Explorer l'aube cosmique et l'époque de réionisation avec le signal 21 cm

Evan Eames

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'Observatoire de Paris

Explorer l'aube cosmique et époche de réionisation avec le signal 21 cm

École doctorale n°127

ASTRONOMIE & ASTROPHYSIQUE D'ÎLE-DE-FRANCE

Spécialité COSMOLOGIE

COMPOSITION DU JURY :

M Philippe ZARKA
LESIA, Président

M Réza ANSARI
LAL, Rapporteur

M Garrelt MELLEMA
Stockholm University, Rapporteur

M Dominique AUBERT
Université Strasbourg, Membre du jury

Mme Paola DI MATTEO
GEPI, Membre du jury

M Andrei MESINGER
SAS Pisa/Italy, Membre du jury

M Benoît Semelin
LERMA, Membre du jury

Soutenue par **Evan EAMES**
le 14 nov 2018

Dirigée par **Benoît Semelin**

l'Observatoire de Paris | PSL RESEARCH UNIVERSITY PARIS

PSL UNIVERSITÉ PARIS

l'Observatoire de Paris | LERMA

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PARIS SCIENCES ET LETTRES

École Doctorale d'Astronomie & Astrophysique d'Île-de-France

LERMA, Observatoire de Paris

EXPLORING THE COSMIC DAWN AND EPOCH OF REIONIZATION USING THE 21 CM SIGNAL

Présentée par

Evan EAMES

Et soutenue publiquement le 14 novembre 2018 devant un jury composé de :

Philippe ZARKA	Président du jury
Benoît SEMELIN	Directeur de thèse
Réza ANSARI	Rapporteur
Garrelt MELLEMA	Rapporteur
Dominique AUBERT	Examinateur
Paola DI MATTEO	Examinatrice
Andrei MESINGER	Examinateur


DON'T
PANIC

Back: Douglas Adams, 1985
The "Hitchhiker's Guide to the Galaxy".

Résumé

Les simulations, de plus en plus, sont capables de saisir la complexité de l'époque de réioisation, durant laquelle l'hydrogène neutre de l'Universe a été ionisé par les premières sources lumineuses. Des bases de données représentatives de la multitude de signaux possibles seront nécessaires pour contraindre les paramètres des modèles quand des observations 21 cm seront disponibles. À cette fin, et en préparation des observations à venir sur des instruments comme SKA, nous avons développé une base de données de cones de lumières EoR haute-résolution (21ssd.obspm.fr), ainsi qu'une modélisation du bruit thermique. Nous avons également développé un formalisme permettant de quantifier la différence entre les modèles de cette base de données, en utilisant le spectre de puissance et la fonction de distribution des pixels. Nous trouvons que les deux diagnostiques sont sensibles à des paramètres différents des modèles, ce qui signifie que les deux peuvent être utilisés ensemble de manière complémentaire pour extraire l'information maximale. De plus, en utilisant le code 21cmFAST, nous avons développé des stratégies pour échantillonner l'espace des paramètres d'une mannière optimale (plus homogène et isotrope), afin de fournir le meilleur point de départ pour l'extraction des paramètres. Finalement, l'échantillonnage amélioré est utilisé pour entrainer un réseau de neurones. Ce réseau retrouve les paramètres du modèle en se basant sur une observable. Nous observons une amélioration modérée dans la précision de ses prédictions quand nous utilisons l'échantillonnage optimisé lors de son entrainement.

Abstract

Simulations are increasingly able to capture the intricacies of the Epoch of Reionization, during which the neutral hydrogen in the Universe was ionized by the first luminous sources. Databases encompassing the range of possible signals will be needed to constrain parameter values when 21 cm observations are available. In preparation for upcoming experiments such as the SKA, we have developed a database of high-resolution EoR lightcones (21ssd.obspm.fr), along with realistic thermal noise modelling. We examine frameworks with which we can quantify the difference between entries in this database, specifically with the power spectrum and pixel distribution function. We find that the two diagnostics are sensitive to different parameters, meaning they can be used together to extract maximumal information. Then, using the 21cmFAST code, we explore how to optimally sample a parameter space (so that it is more homogeneous and isotropic), in order to provide the best set-up for parameter extraction. Finally, the improved sampling is used in training a neural network. The neural network uses observables as input data, and attempts to estimate the corresponding parameter values. When the optimal sampling is used as training data, we find that the neural network is able to estimate parameter values with a modest improvement in accuracy.

Research summary

The period of the Universe during which the first stars and galaxies formed is still not well understood. This epoch, known as the 'Cosmic Dawn', is considered important towards understanding the beginning of large-scale structure. After these first luminous sources came to be created, they gradually ionized the neutral hydrogen in the Universe. Ionized regions grew and combined, until the Universe was more-or-less entirely ionized. This period, which is thought to have taken place roughly 13 billion years ago, is called the Epoch of Reionization, and provides the setting for this manuscript.

Neutral hydrogen emits radiation at ∼21 cm (∼1400 MHz), and this is expected to prove a key tool towards studying the progression of reionization. Theoretical work, in tandem with recent simulations, has attempted to estimate the expected strength of this radiation when emitted in the intergalactic medium during the Universe's infancy. The value is dependant on many factors: the astrophysical properties of the first luminous sources, how exactly each of them (galaxies, quasars, X-ray binaries) individually contributed, the start and duration of reionization, etc. Most now agree that the maximum strength of the signal will be somewhere between a few tens of mK to a few hundred mK. Compare this to the cosmic microwave background, detected in the 1960s, which is orders of magnitude stronger. Nonetheless, current experiments are accurate enough that the signal should be detectable (the issue is now properly modelling all the brighter 'foreground' contamination between us and the distant Universe, and calibrating instrumentation accordingly).

With this context in mind, we set out to prepare for these upcoming observations. Firstly, a database of high-resolution simulated 21 cm signals (lightcones) was created. Low-resolution versions, intended to simulate upcoming observations, were also created, and we included thermal noise (based on SKA specifications) for both. We then looked into defining the 'distance' between different lightcones. Finally, using a semi-numerical code, we explored how best to create a database 'optimally'. The idea is that, when the true 21 cm signal is eventually detected, we will be able to compare this real signal to the simulated signals in our database in order to determine the true nature of physical phenomenon in the early universe (the physics of the first sources, for example). There must be a way to optimally choose how to simulate signals, in order to assure the best chance of understanding the real one.

The database was created using a fully coupled radiative hydrodynamical code named LICORICE, which has been described in a number of papers (Semelin 2016 and references therein). The simulation begins with an initial hydrogen field at redshift ∼100, from which the gravitational, radiative, and baryonic physics are explicitly calculated and advanced with each time step. As the simulation box evolves, thin slices at different redshifts are arranged one-by-one to build a full time-line of the EoR: known as a lightcone. As LICORICE is a computationally intensive simulation, the first version of the database (named 21SSD: 21 cm Simulated Signal Database) contains 45 different simulated signals. Between the signals, three astrophysical parameters are varied: two of which relate to the X-ray properties of early sources, and the third quantifies the Lyman

α emissivity. We also simulate thermal noise, somewhat based on previous modellings (McQuinn et al., 2006; Mellema et al., 2013; Koopmans et al., 2015), although modelling the UV visibilities is carried out with more up-to-date values (Dewdney, 2015). The final database is described in (Semelin et al., 2017).

These lightcones are then simplified and quantified, specifically in the context of comparing them efficiently. The individual lightcones are not only on the order of $\sim$a few tens of Gb in file size, but the placement of their ionization bubbles is arbitrary. This makes them impossible to compare without first being simplified. We first create a routine to calculate the power spectra along the length of the lightcone (corresponding to the line of sight). Each of these has some width along the line of sight, but roughly corresponds to a single redshift. The advantages of allowing for some width are noted in (Morales & Hewitt, 2004; Morales, 2005), and the concept has been used by a number of authors. In addition, we consider a less explored diagnostic for differentiating between the simulated signals: the pixel distribution function. This is effectively a 2D histogram, in which the pixels at each redshift are binned based on their brightness temperature. This second method proves interesting, in that it seems to better capture some of the more complex morphologies of the EoR. We also look at the distances between different simulated signals using both diagnostics. The power spectra appears to be more sensitive to the X-ray efficiency of early sources, while the pixel distribution function picks up more on the Lyman α efficiency. Neither method seems especially sensitive to the third parameter (the ratio of high energy and low energy X-ray photons emitted by sources).

We then go further, asking if there is an optimal choice for parameter values when simulating signals and creating a database. The definition of optimal is non-trivial, but in a nutshell it requires the distribution of the simulated signals (using the above framework for distances) to be as homogeneous and isotropic as possible. To explore this idea of database optimization, we rely on the semi-numerical code 21cmFAST (Mesinger et al., 2011) which, although not as detailed in the physical modelling as LICORICE, can simulate the EoR quickly. To build up the basics of an algorithm to create an optimal database, we require a fast code in order to test many different parameter values. Part of this endeavour also involves understanding the geometry of the space inhabited by the power spectra, and defined by the distances between them. We quantify this geometry using metrics, and the resulting eigenvectors. Ultimately, two different optimization algorithms are tested. Assuming the parameter values initially lie on a grid, the first one rotates and stretches the grid (based on the average metric information) in order to find the best orientation for a homogeneous and isotropic distribution of the simulated signals. The second does not assume any grid, and instead tries to move points in the parameter space iteratively (based on the local metric information) such that they are not overly close to their neighbours. We finally verify whether these methods result in databases that can better train a neural network to estimate the parameter values with which test signals were created. We do find modest improvement, although there is still work to be done to assure all regions of the parameter space reflect this improvement (Eames, 2018; in prep).

Acknowledgements

Ah yes, time to thank the innumerable people who helped me along the three year journey, now culminating, in a way, by writing this. There have been so many people, and I will do my best to convey just how grateful I am (a tricky task — I'm very grateful!).

Yet deciding who to start with is easy. My supervisor Benoît Semelin deserves the main ovation. These three years, he has been the perfect balance between supportive, patient, understanding, and (crucially) able to politely tell me to speed up when needed. Merci.

Then there's Paris, and France in general. Sure, it's perhaps odd to thank a city, but I owe so much to the place in which I've spent the best three years of my life. I really can't overstate how warmly welcomed I've felt here — it simply seems like home.

Surely it's all my friends who come next. Aymeric first, who I befriended a decade ago, having found ourselves stranded in the Tasmanian jungles; and then fate decided that we would be neighbours all these years later. Ane and Isadora, never hesitating to share every ounce of their Brazilian 'goût de vie'; Julien, always ready for fresh new adventures; the Italians, putting up with all my jokes; the Indians, putting up with my Bollywood attempts; the French, putting up with my 'artistic interpretation' of their language.
.
.
.
So many lovely people I've met at my residence. And all nationalities (30 at last count) hanging out, breaking bread, learning one another's languages, sharing our ideas. Truly a tiny little utopia. Another heartfelt merci is due for Édouard, who graciously agreed to let me stay with him without a moment's hesitation. There are also the ∼one hundred lovely travellers I've hosted while in Paris: the world's stories and smiles at my doorstep.

Worth a quick nod are the incredible people I volunteered with at Restos du Cœur, especially Raymond. They made the years a lot of fun, and made Paris a better place.

Getting to the family stuff, my brothers are annoying brats! But I suppose it should be eventually added that I'm so incredibly proud of the amazing adults they've become. Ben too, who, being my oldest friend, is effectively a brother.

Mom and Dad, you're up last. Without the shadow of a doubt, you are the two reasons I am where I am today. Your insistence, from day one, on instilling curiosity and humility really laid down the solid foundation I needed to become...well...me. I am sure you remember countless times when, raising us and faced with big decisions, you were unsure if your choices were the right ones. I hope now, seeing what your boys have accomplished, every doubt has faded. So now it's your turn to celebrate and go enjoy the world! I don't doubt it will me amazing, and I wonder what's in store for the lot of us . . .
.
.
.
?

Contents

List of Figures

Introduction

CHAPTER 0

Introduction

Ever Larger Structure

There is a very very very lot of stuff out there. So much so that it can be rather intimidating to think about. As soon as we as a species could do this — think — we became naturally curious about the heavens. The twinkling stars and the sun and moon were initially attributed to deities, who watched down upon us with their ubiquitous gaze.

Yet, over the millenniums, we have slowly unravelled these mysteries to arrive at a clearer picture of these lights in our sky. Those dots that moved, called wanderers (the Greek meaning of 'planet'), were eventually understood to be other worlds orbiting a common star — our sun. And those distant twinkles were shown to be in fact stars, often very much like our own. They could even host their own planets; at least so much was suspected as early as 1584 by Giordano Bruno, as attested to in *"De l'infinito universo et mondi"* (Bruno, 1584); later to be proven in 1991 when an exoplanet was discovered (Latham et al., 1989; Cochran et al., 1991). In addition to our realization that we comprise neither the only planet, nor the only solar system, our classical sentiment of privileged uniqueness in the Universe was further challenged as time passed. In the 17^{th} and 18^{th} centuries the notion of a 'galaxy' — which has existed since the ancient Greeks — began to form an evidence-based foundation. First with Galileo's studies of distant faint stars (Galilei, 1610), then Thomas Wright's concept of a large body of gravitationally bound stars (Wright, 1750), Immanuel Kant's addition that there could be many of these 'Island Universes' (Kant, 1755), and then an attempt by William Herschel to understand the shape of our galaxy through counting stars in different regions of the sky (Herschel, 1785).

Wright and Kant were not only prescient in their supposition that stars could amass into large 'islands' – they also (correctly) supposed that some of the blurry nebula seen in the night sky could be examples. Herschel did not share this view, and believed that nebulae were phenomena entirely within our own galaxy. In fact, this view was not seriously challenged until the 1900s, in which it was realized that novas in Andromeda

— thought at the time to be a nebula — were much fainter than those occurring in other regions of the Milky Way. On the 26^{th} of April, 1920 a debate between Harlow Shapley and Heber Curtis took place (Shapley & Curtis, 1921). Now remembered as the 'Great Debate', Curtis (who had noted the nova brightness discrepancy) argued that Andromeda and other 'Island Universes' were outside our own Galaxy, while Shapley believed that everything was contained within the Milky Way. But the debate did more than simply highlight contrasting scientific view points. It illustrates the difficulty we faced in conceiving the massive scales required for a multi-galaxy Universe. If Curtis was correct — as would eventually be accepted — then Andromeda would be millions of light years away.

Today we have accepted this vastness; it is common knowledge that there are many galaxies, perhaps on the order of a few trillion in our observable Universe (Conselice et al., 2016). Powerful new probes have peered deeper and further into our cosmos. The most iconic example of this generation is the Hubble Telescope, with which the stunning Hubble Deep Field image was created in 1995 (figure 1). In 2021 the James Webb Space Telescope will raise the bar further, and usher in a new era of deep space astronomy.

Another curious thing has happened as we have, through the decades, peered deeper and deeper into the Universe that surrounds us. On account of the finite speed of light, the further outwards we probe, the further back in time we see. Some of the light particles that created the Hubble Deep Field image had travelled for billions of years, before ending their journeys on the electronic cells of the telescope. This means that we are able to (and moreover, *obliged* to) see distant galaxies as they were billions of years ago.

Through this phenomenon, our understanding of the Universe has evolved to encompass not only the present, but the long chronology that has preceded our infinitesimal present existence. We can see stars being born and dying, young galaxies birthed through massive collapse while others collide catastrophically, ancient black holes voraciously shredding their surrounding stars and gas clouds with the accompanying hyper-luminous accretion disks that shine brightly across the billions of years that separate us, and even the faint afterglow of the big bang itself. The spectacle that is eternally performed in the dark skies above us is one that encompasses fantastical sights playing out for the watchful eyes of curious astronomers on a stage that stretches across, not only the physical heavens, but the chronology of existence itself.

Yet, the spacial and temporal dimensions of our Universe — though intimately paired — have a striking dissimilarity. As far as we can tell, our Universe has no conceivable *physical* border. We can speak of the edge of the 'Observable Universe', however this represents a barrier beyond which matter cannot be imaged, *not* beyond which it cannot exist[1]. Conversely, in the chronological direction, we know that there must be a previous time before which no galaxies nor stars could exist. We know this must be true for a number of reasons.

[1]It cannot be excluded from observations that the universe has a hyperspheric geometry. This would mean no boundaries, yet a finite volume (much larger than the current observable universe). The Omega curvature parameter is currently $\Omega_k = 0.001 \pm 0.002$ (Planck Collaboration et al., 2018), which could be negative, positive, or zero.

Figure 1 – *The Hubble Deep Field. Image Credit: R. Williams (STScI), the Hubble Deep Field Team, NASA.*

In 1929, it was discovered that distant galaxies are moving away from us faster than nearer ones (Hubble, 1929). This suggested that the Universe is expanding (and in fact, in the late 90s it was discovered that this expansion is accelerating; see Perlmutter et al. 1997; Riess et al. 1998; Schmidt et al. 1998). Extrapolating into the past, we expect that the Universe must have initially been much smaller, perhaps originating in a 'Big Bang' (as coined by British astronomer – and occasional radio presenter – Fred Hoyle in 1949, who was in fact an opponent of the model). Repeated observations over the decades have confirmed this phenomenon, most recently via cepheids (variable stars) (Riess et al., 2018) and the Cosmic Microwave Background (CMB) (Planck Collaboration et al., 2016a). Curiously the two methods currently offer different estimates for the speed of this expansion: 73.52 ± 1.62 and 67.74 ± 0.46 (km/s)/Mpc respectively. Some suggestions for the reason behind this disagreement are presented in Planck Collaboration et al. (2018), and it is hoped that upcoming estimations of the constant using gravitational waves will settle the matter (Nair et al., 2018).

The second of the two methods of estimating the constant of expansion — the Cosmic Microwave Background — is unsurprisingly also a 'smoking gun' of said expansion. Considered a possibility since the 1940s (Gamow, 1948), the background radiation would come to be known serendipitously in the mid 1960s when two young instrumentalists working at Bell Telephone Laboratories measured excess temperature they could not account for (Penzias & Wilson, 1965). Penzias and Wilson narrowly superseded other efforts to detect the signal, and this fortuity assured them the 1978 Nobel Prize.

What had been detected was faint radiation emitted when the Universe was a mere ∼379,000 years old ($z \approx 1089$[2]) as well as much hotter and denser. Earlier than this we expect that there was nothing more than a chaos of particles and nuclei[3], too energetic to allow electrons to be captured. As the expansion continued, the temperature fell, and at around 3000 K the particles were sufficiently cooled to form the first neutral elements (primarily hydrogen and helium). Rapidly the Universe went from being opaque to being traversable by photons, which would not be easily captured by the neutral atoms. It is these photons that can be detected everywhere in our Universe: residual radiation that still lingers after over 13 billion years. The radiation corresponds to a temperature brightness of ~ 2.73 K, follows a near perfect blackbody spectrum (with a peak in the microwave regime — hence the name), and is remarkably isotropic (with small anisotropies of root mean square roughly $\Delta T \approx 18\ \mu$K Wright 2004; Smoot 2007). Over the past decades a series of probes have incrementally refined our sky-wide maps of the CMB. Ordered by launch date they were: COBE in 1989, WMAP in 2001, and Planck in 2009.

The discoveries of Hubble expansion and the Cosmic Microwave Background together began to shape and illuminate many of the mysteries of our Universe's history. We now understand that our Universe is expanding, and that — of crucial importance — it was

[2]For the layperson, z is the redshift: a measurement of how stretched the light has become since is emission; this stretching is due to the expansion of the Universe. $z = 1089$ tells us that the light's wavelength is 1089 times longer when it arrives at Earth. Redshifts are the standard measure for referring to previous periods of the Universe's history.

[3]The first multi-nucleon nuclei are thought to have been formed within the first 20 minutes of the Universe's history, during a period known as Big Bang Nucleosynthesis.

once too hot and dense for elements to form. As our current cosmos is replete with uncountable structures and complexities – stars, galaxies, black holes, nebulae, comets, asteroids, planets, etc. — we know that there must have been a period during which the first structures were formed from the first gasses. We therefore refer to this period as that of 'Structure Formation'. Before any complex structure, we must not have had any luminous sources. Our Universe, we presume, consisted of simply neutral gas (at $z \lesssim 1000$). The only light particles would have been the CMB photons, streaming through the darkness with low — and ever lower — energies. This period is referred to, quite literally, as the 'Dark Ages'.

A Summary of What we Know

And so we can begin to construct a timeline. At first, for a few hundred thousand years, there was an extremely hot dense 'soup' of energetic particles, within which atoms could not form. Before this we can say very little with certainty. It seems tempting to continue extrapolating backwards, concluding that the entire Universe was housed within an initial and infinitesimal singularity. Perhaps this *was* the case, yet there may very well be more to the story. If we are someday able to directly detect the 'Cosmic Neutrino Background' (thought to be emitted when the Universe was only 1 second old, and indirectly confirmed in recent years, e.g. Follin et al. 2015), or the even more enigmatic 'Cosmic Gravitational Wave Background' (hypothetically emitted between $10^{-36} - 10^{-32}$ s after the Big Bang, Caprini & Figueroa 2018), then we will be at liberty to comment further on what came before. For now, the Universe's infancy remains shrouded in mystery.

After temperatures dropped sufficiently, neutral atoms were able to form, and the Universe became clear. All of the photons that had been trapped in the initial mess were suddenly free to travel the Universe unimpeded, and begun their long journeys simultaneously – collectively forming the Cosmic Microwave Background. The Dark Ages of the Universe began. Very slight overdensities in the neutral, and otherwise uniform, gas began to gravitate into increasingly dense and hot regions. The first large-scale structures began to appear in the Universe: webs of dense gas framing endless voids of lower density. Structure Formation began. Eventually, the most dense regions of these cosmic gas webs surpassed the threshold temperatures for hydrogen fusion (13 million K). It was then that the very first stars (called Population III stars) were born, a few hundred million years after the Big Bang ($z \approx 20 - 30$) (Bromm et al., 2009). This period is sometimes, quite poetically, referred to as the 'Cosmic Dawn'.

These first stars, entirely metal free, are thought to have been on the order of 100 times larger than our sun (Umeda & Nomoto, 2003), although smaller 'secondary' Population III stars, smaller than our own sun, may have also been produced (Krumholz, 2015). Regardless of their sizes, the first stars introduced high energy ionizing photons into the Universe. In the regions around the stars, the neutral hydrogen was slowly ionized, and these H_{II} (ionized) bubbles[4] continued to grow and eventually merge with other such

[4]When one of these spheres is in equilibrium, it is referred to as a Strömgren Sphere (Stromgren, 1939).

bubbles. And so, over the next billion years ($z \approx 20 - 6$) the Universe was completely ionized[5]. This period is referred to the *Epoch of Reionization* — often abbreviated as EoR — and will set the stage for the research presented in this thesis.

During this period, stars were only one of the sources of ionizing photons. At the beginning of the EoR, it is clear that stars drove reionization (Alvarez et al., 2006), however the exact contributions from other sources are still debated. Active Galactic Nuclei (AGN) and Supernovae are also thought to have contributed significantly (perhaps primarily at certain redshifts). In recent years, studies of extremely distant luminous objects have found that some of them existed in regions of the Universe that were not yet fully ionized. Spectral absorption lines in the quasar ULAS J112001+0641 ($z \approx 7$) indicate that the hydrogen at that time was >10% neutral (Mortlock et al., 2011). Such studies seem to imply that reionization was complete at $z \approx 6$ (Planck Collaboration et al., 2016b; Becker et al., 2001).

After reionization, as of around 1 billion years after the Big Bang, the Universe begins to look more like what we know today. Effectively all of the hydrogen has been ionized in the intergalactic medium, massive stars die in fiery supernovas and litter their neighbourhoods with heavier elements, galaxies merge and evolve, and (relatively late in the game) the expansion of the Universe begins to mysteriously accelerate. Over the past 12 billion years our solar system was born from the metal-rich remnants of previous generations of stars, our planet's chemistry allowed for the formation of liquid water, cyanobacteria provided the oxygen, evolution through natural selection brought about a curious ape, and in the blink of an eye we find ourselves here today to wonder about it all.

Where to go from Here?

Looking back at all that has happened, our inclination is to try to deepen our understanding of periods that are less well understood. With the exception of the CMB, we have yet to see anything older than $z \approx 11$ (the current record being galaxy GN-z11, formed when the Universe was ∼400 million years old Oesch et al. 2016). Although it would be fascinating to peer towards the extremely young Universe, as it existed before the emission of the CMB, the technical challenges involved are enormous. As stated, until we perfect the detection of either cosmic neutrinos, or cosmic gravitational waves, this early epoch will remain beyond our reach. However, the upcoming JWST will be able to push far enough to peer into the Dark Ages, and is expected to see 1 galaxy at $z \approx 11$ per field of view in 10^4 s (Cowley et al., 2018). Currently we have only observed a single galaxy at this distance (Oesch et al., 2016). See figure 2 for a comparison to the Hubble Space Telescope.

However, such early epochs of the Universe begin to present us with a curious problem: lack of sources. Between the emission of the CMB and the period of the first stars our cosmos consisted of nothing but neutral gas. Although there is much information to

[5]Or at least 99.9% ionized at $z = 6$ (Fan et al., 2006a).

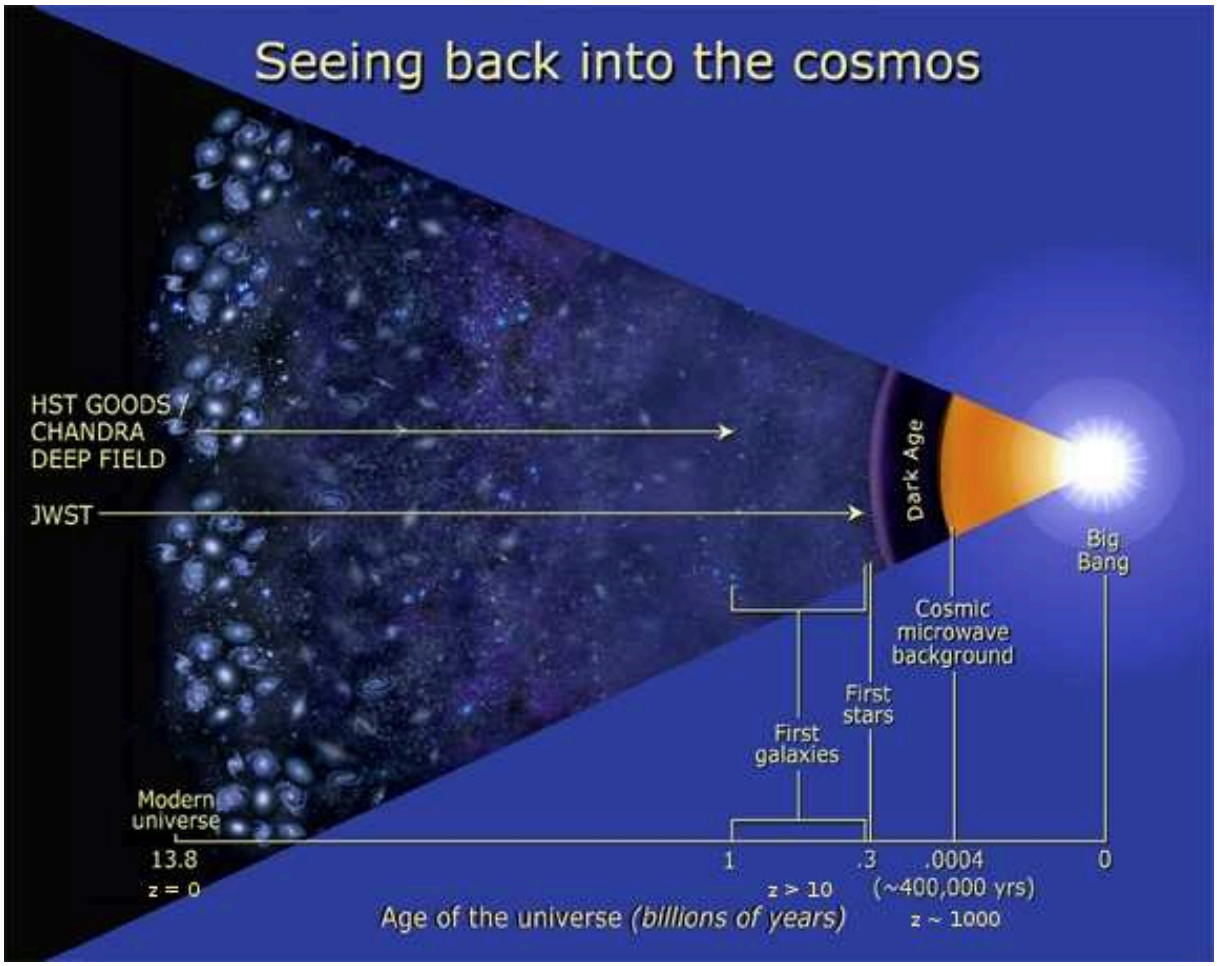


Figure 2 – *Cosmic timeline. Image credit: NASA & Ann Feild*

be gained during the Dark Ages of the Universe — such as how structures developed from slight density variations — optical and IR telescopes simply have no sources with which to study this period. Regardless of how big one envisions the collecting area of the instrument, the problem remains the same. "If only there were a way to study the gas itself," one might lament.

Being able to image the position and density of gas across the ancient Universe would be invaluable not only for studying the Dark Ages. It would also allow us to constrain the duration and properties of the Epoch of Reionization. Many of our theoretical assumptions regarding the EoR find themselves based upon subsequent assumptions about the objects responsible for ionizing. We are unsure about how important early AGN were in ionizing the Universe compared to stars, for example, simply because we are unsure of how early AGN and stars behaved. If we were able to image the structure formation, and eventual ionization of the gas, it would allow us to infer the properties of these first sources. Understanding the astrophysical properties of said sources would, in turn, give us a better understanding of our Universe's fundamental workings. However, this all depends on being able to image neutral gas.

As it happens there is, in fact, a way to do exactly that...

The 21 cm Hydrogen Line

The magic lies in a very specific wavelength[6]: 21 cm (van de Hulst, 1945; Furlanetto et al., 2006). Or 21.1061140549 cm (in free space) to be exact (Dupays et al., 2003). In cosmology there is perhaps no observational tool with more potential than that of the 21 cm hydrogen line, also known simply as 'The Hydrogen Line'. If we properly develop our ability to detect this evasive signal from the early Universe — an endeavour that is on track to becoming a reality in the coming decade — it will herald a new era of profound understanding regarding our place in the cosmos. It is with this signal that we hope to be able to, as hinted at above, explore the periods of our Universe bereft of luminous sources.

As it stands, there still remain many challenges towards achieving this goal. However, instrumental sensitivity has, in recent years, finally arrived at what is expected to be sufficient for detection. The remaining hurdles lie primarily in theoretical and technical considerations, such as proper instrumental calibration, and developing clever methods of dealing with the multitude of phenomena that obscure this distant signal. The three main offenders of this obstruction are: Galactic Synchrotron emission, Galactic Free-free emission, and extragalactic emission. These sources of foreground radiation can be many orders of magnitude brighter than 21 cm radiation. Foregrounds will be discussed in detail

[6]In fact, the neutral gas of the Dark Ages has a number of emission lines due to hyperfine transfers (those due to deuterium or helium-2), however these are much weaker. The 21 cm line is the strongest, and therefore the most studied.

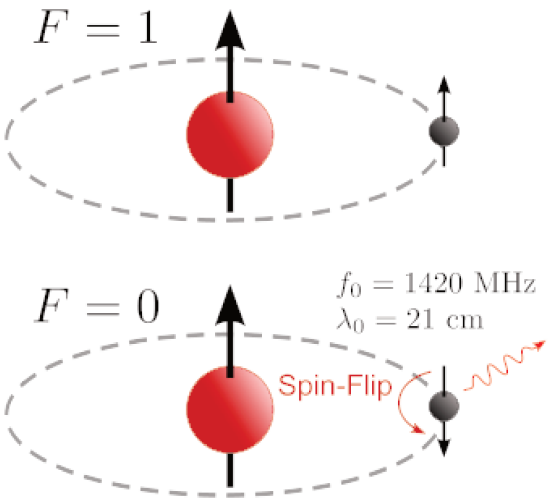


Figure 3 – *A 21 cm photon emitted from a neutral hydrogen atom.*

in section 0.5. Beforehand, to understand and contextualize this hurdle, it is necessary to look towards the history of the subject.

However, perhaps a slight teaser is in order before delving deeper into the richness of the topic. Earlier this year a tentative detection — after decades of searching — could be the first concrete success of 21 cm cosmology (Bowman et al., 2018). Although much (rightfully deserved) scepticism surrounds the announcement, it still incites titillating excitement to know that we are likely on the verge of finally unlocking the secrets of the Dark Ages.

Initial Discovery

The idea that neutral hydrogen is able to radiate may strike one as odd. Yet a very subtle quantum effect makes this so. The 21 cm hydrogen line is a consequence of hyperfine splitting in the energy of the ground state of neutral hydrogen. More straightforwardly, when the spin of the electron and proton are in parallel (pointing in the same direction), the energy is slightly higher (5.9×10^{-6} eV, van de Hulst 1945; Furlanetto et al. 2006) than if the spins are anti-parallel (pointing in opposite directions). The spin of the electron can spontaneously flip to the lower energy configuration[7], which will release the 5.9×10^{-6} eV of energy in the form of a photon with $\nu \approx 1420$ MHz and $\lambda \approx 21$ cm. Conversely, a 21 cm photon can be absorbed by a hydrogen atom in the lower configuration, causing the spin to change from anti-parallel to parallel.

In the 1940s, Dutch physicist Jan Oort began to recognize the possibility of probing the Galaxy with radio waves on account of their ability to penetrate much of the opaque

[7]This transition has an Einstein-A coefficient of $A_{1\to0} = 2.85 \times 10^{-15}$ s^{-1}, which corresponds to a lifetime of 1.1×10^7 years (Zaroubi, 2013).

phenomena that obscure other wavelengths. His student, Hendrik van de Hulst, correctly predicted the theoretical existence of the hydrogen line in 1944 (van de Hulst, 1945), however detection would take the team another 7 years. In 1951 Oort and a young radio astronomer, C. A. Muller, did finally succeed in detecting the Galactic hydrogen line (Muller & Oort, 1951). The two owed their success largely to Harold Ewen and Edward Purcell, who offered technical help after having detected the line earlier that year using their iconic horn antenna (Ewen & Purcell, 1951). In fact, the latter two had held off on publishing until comparing with van de Hulst (although, unlike with the CMB, nobody lost out on a Nobel Prize; the discovery did not lead to any laureates).

After this initial discovery, the mid to late fifties saw a flurry of subsequent interest in the 21 cm line. In 1954 Oort, Muller, and van de Hulst used the hydrogen line to map the spiral structure of the Galaxy for the first time (van de Hulst et al., 1954). The potential for studying extragalactic sources was also quickly recognized, and the focus shifted to the nearby Andromeda galaxy (van de Hulst et al., 1957), other more distant galaxies (Kerr et al., 1954; Chamaraux et al., 1970), and eventually intergalactic space (Field, 1959a).

The Hydrogen Line in Cosmology

Early Interest

At the time of discovery, it was already expected that the early Universe should have contained large quantities of neutral hydrogen. However, interest in cosmological applications was not immediate as it was expected that HI radiation would prove much too weak for detection. Beginning in the mid 70s, the hydrogen line saw a second wave of interest when it was realized that it could likely prove a powerful tool in cosmology for probing Recombination (Dubrovich, 1975), the Dark Ages (Varshalovich & Khersonskii, 1977), as well as the IGM (Watson & Deguchi, 1984), and that the signal may be detectable in the near future.

Tomography

One of the primary reasons that 21 cm radiation is considered as a strong candidate for understanding the EoR and the Dark Ages is that it would allow for the construction of a large scale ‘tomograph’. Most extragalactic radio sources are spectrally smooth, which is to say they emit photons at all frequencies. For example, the Hubble Ultra-Deep Field (figure 1) contains a 2D projection of roughly 10,000 galaxies, all of which are imaged at different times in the Universe’s history. Certainly, imaging the galaxies at a range of frequencies most definitely gives us interesting information about their age, physical properties, and distances. However, we will only see each galaxy as it existed at one specific point in time. Shifting our observations in frequency space will affect only the relative brightnesses of the individual galaxies (redshift evolution will be negligible).

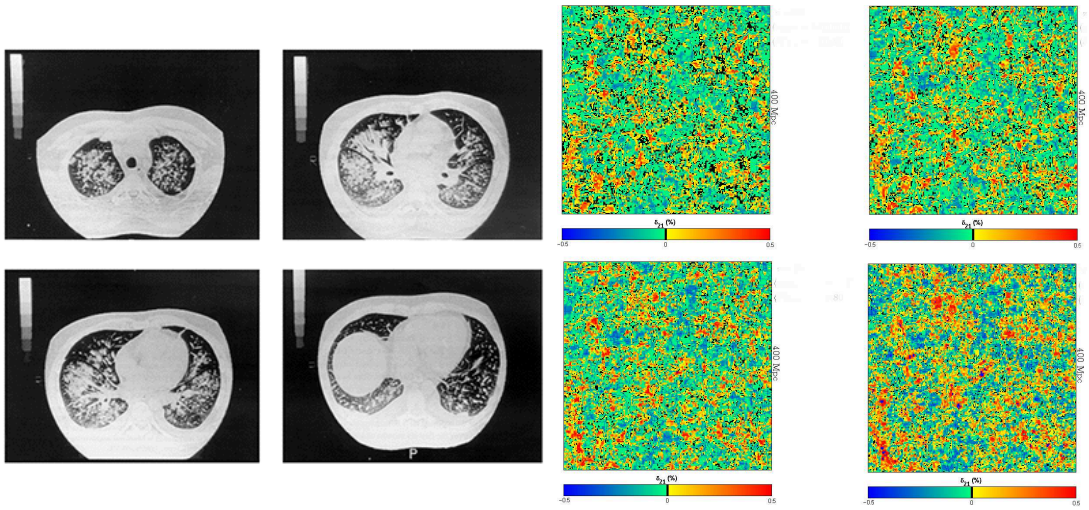


Figure 4 – *In anatomy, an example of tomography is viewing slices through a human body. In cosmology, we can view slices through the Universe; however, each slice refers to a different time period. Image Credit: scielo.br & 21cmFAST (Mesinger et al., 2011).*

Contrary to this, for any instance throughout the Universe's history in which there has been any hydrogen, the accompanying hydrogen line has always been emitted at a single discrete wavelength: 21 cm. Therefore, ignoring any peculiar velocity along the line of sight, differences in the frequency of the hydrogen line as observed from Earth can only be due to the effects of redshift. As we know, redshift is in turn directly related to the age of the radiation. On account of this relationship, observing only a thin band of frequencies is equivalent to focusing on a thin 'slice' of the Universe's history, and varying the frequency moves our observations through the hydrogen's chronology. The process of creating a 3D image of the cosmos and then studying specific slices is called tomography[8]. Performing a tomography, as opposed to a simple projection image, would equate to a more robust and comprehensive glimpse into the early Universe. Some example slices are shown in figure 4, as well as a more detailed slice 14 on page 43.

First Detection Attempts

In the 1970s an attempt was made by Rashid Sunyaev and Yakov Zel'dovich to predict the signal strength of high redshift EoR 21 cm radiation (Zel'dovich, 1970; Sunyaev & Zeldovich, 1975). However, the pair incorrectly assumed that large gas overdensities should form first. The idea was that eventually these large overdnesities would break down into smaller overdensities, giving way to the first structures: filaments, stars, and galaxies. This was known as the 'top-down' view of structure formation. If true, the signal from massive sheets (with masses estimated to be up to $\sim 10^{15} M_{\odot}$) was estimated to be as high as ∼10 K when observed from Earth (Sunyaev & Zeldovich, 1975).

[8]Tomography can be used outside of the astrophysical context, for example to study cross-sections of the human body (MRI, figure 4), foetal development (Ultrasound), or large volumes of water (SONAR).

Another early study exploring the possibility of detecting high redshift 21 cm signals was the work of Craig Hogan and Martin Rees in 1979 (Hogan & Rees, 1979). They predicted correctly that the primordial 21 cm radiation should show structure in both real and redshift space, and were also able to formulate a basic theory for how the brightness temperature should vary with frequency. However, they overestimated the redshifts at which emission would be visible with 1979 technology to be as high as $z \approx 9$. Thus, it is unsurprising that the next decade saw many failed attempts at detecting high redshift 21 cm radiation (Bebbington, 1986; Hardy & Noreau, 1987; de Bruyn et al., 1988).

Thankfully, the 1990s brought fresh insight into why the signal had proven so elusive at these high redshifts. In 1990 Rees re-evaluated the theories of primordial hydrogen structure with Douglas Scott, (Scott & Rees, 1990). In particular, they re-visited the previous failures of high redshift 21 cm radiation detection and were able to prove that structure must indeed form from the 'bottom up', with small clumps of hydrogen giving way to increasingly large clusters. In addition, they presented an early formula for calculating the expected brightness temperature for a given redshift[9]:

$$T_b = 3.66 \times 10^{-23} \times \frac{\rho_0 \Omega_B f \sqrt{1+z}}{\mu_H m_H H_0} \text{ K} \tag{1}$$

where ρ_0 is the present-day density of the Universe (equal to $\frac{3H_0^2}{8\pi G}$ kg m^{-3}), Ω_B is the ratio of baryon density to critical density, f is the fraction of hydrogen that is neutral, $\mu_H m_H$ is the average mass per hydrogen atom, H_0 is Hubble's constant, and z is, of course, the redshift. Inserting modern values from Planck Collaboration et al. (2016a) we expect that near the end of the EoR ($z = 6$) we need sensitivity to T_b fluctuations on the order of 3.3 mK. To probe an early period of the EoR ($z = 9$) these fluctuations would be as small as 2.3 mK, and to probe the Dark Ages (perhaps the birth of the first stars at $z \approx 30$) we would need to detect fluctuations <1 mK. It is therefore unsurprising that, by the early 90s, experiments had failed to detect primordial 21 cm radiation; they simply lacked the necessary sensitivities (sensitivities were on the order of 900 mK, see de Bruyn et al. 1988).

However, thankfully the situation is not as bleak. Equation 1 considers only the emission regime[10] ($T_S \gg T_{\text{CMB}}$), and also does not consider the contribution of the ionized fraction (approximations are listed in section 9 of Scott & Rees 1990). We shall see in section 0.4 that current estimates of the brightness temperature at these redshifts are between 30 and 50 mK (Furlanetto et al., 2006). A more accurate expression[11] for calculating the brightness temperature will be explored in equation 16.

[9] The corresponding equation in the article is eq.7. Here the N_{HI} and $\Delta\nu$ terms have been expanded, hence the different appearance.

[10] At the time, the redshift at which the signal would switch from absorption to emission was more uncertain.

[11] Although Scott & Rees (1990) weren't too far off the mark for the emission regime. Modern simulations still give an emission signal of ∼a few mK with the high T assumption (see Ross et al. 2018, figure 3).

Continued Attempts

Many radio astronomers, including Rees and Scott, were optimistic that the Giant Metrewave Radio Telescope (GMRT) array, built in India in the early to mid 90s, would finally provide the necessary precision to detect primordial hydrogen line radiation. Piero Madau and Avery Meiksin teamed up with Rees in 1997 to stipulate as to whether or not the GMRT would be able to detect the faint T_b fluctuations discussed previously. It was concluded that, if reionization occurred at redshifts between $z = 5$ and $z = 10$, then the GMRT stood a good chance of detection Madau et al. (1997). The GMRT went online the next year, however to date it has not yet detected primordial neutral hydrogen radiation up to $z \approx 8.6$ with sensitivity ≈ 248 mK (Paciga et al., 2011, 2013; Paciga, 2013). The sensitivity of the GMRT is thought to be sufficient, however the calibration, systematics, and foregrounds are the limiting factors that have yet to be overcome.

In the new millennium, many authors have presented an in-depth look into the prospects of 21 cm tomography (Loeb, 2005; Furlanetto et al., 2006; McQuinn et al., 2006; Santos & Cooray, 2006; Mao et al., 2008). If temperature brightness sensitivity limits are sufficiently low, it is shown that the method could be realistic as far back as $z \approx 50$, a period long before the first stars. Taken from Mao et al. (2008), figure 5 gives a rough picture of tomography's predicted reach in terms of the Universe's chronology. Amongst the conclusions presented by the aforementioned paper, and in addition to probing the EoR, it is expected that tomography performed by the upcoming Square Kilometre Array (SKA) could also improve our limits on space curvature and neutrino masses by two orders of magnitude (SKA-Collaboration, 2015; Shao et al., 2015; Sprenger et al., 2018).

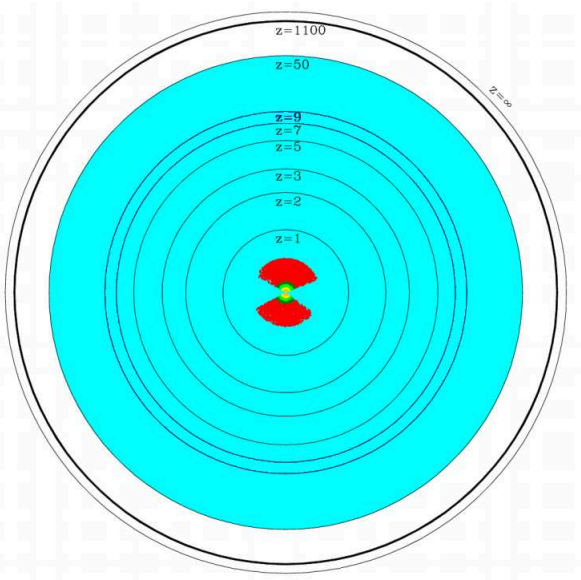


Figure 5 – *21 cm tomography could probe the Universe at all redshifts as far back as $z \approx 50$ (light blue). In comparison, the CMB allows us to see only $z \approx 1089$ (thick line), and the Sloan Digital Sky Survey (Gunn et al., 2006) up to $z \approx 0.7$ (red). Figure taken from (Mao et al., 2008).*

Recent Observational Developments

Construction of the SKA is expected to begin next year, with the first observations coming in the mid 2020s. For the time being, plenty of pathfinders have either already been deployed, or will soon be completed.

21 cm Power Spectrum Detection Experiments

Some of these aim to trace the Spacial/Spectral intensity fluctuations. This would allow us to create a power spectrum of the 21 cm signal.

- Already mentioned, the Giant Metrewave Radio Telescope was one of the earliest efforts, and current constraints are $\Delta^2(k) \leq 248^2$ mK^2 for $z \approx 8.6$ (Paciga et al., 2013).

- The Murchison Widefield Array (MWA) is another such pathfinder operating in Australia (Tingay et al., 2013). Observations began in 2012 for the redshift range $z = 6.2 - 11.7$, and to date the best limit in the range is $\Delta^2(k) \leq 164^2$ mK^2 for $z \approx 7.1$(Dillon et al., 2015; Beardsley et al., 2016). An upgraded Phase II version of the MWA is online as of early 2018, and is expected to push these limits further. Limits have also been placed on the beginning of the EoR ($z \approx 12 - 18$), and are $\Delta^2(k) \leq (10^4)^2$ mK^2 in this redshift interval (Ewall-Wice et al., 2016).

- Another effort is the Precision Array for Probing the Epoch of Reionization (PAPER) (Parsons et al., 2010). Built in South Africa as a collaboration between a group of American Universities, it began collecting data in 2009. PAPER previously set an upper limit at $z = 8.4$ of $\Delta^2(k) \leq 22^2$ mK^2 (Parsons et al., 2014; Ali et al., 2015), which was the lowest at the time, however the result has since been retracted due to insufficient noise considerations[12]. The project has recently wrapped up, and will eventually be integrated in HERA (section 0.3.3).

- One of the more ambitious pathfinder efforts is the Low Frequency Array (LOFAR), which consists of 50,000 antennas (in the 110-250 MHz band of interest here) arranged into 51 stations across Europe (with the central cluster in the Netherlands) (van Haarlem et al., 2013). At present the upper limit is $\Delta^2(k) \leq 79.6^2$ mK^2 in the range $z = 9.6 - 10.6$ (Patil et al., 2017)[13]

[12] See the Erratum addition in Ali et al. (2015).

[13] As of 2018 this has been pushed further to $\Delta^2(k) \leq 45^2$ mK^2 at $z = 9$, which represents (since the retracted PAPER result) the current deepest upper limit (not yet published).

Globally Averaged 21 cm Experiments

As opposed to constructing instrumentation to measure variations of the signal across the sky, which requires multiple receivers and a large collecting area, a cheaper alternative is to measure the 'globally averaged 21 cm signal'. This option is quite common as it is less expensive[14], and the following instruments are examples. However, there are disadvantages. The averaged signal contains much less information, and loses effectively all morphological information relating to the evolution of structure. This in turn can make foreground subtraction more difficult. What we are left with is simply a 'bulk sum' of neutral hydrogen for various epochs.

