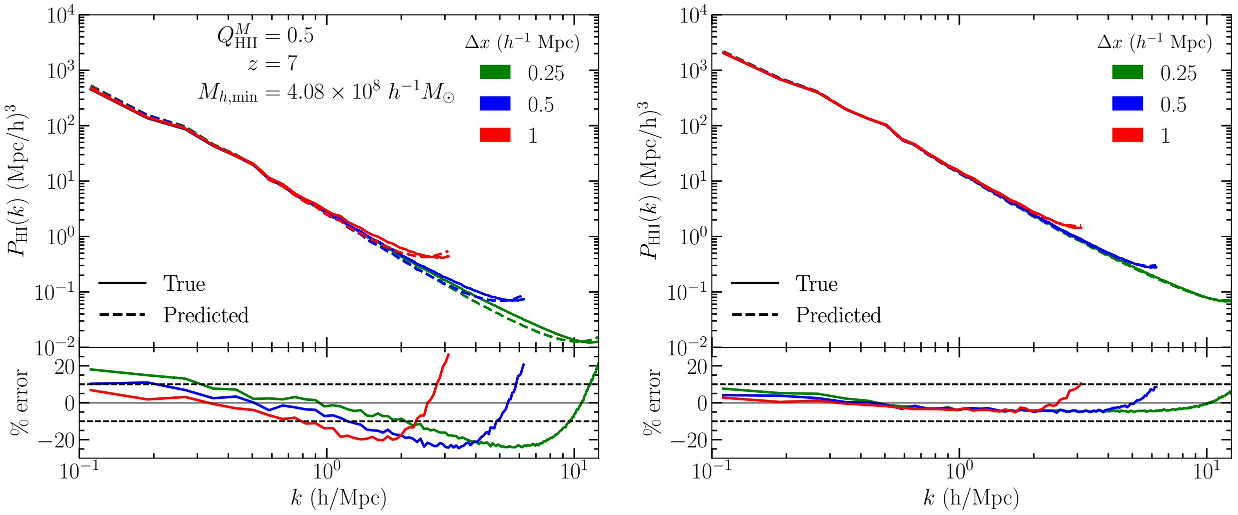


Figure 3.11: HI-HI *(left panel)* and HII-HII *(right panel)* power spectra for truth and prediction, for three different values of grid size Δx used for getting the density and $f_{\rm coll}$ fields, at fixed ionisation fraction $Q^M_{\rm HII} = 0.5$. Each case has been plotted upto its Nyquist frequency. This shows the robustness of our method against the choice of grid size.

Chapter 4

Discussion

Our interpolator draws f_{coll} values for cells taking only their density information into account. This means that the correlation between sampled f_{coll} values across different cells is controlled purely by the correlation between the density values conditioning the CDFs from which these f_{coll} values are sampled. We expect other environmental factors to play a role in the true f_{coll} correlation as well, but these effects are randomised across all cells by our interpolator via picking a uniform random number between 0 and 1 for inverse CDF sampling. A comparison of the recovery of f_{coll} features by our interpolator at different scales, then, is a way of testing the sensitivity of halo formation on the cosmological environment at these scales.

As it turns out, conditioning the f_{coll} CDFs on the density field allows a reasonable recovery of the large-scale structure of the f_{coll} field and consequently, the HI density map. This can be confirmed visually from the full maps in the left part of Figure 3.3 and quantitatively through the power spectra at low k in Figure 3.2 for the f_{coll} field and Figure 3.4 for the HI and HII density fields. Moreover, this large-scale recovery is very similar between the stochastic and deterministic cases, with the latter being marginally better. Therefore, the stochastic variations in the f_{coll} field for a fixed matter density δ, which are precisely due to the effect of other environment variables, do not affect the large-scale distribution of collapse fractions and hence the ionisation bubbles within our tolerance.

However, the f_{coll} value in a particular cell is more strongly influenced by the neighboring density modes than density fluctuations on larger scales (say, over tens of cells). This makes the small-scale distribution of f_{coll} values quite sensitive to the immediate environment and not just the δ value of their parent cell. Consequently, these environmental factors become more important in dictating the small-scale power of the HI density fields. If we focus on the HI density maps in Figure 3.3, the deterministic case completely ignores stochastic fluctuations in the $f_{\rm coll}$ field and ends up producing a relatively smooth HI field outside the ionisation bubbles. When we move to the stochastic case, we transition from this underlying smooth, mean-only $f_{\rm coll}$ field to one that has scatter around the mean $f_{\rm coll}$ incorporated into it. Apart from the correlations introduced due to the δ values, this scatter is uncorrelated cell-wise and leads to an effect similar to adding shot noise over the deterministic $f_{\rm coll}$ and hence the HI map. This explains the increase in small-scale HI power in the stochastic prediction above the deterministic case ($k \gtrsim 2$ in Figure 3.4a). This can also be seen, although to a lesser extent, in the $f_{\rm coll}^{M}$ auto power spectra at the highest k in Figure 3.2a.

The true case accounts for the effect of stochasticity in the correct way, increasing fluctuations in the field but doing so in a way consistent with the full information contained in the environment. One can view the $f_{\rm coll}$ value at a cell as having been sampled from a Dirac delta distribution of $f_{\rm coll}$ conditioned on all the cosmological environment variables that it depends on in principle, denoted by $(\delta, \alpha_1, \alpha_2, \dots)$, where δ is the dark matter overdensity as usual. All of these variables have some particular value at the cell, which dictates the $f_{\rm coll}$ value. In our stochastic sampling, we are only bothering about the δ value and then uniformly sampling from the distribution conditioned only on δ. This essentially amounts to assigning a 'wrong' $f_{\rm coll}$ that is actually associated with the variables $(\delta, \alpha'_1, \alpha'_2, \dots)$, which are found in some other cell. In this sense, our sampling *redistributes* the $f_{\rm coll}$ values from their true spatial distribution, and does so in a randomised manner, washing over structure and its correct spatial correlations at small-scales. The misplaced $f_{\rm coll}$ values sampled this way that are high enough to cross the excursion-set barrier cause the corresponding cells to get flagged as ionised. This leads to the same effect in the HI density field (see the relatively more scattered and uncorrelated tiny bubbles in the middle panel of Figure 3.3) and thus decreases the small-scale power in the stochastic prediction as compared to the truth (Figure 3.4a). On the other hand, ignoring stochasticity turns out to be detrimental to the small-scale HI power of the deterministic case, leading to large errors. While our middle ground is far from the truth, it is still better at recovering the small-scale HI power than the deterministic case.

It is then also interesting to note the behaviour of the ionised field. Not only do the stochastic and deterministic cases recover the HII power spectra almost equally well (right panel of Figure 3.4), the errors are significantly lesser as compared to the HI power spectra (compare left and right panels of Figure 3.7). This can be understood if we look at the HII density map in Figure 4.1. The spurious tiny ionised bubbles are present here as well, but the difference is that the dominant contribution to power at all scales comes from the much stronger density field fluctuations present inside the ionised regions (note that these regions trace the density field since the ionised fraction $x_{\mathrm{HII}}(\mathbf{x})$ there is identically 1). In the case of the HI density field, these regions were masked out and the power spectrum contained complementary information regarding the distribution of less prominent ionised bubbles. These tiny, spurious ionised bubbles are random fluctuations that contribute in tandem to decreasing the power at small scales, but average out when large scales are considered, thereby not contributing much to the large-scale power.

