UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Hydrodynamic Simulations and Tomographic Reconstructions of the Intergalactic Medium

Permalink
https://escholarship.org/uc/item/47q836xd

Author
Stark, Casey William

Publication Date
2015

Peer reviewed|Thesis/dissertation

eScholarship.org
Powered by the California Digital Library
University of California

Hydrodynamic Simulations and Tomographic Reconstructions of the Intergalactic Medium

By

Casey William Stark

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Astrophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:
Professor Peter E. Nugent, Co-chair
Professor Martin J. White, Co-chair
Professor Uros Seljak
Professor Phillip Colella

Spring 2015

Hydrodynamic Simulations and Tomographic Reconstructions of the Intergalactic Medium

Copyright 2015
by
Casey William Stark

Abstract

Hydrodynamic Simulations and Tomographic Reconstructions of the Intergalactic Medium

by

Casey William Stark

Doctor of Philosophy in Astrophysics

University of California, Berkeley

Professor Peter E. Nugent, Co-chair
Professor Martin J. White, Co-chair

The Intergalactic Medium (IGM) is the dominant reservoir of matter in the Universe from which the cosmic web and galaxies form. The structure and physical state of the IGM provides insight into the cosmological model of the Universe, the origin and timeline of the reionization of the Universe, as well as being an essential ingredient in our understanding of galaxy formation and evolution. Our primary handle on this information is a signal known as the Lyman-alpha forest (or Lyα forest) – the collection of absorption features in high-redshift sources due to intervening neutral hydrogen, which scatters H I Lyα photons out of the line of sight. The Lyα forest flux traces density fluctuations at high redshift and at moderate overdensities, making it an excellent tool for mapping large-scale structure and constraining cosmological parameters. Although the computational methodology for simulating the Lyα forest has existed for over a decade, we are just now approaching the scale of computing power required to simultaneously capture large cosmological scales and the scales of the smallest absorption systems. My thesis focuses on using simulations at the edge of modern computing to produce precise predictions of the statistics of the Lyα forest and to better understand the structure of the IGM.

In the first part of my thesis, I review the state of hydrodynamic simulations of the IGM, including pitfalls of the existing under-resolved simulations. Our group developed a new cosmological hydrodynamics code to tackle the computational challenge, and I developed a distributed analysis framework to compute flux statistics from our simulations. I present flux statistics derived from a suite of our large hydrodynamic simulations and demonstrate convergence to the per cent level. I also compare flux statistics derived from simulations using different discretizations and hydrodynamic schemes (Eulerian finite volume vs. smoothed particle hydrodynamics) and discuss differences in their convergence behavior, their overall agreement, and the implications for cosmological constraints.

In the second part of my thesis, I present a tomographic reconstruction method that allows us to make 3D maps of the IGM with Mpc resolution. In order to make reconstructions of large surveys computationally feasible, I developed a new Wiener Filter application with an algorithm specialized to our problem, which significantly reduces the space and time complexity compared to previous implementations. I explore two scientific applications of the maps: finding protoclusters by searching the maps for large, contiguous regions of low flux and finding cosmic voids by searching the maps for regions of high flux. Using a large N-body simulation, I identify and characterize both protoclusters and voids at $z = 2.5$, in the middle of the redshift range being mapped by ongoing surveys. I provide simple methods for identifying protocluster and void candidates in the tomographic flux maps, and then test them on mock surveys and reconstructions. I present forecasts for sample purity and completeness and other scientific applications of these large, high-redshift objects.

Contents

Acknowledgments iv

1 Introduction 1
- 1.1 Cosmological context . . . 1
- 1.2 The intergalactic medium . . . 3
- 1.3 The Lyα forest . . . 6
- 1.4 Flux statistics . . . 11
  - 1.4.1 Mean flux . . . 11
  - 1.4.2 Flux probability density function . . . 12
  - 1.4.3 Flux power spectra . . . 12
- 1.5 Outline . . . 13

2 Numerical convergence of simulated Lyα forest flux statistics 14
- 2.1 Simulations . . . 17
  - 2.1.1 Included physics . . . 20
  - 2.1.2 Optical depth calculations . . . 21
- 2.2 Physical properties of the Lyα forest . . . 23
- 2.3 Resolution study . . . 30
  - 2.3.1 Mean flux . . . 33
  - 2.3.2 Flux PDF . . . 34
  - 2.3.3 Flux 1D power spectrum . . . 35
  - 2.3.4 Flux 3D power spectrum . . . 37
  - 2.3.5 Richardson extrapolation . . . 37
- 2.4 Box size / missing modes . . . 39
  - 2.4.1 n-point flux statistics . . . 41
- 2.5 Splicing . . . 46
- 2.6 Rescaling optical depths . . . 48
- 2.7 Small-scale statistics . . . 53
  - 2.7.1 Line statistics . . . 53
  - 2.7.2 Wavelet statistics . . . 57
- 2.8 Conclusions . . . 58

3 The impact of hydrodynamic methods on Lyα forest flux statistics 62
3.1 Simulations . . . 63
3.1.1 Nyx . . . 63
3.1.2 Gadget . . . 64
3.1.3 Code discussion . . . 64
3.1.4 Simulation suite . . . 66
3.1.5 Comparison data details . . . 68
3.2 Morphological comparison . . . 76
3.3 Density and temperature statistics . . . 80
3.4 Flux statistics . . . 86
3.4.1 Mean flux . . . 86
3.4.2 Rescaling optical depths . . . 88
3.4.3 Flux PDF . . . 89
3.4.4 Power spectra . . . 90
3.4.5 Remaining differences . . . 91
3.5 Large-box results . . . 95
3.6 Conclusions . . . 96

4 Finding protoclusters using Lyα forest tomographic flux maps 98
4.1 Simulations . . . 100
4.1.1 N-body simulation . . . 100
4.1.2 Halo catalogs . . . 100
4.1.3 Lyα flux field . . . 101
4.2 Protoclusters in density and Lyα forest flux . . . 102
4.3 Reconstruction method . . . 109
4.4 Protocluster identification . . . 110
4.5 Mock surveys . . . 114
4.6 Conclusions . . . 120

5 Finding cosmic voids using Lyα forest tomographic flux maps 122
5.1 Simulations and void finding . . . 124
5.1.1 N-body simulations . . . 124
5.1.2 Void finding . . . 124
5.2 Voids at $z = 2.5$ . . . 129
5.3 Finding voids in flux . . . 135
5.3.1 Ideal flux . . . 138
5.3.2 Tomographic flux maps . . . 138
5.4 Discussion . . . 142
5.4.1 Survey prospects . . . 142
5.4.2 High-redshift void cosmology . . . 143
5.5 Conclusions . . . 145

Bibliography **148**

A Notation **156**
A.1 Fourier convention and power spectra 157

B Atomic rates **159**

C Reconstruction derivation and implementation **163**
C.1 Wiener filter 163
C.2 Signal covariance 164
C.3 Numerical implementation and scaling 166
C.4 Error estimation 169
C.5 Alternate smooth map construction 169

Acknowledgments

There is a long list of people I have to thank for their contributions to my dissertation:

- First and foremost, my advisor, Peter Nugent, for all of his support over the past five years. This work would not be possible without the numerous hours spent bouncing ideas around in his office, his guidance and patience when progress stalled, his consistent funding and travel support, and his keen eye for networking and collaboration.
- Martin White for overseeing the scientific goals of our hydrodynamic simulation work, bringing me into the IGM tomography group, and advising me through my last two projects. Anyone who has worked with Martin knows that he is an incredible source of ideas and productivity, not to mention his always amusing, sardonic humor.
- Zarija Lukić, my colleague and cubemate for most of grad school. Zarija has taught me everything I know about cosmological simulations, high-performance computing, and the wisdom of common blocks. He will certainly be missed in my new cube.
- Andreu Font-Ribera, for guiding me through the observational side of the Lyα forest and for his company as my fellow brewmaster.
- Khee-Gan (KG) Lee, for his help and feedback on all of my tomography work. Working on the tomographic map press release with KG was one of the most exciting opportunities I had during grad school.
- Joe Hennawi, for sponsoring my visit to Heidelberg, advising me in the work in Chapter 4, and providing some of the most passionate discussion I've seen about the IGM.
- Anže Slosar, for advising me through the work in Chapter 3 and never letting a telecon be too boring.
- Avery Meiksin, for his encyclopedic knowledge of the IGM and his numerous insightful suggestions for the work in Chapter 2.
- Ann Almgren, for her help understanding the internals of the Nyx code.
- Rollin Thomas and Ted Kisner for their sage advice about scientific computing and for making C^3 a fun, lively office space.

In addition to Peter and Martin, I would also like to thank the rest of my qualifying exam and thesis committee, Uros Seljak and Phil Colella, for their time and feedback.

The work in Chapter 2 was published as Lukić et al. (2015). The work in Chapter 3 is in preparation and will be published as Stark et al. (2015b). The work in Chapter 4 was published as Stark et al. (2014) and also helped the analysis of Lee et al. (2014b). The work in Chapter 5 was published as Stark et al. (2015a).

Almost all of the computing for this thesis was performed on the Hopper Cray XE6 and the Edison Cray XC30 systems at the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research has made use of NASA's Astrophysics Data System and of the astro-ph preprint archive at arXiv.org. This dissertation was typeset using a class file based on the ucastrothesis LaTeX template.

Chapter 1

Introduction

Before diving into the main content, I refer less familiar readers to Appendix A, where I review cosmology conventions and other standard notation. This introduction provides a brief overview of cosmic history and the main focus of my thesis, the Intergalactic Medium, but it is certainly not a complete substitute for other excellent educational resources, e.g. Big History Project.

1.1 Cosmological context

> The Universe is expanding from a hot big bang,
> which synthesized the light elements.
> There was a period of inflation, which lead to a flat Universe today,
> and generated almost scale-invariant, adiabatic Gaussian fluctuations.
> Structure grew from the fluctuations by gravitational instability,
> dominated by (cold) dark matter.
> The expansion of the Universe is accelerating,
> indicating either a breakdown of general relativity
> or that 3/4 of the Universe is dark energy.
>
> *Astro 228, Spring 2011*
> *Martin White*

The epigraph above is a succinct status report of our knowledge of the history of the Universe. In just the past two decades, we have learned an incredible amount about the Universe and answered many of the previous open questions of cosmology. Much of this is due to the signal called the Cosmic Microwave Background (CMB), specifically, from the information encoded in the temperature and polarization anisotropies of the CMB. In the most recent analysis, results from the Planck mission have provided the most precise conclusions from the CMB to date. Planck Collaboration et al. (2015) presents results that, assuming that the Universe is spatially flat and contains cold dark matter and a ‘cosmological constant’ flavor of dark energy (known as the flat ΛCDM model), then the Universe

- is 13.8 billion years old.
- is currently expanding at a rate of $68\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, and the rate is accelerating under the influence of dark energy.
- is made up of 5 per cent normal matter, 27 per cent cold dark matter, and 68 per cent dark energy.

For someone who uses these facts almost every day without hesitation, it is easy to forget how strange this state of affairs is (and amazing *how* we have learned this!).

The history of the Universe can be summed up in terms of the major events from the Big Bang up to today. Much of this is a summary of the timeline described in (Kolb & Turner 1990), with some supplementary information from (Dodelson 2003). The first event that physics has been extrapolated to is the moment of inflation. In the standard picture of slow-roll inflation (Guth 1981; Linde 1982), with an energy scale of 10^{15} GeV, the Universe begins inflation at $t \approx 10^{-34}$ s, increasing the horizon scale by at least e^{60}. Inflation finishes ($t \approx 10^{-32}$ s) and leaves the Universe very smooth and flat on scales larger than our current horizon, famously solving the homogeneity and curvature issues in the hot big bang model. Quantum fluctuations during inflation also create the seeds of large-scale structure. Standard inflationary models predict density perturbations that are Gaussian and nearly scale-invariant. At this time, the Universe is $t \simeq 10^{-32}$ s old ($a \simeq 10^{-28}$) and has a uniform temperature of 10^{15} GeV (10^{28} K). Between energies of 10 MeV and 100 keV, atomic nuclei formed during the era famously known as ‘the first three minutes‘ (Weinberg 1993). This process is also known as Big Bang Nucleosynthesis, and this sets the primordial abundances of hydrogen, helium, lithium, and beryllium. At this time, relativistic components (radition) still dominate the energy density of the Universe. However, since radiation dilutes with the scale factor as $\rho_{\mathrm{r}} \propto a^{-4}$ and matter dilutes as $\rho_{\mathrm{m}} \propto a^{-3}$, at some point the matter density must overtake radiation. This event, known as radiation-matter equality, happens when the Universe is about 100,000 years old. About 400,000 years after the Big Bang ($T \simeq 3000$ K, $z \simeq 1100$), nuclei and electrons combine for the first time, emitting the cosmic microwave background. After photons and baryons have decoupled, baryons begin to fall into the potential wells already seeded by dark matter. Structure continues to evolve and collapse under gravity until the first stars and galaxies somewhere between $z = 30$ and 15 (about 400 million years after the Big Bang) (Loeb & Furlanetto 2013). These galaxies eventually emit sufficient radiation to ionize hydrogen (and singly-ionize helium), making most of the Universe transparent again around $z = 10$. Galaxies continue to form, merge, and evolve under the influence of gravity, eventually reaching the cosmic peak of star-formation activity around $z = 2$ (Madau & Dickinson 2014). At $z \simeq 0.7$, matter dilutes enough to have an energy density equal to that of dark energy, in the second equality event, matter-dark energy equality. Finally, by $z \simeq 0.3$, dark energy reverses the deceleration of the expansion of the Universe and the expansion begins to accelerate.

1.2 The intergalactic medium

The intergalactic medium is loosely defined as ‘everything’ between galaxies and clusters of galaxies. More explicitly, the IGM can be thought of as all regions in the Universe where the density is less than 100 times the cosmic mean ($\rho < 100\,\bar{\rho}$). This exact cutoff is arbitrary, but since galaxy halos form at densities of about $200\,\bar{\rho}$ (the turnaround density in the spherical collapse model is roughly 180; see also White 2001), this order of magnitude is a reasonable choice. The intergalactic medium makes up most of the Universe in terms of mass and volume. In one of our simulations at $z = 2$, for instance, over 99 per cent of the volume has a baryon density $\rho_\mathrm{b} < 10\,\bar{\rho}_\mathrm{b}$. Gas with $\rho_\mathrm{b} < 100\,\bar{\rho}_\mathrm{b}$ also makes up about 80 per cent of the total mass.

At a redshift of about 10, the IGM is reionized (HI and HeI are ionized, but not HeII) by some of the first galaxies in the Universe (Meiksin 2009, and references therein). Before this, the majority of baryons (those not in galaxies) had adiabatically cooled to temperatures as low as 1 K, since $T \propto a^{-2}$ during adiabatic expansion, and there were no sources of heating during the cosmic ‘dark ages’. During reionization, the gas is then heated to approximately 20,000 K (Meiksin 2009; McQuinn et al. 2009). After HI reionization finishes, the IGM is essentially transparent to radiation under the HeII photoionization limit, and the IGM settles to photoionization equilibrium. Soon the IGM is permeated by a uniform metagalactic UV background. Eventually, active galactic nuclei emit enough higher-energy UV photons to finally ionize HeII. In the most common scenario, HeII reionization occurs around $z = 3$, although this is poorly constrained. Recent observations provide evidence that HeII reionization began at much higher redshifts (Worseck et al. 2014). Common UV background prescriptions and their limitations are further discussed in Section 2.1.

The chemical composition of the IGM is close to primordial, thus the dominant radiative processes involve only hydrogen and helium. The combination of photoionization heating and adiabatic cooling due to the expansion of the Universe drives the gas to a tight power-law relation between density and temperature (Katz et al. 1996; Hui & Gnedin 1997). This is typically parameterized as

$$T = T_0 \left(\frac{\rho_\mathrm{b}}{\bar{\rho}_\mathrm{b}}\right)^{\gamma-1} , \tag{1.1}$$

where T_0 is the temperature at mean density and γ is the IGM ‘equation of state’ parameter. This should not be confused with the γ in the common gamma-law equation of state $p = A\rho^\gamma$. Since we only consider atomic gases in this thesis, this γ is always 5/3. For the redshifts we are interested in, $T_0 \approx 10^4$ K and $\gamma \approx 1.6$ (Lee et al. 2015). As explained in Hui & Gnedin (1997), the slope of the power-law steepens in time, asymptotically to $\gamma = 1.6$, rapidly increasing near HI reionization, and more slowly later. This relation is apparent in a density-temperature phase diagram of baryons. In Figure 1.1, I show the volume-weighted density and temperature distribution of gas from one of our cosmological simulations at $z = 2.5$. A significant majority of the points sit on the ρ_b-T line (the colorscale is logarithmic). This phase diagram shows how we typically split the IGM into four regimes based on density and temperature. The low-density, cool gas is known as the diffuse IGM, and the hot gas is the

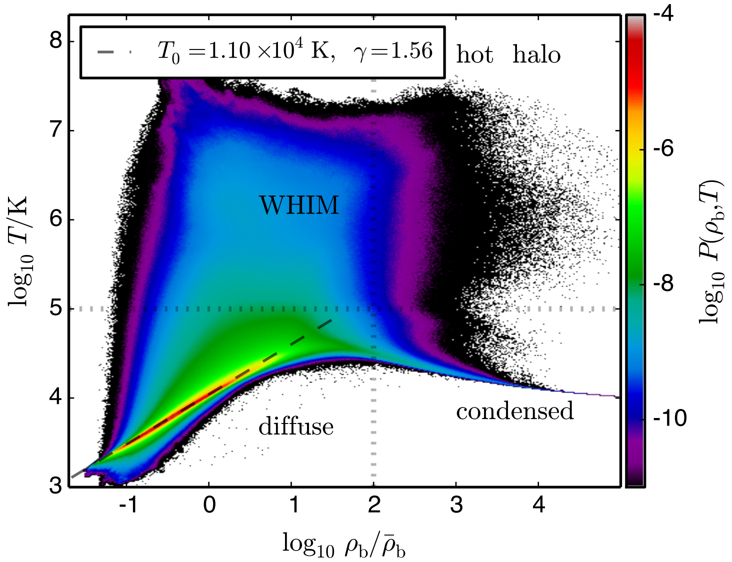

Figure 1.1: The density-temperature distribution of gas (volume-weighted histogram) from a simulation at $z = 2.5$. This phase diagram roughly illustrates the four thermodynamical regimes for baryons.

Warm-Hot Integalactic Medium (WHIM). The high-density gas is broken down into the hot halo phase and the cool, collapsed phase.

The structure of the IGM is largely determined by the Gaussian fluctuations set in place during inflation that are subsequently amplified by gravitational evolution. As soon as photons and baryons decouple at the time of the last scattering, the structure of the IGM begins to fall into place, forming the cosmic web and first galaxies (Bond et al. 1996). Figure 1.2 shows the structure of gas on cosmological scales. For reference, the size of the box is $40\,h^{-1}$Mpc. The different colors used in this figure correspond to the different phases of the gas. The blue structures are the diffuse IGM (in this case, $\rho < 10\,\bar{\rho}$ and $T < 10^6$ K), the green structures are the shock-heated WHIM ($\rho < 10\,\bar{\rho}$ and $T \geq 10^6$ K), and the white structures are the high density hot halo and collapsed phases ($\rho \geq 10\,\bar{\rho}$). In this illustration, it is clear that the diffuse IGM makes up voids and most of the filaments in the cosmic web. These filaments are surrounded by shocked gas, heated to high temperatures when the gas accretes onto the filaments, forming the WHIM. The high-density phases of baryons are sites of galaxy formation, regions where matter has collapsed into virialized structures known as halos. The collapsed phase has a distinct thin tail to higher densities – where galaxies form – while gas in larger halos is heated to high virial temperatures.

While the structure of the IGM is mostly determined by gravitational evolution, there are some differences between the dark matter and baryonic structures on small scales. The standard picture of the structure (see, e.g. Meiksin 2009) is that at moderate overdensities and scales larger than $1\,h^{-1}$Mpc, baryons simply trace dark matter. With a characteristic

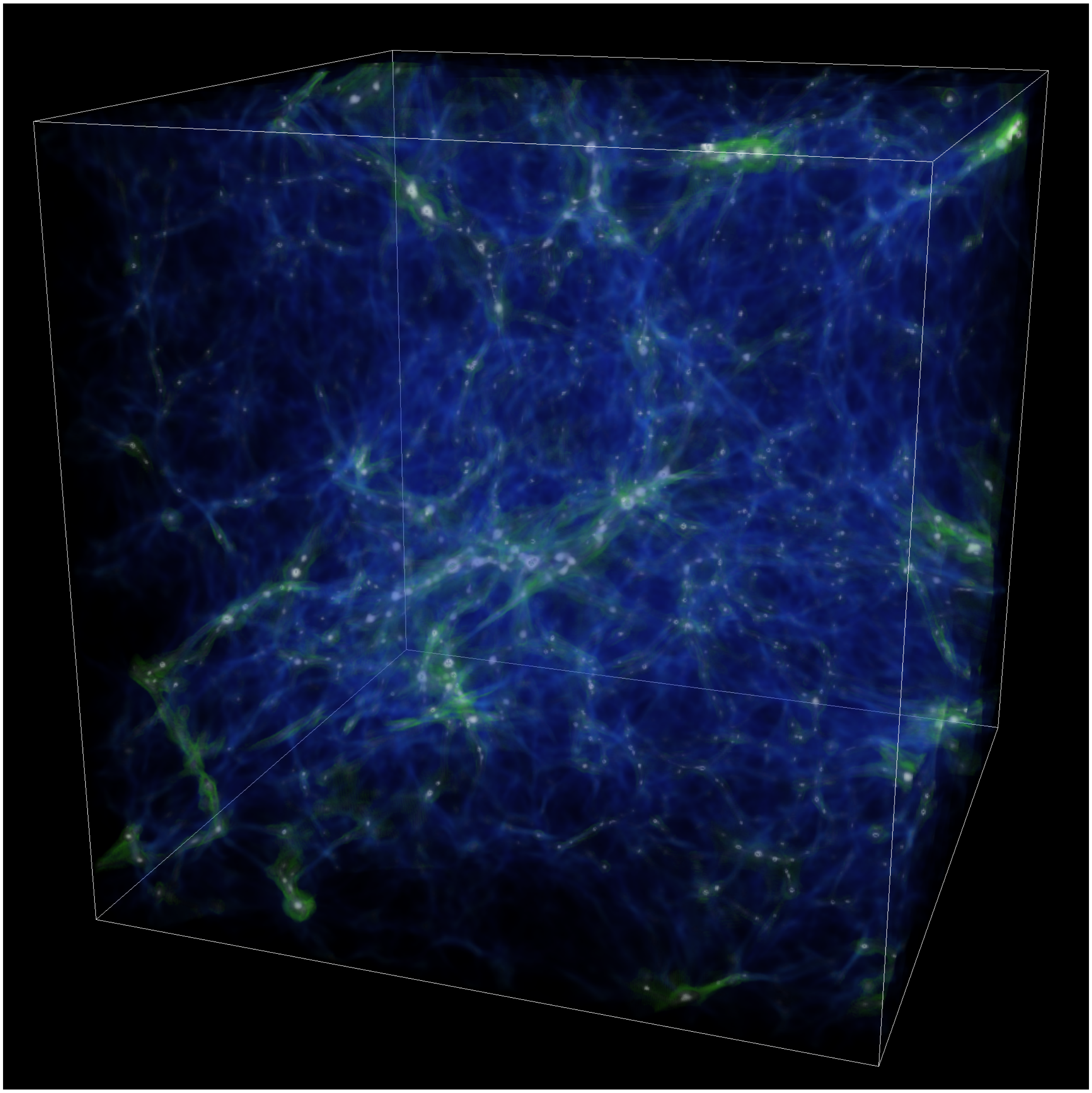

Figure 1.2: A volume rendering of one of our simulations, highlighting the morphology of the phases of the IGM. The diffuse IGM ($\rho/\bar{\rho} < 10$ and $T < 10^6\,\mathrm{K}$) is shown in blue, the WHIM ($\rho/\bar{\rho} < 10$ and $T \geq 10^6\,\mathrm{K}$) is shown in green, and the collapsed and hot halo phases ($\rho/\bar{\rho} \geq 10$) are shown in white.

temperature $T \sim 10^4\,\mathrm{K}$, the baryons in the IGM are pressure supported on scales of $\sim 100\,h^{-1}$kpc and smaller, and the density fluctuations are thus suppressed relative to the dark matter, or as often put, the baryons in the IGM are *filtered* on this scale. Gnedin & Hui (1998) first provided a detailed description of this process in the context of linear theory. The scale at which the gravitational and pressure forces are equal is the Jeans scale

$$\lambda_{\rm J} = (1+z)c_s\sqrt{\frac{\pi}{G\rho}} = 0.783\sqrt{\frac{T/(10^4\,\mathrm{K})}{\Omega_{\rm m}(1+\delta)(1+z)}}\;h^{-1}\mathrm{Mpc}\ . \tag{1.2}$$

In the case of adiabatic expansion (before reionization), $T \propto (1+z)^2$, and the Jeans scale decreases with time. In the case of constant temperature (after reionization), the temperature and Jeans scale increase with time. So at $z = 2$ for instance, the Jeans scale for mean density gas at $T = 10^4\,\mathrm{K}$ is $0.86\,h^{-1}$Mpc.

As defined, the Jeans scale $k_{\rm J} = 2\pi/\lambda_{\rm J}$ is an instantaneous measure that does not take into account the evolution of density or sound speed. Since the amount of filtering at a given epoch also depends on the thermal history of the gas, a more interesting dynamical quantity is the filtering scale $k_{\rm filt}$, the scale at which baryon fluctuations are suppressed relative to cold dark matter. In linear theory, the filtering scale is

$$\frac{1}{k^2_{\rm filt}(t)} = \frac{1}{D_+(t)}\int_0^t dt' a^2(t')\frac{\ddot{D}_+(t') + 2H(t')\dot{D}_+(t')}{k^2_{\rm J}(t')}\int_{t'}^t \frac{dt''}{a^2(t'')} \tag{1.3}$$

The filtering scale in linear theory is always equal to the Jeans scale at an earlier time. This implies that before reionization, the filtering scale is larger than the Jeans scale, and after reionization, the filtering scale is smaller than the Jeans scale. The key point here is that after reionization, in the case of roughly constant temperature, $k_{\rm filt}$ is smaller than the Jeans scale. A rule of thumb is that for typical growth factors and thermal histories, the filtering scale is roughly half the Jeans scale for $2 < z < 4$. In the linear regime, the filtering of baryon power is roughly Gaussian:

$$P_{\rm b}(k) = P_{\rm dm}(k)\exp\left[\frac{-2k^2}{k^2_{\rm filt}}\right]\ . \tag{1.4}$$

Unfortunately, recent studies show evidence that this is of little use since for reasonable thermal histories, as the filtering scale is always too similar to the non-linear scale for this to apply (Lukić et al. 2015; Kulkarni et al. 2015).

1.3 The Lyα forest

Most of what we know about the IGM comes from studying it in absorption. As photons from distant sources propagate through the IGM, there is ample opportunity for the gas and photons to interact, and for the gas to leave a signature in the light that eventually reaches our telescopes. Although most of the Universe is underdense by volume, and the mean

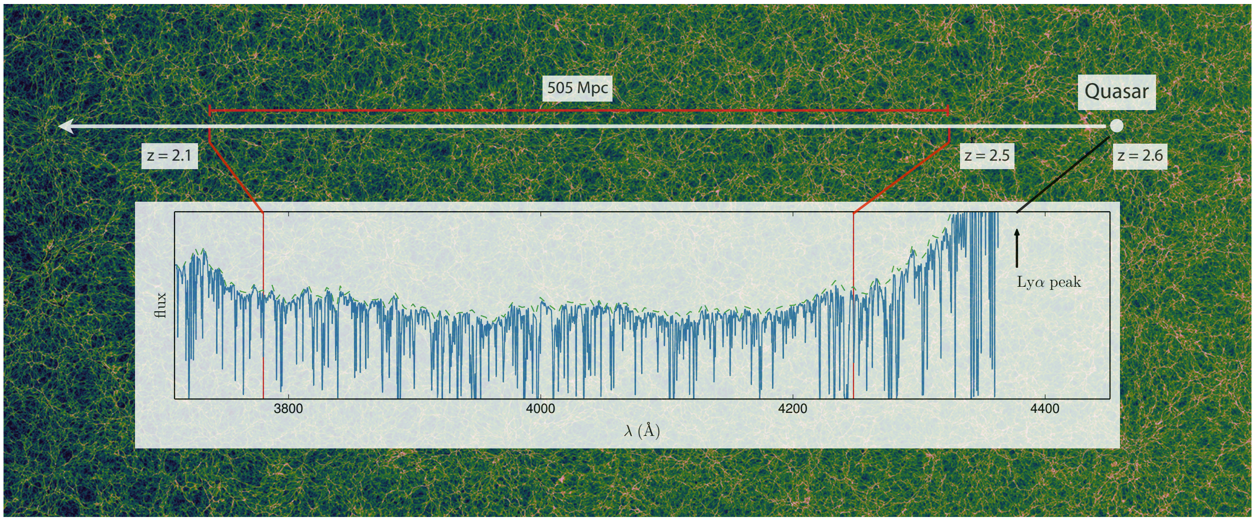

Figure 1.3: An illustration of the Lyα forest in a mock quasar spectrum. The quasar continuum (green dashed line) is taken from Vanden Berk et al. (2001). The background is the HI density (log-scale) from a cosmological simulation. The blue line shows the spectrum after being propagated through the gas shown in the background.

density is incredibly low compared to terrestrial environments, the path lengths covered by high-redshift photons is enormous. For instance, the mean proper density at $z = 2$ is just over $10^{-29}\,\mathrm{g\,cm^{-3}}$ and drops to $5 \times 10^{-31}\,\mathrm{g\,cm^{-3}}$ by $z = 0$, but the distance to $z = 2$ is an enormous 5.2 Gpc, providing an average baryon column of

$$N_\mathrm{b}(z = 2) = \int_0^z n(z')\, dr' = \frac{c\Omega_\mathrm{b}\rho_0}{m_p H_0} \int_0^z \frac{(1+z')^3}{E(z')}\, dz' = 3 \times 10^{22}\,\mathrm{cm^{-2}} \;. \tag{1.5}$$

Such a substantial column makes it easy for low-density gas and relatively weak interactions to scatter a significant amount of light.

Such a phenomenon is readily apparent in the spectra of distant quasars. Quasars are very luminous objects at the center of active galaxies, powered by accretion events onto supermassive black holes (Alexander & Hickox 2012). While very interesting in their own right (Fabian 2012), for the purposes of this thesis, they can be thought of as very bright back lights, with relatively smooth spectra. Between their large, broad HI Lyα and Lyβ emission bumps, quasars have a roughly flat and featureless spectrum which provides a clear continuum from which to measure absorption features. Vanden Berk et al. (2001) compiled a composite quasar spectrum using about 2,200 spectra from the Sloan Digital Sky Survey (SDSS), shown as the green dashed line in Figure 1.3. In this example, I took the composite spectrum, redshifted to $z = 2.6$, and propogated the spectrum through the gas in one of our simulations (detailed later in Section 2.1). The intervening neutral hydrogen creates a multitude of absorption features along the line of sight and these are collectively known as the Lyα forest. Each sightline probes roughly 500 Mpc.

Ikeuchi (1986) and Rees (1986) were the first to suggest the Lyα forest originated from partially ionized hydrogen, in their case confined gravitationally by halos of collisionless or cold dark matter. Using numerical simulations, Cen et al. (1994) demonstrated that Lyα forest systems arise naturally within the framework of theories of structure formation through gravitational instability in CDM dominated cosmologies. The installation of HIRES at Keck (Vogt et al. 1994) made it possible to make precision comparisons with the models, confirming the success of the gravitational instability scenario for the origin of the Lyα forest. Today it is clear that the Lyα forest arises from the smooth, continuous structures in the low-density IGM, rather than the historical collection of dense 'clouds' interpretation.

The connection between large-scale structure and the Lyα forest is clear in a set of relations known as the fluctuating Gunn-Peterson approximation (FGPA). Starting from the baryon density ρ_{b}, we derive an expression for the optical depth to H I Lyα resonant line scattering. As explained above, there is a tight density-temperature relation in the diffuse IGM. The H I density is determined by the overall gas density ρ_{b}, the primordial fractional mass abundance of hydrogen X, and the rates of H II recombination and H I ionization. At typical IGM densities and temperatures, collisional ionization can be ignored, and H I ionization is dominated by photoionization from the UV background. The H II recombination rate is proportional to $n_e n_{\mathrm{H\,II}} T^{-0.7}$, and since the gas is almost entirely ionized, n_e and $n_{\mathrm{H\,II}}$ are proportional to ρ_{b}. Altogether, the H I density $n_{\mathrm{H\,I}}$ is proportional to $\rho_{\mathrm{b}}^2 T^{-0.7} \Gamma_{\mathrm{H\,I}}^{-1}$. Finally, the optical depth is proportional to the H I density (ignoring peculiar velocities and thermal broadening for a moment), and we have:

$$\tau \propto n_{\mathrm{H\,I}} \propto \rho_{\mathrm{b}}^2 T^{-0.7} \Gamma_{\mathrm{H\,I}}^{-1} \propto \rho_{\mathrm{b}}^{2-0.7(\gamma-1)} \Gamma_{\mathrm{H\,I}}^{-1} \; . \tag{1.6}$$

Clearly, this ignores some details, but this is an excellent qualitative guide to the Lyα forest on large scales. Through this relation, we can see how the Lyα forest optical depth traces the underlying density field. One caveat worth mentioning while discussing the FGPA is that some authors use it to describe the entire process of computing the optical depth from N-body simulations, which we will cover in Section 4.1.

Another crucial fact to note is that we do not actually observe the optical depth, but rather, the transmitted flux fraction. Light propagating through the IGM is attenuated such that the transmission (or transmitted flux fraction) is $F = \exp(-\tau)$. It is a bit of a misnomer, but this is typically called the 'flux' in the community. The flux drops to 0 when the absorption is saturated and approaches 1 (the continuum level) when there is almost no absorption. Another quantity of interest is the transmitted flux fraction perturbation $\delta_F = F/\langle F\rangle - 1$, where $\langle F\rangle$ is the mean flux at a given redshift, and this is also sometimes referred to as flux in the Lyα forest community.

In detail, the optical depth to H I Lyα scattering through the IGM is

$$\tau = \int n_{\mathrm{H\,I}} \sigma \, dr = \frac{\pi e^2}{m_e c} f_{12} \int n_{\mathrm{H\,I}}(x) \phi(x) \, a \, dx \; . \tag{1.7}$$

where σ is the resonance line scattering cross section, e is the electron charge, m_e is the electron mass, f_{12} is the Lyα transition oscillator strength, and ϕ is the line profile. In

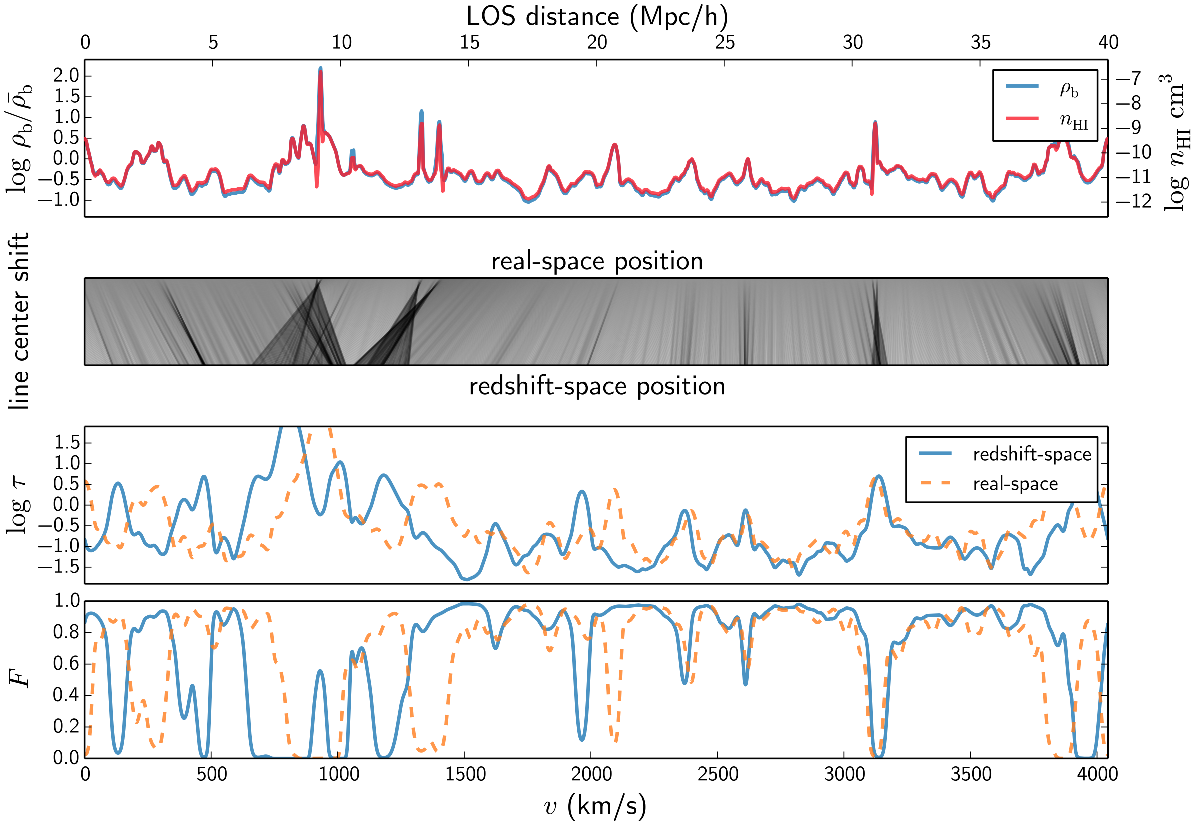

Figure 1.4: A sample skewer, from a 2048^3 simulation in a periodic box of side length $40\,h^{-1}$Mpc at $z = 2.5$, showing the ingredients in the flux calculation. The top horizontal axis gives the (comoving) distance in the line-of-sight direction through the box. The top panel shows the baryon density, on a log scale. The middle panel illustrates how the velocity component along the line of sight $v_{\|}$ shifts the line center going from real- to redshift-space. The lower two panels show the optical depth and flux, respectively. The differences between the real- and redshift-space flux show how the redshift-space distortions do not just shift lines, but also change the blending of the lines.

general, the line profile is a Voigt profile, however, for the optical depths encountered in the forest ($\tau < 100$), the Lorentzian wings are negligible, and we can use the simpler Doppler profile. In this case, the optical depth expression simplifies to

$$\tau(v) = \frac{\pi e^2}{m_e c} f_{12} \int n_{\mathrm{HI}} \frac{1}{\pi^{1/2} b} \exp\left[-\left(\frac{v - v_{\parallel} - v'}{b}\right)^2\right] \frac{dv'}{H} \ , \tag{1.8}$$

where b is the Doppler parameter and is typically equal to the thermal velocity of hydrogen, $v_{\mathrm{th}} = \sqrt{2k_{\mathrm{B}}T/m_{\mathrm{H}}}$, and $v_{\parallel}$ is the peculiar velocity along the line of sight. We use velocity coordinates in this version of the optical depth expression because it is usually most convenient, but it is straightforward to convert to other coordinates using the relations listed in Appendix A. We show an example of the full optical depth calculation from the density and temperature along a line of sight in Figure 1.4.

Observations of the Lyα forest from the ground are limited to roughly $2 \leq z < 5$. The low-redshift cutoff is due to the fact that below $z = 2$, H I Lyα is not redshifted enough out of the UV range blocked by the atmosphere (roughly $\lambda < 3600$ Å). It is possible to observe the Lyα forest at lower redshifts from space (specifically with the Faint Object Spectrograph or Cosmic Origins Spectrograph on the Hubble Space Telescope), although there are only a handful of such observations (Bechtold et al. 2002; Dobrzycki et al. 2002). At high redshifts, observations of the Lyα forest are limited by two facts: quasar targets become increasing rare and faint, and the absorption is often saturated, resulting in a Gunn-Peterson trough rather than a forest (Fan et al. 2006). Broadly speaking, there are two flavors of Lyα forest observations: those with high signal-to-noise and high resolution, meant to resolve individual absorption systems, and those with low signal-to-noise and moderate resolution, meant to probe a greater volume as part of a cosmological survey. The first type of observation is necessary to measure the distributions of absorption systems (column densities and Doppler parameters) and to measure primordial abundances (e.g. Cooke et al. 2014). The second type of observation is a result of the optimal survey strategy for extracting cosmological information from the Lyα forest (McQuinn & White 2011).

The last decade has seen increasing use of Lyα absorption to investigate large-scale structure and cosmology. The Sloan Digital Sky Survey (SDSS) (York et al. 2000) provided an enormous increase in the amount of Lyα forest data with thousands of quasars suitable for 1D analysis, but at the cost of the spectra being low-resolution and fairly noisy. Still, this volume of data allowed a much-improved measurement of the 1D flux power spectrum (McDonald et al. 2006), placing constraints on the large-scale spectral index n_s and the amplitude of fluctuations σ_8. The Baryon Oscillation Spectroscopic Survey (BOSS) experiment of SDSS-III (Dawson et al. 2013) further increased the sky density of suitable quasar lines of sight. The close proximity of large numbers of lines of sight has enabled 3D correlations in the forest to be measured over large scales for the first time using a sample of some 14,000 QSOs (Slosar et al. 2011, 2013; Busca et al. 2013). The 3D flux information has also been cross-correlated with other high redshift tracers (Font-Ribera et al. 2012, 2013). The 1D flux power spectrum has been measured to unprecedented precision (Palanque-Delabrouille et al.

2013). The 3D Lyα absorption correlations are a promising means of constraining the nature of dark energy through the measurements of the angular diameter distance and the Hubble constant at high redshifts by detecting the large-scale Baryon Acoustic Oscillation (BAO) peak (Slosar et al. 2011; Busca et al. 2013; Font-Ribera et al. 2014). At the same time, the measured signal provides a novel test of the gravitational instability origin of the Lyα forest and the large-scale power in the meta-galactic ionizing background (e.g. McQuinn et al. 2011; McQuinn & White 2011). Lee et al. (2014b) demonstrated that it is now possible to map out the IGM on Mpc scales using closely-separated Lyα forest sightlines from star-forming galaxies. The Lyα forest may also be used to constrain galaxy formation models. Galactic winds driven by feedback effects from galaxy formation models can impact statistics of the Lyα forest flux (Viel et al. 2013b) and the circumgalactic medium. Searches for the impact on the circumgalactic medium are underway around Lyman-break galaxies (Crighton et al. 2011; Rudie et al. 2012).

1.4 Flux statistics

As I have described above, the Lyα forest provides an excellent tracer of large-scale structure at moderate overdensities and over $2 \leq z < 5$. Although we are sometimes interested in specific structures found in the Lyα forest (as I explore in Chapter 4 and Chapter 5), we are primarily interested in the statistical information encoded in the forest. In this section, I introduce the first three n-point functions of the Lyα forest and describe what we can learn from them.

1.4.1 Mean flux

The simplest possible flux statistic is the mean transmitted flux fraction $\langle F \rangle$, or equivalently, the effective optical depth $\tau_{\mathrm{eff}} = -\log\langle F \rangle$. This is simply the mean of the transmitted flux fraction of all pixels at a particular redshift. Observations show that the mean flux smoothly evolves from $\langle F \rangle \simeq 0.4$ at $z = 4$, to $\langle F \rangle \simeq 0.9$ at $z = 2$, as expansion gradually lowers the (proper) H I density and the UVB intensity slowly increases (Becker et al. 2013). The mean flux is most sensitive to the metagalactic H I photoionization rate, which in turn constrains average stellar and active galactic nuclei UV radiation (Faucher-Giguère et al. 2008a). In principal, the mean flux also probes the thermal history of the IGM, as $\tau \propto T^{-0.7}$. (Bernardi et al. 2003) reported an increase in the mean flux at $z \sim 3.2$ and interpreted this feature as photoheating from He II reionization. Unfortunately, this feature was not found in later work (e.g. Becker et al. 2013). Measurements of the evolution of the mean flux include Press et al. (1993); Kim et al. (2001); Bernardi et al. (2003); McDonald et al. (2006); Faucher-Giguère et al. (2008b); Becker et al. (2013).