- The Large-Aperture Experiment to Detect the Dark-Ages (LEDA), observing between $\sim 30- \sim 85$ MHz ($z \approx 16 - 34$) in California and New Mexico, has been online since 2011 (Price et al. (2018); www.tauceti.caltech.edu/leda/).
- BIGHORNS (Broadband Instrument for Global Hydrogen Reionization Signal) in Western Australia, probing between $\sim 10- \sim 480$ MHz. Initial results were published in 2015 (Sokolowski et al., 2015).
- SCI-HI (Sonda Cosmológica de las Islas para la Detección de Hidrógeno Neutro) has been operating since 2013 in Mexico in the 40 - 130 MHz range (Peterson et al., 2014; Voytek et al., 2014).
- The Probing Ratio Intensity at high-Z from Marion ($\text{PRI}^{\text{Z}}\text{M}$) experiment recently went online on a small island off the coast of South Africa, and will probe the $\sim$30 - $\sim$200 MHz ($z \approx 6 - 47$) range (Philip et al., 2018).
- SARAS 2 in India operates between $110 - 200$ MHz, with first results in 2017 (Singh et al., 2017).
- Perhaps the most well known is the Experiment to Detect the Global EoR Signature (EDGES), operating in Western Australia (alongside the MWA) in the $\sim 50- \sim 200$ MHz ranges ($z \approx 6 - 27$) (Bowman et al., 2008). In addition to being one of the older apparatuses (online since 2006, although the 50 - 100 MHz low-band capabilities were only added in 2015), as mentioned above EDGES was the first to claim a detection in early 2018 (Bowman et al., 2018), although Hills et al. (2018) warns this could be due to improper foreground modelling (see section 0.5).

Thus, while the EDGES results are awaiting confirmation, the most heavily redshifted neutral hydrogen signal that has been detected to date (in emission) is that originating in nearby galaxies, out to roughly $z \approx 0.1$ (Geréb et al., 2015). Attempts have been made to use gravitational lensing to detect HI line emission out to $z \approx 0.5$ (Hunt et al., 2016). Switzer et al. (2013) have also constrained 21 cm fluctuations out to $z \approx 0.8$ through intensity mapping.

[14] Although calibration is still very difficult.

Upcoming Experiments

In addition to these operating efforts, a number of additional experiments are currently in preparation, or being constructed.

- The upcoming Hydrogen Epoch of Reionization Array (HERA) is being constructed in South Africa (one kilometre from the site of the future SKA). The array aims to bring more sensitivity to the angular and spectral scales where PAPER and MWA have suggested 21 cm signal is most likely to be detected (DeBoer et al., 2017).

- 2019 is also expected to see the completion of NENUFAR (New Extension in Nançay Upgrading LOFAR), which is being built in Nançay, France. The array will operate between $10 - 80$ MHz, and in addition to being a standalone Cosmic Dawn experiment it will act as a second core for LOFAR (Zarka et al. (2012); www.nenufar.obs-nancay.fr).

- There are also two exciting upcoming lunar experiments. The ambitious Netherlands-China Low-Frequency Explorer (NCLE) was put into place at the Earth-Moon L2 point in May 2018, and will operate between $0.08- \sim 80$ MHz. Very recently, a second space based experiment named DARE has been announced. It will observe in the $\sim 40- \sim 120$ MHz range (Burns et al., 2017); www.isispace.nl/projects/ncle/.

The Square Kilometre Array

The above experiments are all considered precursors to the Square Kilometre Array[15]. The SKA, expected to be operational in the mid 2020s, will consist of up to 500,000 antennas. These will be organized into two core sites: SKA-Low (50-350 MHz) in Western Australia (where the MWA and EDGES are currently operating) and a second in South Africa (where MeerKAT[16] and HERA will be operated). In addition, there will be a number of outlying stations situated in other African nations to further increase the baseline coverage. The resolution (in the range of 50 - 350 MHz, with a FoV of $\sim 4°$) is estimated to be ~ 7 arcsec for these long baselines SKA-Low (Huynh & Lazio, 2013). However it should be noted that there are very few of these long baselines, and therefore the noise will be much too high to detect the 21 cm signal at these resolutions. 21 cm detection will depend primarily on the core cluster of antenna[17], and should be achievable on the order of $\sim$ a few arcmins, depending on redshift (for realistic resolution estimates for 21 cm tomography see Koopmans et al. (2015)).

[15] With the exception of HERA, which is expected to also contribute excellent stand-alone tomography, and is therefore more of a competitor.

[16] Holwerda et al. (2012).

[17] This isn't to say that the long baselines are useless. They will allow for more detailed modelling of extragalactic point sources, which will help with foreground subtraction (section 0.5), and will also be useful for other SKA science goals.

In 2012 the Australian Square Kilometre Array Pathfinder (ASKAP) was completed on the future SKA site (specifically the SKA-low site in Australia). The array currently consists of 36 telescopes, and this number will eventually be increased to 60. The key advantage of ASKAP is its use of Phased-Array Feeds (PAFs) to simultaneously observe via a number of different beams, thus increasing the FoV and improving survey speed (Hay & O'Sullivan, 2008). As of 2018, early results and pilot surveys have been published (Hobbs et al., 2016; Kohler, 2017), and the PAHs are currently being upgraded to allow even faster survey speeds (Schinckel & Bock, 2016).

Summary of EoR Experiments

Table 1 – *Summary of EoR experiments*

Name	Type	Location	ν (MHz)	z	Date	Reference
GMRT	Fluctuations	India	~ 50 - ~ 1500	~ 0 - ~ 28	1995	Paciga et al. (2013)
EDGES	Global	Australia	~ 50 - ~ 200	~ 6 - ~ 27	2006	Bowman et al. (2008)
PAPER	Fluctuations	S. Africa	~ 100 - ~ 200	~ 7 - ~ 14	2009	Parsons et al. (2010)
LEDA	Global	USA	~ 30 - ~ 88	~ 16 - ~ 34	2011	Price et al. (2018)
MWA	Fluctuations	Australia	~ 80 - ~ 300	6.2 - 11.7	2012	Tingay et al. (2013)
LOFAR	Fluctuations	Europe	~ 10 - ~ 240	6 - ~ 140	2012	Patil et al. (2017)
SCI-HI	Global	Mexico	~ 40 - ~ 130	~ 11 - ~ 36	2013	Peterson et al. (2014)
BIG-HORNS	Global	Australia	~ 50 - ~ 250	~ 6 - ~ 30	2015	Sokolowski et al. (2015)
PRIZM	Global	Marion Is.	~ 30 - ~ 200	~ 7 - ~ 50	2017	Philip et al. (2018)
SARAS 2	Global	India	~ 110 - ~ 200	~ 7 - ~ 13	2017	Singh et al. (2017)
NCLE	Global	Space	0.08 - ~ 8	~ 18 - large	2019	www.isispace.nl /projects/ncle/
NENU - FAR	Fluctuations	France	10 - 87	~ 18 - ~ 140	2019	www.nenufar.obs -nancay.fr
HERA	Fluctuations Tomography	S. Africa	~ 50 - ~ 250	~ 6 - ~ 30	2020?	DeBoer et al. (2017)
SKA	Fluctuations Tomography	Australia - Africa	~ 50 - ~ 350	~ 4 - ~ 30	2025?	Maartens et al. (2015)
DARE	Global	Space	~ 12 - ~ 120	~ 4 - ~ 35	?	Burns et al. (2017)

Recent Theoretical Developments

In the past two decades, in addition to continued attempts at observing the EoR through 21 cm radiation, there has been significant effort towards perfecting the theoretical framework. This includes developing the mathematics necessary to robustly predict dynamics of the epoch, as well as simulating the expected radiation patterns. The interplay between observation, theory, and simulation is what drives the domain forward.

21 cm Radiation in Absorption and Emission

Throughout the Dark Ages and the Epoch of Reionization, the 21 cm signal is always observed against the background of CMB radiation. That is to say, when we study the blackbody emission of the CMB, 21 cm radiation is measured as a difference from the expected CMB intensity. This difference can manifest itself as either an excess (when the hydrogen clouds through which the CMB is passing contain a suprlus of atoms in the upper hyperfine level, and are emitting 21 cm radiation), or an absence (when the hydrogen clouds are instead absorbing CMB photons). Thus, throughout the cosmic history of the Universe, we describe the signal as being in either emission or absorption for given periods.

The evolution of the 21 cm signal is related to a complex interplay of gas dynamics, the evolution of the CMB, and the expansion of the Universe. Truly preparing for 21 cm cosmology involves understanding this interplay and predicting the evolution of the signal as a function of redshift, and many groups have attempted to do just this (Field 1959a; Gnedin & Shaver 2004; Furlanetto et al. 2006; Iliev et al. 2006; Mellema et al. 2006; Baek et al. 2009; Alvarez et al. 2010; Semelin et al. 2017 — some of these works are theoretical, others are numerical; the list is certainly non-exhaustive). We now describe efforts towards this goal.

Estimating the Brightness Temperature

In 2006, Steven Furlanetto, Peng Oh, and Frank Briggs carried out a comprehensive re-evaluation and exploration of 21 cm cosmology (Furlanetto et al., 2006). Of note was the introduction of a new formula (p.23 in Furlanetto et al. 2006) for calculating the expected 21 cm brightness temperature fluctuations. It is worthwhile outlining the derivation of this formula, as it will form the core of this thesis. Let us therefore start by stating that 21 cm radiation has an associated optical depth τ_{21}. This optical depth will be defined as $\tau_{21} = \ln \frac{\Phi^i_{21}}{\Phi^t_{21}}$, in which Φ^i_{21} and Φ^f_{21} are the initial and final 21 cm radiant fluxes, respectively. Rearranging this relation we arrive at:

$$e^{-\tau_{21}} = \frac{\Phi^t_{21}}{\Phi^i_{21}}. \tag{2}$$

So $e^{-\tau_{21}}$ represents the fraction of transmitted 21 cm radiation. With this result in memory, we turn to the intensity for a given frequency ν, represented as $I(\nu)$. This intensity will be equal to the CMB radiation I_γ that is transmitted through the gas ($I_\gamma e^{-\tau_{21}}$) plus the integral of all attenuated emission in the IGM along the line of sight (which is transmitted). So we arrive at the equation:

$$I(\nu) = I_\gamma e^{-\tau_{21}} + \int dI_{HI} e^{-\tau_{21}(z')} \tag{3}$$

Computing the integral and switching to temperature brightness (using the relationship in the Rayleigh-Jean regime) we arrive at:

$$T_b(\nu) = T_S(1 - e^{-\tau_{21}}) + T_\gamma e^{-\tau_{21}} \tag{4}$$

What is left is now to calculate the 21 cm optical depth. The optical depth can be defined as:

$$\tau_{21} = \int_0^\infty \Phi(\nu)\sigma_{01} n_0 dl \tag{5}$$

where σ_{01} is the cross section of the 21 cm transition, n_0 is the number density for hydrogen in the lower state, and $\Phi(\nu_g)$ is the line profile[18] (peaked at ν_{21} with some slight width to accommodate the thermal motion of the gas clouds, and normalized such that $\int_0^\infty \Phi(\nu_g) d\nu_g = 1$). The integration is carried out along the line of sight (and hence dl is the line element). We also must consider the contribution to the optical depth from spontaneous emission (identical to equation 5, but with $\sigma_{01} \to \sigma_{10}$ and $n_0 \to n_1$). So, combining the absorption and emission within the gas we arrive at:

$$\tau_{21} = \int_0^\infty \Phi(\nu_g)(\sigma_{01} n_0 - \sigma_{10} n_1) dl \tag{6}$$

we can now switch from cross-sections to Einstein B coefficients (Herzberg, 1950; Hilborn, 2002) using the relation:

$$\sigma_{ij} = \frac{h\nu_g B_{ij}}{4\pi} \tag{7}$$

We now arrive at:

$$\tau_{21} = \int_0^\infty \frac{h\nu_g}{4\pi} \Phi(\nu_g)(B_{01} n_0 - B_{10} n_1) dl \tag{8}$$

We wish to integrate in terms of frequency, and must therefore compute the conversion between dl and $d\nu_g$. To begin this, we can start by $dl = c \cdot dt$, and now multiplying the top and bottom by an element of the scale factor we get $dl = (c \cdot dt \cdot da)/da = (c \cdot da)/\dot{a}$. Now, we note that the ν_g used above is the frequency in the *rest frame of the gas* (which is assumed to have peculiar velocity along the line of sight, denoted as $v_\parallel$). Let's define ν_z to be the frequency for a given redshift z, and ν_{obs} as the observed frequency on earth. The two will be related through the scale factor, explicitly as $\nu_z = \nu_{obs}/a$, and deriving this expression gives us $d\nu_z = -(\nu_{obs}/a^2)da$. Substituting da for the above equivalency,

[18]Note that ν_g is the frequency in the rest frame of the gas.

and isolating dl, we arrive at:

$$dl = -\frac{ca^2}{\nu_{obs}\dot{a}}d\nu_z = -\frac{ca}{\nu_{obs}H}d\nu_z \tag{9}$$

We have here substituted the Hubble factor ($H = a/\dot{a}$). Now we must take into account peculiar velocity of the gas to determine the corrected value of ν_g. We begin with $\nu_g = \nu_z(1 - v_\parallel/c)$, which can then be derived to give $d\nu_g = d\nu_z(1 - v_\parallel/c) - (\nu_z \cdot dv_\parallel)/c = d\nu_z(1 - v_\parallel/c) - (\nu_z \cdot dv_\parallel \cdot dl)/(c \cdot dl)$. Rearranging and plugging in equation 9 we arrive at:

$$d\nu_g = d\nu_z\left(1 + \frac{1}{H}\frac{dv_\parallel}{dl}\right) \tag{10}$$

Note that the $v_\parallel/c$ term is assumed to be $\ll 1$. Combining equations 9 and 10 gives:

$$dl = -\frac{ca}{H\nu_{obs}\left(1 + \frac{1}{H}\frac{dv_\parallel}{dl}\right)}d\nu_g \tag{11}$$

And now that the peculiar velocity is also taken into account, we can substitute this into equation 8[19]:

$$\tau_{21} = -\frac{h\nu_g}{4\pi}(B_{01}n_0 - B_{10}n_1)\frac{ca}{H\nu_{obs}\left(1 + \frac{1}{H}\frac{dv_\parallel}{dl}\right)}\int_\infty^{\nu_0}\Phi(\nu_g)d\nu_g \tag{12}$$

Because $\Phi(\nu_g)$ is normalized, the integral solves to 1 (note we are taking the 'CMB to telescope' is positive for the integral bounds). To simplify the B coefficients, we turn attention to the Zeeman effect: when the spin of the electron and proton are in parallel (the higher energy state) the presence of a magnetic field causes the upper hyperfine transition to split into three energy levels (Zeeman, 1897), which results in $B_{01} = 3B_{10}$. The relation is similar for the number densities of the n_0 and n_1 populations, however these will also be effected by the relative strength of the spin temperature to the 21 cm excitation temperature $T_\star$ ($\sim$0.068 K). Explicitly we have $n_1/n_0 = 3e^{-T\star/T_S}$, and because $T_\star \ll T_S$ we then have $n_1/n_0 \approx 3(1 - T_\star/T_S)$. Applying these equalities, we arrive to:

$$\tau_{21} = -\frac{h\nu_g}{4\pi}B_{01}n_0\frac{ca}{H\nu_{obs}\left(1 + \frac{1}{H}\frac{dv_\parallel}{dl}\right)} \tag{13}$$

A few small tweaks are left. $n_{HI} = n_0 + n_1$ (both states), and recall $n_1 \approx 3n_0$, so $n_0 = n_{HI}/4$. As well, the B coefficients are defined as $B_{ij} = c^2/A_{ij}/2h\nu_o bs^3$, $T_\star = h\nu_o bs/k_B$, and $a = 1/(1 + z)$:

$$\boxed{\tau_{21} = -\frac{3hc^3\nu_g n_{HI}A_{10}}{32\pi(1+z)HT_S\nu_{obs}^2}\left(1 + \frac{1}{H}\frac{dv_\parallel}{dl}\right)^{-1}} \tag{14}$$

[19] At this point, we will remove many terms from the integral. This is done by assuming that they vary only negligibly along the line of sight on the spacial scale corresponding to the thermal width of the line profile, Φ. See Appendix A for a verification of this assumption.

Now that we have a definition of 21 cm optical depth, we can differentiate equation 6, and then add a $1/(1+z)$ term to account for the redshift scaling of the T_S and T_γ terms:

$$\delta T_b(z) = \frac{T_S - T_\gamma}{1+z}(1 - e^{-\tau_{21}}) \approx \frac{T_S - T_\gamma}{1+z}\tau_{21} \tag{15}$$

Plugging in τ_{21} from equation 14, using $H(z) \approx H_0\Omega_m^{1/2}(1+z)^{3/2}$, and condensing all constant terms, we arrive at:

$$\boxed{\delta T_b(z) \approx 9x_{\rm HI}(1+\delta)(1+z)^{1/2}\left[1 - \frac{T_\gamma(z)}{T_S}\right]\left[\frac{H(z)/(1+z)}{dv_\parallel/dr_\parallel}\right] \text{ mK}} \tag{16}$$

where $x_{\rm HI}$ is the neutral fraction, $(1+\delta)$ is the fractional overdensity of baryons, $T_\gamma(z)$ is the CMB brightness temperature, $T_S(z)$ is the 21 cm spin temperature[20] for a given redshift, $H(z)$ is the Hubble constant at said redshift, and $dv_\parallel/dr_\parallel$ is the gradient of proper velocity along the line of sight.

Equation 16 highlights the intimate relationship between the CMB temperature and that of the 21 cm radiation. For example, we remark that if $T_S(z) \to T_\gamma(z)$ then $\delta T_b(\nu) \to 0$. To put this conceptually, if the CMB and 21 cm radiation (from the IGM) are observed at equal temperature, then we cannot distinguish the 21 cm signal, which will otherwise appear as either an excess or lack of temperature brightness. We can also explore the two limits. If $T_\gamma(z) \ll T_S(z)$ then $\delta T_b(\nu)$ will approach a fixed positive value. Conversely, if $T_\gamma(z) \gg T_S(z)$ then $\delta T_b(\nu)$ will become increasingly negative.

The spin temperature T_S for a cloud of hydrogen is determined by three factors. Firstly, the kinetic temperature of the cloud (T_K) decides the probability of collisions between atoms, free electrons, and protons. These collisions can induce spin-flips, and hence the release of 21 cm radiation. Secondly, the temperature of the background CMB (T_γ) will effect the 'visibility' of 21 cm radiation, whose intensity is effectively the difference of said radiation from the CMB. Thirdly, a similar absorption/re-emission effect can be caused by the scattering of UV photons from the Lyman Alpha background (T_c, c standing for colour). Together, the interplay of these three effects gives us a succinct expression for spin temperature (p. 24 of Furlanetto et al. 2006):

$$T_S^{-1} = \frac{T_\gamma^{-1} + x_c T_K^{-1} + x_\alpha T_c^{-1}}{1 + x_c + x_\alpha} \tag{17}$$

where x_c is the collisional coupling for T_S, and x_α is the Wouthuysen-Field coupling for T_S. This theoretical framework contributed to many important advances of the next decade (Barkana & Loeb 2007; Mesinger & Furlanetto 2007; Pritchard & Furlanetto 2007; Lidz et al. 2008; Meiksin 2009; Baek et al. 2010; Brandenberger et al. 2010; Mesinger et al.

[20] Also sometimes called the excitation temperature, it can be thought of at follows. For a quantity of hydrogen atoms, in which a certain ratio of the atoms are in the parallel spin configuration and the remainder are in the anti-parallel spin configuration, what is the expected temperature. This temperature will be highest if all atoms are all in the parallel configuration.

2011; Mellema et al. 2013; Fialkov et al. 2014, etc.), and is still a relevant overview of the topic.

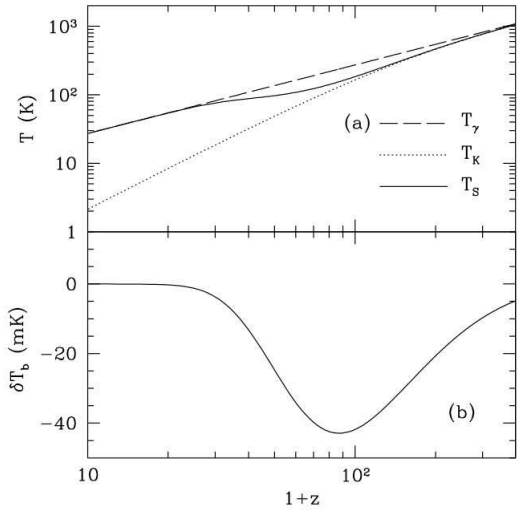


Figure 6 – *Top panel: Evolution of the gas, CMB, and spin temperatures (T_K, T_γ, T_S) as a function of redshift. Bottom panel: The resulting average temperature brightness (δT_b) relative to the CMB as a function of redshift. Reproduced from Furlanetto et al. (2006), page 47.*

The redshift evolution of the spin temperature relative to the kinetic gas and CMB temperatures is shown in figure 6 (reproduced from Furlanetto et al. 2006, p. 47). The authors assumes only adiabatic cooling and Compton heating are involved. At $z \approx 300$ T_S and T_K decouple from T_γ (as Compton heating becomes inefficient) and continue cooling adiabatically. Then, at $z \approx 70$, as the gas density continues to drop, collisional coupling begins to become ineffective in coupling T_S to T_k, and T_S slowly approaches T_γ, with re-coupling essentially complete at $z \approx 30$. Therefore, 21 cm radiation should manifest as absorptions within the CMB at $z \gtrsim 30$, and is predicted to go as deep as a few tens of mK. This initial absorption profile could prove useful for studying the Dark Ages, though it is expected to have ended before the beginning of the EoR, and is expected to be too faint to study in the foreseeable future.

Thankfully, figure 6 presents only half of the story. Recall that only adiabatic cooling and Compton scattering were included. In reality, the period $z \lesssim 30$ introduced many new dynamics to the forming structure of the early Universe. The birth of the first stars must have begun heating the gas, eventually to the point of surpassing the CMB ($T_k < T_\gamma$). As well, at some point, the spin temperature must have recoupled to the gas temperature through Lyα coupling (via the Wouthuysen-Field effect, see Wouthuysen 1952; Field 1959b).

The exact redshift at which these two event occurred remains uncertain. This is because the point at which T_k surpassed T_γ depends on X-ray production rates, and T_S recoupling depends on the star formation rate and escape fraction. All of these quantities are not well understood at high redshifts. If recoupling occurred while $T_K < T_\gamma$, then

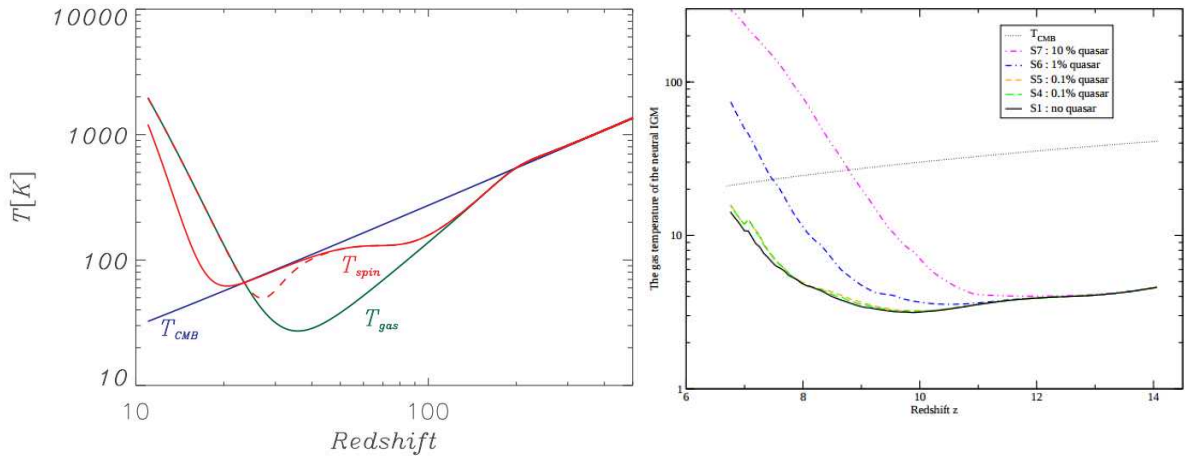


Figure 7 – *Left: Evolution of the gas, CMB, and spin temperatures, with the inclusion of early luminous sources, as a function of redshift. The dashed red line assumes the spin temperature and the gas temperature recoupled while the gas was colder than the CMB. The solid red line assumes the inverse. Note that T_{gas} is T_K, T_{CMB} is T_γ, and T_{spin} is T_S. Reproduced from Zaroubi (2013), page 29. Right: Same, but excluding T_S, zoomed in z to focus on the EoR, and including various quasar luminosity fractions. Reproduced from Baek et al. (2010), page 6.*

it is expected that a second T_S absorption period would have formed, followed by a transition into emission. Alternately, the spin temperature may have recoupled to the gas temperature only when $T_K \geq T_\gamma$, in which case the spin temperature would be visible only in absorption for $z \lesssim 30$. This extended history is illustrated in figure 7, reproduced here from Zaroubi (2013).

Tracing the Neutral Fraction

The existence of 21 cm emission (and possibly a secondary period of absorption) at redshifts below 30 gives us a realistic chance of detection. Although the signal becomes strong in emission as the first stars begin to heat the gas[21], there is another limiting factor. In heating the gas, the first stars and galaxies also begin to reionize the neutral hydrogen. Through this, there is less and less neutral hydrogen from which the signal can be emitted. Hence why, for the recent history of the Universe $z \lesssim 6$ we do not see any 21 cm radiation originating in the intergalactic medium (IGM)[22].

Therefore, we have no way (besides observations) of knowing the *exact* behaviour of the 21 cm brightness temperature, largely due to gaps in our understanding of how exactly reionization played out. Depending on the relative importances of various sources in reionizing the IGM, the neutral fraction will have evolved differently. This will, in turn, effect the strength of the brightness temperature.

[21] Although 'strong' is relative. Emission saturates at a few tens of mK, while absoprtion can be larger than 100 mK in amplitude.

[22] This can also be seen from equation 16. For small z, the neutral fraction $x_{\rm HI} \approx 0$, and hence $\delta T_b \to 0$.

So the goal therefore becomes to constrain the progression of the EoR. This can be carried out through different techniques, some of which we will now present in broad theoretical terms (as well as giving a brief mention to the accompanying observational prospects, despite the aim of the current section being an overview of recent 'theoretical' developments).

Intensity Mapping

As mentioned above, although there is negligible 21 cm in the interstellar medium at low redshifts, there is still 21 cm from neutral hydrogen lingering in galaxies. Capturing HI emission from an individual galaxy, even in the local universe, is challenging (as stated in section 0.3.2). Yet, collectively, the 21 cm contribution from millions of galaxies is measurable (and mappable with large pixels). This procedure of taking a more statistical approach to the neutral hydrogen content (of the low redshift Universe) is called intensity mapping.

The procedure was proposed first for the EoR (Madau et al., 1997), and later for the post-ionization universe (Bharadwaj & Sethi, 2001; Peterson et al., 2009). It has since been applied successfully out to redshift 0.8 (Chang et al., 2010), with upcoming experiments hoping to map Baryon Acoustic Oscillations (Ansari et al., 2012), and push closer towards the EoR (Newburgh et al., 2014; Smoot & Debono, 2017). Even at low redshifts, intensity mapping is an excellent tool for extrapolating what the neutral hydrogen content of the Universe was at higher redshift. If extended to EoR redshifts, the implications would be even more substantial (see Kovetz et al. 2017 for an overview of current developments).

Lyman Alpha Emitters

The early universe can also be studied through Lyman Alpha Emitters (LAE). These consist of young luminous stars in active star forming regions, as well as quasars, which reside within hydrogen clouds. Their radiation ionizes the surrounding hydrogen, yet the electrons are recaptured and trickle down the Lyman lines. The dominant radiation is Lyα, emitted at 1216 Å, which therefore falls in the optical range up to $z \lesssim 6.5$, beyond which it is observed in IR. Because of the strength of the line, LAE have been observed out to the EoR (Nilsson, 2007; Ono et al., 2012; Larson et al., 2018). This allows us to study the LAE luminosity function, which in turn acts as a tracer for reionization. Towards these efforts, sufficiently resolved EoR simulations help us understand the behaviour of early LAE (Zheng et al., 2010; Inoue et al., 2018).

Gunn-Peterson Effect

We can also study the EoR via quasars, specifically via the Gunn-Peterson trough (Gunn & Peterson, 1965), in which the electromagnetic radiation of EoR quasars is damped at

wavelengths below the Lyα line due to neutral hydrogen absorption along the line of sight. Whether or not we see structure past the cut-off is an indication of the ionized fraction at the corresponding redshift. These observations must be bolstered by projects to understand the nature of high redshift quasars (for example, the quasar survey presented in Fan et al. 2006b).

Galaxy Luminosities

Distant galaxies are another piece of the puzzle. The current most distant galaxy is found at $z \approx 11$ (Oesch et al., 2016), well into the expected period of reionization. Yet as we peer into higher redshifts, we see fewer and fewer galaxies. As per the Halo Mass Function towards high-z, they simply did not have enough time to form, especially massive bright ones that could have been detected. Observational data sets of these distant sources — specifically how many we count at each redshift — therefore help us map the EoR. A number of contemporary high-redshift galaxy surveys are rapidly adding to our inventory of EoR galaxies (see Salmon et al. 2018 p.5 for a summary). Figure 8 presents an example of our current estimates on the redshift evolution of the luminosity function (reproduced from Bouwens et al. (2016)).

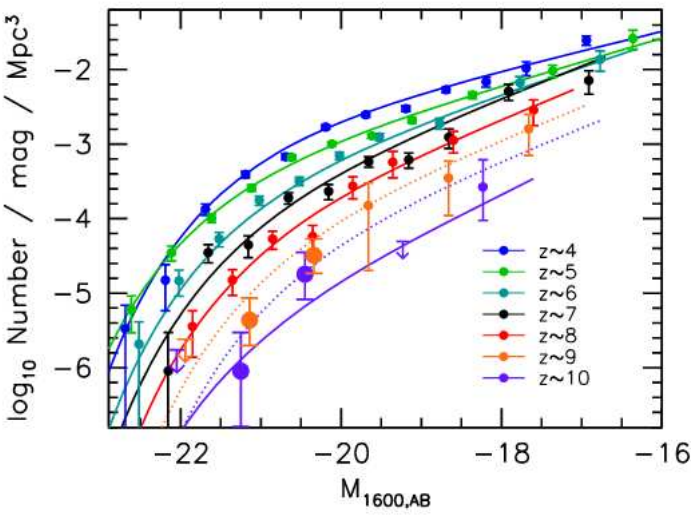


Figure 8 – *Evolution of the galaxy luminosity function as a function of z. For example, we should expect to find ∼1 magnitude -18 galaxy per 10^4 Mpc^3 at redshift $z = 10$. Reproduced from Bouwens et al. (2016), page 14.*

An estimate of the progression of reionization is presented in figure 9. This is based on the above-mentioned evidence (quasar dampening, LAE). 'Dark pixels' refers to the fraction of the Lyα and Lyβ forest that does not emit (McGreer et al., 2015), which is a direct tracer of the neutral fraction. As well, the CMB optical depth also provides rough boundaries on the EoR progression (Planck Collaboration et al., 2016b).

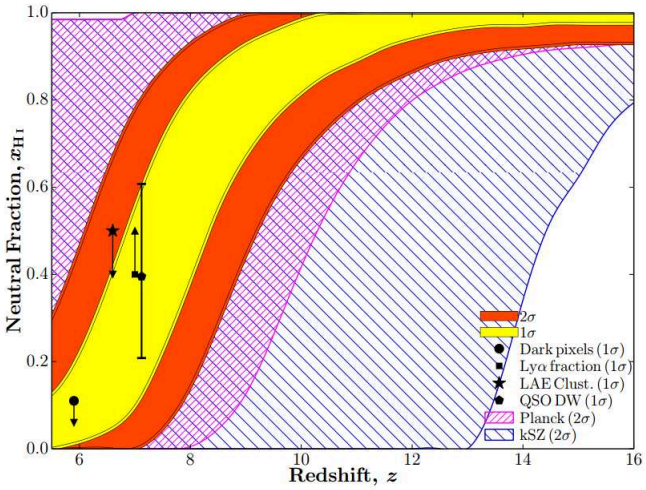


Figure 9 – *Evolution of the neutral fraction of the Universe during the EoR as a function of z. Points indicate known constraints, such as distant Qasars and Lyα clustering. Reproduced from Greig & Mesinger (2017a), page 13.*

Noise Estimation

For the time being, to truly constrain the EoR, and thus the expected intensity of the 21 cm signal, we require deeper surveys (such as those made possible with the JWST, which will not observe the 21 cm signal, but will still help us constrain the EoR through studying early galaxies). Another large hurdle is the amount of unwanted contamination which must be taken into consideration. There are various sources of this contamination, including systematics (such as surface errors and other mechanical imperfections), as well as bright foregrounds.

In the case of a radio-interferometer, the thermal noise can be calculated (see derivation in Taylor et al. 1999, chapter 9) as:

$$\Delta T_{b,noise} = \frac{\text{SEFD}}{\eta_s \sqrt{N(N-1)\Delta\nu t_{\rm int}}} \text{ mK} \tag{18}$$

where SEFD is the 'system equivalent flux density' in Jy (SEFD $= \frac{T_{sys}}{K/Jy}$, where T_{sys} is the system temperature[23]), η_s is the system efficiency, N is the number of substations, $\Delta\nu$ is the frequency bandwidth, and $t_{\rm int}$ is the integration time. To test this theory within the context of real observations, let us refer to Jelić et al. (2008), which used reasonable LOFAR values (importantly, this is at a *resolution of 3"*, very different from EoR 21 cm observations, which will be on the order of arcmins) to predict instrumental thermal noise on the order of 500 mK at 150 MHz with a 1 MHz bandwidth in a single night. However, this could be reduced as low as 52 mK if data was instead collected

[23]Which includes contributions from receiver noise, feed losses, spillover, atmospheric emission, galactic background and cosmic background. See derivation in Crane & Napier 1989.

over one hundred nights. Was this estimation correct? We can now compare with actual noise, which is given in Patil et al. (2016) to be 0.9 mJy for a night of observation at 150 MHz ($\Delta\nu = 195kHz$). This can be converted to temperature brightness using the Rayleigh-Jeans law:

$$T_b = \frac{\lambda^2 \cdot S}{2k_B\Omega} = 1.222 \times 10^3 \frac{I}{\nu\theta^2} \tag{19}$$

where λ is the wavelength, k_B is the Boltzmann constant, Ω is the beam solid angle, S is the flux density, I is the brightness (Jy/beam), ν is the frequency (GHz), and θ is the beam angle in arcsec (we assume here a spherical beam, otherwise $\theta = \theta_{maj} \cdot \theta_{min}$)[24]. Using the above values we arrive at $T_{b,noise} \approx 5.4$ K, but we must remember that this is at a bandwidth of 195 kHz, one fifth of the value used for the theoretical noise calculation. Noise scales as $1/\sqrt{\Delta\nu}$, or roughly $1/\sqrt{5}$ in our case, which brings us to $T_{b,noise} \approx 2.4$ K. This is five times larger than what was predicted, and not only gives an indication of the difficulty of overcoming noise, but also tells us that observations will need to be carried out over many nights.

There are other sources of contamination caused by all that lies between the earth and the primordial Universe. Unfortunately, these tend to be much more complicated to model and understand. Therefore, alongside efforts to improve theory and instrumentation, another branch of EoR research has focused on overcoming the challenge of 'foregrounds'.

Foregrounds

As stated, it was realized quite early on that a major hindrance towards detecting primordial 21 cm radiation is the multitude of much stronger signals that dominate the sky. Notably, synchrotron and free-free emission from our Galaxy are estimated to be up to three orders of magnitude brighter than the 21 cm radiation we hope to measure (Bonaldi et al., 2014; Bowman et al., 2018). Beyond Galactic sources, distant radio galaxies also contribute to the measured signal. Even CMB radiation (though a background rather than a foreground) plays a part[25]. Thus, it should come as no surprise that a considerable effort to understand and remove radio foregrounds has taken shape over the past decade. This section aims to present a quick overview of said efforts, and the list of publications and topics relating to foregrounds is in no way exhaustive.

[24] A full derivation of arriving at the right hand form of the equation can be found at science.nrao.edu/facilities/vla/proposing/TBconv.

[25] Although the CMB peaks in the radio spectrum, the Rayleigh-Jeans tail of the CMB still contributes significant brightness down to low frequencies, and dominates at 1400 MHz. However, for EoR studies, the 21 cm signal has been redshifted to the point that CMB is no longer a significant issue e.g. 233 MHz at $z = 6$.

Diffuse Foregrounds

One of the first investigations into the full range of diffuse radio foregrounds was Oh & Mack (2003). The difficulty expected in dealing with foregrounds was quickly made evident, though it was also noted that the unchanging nature of free-free and synchrotron emission in frequency space could provide a potential tool towards extracting the 21 cm signal. To expand on this point, free-free and synchrotron emissions are released across many frequencies. Conversely, 21 cm radiation is always released at a discrete frequency (1420 MHz), and any changes in this frequency can only be due to redshift[26] (or proper velocity). A visualization of this is shown in figure 10.

de Oliveira-Costa et al. (2008) presents a model of the Galactic radio sky between 10 MHz and 100 GHz. Though not intended specifically for EoR foreground subtraction, they nonetheless prove useful towards this goal. The maps included best estimates for expected synchrotron, dust, and free-free foregrounds, as well as intragalactic point sources. The same year, Jelić et al. (2008) presented simulated radio maps for $5^\circ \times 5^\circ$ patches of sky. They included Galactic free-free, synchrotron, and supernova remnant emissions, as well as that from radio galaxies and radio clusters. In addition, said maps also included instrumental polarization response, which the previous collaboration had not. Chapman et al. (2015) summarizes some of the recent developments on foreground modelling.

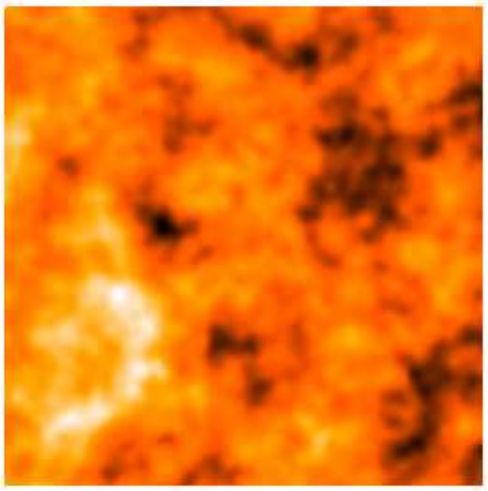
Diffuse Foreground Subtraction Methods

With increasingly accurate foreground and 21 cm simulation maps (Jelić et al. 2008; Santos et al. 2008; Thomas et al. 2009; Mesinger et al. 2011; Iliev et al. 2014; Semelin et al. 2017, etc.) it became possible to superimpose simulated 21 cm maps with foregrounds, and then attempt EoR signal extraction. Among the proposed techniques, some were novel, while others were based on previous subtraction methods that had been developed towards cleaning CMB maps. Presented here is concise overview of a number of different foreground removal techniques.

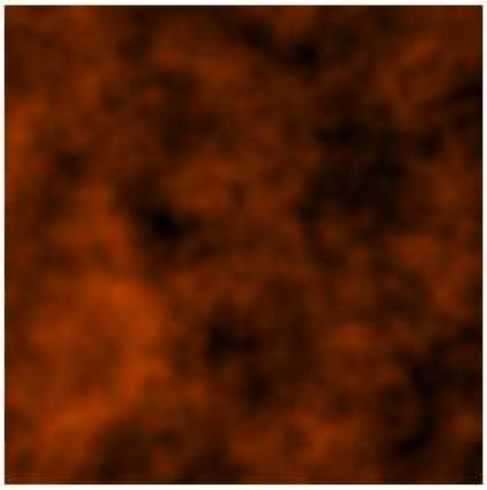
Native Subtraction Models

There have been a number of subtraction methods developed explicitly to deal with 21 cm radiation. In some sense, all techniques are founded on the idea of fitting a smooth frequency function to each line of sight, modelling diffuse foregrounds (free-free and synchrotron), and then subtracting. This method relies on the spectral smoothness of foregrounds, a property not shared by the EoR signal (figure 11, which can be thought of as looking at a single pixel in figure 10 across all frequencies). Based on 21 cm cubes created by Ciardi & Madau (2003), this extraction method was attempted by Jelić et al. (2008).

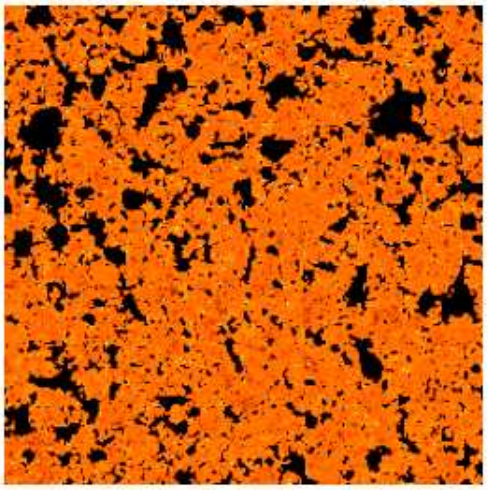
[26] Although it is obviously not this clean cut, with free-free and synchrotron also exhibiting some changes in frequency space (Ali et al., 2008)



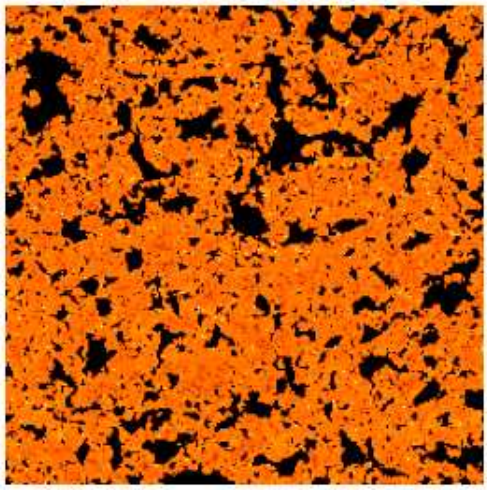
(a) Free-Free - 100 MHz



(b) Free-Free - 200 MHz



(c) HI - 119 MHz



(d) HI - 120 MHz

Figure 10 – *Comparison of the frequency dependence of Free-Free radiation (a,b) and HI radiation (c,d). From 100 to 200 MHz Free-Free displays very little difference in shape, only in brightness. HI has significant differences, even across 1 MHz. Free-Free maps courtesy of Anna Bonaldi. 21 cm maps made using 21cmFAST (Mesinger et al., 2011). All maps 200 cMpc.*

It was assumed that $\Delta T_b = 2$ K for foregrounds and $\Delta T_b = 5$ mK for the EoR signal. The extraction was carried out by first calculating three power laws in frequency space. These power laws represented the expected emission from Galactic synchrotron, Galactic free-free, and extragalactic emissions. The 'dirty map' (comprising EoR signal as well as foregrounds) was then partitioned by fitting a third order polynomial representing the three power laws, as well as a background constant equation, representing the remaining EoR radiation. When the polynomial was fit, it could then be subtracted to reveal the underlying hydrogen radiation. It was shown that the subtraction method could detect the signal down to 52 mK (although limited by noise, such as that of the interferometer). This was an encouraging result, as it achieves a sensitivity on the order of magnitude at which 21 cm radiation is expected to manifest.

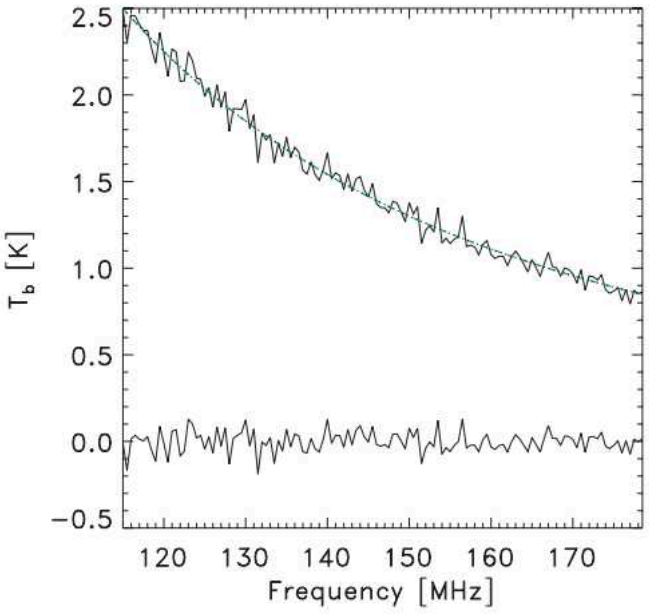


Figure 11 – *Temperature brightness fluctuations along the line of sight. The upper black line is what is observed, the dotted blue and dotted black lines are the smoothed and fitted foregrounds respectively, and the lower black line is the residual signal (EoR + Noise). Reproduced from Jelić et al. (2008).*

Harker et al. (2009) built upon this early effort, concluding that the process of fitting (parametrizing) a power law to data entailed some unwanted consequences. It was argued that choosing a power law description for a foreground source is inherently arbitrary. Specifically, overfitting the data with high order polynomials smooths some of the desired EoR signal, and underfitting is insufficient to properly include all features in the fit. The proposed solution is a method called Wp[27] smoothing, which aims to minimize the change in a fit's curvature, as opposed to simply the distance between the data and the curve.

The author concludes that[28], Wp smoothing easily outperforms other non-parametric cleaning algorithms, and is competitive with many parametric cleaning algorithms – even

[27] As explained in the article, 'Wp' is short for the German 'Wendepunkt', meaning 'Inflection Point'.
[28] When the smoothing parameter λ is correctly chosen, see Harker et al. (2009) section 4.1.

though, in the test cases, these have advanced knowledge of foreground structure. This makes Wp smoothing a strong candidate for handling future foreground subtraction.

More recently, work has been carried out into using wavelet filtering based on the shape of the 21 cm signal's jagged features in frequency space (Gu et al., 2013). There has also been interest in overcoming polarized foregrounds (Moore et al., 2013), a topic that had not been previously studied in much depth, especially with respect to frequency space. After attempting to clean 21 cm maps by individually removing polarized foreground sources, the technique proves to be both inefficient, and insufficient to mitigate contamination temperature brightness below that of the 21 cm radiation. Subsequent efforts have strived to better model and subtract polarized foregrounds (Spinelli et al., 2018; Dillon et al., 2018).

CMB Extraction Techniques

In addition to these foreground subtraction techniques, others have instead focused their efforts on modifying existing ones originally developed to extract the CMB signal. For example, one technique used previously for CMB cleaning is 'Correlated Component Analysis' (CCA), and there has been work to re-purpose it for EoR data (Bonaldi et al., 2014). CCA relies, similarly to polynomial fitting methods, on the fact that the signal to be recovered has structure in frequency space, while foregrounds tend to look the same across the sky at all frequencies (with variation mainly in brightness).

Ultimately the CCA method has been shown to work quite well for simulated primordial hydrogen radiation, and reduced the foregrounds by roughly three orders of magnitude. Some of the shortcomings mentioned by Bonaldi et al. (2014) include the assumption that EoR radiation from two separate redshifts is entirely unrelated. In fact, two EoR slices sampled at redshifts varying by $\Delta z = 0.1$ will be correlated at the $10-20\%$ level, which means there is a risk that some of the EoR signal is lost to the CCA cleaning. It is also assumed that the mixing matrix does not vary rapidly across the sky (which may not be realistic). There is also an omission of discrete (point) radio sources in the foreground layers. However, overall the results paint an optimistic picture for upcoming data collection.

Chapman et al. (2013) introduces a non-parametric cleaning technique called Generalized Morphological Component Analysis (GMCA). While previous techniques largely assumed that foreground components were statistically independent and had smooth frequency spectra, GMCA has avoided these assumptions by instead focusing on the physical morphology of foreground components. This subtraction method does not assume anything about the spectral form of foregrounds, a technique sometimes referred to as 'Blind Source Separation' (BSS). The aim is to study the shape of foregrounds across various frequencies, to infer where small fluctuations stand out.