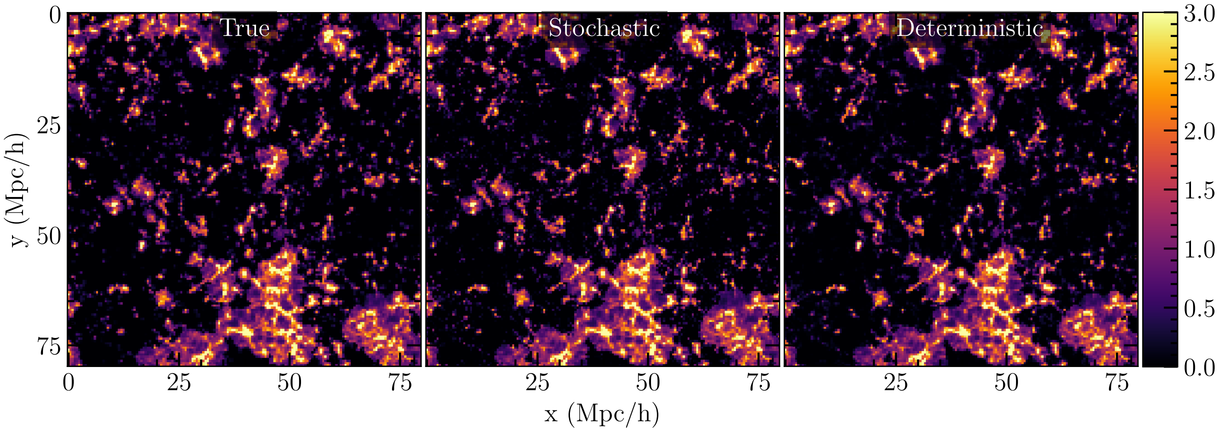


Figure 4.1: The ionised HII density field at $Q^M_{\mathrm{HII}} = 0.5$ in the ground truth *(left)*, as recovered by the stochastic *(middle)* and deterministic *(right)* case at a slice through $z = 50\ h^{-1}$Mpc . The black regions contain neutral hydrogen. The low density regions are masked out and their contribution to the power spectra is subdued by the ionised bubbles that trace the high-density regions of dark matter.

This is apparent from Figure 3.2a, where the large-scale power of the f^M_{coll} field has a constant offset at around 5%. If we plot the un-normalised predicted f^M_{coll} auto power spectrum (that is, without dividing by $\bar{g}^2$ in equation 3.1) then it matches the truth to within 1%. Therefore, the observed $\sim 5\%$ at large scales is mostly due to the error in the global f^M_{coll} mean (squared) made by the interpolator. Since we do not accurately take into account the effect of the environment, the sampled f_{coll} values in nearby cells are incorrectly correlated with each other. This 'mistake' in the f_{coll} sampling, combined with the minor errors in the CDF emulation, implies that the global mass-averaged means of f_{coll} are not constrained to match between truth and prediction, which subsequently leads to the large-scale offset in the properly normalised f_{coll}^M auto power. The effects of an incorrect global f_{coll}^M mean are discussed further in Appendix C.

The relatively large deviation arising in the HI power at the large scales in Figure 3.7 can be understood in the following manner. For the HI (HII) field, we are actually plotting the power spectrum of $\Delta_{\mathrm{HI}}(\mathbf{x})$ ($\Delta_{\mathrm{HII}}(\mathbf{x})$) given by

$$\Delta_{\mathrm{HI}}(\mathbf{x}) = \frac{x_{\mathrm{HI}}^M(\mathbf{x})}{1 - Q_{\mathrm{HII}}^M}\,; \qquad \Delta_{\mathrm{HII}}(\mathbf{x}) = \frac{x_{\mathrm{HII}}^M(\mathbf{x})}{Q_{\mathrm{HII}}^M}\,, \tag{4.1}$$

and these can be related in the following manner:

$$\begin{align} \Delta_{\mathrm{HI}}(\mathbf{x}) &= \frac{x_{\mathrm{HI}}^M(\mathbf{x})}{1 - Q_{\mathrm{HII}}^M} \tag{4.2}\\ &= \frac{x_{\mathrm{HI}}(\mathbf{x})(1+\delta(\mathbf{x}))}{1 - Q_{\mathrm{HII}}^M} \tag{4.3}\\ &= \frac{1+\delta(\mathbf{x}) - x_{\mathrm{HII}}^M(\mathbf{x})}{1 - Q_{\mathrm{HII}}^M} \tag{4.4}\\ &= \frac{1+\delta(\mathbf{x}) - Q_{\mathrm{HII}}^M \Delta_{\mathrm{HII}}(\mathbf{x})}{1 - Q_{\mathrm{HII}}^M}\,. \tag{4.5} \end{align}$$

Following the discussion in Appendix B of [40], if we assume the bias to be scale-free at large scales during the early stages of reionisation, we can relate the HI and HII bias as

$$b_{\mathrm{HI}} = \frac{1 - Q_{\mathrm{HII}}^M b_{\mathrm{HII}}}{1 - Q_{\mathrm{HII}}^M}\,. \tag{4.6}$$

Recall that the square of the bias is simply the power spectrum of the relevant field normalised by the matter power spectrum (equation 3.5), and since the matter power spectrum at large scales is identical between LB and RB, the power spectra error is directly proportional to the bias error. We can write this at fixed Q_{HII}^M as

$$\frac{(b_{\mathrm{HI}})_{\mathrm{predicted}}}{(b_{\mathrm{HI}})_{\mathrm{true}}} = \frac{1 - Q_{\mathrm{HII}}^M (b_{\mathrm{HII}})_{\mathrm{predicted}}}{1 - Q_{\mathrm{HII}}^M (b_{\mathrm{HII}})_{\mathrm{true}}}\,. \tag{4.7}$$

From Figure 3.8b, we can read off the value of $(b_{\mathrm{HII}}^2)_{\mathrm{true}}$ for the smallest k to be around 13.5. This implies $(b_{\mathrm{HII}})_{\mathrm{true}} \approx 3.7$ and so the denominator in the relative error expression above will blow up around $Q_{\mathrm{HII}}^M \approx 1/3.7 \approx 0.27$. Thus, the value of $Q_{\mathrm{HII}}^M = 0.25$ for which we plot the power spectra in Figure 3.7 is also expected to show a large error. The same calculation is confirmed from Figure 3.8a as well, where both the true and predicted HI biases become numerically very small, causing the errors to blow up.

Chapter 5

Conclusion

The advent of more advanced radio interferometer experiments such as the SKA will provide more precise bounds on the 21 cm power spectra, and hence the HI density distribution from the Epoch of Reionisation (EoR). This makes the forward modelling of HI maps during EoR crucial for testing our understanding of the epoch. Efficient methods to do this require the distribution of the fraction of mass in dark matter haloes (collapse fraction field) to be input into excursion-set based semi-numerical models of reionisation [35–40]. Obtaining the collapse fraction field using the semi-analytical formalism of the conditional Press-Schechter [33, 46] and conditional Sheth-Tormen [47, 48] mass functions, while efficient, is an approximation to more accurate results obtained from high-dynamic range N-body simulations [49–53]. The latter are extremely inefficient for parameter estimation due to their high computational cost.