1.4.2 Flux probability density function

The flux probability density function (flux PDF of FPDF) $p(F)$ is simply the distribution of the pixel fluxes. $p(F)$ is normalized in the usual fashion such that the integral of p over the full F range is equal to 1, $\int p(F)\,dF = 1$. In the case of equally-spaced F bins, the $p(F)$ values are thus the appropriately rescaled histogram. The flux PDF is a relatively smooth function, with a shape typically peaked at $F = 0$ and 1, and rising at intermediate fluxes (although the slope will be negative and there will be no $F \simeq 1$ peak at a high enough redshift). In principle, this one-point statistic is a good probe of the thermal history of the IGM and the amplitude of density fluctuations. However, the FPDF is also very sensitive to systematic effects such as the resolution of the spectrograph, determination of the quasar continuum level and/or pixel noise. Recently, the FPDF was measured using a sample of 3,393 high signal-to-noise BOSS quasar spectra (Lee et al. 2015), where they found a good fit to the data with a temperature-density slope of $\gamma = 1.6$ and the results strongly disfavored inverted ρ-T models ($\gamma < 1$). Other measurements of the flux PDF include Jenkins & Ostriker (1991); Rauch et al. (1997); McDonald et al. (2000); Becker et al. (2007).

1.4.3 Flux power spectra

Spatial correlations of the Lyα forest, specifically the δ_F quantity, offer a promising tracer of density fluctuations at moderate overdensities, high redshifts, and small scales. Historically, the two-point function of flux was studied only in 1D, only correlating pixel pairs in the same sightline. Before BOSS observed quasars at a much higher sky density, observed quasar sightlines were so widely separated that it was too inaccurate to correlate across sightlines.

Since the flux field is anisotropic due to redshift-space distortions and thermal broadening, the flux 3D power spectrum is usually defined as

$$P_F(k_\perp, k_\parallel) = V\langle|\hat{\delta}_F(\mathbf{k}_\perp, k_\parallel)|^2\rangle \ . \tag{1.9}$$

The flux 1D power spectrum is the power at a specific line-of-sight mode $k_\parallel$, averaged over all sightlines:

$$P_{F,\mathrm{1D}}(k_\parallel) = L\langle|\hat{\delta}_F(k_\parallel)|^2\rangle \ , \tag{1.10}$$

and is related to the 3D power spectrum as:

$$P_{F,\mathrm{1D}}(k_\parallel) = \frac{1}{2\pi}\int_{k_\parallel}^{\infty} k'_\perp P_F(k'_\perp, k_\parallel) dk'_\perp \ . \tag{1.11}$$

Today, the main strength of the 1D flux power spectrum is that it can probe relatively small scales, down to $\sim 0.1\,h^{-1}$Mpc. Therefore, it is a good test of the nature of dark matter and the mass of neutrinos (Seljak et al. 2006; Viel et al. 2013a). Observational measurements of the flux 1D power spectrum include Croft et al. (1998, 1999); McDonald et al. (2000); Croft et al. (2002); Kim et al. (2004); McDonald et al. (2006); Palanque-Delabrouille et al.

(2013). On large scales, the shape of flux power spectrum is very similar to the matter power spectrum (see e.g. Slosar et al. 2009). As such, the Lyα forest is a tracer of the large-scale structure which can measure the characteristics scale of baryon acoustic oscillations (BAO) and use it as a standard ruler to measure the distances and the expansion rate of the universe. Recently, the first measurements of the cosmological parameters via the location of the BAO peak in the Lyα forest was made with BOSS data (Slosar et al. 2011, 2013; Busca et al. 2013; Delubac et al. 2014).

1.5 Outline

The next two chapters of my dissertation focus on the precision of Lyα forest flux statistics derived from cosmological hydrodynamic simulations. In Chapter 2, I introduce NYX, the code developed by our team at Lawrence Berkeley National Lab to perform large cosmological hydrodynamic simulations. We use a suite of simulations to test how simulation resolution and box size affect flux statistics and set requirements for future simulations. We also discuss the effects of optical depth rescaling and a method called power spectrum splicing. In Chapter 3, I present work on how numerical methods (specific gravity and hydrodynamics schemes) affect flux statistics. Together with collaborators at Brookhaven National Labs, we ran a suite of the same cosmological hydrodynamic simulations in NYX and another code called GADGET. We compare the convergence rates of flux statistics in the two codes and present how well they compare in the converged limit. The last two chapters of my dissertation focus on using the Lyα forest to create 3D maps of the IGM with Mpc-scale resolution. In Chapter 4, I introduce the method we use to tomographically reconstruct the 3D structure of the IGM, and my high-performance implementation of the reconstruction algorithm. I also explore the prospects for using the maps to find protoclusters, the high-redshift progenitors of the most massive structures today. Finally, in Chapter 5, I demonstrate another application of the tomographic flux maps: finding and characterizing cosmological voids.

Chapter 2

Numerical convergence of simulated Lyα forest flux statistics

In principle, it is relatively simple to directly simulate the Lyα forest. The diffuse IGM is responsible for the significant majority of the absorption features, and is only moderately overdense. Numerically, these densities are easier to capture than higher overdensities. The evolution of this gas only depends on gravity and pressure forces, and some simple radiative processes. Since the gas is very close to primordial composition, the background ionizing radiation is relatively spatially uniform, and almost all of the gas (by volume) is optically thin to the radiation, the required simulation machinery is simple indeed. Any cosmological hydrodynamics code modeling the evolution dark matter and gas with uniform radiative heating and cooling should be able to adequately capture the diffuse IGM. Historically, the challenge of simulating the Lyα forest has not been the complexity of the physics or computational techniques, but in the sheer scale of computation required.

The biggest computational challenge for accurately capturing the state of the IGM is the required dynamic range. In order to appropriately model bulk flows, simulations must cover linear scales, on the order of $100\,h^{-1}$Mpc. The bulk flows play an important role in determining the redshift-space distortions (and therefore, the line blending) and the temperature distribution of the IGM via shock heating. At the same time, simulations must resolve the filtering scale, which sets a natural minimum scale. The filtering scale is on the order of $100\,h^{-1}$kpc for the densities of interest and reasonable thermal histories. The required dynamic range ends up being closer to 10^4 than 10^3 however, since adequately resolving a given scale in a simulation means covering it with several resolution elements, so that the simulation element scale must be on the order of $10\,h^{-1}$kpc.

With modern numerical techniques and supercomputers, a dynamic range of 10^4 in a 3D simulation is now practical. Using Lagrangian techniques like Smoothed Particle Hydrodynamics (SPH) or the Eulerian adaptive resolution technique Adaptive Mesh Refinement (AMR), it is straightforward to achieve even larger dynamic ranges. However, these techniques only help simulations focused on resolving small fractions of the total domain volume. The difficulty in simulating the IGM is that it covers almost all of the volume of the do-

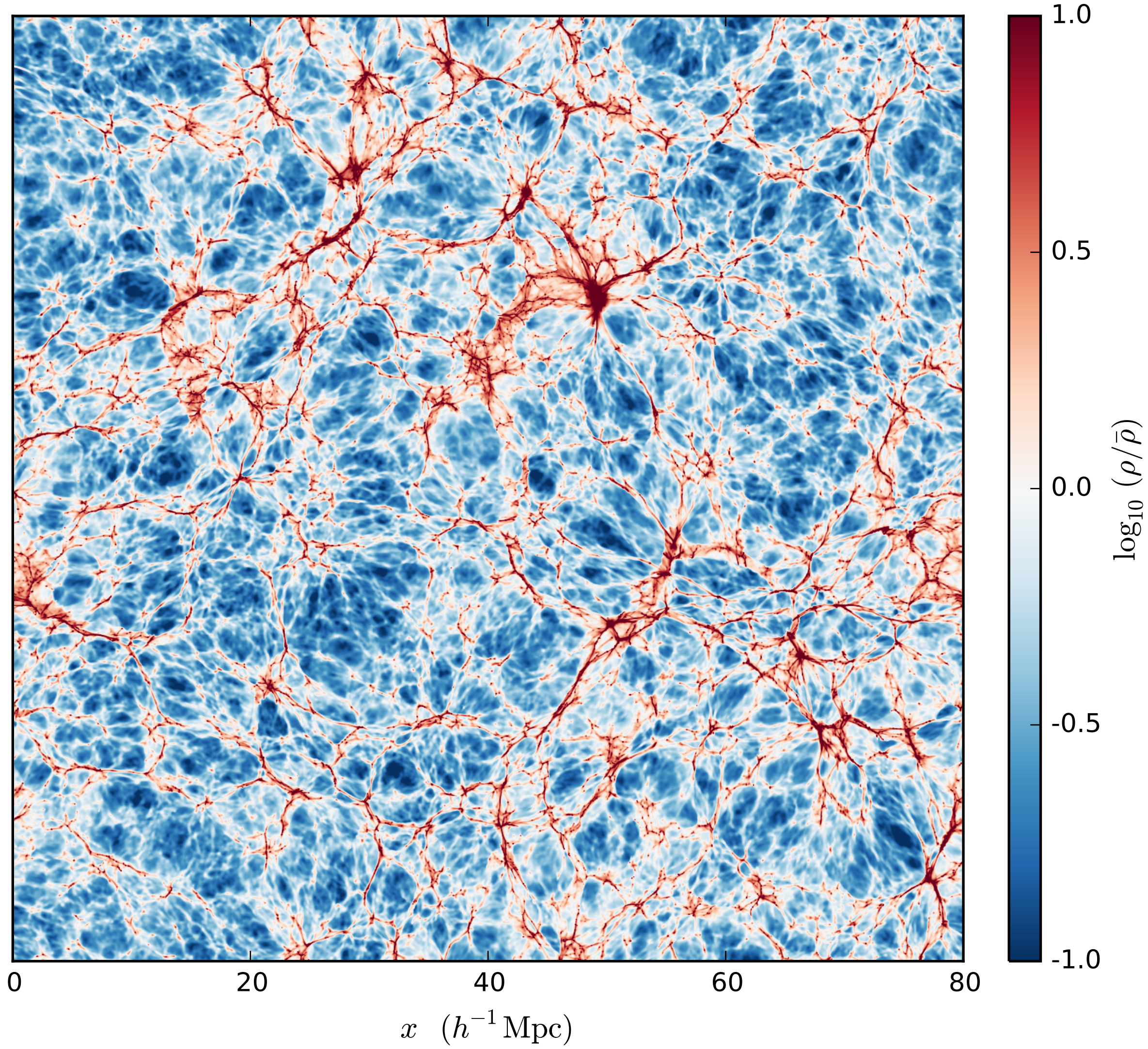

Figure 2.1: A 1 h^{-1}Mpc thick slice of the baryon density from a cosmological simulation at $z = 2.5$. Note how the mean density structures (white) cover a significant fraction of the volume.

main. The gas responsible for the Lyα forest is close to the cosmic mean density rendering Lagrangian methods computationally non-optimal as they spend a majority of the compute cycles evolving dense regions. This is clear in Figure 2.1, where the mean density structures (in white) cover most of the volume. This also gets worse at higher redshifts, e.g. $z \geq 4$, where more of the signal comes from the underdense regions.

It is clear that the Jeans length of the gas ($\sim 500\,h^{-1}$kpc depending on the redshift) must be resolved to recover the correct absorption line widths and small-scale wavelet coefficients, with a suggested resolution of at least $40\,h^{-1}$kpc (comoving) at $z = 2 - 3$ (Bryan et al. 1999; Schaye et al. 2000; Meiksin et al. 2001; Tytler et al. 2009; Lidz et al. 2010). Resolution on this scale is also adequate for converging to better than 5 per cent on the hydrogen ionization rate required to match the measured effective optical depth of the IGM (Meiksin & White 2004; Bolton et al. 2005; Tytler et al. 2009), as well as the effective optical depth itself (Bolton & Becker 2009). At $z > 4$, however, convergence on the effective optical depth diminishes to poorer than 15 per cent at this mean resolution (Bolton & Becker 2009). A resolution of $40\,h^{-1}$kpc is also inadequate for converging on the Doppler parameter and wavelet coefficient distributions at $z = 4$. This is particularly the case for the narrowest features, for which a comoving mean resolution of better than $20\,h^{-1}$kpc appears necessary, with results still not clearly well converged (Bryan et al. 1999; Lidz et al. 2010).

The results are also sensitive to box size. The inferred mean ionizing background is converged to a few percent for comoving box sizes of $30\,h^{-1}$Mpc for $z > 2$ (Meiksin & White 2004; Bolton et al. 2005; Tytler et al. 2009). The line widths increase with box size, possibly not converged to better than 5 per cent at $z = 2$ for a comoving box size as large as $54\,h^{-1}$Mpc (Tytler et al. 2009), although the distribution of smoothed wavelet coefficients appears well converged at this redshift for the smaller box size of $25\,h^{-1}$Mpc (Lidz et al. 2010). At $z > 3$, the wavelet coefficients are not well converged for box sizes as large as $50\,h^{-1}$Mpc (Lidz et al. 2010).

Convergence requirements on the 1D flux power spectrum are also demanding. McDonald et al. (2005b) found better than 5 per cent convergence from the fundamental mode up to $k < 0.025\,\mathrm{km}^{-1}$s for $2 < z < 4$, and up to $k < 0.1\,\mathrm{km}^{-1}$s at $z < 3$, for a resolution of $39\,h^{-1}$kpc, but in a comoving box size of only $5\,h^{-1}$Mpc. In larger boxes ($30\,h^{-1}$Mpc), Viel et al. (2006) found 5 per cent convergence at $k < 0.01\,\mathrm{km}^{-1}$s for a mean resolution of $150\,h^{-1}$kpc, but of only 12 per cent for $k = 0.02\,\mathrm{km}^{-1}$s at $z = 4$. Other work found convergence of up to 10 per cent may be achieved at $k < 0.03\,\mathrm{km}^{-1}$s in $20 - 40\,h^{-1}$Mpc boxes at $z = 2 - 5$ (although possibly as poor as 20 per cent at $z = 5$), with resolutions of 60 – $200\,h^{-1}$kpc, although requiring better than $50\,h^{-1}$kpc resolution for 5 per cent convergence at $k = 0.1\,\mathrm{km}^{-1}$s at $z = 2$, with even this inadequate at $z = 5$ (Meiksin & White 2004; Bolton & Becker 2009; Tytler et al. 2009). Even at this level, the spatial flux correlation function converges to better than 10 per cent over only 3 per cent of the box size (Meiksin & White 2004). The convergence of absorber pair and higher multiple statistics along neighboring lines of sight are not expected to fare better, which is perhaps why they have been largely ignored in simulation comparisons with data.

HI (and HeI) reionization takes place at high redshift, $z > 10$. Therefore the details of this epoch are unimportant for the thermodynamical properties of the gas at redshifts relevant for Lyα forest observations ($z < 4$). In contrast, HeII reionization takes place at an observationally relevant epoch ($3 < z < 5$) although the observational picture is not yet resolved (e.g. Worseck et al. 2014). In addition, the size of fluctuations in the HeII ionizing background are very poorly constrained, varying by an order of magnitude in recent studies (Shull et al. 2010; Syphers & Shull 2014; McQuinn & Worseck 2014). However, the main effect of HeII reionization on the IGM is that the additional photoheating increases the temperature of the IGM. The UV background prescriptions we employ in this study model this increase in the temperature via an increase in photoheating rates and ionize HeII by $z = 3$. We note, however, that HeII reionization could result in higher temperatures with explicit radiative transfer and significant (spatial) fluctuations of the ionizing background. Thus, including HeII reionization correctly in simulations requires incorporating radiative transfer (Tittley & Meiksin 2007), which remains an active area of current research in cosmological hydrodynamics codes (Tittley & Meiksin 2007; McQuinn et al. 2011; Meiksin & Tittley 2012; Compostella et al. 2013).

Modern hydrodynamic simulations recover many of the measured statistical properties of the Lyα forest, such as the HI column density distribution, the pixel flux distribution function and the flux power spectrum at a level capable of distinguishing between plausible variants of the CDM model, with the ΛCDM model the most successful (Zhang et al. 1995; Hernquist et al. 1996; Rauch et al. 1997; Zhang et al. 1997; Croft et al. 1998; McDonald et al. 2000; Meiksin et al. 2001; Croft et al. 2002; Viel et al. 2004). However, some differences are found. Most notable is the distribution of the absorption line widths (Doppler widths) characterized by the b-parameter (typically about 30 km/s). While the line widths are consistent with the amount of broadening characteristic of photoionized gas, the measured distributions show too many broadened lines compared with the predictions of the original simulations. This is likely an indication that HeII was reionized late, at $z \lesssim 4$ (Bryan & Machacek 2000; Ricotti et al. 2000; Schaye et al. 2000; Meiksin et al. 2001; Worseck et al. 2014). Allowing for a late HeII reionization and including radiative transfer during reionization a range of Doppler widths may be achieved consistent with the data (Tittley & Meiksin 2007; Meiksin & Tittley 2012). Since the line widths control the scale and number of features, the distribution of wavelet coefficients is also strongly affected and, to a lesser extent, the column density distribution and the pixel flux distribution (Meiksin et al. 2001). The flux power spectrum is most affected at high wavenumbers.

2.1 Simulations

The simulations we present here are performed with the NYX code (Almgren et al. 2013). NYX follows the evolution of dark matter simulated as self-gravitating Lagrangian particles, and baryons modeled as an ideal gas on a uniform Cartesian grid. NYX includes Adaptive Mesh Refinement (AMR) capabilities, which we can use to extend the simulated dynamic range. We do not make use of AMR in the current work, as the Lyα forest signal spans

Table 2.1: Conversion factors versus redshift

z	λ_α	$H(z)$	$d\lambda/d\chi$	$dv/d\chi$	$10^3\,dz/d\chi$
2.00	3645	285	1.16	95	0.95
2.25	3949	319	1.29	98	1.06
2.50	4253	354	1.43	101	1.18
3.00	4860	428	1.73	107	1.42
3.50	5468	508	2.06	113	1.69
4.00	6075	592	2.40	118	1.98

Conversion factors for the flat ΛCDM cosmology considered here with $h = 0.702$ and $\Omega_{\rm m} = 0.275$. Wavelengths are given in Å, comoving distances in h^{-1}Mpc and velocities in $kmsinv$.

nearly the entire simulation domain rather than isolated concentrations of matter where AMR is more effective. The Eulerian gas dynamics equations are solved using a second-order accurate piecewise parabolic method (PPM) to accurately capture shock waves. Our implementation uses a dimensionally unsplit scheme with full corner coupling (Colella 1990) to better reconstruct the 3D fluid flow. The same mesh structure that is used to update fluid quantities is also used to compute the gravitational field and to evolve the particles via a particle-mesh (PM) method, using Cloud-In-Cell (CIC) interpolation to switch between particle- and mesh-based quantities. The gravitational source terms in the momentum and energy equations are discretized in time using a predictor-corrector approach. The additional physics of radiative heating and cooling is included via source terms in the equations for internal and total energy. As the relevant time scale for heating and cooling can be significantly different from the stability criterion required by the explicit discretization of gas dynamics equations (the Courant-Friedrichs-Lewy or CFL condition), the heating and cooling source terms are integrated in time using VODE (Brown et al. 1989) and coupled to the hydrodynamics using a Strang splitting (Strang 1968) approach. For more details of our numerical methods, see Almgren et al. (2013).

We simulate the *WMAP* 7-yr data constrained flat ΛCDMcosmology, with parameters: $\Omega_{\rm m} = 0.275$, $\Omega_\Lambda = 1 - \Omega_{\rm m} = 0.725$, $\Omega_{\rm b} = 0.046$, $h = 0.702$, $\sigma_8 = 0.816$, and $n_s = 0.96$ (Komatsu et al. 2011). We provide Table 2.1 to help convert between scale, wavelength, and velocity coordinates at redshifts used in this chapter. The latest *Planck* constraints (Planck Collaboration et al. 2015) differ somewhat from WMAP-7 values, most notably in the values for the Hubble constant h and total matter content $\Omega_{\rm m}$. These differences will not play an important role in this work, as we aim to explore numerical prescriptions for achieving 1 per cent accurate Lyα forest statistics. The conclusions here will inform future work for running many *viable* cosmologies and understanding their numerical limitations.

The full set of simulations is listed in Table 2.2. We designed the set of simulations to cover the expected maximum box size and minimum resolution needed to show convergence. All simulations are initialized at $z = 159$, starting from a grid distribution of particles and

Table 2.2: List of simulations

Name	Box size (h^{-1}Mpc)	Elements	Resolution (h^{-1}kpc)	$m_{\rm dm}(\rm M_\odot)$
L10_N128	10	128^3	78	4.3×10^7
L10_N256	10	256^3	39	5.4×10^6
L10_N512	10	512^3	20	6.7×10^5
L10_N1024	10	1024^3	10	8.4×10^4
L20_N256	20	256^3	78	4.3×10^7
L20_N512	20	512^3	39	5.4×10^6
L20_N1024	20	1024^3	20	6.7×10^5
L20_N2048	20	2048^3	10	8.4×10^4
L40_N512	40	512^3	78	4.3×10^7
L40_N1024	40	1024^3	39	5.4×10^6
L40_N2048	40	2048^3	20	6.7×10^5
L80_N1024	80	1024^3	78	4.3×10^7
L80_N2048	80	2048^3	39	5.4×10^6
L80_N4096	80	4096^3	20	6.7×10^5

The simulations used in this chapter. Resolution refers to the cell size, and to ease comparison with SPH simulations we list the mass of dark matter particles in each simulation. See text for details.

Zel'dovich approximation (Zel'dovich 1970). Transfer functions were generated using both analytical approximation Eisenstein & Hu (1999) and Boltzmann code CLASS (Blas et al. 2011). The conclusions presented here have no sensitivity on the particular transfer function used, but it is of course important to maintain the same transfer function accross a series of runs one is comparing to each other. We focus on snapshots in the range $2 \leq z \leq 4$, relevant for most observations. To simplify the comparison, simulations performed in the same box size share the same large-scale modes, the only difference being that higher resolution runs have more modes sampled on small scales.

2.1.1 Included physics

Besides solving for gravity and the Euler equations, we model the chemistry of the gas as having a primordial composition with hydrogen and helium mass abundances of $X = 0.75$, and $Y = 0.25$, respectively. The choice of values is in agreement with the recent CMB observations and Big Bang nucleosynthesis (Coc et al. 2013). The resulting reaction network includes 6 atomic species: H I, H II, He I, He II, He III and e^-, which we evolve under the assumption of ionization equilibrium. The resulting system of algebraic equations is:

$$\begin{aligned}
&\left(\Gamma_{e,\mathrm{HI}} n_e + \Gamma_{\gamma,\mathrm{HI}}\right) n_{\mathrm{HI}} = \alpha_{\mathrm{r,HII}} n_e n_{\mathrm{HII}} \\
&\left(\Gamma_{e,\mathrm{HeI}} n_e + \Gamma_{\gamma,\mathrm{HeI}}\right) n_{\mathrm{HeI}} = \left(\alpha_{\mathrm{r,HeII}} + \alpha_{\mathrm{d,HeII}}\right) n_e n_{\mathrm{HeII}} \\
&\left[\Gamma_{\gamma,\mathrm{HeII}} + \left(\Gamma_{e,\mathrm{HeII}} + \alpha_{\mathrm{r,HeII}} + \alpha_{\mathrm{d,HeII}}\right) n_e\right] n_{\mathrm{HeII}} \\
&\qquad = \alpha_{\mathrm{r,HeIII}} n_e n_{\mathrm{HeIII}} + \left(\Gamma_{e,\mathrm{HeI}} n_e + \Gamma_{\gamma,\mathrm{HeI}}\right) n_{\mathrm{HeI}}
\end{aligned} \tag{2.1}$$

in addition, there are three closure equations for the conservation of charge and hydrogen and helium abundances. Radiative recombination ($\alpha_{\mathrm{r,X}}$), dielectronic recombination ($\alpha_{\mathrm{d,X}}$), and collisional ionization ($\Gamma_{e,\mathrm{X}}$) rates are strongly dependent on the temperature, which itself depends on the ionization state through the mean mass per particle μ

$$T = \frac{2}{3}\frac{m_p}{k_{\mathrm{B}}}\mu\, e_{\mathrm{int}} \tag{2.2}$$

where m_p is the mass of a proton, k_{B} is the Boltzmann constant, and e_{int} is the internal thermal energy per mass of the gas. Here we assume adiabatic index for monoatomic ideal gas. For a gas composed of only hydrogen and helium, μ is related to the number density of free electrons relative to hydrogen by $\mu = 1/\left[1 - (3/4)Y + (1-Y)n_e/n_{\mathrm{H}}\right]$. We iteratively solve the reaction network equations together with the ideal gas equation of state, $p = 2/3\rho e_{\mathrm{int}}$, to determine the temperature and equilibrium distribution of species.

We compute radiative cooling as in Katz et al. (1996), and assume a spatially uniform, but time-varying ultraviolet background (UVB) radiation field from either Faucher-Giguère et al. (2009) or Haardt & Madau (2012). We do not follow radiation transport through the box, nor do we explicitly account for the effects of thermal feedback of stars, quasars, or active galactic nuclei; all cells are assumed to be optically thin, and radiative feedback is accounted for via the UVB model. In addition, we include inverse Compton cooling off the microwave background. For the exact rates used in the NYX code and comparison of two UV backgrounds, we refer the reader to Appendix B.

2.1.2 Optical depth calculations

As described in Section 1.3, the optical depth τ for HI Lyα resonant scattering is

$$\tau_\nu = \frac{\pi e^2}{m_e c} f_{12} \int \frac{n_{\mathrm{HI}}}{\Delta\nu_{\mathrm{D}}} \frac{\exp\left[-\left(\frac{\nu-\nu_0}{\Delta\nu_{\mathrm{D}}}\right)^2\right]}{\sqrt{\pi}}\, dr \tag{2.3}$$

assuming a Doppler profile. We assume that there are no extra kinematic components to the broadening in this work, so $b = \sqrt{2k_{\mathrm{B}}T/m_{\mathrm{H}}}$. In general, the line profile for this process is given by the Voigt profile:

$$\phi_{\nu,\mathrm{V}} = \frac{1}{\pi^{1/2}} H(a,x) = \frac{a}{\pi^{3/2}} \int_{-\infty}^{\infty} \frac{\exp(-y^2)}{(x-y)^2 + a^2} dy \tag{2.4}$$

where $a = \Gamma_{ul}/4\pi\Delta\nu_{\mathrm{D}}$ is the ratio of the damping width to the Doppler width, and $x = (\nu - \nu_0)/\Delta\nu_{\mathrm{D}}$ is the shift from line center. However, for densities and temperatures typical in the IGM we may use the Doppler profile instead, which is just the Gaussian core of the Voigt profile:

$$\phi_{\nu,\mathrm{D}} = \frac{1}{\pi^{1/2}} \exp(-x^2)\ . \tag{2.5}$$

The difference in optical depth computed with a Voigt versus a Doppler profile is very small in the regime we are interested in ($\tau_0 < 100$). Figure 2.2 shows the optical depth to a single absorber with uniform HI density and temperature, computed three different ways. The system spans a comoving scale of roughly 300 h^{-1}kpc, and corresponds to $\Delta v = 30\,\mathrm{km\,s^{-1}}$ for our cosmology's $H(z=2)$. The difference between the left and right panels is the column density, where a system with a typical Lyα forest column density is shown on the left and a weak Lyman Limit System is on the right. The Voigt integral and Doppler integral versions are the full sightline integral, and only differ by the line profile assumed. A third computation approximates the feature as a single line with line center at the center of the absorber ('Voigt line'). The Voigt line version follows the damping wings, but has the wrong shape near line center. It does not account for the change in the line center of the gas across the system and is therefore too narrow. The Doppler integral version correctly traces the Gaussian shape near line center, but misses the damping wings. However, for the low column density lines that make up the Lyα forest, the damping wings add optical depth at the level of $\sim 10^{-3}$. After the transformation to F, such a small τ is far from detectable. For column densities of Lyman-Limit Systems though, the damping wings contribute $\tau \sim 0.1$, which clearly shows up in F. In Lyα forest observations, contamination from LLSs and DLAs is masked out or taken into account in error estimates, so we actually want to avoid

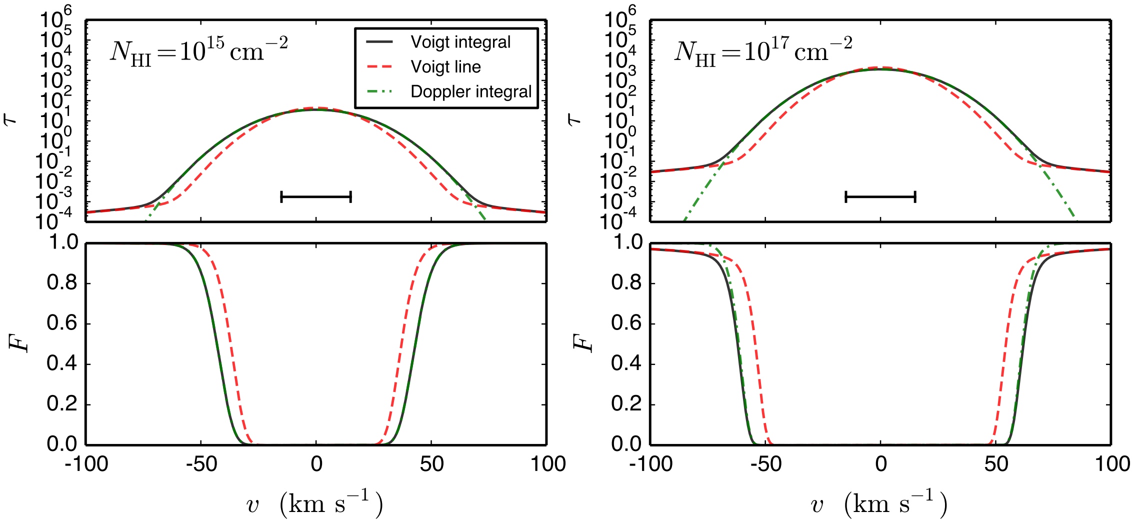

Figure 2.2: A comparison of the optical depth across a uniform absorber computed with three methods, two using the full sightline integral with either a Voigt profile ('Voigt integral') or a Doppler profile ('Doppler integral'), and one approximating the absorber as a static system ('Voigt line'). See the text for more details.

modeling their contamination here. Additionally, our simulations do not properly model the HI density and temperature in the high density regions that give rise to LLSs and DLAs since they do not include radiative transfer. Because the difference in the resulting optical depth is very small, and the Doppler profile is simpler and faster to compute, we use the Doppler line profile in this work.

In order to produce statistics at a single redshift, we also compute the optical depth at a fixed redshift. That is, we do not account for the speed of light when we cast rays in the simulation; we use the gas state at a single cosmic time. The simulated spectra are not meant to look like full Lyα forest spectra, but just recover the statistics of the flux in a small redshift window. The path length in the sightline integration is then $dr = a\,dx = dv/H$, where r is the proper distance, x is the comoving distance, v is the Hubble flow velocity, and H is the Hubble expansion rate at that redshift. In velocity coordinates, the optical depth is

$$\tau_v = \frac{\pi e^2 f_{lu} \lambda_0}{m_e c H} \int n_{\rm X} \frac{1}{\pi^{1/2} b} \exp\left[-\left(\frac{v - v_0}{b}\right)^2\right] dv. \tag{2.6}$$

Although the gas data is fixed at the grid resolution, we can choose an arbitrary spectral resolution $N_{\rm pix}$ along the LOS. We also take the gas values as constant across each cell. With i as the cell index, and j as the pixel index, the discretized version of the optical depth is

$$\tau_j = \frac{\pi e^2 f_{lu} \lambda_0}{m_e c H} \sum_i n_{{\rm X},i} \left[{\rm erf}(y_{i-1/2}) - {\rm erf}(y_{i+1/2})\right] \tag{2.7}$$

where $y = (v_j - v_{\parallel,i} - v)/b$ is the line center shift from the pixel velocity in terms of the broadening scale, $v_{\parallel,i}$ is the component of the gas peculiar velocity parallel to the sightline, and v is the Hubble velocity. The velocity coordinates are also periodic on the domain scale $[0, \dot{a}L)$. It is fortunate the optical depth integration reduces to an analytic expression, as this makes the calculation more robust and straightforward. Previous studies have used the midpoint expression for this integral, but we found that this created too large an error when the sampling scale Δv was $\sim 2\,\mathrm{km\,s^{-1}}$ or larger, whereas the analytic version explicitly conserves the optical depth for any Δv.

We have the choice to evaluate the spectra at an arbitrary resolution. Given a vector of the simulated values $s_i = \left(n_{\mathrm{X,i}}, v_{\parallel,i}, b_i\right)$, at position v_i along a skewer, we can evaluate the optical depth at any v_j. The resolution requirement here is essentially set by the line widths, so that we capture all fluctuations that should be present in the spectra. Given that the most narrow lines have $b \sim 5\,\mathrm{km\,s^{-1}}$, the required spectral sampling should be similar.

We tested that we have adequate spectral resolution by taking the L10_N512 snapshot at $z = 2.5$, recomputing the optical depths at varying spectral resolution, and checked the effect on various flux statistics. We computed the optical depth along the same sightlines (the default 512×512), but with spectral resolutions of 128, 256, 1024, and 2048, spanning a factor of 4 worse to 4 better than the default of 512. We do not plot these results as there is little to no difference in all cases. We see essentially no change in the mean flux (less than 10^{-6}) across all of these resolutions. The difference in the flux PDF is noticeable, but still very small and does not impact our results. The RMS difference between the lowest resolution result and highest resolution result is less than 10^{-3}. For the 1D flux power, the RMS difference is also less than 10^{-3} including points up to $k = 30\,h^{-1}\mathrm{Mpc}$. On very small scales ($k > 50\,h^{-1}\mathrm{Mpc}$), the results do depend heavily on spectral resolution, with high spectral resolution results having tens of per cent larger power. However, this result is of no concern since by this scale, the dimensionless 1D power is already $< 10^{-6}$. It seems that even at the worst resolution, all relevant lines are resolved.

We choose sightlines, or "skewers", crossing the domain parallel to one of the axes of the simulation grid and piercing the cell centers. Computationally, this is the most efficient approach. This choice of rays avoids explicit ray-casting and any interpolation of the cell-centered data, which introduce other numerical and periodicity issues. We cover the entire N^3 grid with skewers, which provides the equivalent of N^2 spectra. Although large-scale modes along different spatial dimensions are statistically independent allowing some gain in statistics from multiple viewing directions, in this work we use a single line-of-sight axis rather than combining together skewers using all 3 axes. The process of going from simulated baryon values to flux F is illustrated in Figure 1.4.

2.2 Physical properties of the Lyα forest

Zhang et al. (1998) discuss the physical properties of the Lyα forest in hierarchical models such as CDM. The discussion in this section can largely be considered as an update of that

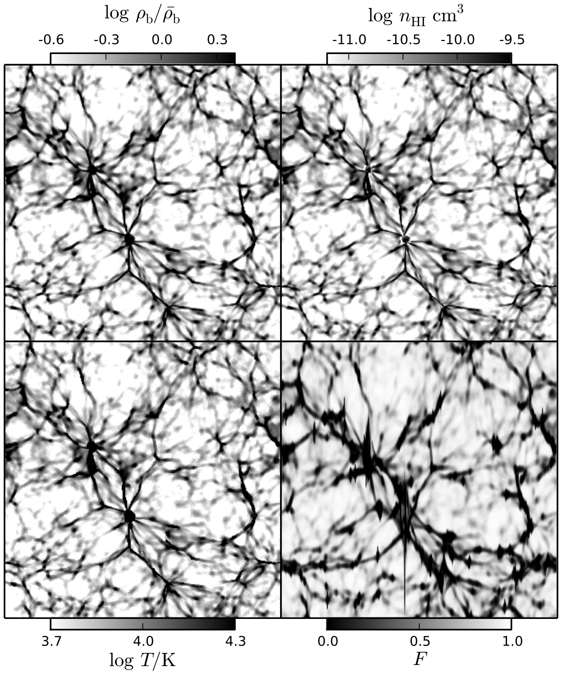

Figure 2.3: A slice of the baryon density, temperature, HI number density, and flux from the L20_N2048 simulation at $z = 2.5$. The slice covers the domain of $20 \times 20\, h^{-1}$Mpc, with a thickness of about $100\, h^{-1}$kpc (10 cells). Note that the F line of sight is the y-axis direction, so that broadened lines show up as vertical black streaks.

work.

As described above, the state of the IGM is relatively simple with a few power laws approximately tying together the spatial distribution of baryon density, temperature, proper HI number density, and optical depth to HI Lyα photon scattering. Figure 2.3 shows a slice of these quantities in one of our high-resolution simulations, except with the optical depth replaced by the transmitted flux. We choose to show flux because it highlights the range of each quantity that is relevant for observations. That is, we want to highlight differences between an optical depth of 1 or 2, which changes the flux drastically, but not between 10 and 100, which is essentially opaque. We adjusted the gray-scale intensity ranges of density, temperature, and HI number density to roughly match the morphology of the flux, which provides a good guide to what ranges of each quantity affects the Lyα forest. We note that over the relevant redshift range, the comoving density and temperature evolve slowly, so that these ranges roughly apply to all redshifts. However, the physical HI density changes drastically primarily due to expansion. The striking morphological similarity between the fields demonstrates how well the usual approximations work. The flux field is clearly the least like the other fields due to two effects: the optical depth is in redshift-space and is therefore distorted by peculiar velocities; in addition it is also thermally broadened, smearing high

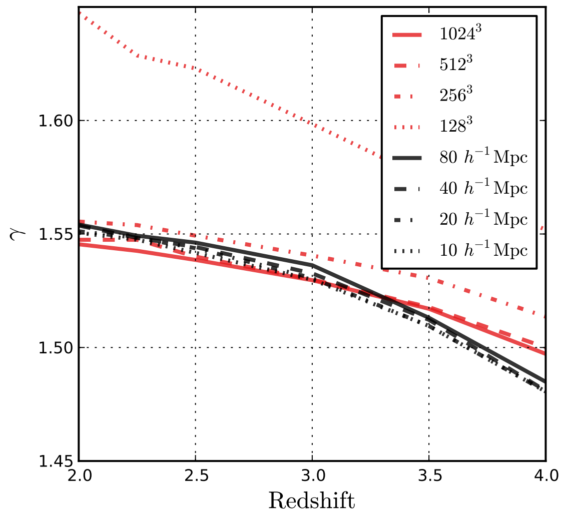

Figure 2.4: Redshift evolution of γ, the slope of density-temperature relation. Lines in black show the weak dependence on the box size (the resolution is kept fixed at $20\,h^{-1}$kpc), while the the red lines show a rapid convergence with respect to the spatial resolution (in $10\,h^{-1}$Mpc box). Black and red lines are simulations with Haardt & Madau (2012) and Faucher-Giguère et al. (2009) UV backgrounds, respectively.

temperature regions across the line-of-sight axis.

We tried fitting the density-temperature relation of the diffuse IGM several ways and found that a linear least-squares fit is sufficient. The number of points in the diffuse IGM phase is very large even for small simulations, so there is very little uncertainty in the fit parameters. However, we noticed a small but systematic difference in the best-fit γ depending on the density range fit. Fitting underdense regions, i.e. points with $-1 < \log_{10}\rho_{\mathrm{b}} < 0$ yields γ values a few percent higher than fitting near the mean density, $-0.5 < \log_{10}\rho_{\mathrm{b}} < 0.5$. Thus even if we neglect the scatter in the ρ_{b}-T relation a single power-law approximation breaks down at a few per cent accuracy.

Figure 2.4 shows the evolution of our best-fit values for γ in the resolution series of simulations and the box size series of simulations. We fit the ρ_{b}-T relation with linear least-squares in $\log\rho_{\mathrm{b}}$ and $\log T$, fitting the range $-0.5 < \log_{10}\rho_{\mathrm{b}} < 0.5$ and $\log_{10} T/\mathrm{K} < 4$. We see that convergence with spatial resolution is rather fast, and that box size does not affect recovered value of γ. In addition, we see that UV background as given by Haardt & Madau (2012), shown in black, exhibits marginally more redshift evolution than that of Faucher-Giguère et al. (2009) (the red lines in Figure 2.4). We find similar results in the fit T_0 values, where there is a small resolution effect for poor resolution, but the fit value remains the same between the L10_N256, 512, and 1024 runs. Box size appears to have no effect on the resulting ρ_{b}-T line, as expected.

A large fraction of the gas lies on the ρ_{b}-T relation line – about 90 per cent by volume and

50 per cent by mass in this case. The significant scatter above the power-law relation line is due to shock heating, whereas the small scatter just below the line is due to a subtlety in the discretization of the gravitational source term in the total energy equation. As discussed in Almgren et al. (2013), the most obvious discretization is to compute the product of the momentum and the gravitational vector. While this is spatially and temporally second-order accurate, it allows gravitational work to change the internal energy since the update to the total energy is no longer numerically equivalent to the update to the kinetic energy calculated using the updates to the momenta. An alternative discretization defines the update to total energy only through the update to kinetic energy as calculated from the momentum equation. This maintains the analytically expected behavior of gravitational work contributing to the kinetic energy only. Through numerical testing we have determined that the latter formulation greatly reduces the number of cells scattered below the line in void regions, while having a negligible effect on the results otherwise. Due to the small number of cells affected, the difference to the flux mean, pdf, and power spectrum at $k \leq 10\,h^{-1}$Mpc is only 0.1 per cent. The fraction of gravitational work in a timestep that directly contributes to the internal energy (thereby increasing the temperature) rather than kinetic energy ranges from 5×10^{-3} at early times to 5×10^{-2} at late times for a run with CFL number of 0.5. These numbers are quite independent of the spatial resolution employed. However, while the two discretizations of the gravitational source produce this difference in ρ_b-T regardless of the spatial resolution, they do converge to the same answer when refining the time-step: in simulations run with a CFL number of 0.05, the two formulations yield indistinguishable ρ_b-T plots, and the fraction of gravitational work that contributes to the internal energy stays below 5×10^{-3} throughout the run.

The overall shape of the ρ_b-T diagram is reproduced in almost any cosmological simulation, even with low-resolution, as long as it includes primordial gas heating and cooling. However, we do find that larger box size simulations produce more shocked gas around filaments (a more significant WHIM). We do not see a significant resolution dependence on the fraction of gas in the WHIM, but we see both that larger boxes have more gas in the WHIM, and that the WHIM is shocked to higher temperatures. This is expected behavior, as small-box simulations miss large-scale velocity components. For the most interesting, diffuse gas region, the ρ_b-T relation and the amount of scatter around it can also be affected by HeII reionization. For instance, McQuinn et al. (2009) found that in their post-processed radiative transfer simulations most of the reionization models increased T_0 and decreased γ while significantly broadening the ρ_b-T relation, mostly due to spatial variations in the ρ_b-T relation from RT effects like shadows. Understanding the full effects of HeII reionization on IGM is beyond the scope of this work.

In Figure 2.5, we show the evolution of the temperature at mean density. This is calculated as an average (in log space) of the gas at mean density for temperatures $T < 10^5$ K at each time step. We also show the effects of different UV backgrounds, from Faucher-Giguère et al. (2009) and Haardt & Madau (2012), and differing atomic rates (see Appendix B). Qualitatively the temperature of the IGM decreases at high redshifts due to the expansion and inverse Coulomb cooling, then rises sharply during HI reionization at $z \sim 10$. We carried

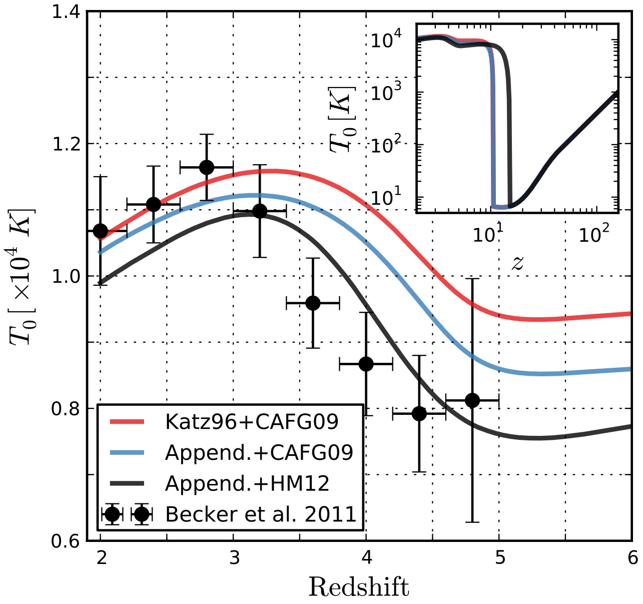

Figure 2.5: The mean temperature of the IGM as a function of redshift in our simulations compared to the observations presented in Becker et al. (2011) (error bars are 2-σ). The red line shows a simulation using Katz et al. (1996) atomic rates and the Faucher-Giguère et al. (2009) UVB. The blue line is obtained using the rates presented in Appendix B and the Faucher-Giguère et al. (2009) UVB. The black line shows a simulation with the rates presented in Appendix B and the Haardt & Madau (2012) UVB. While the main figure shows the T_0 evolution over the observationally relevant redshifts, the inset figure shows the full simulation range starting at $z = 159$, on a logarithmic scale.