Building on previous work (Chapman et al., 2012), as well as GMCA theory developed for CMB map cleaning (Bobin et al., 2008), it has been shown that GMCA applied to

21 cm radiation is still effective in signal recovery. Some sections of 21 cm maps (superimposed with foregrounds) were recovered with correlation coefficients of up to 0.905 when compared to the original clean map (Chapman et al., 2013). Although parametric methods (those that assume spectral shape) perform better, Chapman et al. (2015) encourages more exploration of BSS techniques, which do not present as severe a risk of smoothing EoR features.

Foreground Avoidance

Yet another possibility is 'Foreground Avoidance' (Morales et al., 2012; Chapman et al., 2015, 2016). It has been claimed that most of the EoR signal lies in one region of the Spatial Fourier Space (an area where foregrounds contaminate less). This area, sometimes called the 'EoR Window', is a small wedge in the Fourier Representation of the signal $(k_\perp, k_\parallel)$ (**?**Raut et al., 2018). PAPER employs foreground avoidance, yet there are still uncertainty regarding the degree of leakage polluting such windows. It has been reported that diffuse foregrounds do contaminate these windows, though not unmanageably so (see Pober et al. 2013, section 4). HERA will also make use of the technique (Ali et al., 2015), while the SKA will rely primarily on foreground subtraction (Mellema et al., 2013).

Discrete Radio Sources

Of the research put into 21 cm radiation foregrounds, most of it has focused on Galactic phenomena (free-free, synchrotron) and diffuse radio galaxies/clusters. Discrete radio sources, mostly distant radio galaxies and other Quasi Stellar Objects (QSOs), remains an area that has not been thoroughly explored. In the context of EoR foreground subtraction, there has been some preliminary work concerning point sources (Di Matteo et al., 2002, 2004). However, this remains mainly an attempt to quantify the degree to which point sources contribute to foreground contamination, as opposed to devising a full-fledged subtraction scheme. These initial efforts have shown that current subtraction methods (used for diffuse foregrounds) are incapable of properly overcoming the issue. Part of the reason for this difficulty is the fact that each source has a unique frequency/intensity relationship, making them more complicated. Therefore, most foreground subtraction schemes suggest performing the discrete radio source removal separately (Di Matteo et al., 2004; Trott et al., 2012; Bonaldi et al., 2014). One useful tool towards approaching the challenge is the fact that, like most foregrounds, there is no spatial evolution in a map of radio sources sampled at different frequencies. The only difference is in flux. The complicating factor is that the flux/frequency relationship varies between sources.

A number of teams have focused on developing catalogues of the extragalactic radio sky (Sadler et al., 2002; Jackson, 2005; van Weeren et al., 2014; Williams et al., 2013; Hurley-Walker et al., 2017). This helps us to better estimate the magnitude and distribution of extragalactic sources, and offers the necessary templates for testing point source subtraction techniques. Instruments can also create their own sky maps using

longer baselines, and then subtract these when searching for the EoR on shorter baselines (section 0.3.3) via a process known as self-calibration. LOFAR currently performs self-calibration (Nijboer et al., 2006; Patil et al., 2017). The SKA will also use self-calibration (an overview of self-calibration for the SKA is given in (Repetti et al., 2017)).

Foreground Removal on SKA Pathfinders

It should be noted that much of these frameworks are built around preparing for SKA data intake, and thus the true effectiveness of various foreground removal methods will only become fully evident when tested on actual SKA data. For now, the increasing number of operational EoR experiments (section 0.3) makes it possible to begin preliminary foreground removal testing on real data. This has already been attempted on data from, among others, GMRT (Ghosh et al., 2011), MWA (Pober et al., 2016), and LOFAR (Patil et al., 2016). Such efforts are an imperative step towards understanding what to expect when SKA data is made available, although theoretical results must certainly continue alongside these (Mertens et al., 2018; Zuo et al., 2018).

Also of note is that, although we have consistently referred to the EoR signal here, this is not the only domain for which 21 cm foreground removal must be applied. Similar techniques have also been attempted on 21 cm signal from $z \sim 1$ galaxies to aid with intensity mapping (Ansari et al., 2012). The issue of synchrotron and free-free remains the same, however extragalactic point sources are no longer an issue in the same sense as with 21 cm from cosmological origin.

EoR Simulations

In the domain of EoR research, as in many different avenues of physics, a cooperative triad exists to continually expand and solidify our knowledge. The first aspect is the theoretical front (sections 0.2.1, 0.4, 0.5), imagining and mathematically formalizing the preliminary science. The second aspect is the experimental effort (section 0.3), building instrumentation to test hypothetical predictions. As the technical challenge of building larger and more complex instrumentation has increased over the years, a third aspect has emerged: simulation. Based on theory, simulations now help us predict what observational efforts should realistically hope to see. This, in turn, guarantees that instruments are optimized to verify predictions. Simulations can also directly help us determine astrophysical and cosmological parameters (section 0.6.6).

Boxsize and Resolution

Yet, even with large computing advancements in recent decades, computational power is not an infinite resource. Simulations must balance the physical 'boxsize' with the 'resolution' of each element. Opting for a large box entails – for realistic computation times

– larger pixels (more precisely, 'voxels'). This sacrifices some of the small scale dynamics such as halo formation and evolution. On the other hand, choosing small box and resolution sizes, though better capturing the dynamics of individual halos, results in the loss of larger physics such as the growth of the largest structures, and the accompanying large scale radiative transfer process (e.g. external ionization by distant sources and X-ray heating).

Authors differ on the boxsize necessary for a sufficiently robust 21 cm simulation. It has been suggested that only the largest simulation boxes ($\sim$ a few hundred cMpc) capture the true dynamics of the EoR, and that smaller boxsizes underestimate the size of ionized patches (Iliev et al., 2014).

In terms of resolution, the question is at what mass[29] (and redshift) halos collapse and begin star formation. In the early universe ($z \approx 100$) the IGM was at a temperature of $\sim$a few tens of K, with the first minihalos having a size of $\sim 10^3 M_\odot$ Tegmark et al. (1997a). As these halos grew, the thermal gas pressure and gravitational pressure competed. Cooling effects could lower the former below its virial equilibrium value, and allow the gas in smaller halos to collapse. In the metal-free primordial gas, the only possibilities were atomic hydrogen and molecular hydrogen. However, atomic hydrogen cooling is inefficient under 10^4 K (see figure 12). The corresponding virial mass is $\sim 10^8 M_\odot$, calculated based on the virial temperature and radius[30]:

$$M_{vir} = \frac{5 k_B T_{vir} R_{vir}}{2 G m_p} \tag{20}$$

What this means is that simulations must resolve halos down to *at least* $10^8 M_\odot$ for realistic reionization (Bromm et al., 2002; Furlanetto et al., 2006). Yet smaller halos almost certainly also contributed — the question is to what degree. Molecular hydrogen can cool down to $\sim$a few hundred K, corresponding to a mass of $\sim 10^5 M_\odot$. Yet molecular hydrogen usually forms on dust, which should not have existed in significant quantity until reionization had already began. Any molecular hydrogen formation would have had to go through inefficient channels in gaseous phase. If star formation can indeed begin in $10^5 M_\odot$ halos, reionization may have begun as early as $z \approx 30$ (Gnedin & Hui, 1998; Furlanetto et al., 2006).

In order to deal with this uncertainty, the ability to resolve $10^5 M_\odot$ halos helps to assure reasonable dynamics. However, this resolution is very computationally costly, and makes large boxsizes impossible[31]. Simulations can therefore choose to rely on the constraint that reionization should finish at $z \approx 6$. By strengthening the ionizing efficiency of $10^8 M_\odot$ halos (or larger ones), simulations can assure that this constraint will be satisfied (although reionization will start later, and the small-scale ionized structure will have a different morphology). The solution may ultimately be cooperation between large boxsize

[29] All masses quoted refer to total mass (baryonic + dark matter).

[30] The virial radius and mass can be related when assuming a typical NFW profile (Navarro et al., 1996).

[31] A 300 Mpc boxsize at this resolution would require 10^{14} (100 trillion) resolution elements.

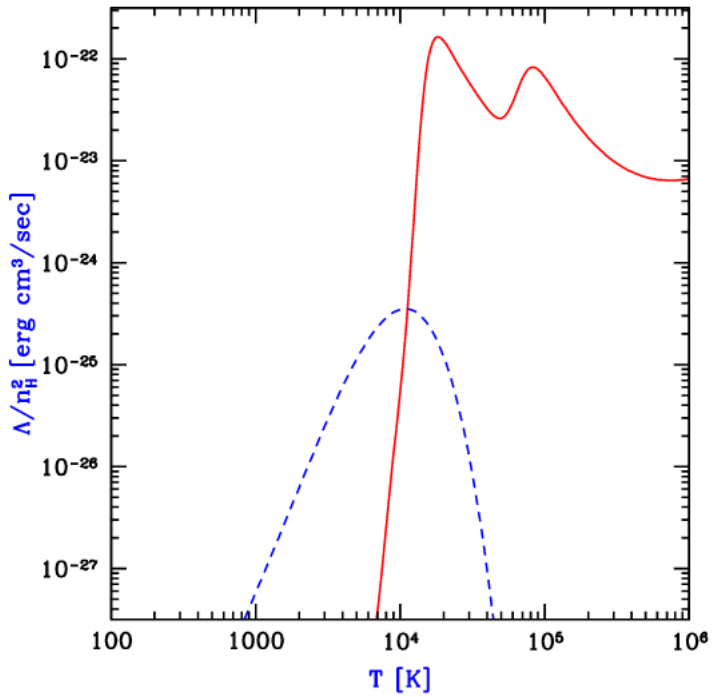


Figure 12 – *Cooling rates for primordial gas (H & He) as a function of temperature. The solid red line shows cooling rates for an atomic gas, while the dotted blue line is for molecular gas (assuming a molecular abundance of 0.1% n_H). Reproduced from Barkana & Loeb (2001).*

simulations, and high-resolution simulations. The latter can answer questions about the star forming efficiency of low-mass minihalos, and this information can be implemented into large boxsize simulations that do not otherwise resolve low-mass halos.

An overview of current EoR codes and simulations is presented in table 2, with accompanying resolution and boxsize.

Eulerian and Lagrangian Specifications

There is also an important distinction to be made that lies at the foundation of simulation work. A code can operate on the principal of dividing up a simulation box into individual fixed cells, within which various information is calculated (density, mass, energy, etc.). This is called the Eulerian specification of the flow field (Cen, 1992; Ryu et al., 1993; Teyssier et al., 1998). Alternately, a simulation can also abandon this fluid approach, and opt instead for tracking various packets of fixed mass. The position and velocity of each packet is recorded and evolved (this can still be calculated against a mesh). This second technique is known as the Lagrangian specification (Gnedin, 1995; Pen, 1995), sometimes referred to as 'meshfree'. Some of the disadvantages are that Lagrangian simulations can experience distortions, and Eulerian simulations tend to have limited dynamical range (Teyssier, 2002).

The Lagrangian specification can also be used as the foundation for Smooth Particle Hydrodynamics (SPH), in which the distances between particles is calculated and evolved (Gingold & Monaghan, 1977). This allows for the interpolation of various fields for regions between particles. The SPH approach will be explored more thoroughly in Chapter 3 (see section 3.3.3). Another point of note is that it is common in both Lagrangian and Eulerian simulations to handle dark matter as particles (e.g. Teyssier 2002). As dark matter is collisionless, treating it as a fluid leads to unrealistic density features.

For the purpose of this thesis the Lagrangian specification is more relevant, as the two codes (to be discussed later) that form the basis of chapters 1 and 3 are both Lagrangian.

Adaptive Mesh Refinement

One interesting compromise — both to the boxsize vs resolution paradigm (applying specifically to the Eulerian specification) — is to use Adaptive Mesh Refinement (AMR). The technique was developed in the 80s (Berger & Oliger, 1984; Berger & Colella, 1989), before being re-purposed for astrophysics (Ruffert 1992; Bryan & Norman 1997; Klein 1999; Teyssier 2002, among others). The idea was to start with a grid of fixed spaces, and then adapt refinement as needed. Regions deemed 'interesting' are subdivided into more finely divided resolution, while other regions are resolved more coarsely. This is often called the 'zoom-in' technique, and the effect is that interesting small-scale dynamics can be explored, while the computational time otherwise wasted on unnecessary regions is economized. For Lagrangian simulations analogous refinement methods may be used, however they must overcome the difficulty of sub-dividing mass-parcels, as opposed to simply the mesh.

Dynamical Complexity

In addition to the size and resolution of the simulation, another consideration is the robustness of the underlying physics. This concerns how the simulation is initialized, and subsequently evolved.

Initial Conditions

Initially, 'packets' of matter (gas, dark matter, etc.) are randomly scatted within the simulation box. Yet, this is an oversimplification, and setting up the initial conditions for a cosmological simulation involves some thought.

The first step is to generate the initial field with some level of fluctuations. This field can be generated at very high redshifts ($z \approx 1000$), and consists of a Gaussian random field (the power spectrum shape and normalization will depend on the assumed cosmology at high-z). In order to compute the density field (and subsequently velocity

and acceleration fields) at redshifts $z \lesssim 1000$, one common approach is known as the Zel'dovich Approximation[32] (Zel'dovich, 1970; Shandarin & Zeldovich, 1989). This is the assumption that a point $\mathbf{q} = (x, y, z)$ will experience density evolution given by:

$$\rho(\mathbf{q}, t) \approx \frac{\bar{\rho}}{(1 - b(t)\alpha(\mathbf{q}))(1 - b(t)\beta(\mathbf{q}))(1 - b(t)\gamma(\mathbf{q}))} \tag{21}$$

Here t is time;$\alpha(\mathbf{q})$,$\beta(\mathbf{q})$, and $\gamma(\mathbf{q})$ are the three eigenvalues of the deformation tensor at point $\mathbf{q}$; and $\bar{\rho}$ is the mean density at time t given in terms of the scale factor as $\bar{\rho} = (a_{z=0}/a)^3\rho_0$. The $b(t)$ factor is the linear fluctuation growth rate, whose full form is bulky (Peebles, 1980; Zeldovich & Novikov, 1983), but can be approximated[33] as (Shandarin & Zeldovich, 1989):

$$b(z) = \frac{b_0}{1 + \frac{2.5z\Omega_0}{1+1.5\Omega_0}} \tag{22}$$

where Ω_0 is the geometry parameter for a Friedmann Universe at $z = 0$, and b_0 is the linear fluctuation growth rate also at $z = 0$. With this we have the mechanism to implement initial density evolution into a homogenous simulation box[34]. Once the density fluctuations are included, the velocity field can be derived (this can also be carried out in reverse). After $z \approx 100$ numerical simulation can be used to capture non-linearities in the growth of density fluctuations.

Note Regarding Adaptive Mesh Refinement

For zoom-in simulations making use of AMR, an additional problem comes to light. When a region of interest is refined to have better resolution, the conditions within these new zoomed regions must be initialized such that there is no discrepancy at the resulting boundary between a finer-sampled region and a courser-sampled region. Poorly handling such boundaries can create shock artefacts. Codes have been developed specifically for the purpose of approaching this issue. Some use discrete Fourier transforms to add small scale perturbations to coarser perturbations (GRAFIC2 code, see Bertschinger 2001), while others use a method called 'real-space convolution kernels' (MUSIC code, see Hahn & Abel 2011). The detailed workings of such codes is not relevant here, however their importance in AMR simulations is worth being stated.

Evolving the Simulation

Once the initial conditions have been set up, each time-step calculates the gravitational attracting between packets, and moves them accordingly. When a region has sufficient density, it is marked as being luminous (the equivalent of stars having formed), and at

[32]The Zel'dovich approximation has now been largely replaced by more advanced approximations.

[33]Accurate to within 15% for $0.01 \lesssim \Omega_0 \lesssim 1$ (Shandarin & Zeldovich, 1989).

[34]Zel'dovich assumed that there would be one eigenvector ($\alpha(\mathbf{q})$,$\beta(\mathbf{q})$, or $\gamma(\mathbf{q})$) larger than the others, and therefore that the collapse would occur first along one eigenvector axis. This explains the presumption that large gas overdensities would form first, as discussed in section 0.2.3.

each subsequent time step photon packets are radiated outwards from said regions. These photon packets can, in turn, be absorbed by matter packets.

Though, as with resolution, full dynamics are computationally expensive. An alternative is Semi-Numerical simulations, in which various short-cuts are taken to approximate EoR dynamics. For example, gravitation may be replaced with linear extrapolation of gas movement from initial conditions. As well, full dynamic simulations generally treat gas differently in dense regions, while semi-numerical methods may neglect this. As for ionisation percentage, the number can also be inferred heuristically based on the photon absorption/emission. With these compromises, semi-numerical codes (Mesinger et al., 2011; Shin et al., 2008; Santos et al., 2008) can simulate reionization in boxes of 1024^3 in a few hours, while Full-RT simulations sometimes require over a hundred thousand (and thus, must be run in parallel on supercomputers).

Radiative Transfer

Another method of reducing computation time is by carrying out certain physical processes after the simulation has run. Radiative transfer (RT), for example, can be carried out in post-processing. This avoids advancing each photon packet at each step. Instead, steps account for only gravitation (and possibly hydrodynamics); the radiation (and temperature) is approximated for each redshift once the simulation has finished. The compromise is that post-processing radiative transfer may neglect important physical processes. Feedback between halos, for example is best captured in simulations where the radiative transfer is 'coupled' to the dynamical steps[35]. Fully coupled radiative transfer also provides a stronger model for the relationship between radiative (X-ray, UV, Lyα) heating and adiabatic cooling (Semelin, 2016; Semelin et al., 2017).

Full coupled RT simulations and semi-numerical simulations can, in a sense, work together. Heavy-duty simulations can help polish the assumptions made by faster semi-numerical codes. In turn, the latter can explore larger parameter spaces, and find 'interesting cases' to be explored in more depth by more robust simulations.

21cmFAST

One code in particular merits a separate word. 21cmFAST is a semi-numerical code that allows fast generation of simulated 21 cm signal from $z = 300$ onwards, and will be used extensively in chapter 3. The code is able to evolve the signal through this time-scale in a few minutes on a single processor (at resolutions of a few cMpc), and although the finer details of full numerical simulations will not be seen (e.g. section 1.2.2), the power spectra are still shown to agree to 10s of percent (Mesinger et al., 2011). 21cmFAST handles initial conditions in the method detailed in section 0.6.2, and then achieves computation time

[35]This will be relevant only for small halos (see section 0.6.1), and at small angular resolutions (∼a few hundred ckpc, see Semelin 2016).

economies via the following: bypassing the halo finding algorithm (which was previous used in Mesinger & Furlanetto 2007), approximating gravitational collapse by moving each particle according to first-order perturbation theory[36] (Zel'dovich, 1970), opting not to treat baryons and dark matter separately, and simplifying baryonic physics[37] (Mesinger et al., 2011). Sample 21cmFAST output is shown in figure 14.

Summary of EoR Simulations

We present here an overview of some recent EoR simulations. As well, figure 13 compares boxsize and resolution between a number of contemporary simulations. The units ckpc and cMpc are comoving kiloparsercs and megaparsecs, respectively.

Table 2 – *Summary of EoR Simulations and Codes*
A star denotes a code, while all other entries are simulations.

Name	Full RT	AMR	Coupled	Resolution ($z = 6$)	Boxsize ($z = 6$)	Reference
21cmFAST*	no	no	no	~ 0.5 cMpc	~ 200 cMpc	(Mesinger et al., 2011)
CoDa[a]	yes	no	yes	~ 3 ckpc	~ 90 cMpc	(Ocvirk et al., 2015)
CubeP^{3}M[b$\star$]	yes	no	no	~ 90 kpc	~ 600 cMpc	(Iliev et al., 2014)
CROC	yes	yes	?	100 cpc	100 cMpc	(Gnedin, 2014)
ENZO[c$\star$]	yes	yes	yes	~ 4 ckpc	20 cMpc	(Norman et al., 2015)
GADGET-2[d$\star$]	yes	?	yes	~ 38 ckpc	~ 9 cMpc	(Finlator et al., 2011)
LICORICE*	yes	yes	yes	~ 10 ckpc	300 cMpc	(Semelin, 2016)
RAMSES-RT[e$\star$]	yes	yes	yes	?	?	(Rosdahl et al., 2013)
SPHINX[e]	yes	yes	yes	~ 0.1 cpc	10 cMpc	(Rosdahl et al., 2018)
Technicolor Dawn	?	?	?	50 ckpc	~ 18 cMpc	(Finlator et al., 2018)
TRAPHIC*	yes	?	?	~ 200 pc	~ 70 cMpc	(Pawlik et al., 2015)
VULCAN[c]	no	yes	yes	~ 0.3 cpc	25 cMpc	(Anderson et al., 2017)
SIMFAST21[f]	no	yes	yes	~ 9 ckpc	100 cMpc	(Santos et al., 2008)
n/a[c]	yes	yes	yes	50 ckpc	~ 14 cMpc	(Chen et al., 2017)

[a]Customized from a previous code detailed in Stranex & Teyssier (2010).
[b]Customized from a previous code detailed in Harnois-Déraps et al. (2013).
[c]Customized from a previous code detailed in Bryan et al. (2014).
[d]Customized from a previous code detailed in Springel (2005).
[e]Customized from a previous code detailed in Teyssier (2002).
[f]Customized from a previous code detailed in Shin et al. (2008).

[36]There is also the option to evolve the density linearly, for an additional speed boost.
[37]Some baryonic physics is included in the parameters (e.g. N_γ and $T_{\rm vir}$; see section 3.3.1).

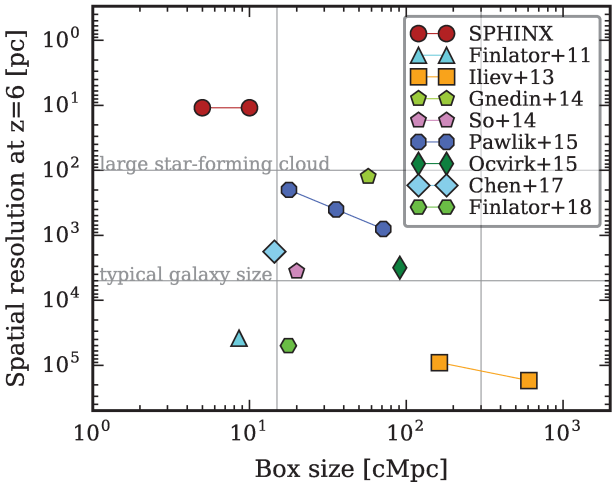


Figure 13 – *A comparison of recent EoR simulations, showing relative boxsize and resolution. Used with permission by Joakim Rosdahl, and originally by Ali Rahmati.*

Visualizing the EoR

Simulations must then be turned into visualizations that represent the structure formation and reionization we hope to understand. One of the most intuitive approaches is to take various slices through the simulation boxes. A slice showing the temperature brightness at each point results in a heat map. An example is shown in figure 14, produced with the semi-numerical code 21cmFAST (Mesinger et al., 2011) at $z \approx 15$. The average 21 cm temperature brightness is 34.8 mK below T_γ, with colour indicating the degree of fluctuations as a percentage of $< \Delta T_b >$. The fraction of neutral hydrogen is ≈ 1, which is to be expected at $z = 15$, at which time reionization is in its earliest stages.

It should be noted that we are not limited to visualizing temperature brightness. Figure 15 shows a number of other EoR diagnostics that can be explored through simulations. These have been produced with the Full Radiative Hydrodynamics SPHINX code (Rosdahl et al., 2018).

Now let us look at how to also visualize the EoR across a wide range of redshifts.

Lightcones

Both figures 14 and 15 represent various aspects of the Universe at a single redshift (a 'snapshot' of the EoR). Instead, we may wish to visualize the chronology and progression of the epoch. We can accomplish this by creating a 'lightcone'[38]. A box is evolved over the course of reionization, and slices are taken at set intervals ($\Delta z = 0.1$, for example).

[38]Lightcones, when visualized, are rectangular prisms, and not cones. The name is somewhat misleading, and comes from the conical shape an expanding wake of photons takes across the chronology. 'Light-rectangular prism' would have been more apt, but lightcone remains the standard.

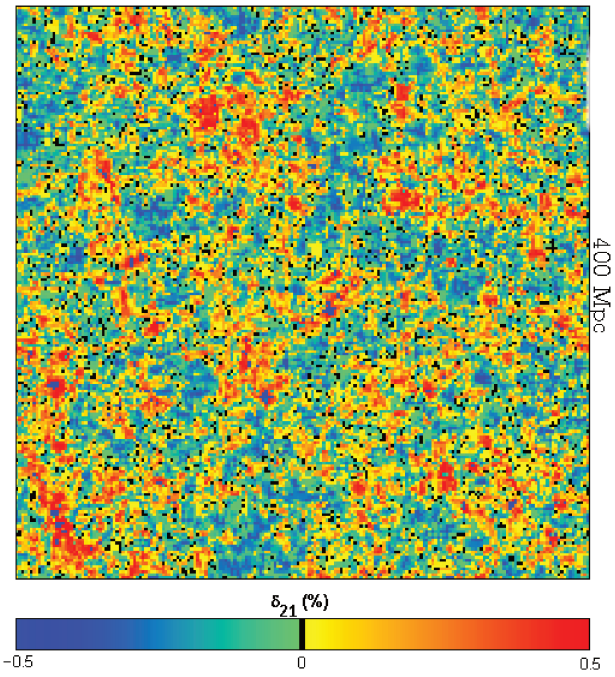


Figure 14 – *An example slice (400×400 Mpc) produced by 21cmFAST (Mesinger et al., 2011) at* $z \approx 15$*, with* $< \Delta T_b > = -34.82$ *mK and* $< x_{HI} > = 0.983$

These slices are then arranged next to one another, such as to create a time line. A visualization of the process is shown in figure 16 and an example is shown in 17, both taken from Zawada et al. (2014)

Parameter Reconstruction

In addition to helping in the development of instrumentation, simulations also play a part in extracting information about astrophysical processes by deriving constraints on the chosen parameters (referred to in future simply as 'parameter extraction'). This is accomplished by first deciding on a number of astrophysical or cosmological parameters for which the values are not well known (e.g. photon escape fraction, Hubble constant, population III star masses, etc.). Each of these parameters is varied within reasonable values (assuming a best estimate can be made), and thus an n-dimensional parameter space can be explored. An EoR observable[39] (observables can be power spectra, snapshots, lightcones, pixel density functions, etc.) is simulated at each point of the parameter space, in order to create a database of the breadth of possible EoR scenarios.

When observational data is collected, databases provide an efficient tool for deducing the true values for explored parameters. Real EoR data can be compared to the different observables (either morphologically, or in terms of some diagnostic such as the power

[39]We use the term observable to refer to the various outputs from a simulation. This is meant to avoid confusion with theoretical models of the universe. In section 3.1.1 we will elaborate further on the nuances of this definition, among others.

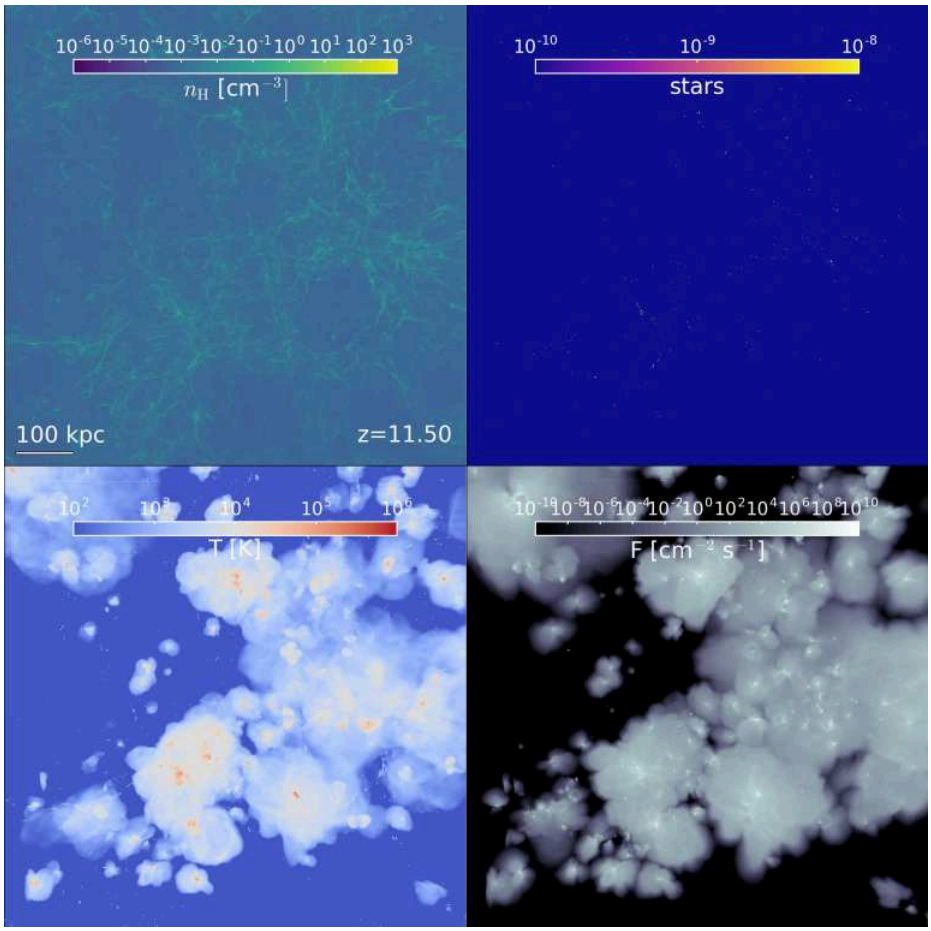


Figure 15 – *An example of various diagnostics produced by SPHINX (Rosdahl et al., 2018). The top left square shows the neutral fraction n_H, the top right is stellar matter (note the scale makes individual stars very small in this image), the bottom left is the temperature, and the bottom right is flux. The box size is 100 kpc.*

spectrum), and the best 'match' gives us an approximation for the true parameter values (based on the values used to simulate the closest match to the real data).

This methodology presents two challenges. Firstly, it will be necessary to quantify the 'difference' (or distance) between reionization scenarios. How should we define such a quantity, and how will our definition affect parameter reconstruction? This is a contemporary issue, with no single solution, and it will be explored in more depth in chapter 2. Secondly, for large databases of detailed observables (>10 Gb each), comparisons for all combinations of observables can become complex and computationally expensive. To overcome this, the last few years have seen an explosion of efforts to realize efficient parameter reconstruction, either via the Bayesian Markov Chain Monte Carlo method (Harker et al., 2012; Greig & Mesinger, 2015; Kern et al., 2017; Greig & Mesinger, 2017b, 2018), or (more recently) via neural networks (Shimabukuro & Semelin, 2017; Kern et al., 2017; Schmit & Pritchard, 2018; Gillet et al., 2018). We add to these efforts, as will be presented in chapter 3.

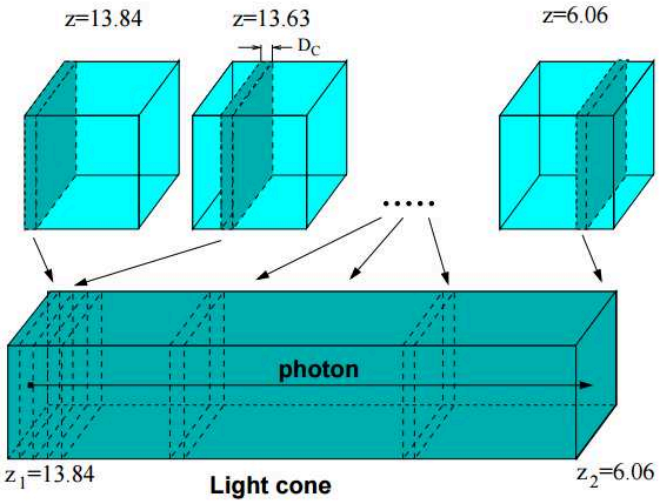


Figure 16 – *The procedure through which snapshots can be used to create a lightcone (Zawada et al., 2014).*

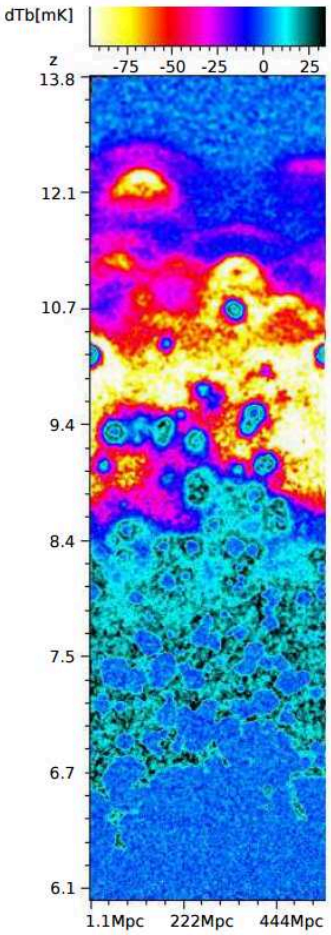


Figure 17 – *An example slice through a lightcone taken from Zawada et al. (2014), showing* ΔT_b *between redshifts* ~ 6 *and* ~ 14.

Bayesian MCMC

The roots for the Monte Carlo method are found in 1940s wartime, during which the probabilistic depth of particle travel through various materials involved tedious manual calculations (Kean 2010; Robert & Casella 2011). Los Alamos scientists (Stanislaw Ulam, and later John von Neumann) realized that randomly sampling initial conditions and calculating the corresponding outcomes could give efficient estimations of the probabilities, as opposed to the tedious calculation of the full numerical solution. The method proved successful, and was further developed after the war (Metropolis & Ulam, 1949; Turing, 1950; Kahn & Theodore, 1951; McKean, 1966).

However, although the Monte Carlo method handled probability functions of one or two variables well, it faltered when faced with high-dimensional functions. Metropolis et al. (1953) focused on an especially difficult probability function for which numerical integration was impossible, and which could not be randomly sampled on account of how small it was for most coordinates. The proposed solution was that, after randomly sampling N points, a random walk could be adopted to move each point iteratively towards a region of higher probability. At each iteration the random walk was represented by a Markov Chain matrix[40], hence the name Markov Chain Monte Carlo (MCMC). This foundation has remained the same, although future developments honed the mathematics and presented improved algorithms (Hastings, 1970; Geman & Geman, 1984; Tanner & Wong, 1987; Gelfand & Smith, 1990).

The MCMC method formed the basis of the BUGS software for Bayesian inference (Gilks & Spiegelhalter, 1994). Subsequent more robust programs used MCMC Bayesian inference to explore n-dimensional parameter spaces initially for the CMB (Christensen et al., 2001; Lewis & Bridle, 2002), and soon for the 21 cm signal as well (Harker et al., 2012; Greig & Mesinger, 2015; Kern et al., 2017; Greig & Mesinger, 2017b, 2018). In sum, for a given observation x we assume some n-dimensional probability distribution function p which depends on n parameters of unknown values ($\vec{\theta} = \theta_1, \theta_2, ..., \theta_n$). Hence we have $p(x|\vec{\theta})$ (the probability of observation x if we have parameters $\vec{\theta}$). Yet we want to know the values of the variables, not the observation! So we flip this to $p(\vec{\theta}|x)$ and plug it into Bayes' theorem:

$$p(\vec{\theta}|x) = \frac{p(\vec{\theta})p(x|\vec{\theta})}{\int p(x|\vec{\theta})p(\vec{\theta})d\vec{\theta}} \tag{23}$$

where the denominator is a normalization constant which must be evaluated. The difficulty in evaluating this n-dimensional integral, as well as the n-1 dimensional integrals required to find $p(\theta_1)$, $p(\theta_2)$, etc., is immense without the MCMC approach (see Christensen et al. 2001 for a full derivation).

[40]The simplest form represents, for a given point, an equal chance of moving to any of the discretized points falling within a certain radius around the initial point.

Principal Component Analysis

One method which has become increasingly common to simplify high-dimensional data, and merits a quick mention, is Principal Component Analysis, or PCA (developed in Pearson 1901 and Hotelling 1936). The PCA method consists of first defining a new set of vectors within a data cluster, each one a linear combination of the associated parameters. The vectors are arranged from the most, to the least, correlated with the data. By ignoring the least correlated vector, it becomes possible to reduce dimensionality while losing the least amount of information. These vectors correspond to the eigenvectors[41] of the Fisher Matrix, for which each entry relates a pair of two parameters to the resulting observable. The full mathematical definition of the Fisher Matrix is not relevant for our purposes (it can be found in Ly et al. 2017, Appendix A).

In the late 90s, cosmological data sets (historically quite small compared to other domains) were large enough to attempt PCA parameter reconstruction (Tegmark et al., 1997b). The method now forms the basis of a number of projects throughout astrophysics (see Ishida & de Souza 2011, and references therein). The theory developed in chapter 3 is tangentially related to PCA, as will be discussed therein (see note at the end of section 3.4).

Coming Challenges

This chapter has presented an overview of the current state of 21 cm Cosmology, and efforts to probe and understand the Epoch of Reionization. At present, there are a number of challenges to be overcome, as well as interdisciplinary partnerships to be strengthened, in order to achieve this goal.

Firstly, experimental efforts must be coordinated and developed further (section 0.3). Upper limits for the brightness temperature must be lowered, and tentative detections must be cross-checked with other experiments.

Secondly, in conjunction with experimentation, previous theoretical frameworks must be revisited and re-evaluated (section 0.4). Limits on 21 cm signal strength must be used to reject models that predicted radiation beyond said limits, and constrained parameter values must be translated into improved theoretical nuance. This is already taking place with recent limits (Ali et al. 2015 now retracted, Patil et al. 2017) which have been used to constrain (for example) the X-ray production at high redshift.

Thirdly, another theoretical consideration — and one so significant it merits separate appraisal — is foregrounds (section 0.5). Increasingly accurate radio sky maps, pushed deeper in order to capture extragalactic sources as well as Galactic ones, must be created

[41] The corresponding eigenvalues are measures of the correlation between a model and data along the eigenvectors.

and used to test various subtraction methods. Real data must also be subject to subtraction testing, to converge on the full cleaning pipeline future experiments such as the SKA will utilize.

Fourthly, both the theoretical and observational efforts must be bolstered by robust and efficient EoR simulations (section 0.6). Simulated EoR signal, built upon theoretical predications, must serve to hone instrumentation in order to maximize the chance of signal detection. Eventually, simulations and the databases they make possible will be invaluable in determining the true parameter values upon which the early Universe operated.

These are the pillars upon which EoR physics has been built. The goals have been made clear, and now, after nearly a century of advances, there is an undeniable sense that we are on the cusp on finally unlocking the secrets of the Cosmic Dawn and Epoch of Reionization.

This manuscript is organized as follows. Chapter 1 discusses the preparation undertaken in building a database of high-resolution 21 cm observables, notably the development of an observable for realistic noise. Chapter 2 explores different methods to quantify the difference between two observables, their advantages and disadvantages, and talks of different ways to use the database. Chapter 3 explores the geometry of the parameter space, with the goal of developing a framework to optimally sample said space (such as to best train neural networks). Finally, Chapter 4 summarizes what has been achieved, how it can help future EoR studies, and highlights avenues that merit further exploration.

Part One

21SSD: Building a Database of EoR Signals

CHAPTER 1

21SSD: Building a Database of EoR Signals

Motivation

In preparation for upcoming EoR experiments, and ultimately the Square Kilometre Array, the need for high-resolution Full-RT simulations has already been stated. Although Bayesian Markov Chain Monte Carlo methods are computationally infeasible, these simulations can still be used for parameter reconstruction methods with neural networks (section 0.6.6). A database of possible EoR 21 cm signals would also provide templates for developing end-to-end simulations for the SKA. This will be a necessary step before the experiment can go online. Another use for simulations with high-resolution and full dynamics is to adjust semi-numerical codes. Specific effects that may only show up in a heavy-duty simulation, such as feedback effects which result from coupling ionization and dynamics, can be imitated by semi-numerical codes through 'fine tuning' the parameters, or with the introduction of new ones. The resolution may not be as fine, but if the resulting reionization scenarios are convergent, this allows semi-numerical codes to replicate the findings of full simulations to obtain a much finer sampling of a parameter space.

Finally, and perhaps the most fundamental reason to create a database, has to do with computational time. Full simulations may take many hundreds of thousands of computer hours to realize, and repeating this multiple times is simply not realistic. Any time anyone wants to carry out *any* science on mock SKA data, it is in their interest to use the highest resolution, and the most physically realistic, simulations available. Re-creating simulated data, and expending immense computational resources, multiple times by different research groups, is simply nonsensical unless the goal is explicitly to test different simulations. Therefore, by making EoR templates publicly available, this wastage can be avoided.

It is for this reason that we sought to prepare 21SSD (21 cm Simulated Signal Database), which can now be found online at 21ssd.obspm.fr.

Our contributions to this goal were as follows. Familiarity with the simulation was gained through studying some aspects of the physics of the resulting models, specifically

with respect to self-shielding effects in cosmological radiative transfer (section 1.2). After this, a review was carried out on the different possible parameter spaces that have been used previously by other groups (section 1.3). Lastly, an additional goal was realistically modelling the thermal noise which the SKA should be subject to at different resolutions and redshifts (section 1.4).

Table 1.1 – *Chapter 1 Contribution Breakdown*

Section	Me	Not Me
1.2	x_α histogram (figure 1.3)	LICORICE Code
1.3	Some discussion and literature review	Final parameter choices and creating 21SSD
1.4	Literature review, initial noise model and noise addition routine	SKA UV-coverage modelling

Corresponding Publication: First half of Semelin et al. 2017.

A Toe in the Water: Self-Shielding

The LICORICE Code

The code used to develop the database is LICORICE. The code has been developed over the past two decades (Semelin & Combes, 2002; Semelin et al., 2007; Baek et al., 2009; Iliev et al., 2009; Baek et al., 2010; Vonlanthen et al., 2011; Semelin, 2016). The simulation allows for fully coupled radiative hydrodynamics at high-resolution (1024^3 or more), and is based on a Tree+SPH method (Semelin & Combes, 2002). Both ionizing UV and X-ray frequencies are coupled to the dynamics, which are advanced with Monte Carlo ray-tracing. The program also makes use of Adaptive Mesh Refinement for the ray-tracing.

Shielding

An interesting feature that appears in the resulting observables[1] is 'shielding'. Two separate effects were observed: X-ray shielding and Lyα shielding. In the first case, some gas regions that are initially warmer than their surroundings at the beginning of the EoR are found to become cooler than neighbouring regions as reionisation proceeds. In the case of Lyα shielding, the result is lower coupling of T_S to T_k, leading to moderately overdense neutral regions in the voids.

[1] We remind the reader that other authors have previously called these 'models'.

X-Ray Shielding

To understand the first shielding effect, we refer to the equation for the comoving mean free path (Semelin, 2016):

$$l = 2 \times (1+\delta)^{-1} \left(\frac{E}{E_0}\right)^3 \left(\frac{10}{1+z}\right)^2 \text{ckpc} \tag{1.1}$$

where $1+\delta$ is the fractional overdensity (as seen in equation 16), E is the photon energy, and E_0 is the ionization threshold energy ($\sim$ 13.6 eV from Mohr & Taylor 2000). The X-ray energies are often taken to be quite high ($\sim$a few kev), resulting in (for overdensities of $\sim$a few, at EoR redshifts) a large mean free path on the order of the Hubble radius. However, the soft X-ray spectrum consists of many photons emitted at energies of a few hundred eV, which corresponds to a much smaller mean free path on the order of a few 100 ckpc. Therefore, when proper X-ray energies, and full radiative transfer, are implemented, X-rays do not efficiently penetrate overdense regions.

An example of this is illustrated in figure 1.1, specifically in the region circled in red. We notice in the left-hand panel that this region is $\sim$5 K warmer (from adiabatic contraction) than neighbouring regions at $z = 14$. Yet at $z = 10$, shown in the right-hand panel, this difference has been inverted. This is because X-rays have heated the surrounding region more than the filament-like structures.

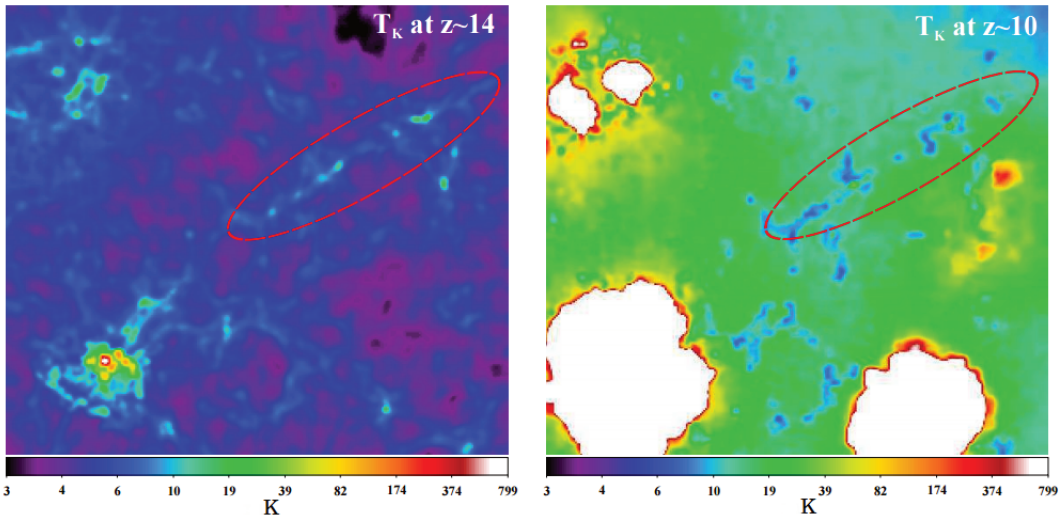


Figure 1.1 – *An example of X-ray shielding within a slice ($\sim$10 cMpc per side, $\sim$140 ckpc thick) produced with the LICORICE code. Reproduced from Semelin (2016).*

Lyα Shielding

A similar, though more complicated, effect exists for Lyα photons. LICORICE is the only code that performs 3D Lyα transfer to compute T_S (Semelin et al., 2007; Baek et al., 2009;

Vonlanthen et al., 2011), and as such this effect has not been observed in other simulations (which will approximate the local Lyα flux that determines T_S as, for example, evolving with a simple $\frac{1}{r^2}$ radial flux profile). Semelin et al. (2007) showed that, when resolution is adequate, photons are often scattered off over-dense regions while still in the wings of the filament (roughly one half may be ejected). Had it not been for this scattering, they would otherwise have been redshifted to Lyα frequency, and scattered in the centre of the filaments (thus contributing therein to the coupling of T_S and T_k). Close to sources, the radial flux profile can steepen to $\sim r^{-\frac{7}{3}}$. This effect leads to Lyα back-scattering in the wings, as shown by the dark regions of low x_α in figure 1.2.

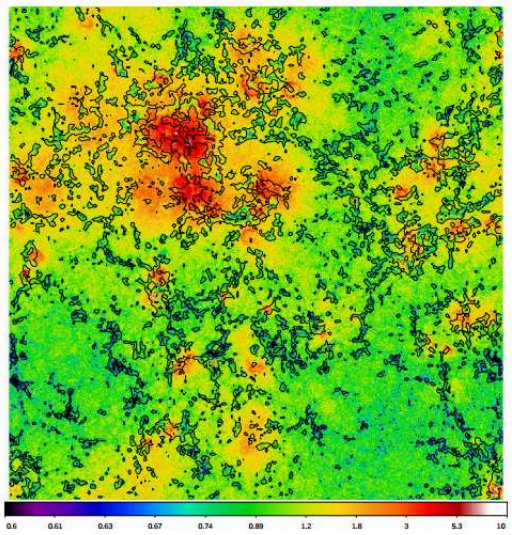


Figure 1.2 – *Lyα coupling coefficient (x_α) in a 140 ckpc thick slice ($z = 10.1$), produced with LICORICE. The contours correspond to $\delta \geq 1.38 \cdot \delta_{avg}$. Reproduced from Semelin (2016).*

To quantify that this was in fact the case, the relationship between density and x_α (the Lyman alpha coupling coefficient) was studied. The goal was to show that regions of higher density do indeed correspond, in general, to regions of low x_α. This is seen in the histogram presented in figure 1.3. The solid black line represents the average x_α value for a given density, and the black dots above and below show the values within which 68% of the cells fall.

Although the average line becomes somewhat erratic at high densities, this is because there are fewer cells which fall into higher density bins, skewing the sample size. Regardless of this, the downwards trend is evident: denser regions tend to have a lower Lyα coupling coefficient. This is, in the essence, what is meant by Lyα shielding, and explains the behaviour seen in figure 1.2. This trend could be used to improve on the $\frac{1}{r^2}$ modeling used in other codes that do not run the fully line transfer.