While there have been attempts to make the prediction of the collapse fraction field more efficient by using hybrid approaches that combine information from low-dynamic range boxes [31, 54–57], they have not taken into account the full stochasticity in f_{coll} for a fixed dark matter density contrast δ, as predicted by N-body simulations. In this work, we build a machine learning model to accurately predict $f_{\mathrm{coll}}(\mathbf{x})$ using a hybrid approach while taking into account the full stochasticity. We use the conditional cumulative distribution functions $\mathrm{CDF}(f_{\mathrm{coll}}|\delta)$ obtained from a set of 7 small-volume, high-resolution simulations (SB) to train the ML model using a methodology based on Gaussian Process Regression (GPR). The density input from a large-volume, low-resolution simulation (LB) is then used to randomly draw samples of $f_{\rm coll}$ values from the emulated CDFs for each cell. This constitutes our *stochastic* case, and we also obtain $f_{\rm coll}(\mathbf{x})$ corresponding to the *deterministic* case, which excludes stochasticity by simply using the conditional means $\langle f_{\rm coll}|\delta\rangle$ computed from the SB.

Upon comparing the auto power spectra of the mass-averaged $f_{\rm coll}(\mathbf{x})$ and its cross with $\Delta \equiv 1 + \delta$ for our fiducial choice of the parameters $z, \Delta x, M_{h,\rm min}$, we find similar levels of agreement between the stochastic and deterministic cases (Figure 3.2). We then compute the HI and HII density fields using the semi-numerical code for reionisation SCRIPT. While the recovery is similar at large scales, the deterministic case performs much worse at smaller scales for the HI density field (Figure 3.4). We then increase the grid size to $\Delta x = 1\ h^{-1}$Mpc to enable a more complete comparison between the simulation-based deterministic and stochastic methods and the semi-analytical conditional mass functions. For the mass-weighted $f_{\rm coll}$, HI and HII power spectra, the simulation-based methods work better and the stochastic case is the best at recovering the small-scale HI power (Figures 3.5 and 3.6). We further test the flexibility of the stochastic case against variations in all the involved parameters, including global ionised fraction, redshift, grid size and minimum halo mass. For almost all the cases, we are able to recover the HI large-scale power ($k \lesssim 1\ h\,{\rm Mpc}^{-1}$) at the $\lesssim 10\%$ level, whereas for the HII density field the errors are well within 10% for the entire range of k values. The accuracy, combined with its significantly lesser RAM requirements of ~ 20 GB for running the SB and LB as compared to ~ 160 GB for running the RB, makes our method a powerful tool for RAM-limited users conducting studies of reionization parameter space exploration who wish to run a single cosmological high-dynamic range simulation.

Using only the dark matter density contrast to condition the distribution of $f_{\rm coll}$, we are able to recover large-scale structures well in the $f_{\rm coll}$ field and the subsequent HI maps. We demonstrate how stochasticity in the $f_{\rm coll}$ predictions can play a critical role in recovering the small-scale structure of the HI maps. However, our specific implementation of stochasticity does not take into account the full information contained in the cosmological environment, and this leads to some spurious small-scale structures in the HI maps. Therefore, further improvements to the ML framework can include finding a set of variables, that can better reflect the environment than δ alone, to condition the distribution of $f_{\rm coll}$. As suggested by [60], the three eigenvalues of the tidal tensor evaluated at each location $\{\lambda_1(\mathbf{x}), \lambda_2(\mathbf{x}), \lambda_3(\mathbf{x})\}$ could be used for such a purpose, and this shall be explored in future work. The GPR machinery set up in this work will become more beneficial in this regard than a simple linear interpolation scheme, due to the high dynamic range of the 3 eigenvalues.

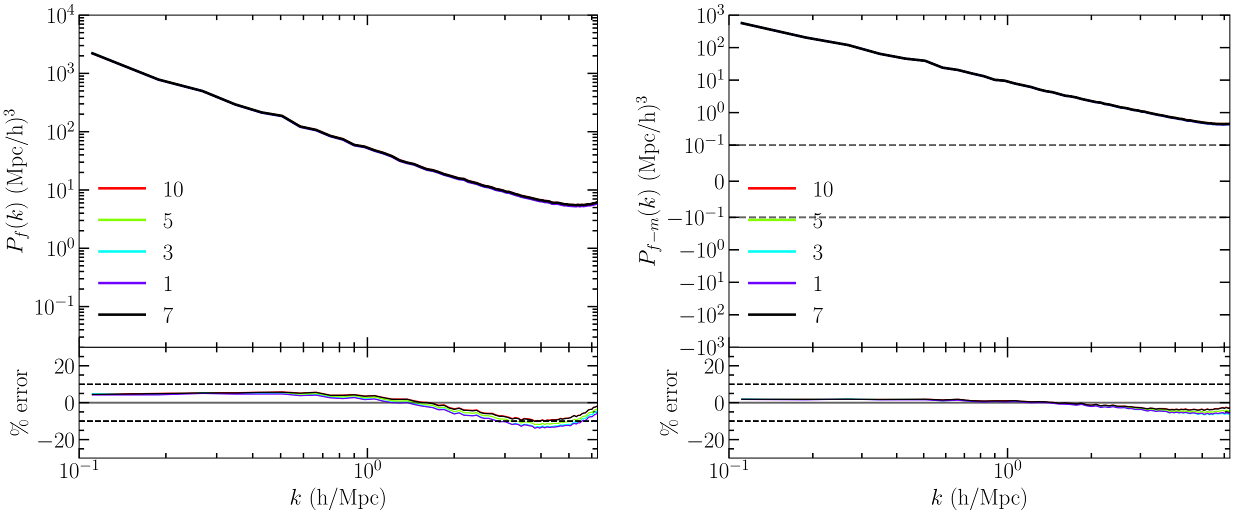
Another possible direction for the future entails increasing the dynamic range gap between SB/LB and RB. Currently, we are using SB simulations that are 8 times smaller in volume than the target RB. We can test the accuracy of the framework for a simulation that is 64 times smaller. One can also explore using our ML model to build a redshift evolution of reionisation by sampling $f_{\rm coll}(\mathbf{x})$ at appropriately spaced redshifts. In conclusion, the method presented in this work can prove to be an efficient yet accurate way to study models of reionisation and also help constrain parameters from upcoming observations.

Appendix A

Convergence of Results

We chose to combine 7 different realisations of SB boxes to get the $(\delta, f_{\mathrm{coll}})$ pairs, from which the training CDFs were constructed. Now, we vary this number to 1, 3, 5 and 10 and observe the effect on the results. The training converges successfully to `cv_thresh` $= 0.015$ for each of these variations and the predicted f^{M}_{coll} auto and cross power spectra are shown in Figure A.1. In the lower panels, we show the error between the power spectra of each case with the truth obtained from RB, but the true power spectra itself is not shown in the upper panels. These results have been obtained for the fiducial $z = 7$ setting. While the variation is not much at large scales, one can clearly notice a trend at the smallest scales, with the error curves of 7 and 10 realisations combined being almost identical. However, the SCRIPT results are quite robust to these differences, and are similarly presented in Figure A.2.

This validates our choice of using 7 realisations to make the training, since any further increase in the number of realisations does not improve the results while any decrease causes the results to change, albeit only for the f_{coll} power.



(a) (b)

Figure A.1: Comparison of (a) $f_{\rm coll}^M$-$f_{\rm coll}^M$ auto and (b) $f_{\rm coll}^M$-Δ cross power spectra (upper panels) for different numbers of SB boxes combined for training, and the relative error of each with the true power spectra (lower panel). The default case that we work with is 7, shown in black. The differences between the errors are very small and mostly visible at small scales. By number of boxes equal to 7, a clear trend of convergence emerges.