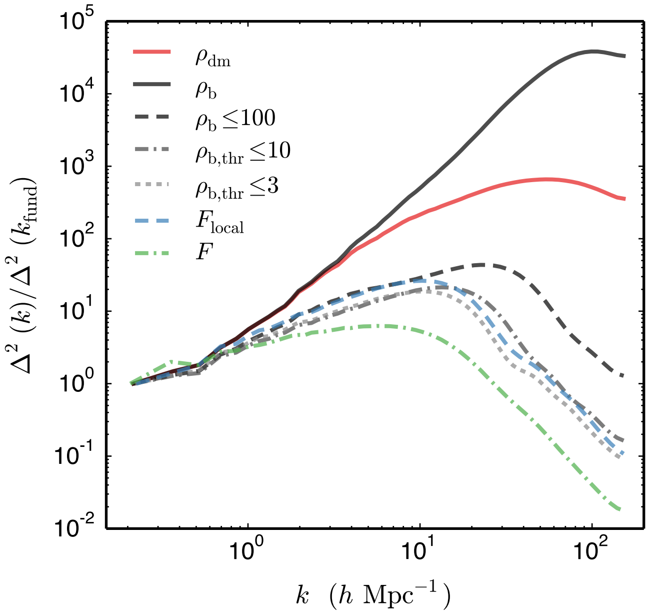

Figure 2.6: Illustration of the effect of the filtering scale on power spectra. Here we show the power spectra of baryon density $\rho_{\rm b}$, dark matter density $\rho_{\rm dm}$, the local flux $F_{\rm local}$, the monopole of the redshift-space (observable) flux F, and the thresholded baryon density $\rho_{\rm b,thr}$ limited to 100, 10, and 3 times the mean density, all from the L40_N2048 simulation at redshift $z = 2$.

out a study spanning several orders of magnitude in initial temperature for our simulations and have determined that, due to adiabatic expansion, Compton cooling and hydrogen reionization, no memory of the initial temperature is retained at $z \lesssim 10$. In Figure 2.5, we also show recent observational results from Becker et al. (2011), which is in good agreement with the $z = 2.4$ measurement recently carried out by Bolton et al. (2014) but lower than the temperatures inferred by Lidz et al. (2010). It is interesting to point out that the differences in temperature evolution that different modern UV backgrounds produce, roughly 10 per cent, are less than observational uncertainties. We also note that both of the UV backgrounds we consider show two visible jumps in the temperature of the IGM, corresponding to H I and He II reionizations.

Due to the direct influence of pressure forces, baryon fluctuations are suppressed compared to dark matter (which is affected by the gas pressure only because the gravitational acceleration has a component due to the gas). Our simulations do not account for the details of star formation, feedback from stars or Active Galactic Nuclei; the regions that should be galaxies are only blobs of overcooled gas. Due to this overcooling inside halos, small-scale fluctuations in the baryonic component are artificially enhanced as shown in the solid black line of Figure 2.6 (the red line shows the dark matter power spectrum for reference). Since we know that our simulations do not realistically represent the gas quantities in high density regions, we can exclude them from our analysis at which point the filtering scale becomes

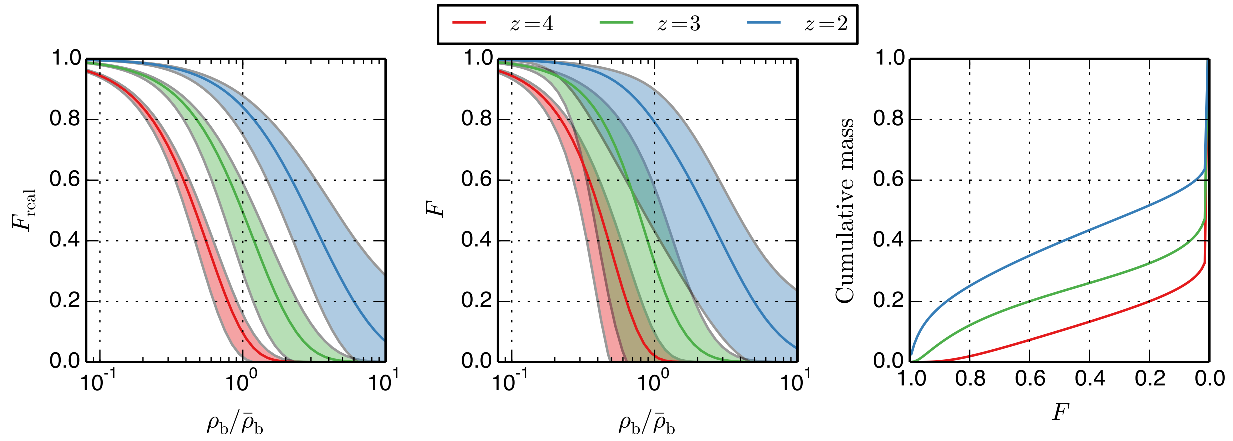

Figure 2.7: Contribution of different density regions to the Lyα forest flux at $z = 2$, 3, 4. Meax flux was rescaled to match Becker et al. (2011). *Left:* the real-space flux vs. gas density. The lines show the medians and the filled regions show the normalized median absolute deviation (normalized to match one standard deviation for a normal distribution). *Middle:* the redshift-space flux vs. gas density. *Right:* the cumulative distribution of mass vs. redshift-space flux.

clear. To highlight this, in Figure 2.6, we show the baryon power spectrum with several density thresholds. These are obtained by 'clipping' the original baryon density field, i.e. resetting the densities higher than the threshold down to the selected threshold value. The clipping is done here only for illustrative purposes. This is qualitatively similar to what happens with the Lyα forest signal – where the flux drops to zero at a certain density, and any higher density has no additional effect. Clipping of the small-scale fluctuations also introduce a linear bias on large scales. For clarity, we have normalized all power spectra to be 1 at the fundamental (box-scale) mode. The different threshold value power spectra also illustrate that there is a density dependence of the filtering scale. As the threshold density decreases, the filtering scale increases. We also show two flux power spectra to see how they probe the filtering and thermal broadening scale. We computed the Lyα forest flux without redshift-space distortions or thermal broadening, which we call the local flux $F_{\rm local}$. In this case, the optical depth for the local flux is just the appropriate rescaling of the HI number density,

$$\tau_{\rm local} = \frac{\pi e^2 f_{lu}\lambda_0}{m_e c H(z)} n_{\rm H\,I} \; ; \; F_{\rm local} = e^{-\tau_{\rm local}} \; . \tag{2.8}$$

The dashed blue line is the local flux spectrum, which shows pressure support smoothing at a scale roughly matching the $\rho_{\rm b,thr} \leq 10$ result. Note the little difference between thresholding at 10 times the mean baryon overdensity and 3 times the mean. We also show the monopole of the 3D flux power spectrum as the dashed green line, which includes smoothing not just from pressure support but also contributions from thermal broadening and redshift-space distortions, giving rise to even more filtering on small scales.

In Figure 2.7, we show relations between the Lyα forest flux and the gas density. In the left panel we plot the real-space flux of cells as a function of gas density. For each density bin we plot the median and normalized median absolute deviation (normalized to equal the standard deviation for a normal distribution) independently above and below. This serves as a qualitative estimate of what density regimes contribute to the Lyα forest signal at different redshifts. For instance, we immediately see that a majority of the signal at high redshift originates in under-dense regions, while at $z = 2$, it lies in the mild overdensities. In the middle panel, we show similar info, but this time we use the redshift-space flux. As redshift-space distortions couple regions several Mpc away and can map different cells to the same redshift-space position (see Figure 1.4), the redshift-space flux is less correlated with density and thus exhibits more scatter than in real space. However, the median lines are similar at all redshifts. In the right panel, we show the cumulative mass of cells with fluxes above some value. The sharp rise in the cumulative mass at $F = 0$ shows the mass fraction in the saturated regions of the forest, filaments and halos. This figure also shows the difficulty of simulating Lyα forest signal at high redshifts, $z \gtrsim 4$: we immediately see that small fluctuations in density produce significant difference in flux. Arguably, this effect is more critical for numerical convergence than the decrease in filtering scale described in Section 2.1.

Historically, the Lyα forest was studied in the context of absorption line systems. However, the process of Lyα forest line finding and fitting is not well-defined and results can vary between implementations. For this reason we will explore line statistics separately in Section 2.7, and in the following sections we will focus on the flux N-point correlation statistics.

2.3 Resolution study

The physical resolution required to model Lyα forest flux statistics varies significantly with redshift, with higher redshifts requiring higher resolution for the same relative error. There are several physical effects which contribute to this behavior. One is the change in the comoving filtering scale, which decreases with increasing z. We further demonstrate the increasing steepness of the flux-density correlation as a function of redshift in Figure 2.7, which means that for the same relavite error in ρ_{b} or $N_{\mathrm{H\,I}}$ the relative error in flux will be larger at $z = 4$ than at $z = 2$. Finally, the gas in voids is ~ 2 times colder than the gas at the mean density (and a factor of ~ 4 colder than in mild overdensity regions), therefore thermal broadening of lines is less at high redshifts than at low ones.

The average transmission sharply increases going to lower redshifts as the physical density of neutral hydrogen decreases, primarily due to the expansion of the Universe, with secondary changes due to the ionizing background radiation. Here we explore the accuracy of simulated Lyα forest flux statistics at $2 \le z \le 4$, the relevant range for most of current and near future observations. We focus on results in our $10\,h^{-1}$Mpc boxes, as they offer the easiest path to an increase in resolution, but we have also explicitly checked that the conclusions presented here are valid in the case of the larger-box simulation series as well (see Table 2.2). In other

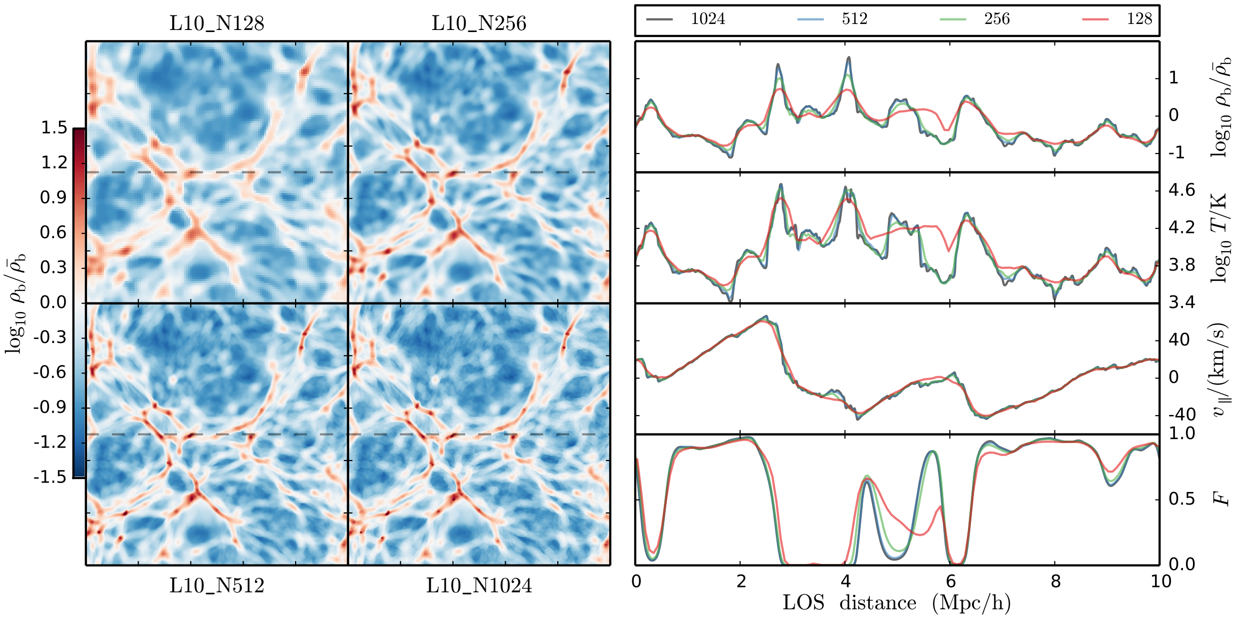

Figure 2.8: Illustration of the effects of resolution in the 10 h^{-1}Mpc simulations at $z = 2.5$. On the left, we plot slices $\sim 150\,h^{-1}$kpc deep (2 cells in the 128^3 simulation) of the baryon density. On the right, we show values along the skewers marked as dashed lines in the slices. The skewers were selected in the same position in each simulation.

words, we do not observe that numerical errors due to missing modes (explored in the next section of this paper) couple with resolution error at more than the per cent level. The same behavior is observed in Gadget simulations presented in Chapter 3. The simulations we present in this section were done with Faucher-Giguère et al. (2009) UV backround.

The results of this section are applicable to grid/Eulerian codes, where the effective resolution of pressure forces is commonly better than the effective resolution of gravitational forces. For example, many tests show that hydrodynamic quantities are already very accurate at 1–2 cells away from discontinuities – see e.g. Almgren et al. (2010) for the case of the hydro algorithm implemented in Nyx but the same is true for schemes used in virtually any other cosmological Eulerian hydro code to date. On the other hand, the gravitational force resolution is much worse. Grid codes use a Particle-Mesh (PM) scheme to compute gravitational forces, which is very fast but suffers from smoothing the density field at small scales. Generally, two particles must be separated by at least 5 cell sizes for the gravitational force to match $1/r^2$ (for example, see Habib et al. 2009). The opposite is true in most SPH Lyα forest simulations presented in the literature. In this case, gravitational resolution is much higher for the same grid size/number of particles, with the gravitational forces computed with a TreePM or particle-particle PM hybrid scheme. This provides a much (roughly 10 times) higher gravitational resolution than the grid codes for the same grid configuration. At the same time, the SPH kernel smooths the hydrodynamic quantities on scales of $\sim 2\times$ the mean inter-particle spacing for gas around mean density. In this regard, the resolution study presented here is not directly applicable to *all* codes. However, we believe that the results of other studies we conduct in this paper are largely code independent.

In Figure 2.8, we provide an illustration of how the grid resolution affects relevant IGM structures and the Lyα forest flux. Here we use our four $10\,h^{-1}$Mpc simulations and plot a slice and skewer in the same position from each simulation. The baryon density slice is on the left, while on the right we show baryon density, temperature, velocity parallel to the line of sight, and transmitted flux along the skewer. In both slices and skewers, we see a clear pattern of converging values. Overall, the L10_N256, L10_N512, and L10_N1024 results agree very well, and the L10_N128 results are similar, but have structural differences. In the baryon density slices, we see that structures in L10_N128 are severely under-resolved. The large cell size prevents the collapse of dense regions, and the solution contains puffy filaments and halos, and less depleted voids. The relatively small number of resolution elements also means that the simulation misses the rare, extremely low and high density regions. In the L10_N256 slice we can see structure that resembles the highest-resolution case much more closely, although the filaments and halos are still a bit puffier. Finally, the differences between the L10_N512 and L10_N1024 slices are minor. The filament widths are essentially the same and the differences noticeable by eye are restricted to the very dense galaxy-like regions. This is fortunate for modeling the Lyα forest signal, since the dense regions are saturated in absorption, rendering those differences undetectable in flux. In the baryon density and temperature skewer values we see the same patterns. The L10_N128 simulation reproduces the broad shapes, but fails catastrophically at the extremes. The other simulations match each other much better, and the L10_N512 and L10_N1024 values are

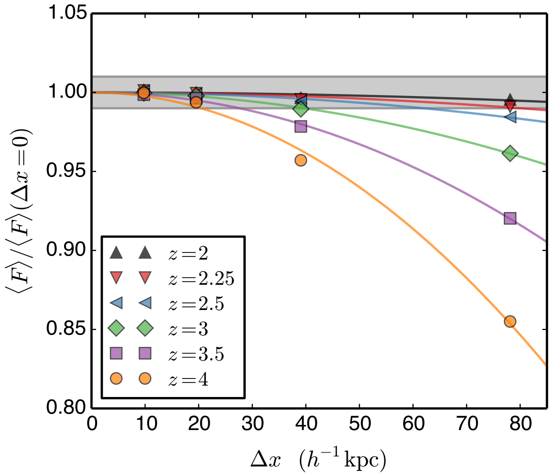

Figure 2.9: Dependence of the mean flux on resolution for 6 redshifts. The lines are the best fits of the form $F(\Delta x) = F(0) - k\Delta x^2$, and the data plotted is normalized to the case $\Delta x \to 0$, i.e. $F(0)$. The gray shaded region shows a ±1 per cent interval.

very close at all positions. One difference is in the dense structure near the LOS distance of 4 h^{-1}Mpc, where the L10_N1024 simulation resolves two temperature peaks, almost certainly accretion shocks. The L10_N512 simulation just barely reproduces the two peaks while this feature is smeared out as one bump in the two lower resolution simulations. The flux field proves to be unaffected by those kinds of details as can be seen in the lowest panel. Interestingly, the parallel velocity values show much less difference between runs. This reinforces the common knowledge that bulk flows are not as sensitive to resolution as they are to the box size. Finally, the most important differences lie in the flux values. Here, we see that L10_N512 and L10_N1024 runs look virtually identical. Further, many of the small differences in baryon quantities between the L10_N256 and higher resolution runs are washed out in the optical depth calculation – the similar velocity shifts and significant broadening provide a fortunate ‘fudge factor’ when only considering the flux. The same cannot be said for the L10_N128 simulation fluxes though, which show significant differences, especially at the LOS distance of 5 h^{-1}Mpc. We have checked this for several random skewers and with all other redshifts available and note that the overall conclusions remain the same, although differences very mildly increase with redshift. In the following sections, we quantify the above differences in resolution.

2.3.1 Mean flux

Figure 2.9 shows the mean flux in four of our 10 h^{-1}Mpc simulations at the snapshot redshifts. Here we immediately see that higher redshifts need higher resolution to maintain the same accuracy. The coarsest resolution run is within 1 per cent of the highest resolution run at $z = 2$, but ∼15 per cent different at $z = 4$. NYX is second-order accurate in both the

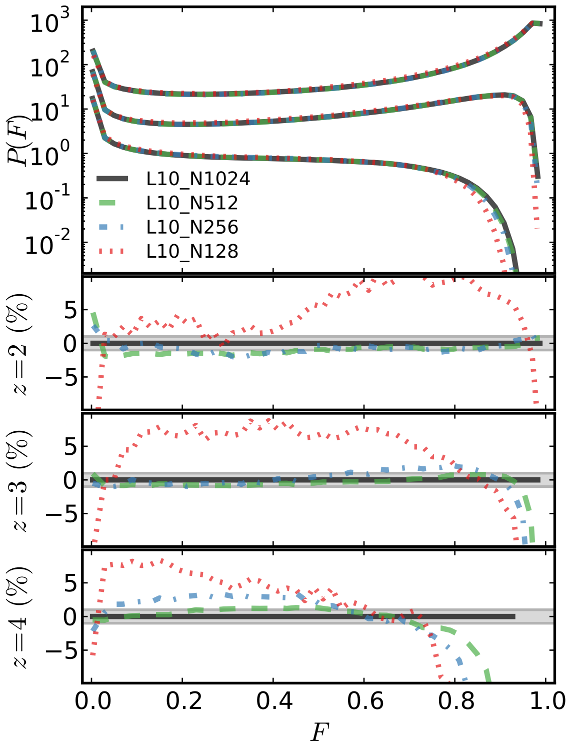

Figure 2.10: Convergence of the flux PDF with respect to grid resolution. For clarity, in the upper panel we have multiplied the $z = 2$ data by a factor of 100, and the $z = 3$ data by 10.

gas dynamics solver and gravity. Although the Lyα forestflux is a heavily processed quantity derived from the density, velocity, and internal energy of the gas, its mean clearly exhibits quadratic convergence, as shown in Figure 2.9. The resolution series allows us to determine $F(0)$ — the simulated mean flux in the limit $\Delta x \to 0$. Understanding the effect of resolution on the mean flux is important for simulation results that rescale optical depths to match an observed mean flux, as we explore later in Section 2.6.

2.3.2 Flux PDF

We consider the convergence of the flux PDF $P(F)$ at redshifts 2, 3, and 4, which we show in Figure 2.10. Again, we note that the resolution requirements increase with redshift. It appears that this is mostly due to the rarity of transmissive regions at high redshift. In the $z = 4$ ratio panel, we can see that even the L10_N1024 simulation does not capture pixels with $F \approx 1$, as the black line cuts short near 0.9. It is instructive to look again at Figure 2.7, which clearly shows how difficult it is to obtain $F = 1$ cells at high redshifts, even in very underdense regions. The L10_N512 simulation does match the highest resolution to a few percent up to $F < 0.6$. At lower redshifts however, L10_N512 is in percent agreement with the 1024^3 run, while the L10_N256 is within a few percent. As expected

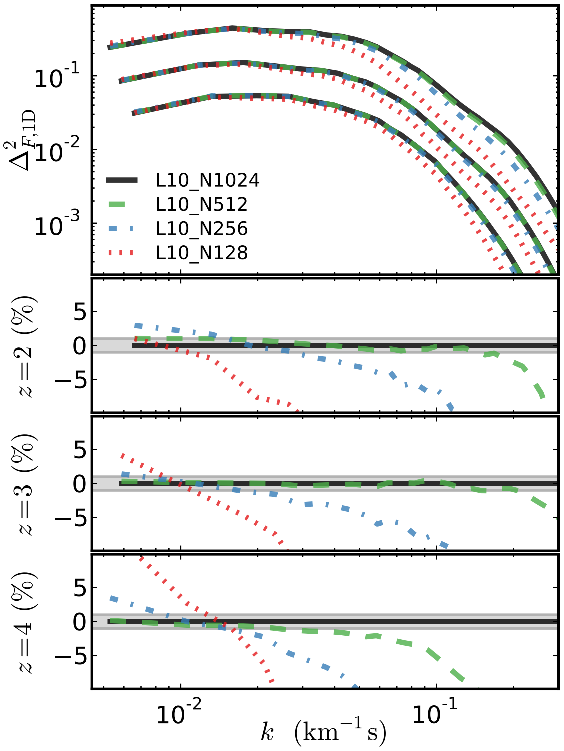

Figure 2.11: Convergence of the 1D power spectrum at redshifts 2, 3, and 4. Here we do not modify the values in the upper panel – the flux power increases with increasing redshift, so from top to bottom are redshifts 4, 3, and 2.

from the qualitative inspection at the beginning of this section, the L10_N128 simulation results are qualitatively different at all redshifts. At high redshift, it severely underestimates transmissive pixels, pushing the low-flux end above the other simulations. At low redshift, it misses the extreme fluxes at both ends, raising the probability at moderate fluxes above other simulations.

2.3.3 Flux 1D power spectrum

Here we consider the resolution convergence of the dimensionless 1D flux power spectrum $\Delta^2_{F,\mathrm{1D}}$ at redshifts 2, 3, 4, as shown in Figure 2.11, leaving other effects for the subsequent sections. We want to emphasize again that while we show results for the 10 h^{-1}Mpc box, we have checked that conclusions are the same in other convergence series in larger box sizes. Figure 2.11 shows that 20 h^{-1}kpc resolution run (L10_N512) agrees with 10 h^{-1}kpc run (L10_N1024) to better than 1 per cent at redshifts $z = 2$ and $z = 3$ even beyond $k = 0.1$ km^{-1}s. Those are much smaller scales than what is usable for making cosmological constrains, as interpreting observations becomes difficult at such small scales due to metal lines and other contaminents. Those scales are also not correctly modeled with the physics

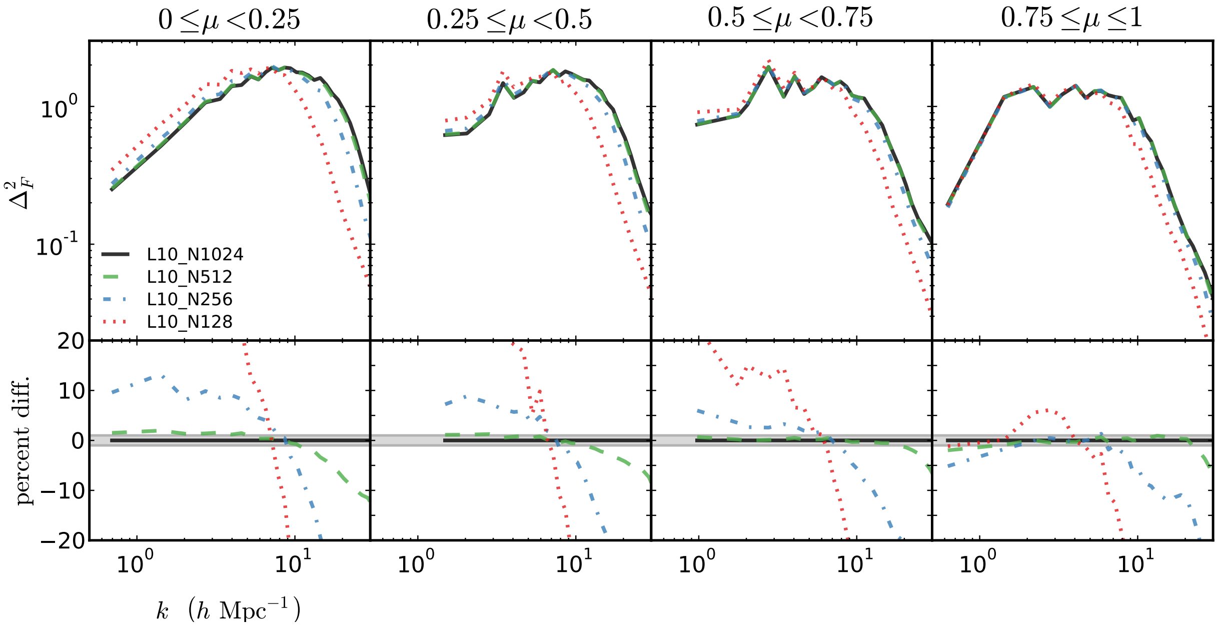

Figure 2.12: Convergence of the 3D flux power spectrum in 4 μ bins at redshift $z = 4$ (the agreement is better at lower redshifts). The leftmost panel shows the power spectrum mostly perpendicular to the line of sight, the rightmost is mostly parallel to the line of sight. The agreement is better along the line of sight due to thermal broadening which erases some of the differences.

included in our simulations. The $z = 4$ ratio panel again shows how hard it is to get flux statistics accurately at very high redshifts: the 20 h^{-1}kpc run departs from the 10 h^{-1}kpc run by 1 per cent around $k = 0.03$ km^{-1}s. This is still sufficiently good for cosmological purposes, especially since the number of observed quasars at such high redshifts is rather small.

One important difference between density and flux fields, is that density is manifestly conserved in our simulations, and its mean value is an input parameter. In contrast, the mean flux will differ — even when the cosmology and physical models for cooling and heating processes are kept constant — due to numerical resolution, box size, and the random realization of the initial density field. Another characteristic feature of the flux field is that it is bounded in value: $0 \leq F \leq 1$. The maximum possible fluctuations around the mean value are therefore also limited. This is again in contrast to the density fluctuations, as density contrast can in principle go to infinity. As a result strong suppression of flux fluctuations on small scales — for example due to numerical effects like lack of resolution — results in increased fluctuations on large scales. This effect is also clearly visible in the Figure 2.11, and is more noticeable when the fluxes in simulations of different resolutions are not rescaled to the same mean (as done here). This effect is the biggest issue to getting the 1D flux power spectrum correct in numerical simulations. Whereas the density power spectrum can, to some extent, be simulated with low resolution simulations using a series of nested-size boxes, each box recovering accurately only a small portion of $P_m(k)$, the flux power spectrum in an under-resolved Lyα forestsimulation will be inaccurate on all scales.

2.3.4 Flux 3D power spectrum

The 3D flux power spectra for 4 μ bins are shown in Figure 2.12. The leftmost panel shows the power spectrum mostly perpendicular to the line of sight, and the rightmost is mostly parallel to the line of sight. Here we show only redshift $z = 4$ data, since this is where the agreement is worst. The agreement is better at lower redshifts, as expected from previous considerations presented in Section 2.1, and Section 2.2. We immediately notice that different resolutions agree more along the line of sight than across it. This is a result of thermal broadening which erases much of the small scale differences, bringing the results of lower resolution runs closer to the high resolution solution. We nevertheless see that the 20 h^{-1}kpc resolution is good enough for most practical purposes, typically 1 per cent away from the 10 h^{-1}kpc result at all redshifts for $k < 10$ h Mpc^{-1}. From the observed rate of convergence we expect that the difference between a 10 h^{-1}kpc and a (hypothetical) 5 h^{-1}kpc run would be sub-percent.

2.3.5 Richardson extrapolation

For a convergent numerical method, it is in principle possible to increase the accuracy of a measured quantity by carefully combining the results from a sequence of simulations where the only difference is the spatial resolution. Here, we discuss the case of combining results

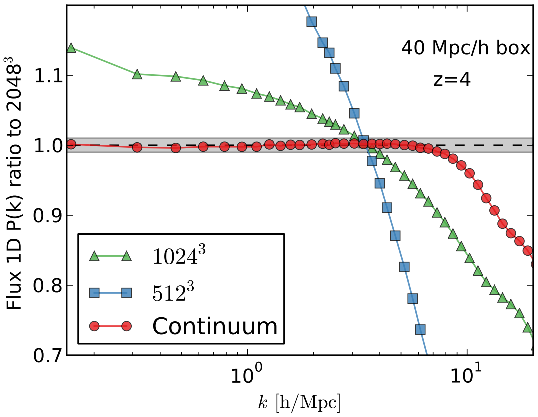

Figure 2.13: Convergence of 1D flux power spectrum at redshift $z = 4$ in a $40\,h^{-1}$Mpc box. We show the results for 512^3 and 1024^3 simulations, together with the Richardson extrapolation from these two runs using the theoretically expected order of convergence, $p = 2$. The shaded band shows ±1 per cent range.

via Richardson extrapolation. A numerical method which is p-th order accurate in space (meaning the error term is proportional to h^p where h is discretization element), produces a numerical approximation $Q(h)$ as

$$Q(h) = Q + Ah^p + O(h^{p+1})\ . \tag{2.9}$$

The first term on the right-hand side Q is the exact value, the second term is the leading error, and the third term is the higher-order error. The leading error can be removed with simulations using two different values of h, for example h and rh, where r is the refinement ratio, giving an extrapolation expression

$$Q_R = \frac{r^p Q(h) - Q(rh)}{r^p - 1}\ . \tag{2.10}$$

The order of accuracy, p, is theoretically known from the algorithm implemented, but can also be determined from actual numerical results. This requires at least 3 simulations, in which case p can be calculated as:

$$p = \frac{\ln\left(\frac{Q(r^2h)-Q(rh)}{Q(rh)-Q(h)}\right)}{\ln r}\ . \tag{2.11}$$

In Figure 2.13 we show one such Richardson extrapolation, applied to our 1D flux power spectrum results. Here we consider the results at $z = 4$, since the differences are largest at this time, and we also use our 40 h^{-1}Mpc box size simulations. We see that the run with 40 h^{-1}kpc resolution (L40_1024) is not as accurate, differing from the 20 h^{-1}kpc resolution run (L40_2048) by up to 15 per cent in the range $k < 10$ h Mpc^{-1}. The run shown in blue

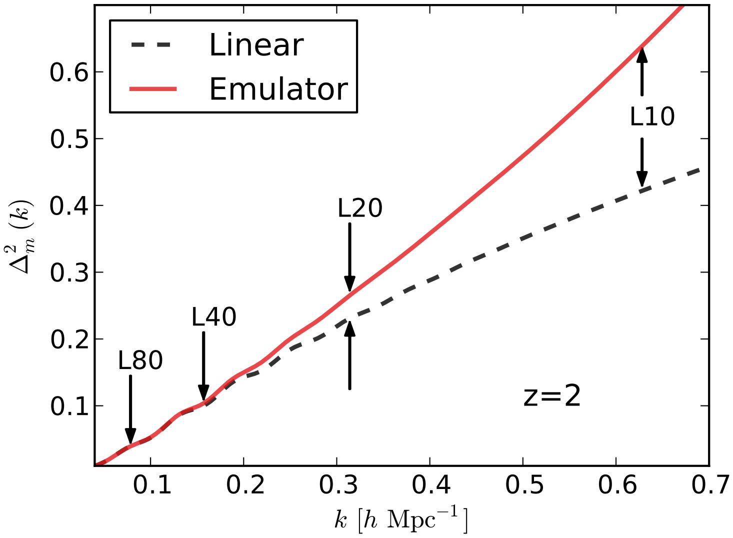

Figure 2.14: Dimensionless linear and non-linear matter power spectrum at $z = 2$. Arrows show the scale of the fundamental mode in our boxes of 10, 20, 40, and 80 h^{-1}Mpc.

(79 h^{-1}kpc resolution) — is even worse. However, the “continuum” value deduced from these two runs via Richardson extrapolation, using Equation 2.10, shows remarkable agreement with the highest resolution reference run. Here, we have used an order $p = 2$, since Nyx is formally second-order. The fact that the theoretical value works so well on the 1D flux power spectrum is very reassuring. For larger k values the extrapolation fails. This is expected as the extrapolation procedure cannot reproduce power that is not present in the underlying low resolution simulations, nor can it work in the regime where the convergence breaks down due to a dramatic loss of accuracy close to the resolution limit. Nevertheless, we see that extrapolation can significantly improve accuracy from low-resolution simulations on scales where convergence does hold. This improvement is a strong evidence that numerical errors beyond the discretization scheme Nyx employs are small to none, and a confirmation of the desired rate of convergence even in a very processed quantity like the 1D flux power spectrum.

2.4 Box size / missing modes

In cosmological simulations, we model a representative, but finite volume of the universe using periodic boundary conditions in all 3 dimensions. As a result, perturbations on scales larger than the box size are identically zeroed out while fluctuations which are smaller, but comparable to the box size, are poorly sampled. A finite box size can compromise Lyα forest results in at least two different ways. First, in cosmological simulations in general, once the non-linear scale of density fluctuations becomes similar to the box size the evolution

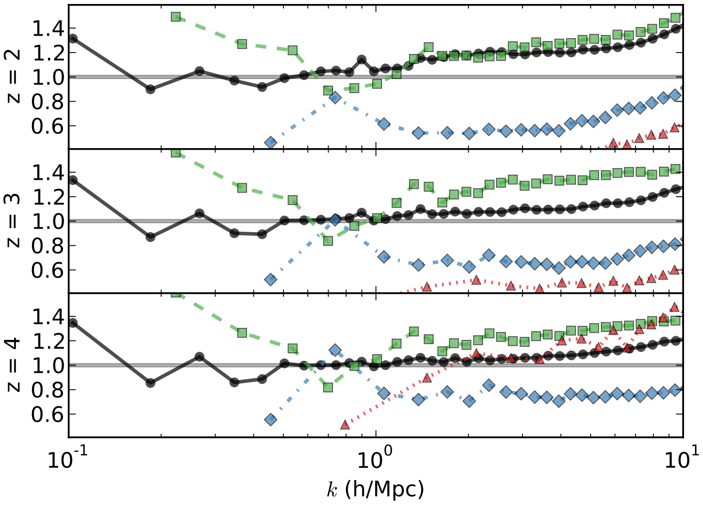

Figure 2.15: The matter power spectrum for different box-sizes. We show ratios to the emulated predictions from a gravity-only simulation (Heitmann et al. 2014) at $z = 2$, 3, and 4. Red triangles are $10\,h^{-1}$Mpc box, blue diamonds are $20\,h^{-1}$Mpc, green squares $40\,h^{-1}$Mpc, and black circles $80\,h^{-1}$Mpc box. The resolution is constant in all runs: $19.5\,h^{-1}$kpc.

of modes is suppressed compared to what would be obtained with a larger simulation box. Second, and relevant to the Lyα forest, a lack of large-scale modes — even linear ones — can lead to an underestimation of the bulk flow velocities of the gas. This, in turn, leads to an underestimation of the heating from accretion shocks. This has a significant impact on both the thermal broadening of lines (via the amount of shock heating) and on the redshift-space distortions of the optical depth. Thus if the simulated box is too small it cannot produce representative Lyα flux statistics for the cosmology of interest.

To estimate the non-linear scale of density perturbations, we use the power spectrum emulator FrankenEmu (Heitmann et al. 2014), shown in Figure 2.14 for the cosmology considered here and at redshift $z = 2$. Since we end the simulations at $z = 2$, this is the worst-case scenario in terms of the required box size. This alone indicates that 40 h^{-1}Mpc is the bare minimum to avoid the box-size mode becoming non-linear, with 80 h^{-1}Mpc being a more comfortable value. In the context of "missing modes" in simulations of the Lyα forest, an important and thorough recent work is that of Tytler et al. (2009). The range of box sizes they consider is even larger than the one presented here: their biggest box (54.5 h^{-1}Mpc) is similar to our largest, while they go to box sizes as small as 1.7 h^{-1}Mpc. Thus most of their simulations are over-evolved at $z = 2$, where the largest – anchoring – mode is deeply in the non-linear regime according to Figure 2.14. Note that their choice of cosmology has $\sigma_8 = 0.9$, therefore non-linearity starts at even larger scales than the cosmology we consider here. In addition, due to the high computational expense, they had to restrict their box size series analysis to a spatial resolution of 53.3 h^{-1}kpc. This resolution is significantly more coarse than the one we find necessary in this paper (20 h^{-1}kpc, Section 2.3), but also coarser than what Tytler et al. would have likely run (13.3 h^{-1}kpc, see their section 11.3) if it were computationally feasible. Here we present a box size convergence study extending to box sizes large enough to sample linear modes even at the end of the simulations ($z = 2$), but also with the desired spatial resolution to capture Lyαstatistics to the per cent level.

Before turning to flux quantities, we will first look at the convergence of the matter power spectrum in our runs as we increase the box size while keeping the resolution constant. This is shown in Figure 2.15. We clearly see the suppression in mode growth in the small-box simulations with respect to 80 h^{-1}Mpc run. The differences in the matter power are rather significant, but as we will show later – and as shown in Tytler et al. (2009) – the differences in the flux power are much less.

2.4.1 n-point flux statistics

In Figure 2.16, we show the mean flux in different box-sizes for a constant spatial resolution. As expected, the 10 h^{-1}Mpc box is significantly inaccurate, while already in the 20 h^{-1}Mpc box we obtain reasonable mean values. As in the resolution study in Section 2.3, we see the same trend of growing differences as we move to higher redshift. This is not immediately intuitive behavior, as one would expect small boxes to be less affected at $z = 4$ rather than at $z = 2$. As we do not have many independent realizations of each box-size, we cannot state with certainty how much this effect is due to runs having different realizations versus an actual physical effect. Another thing to note is that convergence is not one-sided (e.g. as the box size is increased the mean flux does not increase as was the case in the resolution study). Again, this could be just statistical variance. Similar behavior is reported in Tytler et al. (2009), see their Table 5. Overall, we see the behavior one would expect from Figure 2.14 – namely, that there is only a small difference between 40 h^{-1}Mpc and 80 h^{-1}Mpc boxes. The difference increases with a further reduction in box-size, and becomes clearly too large in the 10 h^{-1}Mpc box.

As was done in Section 2.3, we first remove the differences in the mean flux value by rescaling the optical depth in all boxes to the value in our "best" simulation, the 4096^3 run in an 80 h^{-1}Mpc box. Since the rescaling is small for all but the 10 h^{-1}Mpc simulation this plays a rather minor effect, and our conclusions would be the same if we presented unscaled results with the differences being marginally higher. In Figure 2.17, we show the dependence of the flux PDF with respect to the box-size. As in the case of the flux mean, we see that the 40 h^{-1}Mpc and 80 h^{-1}Mpc boxes are in per cent agreement except at the very transmissive end, $F \approx 1$, and at higher redshifts. As commented on in the resolution study, the fully transmissive pixels become very rare at higher redshifts due to the high physical density of neutral hydrogen and therefore the error in determining the relative fraction of such regions decreases. Qualitatively our results at $z = 2$ are in line with those presented in Tytler et al. (2009) (though note most of their boxes are smaller than ours).

Next, we turn to the 2-point correlation function of the Lyα forestflux while first exam-

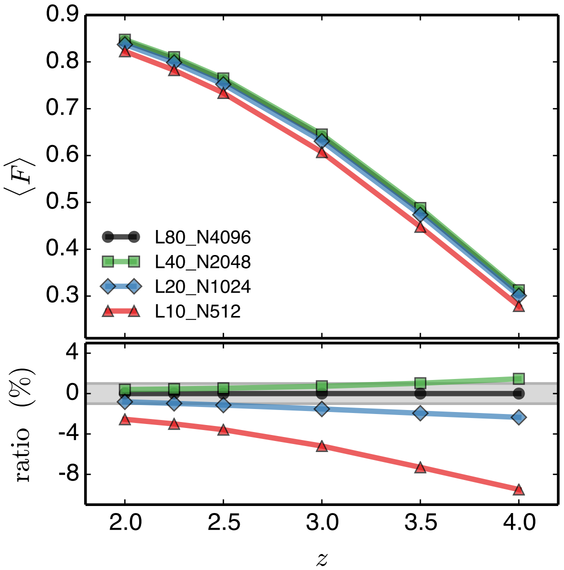

Figure 2.16: Dependence of the mean flux on the box size, for 6 different redshifts. Upper panel shows the mean flux, lower panel presents the ratio to the largest box-size run – 80 h^{-1}Mpc.

ining the 1D $P(k)$. In Figure 2.18, we immediately see that the differences in the flux power are much less than in the matter power. This is not unexpected as the flux comes from only moderate over densities, which are less affected by the sample variance than halo regions. The convergence of the low-k region is difficult to assess due to different realizations of the initial conditions, but overall we again see nice agreement between the 40 and 80 h^{-1}Mpc boxes. Here, however, the 20 h^{-1}Mpc box is noticeably in error (by 5-10 per cent) while the 10 h^{-1}Mpc box has no value for precision cosmology measurements. As before, our results are in good qualitative agreement with Tytler et al. (2009). As we will show next, most of the differences in the 1D power originate from the differences in power along the line of sight. Despite those differences we conclude that for 1D $P(k)$ constraints the 40 h^{-1}Mpc box size is a reasonable one, while 80 h^{-1}Mpc is a safe choice.

Finally, we turn to the 3D P(k), looking at 4 μ bins, going from across the line of sight $0 < \mu < 0.25$ to the power along the line of sight $0.75 < \mu < 1$. This was investigated in McDonald (2003), where they ran simulations with box-size spanning 20 to 80 h^{-1}Mpc, very much like the simulations presented here. However, those were HPM simulations rather than full gas dynamics, and the cell size was kept constant at a rather large value of 156 h^{-1}kpc.

In Figure 2.19 we see good agreement between the 40 and 80 h^{-1}Mpc simulations when the spatial resolution is kept constant. At low k, there is a substantial scatter between simulations as random phases in initial conditions differ in different box-sizes. As a result, we cannot meaningfully compare our boxes at large scales. However, at smaller scales, we see that a 40 h^{-1}Mpc simulation is in percent-level agreement with the 80 h^{-1}Mpc run. Again, just by observing the convergence rate for different box sizes we can be confident that 80

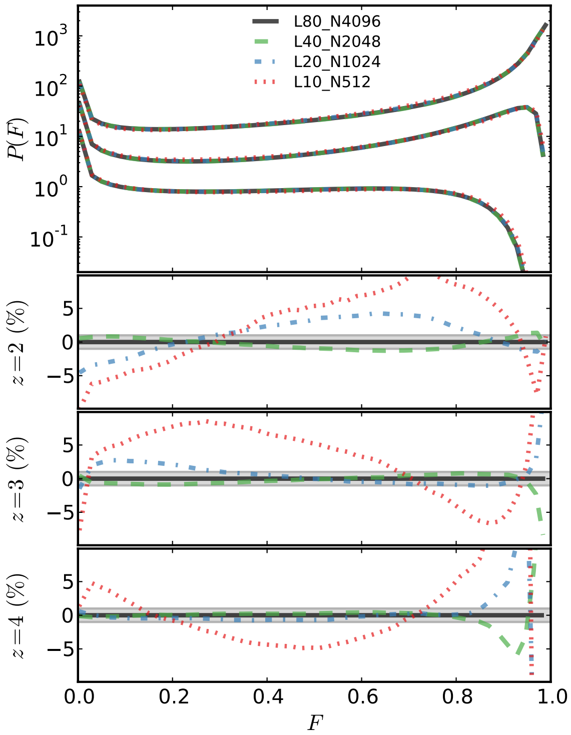

Figure 2.17: Convergence of the flux PDF with respect to box size. For clarity, in the upper panel we have multiplied the $z = 2$ data by a factor of 100, and the $z = 3$ data by 10. The resolution is kept constant at $19.5\,h^{-1}$kpc, and the box-size increases from $10\,h^{-1}$Mpc (dotted red line) to $80\,h^{-1}$Mpc (solid black line).