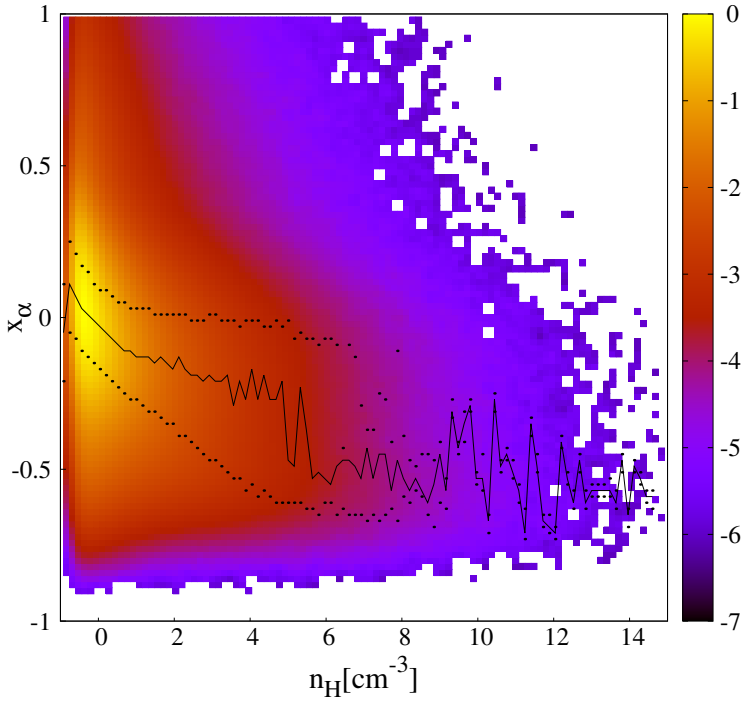


Figure 1.3 – *2D histogram of the density and x_α distribution. The solid black line is average x_α, and the black dots are 68% bounds for each column.*

Defining a Parameter Space

After having developed a familiarity with the code, the next task was to decide which parameters to vary, as well as which values to choose. There are a number of different opinions on how best to parametrize Epoch of Reionization simulations.

Previous Parametrizations

Pritchard et al. (2015) suggests a number of relevant cosmological parameters (Ω density coefficients, Hubble parameter, inflationary parameters, curvature, etc.). Greig & Mesinger (2015) explores a three parameter space consisting of the ionizing efficiency of galaxies ζ, the mean free path of ionizing photons in ionized regions $R_{\rm mfp}$, and the minimum virial temperature for star-forming halos $T_{\rm vir}^{\rm Feed}$. Three more parameters were recently added in Greig & Mesinger (2017b): the integrated soft band luminosity per SFR $L_{X<2{\rm keV}}/SFR$, the X-ray energy threshold for self-absorption by the host galaxies E_0, and the X-ray spectral index α_X. Fialkov et al. (2017) outlines a basic parameter space consisting of an X-ray emissivity term f_X[2], a hard/soft Spectral Energy Distribution

[2]Although this was introduced earlier in Furlanetto et al. (2006).

(SED) term, and a boolean parameter as to whether or not to include Atomic Cooling (equivalent to deciding if small halos should exhibit star formation). Kern et al. (2017) presents an ambitious 11-dimensional parameter space, including the three astrophysical parameters of Greig & Mesinger (2015), five cosmological parameters, as well as three parameters related to the X-ray spectrum. See summary table 1.2.

It can be seen that the choice of parameters varies significantly, and is intimately linked to the corresponding modelling technique. Switching to a new set of parameters can sometimes be very technically challenging in terms of redesigning code previously based on a different set. Regardless, converging on a standardized set of parameters was set out as a task for theoreticians involved in the SKA science working group.

In the context of LICORICE, the cosmological parameters suggested in Pritchard et al. (2015) and Kern et al. (2017)[3] are less relevant: a number of LICORICE's advantages are astrophysical. For example, coupled photon packet propagation, hard X-ray contribution, shielding effects shown above, etc (Semelin et al., 2017). Two parameters from Fialkov et al. (2017) were easy fits for LICORICE: f_X, as well as the SED variable (although the latter was modified slightly for our purpose). The third variable (a 3D parameter space is a reasonable first step for a computational expensive simulation such as 21SSD) relates to the Lyman band emissivity. Ultimately these three are most relevant for the very beginning of the EoR, during the initial absorption regime (see section 1.3.2 for a definition of all 3).

Table 1.2 – *Parametrizations*

Reference	#	Parameters
Greig & Mesinger (2015)	3	ζ, $R_{\rm mfp}$, $T_{\rm vir}^{\rm Feed}$
Fialkov et al. (2017)	3	f_X, SED, Halo SF Cutoff (Atomic Cooling)
Semelin et al. (2017)[a]	3	f_α, f_X, $r_{H/S}$
Kern et al. (2017)	11	ζ, $R_{\rm mfp}$, $T_{\rm vir}^{\rm Feed}$, f_X, α_X, ν_{min}, σ_8, H_0, $\Omega_b h^2$, $\Omega_c h^2$, n_s
Greig & Mesinger (2017b)	6	ζ, $R_{\rm mfp}$, $T_{\rm vir}^{\rm Feed}$, $\mathrm{L}_{X<2\mathrm{keV}}/SFR$, E_0, α_X

[a] 21SSD.

Although the parameters presented in Greig & Mesinger (2015) were not used here, they will be returned to in chapter 3.

Equivalences

There are a few small words to add on these parameter sets. Firstly, the X-ray parameter f_X in Kern et al. (2017) is defined slightly differently than in Fialkov et al. (2017), however

[3] Note that this parametrization, as well as that of Greig & Mesinger (2017b), were published *after* 21SSD went online. Their inclusion here is retrospective.

they are both effectively normalization parameters. The former defines it as:

$$\epsilon_X(\nu) \propto f_X \left(\frac{\nu}{\nu_{min}}\right)^{-\alpha_X} \tag{1.2}$$

where ϵ_X is the source X-ray emissivity, ν_{min} is the obscuration frequency cutoff, and α_X is the spectral slope (also parameters in Kern et al. (2017)). The alternate definition, adopted for 21SSD, is presented below.

The SED parameter in Fialkov et al. (2017) is also roughly equivalent to the $r_{H/S}$ parameter used for 21SSD (described below). Fialkov et al. divides the SED into 'hard' and 'soft' variants, in which each is defined by the luminosity divided by the star formation rate (SFR) (for the frequency bands 0.2-30 keV and 0.5-8 keV). The values for these are given in the article.

Parameter Definitions

The parametrization chosen for 21SSD consists of the following three parameters (in-depth descriptions are given in Semelin et al. 2017).

- *f_α: Lyman band emissivity*
  This value quantifies the Lyman band emissivity efficiency, and encapsulates a number of uncertainties associated with the Lyman emissivity of early luminous sources (choice of the initial mass function, dust, etc.). The definition is somewhat complex, as it should be taken to be a heuristic scaling, however it can be written as

$$f_\alpha = E^{\text{eff}} \left(\int_{\nu_\alpha}^{\nu_{\text{limit}}} \int_M \xi(M) L(M,\nu) T_{\text{life}}(M) dM d\nu \right)^{-1} \tag{1.3}$$

  where E^{eff} is the effective energy emitted in the simulation, $\xi(M)$ is the IMF, $L(M,\nu)$ is the energy emitted per second per Hz by a star of mass M at frequency ν, $T_{\text{life}}(M)$ is the lifetime of a star of mass M, ν_α is the Lyα frequency, and ν_{limit} is the Lyman limit frequency (the ionization energy of neutral hydrogen). Ultimately, everything in the large bracket of equation 1.3 is the theoretical energy that we expect based on the values of the listed variables. We can therefore simply say that $f_\alpha = E^{\text{eff}}/E^{\text{theory}}$. f_α is given values 0.5, 1.0, and 2.0. It should also be noted that this is a computationally 'free' variable: Lyman line transfer has no noticeable feedback on the dynamics, and is therefore carried out in post-processing. Moreover, the resulting local Lyα flux scales linearly with the emissivity of the source.

- *f_X: X-ray emissivity*
  This is a scaling parameter for adjusting the X-ray luminosity of a source (L_X). It is defined as

$$L_X = 4.3 \cdot 10^{40} f_X \left(\frac{SFR}{1 M_\odot \cdot \text{yr}^{-1}}\right) \text{erg} \cdot \text{s}^{-1} \tag{1.4}$$

where SFR is the star formation rate. The X-ray luminosity is very poorly constrained (Furlanetto et al., 2006), and is therefore sampled over a wide range (0.1, 0.3, 1.0, 3.0, 10).

- $r_{H/S}$*: Hard-to-soft X-ray ratio*
  X-ray photons will differ in energy (and therefore in mean free path) depending on whether they originate in AGN or X-ray binaries (Fialkov et al., 2014). The relative importances of these two in reionization is not well constrained, and therefore we define the proportion of them as

$$r_{H/S} = \frac{f_X^{XRB}}{f_X} \tag{1.5}$$

  where $f_X = f_X^{AGN} + f_X^{XRB}$, and has the definition given in equation 1.5. We consider three cases: the X-ray contribution is entirely due to AGN, entirely due to X-ray binaries, and due to both equally ($r_{H/S} = 0, 0.5, 1$).

Table 1.3 – *21SSD Parameter Values*

Parameter	Explored Values
f_α	0.5, 1., 2.
f_X	0.1, 0.3, 1., 3., 10.
$r_{H/S}$	0., 0.5, 1.

Using the different combinations of values listed in 1.3, the resulting sampling consists of 45 observables. Examples of the resulting lightcones included in the 21SSD (slices through the lightcones) are shown in figure 1.4. The initial sampling is sparse, as it represents a first step into eventually constructing a more ambitious set of different observables. For this preliminary sampling, the large computation time ($\sim 2.5 \times 10^6$ computer hours) is incentive to be cautious in the initial scope.

From figure 1.4 we are also able to visualize the general relationship between the three parameter values and the resulting reionization scenarios. Choosing a small value of f_X leads to a later heating, which allows for more time in the strong absorption regime. An increase in f_α means earlier Lyα coupling, and hence a stronger absorption. The effect of $r_{H/S}$ is more subtle, but can been seen between the 3^{rd} and 4^th lightcones (from the top) in figure 1.4. Specifically at redshift 8 we notice how setting $r_{H/S}$ to 1 results in a more diffuse lightcone, with less structure. Soft X-rays heat more locally, while harder X-rays travel farther before heating their surroundings, and hence contribute to a somewhat more 'global' heating.

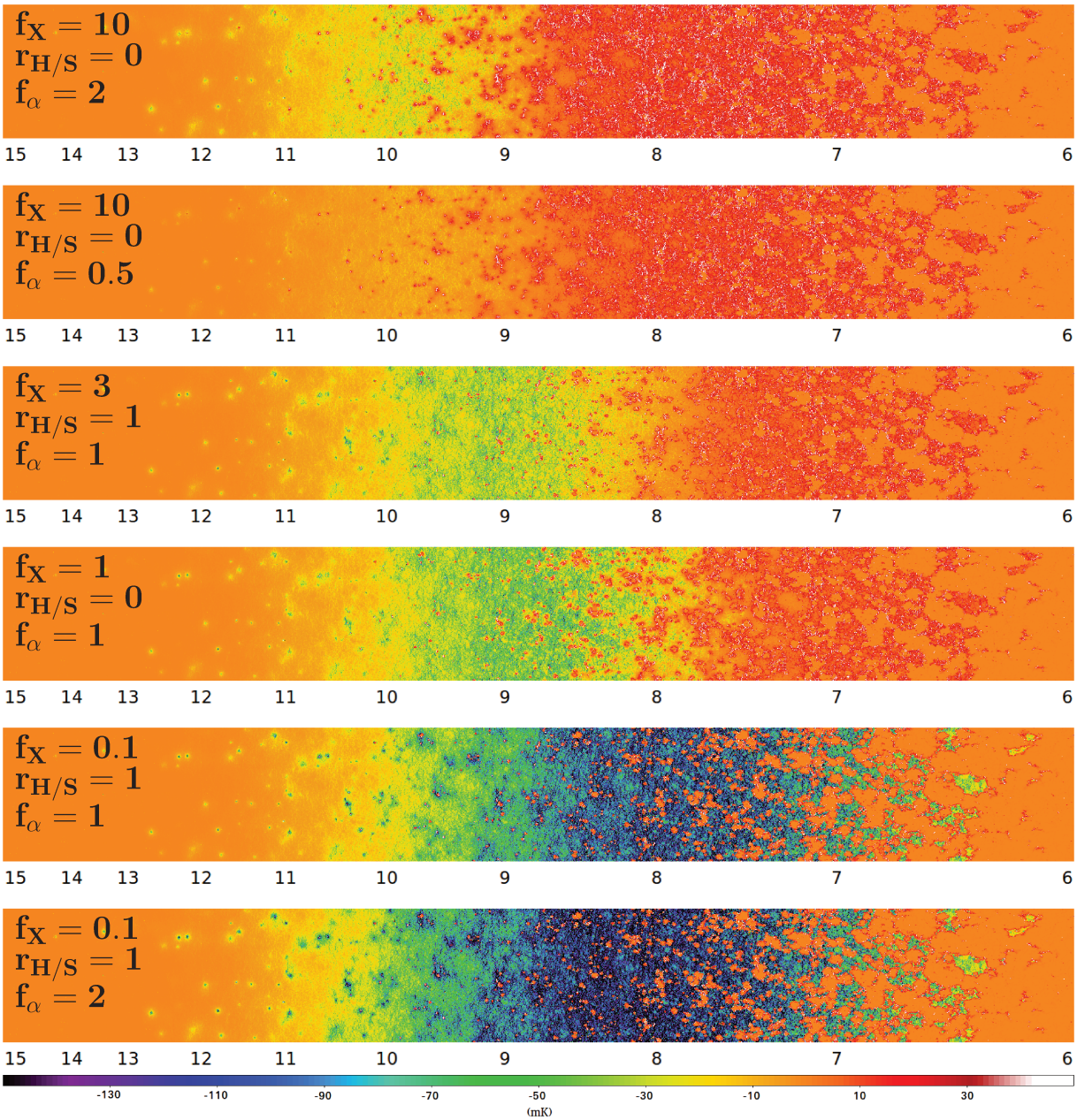


Figure 1.4 – *Example lightcone slices from the 21SSD database, produced with LICORICE. The parameter values used to create each observable are listed in the top right corner.*

Simulating SKA Noise

Having developed these observables, they are ready to be put to use as templates for any number of variety of early-universe research studies. However, they are still lacking two aspects of real SKA data. For one, the highest resolution observables correspond to a level of resolution ($\sim$ 0.3 arcmin for $z \approx 6$) that is unrealistic for SKA-low ($\sim$a few arcmin for $z \approx 6$, as seen in section 0.3.3). Indeed, thermal noise at high resolution will heavily obscure what can be seen, especially at high redshifts.

To simulate these effects, reducing the resolution is straightforward (a simple pixel averaging routine will do the trick[4]). However, modelling the noise is more challenging. One model for tomographic noise scaling is provided by Mellema et al. (2013), based off previous calculations by McQuinn et al. (2006). It gives the expected noise contribution to be:

$$\Delta T_b = \left(\frac{k_\perp}{2\pi}\right)\left(D_c^2 \times \Omega_{FoV}\right)^{1/2}\left(\frac{T_{sys}}{\sqrt{Bt_{int}}}\right)\sqrt{\frac{A_{core}A_{eff}}{A_{coll}^2}}\text{mK} \tag{1.6}$$

where $k_\perp$ is the angular scale, D_c is the comoving distance from the telescope to the source, Ω_{FoV} is the field of view of the smallest beam-formed receiver element, T_{sys} is the system temperature (used previously in equation 18), B is the bandwidth, and t_{int} is the integration time. As well, the three areas A_{core}, A_{eff}, and A_{coll} are the core area, the effective area for a receiver element, and the total collecting area ($A_{eff} \times N_{stat}$, the number of stations) respectively. This equation can be simplified (following the method outlined in Koopmans et al. 2015), as we can express Ω_{FoV} in terms of the wavelength we hope to observer, as $\Omega_{FoV} = \lambda^2/A_{eff}$, which will cancel out the other A_{eff} term. This brings us to:

$$\Delta T_b = \left(\frac{k_\perp}{2\pi}\right)\left(D_c \times \lambda\right)\left(\frac{T_{sys}}{\sqrt{Bt_{int}}}\right)\sqrt{\frac{A_{core}}{A_{coll}^2}}\text{mK} \tag{1.7}$$

As explained in Koopmans et al. (2015), the value $\frac{A_{core}}{A_{coll}}$ is approximately unitary for low resolutions (where the core region of densely packed receivers is the primary collecting area). However, for higher resolutions (higher $k_\perp$), this assumption no longer holds, and equation 1.7 will break down. In addition to this, it should be noted that $k_\perp$, D_c, T_{sys}, and B all have redshift dependence[5]. Therefore, in order to realistically estimate the $\frac{A_{core}}{A_{coll}}$ term, it is necessary to perform a full uv modelling of the sky using SKA specifications (Dewdney, 2015). Through using these, the full uv modelling was carried out by Benoît Semelin, and produced a table of noise values for $\Delta T_b(z, \Delta\theta)$. That is, the expected thermal noise (rms) for various values of redshift and angular resolution. Using this, a program for interpolating these values for any redshift and angular resolution was devised, and then noise could be applied to each redshift slice of the lightcone. Noisy lightcones produced through this method are presented in Semelin et al. (2017).

[4]Although, the ideal approach would have been to convolve the cubes with the instrumental response (a consequence of the station beam and antenna distribution) in Fourier space.

[5]The bandwidth B only has redshift dependence if matched to the angular resolution. One can also choose to keep it fixed.

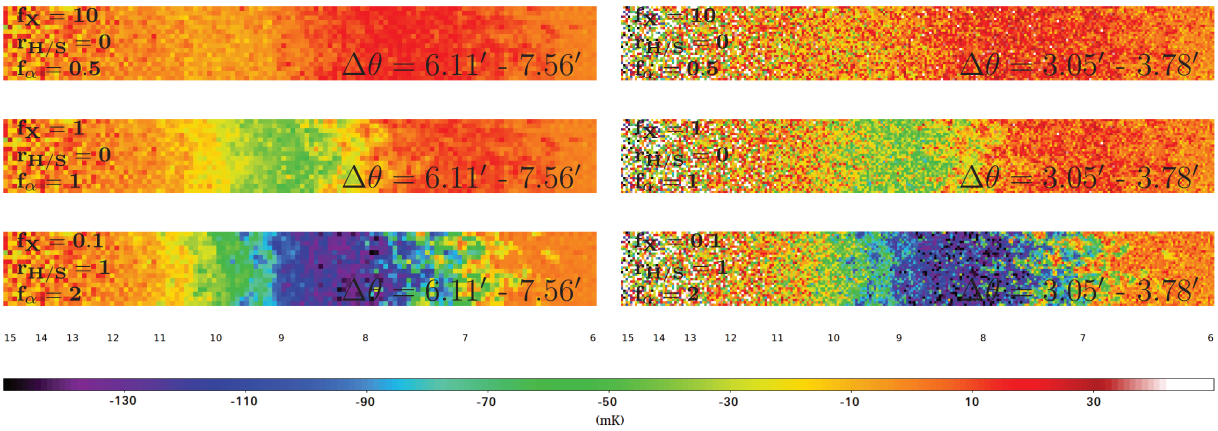


Figure 1.5 – *Example lightcone slices produced at SKA level resolution with realistic thermal noise. These lightcones are also presented in Semelin et al. (2017).*

After both scaling the lightcones to SKA resolution (a few arcmins) and adding noise, we arrive at lightcones that are realistic to what the SKA should be expected to see. These are presented in figure 1.5.

As well, adding SKA level noise to higher resolution lightcones demonstrates why this is unrealistic. Figure 1.6 shows a $64 \times 64 \times 512$ pixel lightcone (corresponding to an angular resolution of 3.13 cMpc, or $\sim$2 arcmin) after having added noise. All of the smaller structure has been lost, and only at the redshifts $\lesssim 9$ can any of the signal be detected (a consequence of the sparsity of long baselines). This could potentially still help trace the evolution of neutral hydrogen across the EoR, but for tomography, this resolution is impractical.

Spatial Noise Correlation

There is an important distinction to make about the manner in which we calculate noise. In figures 1.5 and 1.6 we have applied the noise directly to the image, meaning that the noise is not spatially coorelated. This gives a rough idea of the noise, but is not an ideal representation. A more robust method would be to add noise to the visibilities (Fourier space), and then performing a backwards Fourier transform. This would give a

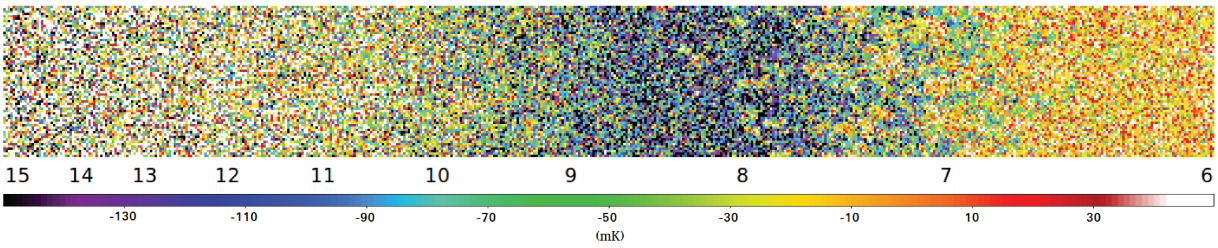


Figure 1.6 – *A higher resolution lightcone (3.13 cMpc resolution, or $\sim$2 arcmin) showing the signal nearly completely obscured by noise. Created with $f_X = 0.1$,$r_{H/S} = 1$,$f_\alpha = 2$.*

noise estimation in which neighbouring pixels are correlated, which will be the case for upcoming observations.

In fact, an updated version of the noise model presented here was developed in which applied noise is added to the Fourier space. This updated version was used in Bolgar et al. (2018) (for which I am 2^{nd} author), however was not finished by the time 21SSD was due to go online.

Finalized Database

With all 45 observables generated, and with a proper noise estimation included, the 21SSD database was made publicly available. All templates can be downloaded from 21ssd.obspm.fr, and for each sampled point in the parameter space the high-resolution observable is provided (1024x1024x8192 pixel lightcones), as well as SKA-low resolution observables (16x16x128 pixel and 32x32x256 pixel lightcones). For all of these, noisy and clean versions are provided, as well as the uv coverage code, and table of thermal noise vs angular resolution at various redshifts. For all lightcones, the procedure to generate them from the snapshots (figure 16) can also be carried out in the x,y, or z spatial dimensions. This allows for the creation of three lightcones for each observable.

The high-resolution templates are unrealistic for the expected capabilities of SKA-Low, though they are still excellent tools for studying the intricate dynamics of the EoR. This can be useful for training neutral networks, or for fine tuning numerical models. High-resolution observables will also be necessary for full end-to-end simulations (in which noise and resolution effects are included to simulate the full SKA pipeline). The SKA-resolution observables, especially those with noise, are more suited to give us an idea of what we should realistically be able to see. They also serve as a sandbox for testing observable diagnostics, as well as parameter reconstruction. For both high and low-resolution observables, we will now demonstrate some preliminary work to showcase the potential of 21SSD.

Part Two

Extracting Knowledge from 21SSD

CHAPTER 2

Extracting Knowledge from 21SSD

Revisiting Parameter Reconstruction

After having dedicated significant time and effort to creating the 21SSD database, it seems only logical to 'take it for a test drive', so to speak. As already discussed in section 0.6.6, a large reason that simulations are such an integral part of the quest to understand the Epoch of Reionization comes from their potential for 'Parameter Reconstruction'. The reader is reminded that the goal here is to use real observational data to uncover the values of astrophysical or cosmological parameters present in our observables. At present, a number of ongoing experiments (table 1) are focused on measuring the global sky-averaged signal, as well as studying the EoR power spectrum, both of which can help with setting limits on parameter values. However, the ideal tool — that which contains the most information — would be a full tomographic map.

In waiting for tomography experiments to go online, primarily the SKA, it is in our interest to devise and test methods to determine parameter values in order to have a framework in place for real data at the time of first light. However, in the absence of real data, simulation data can be readily substituted. In fact, this has already been attempted. Some tests have been made of 21 cm parameter reconstruction using the MCMC method (Harker et al., 2012; Greig & Mesinger, 2015, 2017b, 2018) (which has previously been applied to CMB as well, see Lewis & Bridle 2002, in addition to all Planck and WMAP papers), while others have focused instead on using machine learning and neural networks (Shimabukuro & Semelin, 2017; Kern et al., 2017; Schmit & Pritchard, 2018; Gillet et al., 2018). Yet, currently, most efforts have focused on mock data created by low-resolution semi-numerical simulations. The 21SSD database provides us with a perfect environment to begin to test the prerequisites of parameter extraction on more realistic mock data. Although we do not perform full parameter extraction in this chapter, we will begin to explore the concepts of defining 'distance' between 21SSD observables: the first step towards extracting parameters. This chapter lays the foundation, and sets up a theoretical model, for full parameter reconstruction on both high-resolution observables, or realistic mock SKA data.

To achieve this, section 2.2 lays out the procedure for calculating the multi-redshift power spectrum for a lightcone from the 21SSD database. Section 2.3 presents an alternative, and often overlooked diagnostic, to the power spectrum: the Pixel Distribution Function. Both of these are then used as the basis for calculating the distance between simulations in section 2.4, and the advantages and disadvantages of each are discussed. These methods are then applied to low-resolution mock SKA data in section 2.5, and finally the chapter concludes with a brief discussion on how these results could reconstruct parameters (section 2.6).

Table 2.1 – *Chapter 2 Contribution Breakdown*

Section	Me	Not Me
2.2	Majority	Minor formatting of figure 2.2
2.3	Majority	Adding contours to figure 2.4
2.4	All	—
2.5	Creating low-res lightcones	Noise curves on figure 2.9

Corresponding Publications: Second half of Semelin et al. (2017), and Eames & Semelin (2018)[a].

[a]IAU333 proceedings.

Lightcone Power Spectra

The standard, though by no means the only, technique to quantify the progression of reionization is through the 21 cm power spectrum. Showing the evolution of the power spectra as a function of redshift allows us to examine the dominant scale of fluctuations at different periods throughout the EoR. A simulation may save 'snapshots' of the boxes for various redshifts. A 3D Fourier transform can then be applied to them, to give the power spectra for each redshift. 21SSD saves snapshots between $z = 15 - 6$ at intervals of $\Delta z \approx 0.25$.

Yet, when dealing with real observational data, we will not be simply handed snapshots of the Universe. Our observations will be lightcones (section 0.6.5). Therefore, the question arises of how to derive the power spectra from the lightcones.

The simplest option would be to calculate the 2D power spectrum at each slice (a single pixel in width). When the promise of the power spectra was realized with relation to the EoR, this was also the envisioned approach to handling observational data. It was initially advocated that comparing the 2D power spectra at different frequencies would be an effective method for overcoming noise from foreground sources (Zaldarriaga et al., 2004; Bharadwaj & Ali, 2004). Building on this, it was soon realized that there is a large amount

of information along the line of sight. In fact, constructing the 3D power spectrum was shown to better overcome foregrounds than simply comparing 2D power spectra (Morales & Hewitt, 2004; Morales, 2005). Allowing some width in frequency-space, we can perform a 3D Fourier transform on what is essentially a snapshot. Although it can be argued that the information is relatively similar along the line of sight as perpendicular to it, the fact is that the 3D power spectrum increases statistics and thus decreases the level of noise.

However, there are two other problems with lightcones. Ideally, we would like the 3D space to be isotropic. Yet, peculiar motion along the line of sight creates anisotropy (Kaiser, 1987). This is sometimes called 'redshift space distortion'. With this extra width we also begin to incorporate information about the Universe at different times, with ionized bubbles becoming smaller at higher redshifts (known as the 'lightcone effect'). The relative importance of the lightcone effect will depend on the width of the slice upon which the power spectrum is calculated. Redshift space distortions will as well (although to a lesser extent), yet unlike the lightcone effect they also vary depending on the redshift of the slice. In general it has been shown that anisotropy created by peculiar velocity (redshift space distortion) is dominated by the physical anisotropy of a wide redshift range (Barkana & Loeb, 2005, 2006). This stronger source of anisotropy can, in fact, bear a characteristic imprint of the state of the IGM, and is hence not altogether undesirable (Zawada et al., 2014). Recent work has focused on reconciling or working around these two effects through novel methods of quantifying the EoR (see, for example, Kakiichi et al. 2017; Giri et al. 2018; Majumdar et al. 2018).

Reshaping

Ultimately, it was decided to take a wide slice at each redshift, such that the comoving size along the line of sight is equal to that of the height and width (a cube of equal comoving distance). The decision was motivated by the fact that this both vastly increases the information, as well as significantly facilitates the computation (performing a 3D FFT on a cube is simpler than on a rectangular prism). Yet the issue of scale still stands: the pixels correspond to bins of equal bandwidth, in order to have similar thermal noise. They must therefore be adjusted such that they have equal comoving thickness. This can be accomplished through firstly calculating the comoving distance to each slice along the redshift axis:

$$D_c = \frac{c}{H_0} \int_z^0 \frac{dz}{\sqrt{\Omega_m (1+z)^3 + \Omega_\lambda}} \tag{2.1}$$

where the Ω density parameters have the usual definition (the curvature density Ω_k is taken to be zero), and H_0 is the Hubble constant today, with values taken from Planck Collaboration et al. (2016a). Along the physical axes, the high-resolution lightcone has a width and height of 200 cMpc. Therefore, should we want to calculate the power spectrum at redshift z, we take all slices whose comoving distance falls into the range $D_c(z) \pm 100$ cMpc. Now, these pixels can be stretched such that the temperature brightness is distributed over the correct number of pixels for a cubical slice – 1024 for the high-resolution lightcone (through the standard method of calculating the fractions of the initial pixels that fall into the new grid). Figure 2.1 illustrates this procedure.

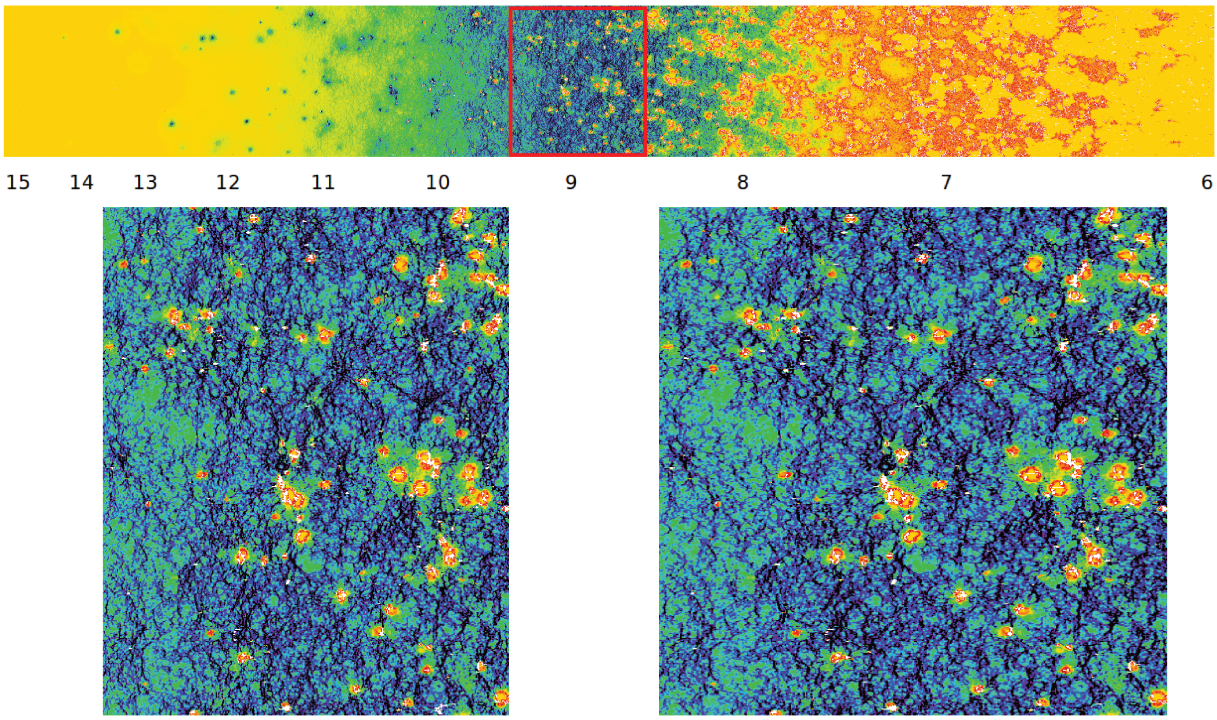


Figure 2.1 – *The process through which lightcones are spliced and reshaped, in preparation for calculating the power spectrum. This shows the process at $z = 9$: the sliced segment on the left, and the stretched-to-cube segment on the right.*

One consequence of taking thick slices as opposed to slices of a single pixel width is that the power spectra cannot be calculated at the extremities of the lightcone. Thin enough slices, even of multiple pixels, could give a good idea of the power spectra at the highest and lowest redshifts, but for the larger segments we use here this is not the case.

Calculating the Power Spectrum

Once a segment has been removed from the lightcone and re-binned into a regular co-moving grid, the Fourier transform can be applied in order to calculate the final power spectra.

Formal Definition

To quickly review the approach, the non-normalized Power Spectrum ($P_{21}(\mathbf{k})$) is related to the temperature brightness fluctuations as follows:

$$\left\langle \widehat{\delta T_b}(\mathbf{k}) \widehat{\delta T_b}(\mathbf{k}') \right\rangle = P_{21}(\mathbf{k})(2\pi)^3 \delta_D(\mathbf{k} - \mathbf{k}') \tag{2.2}$$

where $\widehat{\delta T_b}$ is the Fourier Transform of the field of T_b fluctuations, the angle brackets are the ensemble average, and δ_D is the Dirac function (see derivation in Risken & Frank 1996

section 2.4.3). Assuming spacial isotropy on large scales, $\mathbf{k} \to |\mathbf{k}|$ (a spacial average), and we set the Dirac function to 2π via Fourier series expansion.

$$P_{21}(k,z) = \frac{1}{4\pi} \int \left| \frac{\widehat{\delta T_b}(\mathbf{k},z)}{\mathbf{k}} \right|^2 dS \tag{2.3}$$

$$P_{21}(k,z) = \frac{1}{4\pi k^2} \int |\widehat{\delta T_b}(\mathbf{k},z)|^2 dS \tag{2.4}$$

To apply this to our data, we must first apply the Fourier transform.

Applying the Fourier Transform

This is done with the Intel Math Kernel Library (MKL) DFTI routine[1]. Although the DFTI functions are designed to handle 1D and 2D Fourier transforms, they do not come with out-of-box functionality for 3D Fourier transforms. Thus the data spacing has to be done manually. The 3D functionality, and the system for arranging the output data, was patched together based on information from a number of online forums and examples.

The MKL DFTI routine takes advantage of symmetry, and returns an array with dimensions $[x, y, \frac{z}{2}]$, in which entries are complex. Therefore, it is necessary to apply the following routine (a normalization step) to each entry to arrive at the real Fourier transformed cube:

$$\text{Corrected}(x,y,z) = \sqrt{\frac{(\mathbf{Re}[\text{Output}(x,y,z)])^2 + (\mathbf{Im}[\text{Output}(x,y,z)])^2}{(d_{\text{pix}})^3}} \times (d_{\text{Mpc}})^3 \tag{2.5}$$

in which d_{pix} is the side of a slice in pixels and d_{Mpc} is the size in cMpc. Although we are not the first to use the MKL DFTI for a 3D FT, we present the code nonetheless to save future researchers the trouble of forum hunting (appendix B.1).

Binning

The power spectrum is then created through the standard method of averaging the power in the spherical shells (although, in accordance with the MKL DFTI output, these are hemi-spherical shells). The code used to generate the power spectra, which is included as part of the 21SSD downloadable suite, offers three methods of calculating the bins. Should we want n_b bins, the i^{th} bin limit can be defined such that we assure:

- Equal volume
  i^{th} bin limit $= \frac{d_{\text{pix}}}{2} \sqrt[3]{\frac{i}{n_b}} \cdot k_{min}$

[1]See https://software.intel.com/en-us/mkl-developer-reference-c-dfticomputeforward.

- Linearity in k
  i^{th} bin limit $= (i + 0.5) \cdot k_{min}$
- Logarithmic spacing
  i^{th} bin limit $= 2\left(\frac{d_{\rm pix}}{4}\right)^{\frac{i}{n_b}} \cdot k_{min}$

where $k_{min} = \frac{2\pi}{d_{\rm Mpc}}$. There are advantages to all three methods. Equal volume bins assures that no bin suffers from more sample variance than any of the others. Logarithmic similarly reduces sample variance between bins, however it allows slightly more bins at small-k. Yet we found that neither of these two binning methods sufficiently sampled the small-k region — a region of strong importance for EoR simulations. Therefore, the choice was made to opt for a simple linear binning and accept the increased variance at low k. Although there are many more pixels in large-k bins, the small-k region is well sampled (we used $n_b = d_{\rm pix}/2$). Sample power spectra created through this procedure are shown in figure 2.2. We have included thermal noise following the method outlined in Mellema et al. (2013). The procedure has also been used to generate the power spectra presented in Semelin et al. (2017).

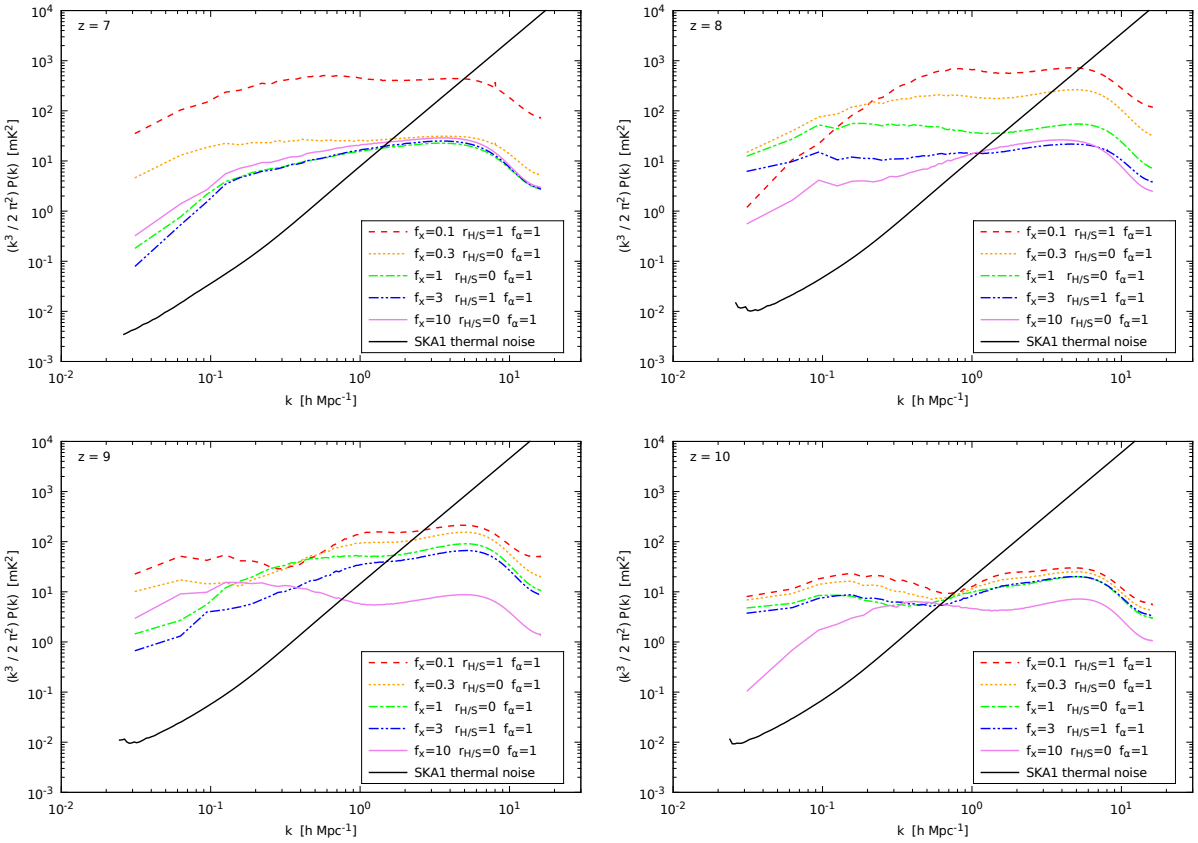


Figure 2.2 – *A selection of five observables (power spectra) from 21SSD. The black line represents thermal noise for a typical SKA survey, created using the procedure in Mellema et al. (2013).*

The ability of the power spectra in categorizing different EoR parameter configurations is evident. The rise of small structure (large k) at $z = 8$, for example. The arrangement of

the different observables in $P(k)$ also reflects the deepness of the 21 cm 'valley' (see figure 1.5). As well, the shape of the power spectra also changes markedly for large structure (small k), which is good news as this is the regime where noise poses less of a problem. We expect peculiar motion would also add more power to the larger k side of the spectra (smaller structure).

The power spectra also serve to illustrate some of the consequences of parameter adjustments that we outlined in discussing the lightcones (figure 1.4). For example, the more diffuse nature of the lightcone at redshifts $8 \gtrsim z \lesssim 9$ we see when $r_{H/S} = 1$ is also seen in the power spectra: the blue dot-dash line is above the green dot-dash line at these redshifts in figure 1.5, but nearly identical at other redshifts. To better see the effects of the parameter values on the power spectra, one can plot them as lines of redshift (each line corresponding to a fixed k value). Examples of this can be found in Santos et al. (2008); Baek et al. (2010).

Quasar Contributions

On an aside, this power spectrum routine was also used to study the contribution of quasars to reionization, with results presented in Bolgar et al. (2018). Figure 2.3, reproduced from said article, shows the quasar contribution to the power spectrum at $z = 8$ and $z = 10$. $f_{\rm duty}$ is the 'duty cycle' of the quasars, which can be thought of as the duration of the period of peak emission, and $f_{\rm corr}$ is a correction factor for the bolometric luminosity. The noise is for SKA1 with 1000 h exposure time.

Pixel Distribution Function

Although the power spectra certainly offer a powerful tool for simplifying and quantifying the progression of the EoR, they do have their shortcomings. The power spectra, on account of the reduction of information to the rms of the modules of the coefficient k-bins, only account for Gaussian. This means that two lightcones with differing non-Gaussian structure could have the same power spectra.

For this reasons, an alternative diagnostic was sought. We therefore explored using the Pixel Distribution Function (PDF): a 2D histogram in which bins are organized in terms of T_b and redshift (more precisely, bins are created such that they are linear in scale factor). Explicitly, the i^{th} bin limits for both variable are:

$$\mathrm{bin}_{a,i} = \left(\frac{i}{n_{b,a}}\right)(a_{final} - a_{initial}) + a_{initial} \tag{2.6}$$

$$\mathrm{bin}_{T_b,i} = \left(\frac{i}{n_{b,T_b}}\right)(T_{b,max} - T_{b,min}) + T_{b,min} \tag{2.7}$$

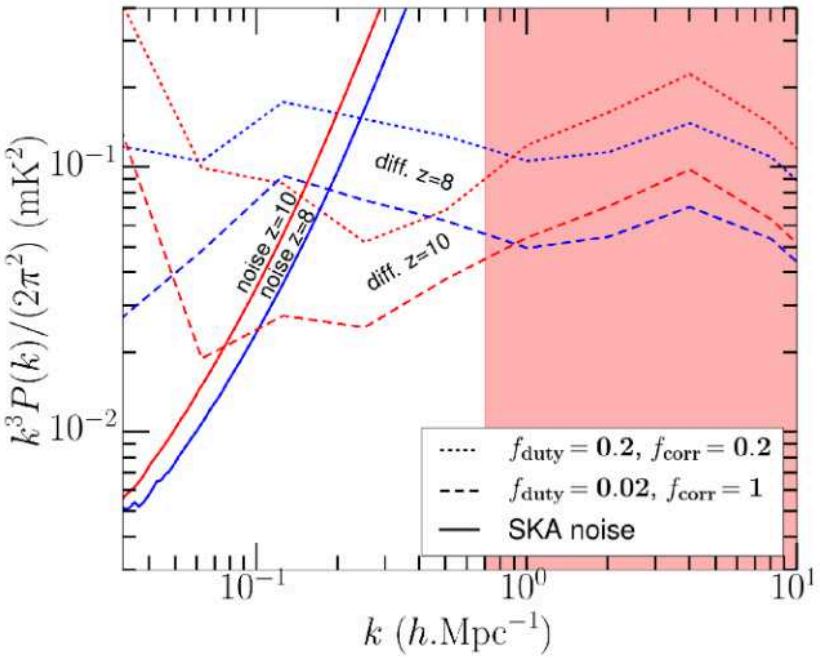


Figure 2.3 – *The contribution of quasars to the power spectrum for redshifts 8 and 10, as well as two different duty cycles. The shaded region is where the noise becomes higher than the EoR power spectrum. Reproduced from Bolgar et al. (2018).*

We then calculate which bin a pixel (x,y,z) falls into via:

$$T_{b,bin} = \text{int}\left(\frac{T_b(x,y,z) - T_{b,min}}{T_{b,max} - T_{b,min}} n_{b,T_b}\right) + 1 \tag{2.8}$$

For the scale factor bins the calculation is identical.

The relevant Fortran code is presented in appendix B.2, and the result are shown in figure 2.4 for 15 observables taken from the 21SSD database.

The PDF also proves to be an interesting diagnostic, whose shape changes noticeably between observables. For more subtle changes, the 1σ and 3σ contour lines help highlight the differences (containing 68.2% and 99.7% of the pixels respectively, for a given redshift). In general, the area of primary pixel concentration roughly traces the values of average T_b for a given redshift. The spikes in the purple (low pixel count) 'wings' are due to cosmic variance; more concretely, a spike corresponds to passing through a bubble which is more coupled in Lyα. If the boxsize was sufficiently high (to remove this sample variance) we would not have any spikes in these purple regions.

We can also adjust the PDF by subtracting the average T_b at each redshift. An example of two 'adjusted' PDF (for both a strong and weak reionization scenario) is given in figure 2.5. Such adjusted PDF correspond to what could be produced should the final SKA be unable to measure average T_b at each redshift (although it is expected that average T_b will be measurable through the antenna signal autocorrelation.).