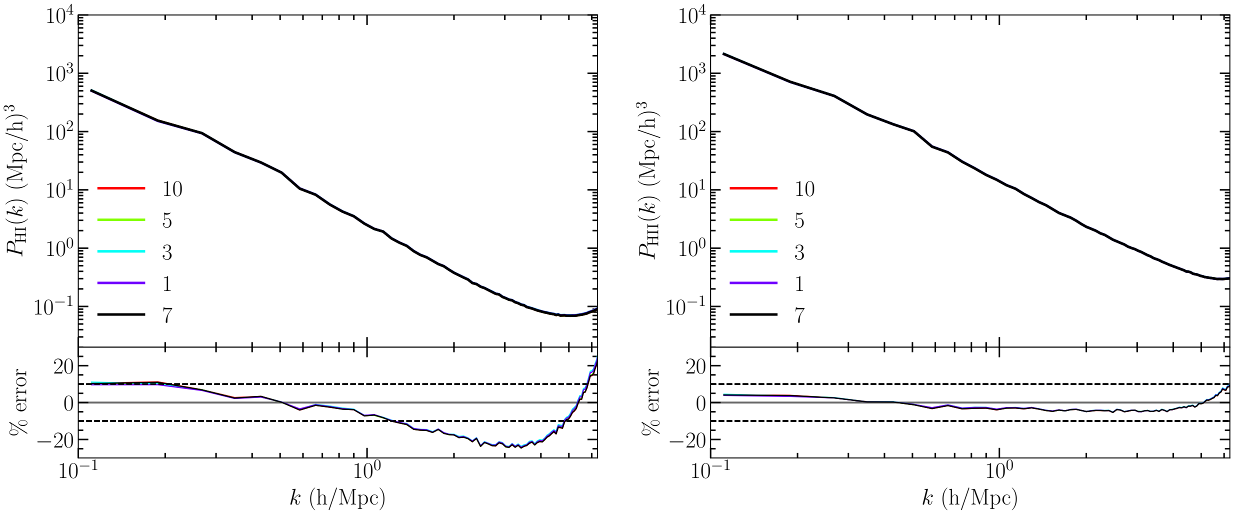


Figure A.2: HI-HI (left panel) and HII-HII (right panel) power spectra for different numbers of SB boxes combined for training, and the relative error of each with the true power spectra (lower panel), at fixed ionisation fraction $Q_{\rm HII}^M = 0.5$. The default case that we work with is 7, shown in black.

Appendix B

Optimisation of Binning

As described in 2.2.1, we adopt a variable binning scheme to construct the training data, defined over $\log(1+\delta)$. For the fiducial case, the bin widths for the first and last bins are 0.06 and 0.2 dex, respectively and a turning point occurs at $\delta = 0$, where it is the minimum at 0.03 dex. We first optimise the bin widths for the $z = 9$ case, where in order to find the last bin width (high-density end), we start with a very small value which we apply to uniformly bin the whole $\log(1+\delta)$ range. We then train the GPR on the empirical CDFs constructed using such a binning. Then, we split the last bin into 3 equal sub-bins and evaluate the empirical CDFs for each sub-bin. These 3 CDFs are then compared with the 3 GPR-predicted CDFs at the central δ of these sub-bins. This procedure is repeated for successively larger parent bin widths. The idea is that at very small bin widths, GPR training at the last bin would suffer from noise and at very large bin widths the training CDFs would systematically deviate from the true underlying CDFs. In either case, the true empirical CDF at the 3 sub-bin centres would substantially differ from the GPR predictions. The parent bin-width that achieves visually the closest match is chosen as the optimal one, and it turns out to be around 0.12 dex for $z = 9$. It must be noted that this was not a very strict choice, and slight deviations from this value do not appreciably change the results. The same procedure was carried out to find the optimal width of the first bin (low-density end) as 0.06 dex and the reference value of 0.03 dex at $\delta = 0$. Given a δ binning, we find that 500 bins in $f_{\rm coll}$ for making the training CDFs are usually enough to achieve accurate results for the joint distribution of δ and GPR predicted $f_{\rm coll}$. Still, we try another case with 900 bins just for comparison and choose the one with the better accuracy in recovering the HI power spectra, which for the

$z = 9$ case is the 900 bins case. Going beyond 900 does not improve the results.

Thus, once the $z = 9$ binning is decided, we use it as a guideline to find the optimal bin-widths of the other cases. For $z = 7$, we linearly scale the last bin width with $\log(1 + \delta_{\mathrm{max}})$, where δ_{max} is the highest density found in the SB at that redshift. This gives us a starting guess which we vary a few times along either direction. While δ_{max} changes substantially across redshifts, the minimum density values δ_{min} are very close to each other and hence we just try out a few variations along either direction of the $z = 9$ first bin width (0.06 dex). These combinations are tested for both 500 and 900 bins in f_{coll} for making the CDFs, and the case that gives the least error in the HI and HII power spectra is considered as the optimal choice.

If we just apply the binning scheme of our fiducial case on all the variations, the results worsen primarily for the redshift and the grid size variations. These are shown in Figures B.1 and B.2. Evidently, in Figure B.1, the $z = 5$ case worsens significantly when compared to its best interpolator, shown in Figure 3.9. The $z = 9$ case in HI and the HII results shows a less significant degradation. For the gridsize variations, the $\Delta x = 1\ h^{-1}$Mpc case shows some noticeable degradation, which is less prominent for all other cases. In general, we see that most of the results are not extremely sensitive to the binning scheme. The minimum halo mass variation cases have identical δ values as the fiducial case, and so applying the fiducial binning scheme over them does not result in any significant degradation, and are thus not shown here.

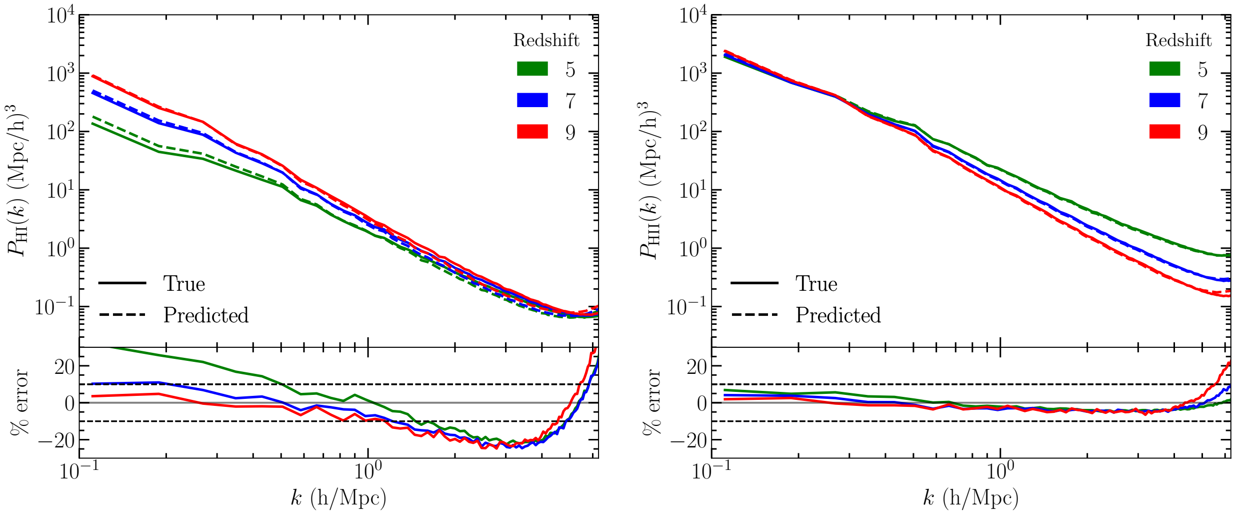


Figure B.1: HI-HI (left panel), HII-HII (right panel) power spectra for truth and prediction using the interpolator made from the fiducial binning scheme. Three different values of redshift are shown at fixed ionisation fraction $Q^M_{\mathrm{HII}} = 0.5$. The $z = 5$ case shows a significant deviation from the case when its binning is separately optimised (compare with Figure 3.9), highlighting the importance of our optimisation procedure.