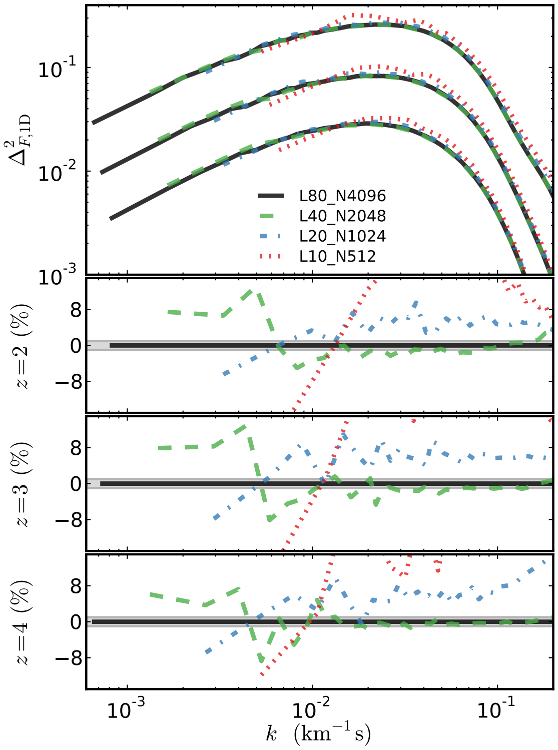

Figure 2.18: Convergence of the flux 1D power spectrum with respect to box size at redshifts 2, 3, and 4. Dimensionless flux power is shown in the upper panel, while ratios to the $80\,h^{-1}$Mpc run at 3 different redshifts are shown in the lower panels. Colors and line styles follow that of Figure 2.17.

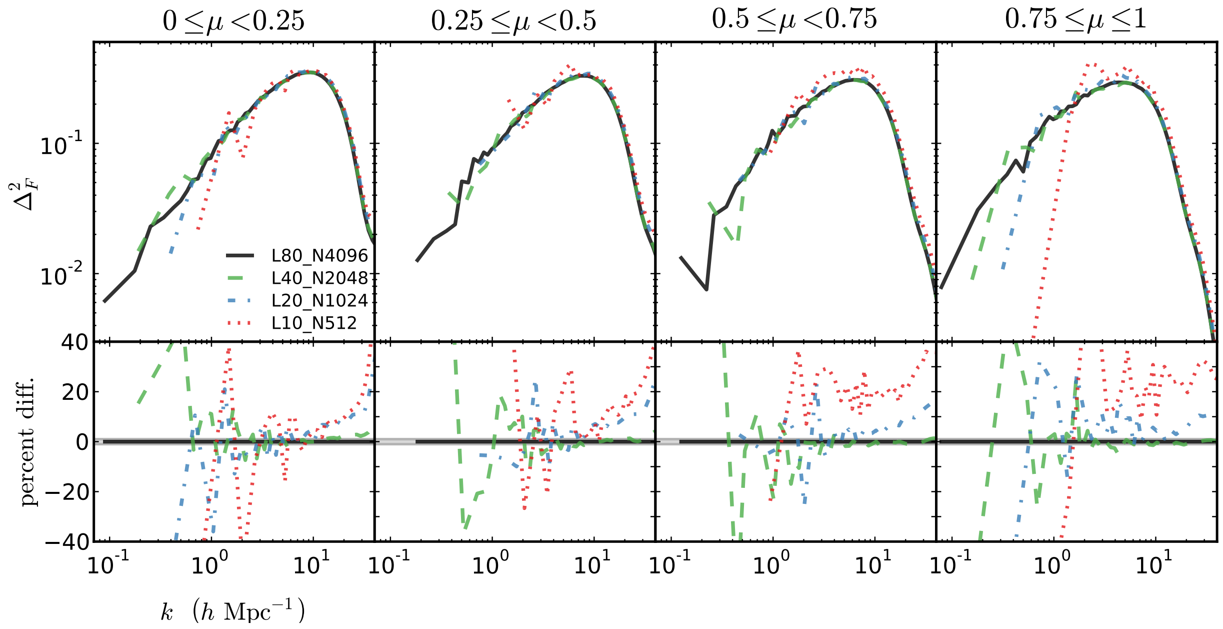

Figure 2.19: Convergence of the flux 3D power spectrum with respect to box size in 4 μ bins. As in Figure 2.12, we show only $z = 4$ data, although here this is representative of the agreement at other redshifts as well. The modes at large scales are different due to sample variance.

h^{-1}Mpc is a very safe value to run at – most likely sub-percent accurate.

Here, we also need to comment on the large-scale Lyα forest bias: on large scales, the Lyα flux is a biased tracer of the matter field, and as such is of great value for cosmology (Slosar et al. 2011, 2013; Busca et al. 2013). At present – or in the near future – running hydrodynamical simulations with box sizes needed to sample the BAO peak, and at the same time obtain the resolution necessary to resolve density fluctuations in the IGM, is not a viable approach. Still, one does not necessarily need to have a BAO-regime simulation box to reach a regime where the redshift-space Lyα flux power is related to a real-space density power via a k-independent (Kaiser 1987) formula:

$$P_F(k, \mu) = b^2(1 + \beta\mu^2)^2 P_m(k) \tag{2.12}$$

We have examined bias in our simulations, and have found that they are all too small for a reliable fit to b and β of Equation 2.12. Our 80 h^{-1}Mpc box barely reaches a regime where the parameters become scale-independent. Thus, while it is possible to obtain those values using different approaches than directly fitting Equation 2.12, that work – and especially the comparison with even smaller box-sizes presented here – would be incongruent with the accuracy carried out in the rest of this paper. For now, we will leave it as a separate topic to be carried out in a future work.

2.5 Splicing

In the previous two sections we have confirmed and quantified both the box-size and resolution requirements for achieving percent-level accuracy for precision Lyα forestcosmology studies. Although possible with today's high-performance computing facilities, as demonstrated with our L80_N4096 simulation, currently performing a large number of such simulations is impossible. In the past, even a single simulation with that dynamical range was impossible. One technique used to compensate for the lack of dynamical range is splicing, first introduced by McDonald (2003), and most recently employed in Borde et al. (2014). Here we will assess the accuracy of splicing on a 1D flux power spectrum. For completeness, we will first briefly review the method itself.

The mechanics behind splicing is to run three simulations, each lacking sufficient dynamic range, and combine them into a result that accurately represents a single full dynamic range simulation. One runs a simulation with enough large-scale power (i.e. a big enough box), but with too coarse a resolution, and another simulation where the box is known to be too small but with good resolution. Finally a simulation is carried out where both resolution and box size are insufficient, the resolution set to the same as in first run, with the box size the same as in the second. The idea is then to use two small-box runs to capture the effect of coarse resolution on the power spectrum, and two runs with coarse resolution to correct for the missing modes in the small-box simulations. Here we will attempt to splice the result of our 4096^3 80 h^{-1}Mpc run, which yields percent accurate results as shown via the box-size and resolution convergence tests in the two previous sections. We will thus splice the L80_N1024, L20_N1024 and L20_N256 runs, and compare the result to L80_N4096. Mathematically expressed, in the regime $k < k_{\mathrm{min},20}$, where $k_{\mathrm{min},20} = 2\pi/(20\,h^{-1}\mathrm{Mpc})$ the flux power is given as:

$$P(k) = P_{\mathrm{L80\_N1024}}(k)\frac{P_{\mathrm{L20\_N1024}}(k_{\mathrm{min},20})}{P_{\mathrm{L20\_N256}}(k_{\mathrm{min},20})} \ , \tag{2.13}$$

in the range $k_{\mathrm{min},20} < k < k_{\mathrm{Nyq},80}/4$ where $k_{\mathrm{Nyq},80} = 1024\pi/80\,h^{-1}\mathrm{Mpc}$ is

$$P(k) = P_{\mathrm{L80\_N1024}}(k)\frac{P_{\mathrm{L20\_N1024}}(k)}{P_{\mathrm{L20\_N256}}(k)} \ , \tag{2.14}$$

and for $k > k_{\mathrm{Nyq},80}/4$ it is

$$P(k) = P_{\mathrm{L20\_N1024}}(k)\frac{P_{\mathrm{L80\_N1024}}(k_{\mathrm{Nyq},80})}{P_{\mathrm{L20\_N256}}(k_{\mathrm{Nyq},80})} \ . \tag{2.15}$$

In Figure 2.20 we show the results of splicing the flux power spectrum at 3 different redshifts. The accuracy of splicing is similar at all redshifts, and is mostly in the ∼10 per cent range. That is in agreement with the accuracy estimated by McDonald (2003), but noticeably above the 2 per cent accuracy claimed by Borde et al. (2014). A possible reason for this discrepancy is the fact that Borde et al. (2014) tested the splicing method on a non-converged simulation (1024^3 particles in 100 h^{-1}Mpc box).

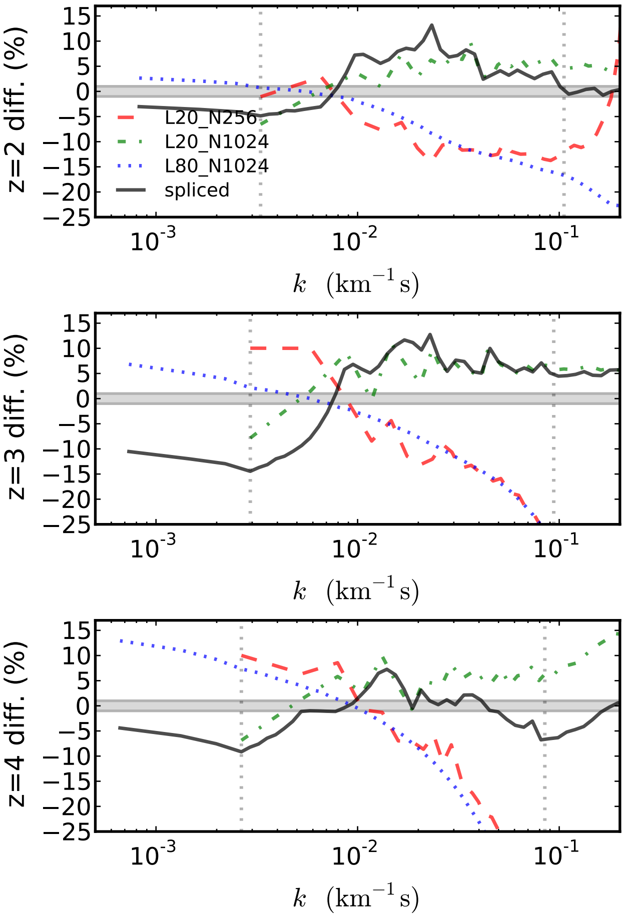

Figure 2.20: Comparison of a spliced 1D power spectrum and the actual one in a high-resolution, large-box simulation. From top to bottom we have redshift $z = 2$, $z = 3$, and $z = 4$. In addition to the spliced run, we show the ratios of power spectra for each of the three runs used for splicing. All ratios are taken with respect to L80_N4096 run.

2.6 Rescaling optical depths

Up to this point we have ignored the common practice of rescaling the simulated optical depths. In most papers presenting the results of Lyα forest predictions from optically-thin simulations, authors multiply the optical depth in each pixel by some factor A such that the simulated mean flux matches the observed mean flux at the same redshift. This is easily done with any root finding method on $\langle \exp(-A\tau) \rangle - \langle F \rangle_{\mathrm{obs}}$ and converges fairly quickly. This fix is well-justified considering how poorly constrained the amplitude, shape, and evolution of the ionizing background are. When we rescale optical depths, it is understood that this is roughly equivalent to adjusting the specific intensity of the UV background used in the simulation. Changes in the photoionization rate, Γ, in the simulation will most directly affect the HI density while sub-dominant changes will come from photoheating rates. The change in photoheating rate affects the temperature and pressure support of the gas at times when hydrogen or helium are not fully ionized.

In order to test the effect of rescaling optical depths, we first tried taking two runs with different UV backgrounds (and otherwise the same input parameters) and rescaling one to the mean flux of the other. We used a run with the Haardt & Madau (1996) UVB (labeled L10_N1024_HM96) and a run with the Faucher-Giguère et al. (2009) UVB (labeled L10_N1024_FG09). One concern with starting from different UV backgrounds is that they can result in different ρ_{b}-T relations, which would leave differences in the flux statistics no matter how the rescaling is done. The HM96 and FG09 UV backgrounds do result in slightly different ρ_{b}-T relations, with very similar slopes but differing T_0 values. In the case of HM96, we fit $T_0 = 9.0 \times 10^3$ K and $\gamma = 1.55$, and in the case of FG09, we fit $T_0 = 1.1 \times 10^4$ K and $\gamma = 1.55$ at redshift $z = 2$. While this is not a significant difference, a temperature difference like this should show up in the flux power spectrum, for instance, as a different thermal cutoff. More importantly, while the two UV backgrounds result in similar instantenious ρ_{b}-T relation, the two have significantly different thermal histories: Haardt & Madau (1996) reionizes hydrogen at $z \approx 6$, while Faucher-Giguère et al. (2009) has this occuring at $z \approx 12$. This will result in two UVBs which have a different filtering scale, even if T_0 and γ at a given redshift are the same. At $z = 2$, the HM96 run has a mean flux $\langle F \rangle = 0.8117$ and the FG09 run has a mean flux $\langle F \rangle = 0.7749$. Rescaling the HM run to the FG mean flux requires $A = 1.403$ (or conversely rescaling the FG run to the HM mean flux requires $A = 0.7138$). We show an example skewer in the top panel of Figure 2.21. The sample spectrum shows that between the HM96 and FG09 runs the flux in regions of high transmission is similar, but the absorption features are generally deeper in FG09 primarily due to the lower photoionization rate. It is difficult to tell if either run has broader features by looking at just a few lines, but overall we found that the HM96 run does have noticeably broader lines. We also show the spectra after rescaling to the other mean flux. It is reassuring to know that while the correction is an average over the entire box, individual features agree well enough that the correction also works well for individual lines.

The flux PDF and flux 1D power from these runs and their rescaled versions are also shown in Figure 2.21. In both statistics the HM96 and FG09 results differ by about 30 per

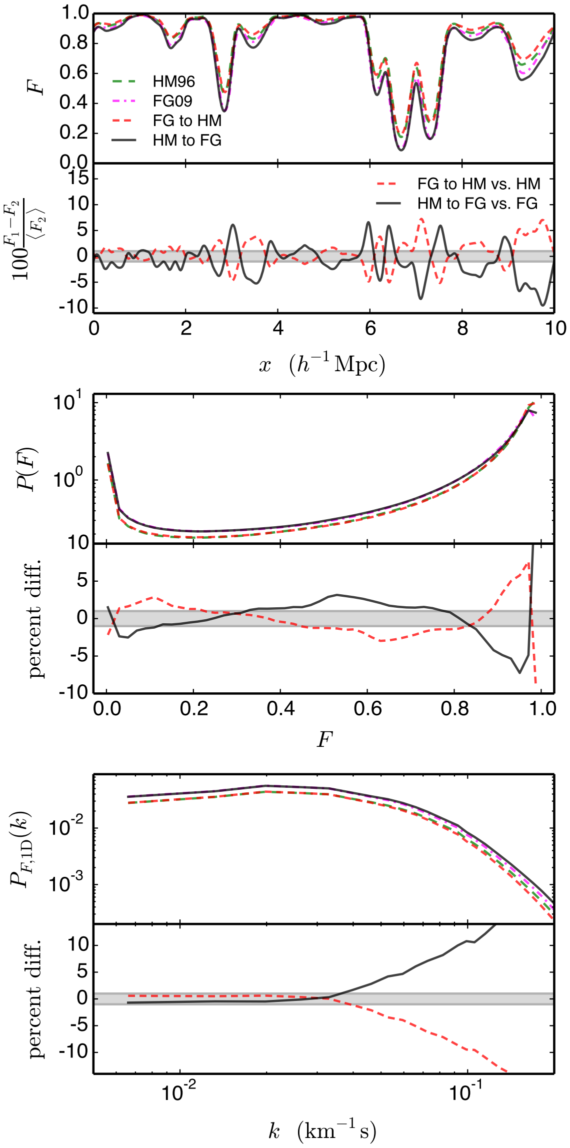

Figure 2.21: *Top:* Flux statistics from the L10_N1024_HM96 and L10_N1024_FG09 runs at $z = 2$. At this redshift, the L10_N1024_HM96 run has $\langle F\rangle = 0.8117$, and the L10_N1024_FG09 run has $\langle F\rangle = 0.7749$. Also shown are the flux statistics of these runs scaled to the mean flux of the other. *Middle:* the flux PDF and percent difference compared to the result with the same mean flux. *Bottom:* the flux 1D power and percent difference compared to the result with the same mean flux.

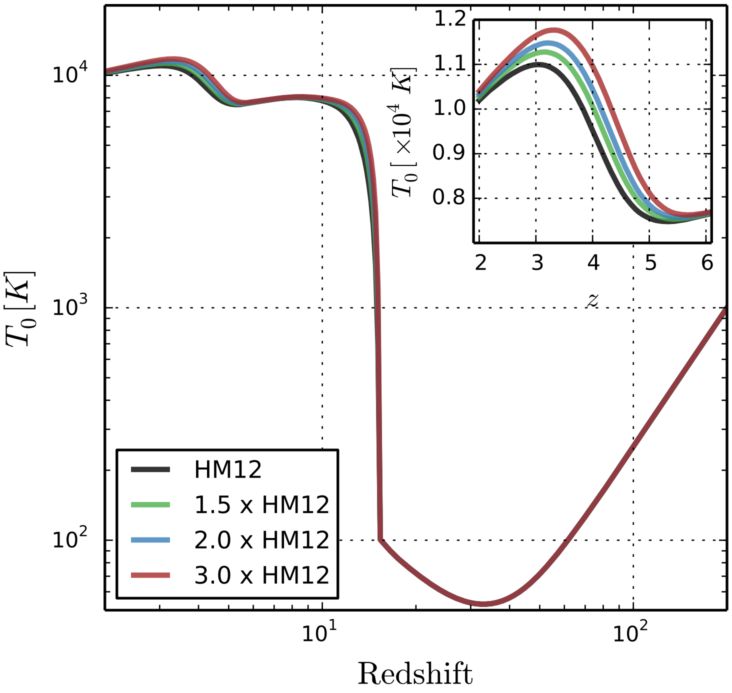

Figure 2.22: The evolution of T_0 for a simulation with Haardt & Madau (2012) UV background (black line), and simulations where photoionization and photoheating rates for all ionic species have been multiplied by the same constant: 1.5 (green line), 2 (blue line), and 3 (red line). The inset panel shows the evolution of T_0 over $2 \leq z \leq 6$.

cent, but the process of rescaling to the other run's mean flux brings them to within several percent. The remaining differences in the flux PDF are not straightforward. The rescaled FG09 run has more pixels at low F, fewer pixels at intermediate F, and more pixels at high F compared to the HM96 result. It appears that the rescaled FG09 rises faster than the HM96 result at $F = 0$ and $F = 1$. In the 1D flux power spectra, the rescaled versions agree very well at large scales. The rescaled FG09 result has slightly more power than the HM96 result, but it is within 1 per cent. On scales below $k > 0.4$ h Mpc^{-1}, the slopes of the rescaled versions start to diverge significantly. This is due to the differing T_0 for each, resulting in a different thermal cutoff. Overall, the rescaling process works remarkably well at removing differences from different UV backgrounds, although one should be careful with results that are sensitive to the ρ_b-T relation.

We also made another test of the UVB rescaling, by running simulations where the photoionization and photoheating rates for all ionic species have been multiplied by the same constant. Here we use the Haardt & Madau (2012) rates and, since we multiply all of them by the same factor, the spectral shape of the original UVB is preserved and only the amplitude changes. In Figure 2.22 we show the IGM temperature evolution. As expected modulating the amplitude of the UVB affects the temperature only when a species is not

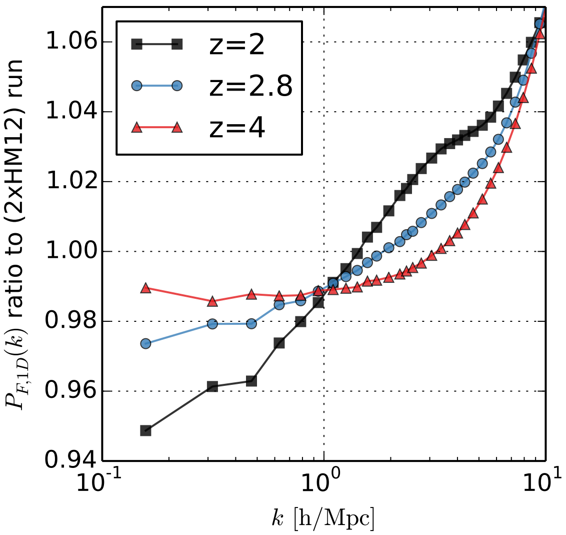

Figure 2.23: The 1D power spectrum from a run using the original Haardt & Madau (2012) UVB rates where the mean flux is rescaled to match that of the run done with doubled UVB rates (2xHM12). We show here the ratio to the actual 2xHM12 run, at 3 different redshifts: $z = 2$ (black, squares), $z = 2.8$ (blue, circles), and $z = 4$ (red, triangles).

fully ionized. Changes in the hydrogen leave no imprints on T_0 at observable redshifts, but the same change in helium photoheating rates do change the temperature due to its partial ionization.

We focus on two runs, one using the original Haardt & Madau (2012) rates, and one where the multiplying factor for all photo-rates is 2, approximately a value needed to recover the observed mean flux in optically thin simulations with this UVB. We compare the run done with doubled photo-rates, labeled 2xHM12, with a flux-rescaled run performed with the original HM12 rates. In other words, the rescaling is done by simply finding the factor A which will bring the mean flux of the original run to that of 2xHM12, a procedure which ignores the differences in the instantenious gas temperature and prior temperature history. The rescaling factor A we recover is close to, but not exactly equal to 0.5, and it shows the tendency to decrease with increasing redshift. For example, it is 1 per cent higher at $z = 2$ ($A = 0.505$), than $A = 0.5$ at $z = 2.4$, and decreases to $A = 0.479$ at $z = 4$. The results of these rescalings are shown in solid lines in Figure 2.23 for 3 different redshifts. The difference on the 1D power spectrum between rescaling the mean flux and actually running the full evolution with different rates is a few per cent, but it is present on all scales. We have also checked the flux PDF, and found smaller differences of approximately 1 per cent. One should note that the difference is smaller at higher redshift, $z = 4$, and is the greatest at $z = 2$. This means that the effect of different temperature evolutions is more significant for the IGM gas at mild overdensities, and not very significant for underdense gas in void regions.

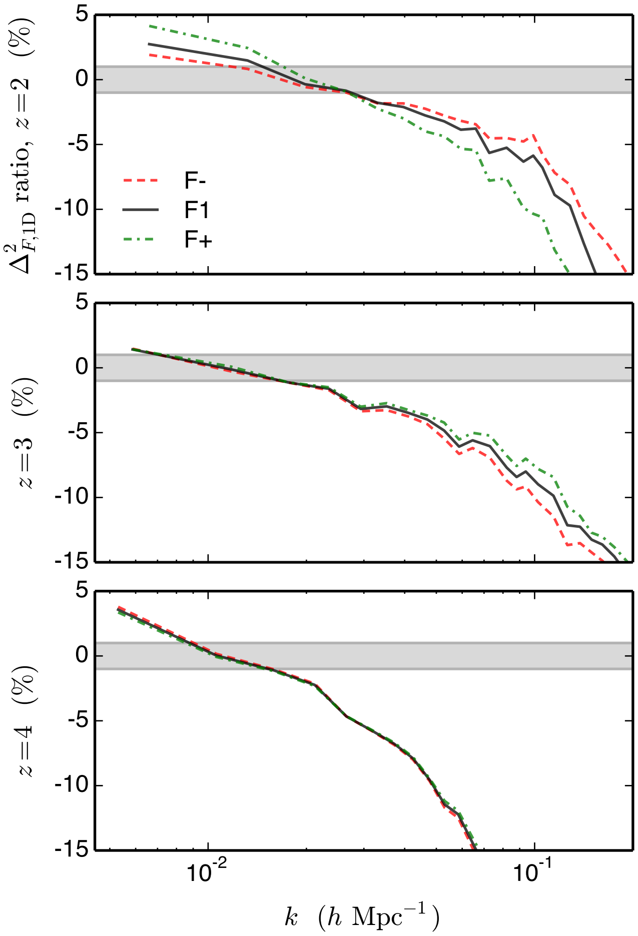

Figure 2.24: The percent difference in the 1D flux power spectra between L10_N256 and L10_N1024 at $z = 2$, 3, and 4, where the optical depths have been rescaled. F- is rescaled to 10 per cent smaller than the L10_N1024 mean flux, F1 is rescaled to the same mean flux, and F+ is rescaled to 10 per cent larger. This shows how the convergence rate depends on the mean flux. At $z = 2$, the F- power converges fastest, but at $z = 3$, the F+ power converges fastest. At $z = 4$, the convergence rate is not noticeably affected by 10 per cent changes in the mean flux.

The final test we performed was to take the $10\,h^{-1}$Mpc resolution series runs and test the convergence rate of the flux statistics when changing the mean flux. We compare the flux statistics computed from the same optical depths, but with three values of the rescaling constant based on the mean flux of the L10_N1024 simulation: one 10 per cent larger, one equal, and one 10 per cent smaller, respectively labeled (F+, F1, and F-). In Figure 2.24, we show the ratio of 1D flux power spectra between the L10_N256 and L10_N1024 runs, computed from the rescaled fluxes. From top to bottom, we show the result for $z = 2$, 3, and 4. At $z = 2$, we see that the 1D power converges faster for a lower mean flux. The resolution error in the 1D flux power shows a characteristic slope difference (the lower resolution result has a more negative slope), and at $z = 2$, the slope difference is smallest for the lower mean flux rescaling, and largest for the higher mean flux rescaling. Rescaling the flux to a lower mean value means shifting the flux contributing regions to a lower gas density, or equivalently, less non-linear structures, and thus it is easier for these structures in the simulation to be converged. At $z = 3$, we see the opposite trend, that the results converge faster with a higher mean flux rescaling. This indicates a 'sweet spot' mean flux (or similarly, a redshift) where it is easiest to resolve Lyα foreststructures. When the mean flux is very high, the forest probes higher densities closer to halos which are harder to resolve. Similarly, when the mean flux is very low, the forest has significant sensitivity to the very underdense regions. Although we are examining purely numerical effects here, note that these conclusions also translate to the question of the importance of additional physics in simulations. From what we have presented in this section, one would clearly expect galactic outflows, AGN feedback, and other processes originating within galaxies to matter much more at e.g. redshift $z = 2$ than $z = 3$. At $z = 4$, we see no difference in the convergence rate with the different mean flux rescalings. At this high of a redshift, the mean flux is low and a rescaling of only ±10 per cent does not significantly affect the density range contributing to the forest (see the red bands in Figure 2.7).

Finally, another issue with the practice of rescaling to the observed mean flux is that few current Lyα forest simulations are converged to the percent level in $\langle F \rangle$ at high redshift. As shown previously, even simulations with a resolution of $\sim 40\,h^{-1}$kpc are not converged to a percent in the mean flux for $z > 3$. Taking a simulation of insufficient resolution and performing a mean flux rescaling will result in the wrong correction. Under-resolving IGM structures results in a mean flux lower than it should be, so a rescaling to a higher mean flux, for instance, will require a smaller rescaling factor A than what would be needed for a higher-resolution run.

2.7 Small-scale statistics

2.7.1 Line statistics

The Lyα forest is classified as systems with $N_{\rm HI} < 10^{17}$ cm^{-2}, known for sitting in the linear portion of the curve of growth. This makes it straightforward to fit the Lyα

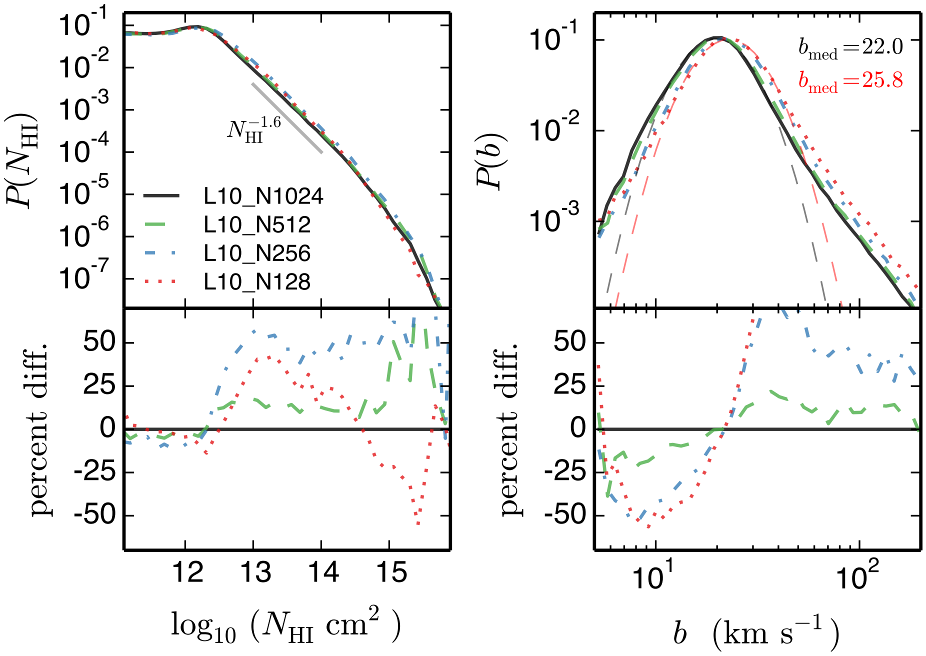

Figure 2.25: The distributions of the line HI column density (left) and the Doppler parameter (right) in the resolution series of the $10\,h^{-1}$Mpc boxes at $z = 2$. The percent difference relative to the L10_N1024 simulation is shown in the bottom panels. In the $P(N_{\rm HI})$ panel, the gray line illustrates the power law slope $\propto N_{\rm HI}^{-1.6}$. In the $P(b)$ panel, the thin dashed red and black lines show lognormal distribution fits to the L10_N128 and L10_N1024 results, and the red and black text gives the corresponding median b-parameter values (also the peak probability). The simulated b-parameter distributions are more skewed than lognormal, but a lognormal fits the core of the simulated distributions well.

forestlines with Doppler profiles. Each line fit provides the column density $N_{\rm HI}$ and Doppler parameter b of the underlying system. However, this neglects the issues of significant line blending in the forest, as well as line broadening dominated by Hubble broadening rather than thermal broadening. Meiksin et al. (2010), for instance, mention that line shapes in simulated spectra ignoring peculiar velocities are qualitatively different than those in the full spectra. This indicates that the $N_{\rm HI}$ and b derived from line fitting may not correspond to the actual column density and temperature distributions of the gas that makes up the forest. However, line parameter distributions are a sensitive measure of line shapes in the forest. Additionally, interpreting line parameter statistics is fairly straightforward — the column density is a proxy for equivalent width and the Doppler parameter is the line width.

In this work, we generate spectra with a fixed resolution of $\Delta v_{\rm pix} = 1\,{\rm km\,s^{-1}}$ to avoid possible issues of the fits depending on spectral resolution and the assumed pixel SNR. We have chosen this resolution because it is sufficiently smaller than what we expect for the minimum line width of gas at $T = 10^3$ K, which corresponds $b = 4\,{\rm km\,s^{-1}}$ for hydrogen.

We want to reduce statistical uncertainty in the distributions as much as possible, which means having many lines in each bin. The line compilation in Haardt & Madau (2012) (their section 3), provides dN/dz of systems above a given $N_{\rm H\,I}$. At $z = 3$, $dN/dz = 24.2$ for $N_{\rm H\,I} > 10^{15}$ cm^{-2}. The corresponding path length required to find 10^4 absorbers (for 1 per cent statistics) is 2.9×10^5 h^{-1}Mpc, or a total spectral length of about $3.1 \times 10^7\,{\rm km\,s^{-1}}$. The L10_N128 simulation is actually smaller than this, so we use the entire box in that case. For all other runs, we evenly distribute the skewers throughout the volume up to the required path length. We use the SPECFIT code (first used in Meiksin et al. 2001) to perform the Voigt line fitting of our simulated spectra. For each spectrum, the code splits the spectrum into regions separated by a threshold value $\tau > \tau_{\rm min}$. In these absorption regions, SPECFIT uses first and second derivatives of the flux to identify line centers and then performs a χ^2 minimization of the line parameters.

We show the effect of simulation resolution on the line parameter distributions in Figure 2.25. We show the probability distribution function of the line column density in the range $11 < \log_{10}(N_{\rm H\,I}\,{\rm cm}^2) < 16$ and the Doppler parameter in the range $5 < b/({\rm km\,s^{-1}}) < 200$ and the per cent difference to the L10_N1024 results. The column density distribution is relatively flat for $N_{\rm H\,I} < 10^{12.5}\,{\rm cm}^{-2}$, and then turns over to a power law $dN/dN_{\rm H\,I} \propto N_{\rm H\,I}^{-1.6}$, where the slope of the power law depends on the UV background. The annotated gray line gives an example of the power law slope. Qualitatively, the different resolutions agree well. The L10_N128 and L10_N256 runs do not peak as much at $N_{\rm H\,I} < 10^{12.5}$ cm^{-2}, and turn over more slowly, resulting in an excess probability of lines in the $12.5 < \log_{10}(N_{\rm H\,I}\,{\rm cm}^2) < 14$ range. This same trend is present in the L10_N512, but it is less significant. The lowest resolution run, L10_N128, shows a deficit of high column density lines $N_{\rm H\,I} > 10^{14}$ cm^{-2} compared to the other runs. The Doppler parameter distribution is close to lognormal with a peak around $b = 20\,{\rm km\,s^{-1}}$. The Doppler parameters distributions show a much clearer convergence pattern. The overall shape of the Doppler parameter distribution does not appear to change with resolution, but the peak b value decreases with increasing resolution. This holds together with our qualitative picture of the resolution study – the lower-resolution runs are like artificially smoothed higher-resolution runs, so the resulting absorption lines are broader as well. We also show two fits of the lognormal distribution $P(b) = A/(b\sigma)\exp[-(\log b - \log b_{\rm med})^2/(2\sigma^2)]$, where A, $b_{\rm med}$, and σ are the free parameters. The red thin dashed line is the fit to the L10_N128 and the black thin dashed line is the fit to the L10_N1024 result. From the lowest resolution to the highest resolution result, the median b value changes from 25.8 km/s to 22.0 km/s, and the corresponding temperatures are 3×10^4 K and 4×10^4 K. This is an important consideration for studies using small-scale statistics to infer the temperature of the IGM. If we were to use the low-resolution runs to infer the IGM temperature, we would be biased to lower temperatures than results based on higher-resolution runs (the fit T_0 values are essentially the same in these runs).

We show the effect of simulation box size on the line parameter distributions in Figure 2.26. Compared to the resolution series, the box size has little effect on the line fits. The different UVB in the box series simulations results in a slightly steeper $N_{\rm H\,I}$ distribution power law, shown with the gray line matching $\propto N_{\rm H\,I}^{-1.7}$. The column density distributions

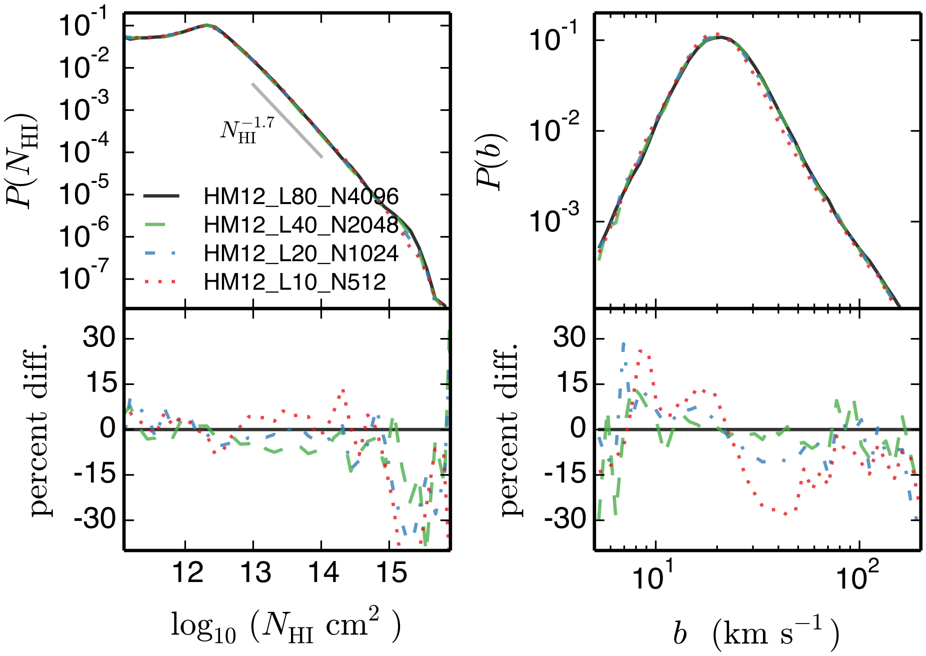

Figure 2.26: The distributions of the line H I column density (left) and the Doppler parameter (right) in the box size series with a grid scale of $20\,h^{-1}$kpc at $z = 2$. The percent difference relative to the HM12_L80_N4096 simulation is shown in the bottom panels. In the $P(N_{\rm HI})$ panel, the gray line illustrates the power law slope $\propto N_{\rm HI}^{-1.7}$.

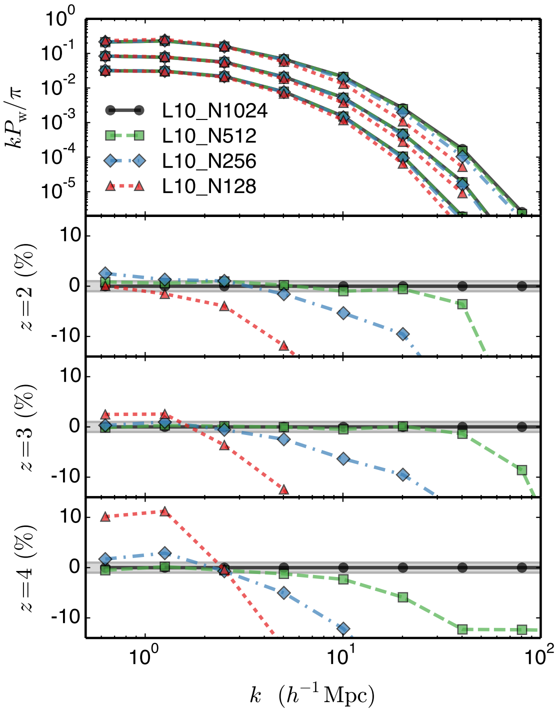

Figure 2.27: Convergence of the flux discrete wavelet transform power with respect to physical resolution in $10\,h^{-1}$Mpc box.

across box size are very close to each other, and differences appear to be in the statistical noise. More significant differences appear above $N_{\mathrm{HI}} > 10^{15}$ cm^{-2}, however, this is also mostly due to the rarity of well-fit high column density systems. The box size has a clearer effect on the Doppler parameter distribution, although it is still much smaller than the resolution effect. Again, the distribution shape across box size is essentially the same, but the peak position increases with increasing box size. This is hard to see in the top $P(b)$ panel, but in the bottom difference panel, we see the curves flattening out around $b = 20$ km/s with increasing box size.

2.7.2 Wavelet statistics

In addition to line fitting, we also performed a wavelet power analysis of our spectra. Wavelets have previously been applied to the Lyα forest as a means of objectively measuring the line widths, primarily in order to probe the IGM temperature. Meiksin (2000) introduced wavelets as a tool for the Lyα forestas a means of data compression and a measure of small-scale power. Zaldarriaga (2002) extended the use of wavelets to search for spatially localized line-width (temperature) fluctuations, an analysis recently repeated with a slightly different use of wavelets in Lidz et al. (2010) and Garzilli et al. (2012). Theuns et al. (2002) used wavelets to search for temperature fluctuations in the IGM associated with HeII reionization.

Wavelet basis functions are orthogonal and complete, providing well-defined transforms into wavelet coefficients and back. More importantly, wavelets are localized in real and Fourier space. The discrete wavelet transform (DWT) is a decomposition of some signal f into discretized wavelet bases $\psi_{jk}(x)$, where we use j as the level (sometimes called stretch) of the wavelet and k as the shift (or position). The DWT provides wavelet coefficients w_{jk} such that $f(x) = \sum_{j,k} w_{jk}\psi_{jk}(x)$, and in this case we are transforming the flux fluctuations δ_F along the line of sight. We use the Daubechies 20 coefficient wavelet, which is the most common choice for wavelet analysis. With the wavelet coefficients in hand, we can compute the wavelet power spectrum P_{w} as average of the squares of the coefficients, just as one would with Fourier coefficients.

$$P_{\mathrm{w}}(k_j) = L\langle w_{jk}^2\rangle \tag{2.16}$$

where the average is taken over all of the shifts k for the level j. The L factor is included to match our previous Fourier convention and the dimensionless wavelet power spectrum is $k_j P_{\mathrm{w}}/\pi$. We associate the level j to the mode $k_j = 2\pi/L2^{j-1}$.

Figure 2.27 shows the impact of spatial resolution on the wavelet power. As expected, we see very similar behavior to the flux power spectrum, namely the percent level agreement between the 20 h^{-1}kpc (L10_N512) and 10 h^{-1}kpc (L10_N1024) runs. As with the power spectrum we see that inadequate resolution to capture the fluctuations responsible for the rise of the Lyα forestproduces an error on *all scales*, not only scales below the resolution limit. We have checked that the box-size behavior is also very similar to that seen in $P_F(k)$, Figure 2.18.

2.8 Conclusions

We have investigated simulated Lyα forest statistics over the redshift range $2 \leq z \leq 4$. A large suite of simulations covering box sizes of 10-80 h^{-1}Mpc, and resolutions 10-78 h^{-1}kpc have enabled us to understand the numerical requirements for future sets of simulations aimed at constraining cosmological parameters using the Lyα forest. While we model gas dynamics using a very accurate finite-volume numerical method, the additional physics which enters as a source (heating) term in the energy equations as well as those used to calculate the ionization state of a primordial chemistry gas is accounted for in a more approximative way. In our optically thin simulations, the gas is rapidly ionized by the assumed UVB at high redshift, and in a short time changes its ionization fraction by order unity – a model of sudden reionization. We neglect the effects of self-shielding in high over-density regions as those do not contribute to the Lyα forest signal. The temperature boost during the sudden reionization depends only on the shape of the spectrum of ionizing radiation, and in fact one expects a range of spectral shapes responsible for ionizing different gas elements. Modeling the details of temperature evolution during and immediately after reionization requires full radiative transfer simulations and is beyond the scope of our needs here, as the thermal memory of the IGM gradually fades after the epoch of reionization (mostly due to Compton cooling, see Hui & Gnedin 1997).

As the IGM fills the simulation box, it is fruitless to try to resolve it with adaptive refinement; similarly, as a large portion of the Lyα forestsignal arises from near mean and under-dense regions (especially at at higher redshifts), Lagrangian methods do not offer any advantage over a fixed grid PDE solver. Needless to say the fixed grid approach is computationally expensive, especially in the 3D case presented here, and thus it is important to determine the minimal resolution requirements needed to bring our simulations to 1 per cent accuracy. We have explored this in Section 2.3 arriving at the conclusion that $\sim 20\,h^{-1}$kpc resolution is good enough over the relevant redshift ranges we wish to consider for the Lyα forest. While in places — for example the flux PDF — a coarser resolution would suffice, the study of 1D power spectra brings with it a more stringent requirement. While we explored resolution convergence, we were also able to show that NYX behaves well on this multi-dimensional, multi-physics problem, exhibiting the expected second order convergence. As shown in Section 2.3, this opens the possibility of achieving a desired accuracy at reduced cost, via extrapolation of lower-resolution runs.

We also explored other numerical artifacts which can easily mask a physical process in the IGM and/or spoil the quality of cosmological predictions. After finding an appropriate resolution, we have explored the effects of finite box-size, i.e. missing modes in Lyα forest-simulations. By running simulations with all cosmological and numerical parameters but the box-size fixed, we were able to show that 40-80 h^{-1}Mpc boxes are large enough for all relevant statistics including the 3D power spectrum in redshift space, i.e. $P(k, \mu)$. For the first time, we were able to perform a Lyα forestsimulation fulfilling both the resolution requirement set by the Jeans / filtering scale and the box-size requirement set by large-scale flows. That enabled us to examine the accuracy of splicing 1D power spectra, a common approach in the case when a full range of simulations are not feasible (McDonald 2003; Borde et al. 2014). We show that accuracy of splicing is only 5-10 per cent, and that the error has a clear scale-dependence.

As the UV background is the largest uncertainty in Lyα forest simulations, it is very common that a cosmological model is evolved with a UVB model, with the resulting optical depth field subsequently rescaled to match desired, observed mean flux value. In Section 2.6, we examined the effects of changing the UVB and rescaling the mean flux. We found that while qualitatively it is possible to change the UVB *post festum* in analysis, one suffers a few percent effect in the UVB rates propagating back into the gas evolution which is most visible in the flux PDF. We note that the 1D $P(k)$ appears largely insensitive to such rescaling for $k \lesssim 3 \times 10^{-2}$ km^{-1}s, the k-range relevant for current observational data.