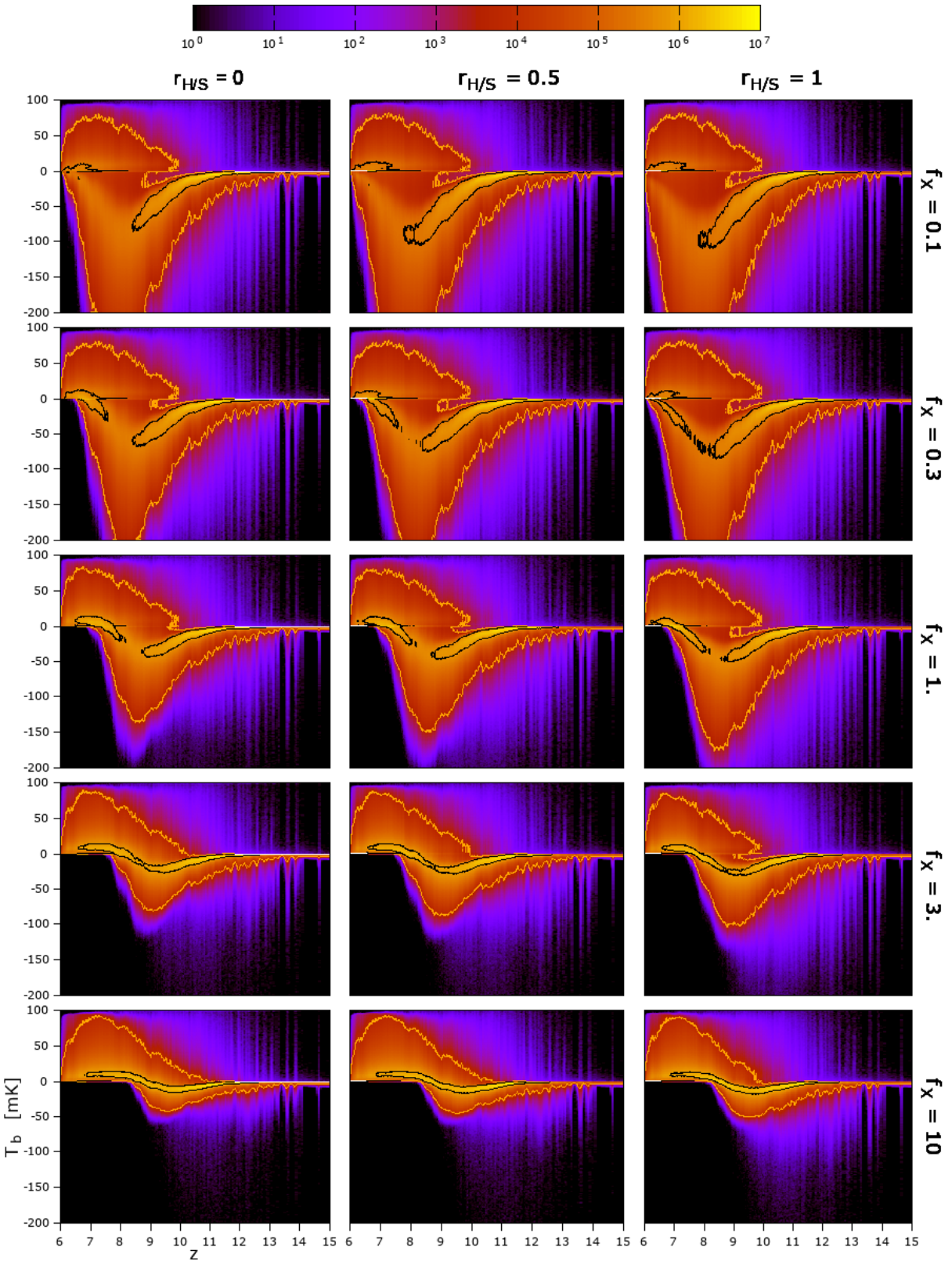


Figure 2.4 – *A selection of pixel distribution functions from 21SSD (those with $f_\alpha = 1$). For each PDF, the black and orange contours contain 68.2% and 99.7% of the pixels (for a given redshift). Note that the contours may seem discontinuous, however in these regions the pixels mostly fall along the 0 mK line. Multiple peaks for a given z can also lead to odd contour shapes.*

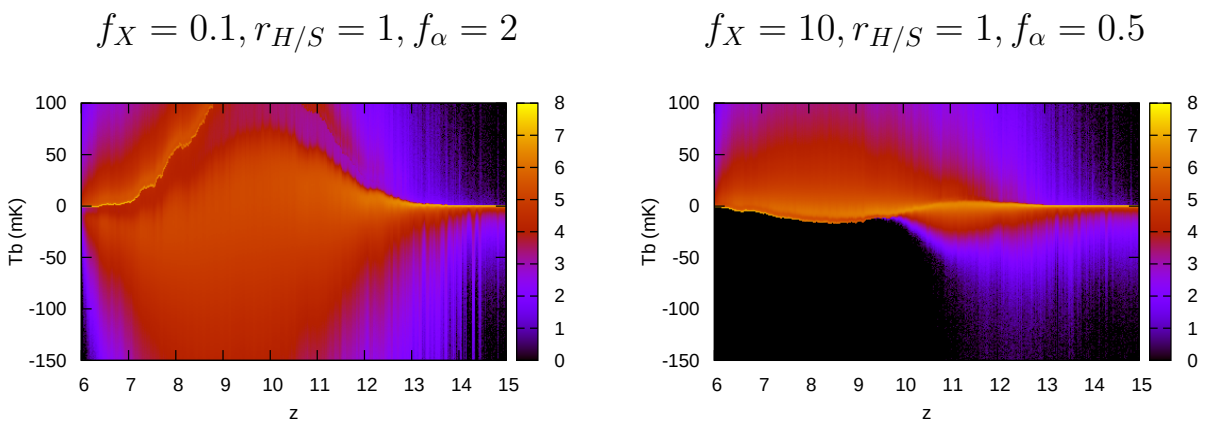


Figure 2.5 – *An example of adjusted PDF, in which the average T_b has been subtracted for each redshift. Two observables have been chosen to represent strong and weak reionization scenarios.*

It should be noted that the thickness of each scale factor bin is still free to be adjusted. Like with the power spectra, wide bins have less noise, but also begin to lose their ability to represent specific redshifts.

Defining Distance

With both the power spectra and the pixel distribution functions offering methods to simplify different EoR progressions, the question then becomes how to quantify the difference between two observables. We adopt a simple L_2 norm for this end. This formalism has been used previously to compare predictions and observations in MCMC implementations, and seems like a logical choice here. Should the power spectra be used, the distance between observables i and j is calculated as:

$$\mathrm{D_{PS}}(i,j) = \sqrt{\int \Big(\mathrm{PS}_i(k,z) - \mathrm{PS}_j(k,z)\Big)^2 \, dkdz} \tag{2.9}$$

where PS_i is the i^{th} observable, k is the inverse distance in $\mathrm{h} \cdot \mathrm{cMpc}^{-1}$, and z is the redshift. This can be thought of as calculating the integrated volume[2] between the two power spectra 'surfaces'. In the case of the PDF, the calculation is similarly:

$$\mathrm{D_{PDF}}(i,j) = \sqrt{\int \Big(log_{10}(\mathrm{PDF}_i(T_b,z)) - log_{10}(\mathrm{PDF}_j(T_b,z))\Big)^2 \, dkdz} \tag{2.10}$$

the only difference being that the logarithm of the PDF is taken beforehand. Unlike the Power Spectrum, the range of each PDF covers 7 or 8 orders of magnitude, making the logarithm useful to give weight to the low-pixel (purple) regions of the distributions.

[2]Or, more exactly, the volume between the spared surfaces of two power spectra.

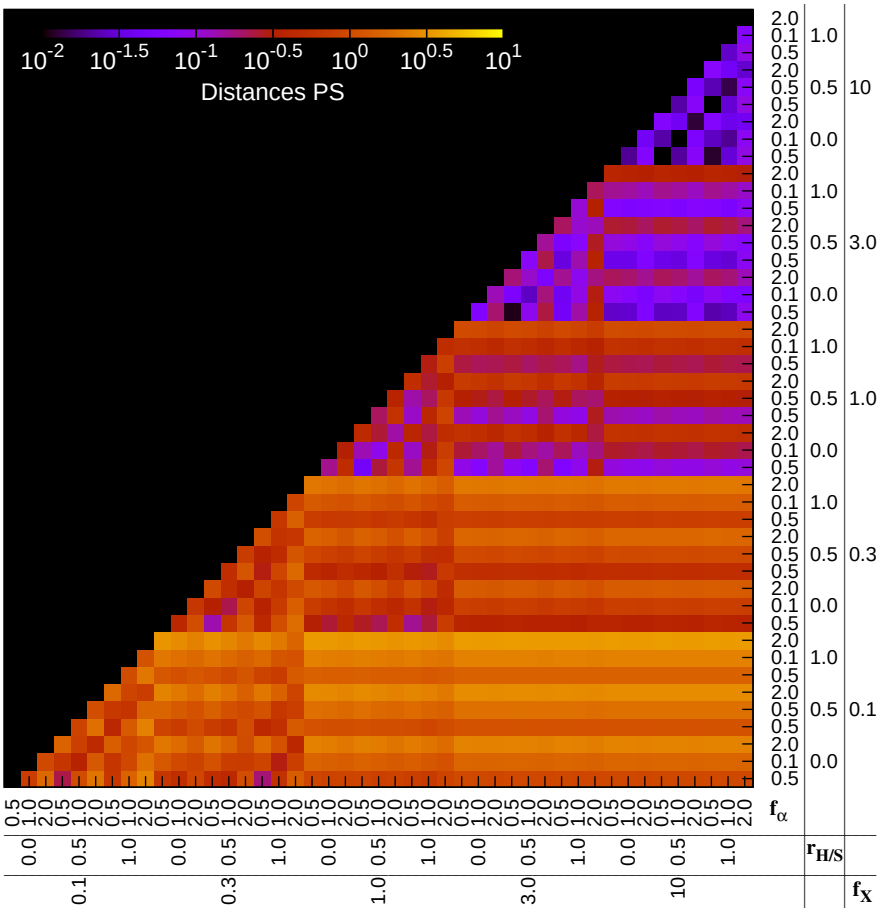


Figure 2.6 – *The distance between the power spectra of the observables in 21SSD.*

One is justified in thinking that it would be logical to take into account the statistical error (noise, e.g.) associated with each bin (both for the PDF and the power spectra) when calculating distance. This is a valid point. We will return to it, and justify our error omission, in the second part of section **??**.

Power Spectrum Distance

The distance between the power spectra of 21SSD simulations is presented in figure 2.6. We see that this distance is the most extreme between the simulation with parameters $[f_X, r_{H/S}, f_\alpha] = [0.1, 1.0, 2.0]$ and simulations made with different values of f_X, as illustrated by the horizontal yellow bar. This is explained by the exceptionally deep EoR in this particular simulation (see figure 1.4). In general, it is seen that variations between each 'block' of simulations made with equal f_X are significant. However, within each of these blocks, there is little variation. This seems to suggest that using the power spectra to calculate distances discriminates well for f_X, but less powerfully for $r_{H/S}$ and f_α.

Pixel Distribution Function Distance

The PDF proves to be a very different diagnostic, which can be easily seen in figure 2.7. Firstly, comparing the PDF leads to a smaller range of values for distance. This is because each individual PDF has the same integrated volume (a result of a constant number of pixels in each simulation)[3]. Therefore the distance between two is somewhat more constrained. Conversely, the power spectra can be orders of magnitude lower or higher than others, depending on the ionization scenario. This difference explains the fact that the power spectra distances fall between three orders of magnitude (10^{-2} - 10^{1}) while the PDF only cover roughly one order ($10^{-0.5}$ - $10^{0.5}$).

In addition to magnitude, the PDF are also very different in form. There is little variation from one 9x9 'block' of constant f_X values to another, nor between 3x3 blocks of constant $r_{H/S}$, yet there is substantial difference between simulations made with different f_α values. This is seen in the different magnitudes of neighbouring pixels.

Comparing Distance Definitions

To properly examine where the different diagnostics excel, a comparison is necessary. For a given pair of simulations i & j we can define the *Distance Ratio* as:

$$\text{Ratio}(i,j) = max\left(\frac{\widetilde{\text{D}}_{\text{PS}}(i,j)}{\widetilde{\text{D}}_{\text{PDF}}(i,j)}, \frac{\widetilde{\text{D}}_{\text{PDF}}(i,j)}{\widetilde{\text{D}}_{\text{PS}}(i,j)}\right) \tag{2.11}$$

To allow for a proper comparison, both distances are divided by their respective average distance. These adjusted distances are denoted $\widetilde{\text{D}}$, and defined by:

$$\widetilde{\text{D}}(i,j) = \frac{\text{D}(i,j)}{\frac{1}{\text{N}}\sum_{i,j=1}^{\text{N}}(\text{D}(i,j))} \tag{2.12}$$

where N is the number of simulations. We can see the result of this in figure 2.8. Simulations made with f_X = 1.0 - 10 are deemed further from other simulations when using the PDF distance as opposed to the PS distance, and the structure in f_α is also seen to stand out, especially at $f_X = 0.1$.

The main information in this map is that the difference in discriminating power is especially strong between observables with high heating levels (upper part of the triangle). As we are taking the maximum of the two ratios, this tells us that one is outperforming the other in this region. Judging from figures 2.6 and 2.7, this appears to be the PDF (the values in this region are noticeably smaller for the power spectra distances).

[3]Although it is true that we are looking at the logarithm of the PDF, this will be true regardless. The nature of the multi-redshift PDF means that any two will always overlap in some places, which is not true for the power spectra.

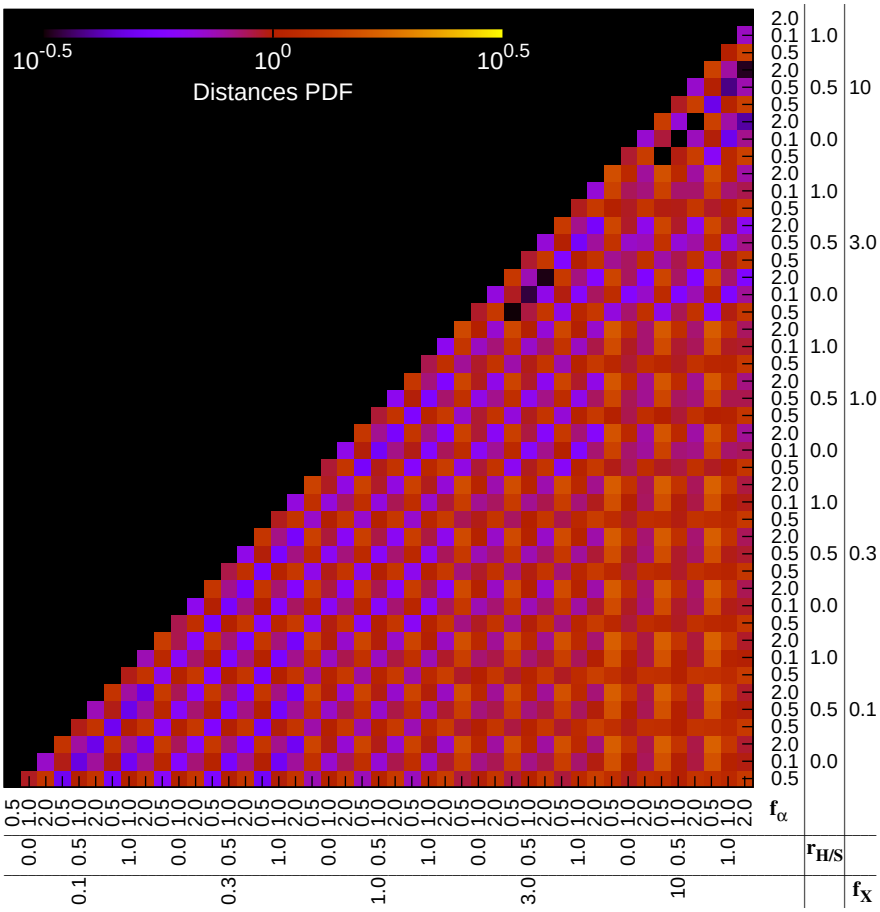


Figure 2.7 – *The distance between the pixel distribution functions of the observables in 21SSD.*

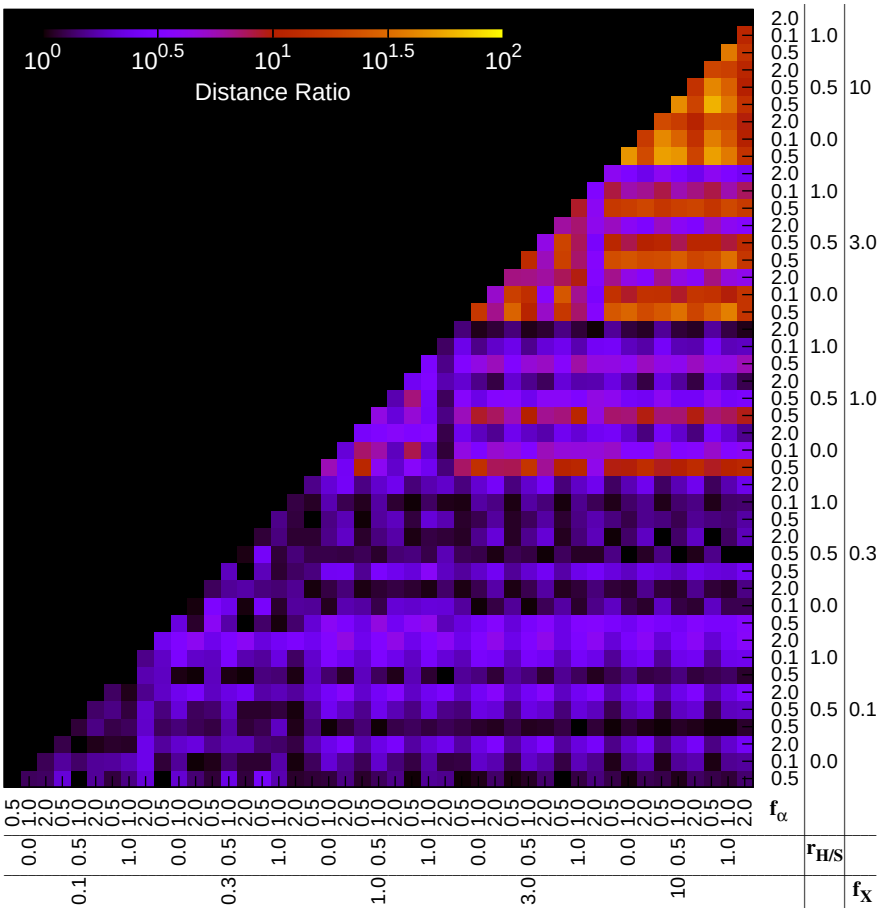


Figure 2.8 – *A comparison of the two methods for defining distance (power spectra and pixel distribution function).*

It should also be remembered that the two diagnostics have different sensitivity to the noise, and even in regions where the PDF appears to perform better, this advantage may be lost to the noise the PDF experiences over the power spectrum. A final conclusion would need to account for the noise.

What is clear is that both methods seem to offer a somewhat complementary pairing. The power spectrum appears to be more sensitive to changes in f_X (the clear differences from one 9×9 block of constant f_X to another), while the PDF weights changes in f_α more heavily (the difference between one pixel to its neighbours). This is fortunate in a way, as it gives us the power to probe two variables more effectively. This being said, neither diagnostic seems to react especially strongly to changes in $r_{H/S}$ (we see little difference between neighbouring 3×3 blocks of constant $r_{H/S}$), indicating that either a new method to constrain this parameter will be needed, or that the variable is simply not sufficiently consequential to reionization observables to realistically be extracted (obviously this needs to be verified with other simulations).

Distance Magnitude

As stated, the multi-redshift power spectra cover roughly 3 orders of magnitude, while the PDF covers only $\sim$1. This leads to an interest effect, where changes in one variable can eclipse changes in the others. For our parametrization changes in f_X cause a large change in order of magnitude of the power spectra (though not necessarily the slope). For regions where f_X is held constant (the points closest to the diagonal in figure 2.6) we see that the f_α differences are apparent (neighbouring pixels are different) as they are when the PDF is used (figure 2.7). This hints at another potential advantage of the PDF, which is that it is less likely for one parameter to dominate. It could perhaps be possible to apply a normalization to the power spectra to account for this effect (we could similarly exponentiate PDF to perhaps better distinguish f_X, however there already appears to be some sensitivity to this parameter).

Mock SKA Data

The advantages of these diagnostics in both categorizing, as well as comparing, different EoR observables, comes across clearly after having seen the results in section 2.4. However, it should be recalled that both the power spectra, as well as the pixel distribution functions, were created using the highest resolution lightcones (a few arcsec). For the power spectra, the noise curve already corresponds to realistic SKA noise, and therefore they already represent a realistic idea of what to expect. However, the PDF will lose their details if a lower resolution lightcone is used, and therefore figure 2.4 represents the idealized potential of such diagnostics.

On account of this, we reproduced the PDF using the low resolution lightcones. This is shown in figure 2.9. Although the shape persists (roughly tracing the average T_b values

for a given redshift), much of the information that may have been contained in the faint wings (seen in figure 2.4) has been lost. Adding noise is the equivalent of smoothing the PDF in the T_b direction. As well, when the thermal noise is included, there is very little hope of tracing the EoR through PDF past $z \approx 12$ at 6' resolution (and perhaps $z \approx 10$ at 3' resolution).

Attempting Parameter Reconstruction

The framework laid out in this chapter sets the stage for future work in parameter reconstruction. The 21SSD observables are sufficiently detailed to capture the intricacies of the Epoch of Reionization, and paired with the two distance measures they could prove powerful tools to add to current parameter inference efforts (section 0.6.6). Here we outline the rough steps through which this goal could be achieved:

1. Ideally, the parameter space should be better sampled, as 45 observables is a relatively sparse sampling with which to form the foundation of parameter extraction. A database of ∼100observables would represent a solid update on the current 21SSD, and likely be sufficient for preliminary parameter extraction attempts.

2. It would also be worthwhile to test other parameters. As was seen in section 2.4.3, the three used here all showed different sensitivities to the PS and PDF methods of comparing observables. Perhaps we could find a choice of parameters which all vary strongly for one of the two methods, which would help us to eventually set tighter constraints.

3. With more observables, the corresponding SKA resolution power spectra and pixel distribution functions could then be created (section 2.5).

4. These would allow for larger distance maps (figures 2.6 and 2.7), using both distance definitions (sections 2.4.1 and 2.4.2).

5. Real incoming tomographic data could then be compared to the observables in the 21SSD database (through calculating the corresponding power spectrum and PDF) using neural networks (recall from section 0.6.6 that MCMC involves recalculating the observables at each step, and is therefore computationally unrealistic for high-resolution codes).

6. Finally the relevant parameter values could be determined.

However, this quick overview is a tremendous oversimplification. In reality there are a number of things to consider. Future versions of the 21SSD database could (and should) also explore other parameters, allowing for corresponding higher dimensional parameter spaces (recall in section 1.3.1 an 11 dimensional sampling has been studied in Kern et al. 2017). There is also the matter of how to consolidate the diagnostic abilities of the power

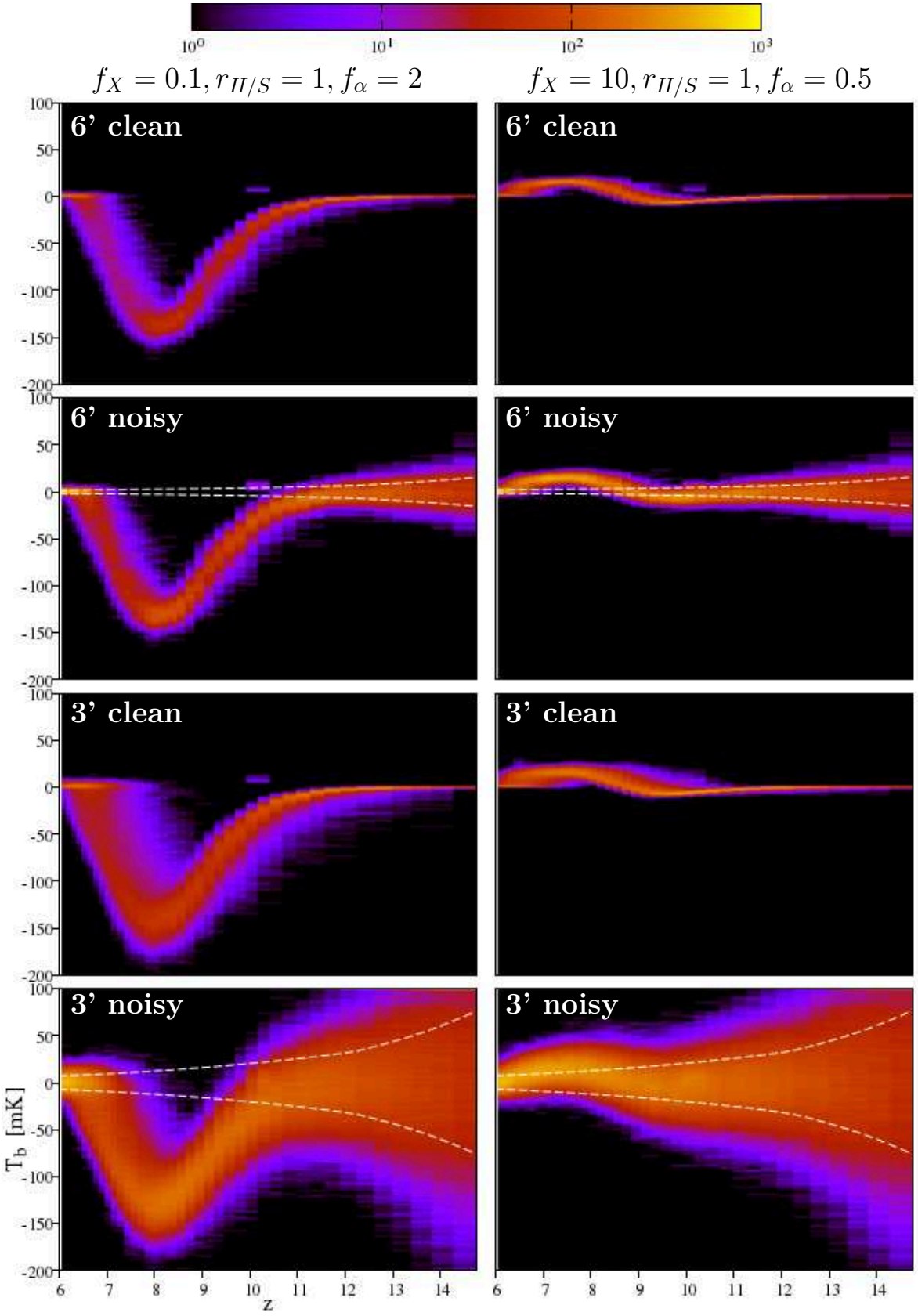


Figure 2.9 – *A selection of pixel distribution functions created at SKA resolution. Thermal noise is represented by the dashed white line (see section 1.4 for an overview of noise modelling).*

spectra and pixel distribution functions. Certainly, as discussed in section 2.4.3, both methods have somewhat complementary advantages, and it seems logical to use both. Yet whether they should be applied separately to real data, or if a hybrid version should be conceived (in which both distances are somehow weighted and combined), remains an open question. There are still other considerations, such as future refinements to the LICORICE code in preparation for SKA data, how to best train neural networks in parameter reconstruction, and the possibility of standardizing a parameter space and coordinating with other contemporary simulations to assure a more thorough sampling.

Needless to say, there is certainly work to be done. For the time being, the astute reader may have noticed that at the core of many of the above-stated questions is the notion of parameter space sampling. It is therefore along this avenue that my research embarked during the final year of my PhD, and thus the following chapter will begin by elucidating exactly how parameter space sampling is intimately tied to resolving many of the above considerations.

Part Three

Finding an Optimal Parameter Space Sampling

CHAPTER 3

Finding an Optimal Parameter Space Sampling

The Problem at Hand

With 21SSD, as with any database of EoR models — or indeed, simulation models in general — there is an understandable interest in sampling a parameter space as thoroughly as possible. This entails sampling as many parameter configurations as realistically feasible. In addition, as mentioned previously, we ideally would like to explore parameter spaces with as many parameters as possible, to eventually constrain a large number at once. Yet properly sampling a high-dimensional parameter space, especially for full dynamics codes (such as LICORICE), requires immense amounts of both time and computational resources. For this reason, simply sampling heuristically and 'hoping for the best' (in terms of capturing the full intricacies of a parameter space) is not desirable. One is justified in asking if there could be an 'optimal' sampling: that is, a sampling (with a set number of points) that optimizes the resulting database towards achieving a predefined goal. This question will form the foundation of this chapter.

Firstly, before attempting to answer this question, we will define what is meant by optimal (section 3.2) and look at previous sampling techniques (section 3.2.1). Section 3.3 will describe the preliminary steps of creating a 'prototype parameter space' on which to test sampling optimization algorithms. The first of these algorithms, which is based on metrics and linear algebra, is outlined in section 3.4. The second, which employs a more physical approach to the problem and does not assume a grid is necessary, is then laid out in section 3.5. These two methods are compared in section 3.5.1, and finally section 3.6 showcases some interesting results in applying these optimized samplings to neural network training.

Table 3.1 – *Chapter 3 Contribution Breakdown*

Section	Me	Not Me
3.2	Majority	Some discussion
3.3	All implementation and figures	Some theoretical input
3.4	All implementation and figures	Theoretical framework
3.5	Creating models and calculating distances	Majority
3.6	Creating various samplings (test data)	Majority

Corresponding Publication: Eames et al. (in prep).

Relevant Definitions

It is worth taking a moment to cement some useful terminology before continuing.

The parameter space consists of a choice of parameters, and the range of values we allow them to assume. A point in this parameter space is effectively a set of coordinates. If we take the example of three parameters $\theta_1, \theta_2, \theta_3$, then our parameter space is all points that fall into the ranges $[\theta_{1,min} : \theta_{1,max}], [\theta_{2,min} : \theta_{2,max}], [\theta_{3,min} : \theta_{3,max}]$.

A sampling is a finite choice of points within a given parameter space. Continuing the above example, the i^{th} point would have the form $\vec{\theta_i} = (\theta_{1,i}, \theta_{2,i}, \theta_{3,i})$, where values fall within the enforced bounds.

A model is a framework (theoretical or numerical) that is used to compute an observable for a given point in the parameter space.

An observable[1] is a quantity that can be computed from the observation. It is a prediction of the observation as computed via a model. The power spectra is an example of an observable, and should we choose this convention then a point $\vec{\theta_i}$ in our parameter space would have corresponding observable $\mathrm{PS}(\vec{\theta_i})$. Other examples of observables could be PDF, lightcones, the global signal, etc.

The space of observables[2] is the space spanned by these observables. For example, a power spectrum estimated in 10 k-bins at a single redshift inhabits a 10-dimensional space of observables.

[1] It is worth restating that other authors have used the word 'model', however this can be misleading, as 'model' traditionally refers to a specific theoretical framework used in simulating a phenomenon. Yet here the 'modelling' does not change, only the parameters used. We therefore feel that 'observable' is a more cautious choice.

[2] This word ordering is intended to bypass the pitfall that would have been 'observable space'; a choice which would have easily been misinterpreted as the space which *can* be observed.

The hypersurface of predictions by the model[3] is a space embedded in the space of observables. The model acts as a map between the the parameter space and this hypersurface (both of which have the same dimension). The geometry can be quite different from that of the parameter space. For example, the two closest points in a sampling may not necessarily transform to give the two closest observables (which also depends on the choice of 'distance' between observables). In section 3.4.1 we will define a metric to quantify the geometry of this hypersurface.

In mathematical terms, we could think of the model (simulation) as a function which maps points from the parameter space to observables on the hypersurface (within the space of observables). The nature of this function will depend on the observable, although we should not expect it to be injective (multiple points in the parameter space could give identical power spectra) nor surjective (there may be regions in the hypersurface that do not have a corresponding point in the parameter space).

Optimal Sampling

When we envision an 'optimal' sampling, we refer to a parameter space sampling for which all observables are equally different from 'neighbouring' observables (if the concept of neighbours exists). Should the points in parameter space be organized on a grid, we would ideally like for any two neighbouring points (along any axis) to map to equally different observables (in the sense that they are at equal distance for a given distance diagnostic). Should we not require a grid system, the definition becomes more nuanced, and we will return to this case in section 3.5.

In the mathematical framework outlined above, we are looking for a sampling of the parameter space that maps onto a homogeneous and isotropic sampling of this hypersurface (according to our distance definition). It should also be noted that the specific configuration of an optimal sampling depends on the definition of distance. For example, a sampling that is optimal when the power spectra is used as the measure of distance, will not necessarily be optimal when the pixel distribution function is substituted. This will be discussed in section 3.7.

As a first step, we will satisfy ourselves with 'optimal' referring to equally distant observables, arranged in a grid (one which is not necessarily Cartesian). The hope is that this will assure a better chance of exploring the full range of EoR scenarios, as well as providing the best training data for neural networks attempting parameter reconstruction.

Previous Sampling Methods

There are a number of ways we may decide to sample a parameter space:

[3]Note that here we use 'model' in the traditional sense: the theoretical framework implicit in a simulation.

- Linear (Regular) Sampling (on a grid)
- Logarithmic Sampling (on a grid)
- Random Sampling
- Latin Hypercube Sampling
- n-Sphere Packing
- Heuristic Sampling

Spacing parameter values linearly or logarithmically along an axis is logical if we expect the observable to change in a similar manner. If we recognize that moving within a region of the parameter space has a strong effect on the resulting observables compared to other regions, it is natural to oversample this region, approximating the sampling heuristically ('by eye'). When the relationship between the parameters and the resulting observables is not immediately obvious, as is often the case for EoR simulations, one can take a random sampling. Yet this risks oversampled and undersampled regions due to both shot noise and intrinsic properties of the parameter space. Therefore, authors have previously used the 'Latin Hypercube' method to avoid this.

Latin Hypercube Sampling

Latin Hypercube Sampling (LHS) — as developed by Eglajs & Audze (1977); McKay et al. (1979); Iman et al. (1981) — requires that no two points in the parameter space share any parameter values. In two dimensions it is equivalent to placing a number of rooks on a chessboard such that no two rooks can capture each other. Maximum spacing[4] between samplings can be paired with LHS to avoid aberrant cases, such as sampling only along one of the diagonals (Morris & Mitchell, 1995). This is called Orthogonal LHS. A basic example can be found in figure 3.1.

LHS has been found to perform marginally better than regular grid sampling in training a statistical emulator[5] on a sparsely sampled parameter space in 3D (Heitmann et al., 2009), or 4D (Urban & Fricker, 2010). For neural network training, Schmit & Pritchard (2018) did not find any notable difference between LHS and grid sampling, however they use the 3D parameter space of Greig & Mesinger (2015), and expect LHS would excel for higher dimensions. The higher dimensional benefits of LHS are shown in Kern et al. (2017).

[4]Note that this is the distance between parameter values, and *not* between the resulting observables.

[5]"A fast proxy for a complex computer model which predicts model output at arbitrary parameter data from a limited ensemble of training data." As described in Urban & Fricker (2010).

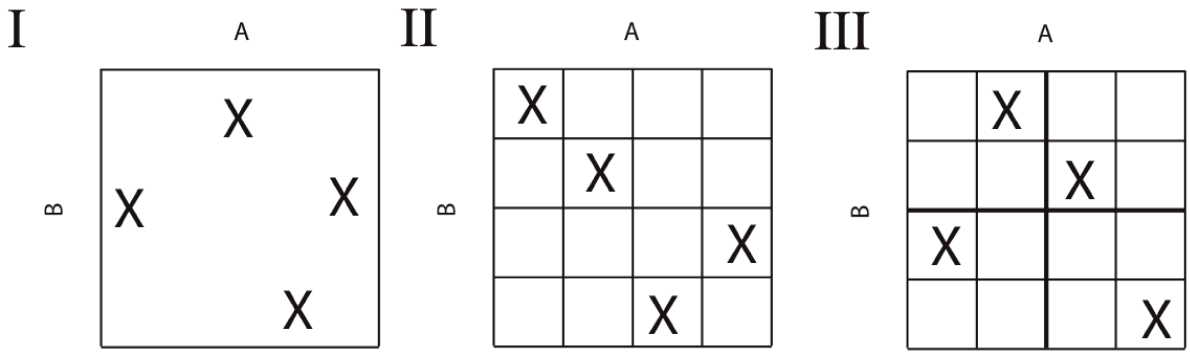


Figure 3.1 – *Visualization of Latin Hypercube Sampling. I – Random Sampling. II – LHS. III – Orthogonal LHS (in which an additional constraint is that each section must be sampled). Image Credit: wikipedia.*

n-Sphere Packing

Another idea is to consider sampled points to be individual spheres (in 3D). The goal is then to pack as many into the space as possible (sometimes called 'Cubic Closed Packing', or CCP). A more precise or coarse sampling is obtained by varying the radius of each sphere. This process has an analogous constriction at higher dimensions (n-spheres in an n-cube), however the mathematics becomes more difficult (Maimbourg et al., 2018). See Belitz (2011); Belitz & Bewley (2013) for an overview.

A sampling method to be outlined in section 3.5 converges to CCP in the case of a Euclidean metric.

Circumscribed Hypersphere

One may be less interested in the dynamics at the corners of the parameters space, which may represent 'extreme cases' in which all parameters have values that are at the limit of what is realistic. To discard these cases, we can inscribe an n-sphere within the n-cube parameter space, and only sample from within this reduced region (using any of the above sampling methods).

This technique was used by Kern et al. (2017), who points out that although the ratio of the area of a 2D square to a circumscribed circle is only $4/\pi$, this increases dramatically along with the number of dimensions. The group uses an 11 dimensional parameter space (section 1.3.1), at which point the ratio of an 11-cube to a circumscribed 11-sphere is over 1000.

Creating a Fiducial Parameter Space

Although the above mentioned sampling methods may have their advantages in certain situations, there is no reason to believe we cannot do better. As well, besides heuristic 'by

eye' sampling techniques, parameter space sampling has always been based on the distance between parameters, as opposed to the distance between the resulting observables. For this reason we will now attempt to develop a sampling technique that is optimal[6] in terms of spacing the resulting observables.

It is clear that we will need to explore large, well sampled parameter spaces. A sparse sampling, such as 21SSD, would be insufficient to truly draw conclusions on the effectiveness of any sampling method put forth. It will also be necessary to move points around within this parameter space, possibly many times. Each time a single point is moved, it requires the simulation to be re-run to create the corresponding observable. Combining this with a satisfactory sample size will require a daunting number of simulation runs, likely on the order of thousands.

Because of this requirement, full numerical codes such as LICORICE are simply out of the question. We therefore use 21cmFAST (Mesinger et al., 2011), described in detail in section 0.6.3. Note that the necessity of a fast semi-numerical code only applies to the development of an algorithm for optimal sampling. Once the algorithm is designed and tested, it can then be easily used to optimally sample with full numerical, high-resolution, codes such as LICORICE, albeit with sparse sampling (recall that a sampling can be optimal even for a sparse number of sampled points). We will return to this in section 3.7.

Parameter Definitions

To enable easy future synergies (e.g. comparing parameter reconstruction results), we choose to use the 3 dimensional parametrization used previously for 21MCMC (Greig & Mesinger, 2015). This was mentioned in section 1.3.1, however here we quickly summarize the three parameters.

- ζ, the ionizing efficiency of high-z galaxies:

$$\zeta = 30 \left(\frac{f_{esc}}{0.3}\right) \left(\frac{f_\star}{0.05}\right) \left(\frac{N_\gamma}{4000}\right) \left(\frac{2}{1+n_{rec}}\right) \tag{3.1}$$

  where, for each bracket moving left to the right, the variables are the ionizing photon escape fraction, the fraction of galactic gas in stars, the number of ionizing photons produced per baryon in stars, and the typical number of times a hydrogen atom recombines. The parameter should be thought of for an 'umbrella parameter' for a number of effects that play into the ultimate reionization rate. In 21cmFAST, this parameter ultimately decides whether, for a given number of sources, the IGM will be ionized. Setting it to a low value will result in reionization taking place at late redshifts (or not at all). High values result in early reionization.

[6] Section 3.2.

- $R_{\rm mfp}$, the mean free path of ionizing photons within ionized regions[7].
  This parameter dictates the speed at which ionized regions around ionizing galaxies grow.
- $T_{\rm vir}$, the minimum virial temperature of star forming halos.
  It can be though of as an ionization switch, controlling when halos begin to ionize their surroundings. Conversely to ζ, setting it to high values will result in a neutral universe up to very late redshifts.

See Greig & Mesinger (2018) for full definitions.

Parameter Ranges

The ranges are taken to be $\zeta \in [10, 200], R_{\rm mfp} \in [10, 75]$ (cMpc), $T_{\rm vir} \in [8 \times 10^3, 10^5]$ (K).

ζ is varied within a range similar to in Greig & Mesinger (2018), which is higher than in previous versions of 21MCMC to include early bright galaxy reionization scenarios (although they go up to $\zeta = 250$, we opt for a lower value for a more closely spaced sampling).

The mean free path in ionized regions has previously been shown to have little effect on the observables above $\gtrsim 15$ Mpc (Sobacchi & Mesinger, 2014). However, our goal here is not to include these findings regarding $R_{\rm mfp}$, but rather rediscover them in the geometry of the hypersurface of models. For this, we take the much higher values of 75 cMpc (compared to 25 in Greig & Mesinger 2018). For the lower bound, however, we take a slightly more conservative value of 10 cMpc (compared to 5), which is the result of discussion on the dynamics at this smaller scale.

For the virial temperature our range ($[8 \times 10^3, 10^5]$ (K)) also differs on for both the higher and lower limits from 21MCMC ($[10^4, 10^6]$ (K)). We allow a slightly lower minimum virial temperature, as radiative cooling can allow for star formation in smaller (cooler) halos (although the star formation efficiency drops to $\sim$20% at $\sim 10^2$ K, it is still likely reasonably at 8×10^3 K; see Kimm et al. 2017). On the other end, there has been hints that the minimum $T_{\rm vir}$ is high ($\sim 10^6$ K) for EoR redshifts on account of metallicity effects suppressing star formation (Kuhlen & Faucher-Giguère, 2012). Yet these higher values lead to very little variation in PS (on account of the very late ionization), and hence we choose a slightly lower one, to better test comparisons between the resulting observables.

The initial 'fiducial' sampling consists of 10 values, spaced logarithmically[8], for each parameter; this results in a total of 10^3 points (table 3.2).

[7] More precisely, this is the 'effective horizon' of photos set by the sub-grid constraints of 21cmFAST. The true mean free path tends to be slightly larger than this value on account of resolution effects (Mesinger et al., 2011).

[8] The choice of logarithmic, as opposed to regular, sampling is based on the expectation that parameter variations will impact the resulting observables more at the lower ends of the bounds. However, this may not be the case. Still, developing an optimal sampling algorithm that functions well on an initially logarithmic sampling, and adapting it to work on a regular sampling is much simpler than going the other way.

Table 3.2 – *Fiducial Sampling*

Parameter	Explored values
ζ	10, 13.95, 19.45, 25.14, 37.14, 52.82, 73.68, 102.78, 143.37, 200
$R_{\rm mfp}$	10, 12.51, 15.65, 19.57, 24.48, 30.63, 38.32, 47.93, 59.96, 75
$T_{\rm vir}/10^4$	0.8, 1.06, 1.40, 1.86, 2.46, 3.25, 4.31, 5.70, 7.55, 10

21cmFAST Cloning and Paralellization

Even with the speed of 21cmFAST, running 1,000 versions of the code is a significant challenge (a second iteration, to be described in table 3.3, necessitated 2,400 runs). Each run takes on the order of a few hours, and thus running thousands of iterations without paralellization would require an unrealistic time frame. Yet, even when running in parallel, the process of copying the 21cmFAST files one-by-one, and manually changing the variables, would be absurd.

To overcome this, a code was developed, written in c, to automate the process. A 'master' 21cmFAST directory was created, and copied into new directories as many times as needed (we nicknamed these copies 'clones'). The parameter file was also altered as required for each clone, and each clone directory was named accordingly (e.g. f10_R10_T0.8e+04). 21cmFAST version 1.2 was used, with all parameters kept at their 'out-of-box' values[9] (with the exception of $z_{max} = 15$, $\Delta z = 1$, and those to be varied in creating the parameter space). We also set $T_S \gg T_{\rm CMB}$ to be true in the parameter files. This is expected to be a safe bet for most of reionization, although the assumption is expected to be incorrect at the beginning of the EoR (the Cosmic Dawn), during which astrophysical parameters are uncertain (Furlanetto et al., 2006; Mesinger et al., 2011). For our purposes, assuming $T_S \gg T_{\rm CMB}$ results in a speed-up of ∼a few, and we remind the reader that our goal is to study sampling optimization, not the effects of different heating scenarios.

Using MPI paralellization, a script was created to launch these thousands of clones simultaneously. The code also deletes all program files and unnecessary files after each run is complete. See Appendix B, section B.3 for the raw code. The code was run on the OCCIGEN supercomputer, maintained at CINES (https://www.cines.fr/calcul/materiels/occigen/).

With this, one thousand iterations of 21cmFAST are run in parallel to create observables at each point of this parameter space. For a sampling of 1,000 points, the runtime is approximately ∼a few hours on OCCIGEN. The distance between observables is then calculated for every pair of points in the parameter space using the process described in

[9]Boxlength = 300 Mpc, Boxsize = 256 (low) and 768 (high), cores = 8, RAM = 16 Gb.

section 2.4.1. The 'out-of-box' version of 21cmFAST already creates and saves the power spectra, so we use this for the distance measure as opposed to the PDF (with power spectra created at intervals of $\Delta z = 1, z \in [6, 15]$, and recall that we are comparing $P(k, z)$, a 2D function). We now need to identify regions of the parameter space which have been over/under-sampled.

Defining Density

To do this we must construct a definition of 'density in the space of observables' based on the power spectrum distance to neighbouring observables.

Inverse Average Distance

Perhaps the most intuitive way to go from a clustering of points to a density at each point is simply to calculate the average distance to this point's neighbours, and then invert this value[10]:

$$\sqrt[3]{\rho_i} = \left(\frac{1}{6} \sum_{j=1}^{6} \mathrm{D}_{i,j} \right)^{-1} \tag{3.2}$$

We consider the neighbours of a point to be those immediately bordering this point on the grid — 6 points in our 3D parameter space (and the 6 resulting observables in the space of observables). Note that these are not necessarily the six nearest points, as defined by the distance measure $\mathrm{D}_{i,j}$. We can imagine a situation in which distances in our 3D space are much larger in one dimension than the other two. This is equivalent to having one parameter which, when varied, changes the form of the power spectra more than the others.

For the observables created using a point sampled on the surfaces, edges, or corners of the parameter space (those without a neighbour in all six directions) there is a slight complication that arises. It seems natural that we could simply average these boundary densities over fewer neighbours (5 for surfaces, 4 for edges, and 3 for corners). However, as mentioned, we are confronted with the fact that there is a large anisotropy in our parameter space — varying one parameter may have more effect on the morphology of the observables, and hence the power spectra, than varying another. For our choice of parameters, the distance between observables along the R_{mfp} direction was habitually between one and two orders of magnitude smaller than in the ζ and T_{vir} directions – a fact which produced a systematic lower density at boundary points. To remedy this, a more satisfying choice is mirroring the neighbours, such that all observables have six. Through this, we created the density cube shown in figure 3.2.

[10] We use the cube root of the the density to increase the contrast in upcoming figures, however we will still use the convention of referring to the quantity as simply the 'density'.

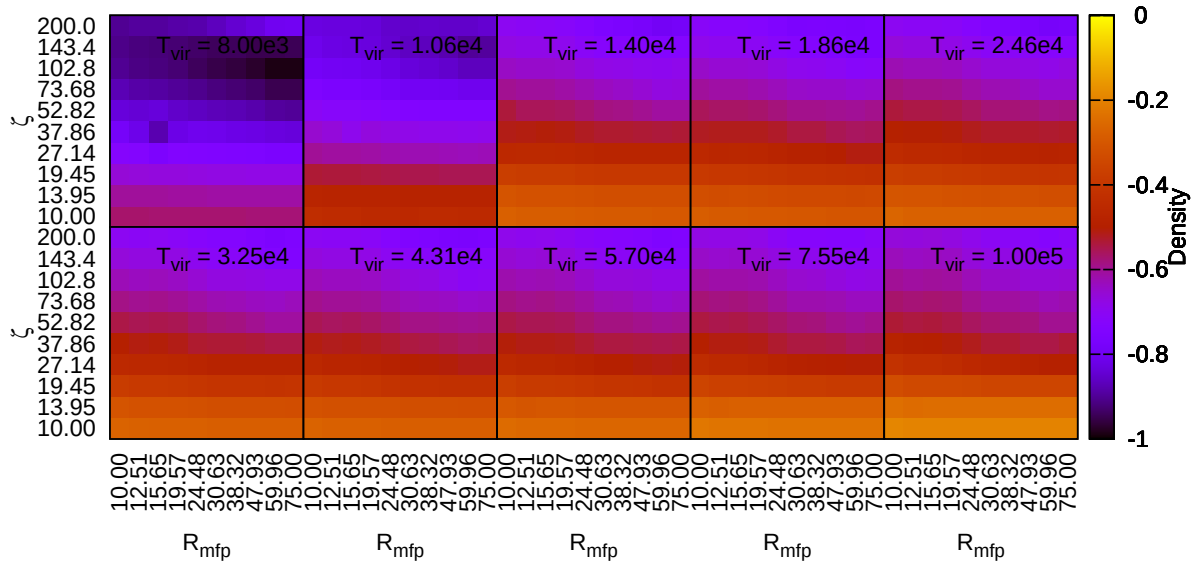


Figure 3.2 – *Densities of points sampled from this parameter space, as calculated using the simple density model of inverse average distance to neighbouring observables (equation 3.2). The observable is the power spectrum.*

We see that the density varies little as a function of $R_{\rm mfp}$, except for low values of $T_{\rm vir}$. Raising $T_{\rm vir}$ above 1.40×10^4 K seems to have nearly no effect on the observables. This seem logical, as the higher we raise $T_{\rm vir}$, the less star formation will occur, and hence the mean free path will become less important. Low values of ζ account for the region where observables are the most similar. Likely at these low ionization efficiencies, ionization is only starting at very late redshifts, and the observables are therefore all nearly identical.

Computing the Densities with Smoothed Particle Hydrodynamics

One issue with the simplistic approach of 'inverse average distance' is that it works best if the topology of this distance-space of observables is flat and isotropic. In actuality, unlike the parameter values, we cannot be sure that the resulting observables are necessary 'next to each other'. We know only the distance between them. As mentioned before, the 6 'neighbouring' observables (with respect to parameter values) used to calculate the average distance are not necessarily the closest in parameter space. A more satisfactory method of calculating distance is preferable.

Smooth Particle Hydrodynamics (SPH), as outlined in (Lucy, 1977; Gingold & Monaghan, 1977) and expanded upon in (Monaghan, 1992) was developed to deal with systems of particles forming an irregular sampling, rather than a grid. It therefore provides a helpful framework with which to approach our transformed parameter space.