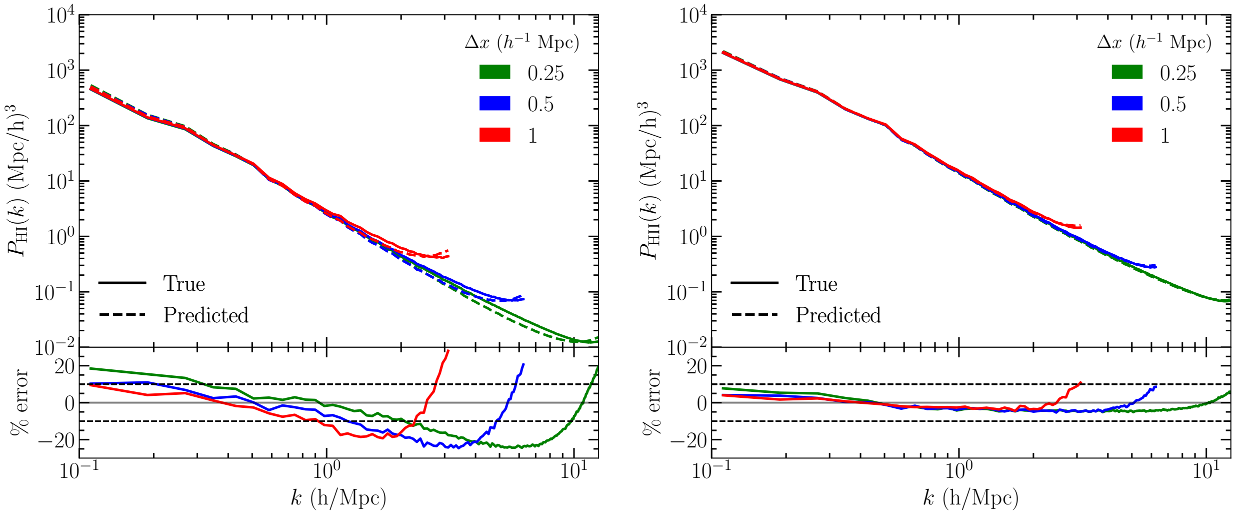


Figure B.2: HI-HI (left panel) and HII-HII (right panel) power spectra for truth and prediction using the interpolator made from the fiducial binning scheme. Three different values of grid size Δx are shown at fixed ionisation fraction $Q^M_{\mathrm{HII}} = 0.5$. Each case has been plotted upto its Nyquist frequency. The $\Delta x = 0.25\ h^{-1}$Mpc case shows the most noticeable worsening of accuracy when compared with its optimal binning scheme (Figure 3.11).

Appendix C

Normalisation Error in f_{coll}

In section 3.1, we saw that the large-scale power of the normalised f_{coll}^{M} field had a constant offset at around 5% (Figure 3.2), and mentioned that this can be traced back to the error in the global mean of f_{coll}^{M} . Instead of using equation 3.1, we can define an unnormalised auto power spectrum (denoted by $\tilde{P}_g(k)$) for a field $g(\mathbf{x})$ as

$$\langle g(\mathbf{k}) g^*(\mathbf{k}') \rangle = (2\pi)^3 \tilde{P}_g(k) \delta_D(\mathbf{k} - \mathbf{k}')\,, \tag{C.1}$$

where the usual notations from equation 3.1 apply. The difference is that we are no longer normalizing g by its mean in position space while computing the Fourier conjugates. Comparing this $\tilde{P}(k)$ for the stochastic and deterministic f_{coll}^{M} in Figure C.1, we see that both the predictions now show a very small error in the large-scale power. This shows that the 5% offset at low k in Figure 3.2, at least for the stochastic case, is predominantly due to the error in recovering the global mean of f_{coll}^{M} .

In our sampling procedure, we are only accounting for the effect of dark matter density on the f_{coll} values, and the effect of other cosmological environmental variables is randomised. This leads to nearby f_{coll} values being incorrectly correlated with each other. This ‘mistake’ in the f_{coll} sampling, combined with the minor errors in the CDF emulation, implies that the global mass-averaged means of f_{coll} are not constrained to match between truth and prediction.

These uncorrelated f_{coll} values show up in small-scale patches of the f_{coll} field, such as the slices compared in Figure C.2. Although it is more apparent in the HI maps, even here one can notice similar features of incoherently high $f_{\rm coll}$ fluctuations in the stochastic case and the lack thereof in the deterministic one, both reflecting the inaccuracy in modelling the small-scale environment. These inaccuracies in the $f_{\rm coll}$ *topology* are a feature of small scales, and we have checked that if we smooth out to larger scales the topology is more or less well-recovered, but the overall mean and thus normalisation is affected due to the small-scale errors.

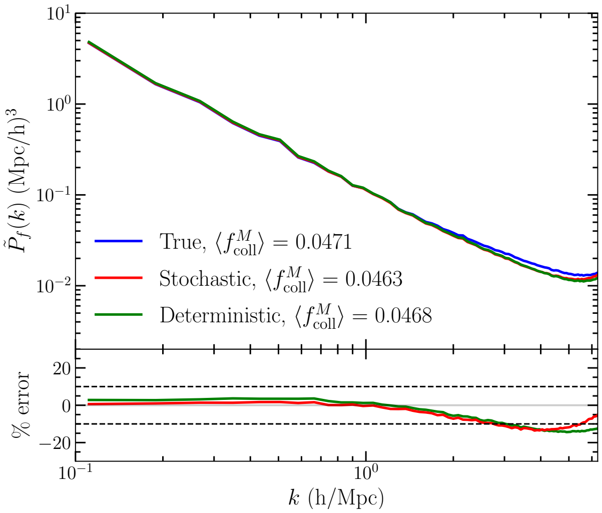


Figure C.1: Comparison of $f^M_{\rm coll}$ *unnormalised* auto power spectra (defined as per equation C.1), between truth and predictions using the deterministic and stochastic cases. The large-scale power matches quite well, implying that the error in Figure 3.2 at large scales is mostly due to error in the mean.