We find that resolution requirements for convergence on the line statistics are much more demanding than on the flux statistics. A resolution of 20 h^{-1}kpc, adequate for reproducing the flux statistics to 1 per cent accuracy as shown in Section 2.3, recovers the distribution of neutral hydrogen column densities only at $\sim$ 10 per cent accuracy in the range $12.5 \lesssim \log_{10}(N_{\rm HI}\,{\rm cm}^2) \lesssim 15$, and even worse at higher column densities. Similarly, the Doppler parameter (b) distribution is converged at the same level, with the peak value decreasing

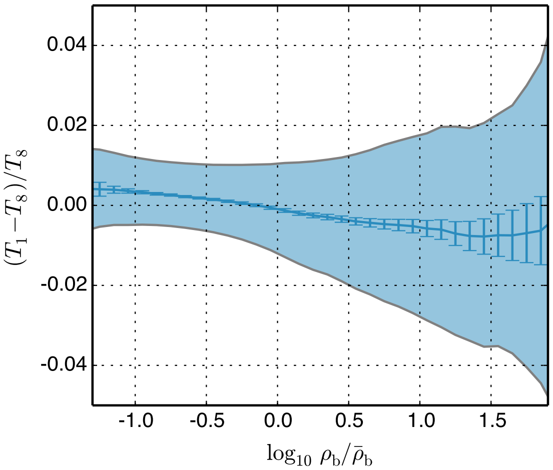

Figure 2.28: Change in the gas temperature in a standard run with 1 particle per cell (T_1), and an 'overloaded' run with with 8 particles per cell on average (T_8). The line shows the difference in the median temperatures and the fill shows the normalized median absolute deviation in temperatures. The median deviation is roughly symmetric and the resulting effect on the Lyα forest statistics is negligible. In addition, we show statistical (Poisson) errorbars in each density bin.

with increasing resolution. This is not directly relevant to modern cosmological studies which do not rely on individual line fitting, but it is important for certain studies of the IGM.

Finally, we have explored the effects that finite sampling of the dark matter particles has on the statistics of the Lyα forest. The expectation is that artificial gravitational collisionality between dark matter particles and gas increases the gas temperature, an effect that should be strongest in void regions. While we indeed notice, on average, temperature increases in void regions, the effect is minor in today's simulations even when using 1 particle per cell and CIC particle deposition. The reason for this is the small particle mass in Lyα forestsimulations and the presence of radiative cooling, which efficiently removes excess heat.

The advent of high performance computing power and scalable numerical algorithms as employed in NYX allows us to make accurate predictions for the Lyα flux statistics, one of the most promising tools for precision cosmology measurements in the redshift range $2 \lesssim z \lesssim 4$. The direct simulation approach, using no *ad hoc* physical assumptions, is possible for this problem. We have made a concentrated effort here in understanding the Lyα forest signal in optically-thin hydrodynamical simulations, and quantifying the accuracy of such simulations with respect to numerous numerical effects. We are by no means the first to attempt this (indeed they go back to at least Cen 1992), however we have been able to consolidate and improve upon earlier studies using modern simulations with the goal of percent-level numerical precision, a level of accuracy required for carrying out precision cosmology over the next decade.

We note that full-range, 4096^3, hydrodynamical simulations like the one presented in this chapter are still computationally demanding today, but will be a fairly typical in coming years. Before one commits to running many such simulations, it is imperative that the precision which can be obtained be understood. Convergence testing is an invaluable tool here, as analytically soluble problems are highly artificial in nature, and experience with them does not necessarily translate to real cosmological runs. In this work, we have used the Nyx code. While Nyx is focused on Lyα forest simulations, the results presented here should be directly applicable to IGM simulations performed by the Enzo, Flash, and Ramses codes. In addition, there are several lessons applicable to a large extent to the Gadget and Arepo codes as well.

Acknowledgments

This work was previously published as (Lukić et al. 2015). We would like to thank Anže Slosar, Pat McDonald, and Matt McQuinn for many useful discussions, and are grateful to Francesco Haardt, Piero Madau, and Claude-Andreè Faucher-Giguére for making their UVB rates publicly available. We also acknowledge the helpful review of an original version of the manuscript by an anonymous referee. ZL and CWS acknowledge the hospitality of Triple Rock where many concepts were refined. This work was supported in part by the Scientific Discovery through Advanced Computing (SciDAC) program funded by the US Department of Energy Office of Advanced Scientific Computing Research and the Office of High Energy Physics.

Chapter 3

The impact of hydrodynamic methods on Lyα forest flux statistics

In structure formation studies, exactly solvable problems either have a reduced dimensionality or are highly symmetric; therefore the accuracy of different codes on such problems (as typically reported in ‘code papers’) is not indicative of an agreement they will achieve when solving realistic cosmological evolution involving complicated physics. Performing convergence tests on an actual problem is one of the most valuable tools for establishing realistic accuracy, but the codes can also converge on an incorrect solution (e.g. due to a constant source of error). For this reason, comparing different codes which employ different discrete algorithms to solve the same physical equations is important for establishing a level of confidence in reported results. In this chapter, I examine the current state of the art in Lyα forest flux statistics as reproduced by two different optically-thin hydrodynamic simulation codes.

Similar comparisons have been done in the past, most notably in Regan et al. (2007), where the authors compared the GADGET-2 and ENZO codes. The authors reported agreement at the 10 per cent level. Due to advances in both computational methods and computing power, it is again of interest to examine how well different codes agree using improved simulations and to consider implications for measurements of cosmological parameters. The version of GADGET we use in this work, GADGET-3 we use here has the same SPH algorithm as GADGET-2, but significantly improved scalability, allowing us to push the method to higher resolutions. In addition to proven scalability to ∼100,000 cores, NYX also has an improved hydrodynamical scheme over ENZO as it uses a dimensionally unsplit hydrodynamics scheme. In Bird et al. (2013), the authors compared the GADGET-3 and AREPO codes. These two codes share the same pedigree and are very similar to each other besides the hydrodynamics solver. While Bird et al. (2013) focuses on the damped Lyα (DLA) and Lyman-limit systems (LLS), they note that typical low-column density systems that are mostly probed by the Lyα forest agree at the per cent level. This is complementary to our work, where we compare substantially different codes, representative of virtually all results reported in recent literature on this topic.

We note that more general comparisons of SPH and Eulerian techniques were performed in Agertz et al. (2007); Mitchell et al. (2009). Both papers argue that the main differences between SPH and Eulerian codes (in an astrophysical context) can be tracked down to the inability of SPH to resolve fluid instabilities such as Kelvin-Helmholtz or Rayleigh-Taylor and their consequential effects on mixing of multi-phase media. While these processes are important in many astrophysical contexts, it is unlikely that they influence the state of the gas making up the Lyα forest signal. The intergalactic gas that creates the Lyα forest signal is in a single phase and smooth, besides accretion shocks onto higher density structures. While differences in the shock structure could create differences down the line in the Lyα flux, we will show that we do not find significant differences.

3.1 Simulations

In principle, it is relatively simple to directly simulate the Lyα forest. The gas that creates the significant majority of the absorption features, known as the diffuse IGM, is at moderate overdensities, $\delta = \rho/\bar{\rho} < 10$. Numerically, these densities are easier to capture than higher overdensities. The evolution of this gas depends on gravity and pressure forces, and some simple radiative processes. Since the gas is very close to primordial composition, the background ionizing radiation is relatively spatially uniform, and almost all of the gas (by volume) is optically thin to the radiation, the required simulation machinery is simple indeed. Any cosmological hydrodynamics code modeling the evolution dark matter and gas with uniform radiative heating and cooling should be able to adequately capture the diffuse IGM. See Meiksin (2009) for a more thorough review of the relevant physics. We note here that there are secondary effects such as inhomogeneity of the ionizing background, heating from helium reionization, and galactic feedback, but they are beyond the scope of this work. We aim to use the simplest physics possible for the problem at hand, and test the level of agreement in this most basic case.

3.1.1 Nyx

NYX is a cosmological Eulerian hydrodynamics code with Adaptive Mesh Refinement capabilities, built on the BOXLIB AMR framework. We summarize the algorithms here, but direct more interested readers to the NYX code paper, Almgren et al. (2013). Nyx implements an Eulerian formulation of hydrodynamics, evolving 5 conserved variables on an adaptive mesh. In addition to the conserved quantities, Nyx also solves for the evolution of the internal energy density, implementing the dual-energy formalism (see Bryan et al. 1995). In our simulations, the dual-energy formalism will play a much smaller role than in adiabatic cosmological simulations, as the UV background imposes a relatively high temperature floor. The hydrodynamics solver is a higher-order Godunov scheme, with a piecewise parabolic method (PPM) reconstruction, and uses unsplit fluxes with full corner-coupling (Colella 1990). The Riemann solver in Nyx iteratively solves the Riemann problem using a two-shock approximation (Colella & Glaz 1985). Nyx models dark matter as N-body particles updated with a standard Particle-Mesh (PM) method with symplectic timestepping. Particles are deposited with the cloud-in-cell (CIC) kernel to the same grid where the baryon quantities are defined. The Poisson equation is solved using a multigrid method. Finally, Nyx finds cell-centered gravitational acceleration using 7-point stencil. The gravitational acceleration is coupled to the hydrodynamics solver via source terms in the momentum and energy equations and the dark matter particle positions and momenta are updated with Kick-Drift-Kick stepping. In this work, we do not use the mesh refinement features of Nyx so the root grid of the simulation is the only grid (typically referred to as 'unigrid' simulations in the literature).

3.1.2 Gadget

GADGET-3 is a cosmological Smoothed Particle Hydrodynamics (SPH) code (Springel 2005), which has become a standard community code over the past decade. Gadget implements the conservative entropy formulation of SPH, as described in Springel & Hernquist (2002). In SPH schemes, each particle serves as a Lagrangian fluid element with a given mass, position, velocity, and internal energy. At each step, the particle's density is estimated by smoothing over a prescribed number of neighboring SPH particles determined by an adaptive SPH kernel. Gadget also models dark matter as N-body particles evolved with a Tree-PM scheme, where the long range force is solved using a PM method, as in Nyx. However, the gravitational acceleration algorithm in Gadget also adds a short-range force computed with a Tree multipole method (Barnes & Hut 1986).

For this work, we run Gadget with the `quicklya` option as described in Springel (2005). With quicklya enabled, Gadget uses an extremely simple star formation method, where any SPH particle that reaches a maximum density setting is converted into a star particle. This process appears to have a violent effect near high-density regions, as we will show later in Figure 3.4, likely due to the sudden drop in pressure support. However, we expect these glitches to be isolated to small volumes, where the flux is already saturated, and will therefore have little impact on flux statistics. The upside is that quicklya runs are computationally much cheaper than runs with no star formation, mainly because the size of the timestep is determined by the high gas-density regions and can become restrictively small without use of this option.

3.1.3 Code discussion

In both codes, we use the primordial composition, optically-thin radiative heating and cooling described in Katz et al. (1996). We have checked that both codes use the same atomic rates for radiative recombination, dielectric recombination, and collisional ionization rates given in (Katz et al. 1996) (corresponding to the `oldrates` option in Gadget). We noticed and corrected for a 4 per cent difference in the Compton cooling rate coefficient in the Gadget source due to 1 per cent difference in the assumed CMB temperature (2.7 K vs. 2.73 K), but the effect should be negligible when considering the post-reionization statistics. As pointed out in Lukić et al. (2015), these atomic rate fits are not very accurate, so they should not be used for precision work, but this should of course not affect the results of relative comparison. The equations for optically-thin radiative heating and cooling are completed by assuming a redshift-dependent photoionization rate and heating per ionization. We use the ionizing background tabulation of Faucher-Giguère et al. (2009). The total heating and cooling is included as a source term in the energy conservation equation. In Gadget, the source term is applied at the end of a timestep. In Nyx, the heating/cooling operator is applied in the middle of the step via Strang splitting.

The codes handle high-density regions differently. Simulations that include optically-thin radiative heating and cooling typically suffer from the so-called overcooling problem Balogh et al. (2001). As gas density rises, the cooling becomes more efficient, allowing the gas collapse more, and the whole process can run away. This is the case in Nyx, since there is no star formation or extra heating processes to prevent this collapse. This produces large overdensities, which in turn create significant small-scale power in the baryon density power spectrum. The quicklya star formation in Gadget was introduced to prevent this situation. Gas particles above the quicklya density cutoff and below a temperature cutoff are converted to star particles, and further collapse is minimal. One issue with the quicklya process is that it creates a sudden drop in pressure support whenever it converts gas particles. The star particle formation appears to disrupt areas of high density, as we will show in a later figure. We believe this difference in subgrid physics could be the main cause of differences between the codes.

The two codes also present an interesting mix of sampling trade-offs due to their different numerical designs. For the same grid and number of particles, the effective spatial resolutions of gravitational and pressure forces can differ significantly. Both codes model the gravitational influence of dark matter with collisionless N-body particles. The Lagrangian nature of N-body methods means that resolution (in phase space) is closely tied to density, sampling high-density regions very well, but undersampling low density regions. Structures near and under mean density are heavily affected by the shot noise resulting from the finite particle sampling, and these structures are the most relevant to the Lyα forest signal, especially at high redshift. The sampling of hydrodynamic quantities between the codes differ much more. In the case of Gadget, SPH has the characteristics of any other Lagrangian particle sampling method and high densities are naturally resolved better than low densities, and low density structures are subject to shot noise. In addition, SPH inherently smooths hydrodynamic quantities on the kernel scale. Most Gadget simulations (including those in this work) use an SPH kernel that weights over the 32 neighboring particles. This means that at mean density, the kernel scale (and pressure resolution) is about twice the initial interparticle spacing. In Nyx, the hydrodynamic resolution is tied to the grid scale, which is fixed and uniform in this work. This means that Nyx should sample underdense regions more evenly, at the cost of resolution in high-density regions and dynamic range. Gadget has a much better spatial resolution of gravity than Nyx, due to the additional short-range force calculation. This will surely affect the collapse of structures with sizes near the grid scale,

Table 3.1: Simulations used in this chapter.

Name	Box size (h^{-1}Mpc)	Elements	Resolution (h^{-1}kpc)	$m_{\rm dm}$ ($\mathrm{M}_\odot$)
G_L10_N128	10	128^3	78	4.3×10^7
G_L10_N256	-	256^3	39	5.4×10^6
G_L10_N512	-	512^3	20	6.7×10^5
G_L10_N512_B64	-	-	-	-
G_L10_N512_B128	-	-	-	-
G_L40_N1024	40	1024^3	39	5.4×10^6
N_L10_N128	10	128^3	78	4.3×10^7
N_L10_N256	-	256^3	39	5.4×10^6
N_L10_N512	-	512^3	20	6.7×10^5
N_L10_N1024	-	1024^3	10	8.4×10^4
N_L40_N2048	40	2048^3	20	6.7×10^5

The top section lists the Gadget simulations (names starting with G). The bottom section lists the Nyx simulations (names starting with N). Dashes indicate a repeated value. The Gadget simulations use a gravitational softening of $\epsilon = 1/3$ of the initial interparticle separation, a quicklya star formation with a threshold density of $1000\,\bar{\rho}_{\rm b}$, and 32 neighbors in the SPH kernel by default. The extra Gadget L10_N512 simulations ending with B64 and B128 use 64 and 128 neighbors, respectively. The Nyx simulations use one particle per cell. All simulations were initialized at $z_{\rm init} = 159$. See text for more details.

implying that Nyx might need more resolution elements to reproduce the same high-density regions seen in Gadget. It is very difficult to say what mix of resolution will more adequately capture the Jeans scale structures near mean density without directly testing. We emphasize that the arguments above are very qualitative and should not be relied on in detail. The resolution effects are much better understood by directly testing the convergence of the flux statistics of interest.

3.1.4 Simulation suite

Our simulation suite consists of 6 Gadget simulations and 5 Nyx simulations, listed in Table 3.1.4. The first letter in the simulation name indicates which code was used – G for Gadget and N for Nyx. In the elements column, we list the number of resolution elements, meaning the number of DM and SPH particles in Gadget and the number of grid cells and DM particles in Nyx. In the resolution column, we list the interparticle spacing in Gadget and Nyx, which is the same as the PM grid in Gadget and the PM and hydro grid in Nyx. We have 5 Gadget simulations and 4 Nyx simulations in a $10\,h^{-1}$Mpc box, with the same large-scale initial perturbations, and varying resolutions. Simulations with the same resolution use the same initial perturbations. These simulations are designed to achieve high resolution in a small box so we can compare the codes' convergence properties and compare directly with converged statistics in the highest resolution simulations. Fixing the initial conditions, cosmology, and the ionizing background between the simulations allows us to isolate differences created by the codes, and primarily the different hydrodynamics algorithms. Additionally, we will present one Gadget simulation and one Nyx simulation in a $40\,h^{-1}$Mpc box, again with matching initial perturbations, that will serve as a best-case test for either code. All simulations were initialized at $z = 159$, and assume a primordial mass fraction of helium $Y = 0.24$. As mentioned above, the Gadget simulations in this work use the quicklya star formation prescription. We use a standard setting for quicklya critical density of 1000 times the mean, above which SPH particles are converted to star particles. In early testing, we found that lower values of around 100 times the mean change flux statistics significantly. We also found that the standard gravitational softening $\epsilon = 1/30$ is unnecessary for this application, and that simulations with $\epsilon = 1/3$ produce indistinguishable flux statistic results. Since simulations with higher gravitational softening values are less expensive, we use $\epsilon = 1/3$ in this work. In 4 out of the 6 Gadget simulations, we use the standard 32 neighbors for the SPH kernel. In order to test the effect of a higher resolution kernel, we reran the G_L10_N512 simulation with 64 and 128 neighbors.

We generated initial conditions matching the WMAP7 ΛCDM power spectrum. The corresponding ΛCDM cosmological parameters are $\Omega_\mathrm{b} = 0.046$, $\Omega_\mathrm{m} = 0.0275$, $\Omega_\Lambda = 0.725$, $h = 0.702$, $\sigma_8 = 0.816$, and $n_s = 0.96$. The positions and velocities of the dark matter particles were initialized using a Zel'dovich approximation. The gas is assumed to trace the dark matter at the initial redshift. This is not quite the case in the Universe, since the baryon and dark matter transfer functions are still significantly different at $z = 159$ and the radiation contribution to the expansion rate is non-negligible at this redshift. However, since the simulations do not contain a radiation component, the initial conditions are generated by calculating the linear power spectrum at $z = 0$, which is then suppressed by the linear growth factor in a radiation-less universe back to the initial redshift. This construction ensures that the linear power spectra at the redshifts of interest are matched (the linear baryon and dark matter transfer functions are indistinguishable at these redshifts). This is a standard procedure in simulations, although it has never been shown explicitly, to our knowledge, that the initialization to the same initial conditions does not produce unwanted artifacts. For numerical reasons, Gadget starts with each mass element being split between an SPH and a dark matter particle. The two are shifted by half of the interparticle spacing in order to avoid coupling between the DM and SPH particles. This is not correct in detail, but we do not modify this as it is the most common way to initialize Gadget simulations and it is very unlikely that correcting this would change our results. We start with an isothermal initial temperature, $T = 1000\,\mathrm{K}$, and results are insensitive to a particular value chosen.

The Nyx simulations were run on Cray XC30 system (Edison) at NERSC. Each Edison node has 24 Intel Ivy Bridge cores running at 2.4GHz and 64GB of RAM, and we ran Nyx in MPI+OMP mode, which is not optimal for small problems (less than tens of thousands of cores). On each node we loaded four 128^3 grids and accompanying particles, meaning that, for example, the N_L10_N1024 simulation was done on 3072 cores, the N_L10_N512 was on 384 cores, and so on. As an example of the total runtime, the N_L10_N1024 run took us about 15,500 core-hours from $z = 159$ to $z = 2$ including I/O, or about 5 hours of walltime.

The Gadget simulations were performed on the Cray XE6 system (Hopper) at NERSC. Each Hopper node has 24 AMD Magny-Cours cores running at 2.1GHz and 32GB of RAM. The Gadget simulations were run in MPI mode only which is effective for small problems described in this work. The G_L10_N512 simulation was run on 768 cores and took a total of 12,596 core-hours which translates to a walltime of 16.4 hours.

We caution readers not to use this benchmark to draw conclusions about relative code performance for general problems. For the simulations in this work, Nyx and Gadget were run with different sets of parameters (typical to each code) which determine the accuracy of the simulation. We include these runtime numbers only to show the relative computational expense expected for science-grade Lyα forest simulations between the codes. The main conclusion is that for the same number of resolution elements, the codes cost about the same in total CPU hours. The difference is that Nyx is able to effectively scale to a much larger number of cores. This makes it possible to run the same problem in less walltime, and also to run larger problems than those possible with Gadget.

3.1.5 Comparison data details

One major difference between the Nyx and Gadget output data is that SPH data is naturally adaptive, while finite-volume data is fixed to the original grid. In order to compare the Nyx and Gadget data at the same positions, we evaluate the SPH data on specific grids and test that our choice of sampling does not degrade the SPH results. Evaluating the SPH data on a fixed grid destroys the adaptivity of the SPH particle data, but the signal we are interested in mostly comes from densities of about 10 times the mean or less. In these regions, the kernel scale is still similar to the mean inter-particle spacing, and we do not require very fine grids (as we will demonstrate later).

We evaluate the SPH data at a position $\mathbf{x}_i$ by first solving for the density ρ_i and kernel scale h_i. The density is

$$\rho_i = \rho(\mathbf{x}_i) = \sum_j m_j W(\left|\mathbf{x}_i - \mathbf{x}_j\right|, h_i) \ , \tag{3.1}$$

where the sum is over all particles, m_j is the mass of particle j, and W is the SPH kernel. Gadget uses a cubic spline for the kernel (Springel 2005):

$$\begin{aligned} W(x, r) &= s(q = x/r) \\ &= \frac{8}{\pi h^3} \begin{cases} 1 - 6q^2 + 6q^3 \ , & 0 \le q < 1/2 \ , \\ 2(1-q)^3 \ , & 1/2 \le q < 1 \ , \\ 0 \ , & q \ge 1 \end{cases} \end{aligned} \tag{3.2}$$

We iteratively solve for the kernel scale h_i so that

$$\rho_i h_i^3 = \frac{3}{4\pi} N_{\rm sph} m_{\rm sph} \ , \tag{3.3}$$

where $N_{\rm sph}$ is the SPH number of neighbors and $m_{\rm sph}$ is the SPH particle mass. One detail to note is that there is an offset of 1 between the typically quoted number of neighbors in Gadget and $N_{\rm sph}$ as used here. When Gadget is quoted as being run with say, 32 neighbors, this means $N_{\rm sph} = 33$, since it is understood as the current particle plus its 32 neighbors. Also note in Equation 3.1 that the scale h used in the kernel is evaluted at the same point, not at the position of the neighbor particle. This is a common mistake which we will return to later in this section. Once we have the density and kernel scale estimates, we can evaluate any other SPH quantity A as

$$A_i = A(\mathbf{x}_i) = \sum_j m_j \frac{A_j}{\rho_j} W(|\mathbf{x}_i - \mathbf{x}_j| \, , h_i) \tag{3.4}$$

The optical depth to HI Lyα scattering is

$$\tau_\nu = \frac{\pi e^2}{m_e c} f_{12} \int n_{\rm HI} \phi_\nu \, dr \tag{3.5}$$

where ν is the observed frequency, e is the electron change, m_e is the electron mass, c is the speed of light, f_{12} is the oscillator strength for the transition, and ϕ_ν is the line profile. In general, the line profile is a Voigt profile, but we use the Doppler profile instead for several reasons. For line center optical depths of less than 1000, the Doppler profile gives identical fluxes. We are only interested in Lyα forest systems, which have line center optical depths less than 10. For LLS and DLAs, our simulations do not produce the correct density and temperature to HI density mapping in any case. The HI density in these systems should have self-shielding corrections, which we cannot evaluate properly without coupled radiative transfer-hydrodynamics in the simulations. If we were to use Voigt profiles with these high column density systems, the damping wings would not only be inaccurate, but those errors would then contaminate nearby regions. In addition, almost all other simulated Lyα forest works have used Doppler profiles for the same reasons, so this makes the comparison to previous work simple. The Doppler profile is

$$\phi_\nu = \frac{1}{\sqrt{\pi}} \exp\left[- \left(\frac{\nu - (1 - \frac{v_\parallel}{c})\nu_0}{\Delta\nu_{\rm D}} \right)^2 \right] \ , \tag{3.6}$$

where ν_0 is the rest-frame frequency, $\nu_{\rm D}$ is the Doppler width, and $v_\parallel$ is the peculiar velocity along the line of sight. We assume that the Doppler broadening is purely thermal ($\nu_{\rm D} = \frac{v_{\rm th}}{c}\nu_0$ and $v_{\rm th} = \sqrt{2kT/m_H}$). Since we evaluate the optical depth at a fixed redshift (that of the snapshot), it is much simpler to work in velocity coordinates, i.e. $a\,dx = dv/H$, where x is the comoving scale. In this case, the full optical depth expression is

$$\tau_v = \frac{\pi e^2 f_{12} \lambda_0}{m_e c} \int \frac{n_{\mathrm{HI}}}{v_{\mathrm{th}}} \exp\left[- \left(\frac{v - v' - v_{\parallel}}{v_{\mathrm{th}}} \right)^2 \right] \frac{dv'}{H} \,, \tag{3.7}$$

where λ_0 is the rest-frame wavelength. We perform the integral as described in Lukić et al. (2015), where we assume the line-of-sight data are piecewise constant, and the integral has an analytic form. This is similar to a simple midpoint evaluation, but more reliable (will not miss contributions from narrow lines, for instance).

In general, the process of calculating the HI Lyα optical depth from simulation quantities is a straightforward process. The steps are choosing the sightlines, sampling the quantities $(n_{\mathrm{HI}}, v_{\parallel}, T)$ on each sightline, and performing the path integral. This is simple in a finite-volume code like Nyx, which already provides these quantities at uniformly spaced points. However, we found that there are many paths going from SPH particle data to Lyα optical depths.

The first possible difference we noticed is that some SPH Lyα forest analyses codes use SPH kernels different from the form used in Gadget. For instance, in an early version of the analysis for this work, our code used a kernel with the form $A_i = \sum_j A_j \frac{m_j}{\rho_j} W(|\mathbf{x}_i - \mathbf{x}_j|, h_j)$. That is, instead of evaluating the kernel with the smoothing length of the point $\mathbf{x}_i$, using the smoothing lengths of the neighbor particles in the sum at $\mathbf{x}_j$. Of course, this is inconsistent with the kernel used to evolve the quantities in the simulation. However, as we inherited this analysis code, we felt we should mention this as a possibility for other works that are not explicit about their SPH evaluation method. This produces significant differences in the flux statistics as shown in Figure 3.1.

We also checked that the tolerance of the iterative solve for the density and smoothing length (ρ_i, h_i) does not affect flux statistics. Since the optical depth is roughly proportional to the square of the density, we thought small errors in the density could produce noticeable differences in flux statistics. The iterative solve for the smoothing length at a point is stopped based on a a tolerance setting for the mass enclosed in a sphere of radius h_i. The default Gadget setting is a mass tolerance of the mass of one SPH particle. However, for most of our analysis we used a mass tolerance of 1 per cent of the SPH particle mass. As we show in Figure 3.1, this change in the density-smoothing length solve tolerance has essentially no effect on the flux statistics, demonstrating that the default Gadget setting is sufficient.

Another possible difference we noticed is the order of operations in evaluating n_{HI} and T. One choice is to evaluate (n_{HI}, T) given (ρ, e_{int}) of the SPH particles, then to SPH interpolate (n_{HI}, T) along the sightlines. This is the method used in Theuns et al. (1998), for example. The other option is to first evaluate (ρ, e_{int}) along the sightlines, and to then solve for (n_{HI}, T). Given fine enough particle sampling, these two methods should recover the same (n_{HI}, T) values, however we find small differences that end up affecting the flux statistics at the level of a few per cent, as shown in Figure 3.1. The first method tends to produce slightly lower temperatures at low densities, resulting in thinner absorption features, and

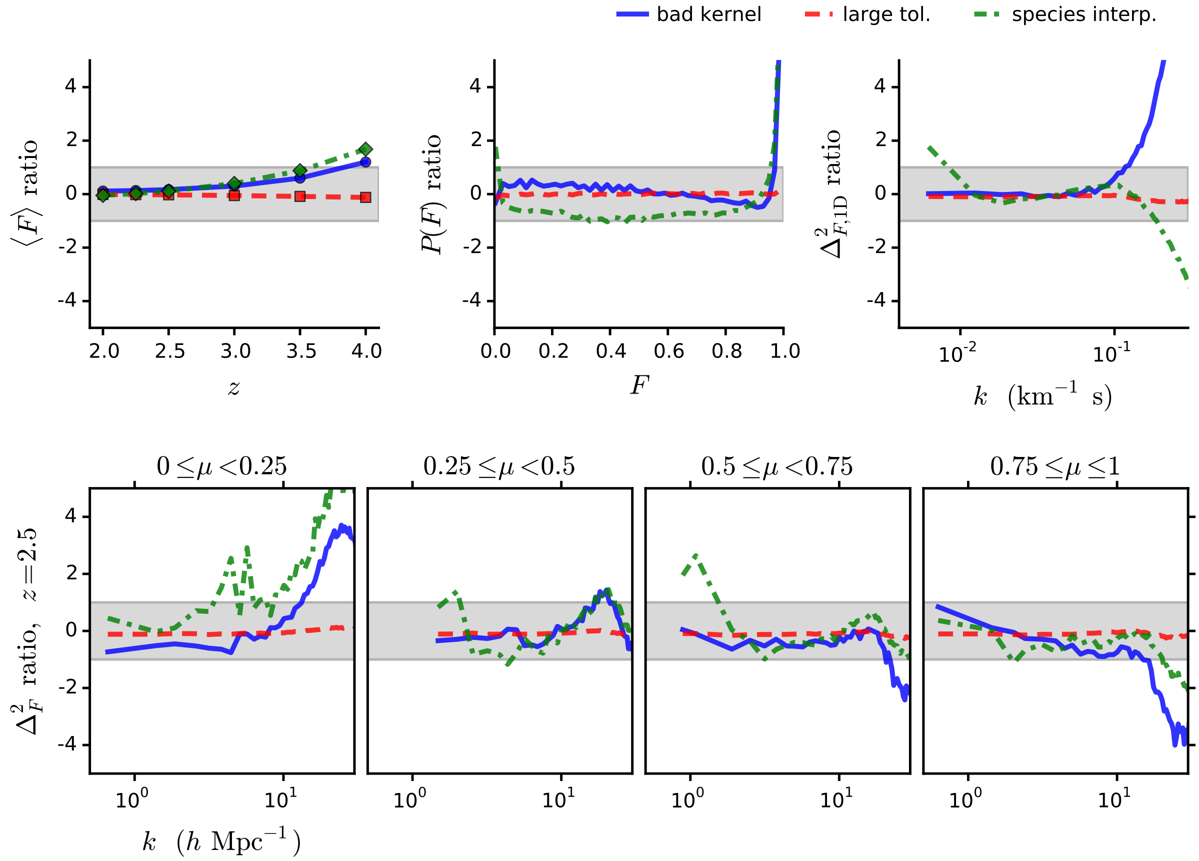

Figure 3.1: Ratios of flux statistics using different methods for computing the optical depths compared to our standard method. 'bad kernel' refers to the case of using the alternate kernel, 'large tol.' refers to using a larger tolerance in the density-smoothing length solve, and 'species interp.' refers to interpolating (n_{HI}, T) rather than (ρ, e_{int}). Top left: mean flux vs. redshift. Top middle: flux PDF. Top right: 1D flux power. Bottom: 3D flux power.

more transmission at high redshift. Although the choice is somewhat arbitrary, we believe it should be more accurate to interpolate the conserved quantities (ρ, e_{int}) rather than the EOS-dependent quantities (n_{HI}, T).

In Figure 3.1, we compare the flux statistics using the three different cases described above, compared to our standard method. In the legend, 'bad kernel' refers to the case of using the alternate kernel, 'large tol.' refers to using a larger tolerance in the density-smoothing length solve, and 'species interp.' refers to interpolating (n_{HI}, T) rather than (ρ, e_{int}). As described above, the tolerance setting makes a noticeable, but extremely small difference. It appears that the tolerance could be increased even more with essentially no impact on the flux statistics accuracy. The bad kernel and species interp. methods both predict a higher mean flux than the standard method, mainly due to lower temperatures in underdense regions. The flux PDF differences are small for $F < 0.9$, but the high transmission tail is much higher in the other methods. In the 1D flux power, the species interp. result appears to have a different slope, with more power on large scales and less power on small scales. The bad kernel result, on the other hand, only has excess power on small scales. In the 3D flux power, the two alternative methods behave similarly, with decreased power along the line of sight, and increased power transverse to the line of sight. While these differences are not very large, it is worrisome that differences of this magnitude can be created from the same Gadget snapshot using slightly different analysis methods. For this reason, we urge future precision SPH Lyα forest work to be explicit about the details of their calculations.

As mentioned above, another issue in using SPH data is how finely the quantities should be sampled. While the Nyx finite-volume data has a natural resolution, the Gadget SPH data is adaptive, and can be sampled at arbitrary resolution. Since we perform most of our analysis on a fixed grid, we must test how the grid resolution affects the Gadget results. There are three resolution parameters that determine the results: the number of simulation elements N, the grid resolution perpendicular to the LOS $N_\perp$, and the grid resolution parallel to the LOS $N_\parallel$. We tested each of these individually and found that $N_\perp$ has essentially no effect (above our minimum resolution), $N_\parallel$ has a small, sometimes negligible effect, and N is the driving factor in differences in the results, as we would expect.

We show the ratios of flux statistics (most at $z = 2.5$ again) using the Gadget L10_N256 simulation evaluated with $N_\perp = 128$ and 256 compared to $N_\perp = 512$. We hold $N_\parallel$ fixed to 256 in this case. The perpendicular resolution appears to make no noticeable difference in the mean flux, flux PDF, and 1D power spectrum. There is some noticeable noise in the $N_\perp = 128$ ratio, but this appears to be driven by the fine binning, as the $N_\perp = 128$ data provides 16 times fewer pixels than the 512 case. In the 3D flux power spectra, we can see the effect of limited perpendicular resolution (as we are now Fourier transforming along the transverse axes), and the lower resolution runs have excess power at small scales. The errors in the $N_\perp = 128$ case are significant, but the errors in the $N_\perp = 256$ case appear to be isolated to small enough scales to not be a concern. This suggests we should use at least $N_\perp = 256$ (or equivalently $\Delta x_\perp = 40\,h^{-1}$kpc) when generating flux statistics.

In Figure 3.3, we show the ratios of flux statistics using the Gadget L10_N256 simulation evaluated with $N_\parallel = 128$ and 256 compared to $N_\parallel = 512$. We hold $N_\perp$ fixed to 256 in this case. In the mean flux and flux PDF, there is a noticeable, but very small error $N_\parallel = 128$ results. The main effect appears to be a small slope difference in the flux PDF, where the $N_\parallel = 128$ result has more pixels at low transmission and fewer satured pixels than the $N_\parallel = 512$ case. However, the differences are much larger in the power spectra. In the k range we are interested in, the $N_\parallel = 128$ and 256 results differ by more than a per cent compared to the $N_\parallel = 512$ results. This suggests we should use at least $N_\parallel = 512$ (or equivalently $\Delta x_\perp = 20\,h^{-1}$kpc) when generating flux statistics.

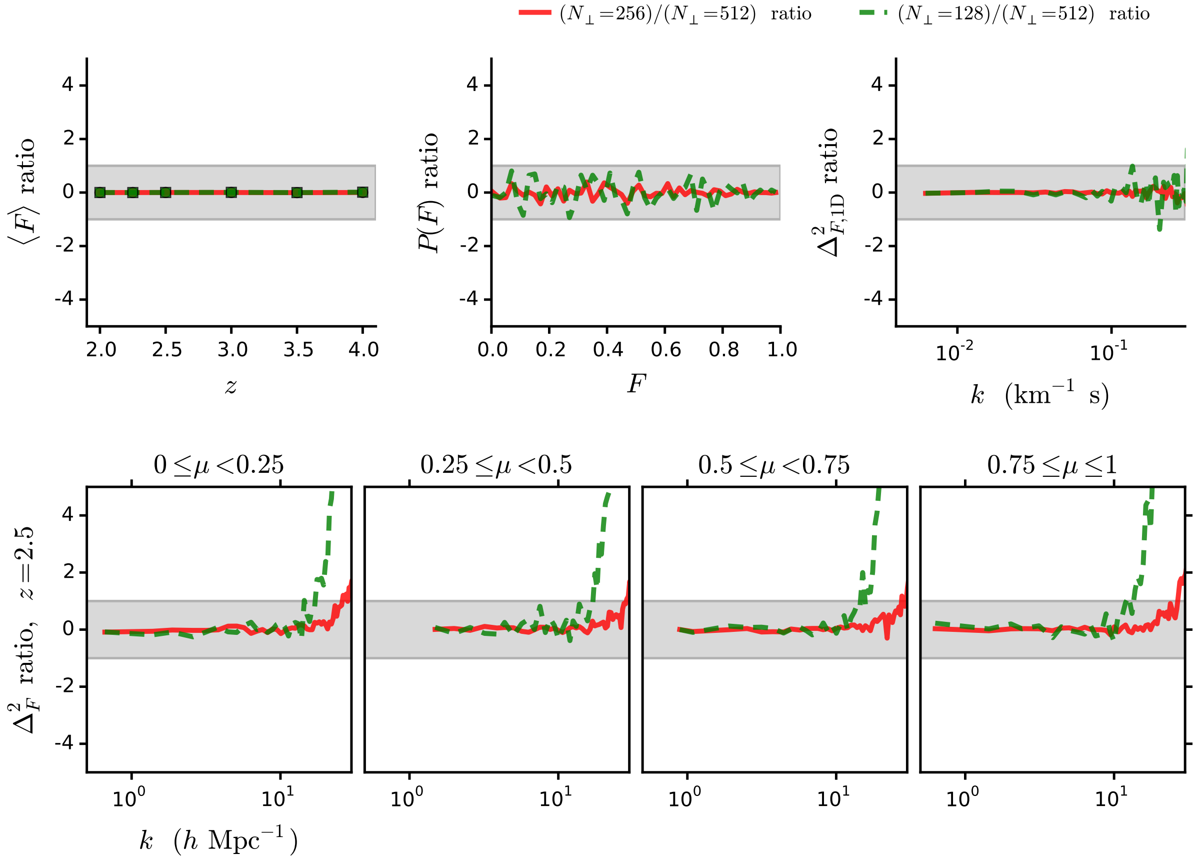

Figure 3.2: Ratios of flux statistics from the Gadget N256 simulation, evaluated on grid with $N_\parallel = 256$ and varying $N_\perp$. Top left: mean flux vs. redshift. Top middle: flux PDF. Top right: 1D flux power. Bottom: 3D flux power.

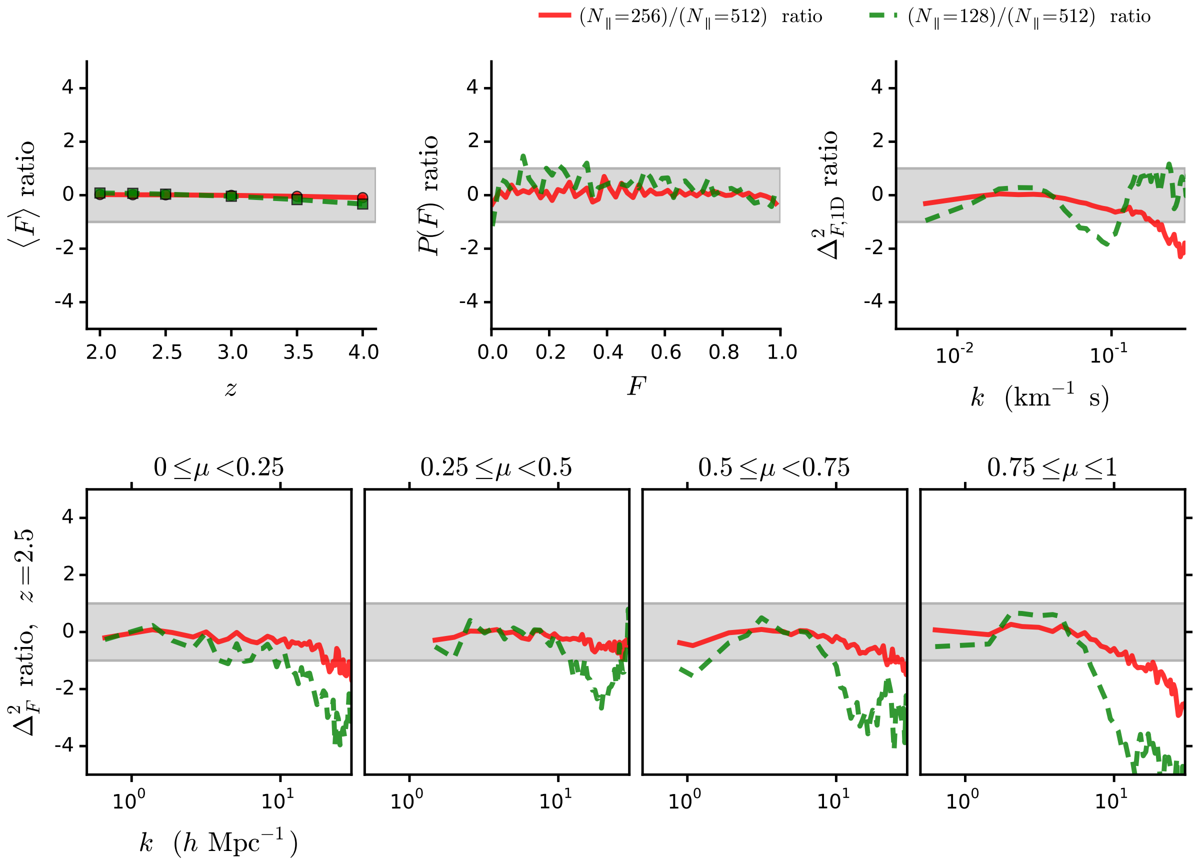

Figure 3.3: Ratios of flux statistics from the Gadget N256 simulation, evaluated on grid with $N_\perp = 256$ and varying $N_\parallel$.

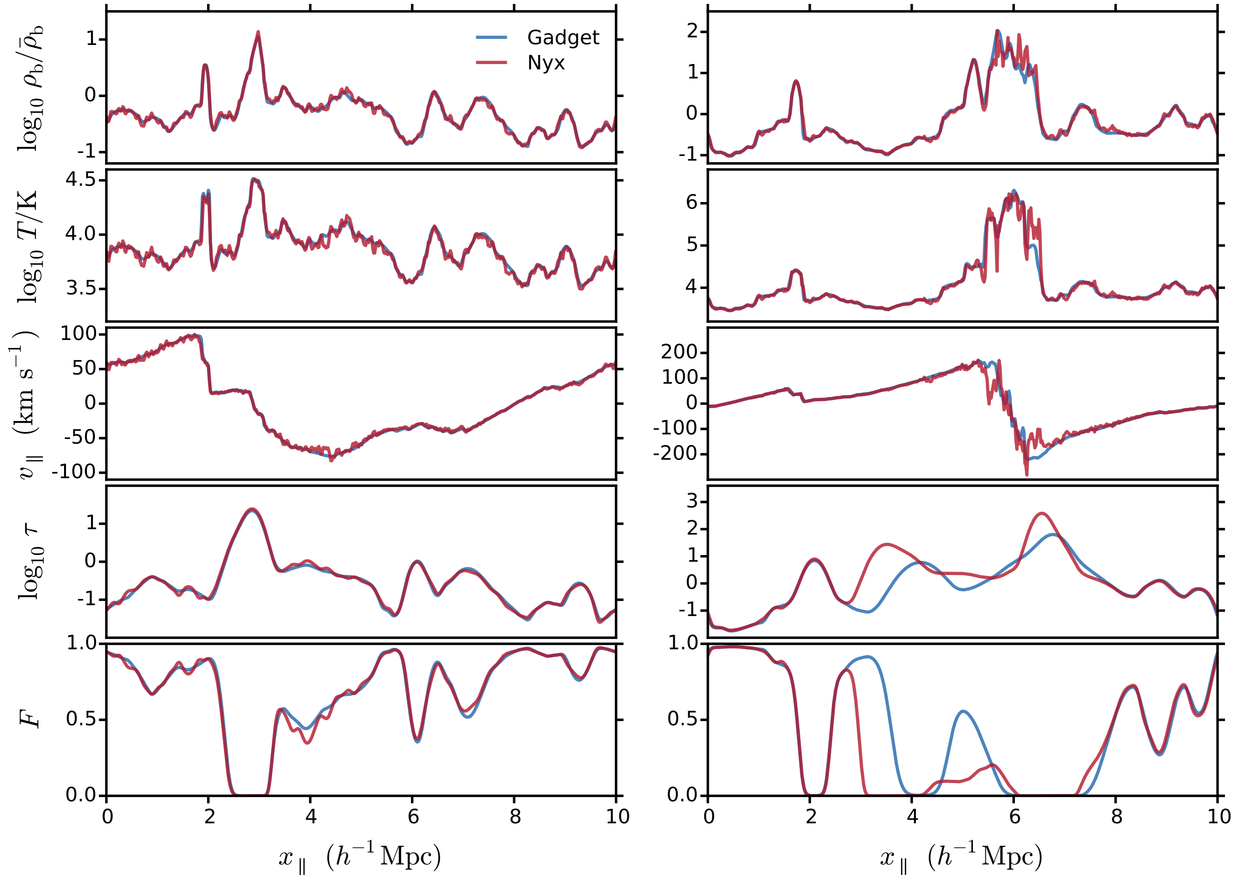

Figure 3.4: Two skewers from the L10_N512 simulations at $z = 2.5$. From top to bottom, we show the baryon density ρ_b, temperature T, velocity parallel to the sightline $v_\parallel$, optical depth τ, and transmitted flux fraction F. The skewer on the left is an example of where the codes agree fairly well passing through moderate densities, while the skewer on the right is an example of where the codes do not agree near high densities (note the difference in the density scales).