For each observable, we first determine the n closest neighbours (where $n < N$, the total number of observables). Of these, we denote the distance to the n^{th} closest neighbour

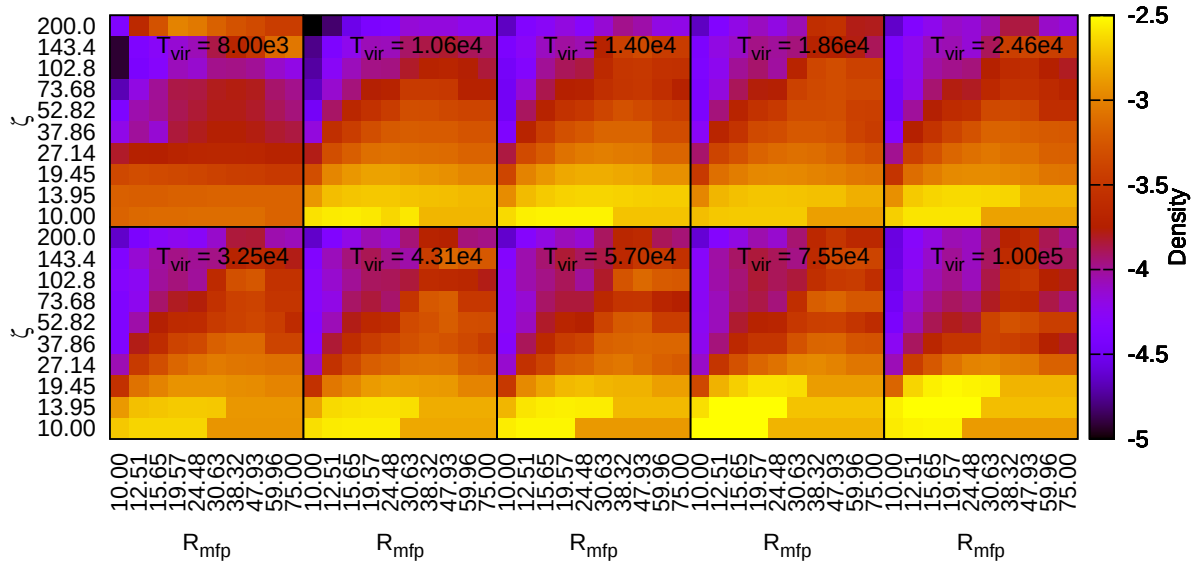


Figure 3.3 – *Densities of points within this parameter space, as calculated using SPH (and using the power spectra as observables).*

h_0, then define a weighting function[11]:

$$W(D, h_0) = \begin{cases} \frac{4-6R^2+3R^3}{4\beta} & \text{for R} < 1 \\ \frac{(2-R)^3}{4\beta} & \text{for } 1 \leq \text{R} < 2 \\ 0 & \text{for } 2 \leq \text{R} \end{cases} \tag{3.3}$$

in which

$$R = \frac{2D}{h_0} \qquad \beta = \pi \left(\frac{h_0}{2}\right)^3 \tag{3.4}$$

The density of the i^{th} point then becomes

$$\rho_{SPH,i} = \frac{1}{n} \sum_{j=1}^{n} W(\mathrm{D}_{i,j}, h_{0i}) \tag{3.5}$$

As before, $\mathrm{D}_{i,j}$ is the distance between the i and j^{th} observables, and $h_{0,i}$ is the distance to the n^{th} closest neighbour from the i^{th} observable. For our purposes, we choose $n = 10$, which gives the densities illustrated in figure 3.3.

Comparing with figure 3.2 we see that the SPH results in an interesting difference. The density cube is similar, and for both density definitions the density of the observable space grows with increasing T_{vir}, while ζ is inversely proportional. The interesting behaviour at $T_{\mathrm{vir}} = 8.00\mathrm{e}3$ is also seen for both density definitions. However, where the two differe is in the R_{mfp} dependency, which has inverted. Explicitly, in figure 3.2 increasing R_{mfp} results in decreased density, while for the SPH case an increase in R_{mfp} results in increasing

[11] W can also be thought of as applying 'Kernel Smoothing'.

density. This tells us that the SPH method is finding the closest observables to differ only in $R_{\rm mfp}$. In short, we have oversampled along the $R_{\rm mfp}$ axis, a parameter which we conclude does not have as strong an effect on the resulting observables as the other two parameters.

We do not want to simply correct the parameter space heuristically, approximating regions where the sampling should be more or less dense 'by eye'. Recall that we ultimately want an algorithm to mathematically resample the parameter space. In fact, when calculating the metric across the parameter space using a finite difference scheme (as will be discussed below) we discovered that the metric in some regions had negative eigenvalues (incompatible with the mathematical definition of distance). This tells us that we simply cannot proceed without first creating a finer sampling.

Therefore, we adjust the fiducial parameter values (table 3.2) by sampling along the ζ and $T_{\rm vir}$ axes more finely (20 values each), and the $R_{\rm mfp}$ axis more coarsely (6 values). The new parameter space sampling consists of 2,400 points, and is presented in table 3.3.

Table 3.3 – *Adjusted Sampling*

Parameter	Explored values
ζ	20.00, 22.58, 25.49, 28.77, 32.48, 36.66, 41.33, 46.71, 52.73, 59.53, 67.20, 75.85, 85.63, 96.66, 109.11, 123.17, 139.04, 156.95, 177.17, 200.00
$R_{\rm mfp}$	5.00, 7.38, 10.89, 16.07, 23.72, 35.00
$T_{\rm vir}/10^4$	0.80, 0.91, 1.04, 1.19, 1.36, 1.56, 1.78, 2.03, 2.32, 2.65, 3.02, 3.45, 3.94, 4.50, 5.14, 5.88 6.71, 7.67, 8.76, 10.00

An Algorithm for Optimal Sampling: The Eigenvector Method

Before using the full 2,400 point sampling, let us first explore the distances between points on a smaller $8\times8\times8$ sampling to study what is going on along the current axes (although we do use the 2,400 sampling in constructing the algorithm described below, 512 points was chosen as the standard number of points for comparing the geometries of the different spaces). Figure 3.4 presents a histogram of the distances between all neighbouring points in the parameter space. A visualization is also provided in figure 3.9.

Looking at the histogram, we are able to confirm our supposition that the space is neither isotropic, nor homogeneous. An isotropic space would have the histograms for each axis overlapping, which is not the case here. We see that the distance between neighbouring points along the $R_{\rm mfp}$ axis is noticeably smaller than for the other two

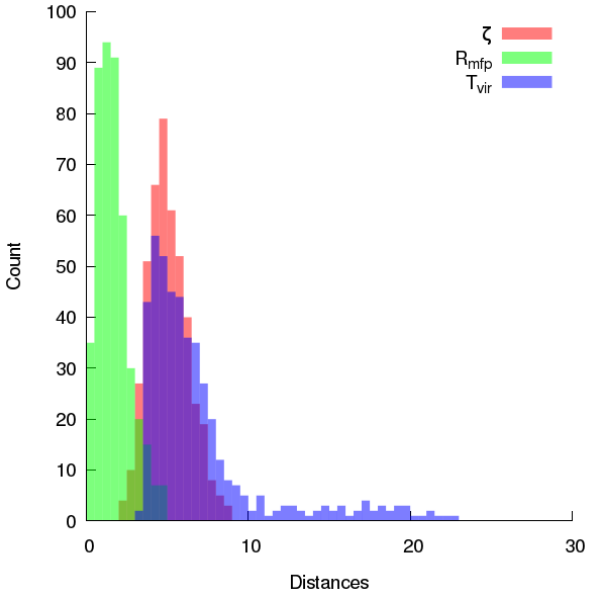


Figure 3.4 – *Distances between neighbouring observables along the three parameter axes for an* $8 \times 8 \times 8$ *logarithmic sampling. See figure 3.9 for a visualization of the sampling.*

directions. As for an equally spaced axis, this would correspond to a Dirac delta function (when represented as a distance histogram). We do not see this here: each histogram has a peak with some width, and the T_{vir} axis has an additional tail with some values scattered at higher distances, extending up to $\sim$23.

The ideal situation would therefore be, for our case, three overlapping Dirac delta functions. In practice this may be impossible, however we can still work towards making the peaks as narrow and close together as possible. Here, we can also see that simply adjusting how finely we sample the three axes will not be enough. This will not remedy the regions of high observable variability. The axes will likely have to be rotated, or (if allowable) the grid could be abandoned altogether.

The idea is then to study the geometry of the space more rigorously, using the eigenvectors (normalized) and eigenvalues at each point (obtained using the metric). These will give us an idea of the rate at which the observables change as we move through the parameter space. Using three orthogonal averaged eigenvectors and eigenvalues, we can reconstruct the grid in steps of equal observable distance (still using L_2 norm on the power spectra).

Working in a Logarithmic Space

It is important to remember that we will continue to use a logarithmic parameter space. More explicitly, the paramater space has been renormalized such that, along each axis, a logarithmic step (base 10) is treated as a difference of coordinate[12] of '1'. This means

[12]This is the length between points in the parameter space, and *not* the power spectrum distance between observables.

that, for example, in the metric definition below (specifically equations 3.8 to 3.13) when we refer to the distance to 'neighbouring' points along a given axis, these points are actually logarithmically spaced. The same holds true for the upcoming definition of volume (equation 3.16). When it comes time to calculate the new vectors along which we will resample, we will account for this decision through the use of constant factors to re-establish the physical value of the parameters with their 'coordinate' values (defined in equation 3.19 below).

Defining the Metric

To begin this, using the distances between all points we can compute the metric at each of these points. The metric of the 3D hypersurface of predictions is defined as:

$$g = \begin{bmatrix} g_{xx} & g_{xy} & g_{xz} \\ g_{xy} & g_{yy} & g_{yz} \\ g_{xz} & g_{yz} & g_{zz} \end{bmatrix} \tag{3.6}$$

such that two points of the hypersurface of prediction separated by an infinitesimal vector (dx, dy, dz) using the above coordinate system, will be at a distance:

$$\mathrm{dl}^2 = \begin{bmatrix} dx & dy & dz \end{bmatrix} \cdot \mathbf{g} \cdot \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} \tag{3.7}$$

In order to compute the metric at each point in our parameter space, we use a simple finite difference scheme. In our coordinate system, the vectors between neighbouring grid points can be written $(\Delta_x, \Delta_y, \Delta_z)$, where Δ_x, Δ_y, and $\Delta_z \in [-1, 0, 1]$. We are thus considering the 26 neighbouring points in 3D (the cube surrounding a point). For a given neighbouring point, the distance $D_{\Delta_x,\Delta_y,\Delta_z}$ is the corresponding distance according to the metric, the relation to the metric terms in the (x, y) plane is as follows:

$$D^2_{1,0,0} = g_{xx} \tag{3.8}$$
$$D^2_{-1,0,0} = g_{xx} \tag{3.9}$$
$$D^2_{1,1,0} = g_{xx} + 2g_{xy} + g_{yy} \tag{3.10}$$
$$D^2_{-1,-1,0} = g_{xx} + 2g_{xy} + g_{yy} \tag{3.11}$$
$$D^2_{1,-1,0} = g_{xx} - 2g_{xy} + g_{yy} \tag{3.12}$$
$$D^2_{-1,1,0} = g_{xx} - 2g_{xy} + g_{yy} \tag{3.13}$$

Should we include the (x, z) and (y, z) planes, 12 more equations can be added. This constitutes an overdetermined set of equations (which could be further expanded to include the corners of the cube) for which we could easily find the least mean squared

approximate solution. The symmetries of the equations suggest a more simple scheme. Using only these 12 equations, we solve for the metric terms to arrive at an approximate solution:

$$g_{xx} = \frac{D^2_{1,0,0} + D^2_{-1,0,0}}{2} \tag{3.14}$$

$$g_{xy} = \frac{D^2_{1,1,0} + D^2_{-1,-1,0} - D^2_{1,-1,0} - D^2_{-1,1,0}}{8} \tag{3.15}$$

The other coefficients have equivalent expressions. This scheme is in the spirit of using centred differences for computing derivatives. Because of this choice, we do not compute the metric at the points located on the faces of our parallelepiped rectangle domain. The accuracy of this estimation based on finite deference obviously relies on our fiducial sampling being dense enough so that the metric varies little between two neighbouring points.

Eigenvector Inversion

For each point, the corresponding metric is diagonalized to give the eigenvalues and eigenvectors. This is accomplished using the 3D diagonalization routine presented in Kopp (2008)[13]. Although fast an efficient, an interesting complication arises. For the i^{th} point, $\vec{v}_i$ and $-\vec{v}_i$ are both equally valid eigenvectors (with eigenvalue λ_i). The diagonalization routine will sometimes switch between signs at arbitrary positions in the parameter space. This, in and of itself, would generally not be an issue. However each point has three eigenvectors, none of which are numbered, and all of which are expected to change direction from one point to the next according to the topology. At regions where the observables change quickly, it can become difficult to discern if the vectors have changed direction on account of the metrics, or if one has accidentally been flipped.

Our goal is to build a set of local orthonormal basis vectors with the *same* orientation. Because of this, the problem must be resolved, as any flipped vectors will bias the average eigenvector we wish to compute. Therefore, a routine was developed to compare and match vectors, to decide if one or more had been flipped. This worked well, but was still seen to fail for regions in which observables changed rapidly. Take the following case, for example:

Point 1: $\vec{v}_1 = \begin{bmatrix} 0.98 & -0.06 & -0.19 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} -0.05 & 0.83 & -0.55 \end{bmatrix}, \vec{v}_3 = \begin{bmatrix} 0.19 & 0.55 & 0.81 \end{bmatrix}$

Point 2: $\vec{u}_1 = \begin{bmatrix} 0.69 & -0.23 & 0.69 \end{bmatrix}, \ \vec{u}_2 = \begin{bmatrix} 0.37 & 0.93 & -0.06 \end{bmatrix}, \ \vec{u}_3 = \begin{bmatrix} -0.63 & 0.30 & 0.72 \end{bmatrix}$

These are the three eigenvectors for two neighbouring points in a region of high observable variability. The one-to-one correspondence between vectors is not immediately

[13]Specifically the DSYEVJ3.c script, which diagonalizes using the Jacobi method.

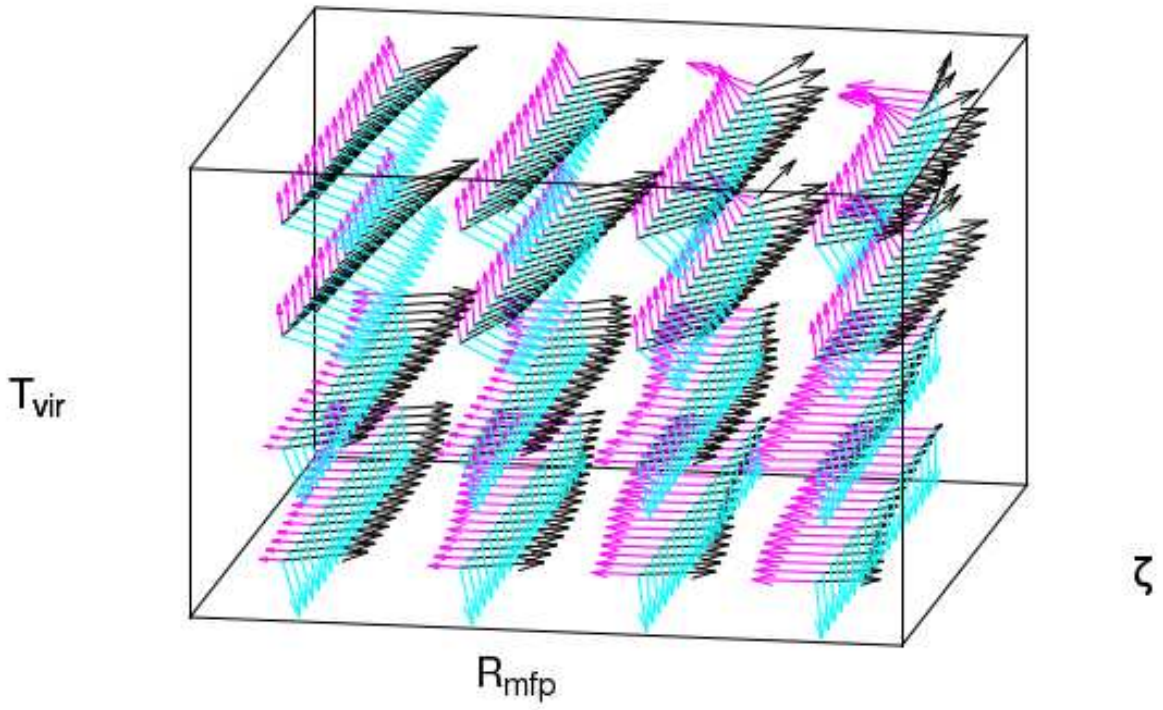


Figure 3.5 – *A region of the parameter space showing the eigenvectors for each point. This image illustrates the rotation of the eigenvectors from one point to the next. The image has been scaled to make this behaviour especially obvious along the ζ axis.*

obvious, nor is whether or not some have been inverted. In fact, given these two points and nothing else, it is impossible to know which eigenvectors correspond (without resorting to a much finer parameter space sampling between the points, at a computational expense). However, we can exploit a feature of the parameter space to resolve this issue.

Figure 3.5 shows a region of the parameter space in which the eigenvectors have been calculated and plotted at each point. This is a section without any inverted eigenvectors, and we see along the ζ axis the eigenvectors appear to rotate across the parameter space. In fact, this rotation exists along all three axes. Although the eigenvector rotation is not linear across the parameter space, it still allows us to estimate the direction of the three eigenvectors at any point by approximating their rotation in the region. To clarify this, let us say we have three points aligned along one of the parameter space axes: i, j, and k. Then, based on a linear extrapolation, we expect the k^{th} point to have eigenvectors $(\vec{v}_{1,k}, \vec{v}_{2,k}, \vec{v}_{3,k})$ given as:

$$\vec{v}_{1,k} = 2\vec{v}_{1,j} - \vec{v}_{1,i}, \vec{v}_{2,k} = 2\vec{v}_{2,j} - \vec{v}_{2,i}, \vec{v}_{3,k} = \vec{v}_{3,j} - \vec{v}_{3,i}$$

This remedies the issue of inverted eigenvectors, and also allows us to group the eigenvalues. A histogram of eigenvalues across the parameter space[14] is shown in figure

[14]In fact, we have not included the eigenvalues nor the eigenvectors that fall on the borders of the parameter space, as they tend to behave erratically on account of the metrics not being properly defined on the borders.

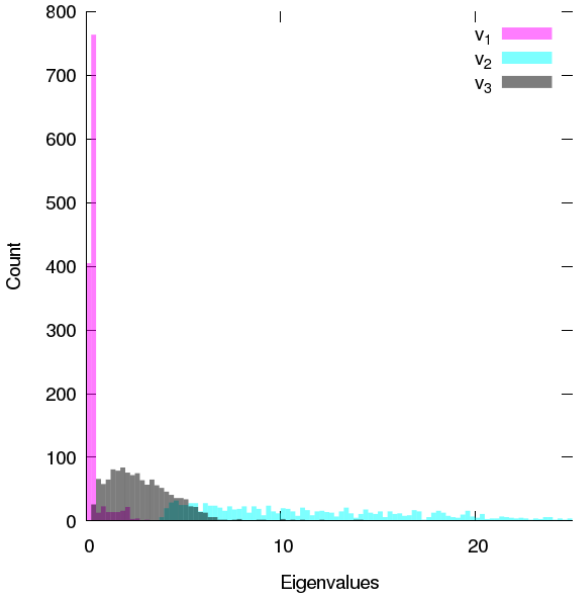


Figure 3.6 – *A histogram of eigenvalues across the parameter space, along with their corresponding eigenvectors. The points on the borders of the parameter space have been excluded.*

3.6. The noticeably different shape for each of the three eigenvalue groups in the figure is an indication of the inhomogeneity and anisotropy of the hypersurface of predictions (as expected from figure 3.4, and the different effects of the parameters on the observables).

It should also be noted that the eigenvectors are each a combination of the three parameter vectors (ζ, $R_{\rm mfp}$, and $T_{\rm vir}$), and there is no direct correspondence between the distances in figures 3.4 and the eigenvalues in 3.6. It is for this reason that different colours have been used.

Resampling

We can now begin to resample by first calculating the average eigenvectors (normalized[15]) and eigenvalues (across the 2,400 point logarithmic sampling). This is shown in figure 3.7.

We see that for all points in this parameter space, the corresponding eigenvectors tend to fall in clustered regions (pointed to by the averaged eigenvectors). However, the eigenvectors v_2 and v_3 (see legend in figure 3.6) have 'tails' wherein the values spread out. v_1 also has a small secondary group of eigenvectors. This is possibly due to $T_{\rm vir}$, for which we observed the eigenvectors to change quickly between 0.91×10^4 K and 1.04×10^4 K. However, understanding the true cause is non-trivial, as rotating eigenvectors is not necessarily an

[15]Be careful, depending on the method used for averaging, it may be necessary to re-normalize (even if the eigenvectors were already normalized). This is because simply averaging the coordinates of the eigenvector's heads does *not* generally result in a normalized eigenvector.

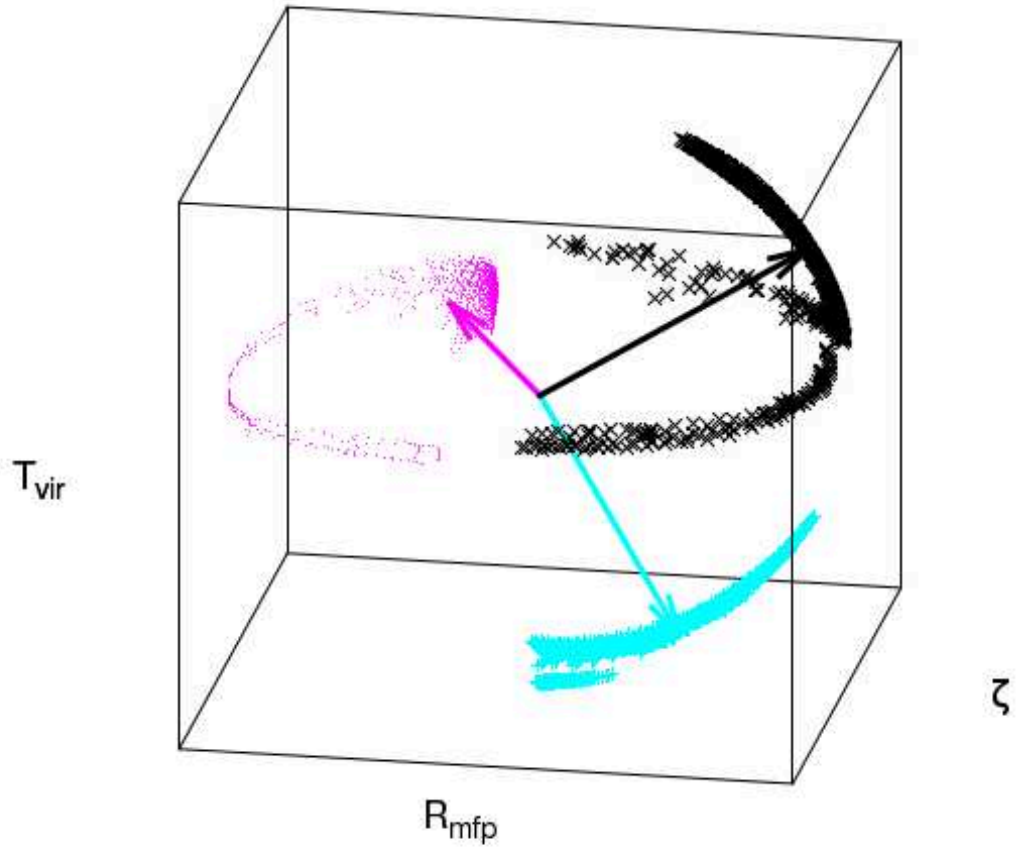


Figure 3.7 – *The three average eigenvectors across the parameter space, showing the clusters of eigenvectors, as well as the 2D projections. The colours are those of the corresponding eigenvalues in figure 3.6.*

indication of changing distances, but curvature of the hypersurface of predictions (section 3.1.1). A full treatment of these effects would require a foray into differential geometry, and is not attempted here.

With these average eigenvectors $\bar{v}_n$ and their corresponding average eigenvalues $\bar{\lambda}_n$ we effectively know the 'average' distance between simulations (eigenvalues) when travelling through this parameter space in three orthogonal directions (eigenvectors). We can therefore use this knowledge to re-sample the parameter space such that, in this new grid, the distances between neighbouring simulations will be closer to constant on average. It is to be expected that, on account of regions where the eigenvectors experience rotations away from the average, there will be some variation in the distances between neighbouring observables in said regions. However, we should still expect a much more isotropic and homogeneous parameter sampling than the logarithmic counterpart.

Starting from the central point of our parameter space, we can slowly expand outwards using three new 'step vectors', which are to be based on the average eigenvectors and their average eigenvalues. To begin, we recall that a point i in our parameter space can be assigned a 'volume' – that is to say, a region of the parameter space closer to this point than to any other (again based on the L2 norm distance):

$$V_i = \sqrt{\det_i} = \sqrt{\lambda_1^i \lambda_2^i \lambda_3^i} \tag{3.16}$$

where λ_n^i is the n^{th} eigenvalue for point i. Assuming N points in our parameter space, we can calculate the average volume to be $\bar{V} = \frac{1}{N}\sum_i V_i$, and therefore the average distance between points will be[16]:

$$\bar{d} = \sqrt[3]{\bar{V}} = \sqrt[3]{\frac{1}{N}\sum_i \sqrt{\lambda_1^i \lambda_2^i \lambda_3^i}} \tag{3.17}$$

We can now use the normalized average eigenvectors $\bar{v}_n = (\bar{x}_n, \bar{y}_n, \bar{z}_n)$ (linear combinations of the three initial parameter axes) and their corresponding eigenvalues $\bar{\lambda}_n$. Starting in the centre, we wish to move along each eigenvector by some distance such that the L2 norm distances of the resulting observables are equal to the average distance. We know that the amount we should move along each eigenvector should depend on that eigenvector's eigenvalue, so let us define a normalization constant that depends on the relevant eigenvalue: $\alpha_n(\bar{\lambda}_n)$. Now, let us take the first average eigenvector $\bar{v}_1$. If we move from the central point to a new point along the vector $\alpha\bar{v}_1$ the distance between the two corresponding observables will be (from equation 3.7):

$$d^2 = \alpha_1 \bar{v}_1 \cdot g \cdot \alpha_1 \bar{v}_1 = \bar{\lambda}_1 \alpha_1^2 \left(\bar{x}_1^2 + \bar{y}_1^2 + \bar{z}_1^2\right) \tag{3.18}$$

To assure that $d = \bar{d}$ we can now set $\alpha_n(\bar{\lambda}_n) = \frac{\bar{d}}{\sqrt{\bar{\lambda}_n}}$. The final step is to account for the fact that the initial sampling was linear in *logarithmic* space, and therefore we must define constants of logarithmic step: $c_1 = \Delta \log \zeta, c_2 = \Delta \log R_{\rm mfp}, c_3 = \Delta \log T_{\rm vir}$ (the difference between the logarithms of neighbouring parameter values). Explicitly, for any parameter

[16]To be clear, this is again a simplification. The true average distance depends on the geometry of the sampling (Cartesian grid, crystal lattice, etc.).

θ, in any number of dimensions, the corresponding constant term can be defined as[17]:

$$c_\theta = \log_{10}(\theta_{i-1}) - \log_{10}(\theta_i) \tag{3.19}$$

where θ_i and θ_{i-1} are any two neighbouring values of the θ parameter ($\theta_i > \theta_{i-1}$). Finally, we can define our new step vectors $\vec{s}_n$:

$$\vec{s}_n = \left(\frac{\bar{d}}{\sqrt{\bar{\lambda}_n}} c_1 \bar{x}_n, \frac{\bar{d}}{\sqrt{\bar{\lambda}_n}} c_2 \bar{y}_n, \frac{\bar{d}}{\sqrt{\bar{\lambda}_n}} c_3 \bar{z}_n \right) \tag{3.20}$$

This formula will give us three step vectors, and starting from the central point we can move outwards along said vectors, until we find ourselves outside the initial bounds defined in section 1.3.2, to re-sample our space. Should we require a specific number of points in our new sampling, we can adjust the $\bar{d}$ value until we have the desired number (a smaller $\bar{d}$ will result in more points, and vice versa).

After applying this method, as expected, the three peaks overlap (figure 3.8). This means that we have arrived at a nearly isotropic and homogeneous sampling of our parameter space. In figure 3.4, we saw that along the $T_{\rm vir}$ axis there were aberrant distance bins extending beyond ∼20. With the new sampling, the outlier distances extend to ∼15, and are distributed along the 3 step vector directions (which we expected, as each step vector is a combination of the initial three axes). These higher distance bins correspond to the previously mentioned regions in which eigenvalues change rapidly. Figure 3.9 shows a comparison of the initial and final parameter space sampling.

We can see that the grid has effectively been rotated, and the space between sampled points have been rescaled for each axis. Yet the grid remains, allowing us to have a clear definition of neighbouring points. For reference, the final step vectors for the 3D parameter space presented here (within the bounds defined in section 3.3.1) were found to be:

$$\vec{s}_1 = \begin{bmatrix} 0.159451 & -0.101265 & 0.135481 \end{bmatrix}$$

$$\vec{s}_2 = \begin{bmatrix} 0.017299 & 0.057789 & -0.023709 \end{bmatrix}$$

$$\vec{s}_3 = \begin{bmatrix} -0.017007 & 0.203757 & 0.042429 \end{bmatrix}$$

when sampling outwards from the point $\begin{bmatrix} 1.827369 & 1.206016 & 4.480007 \end{bmatrix}$ (the format here is $\begin{bmatrix} \log_{10}(\zeta) & \log_{10}(R_{\rm mfp}) & \log_{10}(T_{\rm vir}) \end{bmatrix}$). This is for generating a sampling of 512 points, however the step vectors could be scaled (with the same scaling factor applied to all of them) to increase or decrease the number of sampled points, as needed.

[17] In principle, the log base could be any number, however the corresponding fiducial parameter space sampling (table 3.3) should also have been constructed with the same base.

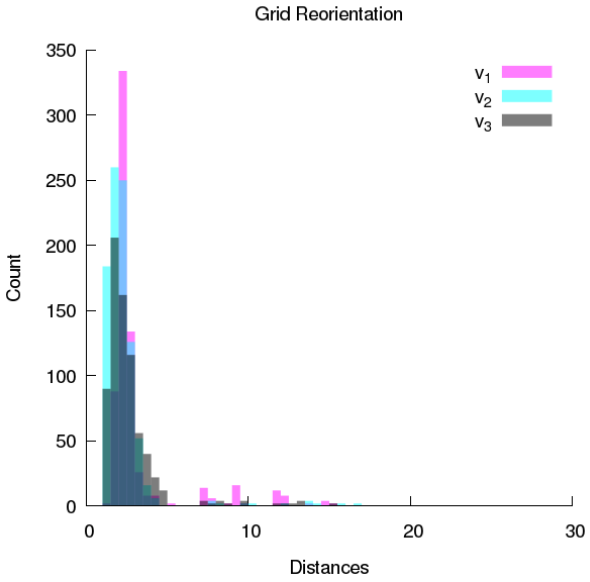


Figure 3.8 – *Histograms of the distances between neighbouring observables after having resampled using the Eigenvector Method. A visualisation of the resampling can be found in figure 3.9.*

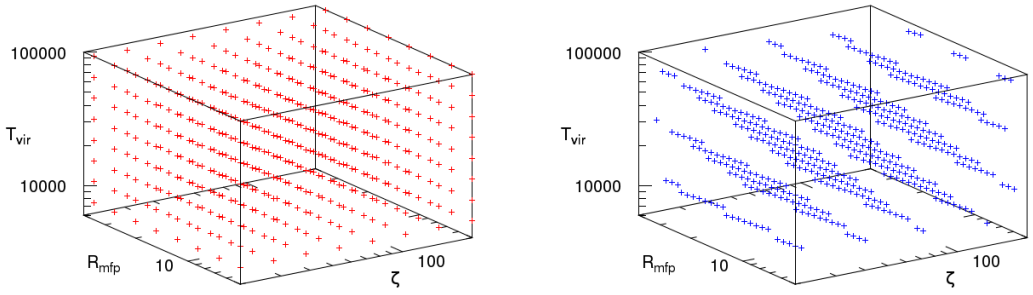


Figure 3.9 – *Visualizations of the initial logarithmic parameter space sampling, and the new sampling after applying the eigenvalue method. Both samplings contain 512 points.*

Note: Comparison with PCA

The method presented here shares some similarities with the PCA technique for parameter reconstruction (section 0.6.6). Both methods, for example, involve creating a new orthonormal set of vectors, and then constructing a grid to reinterpret data. However, the Fisher Matrix used by PCA is based on the correlation of different parameter pairings, and not on the distances between observables. PCA is also often used to reduce the dimension of a data set by eliminating vectors along which the data is weakly correlated, which is not that case here. Finally, although PCA re-grids a data set, new points are not sampled along this improved grid (as is a step in our method). However, with this in mind, it should be noted that there is a strong relationship between the Fisher Matrix and the metric used here, which we will return to in section 3.7.2.

Another Algorithm for Optimal Sampling: The Adaptive Grid-Free Method

As we saw in figure 3.8, the eigenvector method for resampling the parameter space is clearly an improvement over the initial sampling. In particular, the similarity of observables along the $R_{\rm mfp}$ axis (in figure 3.4) has been remedied by reorienting the grid, such that no axis corresponds to a change in only $R_{\rm mfp}$. As well, regions in which observables change rapidly (primarily the $T_{\rm vir}$ outlier bins at distance $\gtrsim 10$ in figure 3.4) are nearly all taken care of after resampling.

Yet there is still room for improvement. All three (new) axes still have some width, and have some very slight tails at higher distances (around distance ~ 4). There are a couple of other issues as well. A few distances, along all three axes, are still abnormally high (distance ~ 5 to ~ 15). Looking closely, the three peaks are also not perfectly overlapping, with most distances along the v_1 axis still slightly higher than along the other two.

The question then becomes how to remedy these effect. If we insist on conserving a grid arrangement for sampled points, it is unlikely we will be able to do much better. This is because, although rotating and stretching the grid can certainly reduce the anisotropy, there may simply be patchy sections of the parameter space within which observables vary differently than the rest.

However, if we abandon the requisite of a grid altogether, then there are other ways to further reduce the anisotropy. To attempt this, we present a new algorithm, which adapts the sampling without the grid constraint (hence the name: Adaptive Grid-Free Method).

Algorithm Overview

Here we provide a brief quantitative description of the algorithm.

Preliminaries

We begin with the sampling in table 3.3, and the metric at each point is computed through the same equations defined in section 3.4.1. The total volume[18] of the parameter space is computed using the metric. Assuming maximum n-sphere packing (section 3.2.1), the typical volume[18] for each n-sphere is calculated. We initiate a maximum interaction distance equal to twice the radius of the n-spheres (D_{max}). Along the extremities of the parameter space, we also designate a 'buffer zone' of set width. A new set of points is randomly scattered (without grid constraints) in this space (points can also fall into the buffer zone).

Iterating

1. Metrics at each of these randomly scattered points are interpolated using the values of the nearest metrics (those that were computed on the fiducial grid sampling). Interpolation is carried out with an SPH scheme (section 3.3.3).

2. Take a pair of points i and j.

   (a) Average the metric between these two points.

   (b) Use this averaged metric to approximate the distance between the observables at these two points ($\mathrm{D}_{i,j}$).

   (c) Repeat this for the next set of points, until the distance between all pairs of points is known.

3. For a point i and its neighbour j, if $\mathrm{D}_{i,j} < \mathrm{D}_{max}$ then we define:

$$\vec{d}_{i,j} = \frac{1}{2}(\mathrm{D}_{max} - \mathrm{D}_{i,j}) \cdot -\vec{r}_{i,j} \tag{3.21}$$

   where $\vec{r}_{i,j}$ is a unitary vector pointing from i to j (and is negative as nearby points should be moved apart). This is repeated for all pairs of points.

4. For each point i, the contribution of all n nearby points is summed, and then the point is moved along this new vector $\vec{d}_i = \sum_{j=1}^{n} \vec{d}_{i,j}$.

5. Points in the buffer zone receive an additional displacement towards the centre (representing a sort of confinement).

6. A new interaction distance D_{max} is evaluated using the current number of points in the parameter space region. The above steps are then repeated.

[18] Or hypervolume.

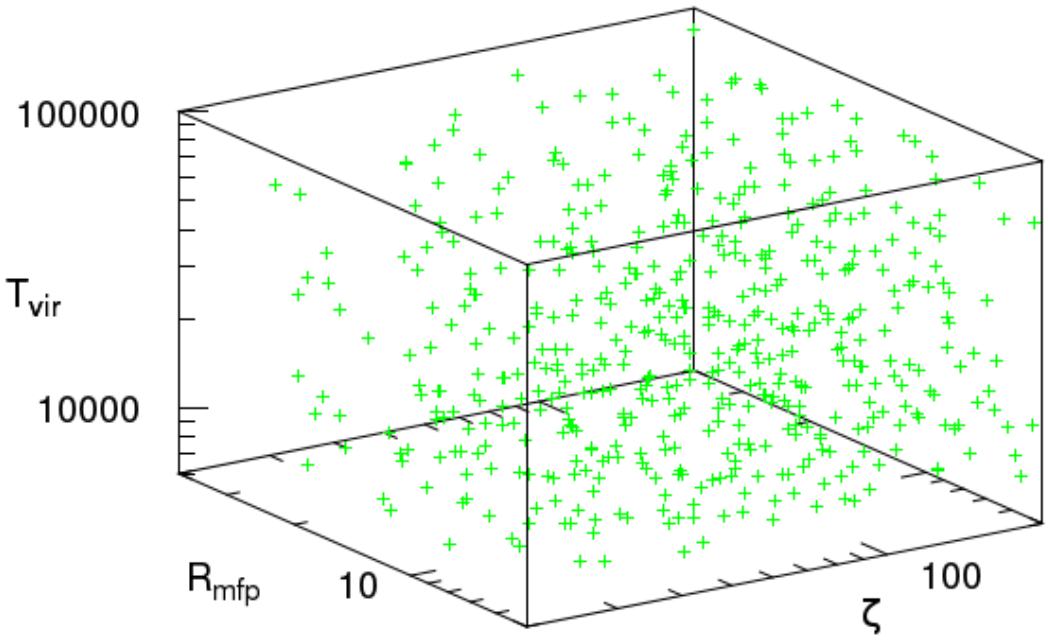


Figure 3.10 – *An example of an Adaptive Grid-free sampling (506 points).*

Final Sampling

The positions of the points after each iteration are recorded. At the end of the loop, the best configuration can be selected from the histogram of the distances (computed using the interpolated metrics between pairs of points) from each point to the N_{kiss} points, where N_{kiss} is the 'kissing' number for a given dimension[19].

Figure 3.10 shows an example of the best sampling after a number of iterations. Notice that the points are more spread out at high $R_{\rm mfp}$ and low ζ. This matches what we found in section 3.3.3 — that this is a region of low observable variability. It's also important to note that, although the sampling was initialized with 512 points, there are *not* 512 points in the final sampling (but rather 506). In our current method, the Adaptive Grid-free algorithm generally will not preserve the number of points (although this depends on initial randomly scattered points). This could possibly be improved in future.

[19]The kissing number is defined as follows. For a given n-sphere, how many n-spheres of the same radius can be made to touch the first one without overlaps. For 2D $N_{kiss} = 6$, for 3D $N_{kiss} = 12$, etc. The kissing number is not known exactly for dimension > 4, except for 8D and 24D, however bounds have been set at $\sim 20\%$ up to 24D (Mittelmann & Vallentin, 2009).

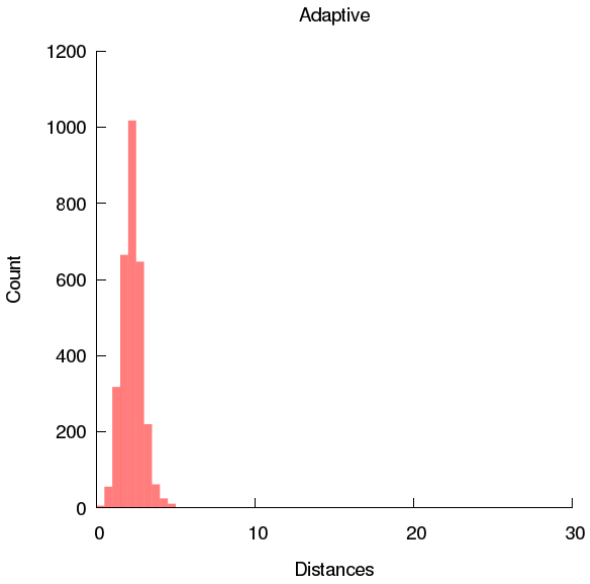


Figure 3.11 – *Histograms of the distances between neighbouring observables after having resampled using the Adaptive Grid-free method. A visualisation of the resampling can be found in figure fig:Adaptive.*

Figure 3.11 shows the result of this algorithm in terms of the distances between observables. The most obvious advantage is that there are no outlier distances at higher bins. In addition, the fact that there are no clearly defined axes means that there is no longer any issue of the distances along each axis being different (effectively we don't have to worry about having multiple peaks). There is still some spread in distances (between distance ~ 1 and ~ 4), which is likely due to border effects, where points (and the resulting observables) may behave differently depending on the dynamics used within the buffer zone.

Algorithm Comparison

To summarize: both algorithms have been showed to improve the homogeneity and isotropy of the parameter space sampling. If we do not require a grid, nor a specific number of points, then the adaptive method is slightly better. The distances cluster slightly tighter, and there are no outlier observables far removed from the rest. Although if a grid system, or fixed number of points, are required, then the eigenvector method is still a vast improvement over the initial sampling.

Implications for Neural Networks

With these new samplings prepared, it is tempting to see if our efforts have indeed improved upon the fiducial sampling. Recall that one of our goals was to provide better

training data for neural networks. We are now in a position to test if we have accomplished this.

Note: Although based on my work, this section presents research largely undertaken by Aristide Doussot and Benoît Semelin.

Network Structure

We will be considering a neural network that takes an observable as input (in our case, the power spectrum generated by 21cmFAST, discretized on 12 wavenumber bins and 10 redshift bins), and the values of the model parameters as output (ζ, $R_{\rm mfp}$,$T_{\rm vir}$). Such a network needs to be trained before it can be used on the observed data. The training set consist of a number of different inputs ($P(k,z)$ in our case), and the associated desired outputs (the corresponding values of ζ, $R_{\rm mfp}$ and $T_{\rm vir}$). The hypothesis to be tested is whether training on an optimal sampling of the parameter space will improve the accuracy of the predictions of the resulting network.

For this test we used the Keras framework[20] to implement a full connected neural network with a single hidden layer. The input-layer contains 120 nodes (12 wavenumber bins $\times$ 10 redshift bins), the hidden layer contains 80 neurons, and the output-layer contains three neurons (one for each parameter).

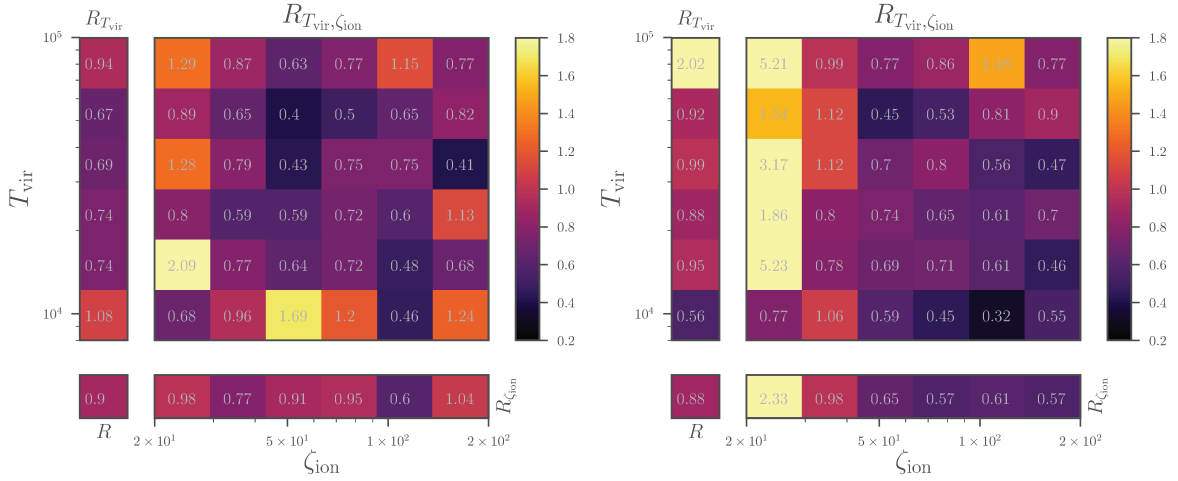
We trained this network with three different training sets, each scattered in the same region of the parameter space defined in section 3.3.1. The first set consisted of a simple 3D grid with 512 points logarithmically spaced (equivalent to the fiducial sampling, however more sparsely sampled). The second, also with 512 points, was produced with the Eigenvalue Method algorithm described above. The third was produced using the Adaptive Grid-free, and instead contains 506 points (explained in section 3.5.1). The number of points in these sets was chosen because given too many points as training data the network performs well regardless of the sampling technique.

Quantifying Performance

The accuracy of the network during, and after, training is evaluated using a different test set consisting of 512 points of the parameter space. The accuracy of the prediction for sample j in the training set is estimated using a quantity known as the 'loss function', and defined as:

$$C_j = \frac{1}{n}\sum_{i=1,n}\left[\log_{10}\left(y_{i,j}^{\rm pred}\right) - \log_{10}\left(y_{i,j}^{\rm true}\right)\right]^2 \tag{3.22}$$

[20] https://keras.io/



(a) *(b)*

Figure 3.12 – *The cost function representing the improvement for parameter reconstruction (in recovering $T_{\rm vir}$ and ζ values) for neural networks trained on the initial logarithmic gridded versus (a) a resampling carried out using the Eigenvector Method, or versus (b) an Adaptive Grid-free resampling. Partial cost functions defined equivalently for a single parameter (thus averaged over all possible values of the other two) are also plotted. The value in the bottom left is the average error difference. Image credit: Aristide Doussot.*

where n is the number of output parameters, $y_{i,j}^{\rm pred}$ is the prediction of the network for parameter i of the sample j, and $y_{i,j}^{\rm true}$ is the true value of parameter i (the value used by the model to predict the observable for sample j). A total cost function $C_{\rm tot}$ can then be defined for a set of samples by averaging the individual cost functions. Training the network is then reduced to the process of minimizing $C_{\rm tot}$ by adjusting the neural network weights. This is typically achieved using various forms of gradient descents.

Using a global cost function defined on the entire test sample was found to discriminate between the three choices of training samples insufficiently. It is expected that it may have been sufficient had the network trained on the Adaptive Grid-free sampling performed globally better compared to that trained on the logarithmic grid sampling, but this was not the case.

To remedy this we define $C_{T_{\rm vir},\zeta}$, a partial cost function defined as the average of the individual cost functions C_j for all samples that fall within some intervals centred on $T_{\rm vir}$ and ζ. Figure 3.12 (a) shows a map of the ratio of the partial cost function for the network trained on the initial logarithmic sampling versus one trained on the Eigenvalue Method resampling, while Figure 3.12 (b) shows the analogous map for the Adaptive Grid-free sampling. Note that, as the power spectrum is less sensitive to $R_{\rm mfp}$, we have focused on the other two parameters. We also plot single parameter cost functions (the leftmost column and bottom row), as well as the overall cost (in the bottom left). To help one read the figure, the bottom left value of (a) is 0.68, which indicates that — in this region of the parameter space — a neural network is 38% more accurate in parameter estimation when trained on the resampled sampling verses one trained on the initial logarithmic sampling.

Comparing Performance

We can see that, for the sampling created using the Eigenvector Method, the cost function is lower across most of the parameter space. Only for some regions on the borders is it higher. This represents a failure of the neural network to properly predict parameter values for these regions. Although the overall performance is only 10% better than that of the neural network trained on the initial logarithmic sampling, when we exclude the border regions the improvement is closer to $\sim$40%.