We can also observe the effect of normalisation on the power spectra of the conditional ST and PS predictions (Figure 3.5), by plotting their unnormalised $\tilde{P}(k)$ in Figure C.3. As can be noted from the values mentioned in the plot, the conditional ST (conditional PS) overestimates (underestimates) the global $f^M_{\rm coll}$ mean, with the difference being much greater than the stochastic and deterministic cases. Given the relation between the relative error in the normalised (P) and unnormalised ($\tilde{P}$) power spectra,

$$\frac{\tilde{P}_{\rm sampled}}{\tilde{P}_{\rm truth}} = \frac{P_{\rm sampled}}{P_{\rm truth}} \frac{\langle f^M_{\rm coll,\ sampled}\rangle^2}{\langle f^M_{\rm coll,\ true}\rangle^2}\,, \tag{C.2}$$

the unnormalised large-scale power has a larger error in the case of conditional ST as compared to PS. When we move to the HI density field calculation, the ζ value required in order to achieve a global ionised fraction of $Q^M_{\rm HII} = 0.5$ is then substantially different for the con-

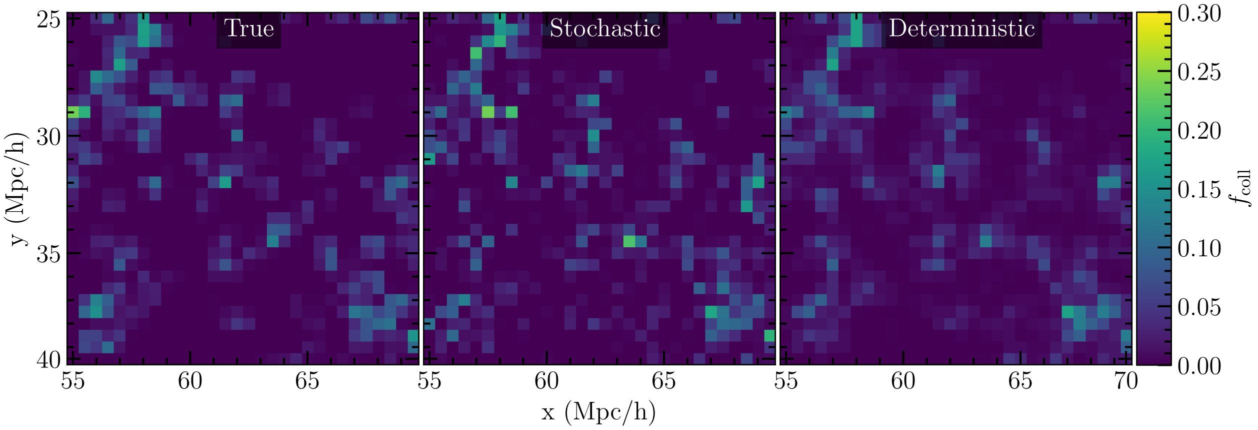


Figure C.2: Slices of the $f_{\rm coll}$ field for the truth, stochastic and deterministic cases for the same zoomed-in region as shown for the HI map in Figure 3.3. The error in small-scale $f_{\rm coll}$ correlations due to not accounting for the full cosmological environment information can be seen in the middle and right panels.

ditional PS and ST cases, which then ends up making a large error in their ionised bubble topology even at large-scales, hence causing the large $\sim 25\%$ error in large-scale power (left panel of Figure 3.6).

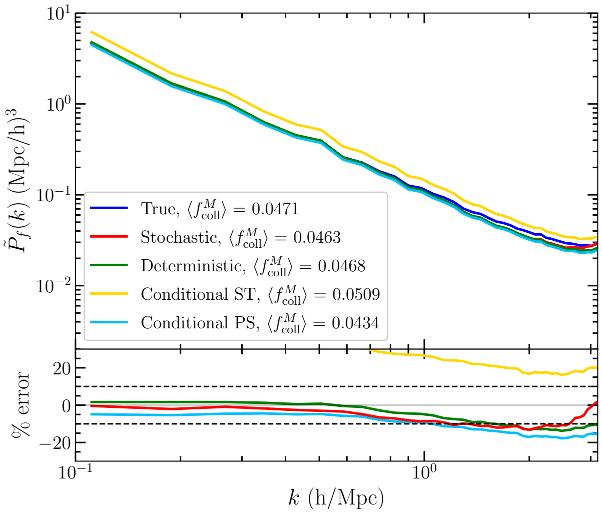


Figure C.3: Comparison of $f_{\rm coll}^M$ *unnormalised* auto power spectra (defined as per equation C.1), between truth, stochastic, deterministic and the semi-analytical predictions. Compared with the normalised case (Figure 3.5), the errors of conditional PS and ST are now very different due their substantially different means.

References

[1] C. Cain and A. D'Aloisio, *FlexRT — a fast and flexible cosmological radiative transfer code for reionization studies. part i. code validation*, *Journal of Cosmology and Astroparticle Physics* **2024** (2024) 025.

[2] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press (2006).

[3] N.Y. Gnedin and P. Madau, *Modeling cosmic reionization*, *Living Reviews in Computational Astrophysics* **8** (2022) 3.

[4] T.R. Choudhury, *A short introduction to reionization physics*, *General Relativity and Gravitation* **54** (2022) .

[5] G. Kulkarni, L.C. Keating, M.G. Haehnelt, S.E.I. Bosman, E. Puchwein, J. Chardin et al., *Large ly opacity fluctuations and low cmb in models of late reionization with large islands of neutral hydrogen extending to z &lt; 5.5*, *Monthly Notices of the Royal Astronomical Society: Letters* **485** (2019) L24 [https://academic.oup.com/mnrasl/article-pdf/485/1/L24/56977861/mnrasl_485_1_l24.pdf]

[6] F. Nasir and A. D'Aloisio, *Observing the tail of reionization: neutral islands in the z = 5.5 lyman- forest*, *Monthly Notices of the Royal Astronomical Society* **494** (2020) 3080 [https://academic.oup.com/mnras/article-pdf/494/3/3080/33128596/staa894.pdf].

[7] T.R. Choudhury, A. Paranjape and S.E.I. Bosman, *Studying the lyman optical depth fluctuations at z 5.5 using fast semi-numerical methods*, *Monthly Notices of the Royal Astronomical Society* **501** (2021) 5782 [https://academic.oup.com/mnras/article-pdf/501/4/5782/58691226/stab045.pdf].

[8] J.S. Bolton, G.D. Becker, J.S.B. Wyithe, M.G. Haehnelt and W.L.W. Sargent, *A first direct measurement of the intergalactic medium temperature around a quasar at z= 6*, *Monthly Notices of the Royal Astronomical Society* **406** (2010) 612 [https://academic.oup.com/mnras/article-pdf/406/1/612/3751819/mnras0406-0612.pdf].

[9] S. Raskutti, J.S. Bolton, J.S.B. Wyithe and G.D. Becker, *Thermal constraints on the reionization of hydrogen by population ii stellar sources*, *Monthly Notices of the Royal Astronomical Society* **421** (2012) 1969 [https://academic.oup.com/mnras/article-pdf/421/3/1969/5835775/mnras0421-1969.pdf].

[10] B. Maity and T.R. Choudhury, *Probing the thermal history during reionization using a seminumerical photon-conserving code script*, *Monthly Notices of the Royal Astronomical Society* **511** (2022) 2239 [https://academic.oup.com/mnras/article-pdf/511/2/2239/42500298/stac182.pdf].

[11] M. Ouchi, Y. Ono and T. Shibuya, *Observations of the lyman- universe*, *Annual Review of Astronomy and Astrophysics* **58** (2020) 617.

[12] T.R. Choudhury, S. Mukherjee and S. Paul, *Cosmic microwave background constraints on a physical model of reionization*, *Monthly Notices of the Royal Astronomical Society: Letters* **501** (2020) L7 [https://academic.oup.com/mnrasl/article-pdf/501/1/L7/54638482/slaa185.pdf].

[13] Gorce, Adélie, Douspis, Marian and Salvati, Laura, *Retrieving cosmological information from small-scale cmb foregrounds - ii. the kinetic sunyaev zel'dovich effect*, *AA* **662** (2022) A122.

[14] S.R. Furlanetto, S. Peng Oh and F.H. Briggs, *Cosmology at low frequencies: The 21cm transition and the high-redshift universe*, *Physics Reports* **433** (2006) 181.

[15] J.R. Pritchard and A. Loeb, *21 cm cosmology in the 21st century*, *Reports on Progress in Physics* **75** (2012) 086901.