3.2 Morphological comparison

We start by comparing the morphological characteristic of quantities in the two codes. This section is not very quantitative and its purpose is mainly to build intuition of how the codes come to somewhat different results. In subsequent sections we will make the comparison more quantitative by considering directly observable quantities, such as power spectra and probability density functions.

Figure 3.4 shows two example skewers through our simulations. We show the baryon density ρ_b, temperature T, velocity along the skewer $v_\parallel$, the optical depth τ, and transmitted flux $F = e^{-\tau}$. The skewer on the left pierces only moderate densities ($0.1 < \rho_\mathrm{b}/\bar{\rho}_\mathrm{b} < 10$), while the skewer on the right pierces a high density region (meaning $\rho_\mathrm{b} \sim 100\,\bar{\rho}_\mathrm{b}$). We chose these two skewers to show a common, good agreement case and a rare, poor agreement case. Overall, the codes agree very well (ignoring the flux on the right for a moment). Although the initial conditions are the same, there is no guarantee that the codes will evolve the structures in the same way, to the same positions at $z = 2$. There was no fine tuning of the skewer position between the codes, just a choice of the same sightline (x, y) position. The most apparent difference between the codes is the noise-like jitter in the Gadget density, temperature, and velocity compared to the relative smoothness in the Nyx data. This does not mean that the Nyx data is inherently more accurate or reliable, but it certainly highlights the difference in the properties of the simulation data between the codes. The jitter in the Gadget skewers is due to the shot noise behavior of particle sampling ('seeing' the particles). If we were to deposit the Gadget particle data to a coarse enough grid, it would appear much smoother, like the Nyx data. However, too coarse a grid would mean undershooting many extrema in the data, and it is much simpler to use the same grid. In the left skewer, the Nyx and Gadget fluxes are overall very similar, with small deviations around $x = 4$ and $7\,h^{-1}$Mpc. These differences appear to originate in differences in the density peaks between the codes, where the Nyx peaks are smooth and the Gadget peaks have more small-scale structure. The skewer on the right intersects a high-density spike around $x = 6\,h^{-1}$Mpc – a region affected by the quicklya processing in Gadget. This produces sharp features in density, temperature, and velocity. Although this region covers only about $1\,h^{-1}$Mpc in real space, the combination of the high temperature and large velocity gradient creates large differences in the flux. The disrupted region distorts the spectrum from $x = 3$ to $8\,h^{-1}$Mpc. After visually inspecting many skewers, we believe that the rare skewers passing through high densities account for all of the major differences in flux between the codes. That is, in low density regions, the two codes produce very similar structures and, qualitatively, the same flux signal with some small deviations in line centers and widths. However, near high densities, the overcooled halos in Nyx and the disrupted halos forming stars in Gadget create significant flux differences. More importantly than the code differences, we want to highlight this contradiction of the standard lore that high-density regions do not affect the forest. The absorption is saturated near the position of the density spikes, but the differences in the H I column and temperature give different line shapes in the spectra. In addition, high-density regions typically have large velocity gradients, meaning that any difference is spread over much larger scales in redshift-space.

Figure 3.5 shows an example slice through the G_L10_N512 and N_L10_N512 simulations. We show the baryon density and temperature, the HI number density and the transmitted flux from the L10_N512 simulations at $z = 2.5$. We also include difference images on the bottom. The line-of-sight (LOS) axis is plotted vertically, so that in the flux images, each column of pixels is a separate sightline. Again, the qualitative agreement is very good and we see the same structures in all panels. In the baryon density slices, we again see that the Nyx image appears smoother than the Gadget image. In particular, looking at the low density regions in the Gadget image, we can see the wavy structure of the initial particle lattice. These wavy structures are also mirrored in the other Gadget fields. The major differences in density (darker patches in the ratio image) come from the higher density regions, such as the halo near the center, the filament to the bottom right, and the smaller halo farther in the bottom right. The highest density region is more extended in Gadget, due to the quicklya processing. Nyx on the other hand, allows baryons to cool and collapse more, producing more concentrated halos. The filament in the bottom right appears a bit wider in Nyx, seen as dark blue streaks in the difference image. In the halo in the lower right, the Gadget density appears much more extended again, creating a red blob in the difference image. In the temperature panel, we see the largest differences mostly in accretion shocks enveloping the filaments and halos. The shocks in Gadget are a bit puffier and spotty. In the difference image, the filaments are outlined in blue then red. This suggests that although filaments have similar density profiles between the codes, the accompanying shocks are somewhat narrower in Nyx. The HI density panel shows the multiplicative differences in density and temperature that end up producing the differences in flux. The flux image is qualitatively different than the rest. Structures appear very similar in the other panels, but in flux, they are distorted in position along the LOS by peculiar velocities and fattened along the LOS by thermal broadening. Where the other panels also show smooth transitions between extreme values, the flux has a limited dynamic range and therefore much sharper features, quickly changing from a saturated region (red) to a transmissive region (blue). The largest differences in flux once again come from the high-density regions. In addition, we noticed a pattern in the flux difference image, where there are blue spots surrounded by red lines. We believe this is a result of the small difference in the temperature profiles around filaments. The sightlines pass through similar HI columns, but the Gadget temperature bumps are wider, resulting in broader lines. Nyx has more absorption near line center and Gadget has more absorption in the edges. Although these are small differences, at some level, this could affect estimates based on linewidths.

In order to check these slight morphological differences between the codes, we also compared cross sections of a single filament. It was more difficult than expected to find an isolated filament – an object that looks like a filament in one projection often turns out to be a sheet-like structure, or in other cases the filament turned out to be merging with another one. However, filamentary structures are abundant and even the small volume of the $10\,h^{-1}$Mpc box provides enough good samples. Figure 3.6 shows cross sections of a single filament found in the L10_N512 simulations at $z = 2$. We found this filament nearly

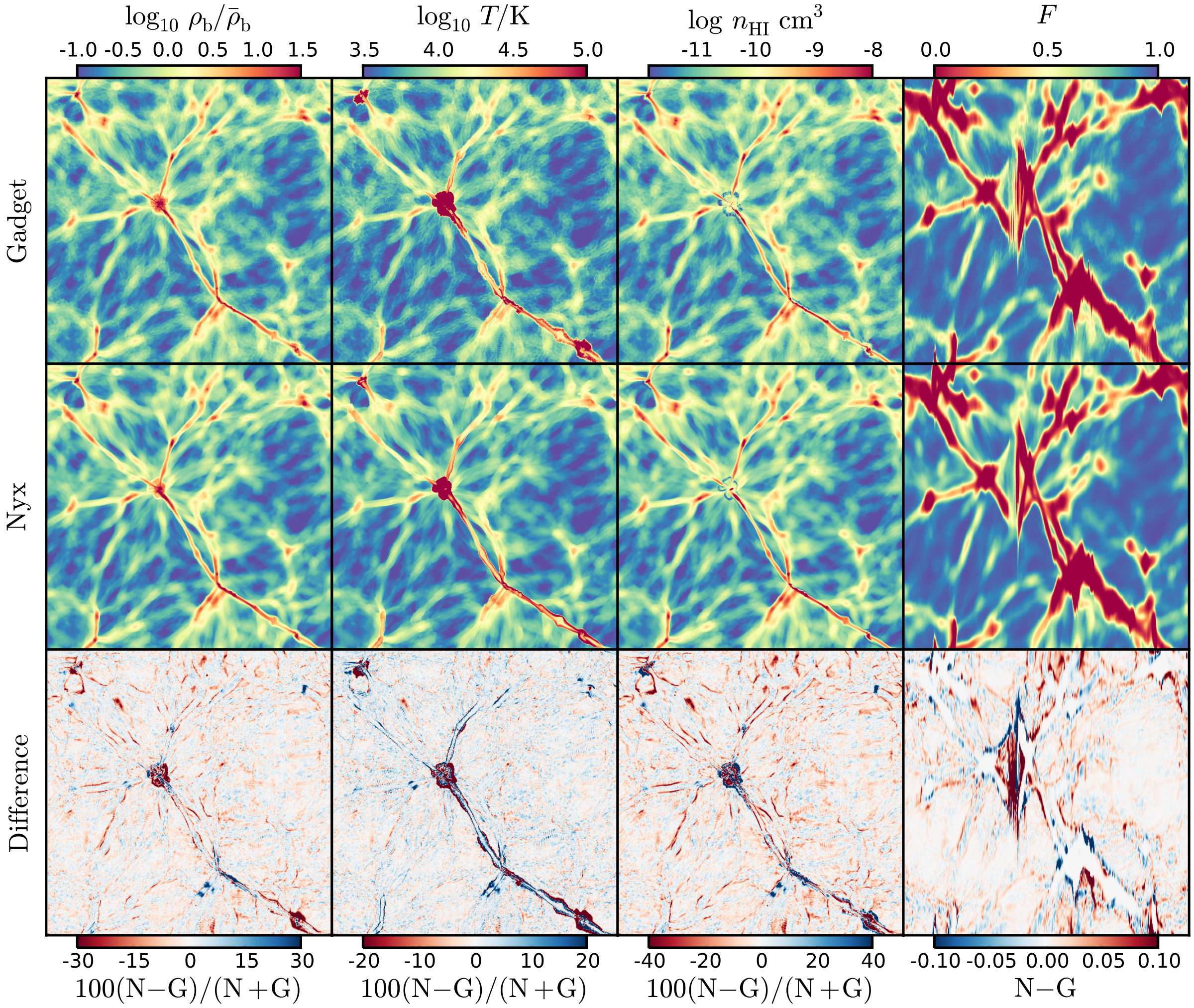

Figure 3.5: A slice through the L10_N512 simulations at $z = 2.5$, showing the baryon density ρ_b, the temperature T, the HI number density n_{HI}, and transmitted flux fraction F. From top to bottom, we show the Gadget data, the Nyx data, and the difference.

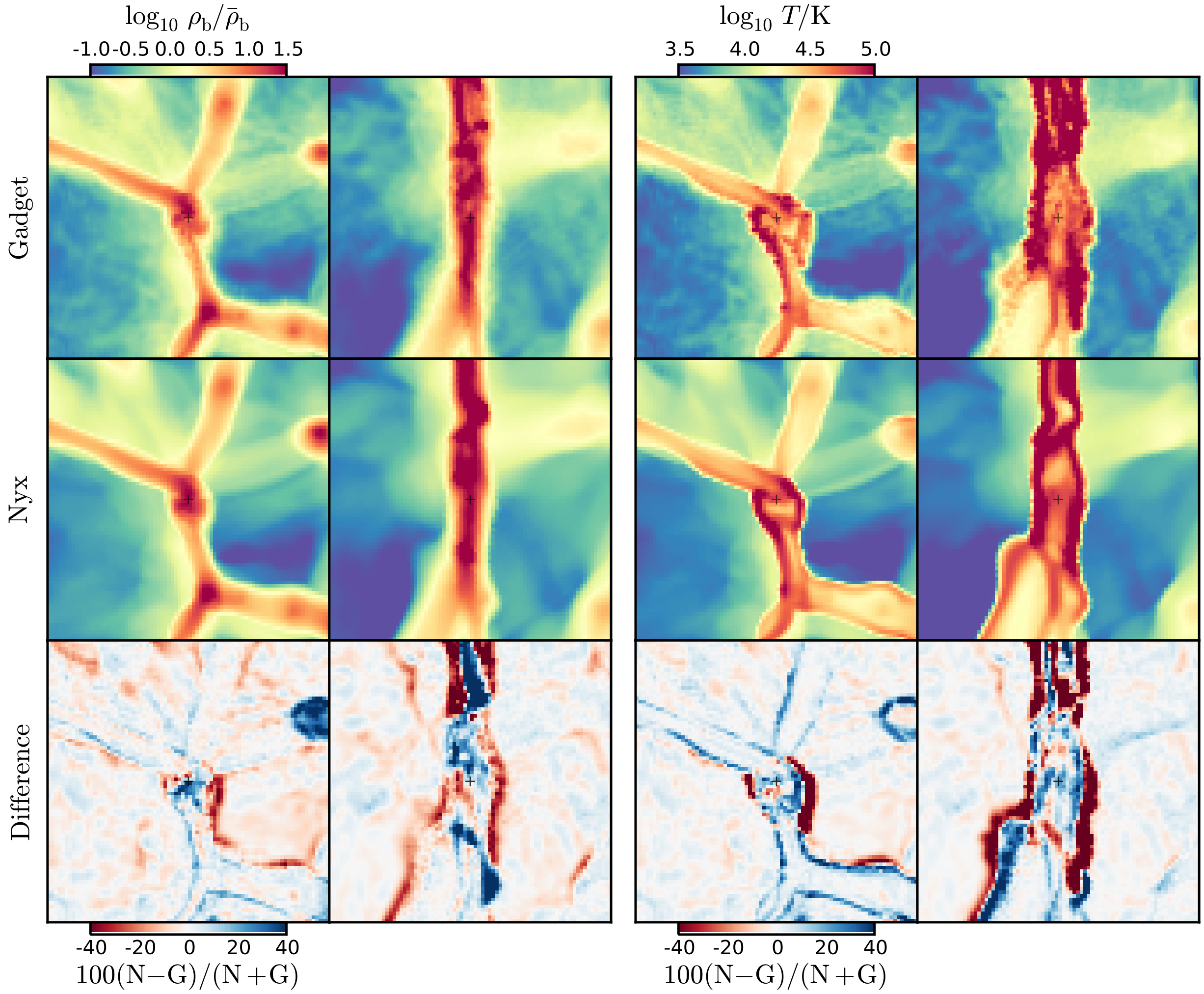

Figure 3.6: Slices along and across a filament in the L10_N512 simulations at $z = 2$. In the left panels, we show the baryon density $\rho_{\rm b}$ and in the right panels, the temperature T. In each panel section, we show a cross section across the filament on the left and along the filament on the right. From top to bottom, we show the Gadget values, Nyx values, and the difference. The slices are about $1.6\,h^{-1}$Mpc across. A black cross marks the center (common point) of the slice in both views.

aligned to the grid in our simulations, but we did slice at an angle in order to look as closely as possible down the axis of the filament. The slice values were computed using trilinear interpolation on the 8 neighboring grid values of each slice point. The figure is split into two vertical panels, showing the baryon density on the left and temperature on the right. Both panels include the two views of the filament with the filament into the page on the left and cutting along the filament on the right, with the common point marked by a black cross in the center of the image. As before, we include the Gadget and Nyx values and their difference from top to bottom. The slices are all about $1.6\,h^{-1}$Mpc across (80×80 grid points). There is clearly more substructure in the Gadget density slices. The blue regions (excess in Nyx) mostly appear in the regions within the filament where gas has evacuated in Gadget. The structural differences in the temperature images are more striking, where the shocks surrounding the filaments have qualitatively different shapes. The shocks in Nyx are sharper overall, creating thin blue lines in the difference images. The shocks in Gadget are wider, producing the dark red patches around the blue lines in the difference. This is particularly clear in the far right column.

3.3 Density and temperature statistics

We start by presenting the one- and two-point functions of the baryon density, the probability density function and the power spectrum. We are interested in the material that contributes to the Lyα forest flux, therefore in the case of Gadget, we ignore the density contribution from star particles. The star formation prescription is far from realistic, and is employed only to speed up the computation by removing SPH particles in high-density and low-temperature regions. While lower density regions should not be affected by this, we want to be clear that ‘baryon density’ in this work refers to the gas density for Gadget, where we have deposited SPH particles onto the same grid as the matching Nyx runs, and we do not include the star particle mass. In order to keep the number of figures manageable, we choose to show statistics at $z = 2.5$. This redshift is in the most observationally relevant redshift range and is fairly representative of the agreement at the other redshifts we considered. Overall, the code differences decrease with decreasing redshift, but the high redshift differences are not so much larger as to be worth showing here.

The volume-weighted baryon density probability density functions (PDF) from the small box simulations are shown in Figure 3.7. To illustrate the effects of convergence and code differences separately, we include two other axes showing the ratios of results between simulations. The Gadget ratio shows the percent differences relative to the highest resolution Gadget simulation, computed as $100(p/p_{\mathrm{hires}} - 1)$. This shows the behavior of Gadget’s convergence with respect to resolution. We also include the highest resolution Nyx simulation ratio on the same plot to compare the code results at the best resolution. We repeat the same process for the Nyx results in the Nyx ratio panel. The codes converge in a similar way for this statistic. The lower resolution runs miss low-density elements and overproduce high-density elements. The codes converge on the low end at about the same rate, but Nyx

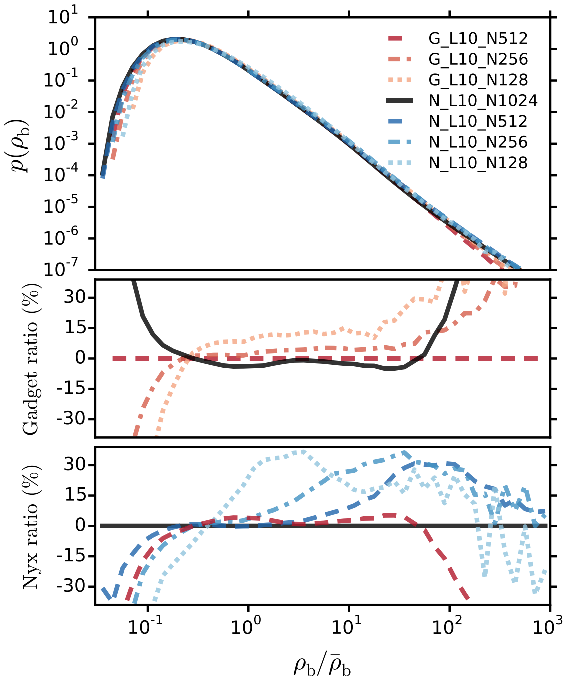

Figure 3.7: The baryon density PDFs $p(\rho_{\rm b})$ in the $10\,h^{-1}$Mpc simulations at $z = 2.5$, and their ratios. *Top:* the baryon density PDFs. *Middle:* the differences relative to the highest resolution Gadget simulation, G_L10_N512. *Bottom:* the differences relative to the highest resolution Nyx simulation, N_L10_N1024. The codes show similar convergence patterns, peaking at lower $\rho_{\rm b}$ and probing the tails of the distribution better with increasing resolution. Gadget converges much faster at higher density due to its Lagrangian nature, and also appears to converge faster to the low-density peak position. The highest resolution runs between the codes agree well in the relevant density regime of $0.1 < \rho_{\rm b}/\bar{\rho}_{\rm b} < 10$, with the G_L10_N512 result falling between the N_L10_N512 and N1024 results. The differences between the codes for $\rho_{\rm b} > 100\,\bar{\rho}_{\rm b}$ are due to different subgrid physics in high-density regions.

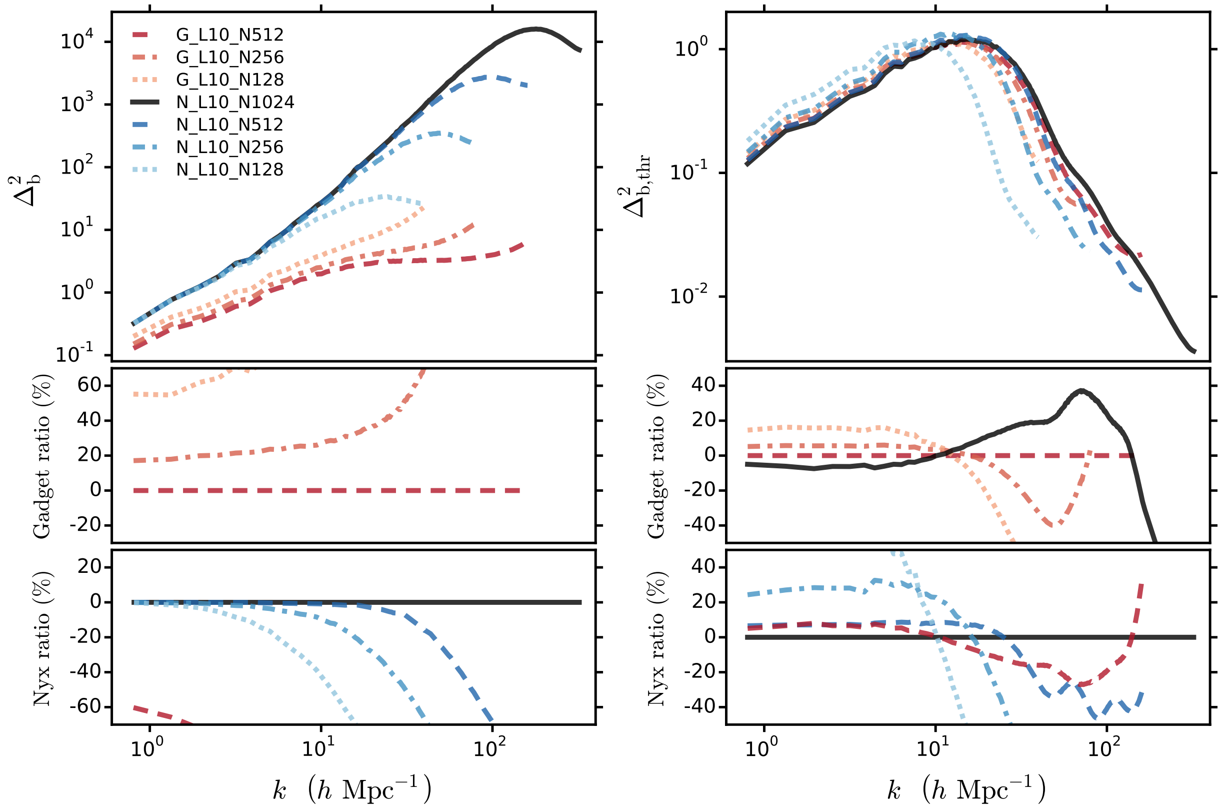

Figure 3.8: The baryon power spectra in the $10\,h^{-1}$Mpc simulations at $z = 2.5$, and their ratios. *Left:* the dimensionless power spectra of baryon overdensity $\Delta^2_{\rm b}(k)$. *Right:* the dimensionless power spectra of the thresholded baryon overdensity $\Delta^2_{\rm b,thr}(k)$, where the baryon density was limited to $10\,\bar{\rho}_{\rm b}$. The large-scale power offsets are due to different amounts of mass lost either to stars or to thresholding.

appears to converge much more slowly at the high end. This is most likely due to the Lagrangian nature of Gadget, which should capture high-density structures more easily. Over the relevant density range of $0.1 < \rho_{\rm b}/\bar{\rho}_{\rm b} < 10$, the highest resolution PDFs have very similar shapes and are within 10%. This is a good sign as this is the density range probed by the Lyα forest. Above densities of 100 times the mean, the high-resolution results diverge, and Nyx has higher probabilities. This is a combination of the quicklya star formation removing high densities in Gadget and overcooling in Nyx.

In Figure 3.8, we show the baryon power spectra in the small-box simulations. We compute the density perturbations $\delta_{\rm b} = \rho_{\rm b}/\bar{\rho}_{\rm b} - 1$, take the power spectrum of the field, and average in bins of width $\Delta k = k_{\rm fund} = 2\pi/L$. We plot the dimensionless power spectra $\Delta^2 = k^3 P/(2\pi^2)$ (for our Fourier convention and being in 3D). The codes behave very differently in this statistic. First, the Gadget power spectra have different large scale values because of the mass lost to stars. The total baryon density is still conserved in Gadget, but we do not include the mass converted to stars in this power spectrum, hence the large-scale offset. The Nyx runs show convergence characteristic of a PM code. The large scales agree, but lower resolution runs have suppressed power at small scales. This is due to the PM gravity solver in Nyx, where we can see the effect of cloud-in-cell (CIC) filtering in the suppressed growth of the baryon perturbations. In addition, the overcooling produces smaller, denser structures than in Gadget, resulting in a significant nonlinear tail in Nyx, which depends much more strongly on resolution. These very high-density regions are unphysical, but these regions should not contribute to the Lyα forest signal anyway. Nevertheless, as their overdensity is very high, they end up dominating the power spectrum shown here. In the same figure, we also show the power spectra of the thresholded baryon density on the right. We limited the baryon density to 10 times the mean (where flux is essentially saturated for $z > 2$), and used the new mean when computing the overdensity. Since the thresholding operation does not conserve mass, the large-scale power varies between runs. The thresholding operation is very effective at removing differences between the Gadget runs. The Gadget runs converge much more quickly than the Nyx runs in this statistic, mostly because a large fraction of the mass removed by the thresholding operation has already been taken care of by the star formation in Gadget. The codes agree well at best resolution, better than 10 percent up to $k \simeq 10\,h\,\mathrm{Mpc}^{-1}$. At smaller scales, Nyx still has more power than Gadget. This suggests that high densities are not as isolated as we typically imagine. Although the high-density regions saturate the flux, they still affect the nonlinear evolution of nearby lower density regions.

We show two samples of the ρ_b-T joint PDFs from the L10_N512 at high and low redshifts in Figure 3.9. We plot a large range in ρ_b and T in order to show all phases of the IGM as captured in both codes. The effect of quicklya processing is apparent in the Gadget images, where there is a sharp cutoff at $\rho_\mathrm{b} = 1000\,\bar{\rho}_\mathrm{b}$. As expected, the diffuse phase is similar between the codes. The ρ_b-T relation overlaps at both redshifts, seen as the sharp white line in the difference images. However, the scatter around the sharp ρ_b-T line differs. Gadget has more points that scatter under the line, while Nyx has more points which scatter above the line. The Gadget contours also seem to have a bump in the density range $0.1 \leq \rho_\mathrm{b}/\bar{\rho}_\mathrm{b} < 1$. At $z = 4$, Gadget appears to extend to much higher temperatures and has a much more substantial warm-hot intergalactic medium (WHIM) phase. At $z = 2$, the WHIM phases are much more similar, although Gadget has more underdense points in the $4 \leq \log_{10} T/\mathrm{K} < 6$ range. At both redshifts, Nyx seems to have many more points in collapsed phase, seen as the dark blue patch in the difference images for $\rho_\mathrm{b} > 10\,\bar{\rho}_\mathrm{b}$.

Next, we focus on small differences in the ρ_b-T relation, since they are the most relevant for the properties of the flux. Figure 3.10 shows boxplots of the temperature distribution vs. density in the $10\,h^{-1}\mathrm{Mpc}$ simulations at $z = 2.5$. We use density bins over $0.1 \leq \rho_\mathrm{b}/\bar{\rho}_\mathrm{b} < 10$ with 0.2 dex spacing. For each density bin, the line in the center of the box shows the median of $\log_{10} T/\mathrm{K}$, the box extends from the 16th to 84th percentiles (roughly the one sigma percentiles for a normal distribution). The whiskers extend down to the 2nd percentile and up to the 98th percentile (roughly two sigma percentiles). The density binning is the same for the Nyx and Gadget data, and the boxes are offset from the center only for

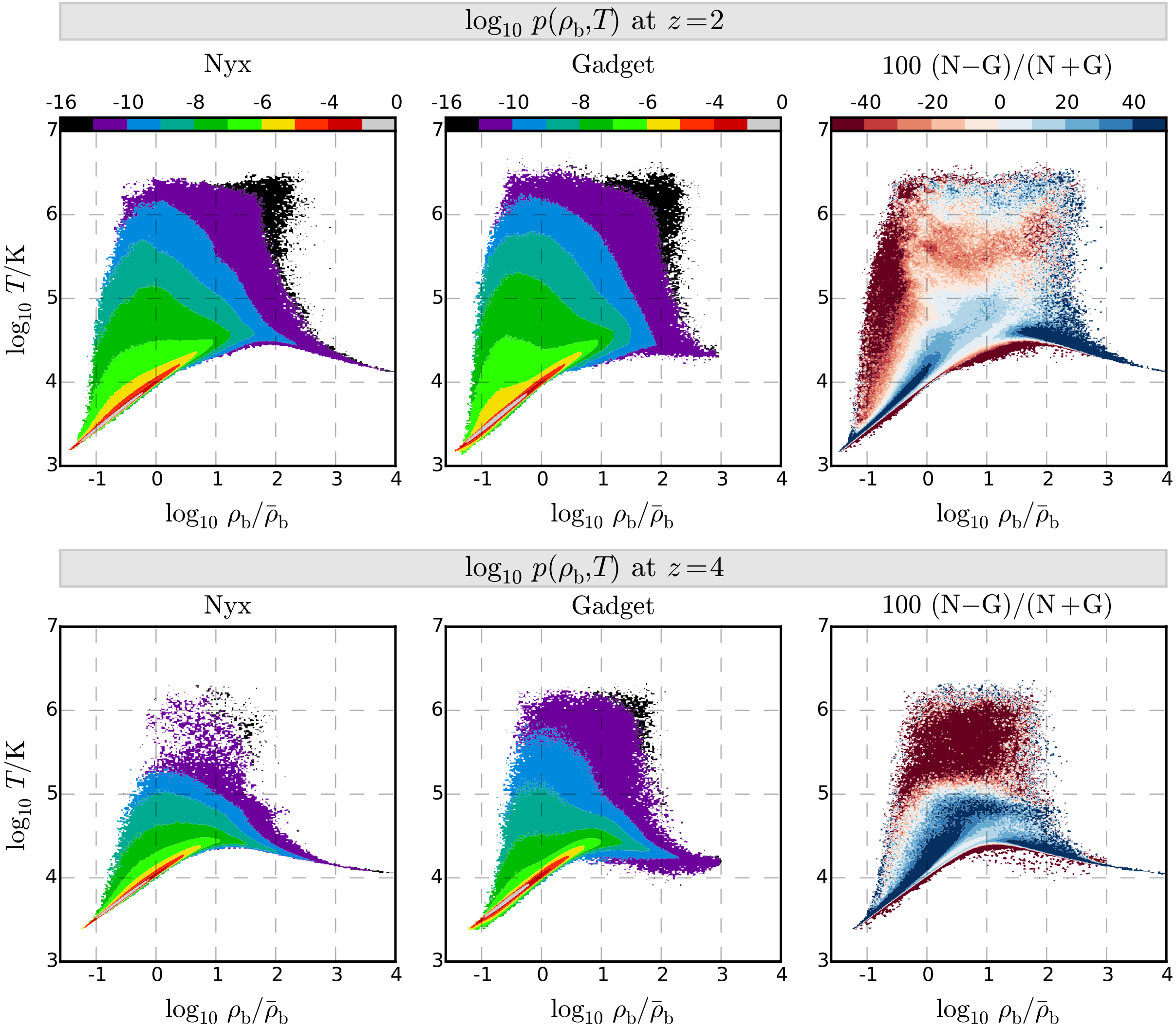

Figure 3.9: The joint PDFs of baryon density and temperature $P(\rho_b/\bar{\rho}_b, T/\mathrm{K})$ in the L10_N512 simulations. From left to right are the Nyx and Gadget PDFs and their ratios, computed as $100(\mathrm{N}-\mathrm{G})/(\mathrm{N}+\mathrm{G})$. *Top:* the distributions at $z=2$. *Bottom:* the distributions at $z=4$.

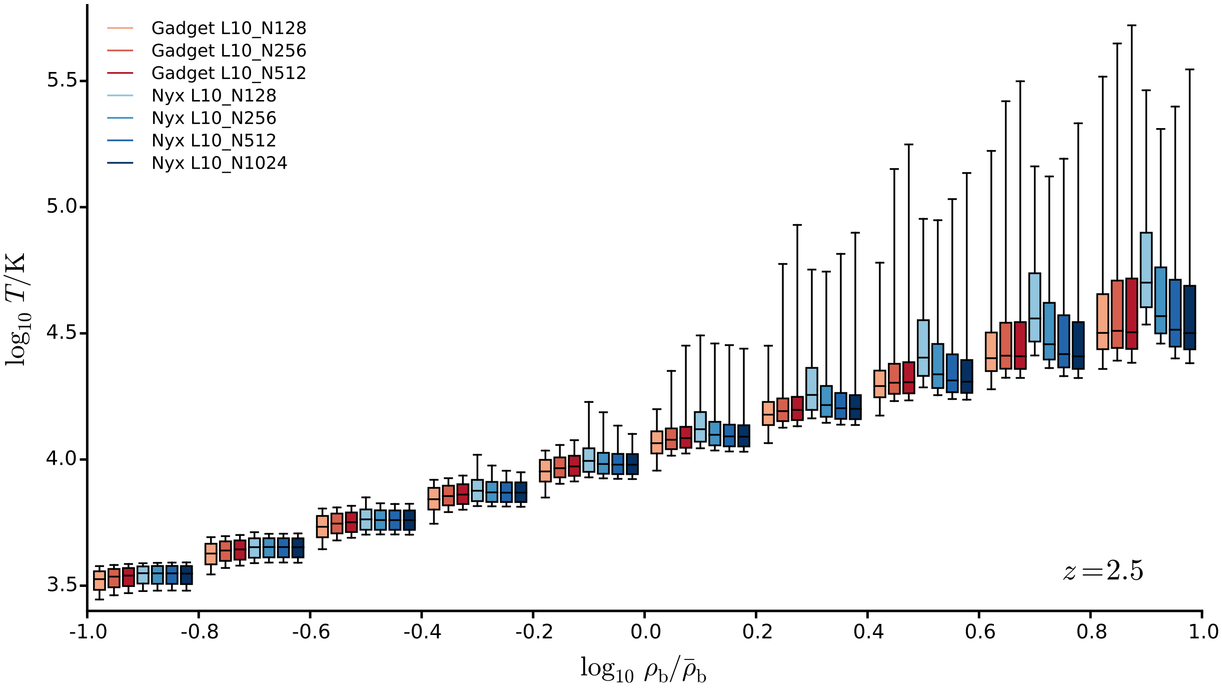

Figure 3.10: Boxplots of the temperature distributions vs. density in the $10\,h^{-1}$Mpc simulations at $z = 2.5$. There are 10 density bins over the range $0.1 \leq \rho_\mathrm{b}/\bar{\rho}_\mathrm{b} < 10$ with 0.2 dex spacing. For each density bin, the box and whiskers show the shape of the $\log_{10} T/\mathrm{K}$ distribution. The line in the middle of the box shows the median, the box extends from the 16th to 84th percentiles, and the whiskers extend to the 2nd and 98th percentiles. The density binning is the same for the Nyx and Gadget data, and the boxes are offset from the bin center only for clarity.

clarity. The exact shape of the ρ_{b}-T relations in the simulations and how they change with simulation resolution is much more apparent in this figure. This was the one statistic we considered in which the codes appear to converge differently. With increasing resolution, the Gadget temperature distributions appear to shift up slightly, while the Nyx temperature distributions shift down. The spread of the Nyx temperature distribution decreases with increasing resolution in the low-density bins ($\rho_{\mathrm{b}} < 0.2\,\bar{\rho}_{\mathrm{b}}$) and increases with resolution for higher densities. The spread of the Gadget temperature distribution increases with resolution for all densities shown. Although the codes seem to converge from different directions, at high resolution, the temperature distributions agree very well. The RMS difference between the medians is 0.012, for instance, with the Nyx median being about 1 percent higher in all bins.

3.4 Flux statistics

In this section, we compare the observationally-relevant zero-, one-, and two-point functions of the Lyα forest flux – the mean flux, the flux probability density function (PDF), and the flux power spectrum – in the small-box simulations. We show the mean flux at all available redshifts, but again we show the flux PDF and flux power only at $z = 2.5$ for the sake of brevity. We did find the expected trend in the flux PDF and power that they agree better at low redshift than at high redshift.

3.4.1 Mean flux

The mean flux $\langle F \rangle$ is the simplest direct flux statistic, which is a measure of the opacity evolution in the IGM. We compute the mean flux using all pixels in the box. Figure 3.11 shows the mean flux $\langle F \rangle$ in the small-box simulations vs. z. We also include the latest observational results in Becker et al. (2013) on the same panel, rescaled to match the simulation value at $z = 2.5$ (note that observationally the well-constrained quantity is the evolution, rather then the absolute value of the mean flux). We also show the ratios between the codes in mean flux and the effective optical depth. Again, the codes appear to converge in a similar way and with similar ratio values – lower resolution runs have more absorption than the best resolution. We also find the expected trend that the mean flux is less converged with increasing redshift. This trend has been shown in previous works (e.g. Lukić et al. 2015), and can mainly be attributed to a constant error in the effective optical depth. This can be seen in the effective optical depth ratio panels, where the 256^3 simulation result has a roughly constant offset from the highest-resolution simulation result. When this error is compared in the mean flux though, the error is much larger for low mean flux values (high redshift). At worse resolution though (128^3), there are clearly other effects. We found that lower resolution runs tend to have more absorption not because they miss dense absorption systems, but because they miss low density extremes.

The mean flux evolution differs between the two codes and both differ significantly from

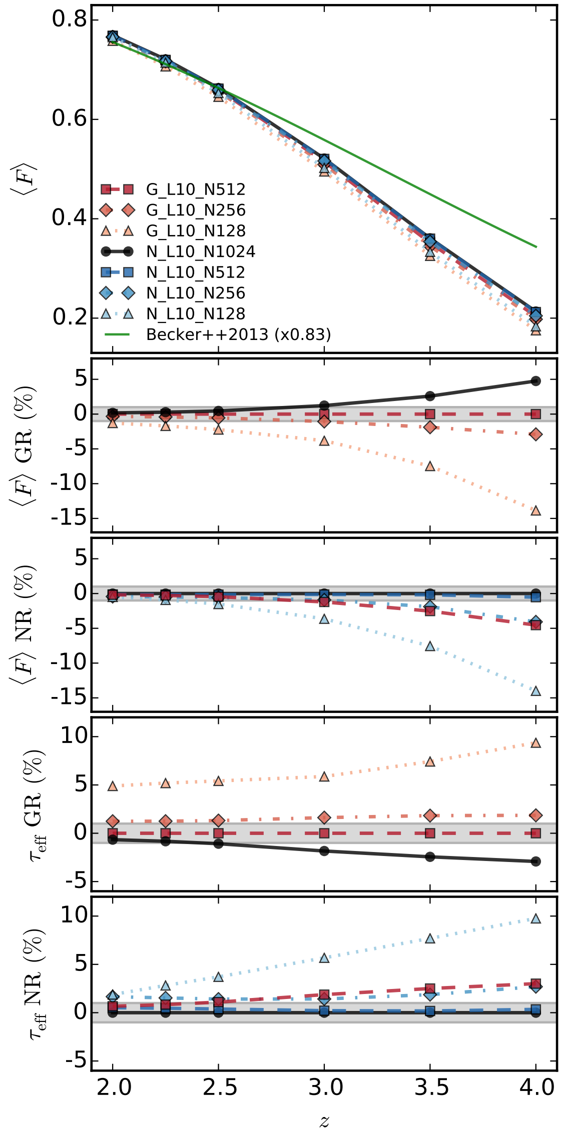

Figure 3.11: The mean flux vs. redshift $\langle F \rangle(z)$ in the $10\,h^{-1}$Mpc simulations, and their ratios. *Top panel:* $\langle F \rangle(z)$ in our simulations and compared to the Becker et al. (2013) observed mean flux fit result, multiplied by a constant to match the $z = 2.5$ value. *Second and third panels:* the percent differences in $\langle F \rangle(z)$ relative to G_L10_N512 and N_L10_N1024 respectively. *Second and third panels:* the percent differences in $\tau_{\rm eff}(z) = -\log\langle F \rangle(z)$ relative to G_L10_N512 and N_L10_N1024 respectively. The codes converge with respect to resolution at all redshifts, but the redshift evolution between the codes only agrees to a few percent. Both codes do not agree well with the observational result. See text for discussion concerning the UV background prescription.

observations. At the best resolution, Nyx starts at a higher mean flux, but evolves less than Gadget, so that they are about one per cent different by $z = 2.5$. Overall, the Gadget runs have more absorption at all redshifts. Because of this, the G_L10_N512 results are between the N_L10_N128 and N_L10_N256 results. We found that this difference mostly comes from differences in the HI density, which we will explore in Subsection 3.4.5. It is surprising how much both codes disagree with the evolution of the observed mean flux. Even with rescaling the observed $\langle F \rangle$, the observed $\langle F \rangle$ evolves less than the simulations at all redshifts. The codes appear to converge at all redshifts to a precision much better than the difference between the simulations and the observation. This points toward problems with the assumed UV background. Again, we use the prescription of Faucher-Giguère et al. (2009), however, in Lukić et al. (2015), we found similar mean flux evolution differences in simulations using the prescription of Haardt & Madau (2012).

3.4.2 Rescaling optical depths

Simulations of the IGM must assume a particular ionizing background. Common prescriptions of the ionizing background include Haardt & Madau (1996); Faucher-Giguère et al. (2009); Haardt & Madau (2012), which provide a table of the photoionization rate and heating per atom for the primordial species HI, HeI, and HeII at many redshifts. Bolton et al. (2008); Becker et al. (2011) also implemented parameterized versions of the ionizing background. The reality is that our constraints on the ionizing background are much weaker than our knowledge of cosmology, and we often parameterize out our ignorance by rescaling the simulated optical depths so that the simulated mean flux matches the observed mean flux. This rescaling operation is roughly equivalent to changing the photoionization rate Γ in the simulation, since $\tau \propto n \propto 1/\Gamma$. The idea is that small changes in Γ should only impact the species fraction to first order and any dynamic effect like changing the density or temperature is subdominant.

Unfortunately, we found that the mean flux in our simulations are very far from the observed mean flux and rescaling to the observed mean flux would not be a small correction. At redshift 2.5 for instance, in our highest resolution runs, Nyx has $\langle F \rangle = 0.664$ and Gadget has $\langle F \rangle = 0.661$, whereas Becker et al. (2013) gives $\langle F \rangle = 0.789$. We found that in order to match the observed mean flux at each redshift, we had to rescale optical depths by about 1/3. This is partly due to the low temperatures in our simulations – because we assume all material is optically thin, the simulations are missing some heating from HI and HeII ionization. A hotter IGM in the simulations would produce a higher mean flux, although the dependence is not so strong, as $\tau \propto T^{-0.7}$. Instead, we believe that most of the difference is due to the low HI photoionization rate in our ionizing background prescription. We found that simulations run with the Haardt & Madau (1996) or Haardt & Madau (2012) prescriptions have mean fluxes much closer to the observed mean flux.

Although we do not need to rescale our optical depths all the way to the observed mean flux, we do want to remove the effect of the different mean fluxes when comparing other flux statistics. We found that having different mean fluxes drives most of the differences in

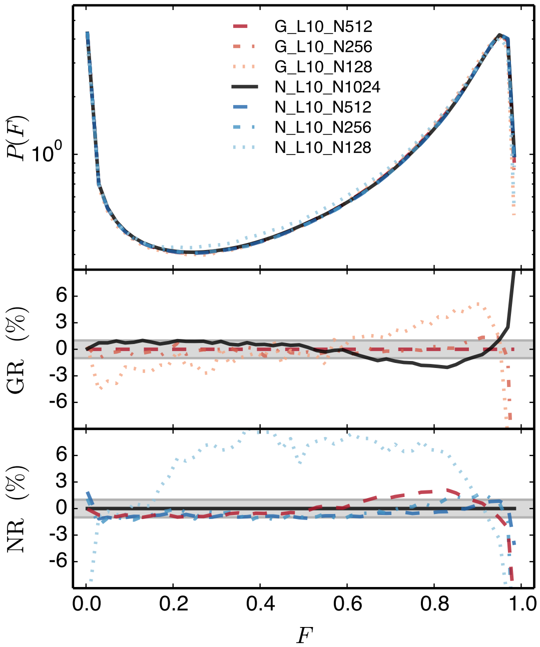

Figure 3.12: The flux PDF $p(F)$ in the $10\,h^{-1}$Mpc simulations at $z = 2.5$, and their ratios. *Top:* the flux PDFs. *Middle:* the percent differences to G_L10_N512. *Bottom:* the percent differences to N_L10_N1024.

the results. In the case of power spectra, for instance, rescaling a small amount changes the amplitude much more than the slope. All comparisons between simulation results and data are done with rescaled fluxes as well, so testing with rescaled fluxes is more relevant in any case. To keep the rescaling constants small, we decided to scale all optical depths so that the mean fluxes match that of the mean flux of the simulation with the highest mean flux (N_L10_N1024 at all redshifts). That is, for the following flux statistics, we use $F' = e^{-\tau'}$, $\tau' = A\tau$, and $|A - 1| \ll 1$.