As for the Adaptive Grid-free method, the cost function is lower over most of the parameter space, with the exception of the lowest ζ values. There could be several possibilities to explain the low performance for these values. It has been verified that the Adaptive Grid-free method generates a CCP-like lattice sampling (section 3.2.1) of the parameter space when applied to a Euclidian geometry for the hypersurface of predictions. As the actual geometry is non-Euclidian, gaps may appear in the lattice in regions where, for a given sampling density, the curvature properties of the hypersurface do not allow for the same number of neighbouring points. Such a gap is caused by generating a *discrete* sampling, and should one appears in a region with a low number of sampled points, it could be responsible for large errors in the performance of the trained network. Another more simple explanation is that, despite the buffer zone included in the Adaptive Grid-free method (section 3.5.1), boundary effects are still not well under control. In any case, the improvement over the initial sampling is still $\sim$12%, marginally better than over the Eigenvector Method resampling (and again better if we exclude the low ζ column).

Future Prospects

We have presented here two algorithms for optimally sampling a parameter space. In addition, these have been tested on neural network training to show their effectiveness. This represents a solid first step into exploring the feasibility of an 'optimal sampling', yet as the concept has been relatively unexplored, there are plenty of opportunities to build upon this research.

Higher Dimensional Parameter Spaces

None of the steps presented thus far have been dimension dependant. That is to say, it would be worthwhile testing them on higher dimensions, such that they can be readily applied to higher dimensional parameter spaces. A logical first step would perhaps be attempting optimal sampling on the six dimensional parameter space tested in Greig & Mesinger (2017b), as this is a relatively simple extension of the sampling this work is based on (Greig & Mesinger, 2017b). More ambitious could be attempting to optimally sample the 11 dimensional space of Kern et al. (2017).

In terms of how to approach this, first recall that the general scaling of the size of a grid sampling is exponential in the dimension of parameter space. Therefore, the fiducial space could be constructed with the Latin Hypercube method, as well as using a circumscribed n-sphere (section 3.2.1). However, it remains to be seen how well these algorithms could be applied upon a LHS initial sampling. Another issue is that in general a Latin Hypercube sampling, unless carefully arranged, would not maintain the concept of neighbouring points. This would prove an issue for the eigenvector method.

Evaluating the Metric in High Dimension

There is a way, however, to account for this. We could redefine the metric such that it minimizes the error when computing the distance to a set of neighbouring points. To begin, in 3D we could consider the 25 neighbouring points (in section 3.4.1 we only considered the 20 points at a distance, in indices, of $< \sqrt{2}$). Then, seeing as all the distances are known, we could start with our current metrics, and define an error function. For example, the error on the metric for the i^{th} point would be:

$$\text{error} = \sum_{n=1}^{25} \sqrt{(\mathrm{D}_{i,n}^{\mathrm{PS}} - \mathrm{D}_{i,n}^{\mathrm{metric}})^2} \tag{3.23}$$

where $\mathrm{D}_{i,n}^{\mathrm{PS}}$ is the standard Power Spectrum distance we have been using up until now between point i and its n^{th} neighbour, and $\mathrm{D}_{i,n}^{\mathrm{metric}}$ is the distance between the same points given when applying the metric. The goal would then to be to reduce the error, and this should assure the best estimation of the geometry of the hypersurface of predictions. This would result in a set of 25 equations with 6 independent metric coefficients as variables. It is simple linear algebra to find the set of coefficients that minimize the error (e.g. via the least square method). In fact, this has been implemented, but the result has not yet been compared to the simple 'finite distance' method.

Note that Matrix Diagonalization is not computationally expensive for any dimension we would realistically be exploring (Nebot-Gil, 2015).

The Fisher Information Metric

Currently, for Bayesian inference a flat prior (in which all parameter values, or the logarithm of all parameter values, are equally probable) is often used. One improvement would a prior which also takes noise into consideration, such as Jeffreys' prior (Jeffreys, 1946). Calculating Jeffreys' prior is dependent on computing the Fisher Information Metric, whose terms are defined as:

$$I_{\theta_i,\theta_j} = E\left[\frac{\partial \log f}{\partial \theta_i}\frac{\partial \log f}{\partial \theta_j}\right] \tag{3.24}$$

where θ_i and θ_j are parameters of our model, E is the expected value (for a variable X on a probability space (Ω, Σ, P) $E[X] = \int_\Omega X(\omega) dP(\omega)$), and f is the distribution of possible power spectra, in turn defined as:

$$f(P_N, \vec{\theta}) \approx \prod_{k_i} \exp\left[-\frac{(P_N(k_i, z) - P(k_i, z, \vec{\theta}))^2}{2(\sigma(k_i, z))^2} \right] \tag{3.25}$$

where the parameters of the model are $\vec{\theta}$, P_N is the power spectrum with noise (binned at wavenumbers k_i and redshifts z), and $(\sigma(k_i, z))^2$ is the variance of the combined sources of noise for the i_{th} wavenumber bin at redshift z. Inserting this definition into equation 3.24 we arrive at:

$$I_{\theta_i, \theta_j} = \sum_{k_i, z} \frac{1}{(\sigma(k_i, z))^2} \frac{\partial P(k_i, z, \vec{\theta})}{\partial \theta_i} \frac{\partial P(k_i, z, \vec{\theta})}{\partial \theta_j} \tag{3.26}$$

In which the noise term $(1/\sigma^2)$ is applied to each power spectra bin (in wavenumber and redshift). If we ignore this noise variance term:

$$I_{\theta_i, \theta_j} = \sum_{k_i, z} \frac{\partial P(k_i, z, \vec{\theta})}{\partial \theta_i} \frac{\partial P(k_i, z, \vec{\theta})}{\partial \theta_j} \tag{3.27}$$

which is equivalent to equation 3.4.1 (in our case $\vec{\theta} = \zeta, R_{\rm mfp}, T_{\rm vir}$). We chose not to use the noise term (and therefore, not the true Fisher Information Metric) primarily because our current neural network configuration does not handle noisy power spectra very well. We therefore consider this theoretical framework as a first (idealized) step towards optimal parameter space sampling. A planned next step is to verify that our noise model is correct, implement it to arrive at the full Fisher Metric, and then use Jeffreys' prior towards Bayesian inference.

Combination with 21cmMCMC

We have taken a tiny first step into exploring what optimal parameter sampling means for neural networks, it would also be worthwhile using one of the generated resamplings as prior for MCMC parameter reconstruction. This would be especially easy in the case of 21CMMC, as we have used exactly the same code and parameters as in (Greig & Mesinger, 2017b). Therefore, testing the effectiveness would simply be a matter of re-running the code with the optimized sampling as a prior (possibly Jeffreys' prior).

Alternate Distance Definitions

One of the more important avenues to be explored — especially pertinent to this work — is the effect of distance definition on the resulting parameter space. As an example of this, consider the following. The power spectra were found to be only weakly sensitive to the $R_{\rm mfp}$ parameter. This largely defined the resulting optimal sampling, as the eigenvectors

were ultimately arranged to accommodate the sensitivity of the power spectra to the three axes. Should we have defined distance differently — say, in terms of the L_2 norm of the pixel distribution functions — there is no reason to believe the $R_{\rm mfp}$ parameter would again be weakly correlated. Recall that in section 2.4 we saw how, for the observables of 21SSD, the power spectra and the PDF were found to exhibit sensitivities towards different parameters.

Therefore, throughout this chapter, when we have referred to the sought after 'optimal sampling' it bears remembering that this is in fact the 'optimal sampling with respect to the power spectra'. We have no guarantee that another diagnostic wouldn't result in an entirely different optimal sampling (in fact, we can almost be sure that it would).

Optimally Sampling 21SSD

Another stated goal of developing this methodology was to make an informed decision on how to better sample a sparsely sampled high-resolution database. Now, we are prepared to do just that. In the case of 21SSD, currently sampled at only 45 points (table 1.3), we now have a framework for seeking out worthwhile new points at which to create observables. As we already have the power spectra and PDF observables for each point, and have already explored using both as distance measures, we have a good foundation for intelligently expanding the database.

One may note that, for the fiducial sampling used with 21cmFAST, we began with 1,000 points (table 3.2), and later increased to 2,400. So it may seem unrealistic to begin with a mere 45 points. It is true that applying the resampling algorithms to 45 points is unlikely to generate an optimal sampling. Regardless, we can be confident that the new sampling will be a strong choice to assure all resulting observables will be different, likely better than resorting to LHS or other standard sampling methods. One final point worth mentioning is that we may not wish to recreate the 45 observables already created (to save computation time). Therefore, it would be possible to fix these 45 points, and apply the Adaptive method to any new points being added.

Conclusion

CHAPTER 4

Conclusion and Perspectives

In this work we have presented efforts towards adding to the foundation of Epoch of Reionization studies. This includes framework relating directly to current observations, as well as aimed at preparing for upcoming experiments. In both cases, the setting is theoretical, and built upon the use of simulations. We will now briefly summarize what has been accomplished, and imagine the first steps towards taking these developments further, with the ultimate goal of preparing for increasingly ambitious EoR experiments.

In the first part, a database of simulated 21 cm tomographies was created. Much of what has been presented therein served to gain familiarity with the LICORICE code, and better understand the nature of simulated EoR lightcones. This included a quick look into an effect known as 'self-shielding', in which dense regions of neutral hydrogen were shown to scatter less Lyα photons. We then discussed different parametrizations, their merits and shortcoming, and our parameter choices for the 21SSD database ($f_X, f_\alpha, r_{H/S}$). Another addition to the database was realistic thermal noise modelling, added to the lightcones for various resolutions, and created by modelling in UV visibility space. SKA resolution lightcones, created with full dynamics, as well as realistic noise, provide some of the best prototypes for what we should realistically expect to see in the coming years. We arrived at a reasonably optimistic outlook for thermal noise, with structure being recoverable down to $z \approx 9 - 12$, depending on the resolution and reionization scenario.

In the second part, preliminary steps were taken towards making use of this database. The primary objective, planned for a less sparse future incarnation of 21SSD, is parameter extraction. With this goal in mind, we first examined different ways to characterise the different reionization scenarios, through simplifying and quantifying the corresponding lightcones. We examined the power spectra, and a less explored diagnostic: the Pixel Distribution Function. We found the PDF to be a strong tool for this purpose, with some unique attributes not present in the power spectra (principally the ability to encompass non-gaussianity). To build upon this, we examined how these two diagnostics handle the concept of 'distance' between different observables. We found that the PS and PDF are both quite different in this respect: the power spectra seem more sensitive to changes in f_X, while the PDF responded more strongly to changes in f_α. This leads us to believe both

could — and should — be used to maximize the information available when attempting parameter extraction.

In the third part, we approached parameter extraction from another angle: asking if there is an 'optimal' manner in which to sample a parameter space when creating a database. We defined the relevant terminology, and then examined two different methods for improving a parameter space sampling. The first method constructs metrics at all sampled points (based on the distances to neighbouring points), and then reorients the grid based on the corresponding eigenvectors. The second method interpolates the metric at any point on the hypersurface, and iteratively moves points towards regions identified as overdense by their metric properties. This second method also does not constrain points to a grid (hence the name 'adaptive grid-free' method). We showed that both of these methods succeeded in reducing the anisotropy and inhomogeneity of the parameter space sampling. As expected, the adaptive grid-free method works slightly better, as points were less constrained within the space. The grid-free method was therefore used to train neural networks, and we find that this method works modestly better in the centre of the parameter space, and on the low $T_{\rm vir}$ edge, although actually performs worse on the other extremities of the parameter space. This is possibly due to the adaptive grid-free algorithm encountering difficulties determining how to displace sampled points on the boundaries.

This represents a first foray into the notion of exploring the geometry of what we have referred to as the 'space of observables'. Although the advantages we have demonstrated for neural network training have not proven revolutionary, it is nonetheless satisfying to confirm that a slight benefit does indeed exist. The boundary effects could perhaps be managed by expanding the bounds of the parameter space, such that the region within which the neural network performs well corresponds to the region of reasonable estimates on EoR parameter values. Regardless of the remedy, the logical 'next step' seems to be to explore the source of these border effects in neural network training. The hope is that the benefits of the resampling methods could be extended to the entire parameter space. It is also possible that the improvements seen from the adaptive grid-free method may be more substantial when applied to a higher dimensional parameter space.

There are also a number of other directions in which this foundation could be expanded. Both methods were tested on a relatively simple 3D parameter space, and the power spectrum was used in the definition of distance between observables. Once the optimization algorithms are perfected, it would be worthwhile testing them on higher dimensional parameter spaces, as well as examining how the PDF (or other observables in general) effect the performance. Trying to optimize a parameter space constructed with a full-dynamical code (such as LICORICE) would also be interesting. A large part of the choice of using 21cmFAST as the 'sandbox code' with which to test the optimization algorithms is the speed at which it runs. This allowed us to run it thousands of times in parallel over the period of only a few days. When, after having created a space of observables, an error was found, this realization was not 'catastrophic' as the observables could be recreated in a reasonable amount of time. We would lose this liberty with a more heavy-duty code, and should be sure of our decision before creating such a database (for example a future updated 21SSD).

Besides pushing towards more ambitious and complex data sets and algorithms, much of what has been accomplished here should also be revisited and updated to be SKA realistic. The noise model outlined in the first part should be compared and standardized with other authors (this is already under way). The resulting noisy lightcones (at SKA resolutions) could then be used to re-test the distance diagnostics (power spectra and PDF) used in the second part (recall that, although we created noisy SKA resolution power spectra and PDF, they were not used in calculating the L_2 norm distances). Much of the framework in the third part could also be improved by taking into account the SKA-context within which these ideas are intended to operate. As we advance along the SKA's construction time line, the technical specifications of the project will become clearer, making this task easier. One of the issues we encountered occasionally when working on our noise model was the difficulty in tracking down definitive standardized values, in particular information relating to the dipoles and their arrangement into stations

Our focus on EoR experiments should not be so narrow as to ignore the ever-growing number of other excellent instrumental efforts. Much of the framework presented here could (and should) be tested on existing SKA pathfinders (LOFAR, for example). Even with the lack of a definitive EoR 21 cm detection (as the EDGES detection concerns the Cosmic Dawn, and remains speculative), and even though it may be a number of years before we have EoR power spectra and tomographies, noise models can still be tested, and parameter reconstruction techniques can still be attempted with the global signal measurements (which should be coming in the very near future). The next year will be an eventful one in this respect. Three powerful new instruments are expected to go online (NCLE, NenuFAR, and HERA), any of which could very well be the first to definitively detect the 21 cm signal (if LOFAR does not do so before).

Turning to the more personal now, and allowing for a moment of retrospection, I'm pleased with how this PhD has turned out. At the very beginning, the prospect of training neural networks to better extract EoR parameter values (how the project was advertised) was quite exciting. Over the course of the three years — as is perhaps the hallmark of research — unexpected hurdles sprung up en route. Three in particular stand out as the most 'memorable' (similar to how walking over hot coals would be memorable). I include them here as a warning to others in the domain.

Properly simulating noise was the first of these. Digging through articles, trying to interpret formulae written with different variable formalisms, and pinning down current SKA specifications, was trying (to say the least).

The second was a small 'clean-up' routine I built into my 21cmFAST cloning code, intended to deal with the millions of files generated when running thousands of clones in parallel. When run on the supercomputer, the overambitious code — a mere three lines — eagerly removed all files in my directory (including itself, in fact). It took three weeks to re-code all that had been lost (and foolishly not backed up).

The third, and easily the worst, was the issue of eigenvector inversion. In this thesis, eigenvalue inversion inhabits but a single page. Yet outside of the thesis it inhabited many *many* long nights, and was responsible for a number of literal headaches. Like a hydra,

every time I had the hubris to think I'd solved the problem and finally oriented all of the eigenvectors properly, three new regions of the parameter space would pop-up and rear their improperly aligned heads. Finally, extrapolating the rotation bested the beast, and when staring at the properly aligned eigenvector field for the first time, the overwhelming beauty may have provoked a tear.

Suffice to say, to have seen the prospect of testing neural networks growing more distant with each hurdle, having my work finally being used for exactly this was extremely satisfying (perhaps redeeming is a better word). Although I personally had no part in building the neural networks themselves, receiving the image that quantified their performance when applied to my samplings left me with a sense of having 'accomplished my goal' (and not a moment too soon, as I was well into writing this manuscript at the time). In hindsight, what I'm most proud of is the code to initiate, run, simplify, and compare thousands of 21cmFAST clones in parallel. After fixing the renegade clean-up routine, the code worked fast and efficiently. All-in-all I estimate over 10,000 clones were created and run in testing various samplings.

And so I leave on a happy note, pleased with what I've accomplished in Paris (both inside and outside of my research), and thankful to the city for having provided me these experiences, as well as the chance to add a minuscule iota to human knowledge.

Appendices

APPENDIX A

Thermal Width of the 21 cm Line Profile

In equation 12 we removed a number of terms from the integral $\int_{\infty}^{\nu_0} \Phi(\nu_g) d\nu_g$, making the assumption that the terms should not vary along the line of sight on the spatial scale of the line profile thermal width. We now test this by equating the translational and kinetic energies:

$$\frac{1}{2} m_p v^2 = \frac{3}{2} k_B T \tag{A.1}$$

where m_p is the mass of hydrogen, v is the thermal velocity, k_b is the Boltzmann constant, and T is the temperature of the IGM. Using a rough estimate for the temperature of $T = 10$ K, we find that $v \approx 0.5$ km/s. This will give a line width of:

$$\Delta l = \frac{v}{H(z)} \tag{A.2}$$

Assuming $H(z) \approx 1000$ km/s/Mpc for EoR redshifts, we conclude that $\Delta l \approx 0.5$ kpc. Recall that we are using proper velocity (km/s), so this is a proper distance. Switching to comoving distance, we arrive at a result closer to ∼a few ckpc. This is small enough that our assumption will, indeed, hold. However, it will no longer be valid when considering mini-halos, which can also be ∼a few ckpc in size.

Note: In principle, the temperature and Hubble expansion may be slightly larger, depending on the redshift. However, even a slightly larger Δl will remain below the scale at which the properties of the IGM fluctuate.

APPENDIX B

Sample Code

3D Fourier Transform for Power Spectra

This Fortran code generalizes the MKL Fourier transform routine (DFTI) to 3D, and properly normalizes the output.

```
Use MKL_DFTI
implicit none

Mpc = 200 !Physical size of slices in cMpc
grid_size = 1024 !Size of slice in pixels
complex(WP), dimension(grid_size/2+1,grid_size,grid_size) :: ps_complex
real(WP), dimension(grid_size/2,grid_size,grid_size) :: real_cube

status = DftiCreateDescriptor(plan_forwards, DFTI_SINGLE, DFTI_REAL, 3, L)
status = DftiSetValue(plan_forwards, DFTI_PLACEMENT, DFTI_NOT_INPLACE)
status = DftiSetValue(plan_forwards, DFTI_CONJUGATE_EVEN_STORAGE,
DFTI_COMPLEX_COMPLEX)
       !MKL doesn't set up the required data spacing for 3D FT,
       so it has to be done manually
       cstrides = [0, 1, INT(L(1)/2.0)+1, L(2)*(INT(L(1)/2.0)+1)]
       rstrides = [0, 1, L(1), L(2)*L(1)]
status = DftiSetValue(plan_forwards, DFTI_INPUT_STRIDES, rstrides)
status = DftiSetValue(plan_forwards, DFTI_OUTPUT_STRIDES, cstrides)
status = DftiCommitDescriptor(plan_forwards)
status = DftiComputeForward(plan_forwards, ps_cube(:,1,1), ps_complex(:,1,1))
status = DftiFreeDescriptor(plan_forwards)

!ORGANIZE THE FOURIER TRANSFORM OUTPUT
print*,"Rearranging Fourier Parameter Space..."
do i = 1, grid_size
       do j = 1, grid_size
```

```
            do k = 1, grid_size/2
                   !Take the modulus of each value and normalize (divide by
                      cube volume)
                   real_cube(k, j, i) = sqrt(real(ps_complex(k,j,i))**2 +
                      aimag(ps_complex(k,j,i))**2)/sqrt(real(grid_size**3))*
                      Mpc**3
            enddo
      enddo
enddo
```

Pixel Distribution Function Code

This Fortran code creates the PDF.

```
      !Set up the bin limits.
      do i = 1, a_bin_nb
            a_bins(i) = (real(i)/a_bin_nb)*(a_f-a_i)+a_i
      enddo
      do i = 1, Tb_bin_nb
            Tb_bins(i) = (real(i)/Tb_bin_nb)*(Tb_max-Tb_min)+Tb_min
      enddo
      do j = 1, Tb_bin_nb
            do i = 1, a_bin_nb
                   !Set all values of the histogram to 1 (after we take the
                      log they will be zero).
                   Tb_histogram(i,j) = 1
            enddo
      enddo

      !Calculate the average density if necessary (if not just set the whole
         density array to zero)
      avg_dens = 0
      do i = 1, nz
            do j = 1, grid_size
                   do k = 1, grid_size
                         if (dens_cutoff) avg_dens = avg_dens + dens(k,j,i)
                         if (.not. dens_cutoff) dens(k,j,i) = 0
                   enddo
            enddo
      enddo
      if (dens_cutoff) avg_dens = avg_dens/real(grid_size*grid_size*nz)

      !For each slice, find the average Temperature
      do k = 1, nz
            T_avg = SUM(tk(:,:,k))/grid_size**2
            a_bin = 1
```

```
        !Find the expansion factor of the current slice, and then which
           expansion factor bin it falls into
        a = (real(k)/real(nz))*(a_f-a_i)+a_i
        do while (a > a_bins(a_bin))
               a_bin = a_bin + 1
        enddo
        do j = 1, grid_size
               do i = 1, grid_size
                      !For the current slice, find all pixels that are
                         under the density cutoff
                      if (.not. isnan(tk(i,j,k)) .and. dens(i,j,k) <=
                         avg_dens*100) then
                          !Find which Tb bin each pixel fals into, and
                               augment that bin in the histogram
                          if (subtract_avg) then
                                 Tb_bin = int((tk(i,j,k)-T_avg - Tb_min
                                    )/(Tb_max-Tb_min)*Tb_bin_nb) + 1
                                 if (Tb_bin <= Tb_bin_nb .and. Tb_bin >
                                      0)
                                 Tb_histogram(a_bin,Tb_bin) =
                                    Tb_histogram(a_bin,Tb_bin) + 1
                          else
                                 Tb_bin = int((tk(i,j,k) - Tb_min)/(
                                    Tb_max-Tb_min)*Tb_bin_nb) + 1
                                 if (Tb_bin <= Tb_bin_nb .and. Tb_bin >
                                      0) Tb_histogram(a_bin,Tb_bin) =
                                    Tb_histogram(a_bin,Tb_bin) + 1
                          endif
                      endif
               enddo
        enddo
enddo

!Create filename and save out the data
print*,'Writing file ',filename_out
open(20,file=filename_out,status='replace',form='formatted')
do ix=1,a_bin_nb
        do iy=1,Tb_bin_nb
               write(20,*) 1/((ix-0.5)/real(a_bin_nb)*(a_f-a_i)+a_i) - 1,
                    &
                      (iy-0.5)/real(Tb_bin_nb)*(Tb_max-Tb_min)+Tb_min, &
                      log10(real(Tb_histogram(ix,iy)))
        enddo
        write(20,*)
enddo
```

21cmFAST MPI Cloning

This code initializes and runs clones of 21cmFAST (each clone has its parameters changed first).

```
//Compile with mpicc run_clones_MPI.c -o run_clones_MPI
//Run with mpirun -np # run_clones_MPI
//Set mode and input parameters below
//Makes ure there is a '21cmFAST-clones' directory, and that there's nothing
    important in it. That's where the program will write to.

int main(){
  MPI_Init(NULL, NULL);

  //This is for an 8x8x8 sampling. Change as required.
  float HII_EFF_FACTOR[] = {20, 27.79, 38.61, 53.65, 74.55, 103.59, 143.94,
      200};
  float R_BUBBLE_MAX[] = {5, 6.60, 8.72, 11.51, 15.20, 20.07, 26.51, 35};
  float ION_Tvir_MIN[] ={8e3, 1.15e4, 1.65e4, 2.36e4, 3.39e4, 4.86e4, 6.97e4, 1
      e5};

 char cmnd[1000];
  char filename[1000];
  int i,j,l,i_max,j_max,l_max,threads,id,clone,total_clones;
  int nb_tasks, my_rank, err_mpi, nb_lines, line;
  float v1,v2,v3,w1,w2,w3;
  char w1_char[25], w2_char[25], w3_char[25];
  FILE *file;

//=============================================================================

//=============================================================================

  //USE THIS TO DECIDE HOW THE CODE WORKS!!
  int NORMAL = 0; //Uses the parameter space defined above
  int RANDOM_CHOICE = 0; //Randomly choose parameter values
  int UNIFORM_IN_LOGSPACE = 1; //Parameters will be uniformly chosen in log
      space, as opposed to linear
  int READ_IN = 1; //Reads in parameter values (specify filename below)
  int ADAPTIVE = 0; //Read in the adaptive mesh file (READ_IN should also equal
       1)
  if (ADAPTIVE) READ_IN = 1;
  sprintf(filename,"resampled_sqrtlambda.dat");
//=============================================================================

//=============================================================================


  err_mpi = MPI_Comm_size(MPI_COMM_WORLD, &nb_tasks);
```

```
  err_mpi = MPI_Comm_rank(MPI_COMM_WORLD, &my_rank);

  i_max = sizeof(HII_EFF_FACTOR)/sizeof(HII_EFF_FACTOR[0]);
  j_max = sizeof(R_BUBBLE_MAX)/sizeof(R_BUBBLE_MAX[0]);
  l_max = sizeof(ION_Tvir_MIN)/sizeof(ION_Tvir_MIN[0]);

  total_clones = i_max*j_max*l_max;
  if (RANDOM_CHOICE) total_clones = nb_tasks;
  if (ADAPTIVE) total_clones = nb_tasks;

  clone = my_rank;
  i = clone % i_max;
  j = clone / i_max % j_max;
  l = clone / (i_max*j_max);

if (RANDOM_CHOICE){
       srand(my_rank*1000);
       if (UNIFORM_IN_LOGSPACE) {
              HII_EFF_FACTOR[i] = pow(10,log10(HII_EFF_FACTOR[0])+((double)
                  rand() / (double) RAND_MAX)*(log10(HII_EFF_FACTOR[i_max-1])-
                  log10(HII_EFF_FACTOR[0])));
              R_BUBBLE_MAX[j] = pow(10,log10(R_BUBBLE_MAX[0])+((double) rand()
                  / (double) RAND_MAX)*(log10(R_BUBBLE_MAX[j_max-1])-log10(
                  R_BUBBLE_MAX[0])));
              ION_Tvir_MIN[l] = pow(10,log10(ION_Tvir_MIN[0])+((double) rand()
                  / (double) RAND_MAX)*(log10(ION_Tvir_MIN[l_max-1])-log10(
                  ION_Tvir_MIN[0])));
              fprintf(stderr,"%f\t%f\t%f\n",HII_EFF_FACTOR[i],R_BUBBLE_MAX[j],
                  ION_Tvir_MIN[l]);
       } else {
              HII_EFF_FACTOR[i] = ((double) rand() / (double) RAND_MAX)*(
                  HII_EFF_FACTOR[i_max-1] - HII_EFF_FACTOR[0]) + HII_EFF_FACTOR
                  [0];
              R_BUBBLE_MAX[j] = ((double) rand() / (double) RAND_MAX)*(
                  R_BUBBLE_MAX[j_max-1] - R_BUBBLE_MAX[0]) + R_BUBBLE_MAX[0];
              ION_Tvir_MIN[l] = ((double) rand() / (double) RAND_MAX)*(
                  ION_Tvir_MIN[l_max-1] - ION_Tvir_MIN[0]) + ION_Tvir_MIN[0];
       }
  }

  //If the file of resampled points is to be read:
  nb_lines = 0;
  if (READ_IN){
       if (!ADAPTIVE){
              file = fopen(filename,"r");
              while (fscanf(file, "%i\t%i\t%i\t%f\t%f\t%f\t%f\t%f\t%f\t%f\n",&i,&j
                  ,&l,&v1,&v2,&v3,&w1,&w2,&w3) == 9){
                     if (w1 != 0 || w2 != 0 || w3 != 0){nb_lines++;}
              }
```

```
            fclose (file);
      }else{
            file = fopen("adaptive.dat","r");
            while (fscanf(file, "%16s%16s%16s",w1_char,w2_char,w3_char) == 3)
                {nb_lines++;}
            fclose (file);
      }
}
float resampled[nb_lines][6];
if (READ_IN){
      line = 0;
      if (!ADAPTIVE){
            file = fopen(filename,"r");
            while (fscanf(file, "%i\t%i\t%i\t%f\t%f\t%f\t%f\t%f\t%f\t%f\n",&i,&j
                ,&l,&v1,&v2,&v3,&w1,&w2,&w3) == 9){
                   if (w1 != 0 || w2 != 0 || w3 != 0){
                         resampled[line][0] = (float) i;
                         resampled[line][1] = (float) j;
                         resampled[line][2] = (float) l;
                         resampled[line][3] = w1;
                         resampled[line][4] = w2;
                         resampled[line][5] = w3;
                         line++;
                   }
            }
            fclose (file);
      }else{
            file = fopen("adaptive.dat","r");
            while (fscanf(file, "%16s%16s%16s",w1_char,w2_char,w3_char) == 3)
                {
                   resampled[line][0] = resampled[line][1] = resampled[line
                       ][2] = 0.;
                   resampled[line][3] = atof(w1_char);
                   resampled[line][4] = atof(w2_char);
                   resampled[line][5] = atof(w3_char);
                   //if (my_rank == 1) fprintf(stderr,"%f\t%f\t%f\n",
                       resampled[line][3],resampled[line][4],resampled[line
                       ][5]);
                   line++;
            }
      }
} else {float resampled[1][6];}

//Make sure the clone directory is empty
//system("rm -rf 21cmFAST-clones/*");

if (NORMAL) fprintf(stderr, "Creating clone %i/%i with parameters f = %.0f, R
     = %.0f, T = %.1e\n",clone, total_clones, HII_EFF_FACTOR[i], R_BUBBLE_MAX
    [j], ION_Tvir_MIN[l]);
```

```
//if (RANDOM_CHOICE) fprintf(stderr, "Creating clone %i/%i with parameters f
    = %.0f, R = %.0f, T = %.1e\n",clone, nb_tasks, HII_EFF_FACTOR[i],
    R_BUBBLE_MAX[j], ION_Tvir_MIN[l]);
if (READ_IN) fprintf(stderr, "Creating clone %i/%i with parameters f = %.0f,
    R = %.0f, T = %.1e\n",clone, nb_lines, resampled[clone][3], resampled[
    clone][4], resampled[clone][5]);
//Copy 21cmFAST-master into the clones directory and rename the clone with
    the relative parameters
if (!READ_IN) sprintf(cmnd,"cp -R 21cmFAST-master 21cmFAST-clones/f%.0f_R%.0
    f_T%.1e", HII_EFF_FACTOR[i], R_BUBBLE_MAX[j], ION_Tvir_MIN[l]);
if (READ_IN) sprintf(cmnd,"cp -R 21cmFAST-master 21cmFAST-clones/f%.0f_R%.0
    f_T%.1e", resampled[clone][3], resampled[clone][4], resampled[clone][5]);
system(cmnd);
//Update the parameter files (The values in the master should be 15, 30, 1e4
    respectively for this to work)
if (!READ_IN){
     sprintf(cmnd, "sed -i 's/HII_EFF_FACTOR (float) (30)/HII_EFF_FACTOR (
        float) (%.2f)/g' 21cmFAST-clones/f%.0f_R%.0f_T%.1e/Parameter_files/
        ANAL_PARAMS.H", HII_EFF_FACTOR[i], HII_EFF_FACTOR[i], R_BUBBLE_MAX[j
        ], ION_Tvir_MIN[l]);
     system(cmnd);
     sprintf(cmnd, "sed -i 's/R_BUBBLE_MAX (float) (50)/R_BUBBLE_MAX (float)
        (%.2f)/g' 21cmFAST-clones/f%.0f_R%.0f_T%.1e/Parameter_files/
        ANAL_PARAMS.H", R_BUBBLE_MAX[j], HII_EFF_FACTOR[i], R_BUBBLE_MAX[j],
         ION_Tvir_MIN[l]);
     system(cmnd);
     sprintf(cmnd, "sed -i 's/ION_Tvir_MIN (double) (3e4)/ION_Tvir_MIN (
        double) (%.2e)/g' 21cmFAST-clones/f%.0f_R%.0f_T%.1e/Parameter_files/
        ANAL_PARAMS.H", ION_Tvir_MIN[l], HII_EFF_FACTOR[i], R_BUBBLE_MAX[j],
         ION_Tvir_MIN[l]);
     system(cmnd);
}else{
     sprintf(cmnd, "sed -i 's/HII_EFF_FACTOR (float) (30)/HII_EFF_FACTOR (
        float) (%.2f)/g' 21cmFAST-clones/f%.0f_R%.0f_T%.1e/Parameter_files/
        ANAL_PARAMS.H", resampled[clone][3], resampled[clone][3], resampled[
        clone][4], resampled[clone][5]);
     system(cmnd);
     sprintf(cmnd, "sed -i 's/R_BUBBLE_MAX (float) (50)/R_BUBBLE_MAX (float)
        (%.2f)/g' 21cmFAST-clones/f%.0f_R%.0f_T%.1e/Parameter_files/
        ANAL_PARAMS.H", resampled[clone][4], resampled[clone][3], resampled[
        clone][4], resampled[clone][5]);
     system(cmnd);
     sprintf(cmnd, "sed -i 's/ION_Tvir_MIN (double) (3e4)/ION_Tvir_MIN (
        double) (%.2e)/g' 21cmFAST-clones/f%.0f_R%.0f_T%.1e/Parameter_files/
        ANAL_PARAMS.H", resampled[clone][5], resampled[clone][3], resampled[
        clone][4], resampled[clone][5]);
     system(cmnd);
}
//Go into the new clone's Program directory and run 'make'
```

```
  if (!READ_IN) sprintf(cmnd,"21cmFAST-clones/f%.0f_R%.0f_T%.1e/Programs",
     HII_EFF_FACTOR[i], R_BUBBLE_MAX[j], ION_Tvir_MIN[l]);
  if (READ_IN) sprintf(cmnd,"21cmFAST-clones/f%.0f_R%.0f_T%.1e/Programs",
     resampled[clone][3], resampled[clone][4], resampled[clone][5]);
  chdir(cmnd);
  system("make");
  //system("rm *.c");
  //system("rm PROGRAM_LIST");
  //system("rm Makefile");
  //chdir("../../..");

  if (NORMAL) fprintf(stderr, "Treating clone = %i/%i with thread %i\n", clone,
     total_clones,clone);
  if (RANDOM_CHOICE) fprintf(stderr, "Treating clone = %i/%i with thread %i\n",
      clone,nb_tasks,clone);
  if (READ_IN) fprintf(stderr, "Treating clone = %i/%i with thread %i\n", clone
     ,nb_lines,clone);

  //Go to the clone directory
  if (!READ_IN) sprintf(cmnd,"21cmFAST-clones/f%.0f_R%.0f_T%.1e/Programs",
     HII_EFF_FACTOR[i], R_BUBBLE_MAX[j], ION_Tvir_MIN[l]);
  if (READ_IN) sprintf(cmnd,"21cmFAST-clones/f%.0f_R%.0f_T%.1e/Programs",
     resampled[clone][3], resampled[clone][4], resampled[clone][5]);
  chdir(cmnd);
  //Run 21cmFAST
  system("./drive_zscroll_noTs");
  chdir("..");
  system("mv Output_files/Deldel_T_power_spec/* .");
  //system("rm -r */");
  chdir("../..");

  fprintf(stderr, "Clone %i/%i complete\n **************************\n",
     clone, total_clones);
  MPI_Finalize();

  return 0;
}
```

APPENDIX C

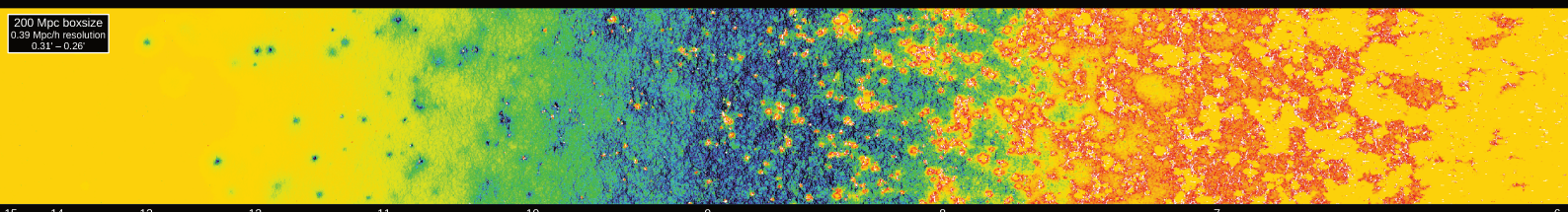
Conference Posters

Grooving to SKA Noise

Evan Eames[1*], Florian Bolgar[1], and Benoît Semelin[1]

[1]*LERMA, Observatoire de Paris*

**evan.eames@obspm.fr*



Preparing for SKA Data Intake

With the date of SKA's first light rapidly approaching, much work must be done in order to prepare for SKA data intake. Here we examine the effects of realistic SKA level noise on 21 cm tomography. This helps us to predict how deep, and at what resolution, we should realistically expect to probe with the SKA. Lightcones are made using LICORICE, a numerical particle-based code which simulates the cosmological evolution of gas, DM, and radiation (Semelin 2016). We use realistic SKA-1 parameters, which are listed below. These lightcones are then postprocessed to add thermal noise as a function of redshift and angular resolution. By looking at slices of these lightcones we can see at which redshifts noise overpowers for a given angular resolution.

The ultimate goal of this research is threefold. Firstly, it will help us to better understand the limitations of the SKA. Secondly, we are currently building a database of noisy lightcones created using simulations with different astrophysical parameters, which will serve to train algorithms for parameter reconstruction. Thirdly, we are beginning to conduct preliminary studies into using this database to train a neural network, which we hope will help extract astrophysical parameters from upcoming SKA tomographies.

Simulating SKA Level Noise

To realistically approximate SKA noise we refer to the methodology as it is presented in Koopmans et al 2015, which is based off the word in Mellema et al 2013 and McQuinn et al 2006. We use a preliminary fit, meaning it holds for resolutions $\Delta\theta < \sim 10'$.

$$\Delta T_b = 8.67 \times 10^{-3} \times \frac{z^{4.105}\Delta\theta^{-1.71}}{\sqrt{B}} \text{ mK}$$

Where B is bandwidth in MHz, z is redshift, and $\Delta\theta$ is resolution in arcminutes. The lightcones presented here are made such that expansion factor change from one slice to the next 'da' is constant, so the bandwidth term is simply B = 1420 MHz x da. $\Delta\theta$ can also be expressed in terms of redshift, pixel resolution, and boxsize:

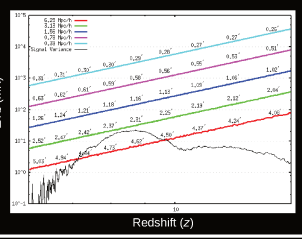
$$\Delta\theta = \frac{d_{\text{Mpc}} \times h^{-1}}{2.02 \times d_{\text{pix}}} \left(\frac{10}{1+z}\right)^{0.277}$$

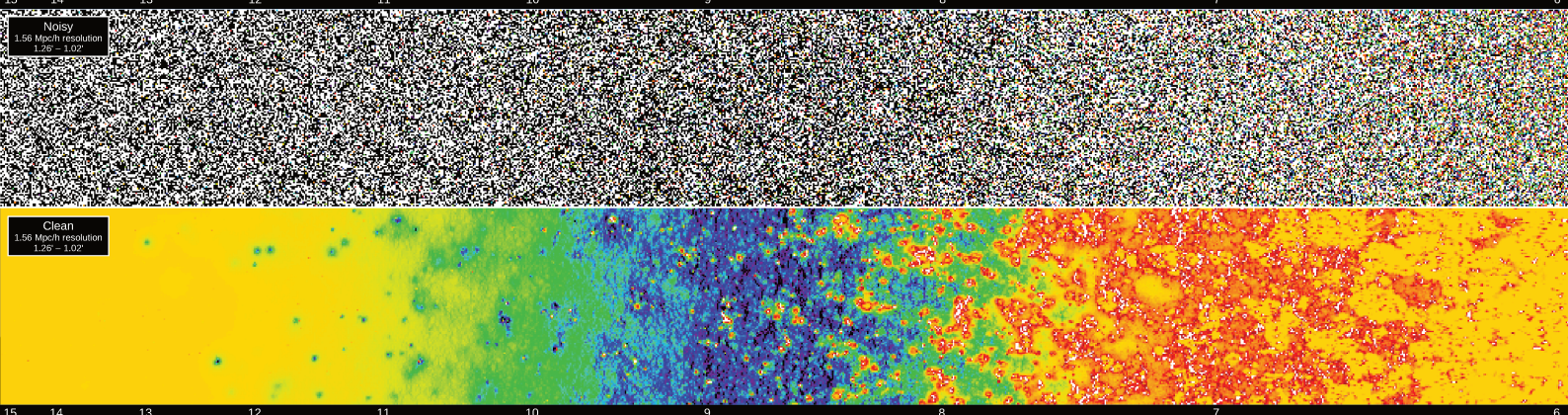
Where the d values are the lightcone box sizes in Mpc and pixels, and the result is in arcminutes. With this formula in mind we can study the severity of the noise as we adjust the resolution and redshift, as can be seen in the following figure.

T_b Sensitivity

Noise levels for various resolutions and redshifts. The numbers superimposed on the curves show the corresponding angular resolution for a given redshift, in arcminutes.

The black line shows the variance of the simulated signal over the course of its evolutionary history. Even in regions where the signal is below the noise level, it is sometimes still possible to make out strong features, such as can be seen in the low-resolution lightcones (6.25 Mpc/h) at low-redshifts ($z \sim 6$).





Recovering Signal

The full, high-resolution lightcone shown at top is an unachievable level of precision. On account of the unrealistically small angular resolution at this level (0.31' – 0.26') the $\Delta\theta$ term in the ΔT_b equation causes the noise to skyrocket. Even at z = 6 it is impossible to make out any details. Directly above is a lightcone at a resolution of 1.56 Mpc/h. Even here the noise dominates at EoR redshifts.

The only reasonable way to get around this is to drop to lower resolutions, such as the lightcones shown below, which were made at 6.23 Mpc/h. At this point we can see some structure begin to emerge. Down to z = 8.5 some form is visible, however much of the large scale structure is lost at earlier redshifts, where the signal is weaker.

One solution could be to push the integration time past 1000 h. The bottom-most lightcone has been made identically to the one directly below, however the integration time was 4000 h (effectively halving the noise). Here we can see structure down to nearly $z \approx 10$, which corresponds to the beginning of the EoR.

Improving the Noise Model

At present, as was mentioned before, the model we use for noise is an approximation based on a model created by Koopmans et al. Our simplified fit is only known to hold between $-1' < \Delta\theta < -10'$. Therefore, there is no way to guarantee that the noise presented for higher resolution models is accurate. However, SKA tomography will fall within these bound. With this in mind, the coming steps will be as follows.

Implement Koopmans' model fully, to assure all resolutions are accurate. → Decide upon a fiducial resolution which best suits the SKA project. → Reproduce this lightcone with different parameters to build a database.

Note: It is also important to keep in mind that the field of view of the SKA will be much bigger than what we have presented here (~10 times). Although this will not improve the resolution, it could allow us to see very large-scale structures not reachable here. Station layout is also a crucial factor, and other designs may further reduce noise.

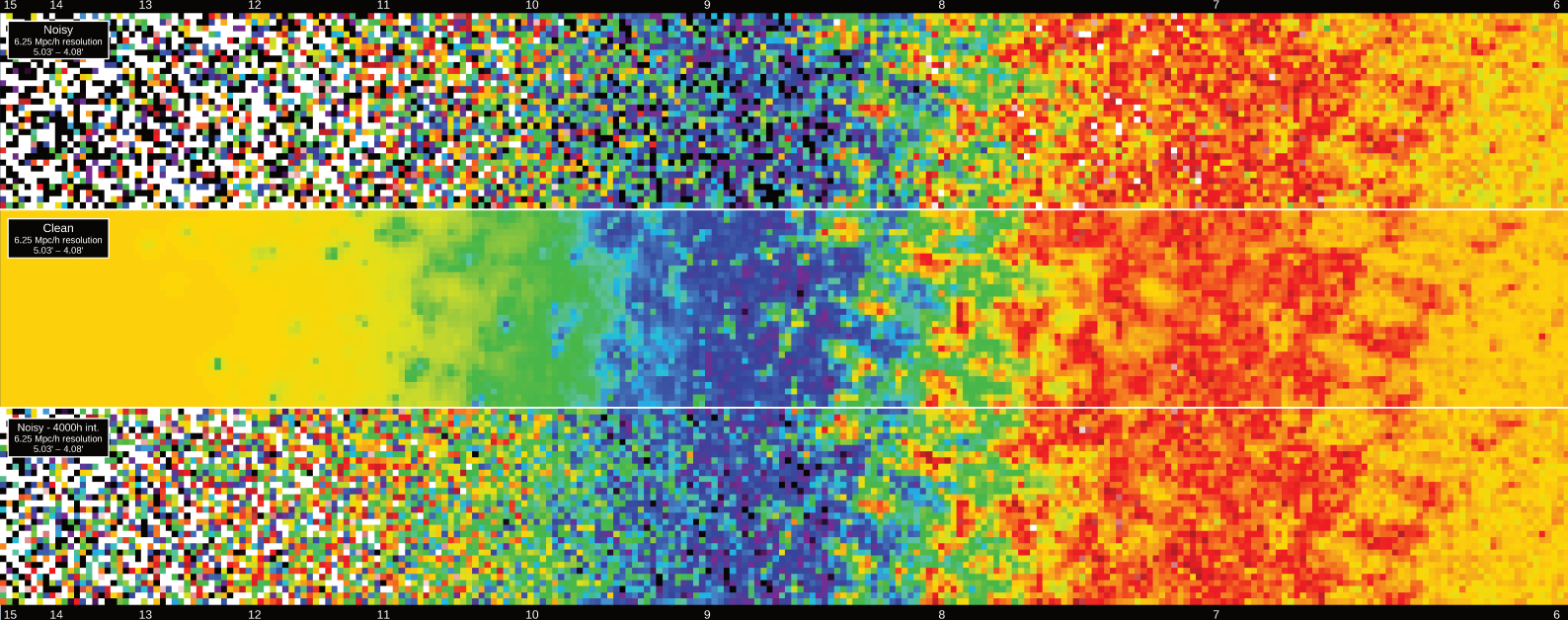
Building a Database

We are currently working on a database of similar simulations, with the hope of training a neural network for parameter reconstruction. By next year we hope to have ~100 simulations at full resolution, which will be freely available.

The LICORICE Simulation:
- 200 Mpc/h
- 1024^3 resolution elements
- Coupled-Radiative Hydrodynamics
- 3D radiation transfer for L-α coupling
- X-ray heating

The Parameter Space:
- f_x = 0.1, 1.0, 10 (x-ray production efficiency)
- f_α (L-α coupling efficiency)
- $f_{H/S}$ = 0.0, 0.5, 1.0 (Hard/Soft x-ray fraction)
- Potential parameters (Greig 2015):
  - R_{mfp}, f_{esc}

Although the database is unfinished, we have begun preliminary testing of neural network training on 21 cm power spectra generated with 21cmFAST. The initial results, and a discussion of the feasibility of the concept, are expected to be published in 2017.



Parameters

t_{int}	1000 h	Integration time	Dewdney 2015
d_{core}	400 m	Core diameter	Dewdney 2015
N_{ant}	131072	Number of antennas	Dewdney 2015
A_{eff}	$13E4(\min(\lambda^2/3,2.56))$	Effective area	Semelin 2016
A_{coll}	$N_{stat} \times A_{eff}$	Collecting area	Koopmans 2015
T_{sys}	$100+400(\nu/150)^{-2.55}$	System temperature	Koopmans 2015

Conclusions

Here we have presented a first look at the severity of noise in SKA tomography. From this preliminary study we have seen how deep we can expect to probe. Reducing resolution to be on the order of ~5' and pushing the integration to the extremely ambitious value of 4000 h, structure is somewhat visible at z = 9.5. Small-scale details are unobtainable, however, even at $z \approx 6$. Nonetheless, this would still provide us with a glimpse of large-scale structure during the EoR, and could be sufficient to constrain many parameters associated with this epoch. It is also worth keeping in mind that station layout is a huge factor that needs to be explored further.

These findings are expected to be published in more detail in early 2017.