[16] A. Mesinger, ed., *The Cosmic 21-cm Revolution*, 2514-3433, IOP Publishing (2019), 10.1088/2514-3433/ab4a73.

[17] J.D. Bowman, A.E.E. Rogers, R.A. Monsalve, T.J. Mozdzen and N. Mahesh, *An absorption profile centred at 78 megahertz in the sky-averaged spectrum*, *Nature* **555** (2018) 67–70.

[18] S. Singh, R. Subrahmanyan, N.U. Shankar, M.S. Rao, A. Fialkov, A. Cohen et al., *First results on the epoch of reionization from first light with saras 2*, *The Astrophysical Journal Letters* **845** (2017) L12.

[19] N. Patra, R. Subrahmanyan, A. Raghunathan and N. Udaya Shankar, *Saras: a precision system for measurement of the cosmic radio background and signatures from the epoch of reionization*, *Experimental Astronomy* **36** (2013) 319–370.

[20] L.J. Greenhill and G. Bernardi, *Hi epoch of reionization arrays*, 2012.

[21] N.Y. Gnedin, *Cosmological reionization by stellar sources*, *The Astrophysical Journal* **535** (2000) 530.

[22] B. Ciardi, A. Ferrara and S.D.M. White, *Early reionization by the first galaxies*, *Monthly Notices of the Royal Astronomical Society* **344** (2003) L7 [https://academic.oup.com/mnras/article-pdf/344/1/L7/18652206/344-1-L7.pdf].

[23] I.T. Iliev, G. Mellema, U.-L. Pen, H. Merz, P.R. Shapiro and M.A. Alvarez, *Simulating cosmic reionization at large scales – i. the geometry of reionization*, *Monthly Notices of the Royal Astronomical Society* **369** (2006) 1625 [https://academic.oup.com/mnras/article-pdf/369/4/1625/3799304/mnras0369-1625.pdf].

[24] H. Trac and R. Cen, *Radiative transfer simulations of cosmic reionization. i. methodology and initial results*, *The Astrophysical Journal* **671** (2007) 1.

[25] M. Petkova and V. Springel, *An implementation of radiative transfer in the cosmological simulation code gadget*, *Monthly Notices of the Royal Astronomical Society* **396** (2009) 1383 [https://academic.oup.com/mnras/article-pdf/396/3/1383/5796731/mnras0396-1383.pdf].

[26] N.Y. Gnedin, *Cosmic reionization on computers. i. design and calibration of simulations*, *The Astrophysical Journal* **793** (2014) 29.

[27] A.H. Pawlik, A. Rahmati, J. Schaye, M. Jeon and C. Dalla Vecchia, *The aurora radiation-hydrodynamical simulations of reionization: calibration and first results*, *Monthly Notices of the Royal Astronomical Society* **466** (2016) 960 [https://academic.oup.com/mnras/article-pdf/466/1/960/10864970/stw2869.pdf].

[28] J. Rosdahl, H. Katz, J. Blaizot, T. Kimm, L. Michel-Dansac, T. Garel et al., *The sphinx cosmological simulations of the first billion years: the impact of binary stars on reionization*, *Monthly Notices of the Royal Astronomical Society* **479** (2018) 994 [https://academic.oup.com/mnras/article-pdf/479/1/994/25129300/sty1655.pdf].

[29] R. Kannan, E. Garaldi, A. Smith, R. Pakmor, V. Springel, M. Vogelsberger et al., *Introducing the thesan project: radiation-magnetohydrodynamic simulations of the epoch of reionization*, *Monthly Notices of the Royal Astronomical Society* **511** (2021) 4005 [https://academic.oup.com/mnras/article-pdf/511/3/4005/42579816/stab3710.pdf].

[30] J.S.W. Lewis, P. Ocvirk, J.G. Sorce, Y. Dubois, D. Aubert, L. Conaboy et al., *The short ionizing photon mean free path at $z = 6$ in cosmic dawn iii, a new fully coupled radiation-hydrodynamical simulation of the epoch of reionization*, *Monthly Notices of the Royal Astronomical Society* **516** (2022) 3389 [https://academic.oup.com/mnras/article-pdf/516/3/3389/45882789/stac2383.pdf].

[31] I.T. Iliev, G. Mellema, K. Ahn, P.R. Shapiro, Y. Mao and U.-L. Pen, *Simulating cosmic reionization: how large a volume is large enough?*, *Monthly Notices of the Royal Astronomical Society* **439** (2014) 725 [https://academic.oup.com/mnras/article-pdf/439/1/725/5599101/stt2497.pdf].

[32] H.D. Kaur, N. Gillet and A. Mesinger, *Minimum size of 21-cm simulations*, *Monthly Notices of the Royal Astronomical Society* **495** (2020) 2354 [https://academic.oup.com/mnras/article-pdf/495/2/2354/33323159/staa1323.pdf].

[33] J.R. Bond, S. Cole, G. Efstathiou and N. Kaiser, *Excursion Set Mass Functions for Hierarchical Gaussian Fluctuations*, **379** (1991) 440.

[34] S.R. Furlanetto, M. Zaldarriaga and L. Hernquist, *The growth of HII regions during reionization*, *The Astrophysical Journal* **613** (2004) 1.

[35] A. Mesinger and S. Furlanetto, *Efficient simulations of early structure formation and reionization*, *The Astrophysical Journal* **669** (2007) 663.

[36] O. Zahn, A. Lidz, M. McQuinn, S. Dutta, L. Hernquist, M. Zaldarriaga et al., *Simulations and analytic calculations of bubble growth during hydrogen reionization*, *The Astrophysical Journal* **654** (2007) 12.

[37] T.R. Choudhury, M.G. Haehnelt and J. Regan, *Inside-out or outside-in: the topology of reionization in the photon-starved regime suggested by lyα forest data*, Monthly Notices of the Royal Astronomical Society **394** (2009) 960 [https://academic.oup.com/mnras/article-pdf/394/2/960/3710519/mnras0394-0960.pdf].

[38] A. Mesinger, S. Furlanetto and R. Cen, *21cmfast: a fast, seminumerical simulation of the high-redshift 21-cm signal*, Monthly Notices of the Royal Astronomical Society **411** (2011) 955 [https://academic.oup.com/mnras/article-pdf/411/2/955/4099991/mnras0411-0955.pdf].

[39] Y. Lin, S.P. Oh, S.R. Furlanetto and P.M. Sutter, *The distribution of bubble sizes during reionization*, Monthly Notices of the Royal Astronomical Society **461** (2016) 3361 [https://academic.oup.com/mnras/article-pdf/461/3/3361/8112292/stw1542.pdf].

[40] T.R. Choudhury and A. Paranjape, *Photon number conservation and the large-scale 21 cm power spectrum in seminumerical models of reionization*, Monthly Notices of the Royal Astronomical Society **481** (2018) 3821 [https://academic.oup.com/mnras/article-pdf/481/3/3821/25844366/sty2551.pdf].

[41] S. Majumdar, G. Mellema, K.K. Datta, H. Jensen, T.R. Choudhury, S. Bharadwaj et al., *On the use of seminumerical simulations in predicting the 21-cm signal from the epoch of reionization*, Monthly Notices of the Royal Astronomical Society **443** (2014) 2843 [https://academic.oup.com/mnras/article-pdf/443/4/2843/6274877/stu1342.pdf].