3.4.3 Flux PDF

We compute the flux PDFs using the full box, with 50 bins evenly distributed in $0 \le F < 1$. We show the flux PDFs at $z = 2.5$ in Figure 3.12. The flux PDF is generally peaked at the tails of the distribution, although the transmissive end of the distribution is not peaked at high enough redshift. For all redshifts considered in this chapter, the PDF shows a peak in the lowest bin due to saturated pixels. At high redshifts ($z \ge 3.5$), the distribution is heavily skewed to saturated fluxes and transmissive $F > 0.5$ pixels are very rare due to the larger opacity of the IGM. By a redshift of about 3, the distribution turns over, and there is another peak at $F \sim 1$, with a drop in the highest bin. The falloff at very high transmission is due to the fact that the optical depth cannot reach 0 anywhere. In both codes, the flux

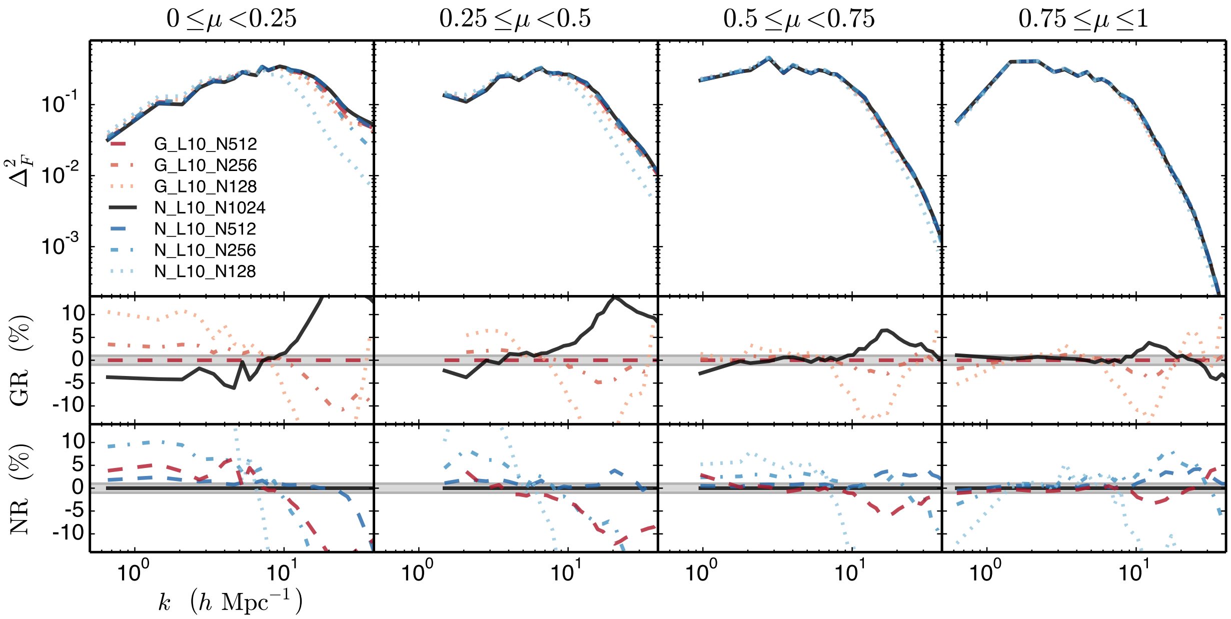

Figure 3.13: The dimensionless 3D flux power spectra $\Delta_F^2(k, \mu)$ in the 10 h^{-1}Mpc simulations at $z = 2.5$, and their ratios. *Top:* the dimensionless power spectra, binned in k and 4 μ bins (each μ bin plotted separately from left to right vs. k). *Middle:* the percent differences to G_L10_N512. *Bottom:* the percent differences to N_L10_N1024.

PDF values converge surprisingly quickly, but the codes converge to slightly different values, mostly driven by differences at the edges. The 256^3 runs appear to agree with the high-resolution result to a percent in almost all bins, with the most disagreement at the high-F end, where P falls off quickly. Between the codes, Nyx has a larger probability of $F \sim 0$ and $F > 0.9$, and the integral constraint pushes the intermediate probabilities down relative to the Gadget values. The difference at high F comes from a small difference in position of the probability peak at $F \sim 0.9$ between the codes, where the Gadget P falls off at lower F.

3.4.4 Power spectra

Spatial correlations of the Lyα forest flux provide an excellent probe of matter fluctuations on relatively small scales (compared to other cosmological probes). The main reason for this is that we can hope to study them *ab-initio*, since majority of flux fluctuations do not depend on complex small-scale astrophysics like galaxy formation, but is instead driven by weakly non-linear gas dynamics. Flux perturbations are usually defined as $\delta_F = \frac{F}{\langle F \rangle} - 1$, where $\langle F \rangle$ is averaged over all sightlines. We compute the power spectra using this definition of the flux perturbation, which is the standard used in observations.

The flux field is not isotropic, so it is not appropriate to simply average over k-space shells to compute a power spectrum. There are many ways to visualize the anisotropy of the power spectrum, but we follow the convention of using $P(k, \mu)$ where μ is the cosine of the angle between the mode and the LOS, i.e. $\mu = k_{\parallel}/k$. We chose to average the power over the same k bands as before and 4 μ bins. The resulting 3D flux power spectra are shown in Figure 3.13. From left to right, the power μ wedges shown are increasingly parallel to the LOS. That is, the panel on the left can be thought of as transverse power and the panel on the right as LOS power. Lower-resolution runs tend to have more large-scale power, with a larger tilt, so that they fall off at smaller k than higher-resolution runs. The rescaling operation changes this trend a bit, so that along the LOS, some of the lower-resolution runs do not have more large-scale power. The LOS power tends to agree much better, between resolutions and codes. This is due to the thermal broadening along the LOS, which acts to smooth out fluctuations in this direction. Still, it is reassuring that the thermal cutoff between the codes agrees well. The perpendicular perturbations do not have such a fudge factor, and show much more small-scale power, which acts to separate the different resolution simulations much more. The 512^3 simulations runs are converged to percents up to $k = 10$ h Mpc^{-1}, and agree with each other at large scales. Parallel to the line of sight, the codes agree to 2 percent, and perpendicular to the line of sight, the codes agree to 6% up to $k = 10$ h Mpc^{-1}. The main difference between the codes is that Nyx has more small-scale power, and the small-scale difference is more significant in transverse power. The position and magnitude of the excess power here matches the result in the thresholded baryon power in Figure 3.8. We believe this is a result of the high-density regime differences between the codes, which results in more small-scale growth in Nyx.

Another key Lyα forest observable is the 1D flux power spectrum. Although the 1D flux power spectrum can be computed by an appropriate integral over the three- dimensional power spectrum, it is easier and numerically more robust to extract it directly from the simulations. We compute the 1D power by taking the Fourier transform of δ_F along each skewer and averaging the power in bins of $k_{\parallel}$. We show the dimensionless 1D flux power spectra $\Delta_{F,\mathrm{1D}} = k_{\parallel}P/\pi$ in Figure 3.14. Again, lower-resolution runs tend to be steeper and fall off at smaller k. The best resolution results between codes agree decently well up to about 0.1 km^{-1} s (to 6%), although they clearly have different slopes. The G_L10_N512 result, for example, looks most like the unconverged N_L10_N256 result. The disagreement in the 1D power is driven by the disagreement in the transverse 3D power, which can be seen considering $P_{\mathrm{1D}}(k_{\parallel}) = \frac{1}{2\pi}\int_0^{\infty} P(k_{\parallel}, k_{\perp})k_{\perp}dk_{\perp}$.

3.4.5 Remaining differences

Our highest-resolution Nyx and Gadget simulation results agree well, but there are some remaining differences at the level of a few percent in all common flux statistics. We aim to find the origin for the small differences in this section. Many of the remaining differences are mentioned in the flux statistic sections, but we summarize here. In the mean flux, Gadget is lower at high redshift and evolves more than Nyx, so they are within a percent by $z = 2.5$. In the flux PDF, Gadget rises slightly faster than Nyx approaching high transmission, and Nyx has a higher $F \sim 1$ probability. In the flux power, Nyx has more transverse power on

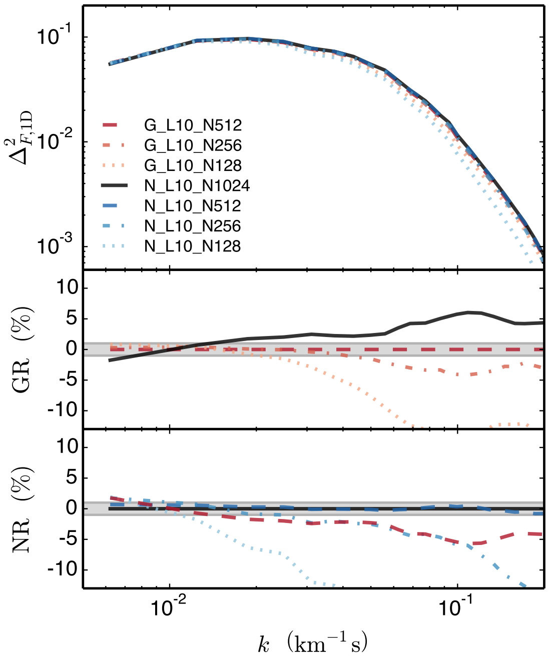

Figure 3.14: The dimensionless 1D flux power spectra $\Delta^2_{F,1\mathrm{D}}(k_\parallel)$ in the 10 h^{-1}Mpc simulations at $z = 2.5$, and their ratios. *Top:* the dimensionless 1D power spectra,. *Middle:* the percent differences to G_L10_N512. *Bottom:* the percent differences to N_L10_N1024.

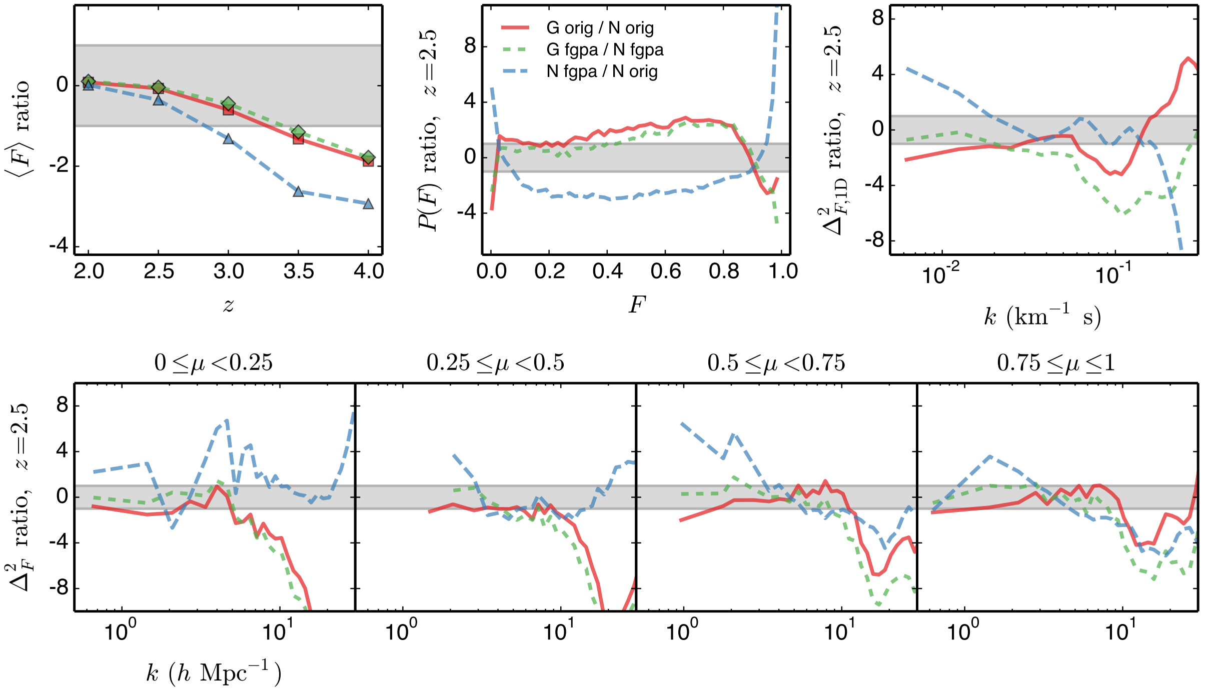

Figure 3.15: Flux statistics in the L10_N512 simulations using the original fluxes and FGPA-like fluxes. *Top:* the dimensionless 3D flux power spectra $\Delta^2_F(k,\mu)$, and percent differences. *Bottom:* the mean fluxes vs. z, the flux PDFs at $z = 2.5$, and the 1D flux power spectra at $z = 2.5$.

small scales ($> 0.6\ h^{-1}$Mpc).

In order to reduce the number of variables as much possible, we also tried making new, simplified flux fields with the L10_N512 simulations. We first removed the effect of high-density regions by thresholding to densities less than 30 times the mean. We chose a value of $\rho_\mathrm{b} = 30$ again based on the ρ_b PDF, since the codes have a similar distribution of densities below 30. For each cell in both simulations, if the density in both simulations exceeded 30, we set $\rho_\mathrm{b} = 30$. If the density in just one simulation exceeded 30, we used the minimum of either $\rho_\mathrm{b} = 30$ or the average of the two original densities. By considering the densities of both simulations at one time, we ensure the thresholded cells match between the codes. We know the high-density regions in Nyx are more concentrated and we wanted to prevent this from creating any differences. We then set the temperature of the gas using a fluctuating Gunn-Peterson approximation (FGPA; Croft et al. 1998; Gnedin & Hui 1998; Meiksin 2009), with an equation of state $\log T = \log T_0 + (\gamma - 1) \log \rho_\mathrm{b}$, with parameters matching what was found in the simulations, $T_0 = 11000$ K and $\gamma = 1.55$. We then recalculated $n_{\mathrm{H\,I}}$ and computed τ the normal way. This process of constructing what we call the 'FGPA-like' flux removes differences originating in high-density regions, differences in the ρ_b-T scatter and even the ρ_b-T line since they are now the same. One possible remaining difference is the effect of the peculiar velocity, since we did not mask that field. We tried repeating the following analysis with the real-space optical depths ($v_{\parallel} = 0$) and found virtually no difference from the redshift-space results, indicating that the $v_{\parallel}$ differences are insignificant.

In Figure 3.15, we compare the original flux stats with the FGPA-like flux stats. In the top panel, we show the 3D flux power $\Delta_F^2(k, \mu)$ and in the bottom panel, from left to right, we show the mean flux at $z = 2$, 2.5, 3, and 3.5, the flux PDF at $z = 2.5$, and the 1D flux power at $z = 2.5$. For each of the statistics, we also show ratios between the original flux results and the ratios between the FGPA-like flux results. The reason for this is to check if the differences between codes remain the same after simplifying to FGPA-like flux, which should only depend on the density. In the 3D flux power ratio panels, for instance, we show the Gadget over Nyx original ratios, the Gadget over Nyx FGPA-like ratios, and the Nyx FGPA-like over Nyx original ratios. The original flux ratio and FGPA-like ratio lines tell us if the code differences are preserved between the two flux fields, while the Nyx original to FGPA-like flux ratio line shows us the difference the FGPA-like flux makes.

The two Gadget to Nyx ratio lines are very similar, indicating that the FGPA process has not removed the differences between the codes. Although we have removed high-density regions in the FGPA-like flux, we cannot remove the effect of high-density regions on the evolution of lower densities. Especially in the transverse power, we see that Nyx FGPA-like still has about 10 per cent more power than Gadget FGPA-like on small scales around $k = 20\,h\,\mathrm{Mpc}^{-1}$. In the mean flux, flux PDF, and 1D flux power panels, we see the same pattern that the original and FGPA-like ratios lines agree very well. The FGPA-like flux fields reproduce the mean flux evolving more in Gadget, and has almost the same amplitude. In the flux PDF, the Gadget FGPA-like flux also has lower edge probabilities, although the $F \sim 0$ probability is a bit higher. We found that the agreement around $F = 0.2$ depends on the density thresholding, and that the ratios agree better as we remove the density threshold. Although high-density regions are mostly hidden in saturated flux regions, they appear to have a small contribution to the shape of the low-transmission probability. The agreement in the 1D flux power is not as clear as in other panels, but we see similar features of the lower power on large scales and the dip at $k = 10\,h\,\mathrm{Mpc}^{-1}$. The fact that the ratio between Nyx and Gadget results changes very little going from the full fluxes to the FGPA-like fluxes suggests that differences between the codes is driven by the differences in the baryon density, rather than differences in temperatures and velocities.

Overall, we found that the agreement between the original and FGPA-like flux statistics is increasingly better at high redshift. This agrees with the intuition that there is less scatter in the density-temperature relation and less mass at high densities at higher redshifts. In addition, the flux signal at higher redshift is sensitive to lower densities, so that the high-density regions contribute much less to the signal. As the mean flux increases, the flux signal becomes more sensitive to high densities, bringing out the differences in how the codes handle those regions.

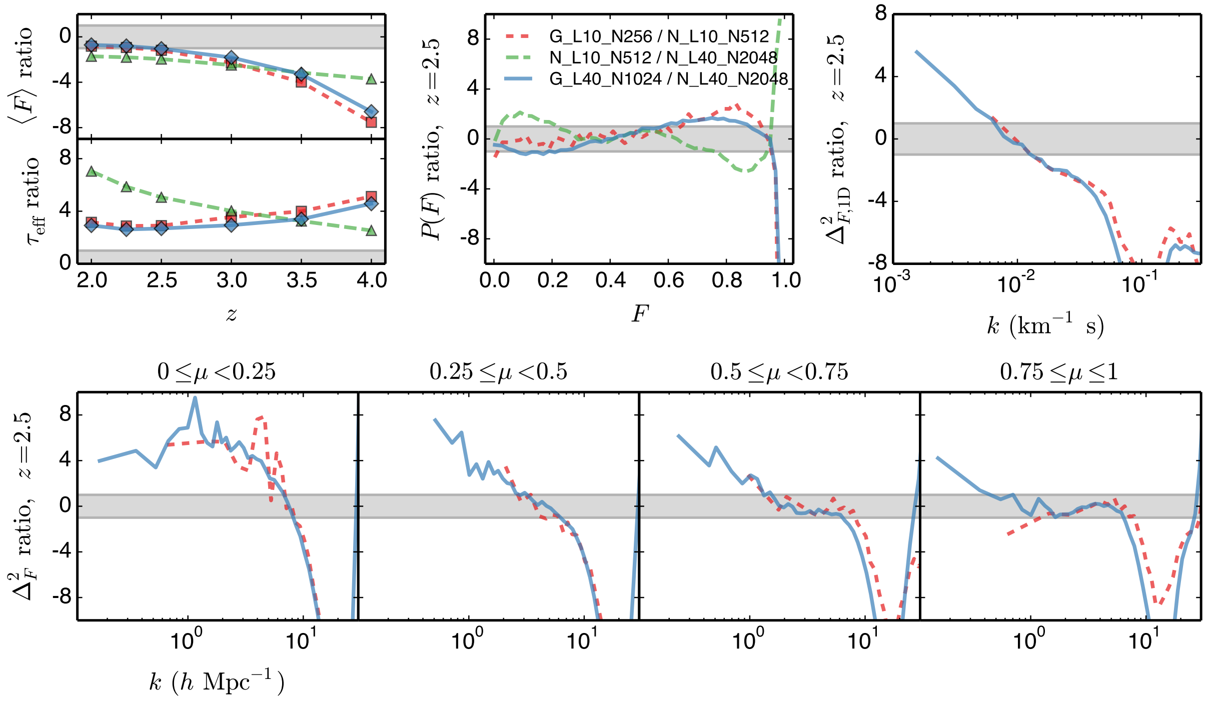

Figure 3.16: Combined flux statistics from the 40 h^{-1}Mpc runs, along with results from the 10 h^{-1}Mpc of the same spatial resolution. Top left: mean flux vs. redshift. Top middle: flux PDF. Top right: 1D flux power. Bottom: 3D flux power.

3.5 Large-box results

So far, we have only used flux statistics from our small box simulations. Although the 512^3 simulations appear converged with respect to resolution, they are definitely not converged in box size. By $z = 2$, the box scale mode is significantly nonlinear, and the simulations are missing growth from modes larger than the box size. Missing large scale modes outside the small box scale also has significant effects on the bulk flows in the simulations. This is generally the least converged quantity in small box simulations, and results in a lower temperature distribution of shocked gas and smaller peculiar velocities (affecting the redshift-space distortions). In order to compare more representative predictions from both codes, we also compare flux statistics from larger box simulations at the best resolutions available for each code. In this case, we have a 40 h^{-1}Mpc simulation from both codes, where the Gadget run has 2×1024^3 particles, and the Nyx run has a 2048^3 grid.

In Figure 3.16, we show the combined flux statistics in the 2 large box simulations, along with the matching resolution small box simulations for comparison. The codes agree to 3 percent at $z = 3.5$ and are within a percent by $z = 2.5$, but the large box simulations are 2 to 3 percent higher at all redshifts. The 40 h^{-1}Mpc simulations have a larger mean flux than the 10 h^{-1}Mpc, and the box size error appears to be roughly constant over redshift (a bit larger at high redshift). This is clear in the mean flux ratio plot, where the 10 h^{-1}Mpc G/N values (red) and the 40 h^{-1}Mpc G/N values (blue) are similar. The Nyx 10 h^{-1}Mpc over Nyx h^{-1}Mpc ratio values (green) show the magnitude of the box size error. In the flux PDF, we see a similar trend that the difference between the codes at 10 and 40 h^{-1}Mpc is similar. In both cases we see that the code error, namely difference between the results of two codes taken as an estimate of the difference with respect to the truth, is independent of box size.

When comparing the power spectra, we find, both for 1D and 3D that at the scales shared between the 10 h^{-1}Mpc and 40 h^{-1}Mpc, the ratio between the two codes is the same to better than percent, again indicating that the code error is effectively independent of the box size. More worringly, however, it appears that the large scales in our $10h^{-1}$Mpc simulations just happen to lie correspond to the scales where the two codes agree very well and that the differences are again more pronounced for the largest scales probed by $40h^{-1}$ Mpc boxes. The general trend is that Gadget has less small-scale power and more large-scale power. The codes are about 4 per cent different for the largest scales, but the difference might be even bigger had we had resources to run even larger boxes.

3.6 Conclusions

We have performed a detailed comparison of the IGM as captured in the cosmological hydrodynamic codes Gadget and Nyx. The simulations performed in this work were not chosen to fit the data, but instead to be a typical problem with which we can thoroughly assess the numerical accuracy of the two codes under investigation.

Our results can be summarized as follows:

- The two codes agree within several percent precision, when converged results are compared. However, they can approach convergence from different directions and discrepancies can appear to be larger for runs with insufficient precision.
- The differences in the flux probability distribution are dominated by differences at the extreme fluxes, but the shapes of PDF agree better than per cent level for $0.1 < F < 0.9$.
- The main differences are mostly associated with different behavior of the two codes in the very dense regions. For Gadget, the particles are transformed to ‘star particles’ with an unrealistic precision, while in Nyx the lack of feedback results in a run-away overcooling process. While the real-space flux is zero in regions where this effect is significant, the surrounding regions are affected and the real- to redshift-space transformation accentuates these differences.
- When performing a FGPA-like transformation on the field, we noted that relative differences remained the same between the codes even though this uses a different

optical depth – this suggests that the differences seen in the flux statistics originate in the distribution of baryons.

- Both three and one-dimensional power spectra show differences that are around a few percent on scales around $k = 1\,h^{-1}$Mpc, but can be considerably larger (10 per cent level) for much smaller scales. The same is true for larger scales. In the $40\,h^{-1}$Mpc boxes, the differences are around 4 per cent on large scales but naive extrapolation would indicate it would be even larger on larger scales. This implies that the inferred large-scale density and velocity bias parameters for the Lyα forest are inferred with similar percent-level accuracy.

We have shown that choice of the code and hence the numerical method leads to irreducible uncertainties that are not negligible compared to the current level of precision in the measurements of the one-dimensional power spectrum. These uncertainties are mostly associated with how the codes treat the high-density regions and in data-fitting comparison they will be naturally marginalized over as part of marginalization over unknown astrophysics. However, at the same time, results in this work can and should be used to create new templates that will allow a careful fitting operation to marginalize over the uncertainties in prediction associated with a choice of a given code.

Acknowledgments

This work is in preparation as Stark et al. (2015b). In addition to thanking my collaborators on this project, Nishikanta Khandai, Zarija Lukić, Peter Nugent, Anže Slosar, and Martin White, I would like to thank Pat McDonald and Matt McQuinn for useful discussions. This work was supported in part by the Scientific Discovery through Advanced Computing (SciDAC) program funded by the US Department of Energy Office of Advanced Scientific Computing Research and the Office of High Energy Physics.

Chapter 4

Finding protoclusters using Lyα forest tomographic flux maps

Galaxy clusters are the largest and most massive gravitationally-bound structures in the Universe, the endpoint of a long process of hierarchical structure formation. Due to their large mass, deep potential wells and dynamic formation histories they are important laboratories for studying galaxy evolution, plasma physics, and our models of gravity and cosmology (Fabian 1994; Kravtsov & Borgani 2012; Feretti et al. 2012). Despite keen interest in how clusters form, the study of early cluster formation, at high z, is observationally limited: clusters are rare and surveying large volumes is expensive. Indeed, the total comoving volume of even the largest surveys for distant galaxies at $z \sim 2$–3 (e.g. KBSS, Rudie et al. 2012) is only $\sim 10^7\,\mathrm{Mpc}^3$, which would barely contain a single rich cluster locally. In the past decade small samples of protoclusters have been compiled (see e.g. Chiang et al. 2013, 2014, for a recent compilation) but important questions regarding the formation of clusters and the evolutionary tracks of member galaxies remain unresolved (e.g. Peterson & Fabian 2006; Dolag et al. 2009; Martizzi et al. 2014). There has been progress in the theoretical understanding of cluster formation through the use of N-body simulations and semi-analytic galaxy formation models (Baugh 2006; Benson & Bower 2010; Benson 2012) and hydrodynamical simulations (Sijacki & Springel 2006; McCarthy et al. 2010; Yang et al. 2012; Skory et al. 2013; Vogelsberger et al. 2013; Genel et al. 2014), although it is a notoriously difficult problem to predict member galaxy properties from first principles. It is an area of ongoing research to validate and extend the numerous assumptions and subgrid recipes which are made in these works.

With the advent of large surveys in the optical and near-IR (Postman et al. 1996; Kochanek et al. 2003; Gladders & Yee 2005; Koester et al. 2007; Wilson et al. 2009; Muzzin et al. 2009; Hao et al. 2010; Szabo et al. 2011; Murphy et al. 2012; Rykoff et al. 2014; Bleem et al. 2014), sub-mm (Marriage et al. 2011; Reichardt et al. 2013; Planck Collaboration et al. 2013b) and X-ray (Ebeling et al. 2000; Böhringer et al. 2004; Burenin et al. 2007; Pierre et al. 2006; Finoguenov et al. 2007) bands, we now have large samples of clusters, with a tail extending beyond $z \simeq 1$–2. These surveys leverage the fact that ‘mature’ clusters contain large overdensities of (typically red) galaxies and a hot intracluster medium. Protoclusters, at $z = 2$ or earlier, lack these signatures making them more difficult to find and study. At the time of writing only a few tens of protocluster candidates are known at $z > 2$ (see Chiang et al. 2013, 2014; Finley et al. 2014; Cucciati et al. 2014), and most candidates were found via the the signpost technique, i.e. using a radio-galaxy, Lyα blob, or another source as a marker.

Assuming a mean interior density of 200 times the background density, the linear size of the mean-density region from which material is accreted into a halo should be about 5–6 times the virial radius of the final halo. For protoclusters this can be up to $10\,h^{-1}$Mpc, i.e. we expect that the $z \sim 2$ progenitors of massive clusters should lie in overdense regions many (comoving) Mpc in radius. This expectation is born out of numerical simulations (e.g. Chiang et al. 2013) which also show that the most massive clusters today form not from the most overdense regions at high z but from large, possibly only moderately overdense regions (Angulo et al. 2012). The progenitor regions of massive low-z clusters should thus be identifiable in relatively low-resolution large-scale structure maps of the high-z Universe. Systematic searches in large, deep, galaxy redshift surveys or multi-band photometric surveys are one promising way to find protoclusters (e.g. Chiang et al. 2014; Diener et al. 2014; Yuan et al. 2014), although projection effects pose a challenging problem. Spectroscopic surveys with sufficient sampling of Mpc-scales take care of this problem, although redshift errors can still be significant and covering large volumes with such high resolution is prohibitively expensive.

An alternative is tomographic mapping using Lyα absorption from neutral Hydrogen in the intergalactic medium (IGM) (Caucci et al. 2008; Cai et al. 2014; Lee et al. 2014a). Lee et al. (2014a) demonstrated that IGM tomography allows large volumes of the Universe to be efficiently searched for protoclusters in the $z \simeq 2$–3 range using existing facilities. By targeting star-forming Lyman-break galaxies (LBGs) as well as quasars at $g \gtrsim 24.5$, at signal-to-noise ratios achievable with existing facilities, we can obtain hundreds of sightlines per deg^2. This sightline density corresponds to average spacings of several Mpc, which is also the correlation scale of the Lyα forest. By sampling the IGM absorption along and across sightlines with Mpc spacing, we are able to tomographically reconstruct the 3D Lyα forest flux field. These tomographic maps have a resolution similar to the average transverse sightline spacing and naturally avoid projection effects or redshift errors. In Lee et al. (2014b), we constructed a tomographic IGM map using 24 LBG spectra with an average sightline separation $\langle d_\perp \rangle = 2.3\,h^{-1}$Mpc, obtained with two 2-hour exposures on Keck LRIS. These observations made up the pilot data of the COSMOS Lyman-Alpha Mapping And Tomography Observations (CLAMATO) survey, which we plan to extend to cover 1 deg^2. These observations will result in a tomographic map with a volume of roughly $70 \times 70 \times 230\,h^{-1}$Mpc and $2\,h^{-1}$Mpc resolution. Such a map will provide an unprecedented view of the intergalactic medium and provide a large volume to search for protoclusters.

Given the diversity of protoclusters, the ability to construct large samples is important if we are to draw robust conclusions about cluster formation. Optical, sub-mm and X-ray facilities could then be used to follow up the most promising candidates looking for galaxy over densities, Compton decrements or faint, diffuse X-ray emission. The HETDEX (Hill et al. 2004) and Subaru Prime Focus (Takada et al. 2014) spectrographs would be particularly powerful for following up such candidates in the optical. As we shall show later, the most massive progenitors of the most massive clusters today ($M > 3 \times 10^{14}\,h^{-1}\mathrm{M}_\odot$) can reach the rich group scale ($M \sim 3 \times 10^{13}\,h^{-1}\mathrm{M}_\odot$) before $z \sim 2$. Such structures could well have observable galaxy overdensities and a hot gas component at early times.

4.1 Simulations

In order to validate our protocluster search strategy, and to study the purity and completeness of the sample we obtain, we make use of cosmological N-body simulations. We require a simulation which simultaneously covers a large cosmological volume (to have a statistically fair sample of the rare clusters and protocluster regions) while having a sufficiently small inter-particle spacing to model transmission in the IGM. The requirements are sufficiently demanding that we have used a pure N-body simulation, augmented with the fluctuating Gunn-Peterson approximation (FGPA; Petitjean et al. 1995; Croft et al. 1998; Meiksin & White 2001; Meiksin 2009). This simulation was also used in Lee et al. (2014b).

4.1.1 N-body simulation

Our simulation employed 2560^3 equal mass ($8.6 \times 10^7\,h^{-1}M_\odot$) particles in a $256\,h^{-1}$Mpc periodic, cubical box leading to a mean inter-particle spacing of $100\,h^{-1}$kpc. This is sufficient to model the large-scale features in the IGM at $z \simeq 2-3$ using the FGPA (Meiksin & White 2001; Rorai et al. 2013) and more than sufficient to find clusters at $z = 0$. The assumed cosmology was of the flat ΛCDM family, with $\Omega_\mathrm{m} \simeq 0.31$, $\Omega_\mathrm{b}h^2 \simeq 0.022$, $h = 0.6777$, $n_s = 0.9611$, and $\sigma_8 = 0.83$, in agreement with Planck Collaboration et al. (2013a). The initial conditions were generated using second-order Lagrangian perturbation theory at $z_\mathrm{ic} = 150$, when the rms particle displacement was 40 per cent of the mean inter-particle spacing. The particle positions and velocities were advanced to $z = 0$, using a TreePM code (White 2002) assuming a spline-softened force with a Plummer equivalent smoothing length of $3\,h^{-1}$kpc. This TreePM code has been compared to a number of other codes and shown to perform well for such simulations (Heitmann et al. 2008).

4.1.2 Halo catalogs

At $z = 0$ and $z = 2.5$, we generated halo catalogs using a friends-of-friends (FoF; Davis et al. 1985) algorithm with a linking length $b = 0.168$. This algorithm partitions particles into groups bounded approximately by isodensity contours of roughly 100 times the mean density (e.g. Lacey & Cole 1994; White 2001, and references therein). Since we focus only on the most massive objects in our simulations, FoF halos are sufficient – more sophisticated halo finding methods will recover more detailed halo and subhalo properties, but with increased complexity and computational cost[1].

4.1.3 Lyα flux field

For the output at $z = 2.5$, we also generated mock Lyα forest spectra on a 2560^3 grid with the FGPA. This approximation makes use of the fact that adiabatic cooling of the gas in the presence of a photoionizing ultraviolet background leads to a tight density-temperature relation in the low density gas responsible for the Lyα forest seen in absorption against bright objects (Gnedin & Hui 1998; Meiksin 2009). The approximation has been shown to match more detailed hydrodynamical computations at the ten percent level (Meiksin et al. 2001; Viel et al. 2002; McDonald 2003; Viel et al. 2006), and is certainly sufficient for our purposes.

The dark matter particle positions and velocities were deposited onto the grid using CIC interpolation (Hockney & Eastwood 1988). We then Gaussian filtered the density and velocity on the grid in order to approximate the pressure smoothing which affects the gas density. We assumed an IGM temperature at mean density $T_0 = 2 \times 10^4\,\mathrm{K}$, which gives a filtering scale of about $100\,h^{-1}$kpc at the redshifts of interest here (e.g. Gnedin & Hui 1998; Viel et al. 2002; White et al. 2010; Rorai et al. 2013). Our results are largely insensitive to the details of this pressure smoothing procedure, since we are probing fluctuations on much larger scales (Mpc). We set the temperature according to the density-temperature relation $T = T_0(\rho/\bar{\rho})^{\gamma-1}$, with a standard choice for the equation of state parameter $\gamma = 1.6$ (Lee et al. 2015). We compute the optical depth to H I Lyα scattering τ and the transmitted flux fraction $F = e^{-\tau}$ assuming the H I density is proportional to the ratio of the recombination and photoionization rates $n_{\mathrm{HI}} \propto \rho^2 T^{-0.7} \Gamma^{-1}$ and that the line profile is a Doppler profile, and we normalize the optical depth such that the mean flux $\langle F \rangle = 0.8$, matching the recent observational result in Becker et al. (2013) for this redshift. This scheme ignores several phenomena that could affect the Lyα forest including spatial fluctuations in the temperature of the IGM due to reionization inhomogeneities, spatial fluctuations in the ultraviolet background due to the shot noise of sources, and galactic outflows. Fortunately, at the Mpc scale, the effects of galactic outflows and temperature fluctuations on flux should be rather small, while we expect the ultraviolet background to fluctuate on scales of several hundred Mpc (McDonald et al. 2005a; Greig et al. 2014; Pontzen 2014; Gontcho et al. 2014). In the remainder of the chapter, when we refer to flux, we mean the Lyα forest transmitted flux fraction perturbation $\delta_F = F/\langle F \rangle - 1$.

The final products we use from the simulation, then, are the halo catalogs at $z = 0$ and 2.5, including the positions of the particles within those halos at $z = 2.5$, and 3D grids of density and flux. We begin by studying the relationship between this ideal flux field and the halos and protoclusters. In Section 4.5, we will look at the impact of finite sightline density, resolution and noise on the recovery of the flux field.

---

[1] For a recent review and comparison of halo finding methods see Knebe et al. (2011).

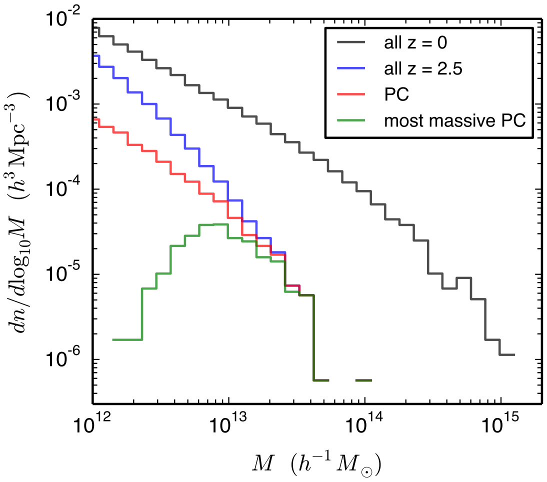

Figure 4.1: The halo mass functions for all halos at $z = 0$ (black), all halos at $z = 2.5$ (blue), protocluster halos (red), and the most massive halos in each protocluster (green). The massive end of the high redshift mass function is dominated by protocluster halos. The most massive halo in a protocluster is typically $10^{13}\,h^{-1}\mathrm{M}_\odot$ at this redshift.

4.2 Protoclusters in density and Lyα forest flux

The boundary between a rich group and a cluster is somewhat arbitrary, but we shall define a cluster at $z = 0$ as a halo with a FoF mass larger than $10^{14}\,h^{-1}\mathrm{M}_\odot$. We have 425 halos above this mass in the simulation at $z = 0$ which form our sample. A protocluster is the high-redshift progenitor of such massive halos, but due to the hierarchical process by which halos form, there is some ambiguity as to what constitutes the progenitor. At $z \simeq 2-3$, the mass which will eventually lie within the $z = 0$ halo is spread among several relatively large progenitor halos and in the nearby IGM, spread over tens of (comoving) Mpc. We tracked the cluster progenitor halos by finding all halos at $z = 2.5$ that contributed half or more of their mass to the resulting cluster. We show the mass functions of all $z = 2.5$ halos, of protocluster halos, and of the most massive halo in each protocluster in Figure 4.1. The high-mass end of the mass function is dominated by the halos that form clusters, but the protocluster halos do not make up all of the high mass halos. Protocluster halos only make up about half of the halos near $10^{13}\,h^{-1}\mathrm{M}_\odot$ for instance. We found that the most massive progenitor halo is typically about $10^{13}\,h^{-1}\mathrm{M}_\odot$, with more massive clusters having more massive progenitor halos on average. Only the most massive such halos are likely to host a hot, X-ray emitting ICM or be found as significant overdensities of galaxies. We also computed the second moment of the progenitor halo positions $\sqrt{[\sum_i m_i(\mathbf{x}_i - \bar{\mathbf{x}})^2]/[\sum_i m_i]}$, where $\bar{\mathbf{x}}$ is the average position and m_i and $\mathbf{x}_i$ are the halo masses and centers, as done in Chiang et al. (2013) to confirm the extent of the halos they found at this redshift. We found that the progenitor halos are spread over $4-8\,h^{-1}$Mpc, in good agreement with their values at $z = 2-3$. However, in contrast to Chiang et al. (2013), we are interested less in the

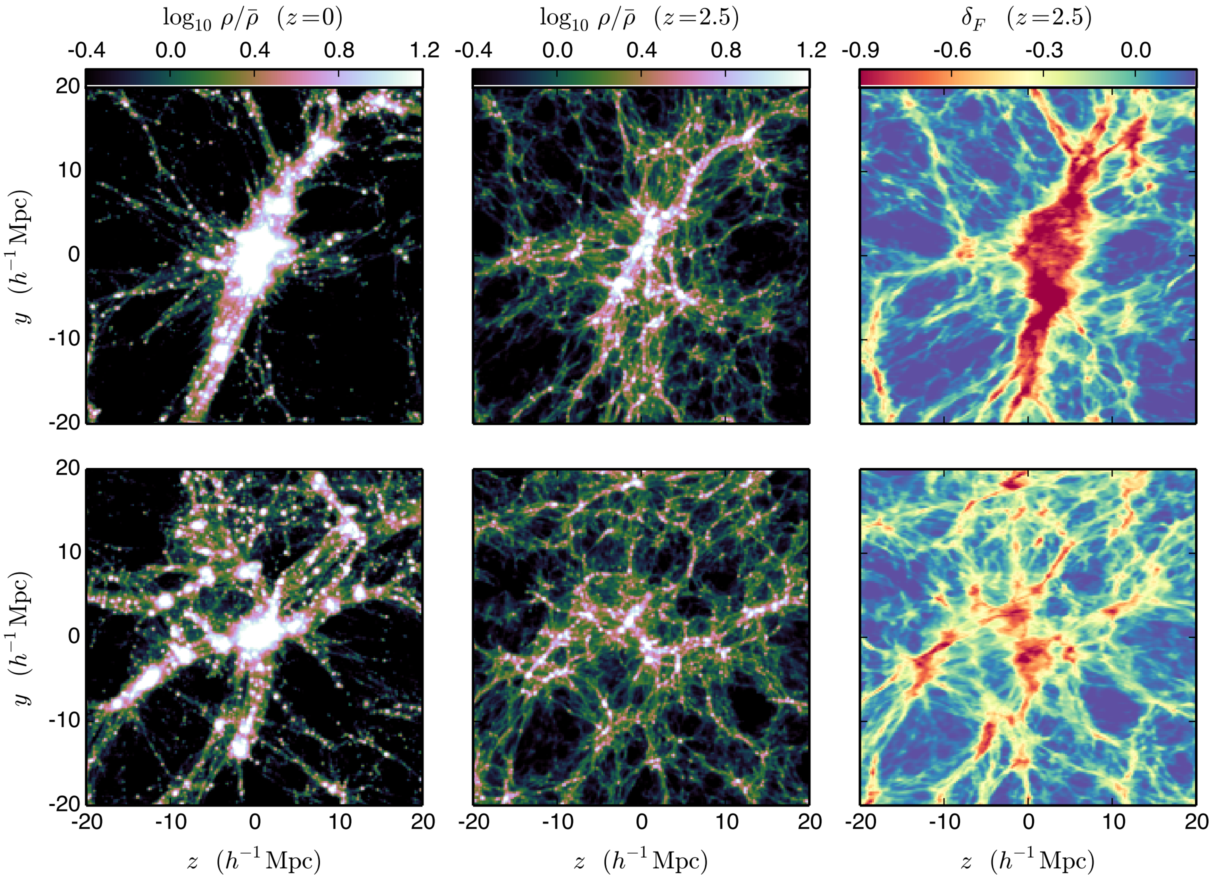

Figure 4.2: Slices through the density and flux fields centered on two protoclusters. The line-of-sight direction is horizontal. The upper row shows a cluster/protocluster which is easily found by our method, while the lower row shows a more problematic case. The upper row cluster has a mass $M = 9 \times 10^{14}\, h^{-1}\mathrm{M}_{\odot}$, while the lower row cluster has a mass $M = 3 \times 10^{14}\, h^{-1}\mathrm{M}_{\odot}$. We chose these sample protoclusters based on the δ_F value at the protocluster center of mass (COM), where the top protocluster has the smallest δ_F value, and the bottom protocluster has the largest. The slices are $40 \times 40\, h^{-1}$Mpc on a side and $5\, h^{-1}$Mpc thick. In each row, the color scale shows the log overdensity or flux perturbation: (Left) the $z = 0$ density (centered on the cluster COM), (Middle) the $z = 2.5$ density (centered on the protocluster COM), (Right) the $z = 2.5$ flux perturbation, δ_F. Note that overdense regions correspond to regions of increased absorption, or more negative δ_F, and that correlation is quite strong on these scales. The small differences in the $z = 2.5$ density and flux fields are due to the density is shown in real-space while the flux is in redshift-space.

progenitor halos and more in the large-scale overdense region from which the mass of the cluster will be assembled.