References

- Dewdney, 2015 SKA doc #: SKA-TEL-SKO0000308
- Greig & Mesinger, 2015 arxiv.org/abs/1501.06576
- Koopmans, 2015 arxiv.org/abs/1505.07568
- McQuinn, 2006 arxiv.org/abs/astro-ph/0512263
- Mellema, 2013 arxiv.org/abs/1210.0197
- Semelin, 2016 arxiv.org/abs/1510.02296

Optimally Sampling a 21-cm Parameter Space

Evan Eames[1*], Aristide Doussot[1], and Benoît Semelin[1]

[1]LERMA, Observatoire de Paris

**evan.eames@obspm.fr*

Step 0 – Why are we doing this?

In the context of 21-cm cosmology, there is a pressing need for a solid theoretical framework with which future experimental data can be understood. Much of this work involves effectively constraining parameter values based on the nature of the 21-cm signal. Towards this goal, a number of simulations have attempted to replicate the expected signal, and examine the range of shapes the signal can take. This entails creating databases of expected tomographies (21ssd.obspm.fr).

Yet, for any such database, the choice of parameter values sampled is somewhat arbitrary. This begs the question of whether we can decide upon an '*optimal*' sampling of parameter values - optimal in that the resulting models are all unique, and therefore efficiently span the space of reionization scenarios. To arrive at such a sampling, a 3-dimensional parameter space was explored, and 2,400 points were selected. The resulting 2,400 models are then compared, and various methods are used to achieve an optimal sampling.

Step 1 – Create a Parameter Space

21cmFAST is used, and the parameters to be explored are (as seen previously in Greg and Mesinger, 2015):

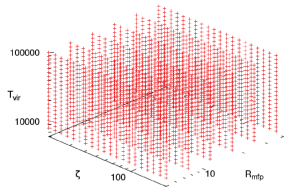
- ζ – The Ionization Efficiency
- Rmfp – The Mean Free Path
- Tvir – The Minimum Virial Temperature

These values are sampled logarithmically. Initial exploration of a 10x10x10 parameter space showed that the physical morphology of the models is much less sensitive to changes in the Rmfp than to the other two variables, hence why it is sampled more sparcely in the final 20x6x20 sampling. The complete list of parameter values is given in Table 1, and illustrated in the accompanying figure.

Table 1 – Initial Parameter Sampling

Paramter	Explored values
ζ	20.00, 22.58, 25.49, 28.77, 32.48, 36.66, 41.33, 46.71, 52.73, 59.53, 67.20, 75.85, 85.63, 96.66, 109.11, 123.17, 139.04, 156.95, 177.17, 200.00
$R_{\rm mfp}$	5.00, 7.38, 10.89, 16.07, 23.72, 35.00
$T_{\rm vir}/10^4$	0.80, 0.91, 1.04, 1.19, 1.36, 1.56, 1.78, 2.03, 2.32, 2.65, 3.02, 3.45, 3.94, 4.50, 5.14, 5.88 6.71, 7.67, 8.76, 10.00

Figure 1 – Initial Parameter Space



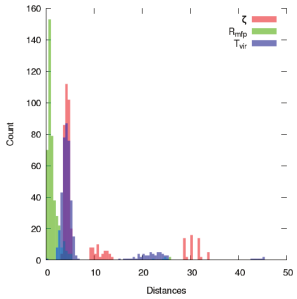
Step 2 – Calculate the Distances Between Models

We can now define a distance measure between the models. This is done with the L2 norm on the multi-redshift power spectra P(k,z), which is a function of both k-space and redshift. For our purposes 6 ≤ z ≤ 15 and Δz = 1. With this, the distance between two models 'i' and 'j' is given as:

$$D_{i,j} = \sqrt{\int (P_i(k,z) - P_j(k,z))^2 dk dz}$$

To understand the geometry of our current parameter space, we can ask what the distances between neighbouring models are in the three dimensions of our Parameter space.

Figure 2 – Distance Histogram



This histogram reveals a high degree of anisotropy and inhomogeneity in our parameter space. Along the Rmfp axis, simulations are very similar to one another compared to the other two axes. In addition, some models are an order of magnitude more different than their neighbours, corresponding to regions where the morphologies of the models are changing rapidly.

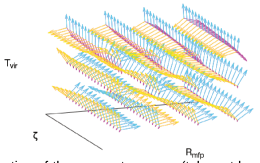
Step 3 – Find Eigenvectors and Eigenvalues

In order to proceed, we can calculate the eigenvectors and their accompanying eigenvalues. To do this we must first calculate the metrics at each point using the distances between all immediate neighbours. For a point 'i', the metric for a given point is calculated using the standard definition:

$$g = \begin{bmatrix} a_{xx} & a_{xy} & a_{xz} \\ a_{xy} & a_{yy} & a_{yz} \\ a_{xz} & a_{yz} & a_{zz} \end{bmatrix} \qquad a_{uv} = \begin{cases} \frac{D^2_{i,i+u} + D^2_{i,i-u}}{2} & \text{if } u = v \\ \frac{D^2_{i,i+u+v} + D^2_{i,i-u-v} - D^2_{i,i+u-v} - D^2_{i,i-u+v}}{8} & \text{if } u \neq v \end{cases}$$

Where u and v can be x, y, or z (corresponding to the three parameters), and $D_{i,i+u}$ implies the distance between model 'i' and and it's immediate neighbour in the positive 'u' direction. Using the metrics to find the eigenvectors at each point, we can begin to understand the geometry of the parameter space.

Figure 3 – Eigenvectors



This is a small section of the parameter space (taken at low Tvir). Regions where the eigenvectors are rotating rapidly represent areas where the morphologies of the models change rapidly. We can also find the corresponding eigenvalues.

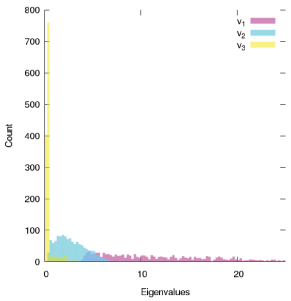


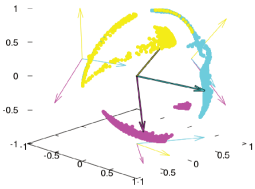
Figure 4 – Eigenvalues

The colours correspond with the eigenvectors above. We notice the three eigenvalues all display interesting differences in their form, with the smallest of the three showing much less variance than the other two. These can now be used to re-sample the parameter space.

Step 4 – Average the Eigenvectors and Create 'Step Vectors'

With the Eigenvalues and Eigenvectors, we have an understanding of the geometry of our parameter space, with which we can infer three new orthogonal 'step vectors'. Starting from the middle point, these step vectors will take us outwards in steps of equal L2 distance. We can construct these step vectors by first averaging the eigenvectors. The figure on the right shows the heads of all the eigenvectors (the three coloured clusters), the average eigenvectors* (thick vectors), and the 2D projections (thin vectors). The eigenvectors on the borders are not included to avoid edge effects.

Figure 5 – Averaged Eigenvectors



We must also know the average distance *d* between the models, which can be calculated from the individual eigenvalues, via the average volume:

$$V_i = \sqrt{\det i_i} = \sqrt{\lambda_1 \lambda_2 \lambda_3} \quad \text{and thus} \quad \bar{d} = \sqrt[3]{\bar{V}} = \sqrt[3]{\frac{1}{N} \sum_i \sqrt{\lambda_1 \lambda_2 \lambda_3}}$$

For an averaged eigenvector *vn* = *(xn,yn,zn)* with corresponding averaged eigenvalue* λ, we can therefore compute a new step vector as:

$$\vec{s}_n = \left(\frac{\bar{d}}{\sqrt{\lambda_n}} c_1 \bar{x}_n, \frac{\bar{d}}{\sqrt{\lambda_n}} c_2 \bar{y}_n, \frac{\bar{d}}{\sqrt{\lambda_n}} c_3 \bar{z}_n \right)$$

The *d*/√λ factor normalizes the resulting distances to assure each step vector corresponds to an L2 displacement of *d*, and the *c* factors are necessary because our initial sampling was linear in *logarithmic* space. They are factors of constant distance in log space, defined as:

$$c_1 = \Delta \log \zeta, c_2 = \Delta \log R_{\rm mfp}, c_3 = \Delta \log T_{\rm vir}$$

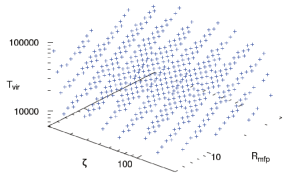
Step 5 – Re-sample the Parameter Space

Using the step vectors defined in the previous step, we arrive at a re-sampling of our parameter space. We are also free to adjust the average distance factor (*d*) to create a new parameter sampling with more or less points. We chose to create a number of different re-samplings, each with 512 points. These explore the effects of, for example:

- Taking the medians of the eigenvectors and eigenvalues, as opposed to the average.
- Ignoring regions of the parameter space where the eigenvectors change rapidly (as seen in Figure 3) in calculating the eigenvectors and eigenvalues.
- Averaging (or taking the median) of the quantity of 1/√λ for each point, as opposed to λ.

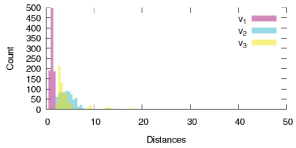
Finally it was decided that the most satisfactory method was to take the averages of the eigenvectors, the averages of the quantity 1/√λ for the eigenvalues, and the entire region (including areas where the eigenvectors change rapidly, however the borders were s

Figure 6 – Re-sampled Parameter Space



In this re-sampled space, the concept of neighbouring models is not as visually clear, however, it is still fairly straightforward. Starting from the central point (from which the re-sampling was carried out along the three step vectors), we can consider any two points that are separated by a single step vector (or the inverse of a step vector) to be neighbours. With this we are able to again calculate the average distance to the neighbouring models along the step vectors (v1,v2,v3).

Figure 7 – Distance Histogram After Re-sampling



See Step 6 for commentary on this figure.

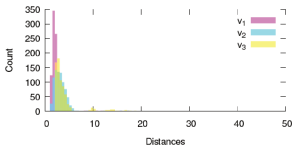
Step 6 – Adjust the Step Vectors

Looking at Figure 7, and comparing it to Figure 2, it is encouraging to see that the discrepancy between the distances has improved. The errant high-value eigenvalues between ~8 and ~45 are noticeably better after the re-sampling, with only a few left at ~13. As well, the lowest peak is now slightly closer to the other two.

However, the three peaks still do not overlap exactly, which tells us the parameter space is still not perfectly isotropic. We are still exploring the reason behind this. Somehow, the quantities of average eigenvectors and eigenvalues are unable to capture the true geometry of the space. Until this is understood, a straightforward additional step provides a solution. Each step vector is adjusted by the ratio of the average distance along the corresponding step vector before and after re-sampling:

$$\vec{s}_n = \vec{s}_n \cdot \frac{\bar{D}_{n,initial}}{\bar{D}_{n,re-sampled}}$$

Figure 8 – Distance Histogram After Step Vector Corrections

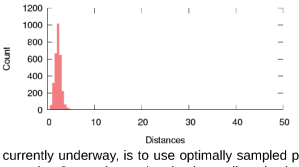


After applying this correction, as expected, the three peaks overlap. This means that we have arrived at a nearly isotropic and homogeneous sampling of our parameter space. The slight tails towards higher values of v2 and v3 correspond to the previously mentioned regions where eigenvectors change rapidly.

Step 7 – Explore Further

We have shown here an effective and relatively simple method through which a parameter space can be re-sampled such that all models are equally 'unique' from their neighbours. However, there is no guarantee that this is the best method of approaching this problem. One issue is that enforcing a strict griding along the step vectors does not take into account the fact that in some regions of the parameter space the morphologies of the models may change faster than in other regions. To remedy this, one can re-sample using an 'adaptive' method. The points are initially generated at random positions, and then moved iteratively, based on a repulsive force between points at a distance (computed using a metric) smaller than a target value. This results in a sort of crystalline lattice that adapts the geometry of the parameter space**. This method seems promising, as shown below.

Figure 8 – Distance Histogram After Adaptive Re-sampling



The next step, currently underway, is to use optimally sampled parameter spaces to train neural networks. Computing and optimal sampling also implies deriving the probability density function of the chosen observable in the parameter space. This PDF can then be used as a prior in Bayesian MCMC methods. The expectation is that they will result in faster and more precise networks, better able to extract astrophysical parameters, and prepare us for the coming era of Epoch of Reionization cosmology.

It is also worth noting that these results are in no way limited to simply astrophysics. Any realm of science that involves parameter space exploration and machine learning stands to benefit from this procedure.

Bonus Step – Use Optimized Parameter Spaces to Better Train Killer Robots and Assure World Domination

What have I done...

*It should be noted that these eigenvectors must be re-normalized, as the average of normalized eigenvectors will not, in general, be itself normalized.

**Using this technique, there are no clearly defined 'neighbours', so as a compromise, for a given point, the nearest six points are considered the neighbours.

Results expected to be published July 2018.

References:

Furlanetto et al. (2006), arXiv:astro-ph/0604080

Greig & Mesinger (2015), arXiv:1501.06576

Semelin et al. (2017), arXiv:1707.02073

Schmit & Pritchard (2018), arXiv:1708.00011

APPENDIX D

Other Activities

Courses

IYAS 2015	15h
Parallel Computing	15h
Galaxy Formation	15h
Formation MT180	15h
Natural Space Risks	30h
Total	90h

Conferences & Meetings

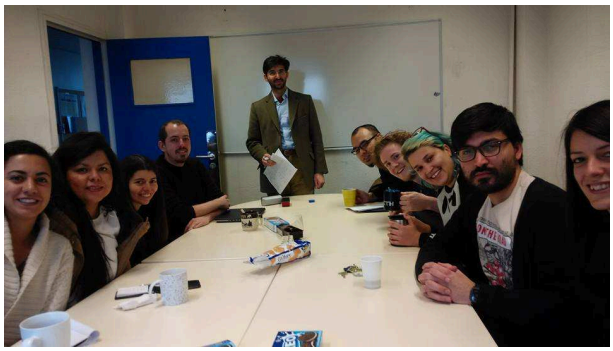
October 2015	Roscoff	ORAGE Meeting 2015
November 2015	Paris	Elbereth 2015
December 2015	Paris	Cosmology and First Light
June 2016	Lyon	SF2A 2016
September 2016	Strasbourg	ORAGE Meeting 2016
November 2016	Goa	Science for the SKA Generation
November 2016	Paris	Elbereth 2016
September 2017	Lyon	ORAGE Meeting 2017
October 2017	Dubrovnik	IAU 333
November 2017	Paris	Elbereth 2017
June 2018	Strasbourg	Rise & Shine

Miscellaneous

Some other things that I've been up to alongside my research:

Showcasing 21 cm lightcones in the 'Apparitions' art show at the École nationale supérieure des beaux arts.

Organizing a weekly journal club for the doctoral students.



Organizing an annual trivia night for the observatory.

Presenting my research at the 'Ma thèse en 180 seconds' competition.

Teaching a college course on Math and Physics (with 72 students).

General Topics in Math & Physics
or *Great Mysteries of the Natural World*
ESME Sudria – Spring 2018

Evan Eames
evaneames8@gmail.com

The goal of this course is to explore some of the big mysteries of physics and mathematics. More than teaching you, I would like you to develop a personal sense of curiosity about the topics presented here.

Grading:
8 tests, each at the beginning of the class → 60 %
1 short paper (2 pages) due on the 2^{nd} to last day of class → 40 %
Youtube videos to be watched on your own → 0%[1]

Schedule[2]

1 Feb 14^{th} - Prime Number Theory
(How to Get Rich Quick)

- Introduction (Infinite Primes, Primality Tests)
- Encryption (Modulo Arithmetic, RSA Encryption, Future Prospects)
- Curiosities (Twin Primes, Golbach's Conjecture, Riemann-Zeta function)

2 Mar 7^{th} - Relativity
(How to Travel in Time)

- Special Relativity (The Equivalence Principle, Mathematics)
- Minkoski Diagrams (Applications)
- General Relativity (Intro)

[1]But please watch them! They are meant to be interesting, and may feature in the tests.
[2]Very tentative!

3 Mar 14^{th} - Particle Physics 1
(How to Create the Universe)

- History
- Atomic Structure (Elements, Ions, Isotopes)
- How to Make an Atomic Bomb
- Neutrinos, Muons, Tauons, Antimatter
- The Quark Model

4 Mar 21^{st} - Particle Physics 2
(How to Maintain the Universe)

- Quantum (Intro)
- Force Carriers
- Feynmann Diagrams
- Current Prospects (Higgs, Proton Decay, SuSy, Neutrino Mass)

5 Mar 28^{th} - Computational Science 1
(How to Win a Nobel Prize)

- LaTeX

6 Apr 11^{th} - Computational Science 2
(How to Create Killer Robots)

- O(n) Complexity Classes
- Computer Science Problems
- P = NP
- Neural Networks & Machine Learning

7 May 2^{nd} - Astrophysics 1
(How to Make Gold)

- Solar System (Definition of a Planet)
- Stars (Birth, Populations, Evolution, Death)
- Astrophysical Tools (Telescopes, Redshift, Spectral Lines, Standard Candles)

8 May 9^{th} - Astrophysics 2
(How to Feel Insignificant)

- Cosmology (The Ancient Universe, Dark Matter, Dark Energy)
- Galaxy Formation (Feedback System, the H-R Diagram, Morphologies)
- Current Prospects (SKA, JWST, ELT, DES)

9 May 16^{th} - Mathematical Philosophy
(How to Create Beauty)

PAPER IS DUE TODAY!

- The Math-Reality Divide
- Imaginary Numbers & Fractals
- Euclids 5^{th} Postulate
- Gödel's Incompleteness Theorem

10 May 23^{rd} - The End of Humanity
(How to Kill Everyone)

- Terrestrial Threats (Supervolcano, Glaciation)
- Space Threats (Asteroids, GRB, Solar Flares)
- Distant Threats (Red Giant, Tidal Locking Heat Death)
- Concluding Remarks

11 June 1^{st} - Optional Class

- Essay & Homework Pick-up
- Open Discussion
- Possibly Wine

References

References

Ali, S. S., Bharadwaj, S., & Chengalur, J. N. 2008, MNRAS, 385, 2166

Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2015, ApJ, 809, 61

Alvarez, M. A., Bromm, V., & Shapiro, P. R. 2006, ApJ, 639, 621

Alvarez, M. A., Pen, U.-L., & Chang, T.-C. 2010, ApJ, 723, L17

Anderson, L., Governato, F., Karcher, M., Quinn, T., & Wadsley, J. 2017, MNRAS, 468, 4077

Ansari, R., Campagne, J. E., Colom, P., et al. 2012, A&A, 540, A129

Baek, S., Di Matteo, P., Semelin, B., Combes, F., & Revaz, Y. 2009, A&A, 495, 389

Baek, S., Semelin, B., Di Matteo, P., Revaz, Y., & Combes, F. 2010, A&A, 523, A4

Barkana, R. & Loeb, A. 2001, Phys. Rep., 349, 125

Barkana, R. & Loeb, A. 2005, ApJ, 624, L65

Barkana, R. & Loeb, A. 2006, MNRAS, 372, L43

Barkana, R. & Loeb, A. 2007, Reports on Progress in Physics, 70, 627

Beardsley, A. P., Hazelton, B. J., Sullivan, I. S., et al. 2016, ApJ, 833, 102

Bebbington, D. H. O. 1986, MNRAS, 218, 577

Becker, R. H., Fan, X., White, R. L., et al. 2001, AJ, 122, 2850

Belitz, P. 2011, PhD thesis, UC San Diego

Belitz, P. & Bewley, T. 2013, Journal of Global Optimization, 56, 61

Berger, M. J. & Colella, P. 1989, Journal of Computational Physics, 82, 64

Berger, M. J. & Oliger, J. 1984, Journal of Computational Physics, 53, 484

Bertschinger, E. 2001, ApJS, 137, 1

Bharadwaj, S. & Ali, S. S. 2004, MNRAS, 352, 142

Bharadwaj, S. & Sethi, S. K. 2001, Journal of Astrophysics and Astronomy, 22, 293

Bobin, J., Moudden, Y., Starck, J.-L., Fadili, J., & Aghanim, N. 2008, Statistical Methodology, 5, 307

Bolgar, F., Eames, E., Hottier, C., & Semelin, B. 2018, MNRAS, 478, 5564

Bonaldi, A., Ricciardi, S., & Brown, M. L. 2014, MNRAS, 444, 1034

Bouwens, R. J., Oesch, P. A., Labbé, I., et al. 2016, ApJ, 830, 67

Bowman, J. D., Rogers, A. E. E., & Hewitt, J. N. 2008, ApJ, 676, 1

Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J., & Mahesh, N. 2018, Nature, 555, 67

Brandenberger, R. H., Danos, R. J., Hernández, O. F., & Holder, G. P. 2010, J. Cosmology Astropart. Phys., 12, 028

Bromm, V., Coppi, P. S., & Larson, R. B. 2002, ApJ, 564, 23

Bromm, V., Yoshida, N., Hernquist, L., & McKee, C. F. 2009, Nature, 459, 49

Bruno, G. 1584, De l'infinito universo et mondi

Bryan, G. L. & Norman, M. L. 1997, in Astronomical Society of the Pacific Conference Series, Vol. 123, Computational Astrophysics; 12th Kingston Meeting on Theoretical Astrophysics, ed. D. A. Clarke & M. J. West, 363

Bryan, G. L., Norman, M. L., O'Shea, B. W., et al. 2014, ApJS, 211, 19

Burns, J. O., Bowman, J. D., Bradley, R. F., et al. 2017, in American Astronomical Society Meeting Abstracts, Vol. 229, American Astronomical Society Meeting Abstracts #229, 306.04

Caprini, C. & Figueroa, D. G. 2018, ArXiv e-prints

Cen, R. 1992, ApJS, 78, 341

Chamaraux, P., Heidmann, J., & Lauqué, R. 1970, A&A, 8, 424

Chang, T.-C., Pen, U.-L., Bandura, K., & Peterson, J. B. 2010, Nature, 466, 463

Chapman, E., Abdalla, F. B., Bobin, J., et al. 2013, MNRAS, 429, 165

Chapman, E., Abdalla, F. B., Harker, G., et al. 2012, MNRAS, 423, 2518

Chapman, E., Bonaldi, A., Harker, G., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 5

Chapman, E., Zaroubi, S., Abdalla, F. B., et al. 2016, MNRAS, 458, 2928

Chen, P., Norman, M. L., Xu, H., & Wise, J. H. 2017, ArXiv e-prints

Christensen, N., Meyer, R., Knox, L., & Luey, B. 2001, Classical and Quantum Gravity, 18, 2677

Ciardi, B. & Madau, P. 2003, ApJ, 596, 1

Cochran, W. D., Hatzes, A. P., & Hancock, T. J. 1991, ApJ, 380, L35

Conselice, C. J., Wilkinson, A., Duncan, K., & Mortlock, A. 2016, ApJ, 830, 83

Cowley, W. I., Baugh, C. M., Cole, S., Frenk, C. S., & Lacey, C. G. 2018, MNRAS, 474, 2352

Crane, P. C. & Napier, P. J. 1989, in Astronomical Society of the Pacific Conference Series, Vol. 6, Synthesis Imaging in Radio Astronomy, ed. R. A. Perley, F. R. Schwab, & A. H. Bridle, 139

de Bruyn, A. G., Wieringa, M. H., Katgert, P., & Sancisi, R. 1988, in IAU Symposium, Vol. 130, Large Scale Structures of the Universe, ed. J. Audouze, M.-C. Pelletan, A. Szalay, Y. B. Zel'dovich, & P. J. E. Peebles, 211

de Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., et al. 2008, MNRAS, 388, 247

DeBoer, D. R., Parsons, A. R., Aguirre, J. E., et al. 2017, PASP, 129, 045001

Dewdney, P. E. 2015, SKA1-LOW CONFIGURATION COORDINATES, Tech. rep., SKA

Di Matteo, T., Ciardi, B., & Miniati, F. 2004, MNRAS, 355, 1053

Di Matteo, T., Perna, R., Abel, T., & Rees, M. J. 2002, ApJ, 564, 576

Dillon, J. S., Kohn, S. A., Parsons, A. R., et al. 2018, MNRAS, 477, 5670

Dillon, J. S., Neben, A. R., Hewitt, J. N., et al. 2015, Phys. Rev. D, 91, 123011

Dubrovich, V. K. 1975, Soviet Astronomy Letters, 1, 196

Dupays, A., Beswick, A., Lepetit, B., Rizzo, C., & Bakalov, D. 2003, Phys. Rev. A, 68, 052503

Eames, E. & Semelin, B. 2018, in IAU Symposium, Vol. 333, IAU Symposium, ed. V. Jelić & T. van der Hulst, 30–33

Eglajs, V. & Audze, P. 1977, Problems of Dynamics and Strengths, 104

Ewall-Wice, A., Dillon, J. S., Hewitt, J. N., et al. 2016, MNRAS, 460, 4320

Ewen, H. I. & Purcell, E. M. 1951, Nature, 168, 356

Fan, X., Carilli, C. L., & Keating, B. 2006a, ARA&A, 44, 415

Fan, X., Strauss, M. A., Richards, G. T., et al. 2006b, AJ, 131, 1203

Fialkov, A., Barkana, R., & Visbal, E. 2014, Nature, 506, 197

Fialkov, A., Cohen, A., Barkana, R., & Silk, J. 2017, MNRAS, 464, 3498

Field, G. B. 1959a, ApJ, 129, 525

Field, G. B. 1959b, ApJ, 129, 536

Finlator, K., Davé, R., & Özel, F. 2011, ApJ, 743, 169

Finlator, K., Keating, L., Oppenheimer, B. D., Davé, R., & Zackrisson, E. 2018, ArXiv e-prints

Follin, B., Knox, L., Millea, M., & Pan, Z. 2015, Phys. Rev. Lett., 115, 091301

Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, Phys. Rep., 433, 181

Galilei, G. 1610, 15

Gamow, G. 1948, Physical Review, 74, 505

Gelfand, A. E. & Smith, A. F. M. 1990, Journal of the American Statistical Association, 85, 398

Geman, S. & Geman, D. 1984, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, 721

Geréb, K., Morganti, R., Oosterloo, T. A., Hoppmann, L., & Staveley-Smith, L. 2015, A&A, 580, A43

Ghosh, A., Bharadwaj, S., Ali, S. S., & Chengalur, J. N. 2011, MNRAS, 418, 2584

Gilks, W. R. & Spiegelhalter, D. J. 1994, Journal of the Royal Statistical Society. Series D (The Statistician), 43, 169

Gillet, N., Mesinger, A., Greig, B., Liu, A., & Ucci, G. 2018, ArXiv e-prints

Gingold, R. A. & Monaghan, J. J. 1977, MNRAS, 181, 375

Giri, S. K., Mellema, G., Dixon, K. L., & Iliev, I. T. 2018, MNRAS, 473, 2949

Gnedin, N. Y. 1995, ApJS, 97, 231

Gnedin, N. Y. 2014, ApJ, 793, 29

Gnedin, N. Y. & Hui, L. 1998, MNRAS, 296, 44

Gnedin, N. Y. & Shaver, P. A. 2004, ApJ, 608, 611

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

Greig, B. & Mesinger, A. 2017a, MNRAS, 465, 4838

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

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

Gu, J., Xu, H., Wang, J., An, T., & Chen, W. 2013, ApJ, 773, 38

Gunn, J. E. & Peterson, B. A. 1965, ApJ, 142, 1633

Gunn, J. E., Siegmund, W. A., Mannery, E. J., et al. 2006, AJ, 131, 2332

Hahn, O. & Abel, T. 2011, MNRAS, 415, 2101

Hardy, E. & Noreau, L. 1987, AJ, 94, 1469

Harker, G., Zaroubi, S., Bernardi, G., et al. 2009, MNRAS, 397, 1138

Harker, G. J. A., Pritchard, J. R., Burns, J. O., & Bowman, J. D. 2012, MNRAS, 419, 1070

Harnois-Déraps, J., Pen, U.-L., Iliev, I. T., et al. 2013, MNRAS, 436, 540

Hastings, W. K. 1970, Biometrika, 57, 97

Hay, S. G. & O'Sullivan, J. D. 2008, Radio Science, 43, RS6S04

Heitmann, K., Higdon, D., White, M., et al. 2009, ApJ, 705, 156

Herschel, W. 1785, Philosophical Transactions of the Royal Society of London Series I, 75, 213

Herzberg, G. 1950, 4, Vol. 1, Molecular Spectroscopy and Molecular Structure, 2nd edn. (New York: Von Nostrand)

Hilborn, R. C. 2002, ArXiv Physics e-prints

Hills, R., Kulkarni, G., Meerburg, P. D., & Puchwein, E. 2018, ArXiv e-prints

Hobbs, G., Heywood, I., Bell, M. E., et al. 2016, MNRAS, 456, 3948

Hogan, C. J. & Rees, M. J. 1979, MNRAS, 188, 791

Holwerda, B. W., Blyth, S.-L., & Baker, A. J. 2012, in IAU Symposium, Vol. 284, The Spectral Energy Distribution of Galaxies - SED 2011, ed. R. J. Tuffs & C. C. Popescu, 496–499

Hotelling, H. 1936, Biometrika, 28, 321

Hubble, E. 1929, Contributions from the Mount Wilson Observatory, vol. 3, pp.23-28, 3, 23

Hunt, L. R., Pisano, D. J., & Edel, S. 2016, AJ, 152, 30

Hurley-Walker, N., Callingham, J. R., Hancock, P. J., et al. 2017, MNRAS, 464, 1146

Huynh, M. & Lazio, J. 2013, ArXiv e-prints

Iliev, I. T., Mellema, G., Ahn, K., et al. 2014, MNRAS, 439, 725

Iliev, I. T., Mellema, G., Pen, U.-L., et al. 2006, MNRAS, 369, 1625

Iliev, I. T., Whalen, D., Mellema, G., et al. 2009, MNRAS, 400, 1283

Iman, R. L., Helton, J. C., & Campbell, J. E. 1981, Journal of Quality Technology, 13, 174

Inoue, A. K., Hasegawa, K., Ishiyama, T., et al. 2018, PASJ, 70, 55

Ishida, E. E. O. & de Souza, R. S. 2011, A&A, 527, A49

Jackson, C. 2005, PASA, 22, 36

Jeffreys, H. 1946, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 186, 453

Jelić, V., Zaroubi, S., Labropoulos, P., et al. 2008, MNRAS, 389, 1319

Kahn, H. & Theodore, H. 1951, Natl. Bur. Stand. Appl. Math, 12, 27

Kaiser, N. 1987, MNRAS, 227, 1

Kakiichi, K., Majumdar, S., Mellema, G., et al. 2017, MNRAS, 471, 1936

Kant, I. 1755, Allgemeine Naturgeschichte und Theorie des Himmels

Kean, S. 2010, The disappearing spoon : and other true tales of madness, love, and the history of the world from the periodic table of the elements (New York: Little, Brown and Co)

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

Kerr, F. J., Hindman, J. F., & Robinson, B. J. 1954, Australian Journal of Physics, 7, 297

Kimm, T., Katz, H., Haehnelt, M., et al. 2017, MNRAS, 466, 4826

Klein, R. I. 1999, Journal of Computational and Applied Mathematics, 109, 123

Kohler, S. 2017, AAS Nova Highlights, 2337

Koopmans, L., Pritchard, J., Mellema, G., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 1

Kopp, J. 2008, International Journal of Modern Physics C, 19, 523

Kovetz, E. D., Viero, M. P., Lidz, A., et al. 2017, ArXiv e-prints

Krumholz, M. R. 2015, in Astrophysics and Space Science Library, Vol. 412, Very Massive Stars in the Local Universe, ed. J. S. Vink, 43

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

Larson, R. L., Finkelstein, S. L., Pirzkal, N., et al. 2018, ApJ, 858, 94

Latham, D. W., Mazeh, T., Stefanik, R. P., Mayor, M., & Burki, G. 1989, Nature, 339, 38

Lewis, A. & Bridle, S. 2002, Phys. Rev. D, 66, 103511

Lidz, A., Zahn, O., McQuinn, M., Zaldarriaga, M., & Hernquist, L. 2008, ApJ, 680, 962

Loeb, A. 2005, in Bulletin of the American Astronomical Society, Vol. 37, American Astronomical Society Meeting Abstracts, 1232

Lucy, L. B. 1977, AJ, 82, 1013

Ly, A., Marsman, M., Verhagen, J., Grasman, R., & Wagenmakers, E.-J. 2017, ArXiv e-prints

Maartens, R., Abdalla, F. B., Jarvis, M., Santos, M. G., & SKA Cosmology SWG, f. t. 2015, ArXiv e-prints

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

Maimbourg, T., Sellitto, M., Semerjian, G., & Zamponi, F. 2018, ArXiv e-prints

Majumdar, S., Pritchard, J. R., Mondal, R., et al. 2018, MNRAS, 476, 4007

Mao, Y., Tegmark, M., McQuinn, M., Zaldarriaga, M., & Zahn, O. 2008, Phys. Rev. D, 78, 023529

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

McKay, M. D., Beckman, R. J., & J., C. W. 1979, Technometrics, 21, 239

McKean, H. P. 1966, Proceedings of the National Academy of Science, 56, 1907

McQuinn, M., Zahn, O., Zaldarriaga, M., Hernquist, L., & Furlanetto, S. R. 2006, ApJ, 653, 815

Meiksin, A. A. 2009, Reviews of Modern Physics, 81, 1405

Mellema, G., Iliev, I. T., Pen, U.-L., & Shapiro, P. R. 2006, MNRAS, 372, 679

Mellema, G., Koopmans, L. V. E., Abdalla, F. A., et al. 2013, Experimental Astronomy, 36, 235

Mertens, F. G., Ghosh, A., & Koopmans, L. V. E. 2018, MNRAS, 478, 3640

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

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087

Metropolis, N. & Ulam, S. 1949, Journal of the American Statistical Association, 44, 335

Mittelmann, H. D. & Vallentin, F. 2009, ArXiv e-prints

Mohr, P. J. & Taylor, B. N. 2000, in American Institute of Physics Conference Series, Vol. 543, Atomic and Molecular Data and their Applications, ICAMDATA, ed. K. A. Berrington & K. L. Bell, 3–16

Monaghan, J. J. 1992, ARA&A, 30, 543

Moore, D. F., Aguirre, J. E., Parsons, A. R., Jacobs, D. C., & Pober, J. C. 2013, ApJ, 769, 154

Morales, M. F. 2005, ApJ, 619, 678

Morales, M. F., Hazelton, B., Sullivan, I., & Beardsley, A. 2012, ApJ, 752, 137

Morales, M. F. & Hewitt, J. 2004, ApJ, 615, 7

Morris, M. D. & Mitchell, T. J. 1995, Journal of Statistical Planning and Inference, 43, 381

Mortlock, D. J., Warren, S. J., Venemans, B. P., et al. 2011, Nature, 474, 616

Muller, C. A. & Oort, J. H. 1951, Nature, 168, 357

Nair, R., Bose, S., & Saini, T. D. 2018, Phys. Rev. D, 98, 023502

Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563

Nebot-Gil, I. 2015, Journal of Chemical Theory and Computation, 11, 472, pMID: 26580907

Newburgh, L. B., Addison, G. E., Amiri, M., et al. 2014, in Proc. SPIE, Vol. 9145, Ground-based and Airborne Telescopes V, 91454V

Nijboer, R. J., Noordam, J. E., & Yatawatta, S. B. 2006, in Astronomical Society of the Pacific Conference Series, Vol. 351, Astronomical Data Analysis Software and Systems XV, ed. C. Gabriel, C. Arviset, D. Ponz, & S. Enrique, 291

Nilsson, K. K. 2007, PhD thesis, Dark Cosmology Centre, Niels Bohr Institute Faculty of Science, University of Copenhagen

Norman, M. L., Reynolds, D. R., So, G. C., Harkness, R. P., & Wise, J. H. 2015, ApJS, 216, 16

Ocvirk, P., Gillet, N., Shapiro, P., et al. 2015, IAU General Assembly, 22, 2255292

Oesch, P. A., Brammer, G., van Dokkum, P. G., et al. 2016, ApJ, 819, 129

Oh, S. P. & Mack, K. J. 2003, MNRAS, 346, 871

Ono, Y., Ouchi, M., Mobasher, B., et al. 2012, ApJ, 744, 83

Paciga, G. 2013, PhD thesis, University of Toronto (Canada)

Paciga, G., Albert, J. G., Bandura, K., et al. 2013, MNRAS, 433, 639

Paciga, G., Chang, T.-C., Gupta, Y., et al. 2011, MNRAS, 413, 1174

Parsons, A. R., Backer, D. C., Foster, G. S., et al. 2010, AJ, 139, 1468

Parsons, A. R., Liu, A., Aguirre, J. E., et al. 2014, ApJ, 788, 106

Patil, A. H., Yatawatta, S., Koopmans, L. V. E., et al. 2017, ApJ, 838, 65

Patil, A. H., Yatawatta, S., Zaroubi, S., et al. 2016, MNRAS, 463, 4317

Pawlik, A. H., Schaye, J., & Dalla Vecchia, C. 2015, MNRAS, 451, 1586

Pearson, K. F. R. S. 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559

Peebles, P. J. E. 1980, The large-scale structure of the universe

Pen, U.-L. 1995, ApJS, 100, 269

Penzias, A. A. & Wilson, R. W. 1965, ApJ, 142, 419

Perlmutter, S., Gabi, S., Goldhaber, G., et al. 1997, ApJ, 483, 565

Peterson, J. B., Aleksan, R., Ansari, R., et al. 2009, in ArXiv Astrophysics e-prints, Vol. 2010, astro2010: The Astronomy and Astrophysics Decadal Survey

Peterson, J. B., Voytek, T. C., Natarajan, A., Jaregui Garcia, J. M., & Lopez-Cruz, O. 2014, ArXiv e-prints

Philip, L., Abdurashidova, Z., Chiang, H. C., et al. 2018, ArXiv e-prints

Planck Collaboration, Adam, R., Ade, P. A. R., et al. 2016a, A&A, 594, A1

Planck Collaboration, Adam, R., Aghanim, N., et al. 2016b, A&A, 596, A108

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2018, ArXiv e-prints

Pober, J. C., Greig, B., & Mesinger, A. 2016, MNRAS, 463, L56

Pober, J. C., Parsons, A. R., Aguirre, J. E., et al. 2013, ApJ, 768, L36

Price, D. C., Greenhill, L. J., Fialkov, A., et al. 2018, MNRAS, 478, 4193

Pritchard, J., Ichiki, K., Mesinger, A., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 12

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

Raut, D., Choudhury, T. R., & Ghara, R. 2018, MNRAS, 475, 438

Repetti, A., Birdi, J., Dabbech, A., & Wiaux, Y. 2017, MNRAS, 470, 3981

Riess, A. G., Casertano, S., Yuan, W., et al. 2018, ArXiv e-prints

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, AJ, 116, 1009

Risken, H. & Frank, T. 1996, The Fokker-Planck Equation/ Methods of Solution and Applications (Berlin, Germany: Springer-Verlag)

Robert, C. & Casella, G. 2011, Statist. Sci., 26, 102

Rosdahl, J., Blaizot, J., Aubert, D., Stranex, T., & Teyssier, R. 2013, MNRAS, 436, 2188

Rosdahl, J., Katz, H., Blaizot, J., et al. 2018, MNRAS, 479, 994

Ross, H. E., Dixon, K. L., Ghara, R., Iliev, I. T., & Mellema, G. 2018, ArXiv e-prints

Ruffert, M. 1992, A&A, 265, 82

Ryu, D., Ostriker, J. P., Kang, H., & Cen, R. 1993, ApJ, 414, 1

Sadler, E. M., Jackson, C. A., Cannon, R. D., et al. 2002, MNRAS, 329, 227

Salmon, B., Coe, D., Bradley, L., et al. 2018, ArXiv e-prints

Santos, M. G., Amblard, A., Pritchard, J., et al. 2008, ApJ, 689, 1

Santos, M. G. & Cooray, A. 2006, Phys. Rev. D, 74, 083517

Schinckel, A. E. T. & Bock, D. C.-J. 2016, in Proc. SPIE, Vol. 9906, Ground-based and Airborne Telescopes VI, 99062A

Schmidt, B. P., Suntzeff, N. B., Phillips, M. M., et al. 1998, ApJ, 507, 46

Schmit, C. J. & Pritchard, J. R. 2018, MNRAS, 475, 1213

Scott, D. & Rees, M. J. 1990, MNRAS, 247, 510

Semelin, B. 2016, MNRAS, 455, 962

Semelin, B. & Combes, F. 2002, A&A, 388, 826

Semelin, B., Combes, F., & Baek, S. 2007, A&A, 474, 365

Semelin, B., Eames, E., Bolgar, F., & Caillat, M. 2017, MNRAS, 472, 4508

Shandarin, S. F. & Zeldovich, Y. B. 1989, Reviews of Modern Physics, 61, 185

Shao, L., Stairs, I., Antoniadis, J., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), 42

Shapley, H. & Curtis, H. D. 1921, Bulletin of the National Research Council, Vol. 2, Part 3, No. 11, p. 171-217, 2, 171

Shimabukuro, H. & Semelin, B. 2017, MNRAS, 468, 3869

Shin, M.-S., Trac, H., & Cen, R. 2008, ApJ, 681, 756

Singh, S., Subrahmanyan, R., Udaya Shankar, N., et al. 2017, ApJ, 845, L12

SKA-Collaboration. 2015, 2, Vol. 1, Advancing Astrophysics with the Square Kilometre Array, 1st edn. (SKA Organisation, Jodrell Bank, Lower Withington, Macclesfield, Cheshire, SK11 9DL, UK: Ddolman Scott Ltd)

Smoot, G. F. 2007, Rev. Mod. Phys., 79, 1349

Smoot, G. F. & Debono, I. 2017, A&A, 597, A136

Sobacchi, E. & Mesinger, A. 2014, MNRAS, 440, 1662

Sokolowski, M., Tremblay, S. E., Wayth, R. B., et al. 2015, PASA, 32, e004

Spinelli, M., Bernardi, G., & Santos, M. G. 2018, MNRAS, 479, 275

Sprenger, T., Archidiacono, M., Brinckmann, T., Clesse, S., & Lesgourgues, J. 2018, ArXiv e-prints

Springel, V. 2005, MNRAS, 364, 1105

Stranex, T. & Teyssier, R. 2010, Master's thesis, University of Zurich, https://www.uzh.ch/cmsssl/physik/dam/jcr:5974b404-4eee-41cb-a3db-7e29101033d0/Stranex_2010.pdf

Stromgren, B. 1939, ApJ, 89, 526

Sunyaev, R. A. & Zeldovich, I. B. 1975, MNRAS, 171, 375

Switzer, E. R., Masui, K. W., Bandura, K., et al. 2013, MNRAS, 434, L46

Tanner, M. A. & Wong, W. H. 1987, Journal of the American Statistical Association, 82, 528

Taylor, G. B., Carilli, C. L., & Perley, R. A., eds. 1999, Astronomical Society of the Pacific Conference Series, Vol. 180, Synthesis Imaging in Radio Astronomy II

Tegmark, M., Silk, J., Rees, M. J., et al. 1997a, ApJ

Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997b, ApJ

Teyssier, R. 2002, A&A, 385, 337

Teyssier, R., Chièze, J.-P., & Alimi, J.-M. 1998, ApJ, 509, 62

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

Tingay, S. J., Oberoi, D., Cairns, I., et al. 2013, in Journal of Physics Conference Series, Vol. 440, Journal of Physics Conference Series, 012033

Trott, C. M., Wayth, R. B., & Tingay, S. J. 2012, ApJ, 757, 101

Turing, A. M. 1950, Mind, LIX, 433

Umeda, H. & Nomoto, K. 2003, Nature, 422, 871

Urban, N. M. & Fricker, T. E. 2010, Computers & Geosciences, 36, 746

van de Hulst, H. C. 1945, 1945, The Origin of Radio Waves from Space, ed. W. T. Sullivan, III, 302

van de Hulst, H. C., Muller, C. A., & Oort, J. H. 1954, Bull. Astron. Inst. Netherlands, 12, 117

van de Hulst, H. C., Raimond, E., & van Woerden, H. 1957, Bull. Astron. Inst. Netherlands, 14, 1

van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2

van Weeren, R. J., Williams, W. L., Tasse, C., et al. 2014, ApJ, 793, 82

Varshalovich, D. A. & Khersonskii, V. K. 1977, Soviet Astronomy Letters, 3, 155

Vonlanthen, P., Semelin, B., Baek, S., & Revaz, Y. 2011, A&A, 532, A97

Voytek, T. C., Natarajan, A., Jáuregui García, J. M., Peterson, J. B., & López-Cruz, O. 2014, ApJ, 782, L9

Watson, W. D. & Deguchi, S. 1984, ApJ, 281, L5

Williams, W. L., Intema, H. T., & Röttgering, H. J. A. 2013, A&A, 549, A55

Wouthuysen, S. A. 1952, AJ, 57, 31

Wright, E. L. 2004, Measuring and Modeling the Universe, 291

Wright, T. 1750, An original theory or new hypothesis of the universe.

Zaldarriaga, M., Furlanetto, S. R., & Hernquist, L. 2004, ApJ, 608, 622

Zarka, P., Girard, J. N., Tagger, M., & Denis, L. 2012, in SF2A-2012: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, ed. S. Boissier, P. de Laverny, N. Nardetto, R. Samadi, D. Valls-Gabaud, & H. Wozniak, 687–694

Zaroubi, S. 2013, in Astrophysics and Space Science Library, Vol. 396, The First Galaxies, ed. T. Wiklind, B. Mobasher, & V. Bromm, 45

Zawada, K., Semelin, B., Vonlanthen, P., Baek, S., & Revaz, Y. 2014, MNRAS, 439, 1615

Zeeman, P. 1897, ApJ, 5, 332

Zeldovich, I. B. & Novikov, I. D. 1983, Relativistic astrophysics. Volume 2 - The structure and evolution of the universe /Revised and enlarged edition (Chicago, IL: University of Chicago Press, 751 p. Translation.)

Zel'dovich, Y. B. 1970, A&A, 5, 84

Zheng, Z., Cen, R., Hy, T., & Miralda-Escude, J. 2010, in Bulletin of the American Astronomical Society, Vol. 42, AAS Meeting Abstracts #215, 305

Zuo, S., Chen, X., Ansari, R., & Lu, Y. 2018, ArXiv e-prints

Résumé

Les simulations, de plus en plus, sont capables de saisir la complexité de l'époque de réioisation, durant laquelle l'hydrogène neutre de l'Universe a été ionisé par les premières sources lumineuses. Des bases de données représentatives de la multitude de signaux possibles seront nécessaires pour contraindre les paramètres des modèles quand des observations 21 cm seront disponibles. À cette fin, et en préparation des observations à venir sur des instruments comme SKA, nous avons développé une base de données de cones de lumières EoR haute-résolution (21ssd.obspm.fr), ainsi qu'une modélisation du bruit thermique. Nous avons également développé un formalisme permettant de quantifier la différence entre les modèles de cette base de données, en utilisant le spectre de puissance et la fonction de distribution des pixels. Nous trouvons que les deux diagnostiques sont sensibles à des paramètres différents des modèles, ce qui signifie que les deux peuvent être utilisés ensemble de manière complémentaire pour extraire l'information maximale. De plus, en utilisant le code 21cmFAST, nous avons développé des stratégies pour échantillonner l'espace des paramètres d'une mannière optimale (plus homogène et isotrope), afin de fournir le meilleur point de départ pour l'extraction des paramètres. Finalement, l'échantillonnage amélioré est utilisé pour entrainer un réseau de neurones. Ce réseau retrouve les paramètres du modèle en se basant sur une observable. Nous observons une amélioration modérée dans la précision de ses prédictions quand nous utilisons l'échantillonnage optimisé lors de son entrainement.

Mots Clés

Cosmologie, simulation, EoR, 21 cm

Abstract

Simulations are increasingly able to capture the intricacies of the Epoch of Reionization, during which the neutral hydrogen in the Universe was ionized by the first luminous sources. Databases encompassing the range of possible signals will be needed to constrain parameter values when 21 cm observations are available. In preparation for upcoming experiments such as the SKA, we have developed a database of high-resolution EoR lightcones (21ssd.obspm.fr), along with realistic thermal noise modelling. We examine frameworks with which we can quantify the difference between entries in this database, specifically with the power spectrum and pixel distribution function. We find that the two diagnostics are sensitive to different parameters, meaning they can be used together to extract maximumal information. Then, using the 21cmFAST code, we explore how to optimally sample a parameter space (so that it is more homogeneous and isotropic), in order to provide the best set-up for parameter extraction. Finally, the improved sampling is used in training a neural network. The neural network uses observables as input data, and attempts to estimate the corresponding parameter values. When the optimal sampling is used as training data, we find that the neural network is able to estimate parameter values with a modest improvement in accuracy.

Keywords

Cosmology, simulation EoR, 21 cm