[42] O. Zahn, A. Mesinger, M. McQuinn, H. Trac, R. Cen and L.E. Hernquist, *Comparison of reionization models: radiative transfer simulations and approximate, seminumeric models*, Monthly Notices of the Royal Astronomical Society **414** (2011) 727 [https://academic.oup.com/mnras/article-pdf/414/1/727/3838580/mnras0414-0727.pdf].

[43] S.J. Mutch, P.M. Geil, G.B. Poole, P.W. Angel, A.R. Duffy, A. Mesinger et al., *Dark-ages reionization and galaxy formation simulation – iii. modelling galaxy formation and the epoch of reionization*, Monthly Notices of the Royal Astronomical Society **462** (2016) 250 [https://academic.oup.com/mnras/article-pdf/462/1/250/18468891/stw1506.pdf].

[44] H.-S. Kim, J.S.B. Wyithe, S. Raskutti, C.G. Lacey and J.C. Helly, *The structure of reionization in hierarchical galaxy formation models*, Monthly Notices of the Royal

Astronomical Society **428** (2012) 2467 [https://academic.oup.com/mnras/article-pdf/428/3/2467/3698435/sts206.pdf].

[45] M.G. Santos, L. Ferramacho, M.B. Silva, A. Amblard and A. Cooray, *Fast large volume simulations of the 21-cm signal from the reionization and pre-reionization epochs*, *Monthly Notices of the Royal Astronomical Society* **406** (2010) 2421 [https://academic.oup.com/mnras/article-pdf/406/4/2421/3339431/mnras0406-2421.pdf].

[46] W.H. Press and P. Schechter, *Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation*, **187** (1974) 425.

[47] R.K. Sheth and G. Tormen, *Large-scale bias and the peak background split*, **308** (1999) 119 [astro-ph/9901122].

[48] R.K. Sheth and G. Tormen, *An excursion set model of hierarchical clustering: ellipsoidal collapse and the moving barrier*, *Monthly Notices of the Royal Astronomical Society* **329** (2002) 61 [https://academic.oup.com/mnras/article-pdf/329/1/61/3882215/329-1-61.pdf].

[49] D.S. Reed, R. Bower, C.S. Frenk, A. Jenkins and T. Theuns, *The halo mass function from the dark ages through the present day*, *Monthly Notices of the Royal Astronomical Society* **374** (2006) 2 [https://academic.oup.com/mnras/article-pdf/374/1/2/2835466/mnras0374-0002.pdf].

[50] J. Tinker, A.V. Kravtsov, A. Klypin, K. Abazajian, M. Warren, G. Yepes et al., *Toward a halo mass function for precision cosmology: The limits of universality*, *The Astrophysical Journal* **688** (2008) 709.

[51] J. Courtin, Y. Rasera, J.-M. Alimi, P.-S. Corasaniti, V. Boucher and A. Füzfa, *Imprints of dark energy on cosmic structure formation – ii. non-universality of the halo mass function*, *Monthly Notices of the Royal Astronomical Society* **410** (2011) 1911 [https://academic.oup.com/mnras/article-pdf/410/3/1911/2864964/mnras0410-1911.pdf].

[52] M. Crocce, P. Fosalba, F.J. Castander and E. Gaztañaga, *Simulating the universe with mice: the abundance of massive clusters*, *Monthly Notices of the Royal Astronomical Society* **403** (2010) 1353 [https://academic.oup.com/mnras/article-pdf/403/3/1353/6170753/mnras0403-1353.pdf].

[53] S. Bhattacharya, K. Heitmann, M. White, Z. Lukić, C. Wagner and S. Habib, *Mass function predictions beyond λcdm*, *The Astrophysical Journal* **732** (2011) 122.

[54] K. Ahn, I.T. Iliev, P.R. Shapiro, G. Mellema, J. Koda and Y. Mao, *Detecting the rise and fall of the first stars by their impact on cosmic reionization*, *The Astrophysical Journal Letters* **756** (2012) L16.

[55] M. McQuinn, A. Lidz, O. Zahn, S. Dutta, L. Hernquist and M. Zaldarriaga, *The morphology of h ii regions during reionization*, *Monthly Notices of the Royal Astronomical Society* **377** (2007) 1043 [https://academic.oup.com/mnras/article-pdf/377/3/1043/5678910/mnras0377-1043.pdf].

[56] M.G. Santos, L. Ferramacho, M.B. Silva, A. Amblard and A. Cooray, *Fast large volume simulations of the 21-cm signal from the reionization and pre-reionization epochs*, *Monthly Notices of the Royal Astronomical Society* **406** (2010) 2421 [https://academic.oup.com/mnras/article-pdf/406/4/2421/3339431/mnras0406-2421.pdf].

[57] Doussot, Aristide and Semelin, Benoît, *A bubble size distribution model for the epoch of reionization*, **667** (2022) A118.

[58] R.K. Sheth and G. Lemson, *Biasing and the distribution of dark matter haloes*, *Monthly Notices of the Royal Astronomical Society* **304** (1999) 767 [https://academic.oup.com/mnras/article-pdf/304/4/767/3558836/304-4-767.pdf].

[59] R.K. Sheth and G. Lemson, *The forest of merger history trees associated with the formation of dark matter haloes*, *Monthly Notices of the Royal Astronomical Society* **305** (1999) 946 [https://academic.oup.com/mnras/article-pdf/305/4/946/18635251/305-4-946.pdf].

[60] A. Barsode and T.R. Choudhury, *Efficient hybrid technique for generating sub-grid haloes in reionization simulations*, 2407.10585.

[61] V. Springel, *The cosmological simulation code gadget-2*, *Monthly Notices of the Royal Astronomical Society* **364** (2005) 1105 [https://academic.oup.com/mnras/article-pdf/364/4/1105/18657201/364-4-1105.pdf].

[62] M. Davis, G. Efstathiou, C.S. Frenk and S.D.M. White, *The evolution of large-scale structure in a universe dominated by cold dark matter*, **292** (1985) 371.

[63] A. Paranjape, *A simulated annealing approach to parameter inference with expensive likelihoods*, *arXiv e-prints* (2022) arXiv:2205.07906 [2205.07906].

[64] B. Maity, A. Paranjape and T.R. Choudhury, *A fast method of reionization parameter space exploration using gpr trained script*, *Monthly Notices of the Royal Astronomical Society* **526** (2023) 3920 [https://academic.oup.com/mnras/article-pdf/526/3/3920/52108319/stad2984.pdf].

[65] T.R. Choudhury, A. Paranjape and B. Maity, *A gpr-based emulator for semi-numerical reionization code script: parameter inference from 21 cm data*, *Journal of Cosmology and Astroparticle Physics* **2024** (2024) 027.

[66] D. Baumann, *Cosmology*, Cambridge University Press (2022).

[67] H.J. Mo and S.D.M. White, *An analytic model for the spatial clustering of dark matter haloes*, **282** (1996) 347 [astro-ph/9512127].

[68] A. Paranjape, T.R. Choudhury and H. Padmanabhan, *Photon number conserving models of Hii bubbles during reionization*, *Monthly Notices of the Royal Astronomical Society* **460** (2016) 1801 [https://academic.oup.com/mnras/article-pdf/460/2/1801/8116027/stw1060.pdf].

[69] O. Zahn, A. Mesinger, M. McQuinn, H. Trac, R. Cen and L.E. Hernquist, *Comparison of reionization models: radiative transfer simulations and approximate, seminumeric models*, *Monthly Notices of the Royal Astronomical Society* **414** (2011) 727 [https://academic.oup.com/mnras/article-pdf/414/1/727/3838580/mnras0414-0727.pdf].