In order to define the protocluster center, we tracked particles that form the core of the $z = 0$ cluster back to $z = 2.5$, and computed their center of mass (COM). The choice of particles that constitute the 'core' of the cluster is arbitrary, but the exact choice of particles does not matter as long as the resulting COM does not change significantly. We chose to select the particles within $200\,h^{-1}$kpc from the most bound (densest) cluster particle at $z = 0$. We refer to this collection of particles that makes up the cluster core as the N-densest particles. We found that changing the cutoff radius from 100 to $500\,h^{-1}$kpc results in small changes to the protocluster center, on the level of $100\,h^{-1}$kpc, which is negligible for objects spanning several Mpc. Inspired by Chiang et al. (2013), we define the protocluster radius r_{pc} as the radius of a sphere, centered on the protocluster center, enclosing 50 per cent of the particles which belong to the halo at $z = 0$. We found the expected trend that more massive clusters have larger protocluster sizes. The 10th percentile radius is $3.3\,h^{-1}$Mpc, the 50th percentile is $4.1\,h^{-1}$Mpc, and the 90th percentile is $5.4\,h^{-1}$Mpc. The largest half-mass radius we found in the simulation is $8.9\,h^{-1}$Mpc, and this protocluster forms a $10^{15}\,h^{-1}\mathrm{M}_{\odot}$ cluster. This, in combination with the moment of the progenitor halo positions, gives us good reason to believe that protoclusters will stand out on scales of $\sim 4\,h^{-1}$Mpc at this redshift.

We show two examples of protoclusters, as seen in density and Lyα forest flux, in Fig. 4.2. The upper row shows a protocluster with a large coherent structure which will be easily found by our method, while the lower row shows a case where the protocluster is spread out and will prove much more difficult to find. The upper row cluster has a mass $M = 9 \times 10^{14}\,h^{-1}\mathrm{M}_{\odot}$, while the lower row cluster has a mass $M = 3 \times 10^{14}\,h^{-1}\mathrm{M}_{\odot}$. From left to right, we show the $z = 0$ density, $z = 2.5$ density, and the flux in a slice $40\,h^{-1}$Mpc across and $5\,h^{-1}$Mpc thick. Due to the physics of the IGM, the flux is tightly correlated with the matter density on large scales, with overdense regions leading to more absorption (low flux). In the protocluster in the upper row, the progenitor halos that merge to form the cluster can be easily seen in the middle column and lead to a large, coherent flux decrement in the right column. The flux decrement in the lower row is still visible, but it is not as pronounced, because the halos making up the protocluster are more diffuse. We compared the progenitor halos of these clusters and found that at fixed mass, the protocluster in the upper row has three times as many halos and that the most massive halo is twice as massive, indicating that the upper row cluster forms earlier. The most massive progenitor halo in the upper row cluster has a mass of $3 \times 10^{13}\,h^{-1}\mathrm{M}_{\odot}$ – a typical rich group mass – meaning that it should be easier to follow up at high redshift. Overall, we found that 40 per cent of the protoclusters contain a halo with a mass $M > 10^{13}\,h^{-1}\mathrm{M}_{\odot}$.

Not surprisingly, all of the protocluster regions lie on the high-density tail of the density distribution. We smoothed the density field with Gaussian filters of scales 2, 4, and $8\,h^{-1}$Mpc (labeled ρ_2, ρ_4, and ρ_8 respectively) and compared the distributions of the full field and the protoclusters. We smooth the fields for two reasons: to mimic the characteristic resolution of our tomographic maps and because protoclusters should stand out most on scales of several Mpc. The top panel of Figure 4.3 shows the probability density function of the density $p(\rho)$

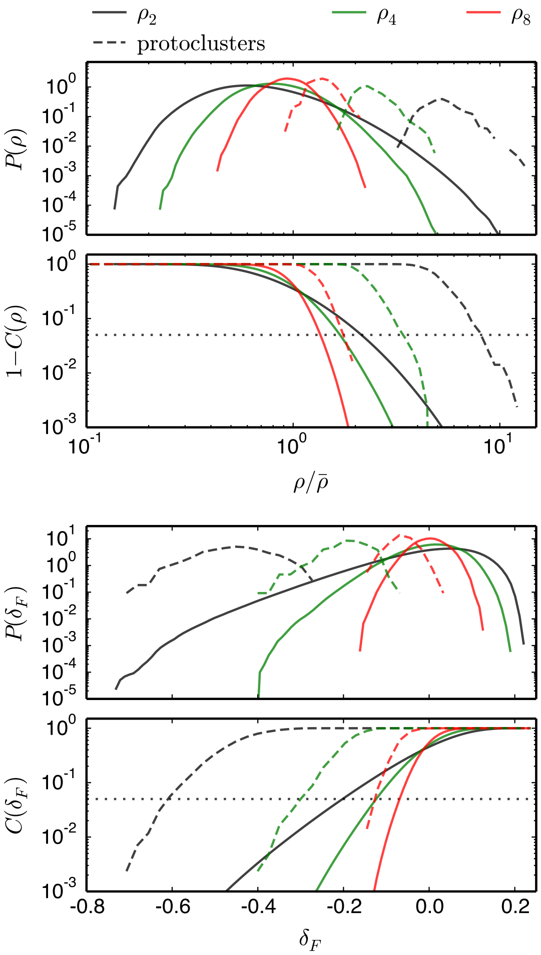

Figure 4.3: Distributions of the matter density and flux, smoothed with Gaussians of $\sigma = 2$, 4, and $8\,h^{-1}$Mpc (labeled ρ_2, ρ_4, and ρ_8 with the broadest distributions having the smallest σ). Solid lines show the PDF for the entire volume while the dashed lines indicate the densities or fluxes at the protocluster positions. The top two panels show the matter density PDF, $P(\rho)$, and the cumulative distribution, $C(\rho)$, plotted as $1-C$ to highlight regions of high density. The horizontal dotted black line shows the 95th percentile. We see that protoclusters preferentially lie in the highest density regions of the density field, smoothed on Mpc scales. The lower two panels show the PDF and cumulative distribution for the flux perturbation, δ_F. We see that protoclusters preferentially lie in the negative tails of the distribution.

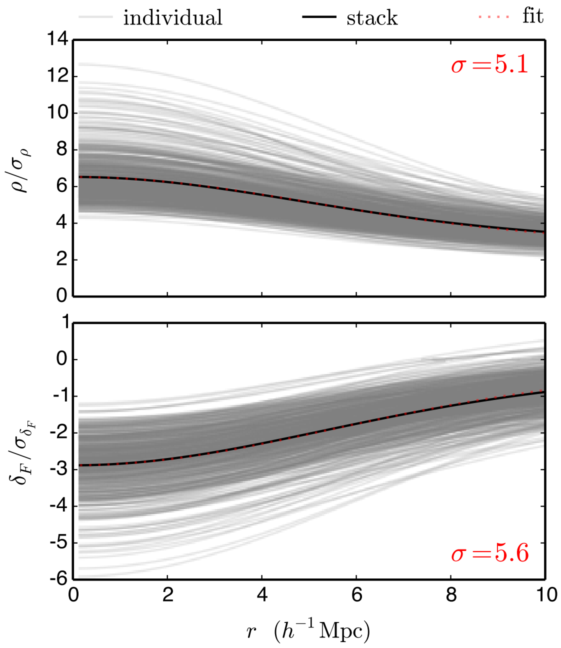

Figure 4.4: The radial density and flux profiles of the protocluster regions (from the $4\,h^{-1}$Mpc smoothed fields). We show the individual profiles in grey, the average profiles in black, and Gaussian fits in dotted red. (Top) the density profiles. We plot ρ/σ_ρ, where σ_ρ is the standard deviation of the field, since we are interested in how extreme the protocluster regions are. (Bottom) repeated with flux. The fit Gaussian scale σ is annotated in red.

for random positions (solid) and for the protocluster regions (dashed). The majority of the protoclusters have densities exceeding the 95th percentile of the density distribution. This is clearer in the second panel showing the cumulative distribution $C(\rho)$, plotted as log-scaled $1 - C$ to highlight the high-density tail. Here, it is easy to see the 95th percentile density for the field, and compare to the protocluster distribution. Regardless of the smoothing scale, nearly all protoclusters have densities in the 95th percentile tail. In the bottom two panels, we show the probability density and cumulative distribution of the flux. Since the large-scale flux is so tightly correlated with the density, we find that the majority of protoclusters similarly lie in the low-flux tail of the distribution. Protoclusters can thus be found quite efficiently by searching for large-scale flux decrements (see also Cai et al. 2014). In 1D, large-scale flux decrements can also be created by damped Lyα systems (DLA) (Meiksin 2009). However, DLAs have physical extents of $< 100\,h^{-1}$kpc, much smaller than our transverse scales, which make it very unlikely for DLAs to contaminate several nearby sightlines at the same redshift.

The radial profiles of the protocluster in density and flux are shown in Figure 4.4. These profiles were constructed by radially binning the $4\,h^{-1}$Mpc smoothed fields, from the center of each protocluster (grey lines) and by stacking all protocluster profiles (black lines). Again, we use the smoothed fields to mimic the tomographic map resolution and to highlight protocluster scales. On the y-axis in both panels, we plot the standard-deviation normalized values (where we use the standard deviation of the smoothed field) to see how much protocluster profiles stand out relative to other fluctuations at this scale. The overdensity and flux decrement near the center is significant. We found that the profiles are even more pronounced in the $2\,h^{-1}$Mpc smoothed fields, while in the $8\,h^{-1}$Mpc smoothed fields, the profiles are shallow, and do not stand out significantly in the center. This indicates that smoothing at a scale of $8\,h^{-1}$Mpc is likely too aggressive for our application. We fit Gaussian profiles to the average density and flux profiles, and show the fits with dotted red lines. We also annotated the fit Gaussian σ values, which indicate that the protoclusters are overdense/under-fluxed over several Mpc. These results validate our strategy for finding protoclusters by looking for large-scale flux decrements in the Lyα forest.

In Figure 4.5, we show three protocluster properties vs. the resulting cluster mass $M(z = 0)$. We plot the individual protocluster values with light gray dots, and $M(z = 0)$ binned results (with std. dev. error bars) in black. The red dashed lines show approximate scalings for each quantity. The top panel shows the protocluster half-mass radii, which scales with the cluster mass. We expect the half-mass and virial radii to scale similarly with mass. The red line shows the $r \propto M^{1/3}$ relation, which fits the protocluster sizes well. This falls in line with the expectation that more massive clusters form from larger overdense regions. The second panel shows the protocluster flux decrement $\delta_F/\sigma_{\delta_F}$, evaluated at the protocluster centers from the $4\,h^{-1}$Mpc smoothed flux field. In this case, the red line is entirely empirical. We noticed that the flux decrement scales roughly linearly with $\log M(z = 0)$ and found a good fit using $\delta_F \propto -2.9 \log M(z = 0)$. This means that more massive clusters stand out more significantly in the flux field, although the flux decrement from low mass clusters is not very significant. Some low mass clusters have decrements of only 1 or 2 σ, which are probably too difficult to distinguish from other background fluctuations. Clusters with a mass greater than $3 \times 10^{14}\,h^{-1}\mathrm{M}_{\odot}$, however, mostly originate in regions that are greater than 3σ flux decrements. For this reason, we expect to focus on finding more massive protoclusters. Finally, in the bottom panel, we show the mass of the most massive protocluster halo. The red line shows the linear scaling $M(z = 0) \propto M(z = 2.5)$, although the cluster masses appear to grow a bit faster than this. Although there is significant scatter in this relationship, this confirms that more massive progenitor halos form the more massive clusters. This is similar to what Conroy et al. (2008) found, where halos roughly maintain mass rank order as they evolve from $z = 2$ to $z = 0$. Altogether, these trends suggest that finding progenitors of the most massive clusters will be easiest, because they host the most massive halos, their flux decrement is more significant, and because the decrement covers a larger volume. We check if this expectation holds up in Section 4.5.

This section contains a basic characterization of protocluster environments, but it is important to note that our protoclusters have a wide range of sizes, profiles and overdensities (see also Chiang et al. 2013). We have presented a simplified view of protoclusters focused on properties that will allow us to identify them in flux maps. The full picture of these environments is probably much more complex, as illustrated by the examples in Figure 4.2. Large

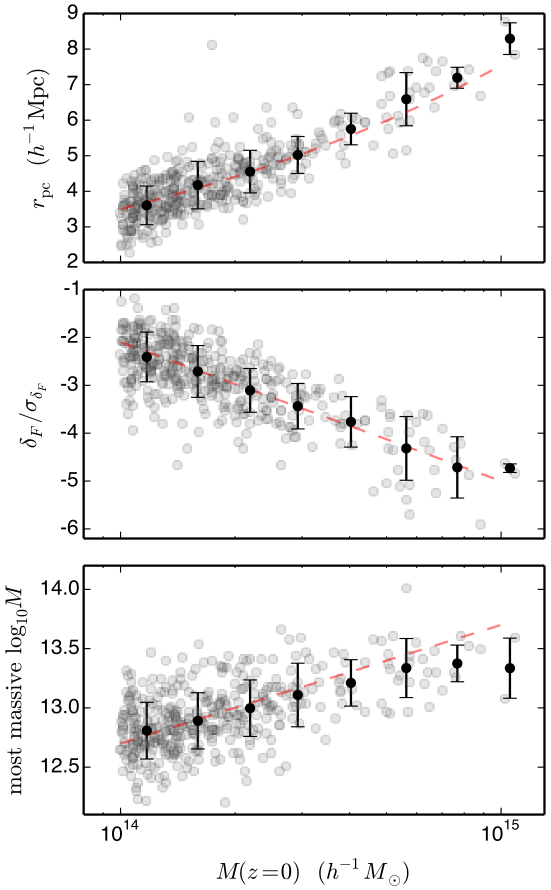

Figure 4.5: Several protocluster properties vs. the resulting cluster mass $M(z = 0)$. In each panel, we plot each protocluster as a gray dot, and the $M(z = 0)$ binned result with std. dev. error bars in black. The red dashed lines show an approximate scaling. Top: the protocluster half-mass radius $r_{\rm pc}$. The red line is the $r \propto M^{1/3}$ relation, which fits well. Middle: the flux decrement $\delta_F/\sigma_{\delta_F}$, evaluated at the protocluster center, smoothed with a 4 h^{-1}Mpc Gaussian. We show $\delta_F/\sigma_{\delta_F}$ on the y-axis to show how extreme the protocluster regions are. The protoclusters that stand out the most form the most massive clusters. The red line is an empirical fit of $\delta_F \propto -2.9 \log M(z = 0)$. Bottom: the mass of the most massive halo in the protocluster. The red line assumes linear growth, $M(z = 0) \propto M(z = 2.5)$. The high mass clusters appear to grow faster than the linear scaling, although this could be due to small numbers.

statistical samples are required to obtain a representative view of protocluster formation and the impact of the protocluster environment on galaxy formation and evolution.

4.3 Reconstruction method

We have argued that an efficient method for finding protoclusters is to look for large-scale decrements in the flux field. In this section we discuss how to make intermediate-resolution maps of the flux field, suitable for protocluster searches, from observations of closely-separated sightlines.

We use a Wiener filter (Wiener 1949; Press et al. 1992) to estimate the 3D flux field from the noisy observations along multiple sightlines, as advocated by Caucci et al. (2008); Lee et al. (2014a). The Wiener filter provides the minimum variance, unbiased linear estimator of the field (under the assumption of a normal distribution) and can be used to interpolate the data into regions which are not directly sampled, making it ideal for our purposes[2]. We briefly review the derivation of the Wiener filter, as we use it, in Section C.1, where we also describe our efficient numerical implementation. Collecting all of the observations of normalized flux into a data vector, $\mathbf{d}$, which is the sum of a signal and noise $\mathbf{d} = \mathbf{s} + \mathbf{n}$, the Wiener filter estimate of the signal at an arbitrary position is $\hat{\mathbf{s}} = \mathbf{L}\mathbf{d}$ with $\mathbf{L} = \mathbf{S}_{mp}(\mathbf{S}_{pp} + \mathbf{N})^{-1}$. Here $\mathbf{S}$ is the assumed signal covariance, where m and p indicate map or pixel coordinates, and $\mathbf{N}$ is the noise covariance. The reconstructed map is thus

$$\hat{\mathbf{s}} = \mathbf{S}_{mp}(\mathbf{S}_{pp} + \mathbf{N})^{-1}\mathbf{d} \tag{4.1}$$

Following Caucci et al. (2008), we model $\mathbf{S}$ as the product of two Gaussians for separations along and transverse to the line-of-sight:

$$S_{ij} = \sigma_F^2 \exp\left[-\frac{(\mathbf{x}_{\perp,i} - \mathbf{x}_{\perp,j})^2}{2l_\perp^2} - \frac{(x_{\parallel,i} - x_{\parallel,j})^2}{2l_\parallel^2}\right] \tag{4.2}$$

For the noise covariance, we assume that the pixel-to-pixel noise is independent, so that $N_{ij} = n_i^2\delta_{ij}$. These assumptions are approximations, but they are reasonably accurate in the context of the Lyα forest and the reconstruction is not sensitive to the form assumed (see tests in Section C.2). Assuming this form for the signal covariance and that the noise covariance is diagonal provides a huge advantage computational advantage, as it allows us to never store the matrices directly and instead compute them as needed. This reduces the space complexity of the algorithm from N^2 to N so that we can still fit large problems on a single node. We provide more details of our implementation in Section C.3.

In this work, we only discuss reconstructing the flux field since it is sufficient for our application of finding protoclusters. However, we note here that other authors have considered schemes to reconstruct the matter density in the context of galaxies as tomographic tracers

---

[2] See Pichon et al. (2001) for a more general method than Wiener filtering and Cisewski et al. (2014) for a non-parametric method

(Willick 2000; Kitaura et al. 2009; Courtois et al. 2012) and the Lyα forest (Kitaura et al. 2012) and how to account for redshift-space distortions in the reconstruction.

4.4 Protocluster identification

As shown in Section 4.2, protoclusters are significant outliers in density and flux on scales of several Mpc. In this section, we show how we can exploit this fact to identify protoclusters in the flux maps.

There are many ways we could test for large-scale outliers, but we start with a simple process of smoothing with a preferred scale and applying a threshold. We smooth the flux field with a 3D Gaussian filter, typically with a scale $\sigma = 4\,h^{-1}$Mpc. Since the protocluster profiles are roughly Gaussian with a similar scale, this acts much like a matched filter. We tried running this procedure with the different σ values of 2, 4, and 8 h^{-1}Mpc and found that the 4 h^{-1}Mpc version performs best. Next, we select all points below some threshold, and group nearby points together. The grouping process is also simple, where we merge points within 4 h^{-1}Mpc. This merging process ensures that we do not mistakenly break up low-flux regions and also that each region has a buffer from other regions. Finally, for each group of points, we define a protocluster candidate as a 4 h^{-1}Mpc sphere centered on the minimum flux point in the group. In principle, we could adjust the choices of the smoothing scale, merging distance, and candidate radius independently to optimize the candidate selection, but we found it was not necessary for our purposes, where this simple procedure already performs well.

We also tested a more advanced procedure for identifying protoclusters in the flux maps to be sure our protocluster efficiency was not too limited by our simple method though. We exploit the fact that we know the shape of the protocluster signal. We assume the flux map $d(\mathbf{x})$ is a combination of the protocluster signal and the background fluctuations of the Lyα forest. That is, $d(\mathbf{x}) = A\tau(\mathbf{x}) + \delta_F(\mathbf{x})$, where $\tau(\mathbf{x})$ is the shape of the protocluster profile and A is the strength of this signal.

In this case, the derivation of the optimal filter is shown in Appendix A of Haehnelt & Tegmark (1996), which we briefly review. We estimate the protocluster signal by convolving with a filter $\psi(\mathbf{x})$, so that $\hat{A} = \int \psi(\mathbf{x}) d(\mathbf{x}) d^3x$. The filter is normalized such that the estimate is unbiased, requiring $\int \psi(\mathbf{x})\tau(\mathbf{x})d^3x = 1$. In Fourier space, the unbiased, minimum variance estimator is then $\tilde{\psi}(\mathbf{k}) = C\tilde{\tau}(\mathbf{k})/P(\mathbf{k})$, where tildes indicate the Fourier transform of a quantity, $P(\mathbf{k})$ is the power spectrum of $\delta_F(\mathbf{x})$, and C is the normalization constant.

The optimal filter requires models for the protocluster profile and Lyα forest power spectrum. We model the protocluster profile as a Gaussian product parallel and perpendicular to the line of sight. This is similar to the data and model shown in Figure 4.4, although in this case we break spherical symmetry into the perpendicular and parallel components. We found that the average protocluster has a Gaussian σ scale of about 7 h^{-1}Mpc perpendicular to the line of sight, and is reduced to about 4 h^{-1}Mpc along the line of sight due to redshift-space distortions. We fit the Lyα forest power spectrum with a Kaiser and isotropic Gaussian-

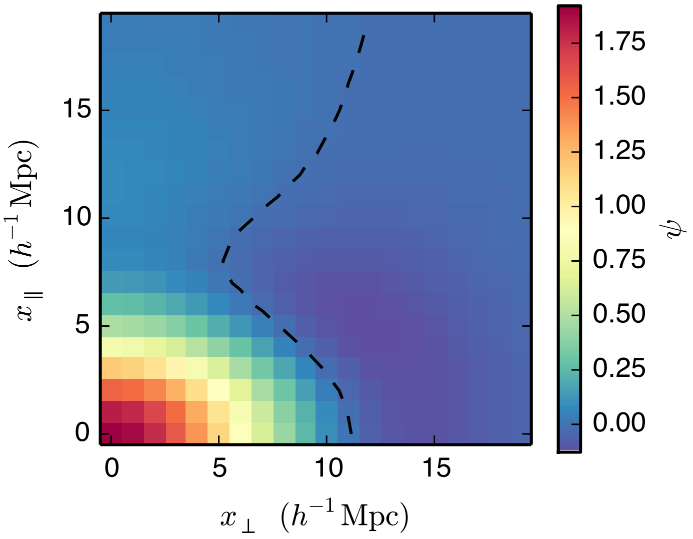

Figure 4.6: The optimal filter ψ in the $x_\perp, x_\parallel$ plane. The black dashed line shows the $\psi = 0$ contour.

damped redshift-space power spectrum model, $P(k_\perp, k_\parallel) = ak^\alpha(1 + \beta k_\parallel^2/k^2)^2 \exp(-k^2\sigma^2)$. The normalization of the power is set by a combination of the bias of the Lyα forest the normalization of the primordial power spectrum. The k^α term accounts for a simple form of the primordial power spectrum scaling, which is sufficient for the scales in the simulation. The Kaiser term $(1+\beta k_\parallel^2/k^2)$ handles the effects of redshift-space distortions on large scales. We include an isotropic, Gaussian damping term in order to capture suppression of small-scale fluctuations either due to pressure support or the smoothing effect of the Wiener filter. We found that the values $\alpha = -1.85$, $\beta = 1.07$, $\sigma = 2.06\,h^{-1}$Mpc provided a good fit. The resulting filter, in configuration space, is shown Figure 4.6. It is encouraging to see a negative region in the plot of $\psi(x_\perp, x_\parallel)$. This means that the filter will naturally downweight modes which are dominated by background Lyα forest fluctuations. This is an improved filter compared to the 3D Gaussian filter used in the rest of the text, but we found that it did not make a significant difference in the candidate identification result.

To get an idea of how this identification procedure performs, we first tested identifying protocluster candidates from an ideal flux field. We took the high-resolution flux field from the simulation, smoothed with a $4\,h^{-1}$Mpc Gaussian, and downsampled to a typical map resolution (grid spacing) of $1\,h^{-1}$Mpc. We chose a threshold of -3.5 times the standard deviation of the field, because we found this value performed best for finding protoclusters forming $> 3 \times 10^{14}\,h^{-1}\mathrm{M}_\odot$ clusters (see Figure 4.5). When we used more negative threshold values, we only found the most massive protoclusters, and when we used more positive threshold values, the protocluster purity decreased and very large protoclusters were mis-

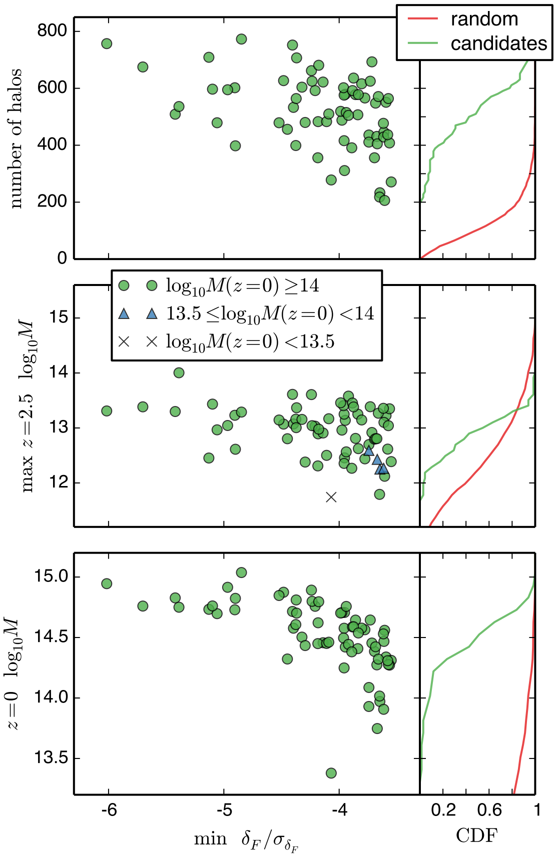

Figure 4.7: Halo statistics in $4\,h^{-1}$Mpc spheres centered on protocluster candidates (selected by flux decrement) compared to centered on random points. In each row, we plot the candidates as points on the left and on the right, we plot the cumulative distribution of the candidates and random points. Top: The number of halos in the sphere. Middle: The mass of the maximum mass halo, where the marker indicates the candidate category based on the $z = 0$ mass (green dots for clusters, blue triangles for rich groups, and black crosses for anything smaller). Bottom: The $z = 0$ mass of the maximum mass halo.

takenly merged. These threshold points make up about 10^{-3} of the simulation volume, and were grouped into 68 candidates.

For each candidate, we computed the number of halos in the $4\,h^{-1}$Mpc radius sphere and found the maximum mass halo within the sphere. We assigned each maximum mass halo to a $z = 0$ halo by tracking its particles to $z = 0$ and checking which $z = 0$ halo contained the most of its particles. The candidate is a protocluster if this $z = 0$ halo mass is $> 10^{14}\,h^{-1}\mathrm{M}_{\odot}$. We also computed these basic halo statistics for randomly positioned spheres to compare to a field distribution. The results are shown in Figure 4.7, where we plot the candidate number of halos, maximum $z = 2.5$ halo mass, and resulting $z = 0$ halo mass vs. the candidate $\delta_F/\sigma_{\delta_F}$ value. These panels clearly show that the low-flux selected candidates have large halo overdensities and are almost all protoclusters. In the top panel, we plot the number of halos in the protocluster candidate regions which shows that the candidates are all fairly rich environments. The candidate regions are on average 5 times the field median value and the cumulative distributions are well-separated. We note that our minimum halo mass is about $4 \times 10^{9}\,h^{-1}\mathrm{M}_{\odot}$, corresponding to the requirement that an FoF halo contains at least 50 particles. The middle panel shows the candidates' maximum-mass halo masses, with markers indicating the candidate category based on the $z = 0$ mass. The green circles are clusters, the blue triangles are nearly clusters, and the black cross is a failure. We also plot the $z = 0$ halo masses in the bottom panel. The four 'nearly' protocluster candidates have $z = 0$ masses of 5.6, 8.1, 8.5, and $9.4 \times 10^{13}\,h^{-1}\mathrm{M}_{\odot}$, while the failure candidate has a $z = 0$ mass of $2.4 \times 10^{13}\,h^{-1}\mathrm{M}_{\odot}$. In all panels, we see the expected trend that the more significant candidates (in terms of the minimum flux value) have richer environments and result in more massive clusters. This also illustrates how the identified protoclusters are more than overdense regions – they already host many galaxies and massive galaxies that can be followed up.

The candidate and random sphere cumulative distributions for the maximum $z = 2.5$ mass halos are particularly interesting. Half of the candidate maximum mass halos are in the mass range of $10^{12} - 10^{13}\,h^{-1}\mathrm{M}_{\odot}$ and the remaining half are in the $10^{13} - 10^{14}\,h^{-1}\mathrm{M}_{\odot}$ range. In the random distribution, half of the maximum mass halos are $< 10^{12}\,h^{-1}\mathrm{M}_{\odot}$, but there is a significant tail to high masses and the distributions cross at $3 \times 10^{13}\,h^{-1}\mathrm{M}_{\odot}$. We checked the total population of $3 \times 10^{12}\,h^{-1}\mathrm{M}_{\odot}$ halos at $z = 2.5$ and found that only 30 per cent end up in clusters by $z = 0$. This suggests that our identification procedure is not just picking out the most massive halos, but finds massive halos with the right environments to form clusters. This is supported by the cumulative distributions in the bottom panel, where the candidate and random position distributions are well-separated again. Despite our simple identification procedure, these results demonstrate that searching flux maps for large flux decrements is very effective for finding protoclusters.

We used a fairly conservative threshold value ($-3.5\sigma_{\delta_F}$) in order to achieve a high candidate sample purity of 93 per cent, compared to the random sample purity of 5 per cent. However, this comes with the cost of missing many of the low-mass protoclusters. We checked the candidate completeness vs. cluster mass, and found that above $3.5 \times 10^{14}\,h^{-1}\mathrm{M}_{\odot}$, the completeness is constant and around 80 per cent. Below this mass, the completeness falls off, reaching 50 per cent around $2.4 \times 10^{14}\,h^{-1}\mathrm{M}_\odot$, and 25 per cent around $1.5 \times 10^{14}\,h^{-1}\mathrm{M}_\odot$. As shown in Figure 4.5, we expect only more massive protoclusters (those forming $> 3 \times 10^{14}\,h^{-1}\mathrm{M}_\odot$ clusters) to stand out significantly in the flux maps, using a simple threshold at least. At the same time, the number of clusters quickly increases as we lower the mass, since these objects are on the tail of the mass function. This unfortunate combination drives our sample completeness to very small numbers for protoclusters forming low-mass clusters. However, for moderate-mass protoclusters (forming $> 3 \times 10^{14}\,h^{-1}\mathrm{M}_\odot$ clusters), the method performs well and successfully identifies 70 – 80 per cent of the population.

4.5 Mock surveys

In this section, we construct several tomographic mock surveys and run reconstructions on the synthetic data to test how our protocluster identification will perform on realistic data. Specifically, we are interested in what we can achieve with different values of the average sightline separation $\langle d_\perp \rangle$, as Lee et al. (2014a) demonstrated that this is the most important factor in determining the quality (effective SNR) in the reconstructed maps. Lee et al. (2014a) provides a simple relation between the exposure time t_exp, the minimum signal-to-noise ratio (SNR) per Å $\mathrm{SNR}_\mathrm{min}$, and the average sightline separation: $t_\mathrm{exp} \propto {\mathrm{SNR}_\mathrm{min}}^2 \langle d_\perp \rangle^{-1.6}$. We assume a fixed $\mathrm{SNR}_\mathrm{min}$ of 1.5, similar to the recent observations of Lee et al. (2014b), so that the exposure time is just a proxy for the desired average sightline separation. In principle, we could vary the sightline density and the SNR independently, but in practice this is not a useful test. If we increase the exposure time to build up the SNR, it is more advantageous (in terms of the reconstruction quality) to target fainter sources and increase the sightline density. We initially chose values of $\langle d_\perp \rangle = 2$, 2.5, 3, 4, and $6\,h^{-1}$Mpc. We expect that a resolution of $2\,h^{-1}$Mpc will be difficult but possible with existing instruments, while a spacing of $4\,h^{-1}$Mpc is fairly coarse, and we expected $6\,h^{-1}$Mpc to perform poorly for our application. We note that the sightlines in Lee et al. (2014b) have an average separation of $2.3\,h^{-1}$Mpc. When we found that the $\langle d_\perp \rangle = 6\,h^{-1}$Mpc separation run still performed decently, we added a survey configuration meant to mimic the Baryon Oscillation Spectroscopic Survey (BOSS) survey (Dawson et al. 2013). For the BOSS-like configuration, we chose an average sightline separation of $15\,h^{-1}$Mpc, which is roughly the spacing for the $200\,\mathrm{deg}^2$ of the survey with a source density of 1.5 – 2 times the mean.

We construct mock surveys using our full $(256\,h^{-1}\mathrm{Mpc})^3$ box. We first choose skewer positions by drawing random (x, y) coordinates in the box. We take the ideal F values along the skewer, smooth the signal based on a typical instrumental resolution $R = 1100$, and bin in pixel widths of 1.2 Å. We call this smoothed and binned flux F_inst. For each spectrum, we choose a constant per pixel SNR. We draw a random SNR value from a simple SNR distribution described below. Next, we realize noise for each spectrum based on its per pixel SNR value. For each pixel, we draw a random noise value from a normal distribution with scale $\sigma = \langle F \rangle / \mathrm{SNR}$. We add the noise vector F_noise to F_inst to get the final mock fluxes F_syn. Altogether, the input to the reconstruction includes the pixel positions $\mathbf{x}$, the data vector

$\mathbf{d} = \mathbf{F}_{\rm syn}/\langle F \rangle - 1$, and the noise vector $\mathbf{n} = \mathbf{1}/{\rm SNR}$.

We model the sightline SNR distribution as a power law, with a scaling based on the LBG luminosity function and the observed distribution in Lee et al. (2014b). We define the number of sightlines per deg^2 as $n_{\rm los} = (70\,h^{-1}{\rm Mpc}/\langle d_\perp \rangle)^2\,{\rm deg}^{-2}$ for our cosmology and $z = 2.5$. Our model is $dn_{\rm los}/d{\rm SNR} \propto {\rm SNR}^{-\alpha}$, and we want to determine values of α. Based on fits to the LBG luminosity function, Lee et al. (2014a) found that $d\log n_{\rm los}/dg$ is close to unity for the sources we are interested in, where g is the source g-band magnitude. Combined with the relation $d\log{\rm SNR}/dg = -2.5$, we have $\alpha = -d\log n_{\rm los}/d\log{\rm SNR} = 2.5$. This is a good approximation, but as we probe brighter in the luminosity function and sit more on the exponential tail, we know that $|d\log n_{\rm los}/dg|$ must increase. To correct for this, we take the SNR distribution from our pilot observations, rescale them based on ${\rm SNR}_{\rm new}/{\rm SNR}_{\rm obs} = (\langle d_\perp \rangle_{\rm new}/\langle d_\perp \rangle_{\rm obs})^{-0.8}$, and fit a power law. For our choices of $\langle d_\perp \rangle = 2$, 2.5, 3, and $4\,h^{-1}$Mpc, we found $\alpha = 2.7, 2.9, 3.5, 3.6$. For larger separations, we did not have enough bright sources in the pilot observations to reliably estimate α, so we kept $\alpha = 3.6$. We note that for large separations, we would also target more QSOs, which have a smaller $|d\log n_{\rm los}/dg|$ value at these magnitudes, and provides a natural maximum value for α.

Altogether, we ran 30 mock surveys and reconstructions. For each choice of $\langle d_\perp \rangle$, we ran 5 reconstructions to check how the results varied with a fixed ideal δ_F, but different skewer sampling and noise realizations. For all reconstructions, we fixed $\sigma_F^2 = 0.05$, $l_\parallel = 2\,h^{-1}$Mpc, and $l_\perp = \langle d_\perp \rangle$ as done in Lee et al. (2014a). The small-separation runs were much more time consuming than the large-separation runs since $N_{\rm pix} \propto \langle d_\perp \rangle^{-2}$ and the reconstruction algorithm scales with $N_{\rm pix}^2$, so that a run with a half the average sightline separation takes 16 times longer.

We tested the success of the surveys by running the protocluster identification procedure on the mock maps and comparing to the halo catalog, just as we did for the ideal field in the previous section. Again, we used a smoothing scale of $4\,h^{-1}$Mpc, a threshold of -3.5 times the standard deviation, and a region size of $4\,h^{-1}$Mpc. Overall, we found an good agreement between protocluster candidates in the ideal and reconstructed fields, and that the success rates decrease with increasing average sightline spacing, as expected. In Table 4.1, we list the number of candidates identified in each map, and the fraction of candidates that fell into classes of protoclusters (PC), nearly protoclusters (NPC), and failures (fail). These classes follow the definitions used earlier in Figure 4.7, where protoclusters form clusters ($M \geq 10^{14}\,h^{-1}{\rm M}_\odot$), nearly protoclusters almost form clusters ($10^{13.5}\,h^{-1}{\rm M}_\odot \leq M < 10^{14}\,h^{-1}{\rm M}_\odot$), and failures are anything less massive ($M < 10^{13.5}\,h^{-1}{\rm M}_\odot$). The mock results are averaged over the 5 survey realizations for each configuration. The number of candidates in the reconstructed maps is consistent with the result for the ideal field, although slightly higher, except for the BOSS-like survey which is much lower. If we scale the number of candidates ($N_{\rm cand} \sim 70$) found in the simulation volume of $(256\,h^{-1}{\rm Mpc})^3$ to the final CLAMATO volume of $70 \times 70 \times 230\,(h^{-1}{\rm Mpc})^3$, we should find 5 candidates. However, using a smaller threshold will yield many more candidates, if the decrease in purity can be accommodated.

There is a clear trend of the success rates vs. the average sightline separation. As the sightline separation increases, the map quality decreases, and the sightlines begin to miss

Table 4.1: Protocluster candidates and success rates

Map	t_{exp} (hrs)	N_{cand}	f_{PC}	f_{NPC}	f_{fail}
ideal	N/A	68	0.93	0.06	0.01
random spheres	N/A	1000	0.05	0.07	0.88
$\langle d_\perp \rangle = 2\,h^{-1}$Mpc	2.7	73	0.89	0.08	0.03
$\langle d_\perp \rangle = 2.5\,h^{-1}$Mpc	1.9	68	0.89	0.09	0.01
$\langle d_\perp \rangle = 3\,h^{-1}$Mpc	1.4	76	0.84	0.10	0.06
$\langle d_\perp \rangle = 4\,h^{-1}$Mpc	0.90	77	0.78	0.15	0.07
$\langle d_\perp \rangle = 6\,h^{-1}$Mpc	0.47	72	0.61	0.20	0.20
$\langle d_\perp \rangle = 15\,h^{-1}$Mpc	N/A	26	0.35	0.10	0.55

Protocluster identification success rates for the ideal δ_F field and randomly-positioned spheres compared to the mock survey reconstructions. t_{exp} is the corresponding exposure time to achieve the desired sightline spacing (rescaled from the Keck/LRIS setup in Lee et al. (2014b)). N_{cand} is the number of candidates found in the map and the f values are the fractions of candidates broken into three class: protoclusters (PC), nearly protoclusters (NPC), and failures (fail). The numbers reported for the mock reconstructions are averages over the 5 realizations of sightline positions and noise. The $\langle d_\perp \rangle = 15\,h^{-1}$Mpc configuration is meant to reproduce the relatively high sightline density areas of the BOSS survey.

protocluster structures leading to the decline in success. Additionally, as the noise in the map increases, the false positive rate increases. When we increase the sightline separation to larger than $10\,h^{-1}$Mpc, the quality of the map degrades significantly, which is reflected in the BOSS-like success rates and lower number of candidates. For small separations, the protocluster identification success rate is close to ideal – 93 per cent in the ideal case and 89 per cent for $\langle d_\perp \rangle = 2$ and $2.5\,h^{-1}$Mpc. Even with a coarse sightline separation of $4\,h^{-1}$Mpc, the success rate is 78 per cent, and this only drops to 60 per cent with the $6\,h^{-1}$Mpc separation that we thought might be catastrophic.

In the BOSS-like separation surveys, the candidate purity is much lower. This is expected since the average spacing in this case is larger than all but the largest protoclusters. However, with random positions, it is possible for several sightlines to overlap with a protocluster and this configuration still performs significantly better than random. We believe the purity in the BOSS-like configuration could also be improved if we considered sightline positions, and only saved candidates with many overlapping sightlines.

In Figure 4.8, we show the candidate completeness and failure rates for the various survey configurations. On the left, we plot the fraction of candidates that did not form clusters. The mock map values are averages over the 5 realizations and we show the std. dev. error bars. We see a steady increase in the candidate contamination as the sightline separation increases. On the right, we plot the candidate completeness measured in four cluster mass bins. For reference, the numbers of protoclusters from the full sample in these bins are

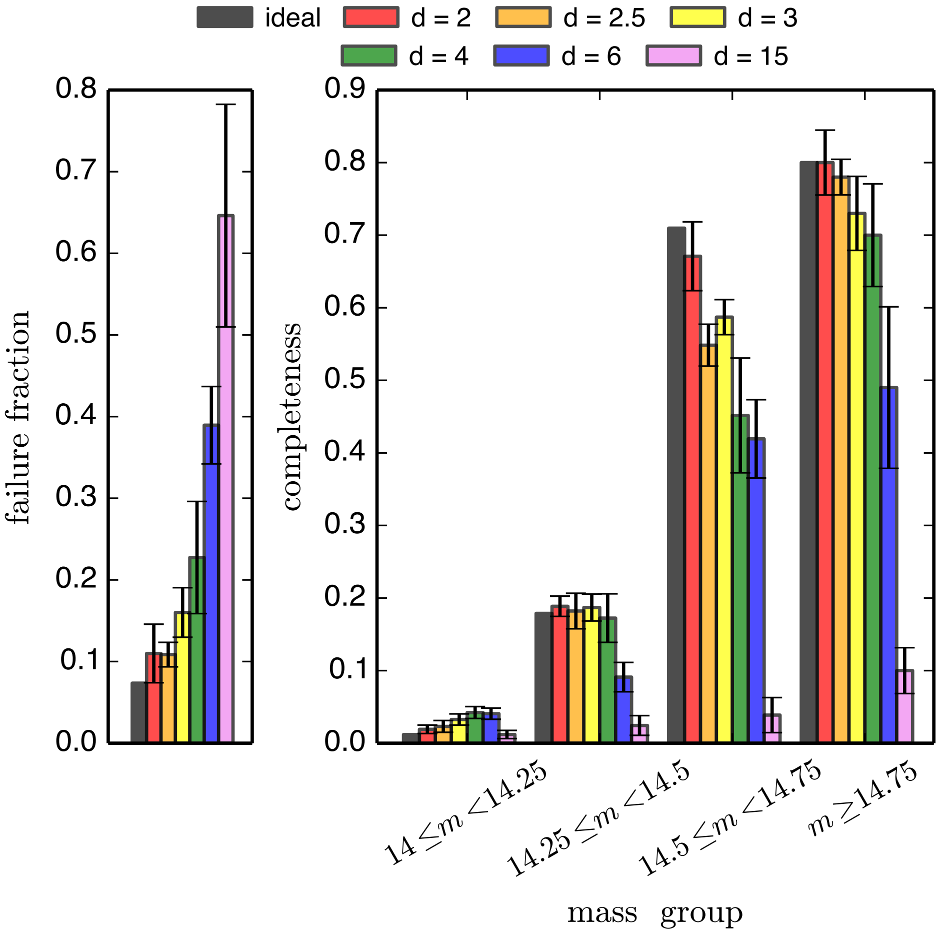

Figure 4.8: Candidate failure fractions and completeness identified in the ideal and mock maps. Left: Fraction of candidates that are failures ($M(z = 0) < 10^{14}\,h^{-1}\mathrm{M}_{\odot}$). Right: Candidate completeness measured in 4 cluster mass bins, where the mass range in $m = \log_{10}[M/(h^{-1}\mathrm{M}_{\odot})]$ is indicated on the x-axis. The mock map counts are averages over 5 realizations, with std. dev. error bars. As the sightline separation increases, we see a steady increase in contamination, and the success fraction decreases. In the low-mass cluster bins, the mock survey completeness is sometimes higher than the noiseless map. This is a result of noise in the mock reconstructions pushing some less significant protoclusters over the chosen threshold value.

251, 123, 31, and 20. In the two high mass bins, the completeness of the $\langle d_\perp \rangle = 2$, 2.5, and $3\,h^{-1}$Mpc surveys is similar to the ideal result. The completeness decreases for larger separations, although it is still about 50 per cent for the $\langle d_\perp \rangle = 4$ and $6\,h^{-1}$Mpc, but only 5 – 10 per cent for the BOSS-like survey. For the two low mass bin, the completeness overall is much lower, as discussed in the previous section. The completeness falls off for very large separations, as before, but for small separations, the completeness is sometimes larger than the ideal map. This is due to the noise in the reconstructions scattering some low mass protoclusters over the threshold value. That is, for some protoclusters that did not make the cut in the noiseless map, the reconstruction noise fortunately pushes them over the edge. Overall, this result makes us confident that we can find a large fraction of the protoclusters that form $> 3 \times 10^{14}\,h^{-1}\mathrm{M}_\odot$ clusters with a CLAMATO-like survey.

In order to understand the cases where our protocluster identification method failed (either missing protoclusters or selecting false positives), we looked at many slices of individual candidates. We performed a union of all candidates identified in the ideal map and in the mock reconstructions, based on the candidate's $z = 0$ halo ID, and tracked which candidates were identified in which maps. After visually inspecting many candidates, we found that we could group the failures into four categories which we called dropout, bad merge, false positive, and borderline protocluster. We illustrate these cases with example candidates in Figure 4.9. Each row is a separate candidate, and the columns show the same slice from the ideal map and the $\langle d_\perp \rangle = 2$, 3, 4, and $6\,h^{-1}$Mpc mock maps. If the candidate was identified in the map, we marked the center with a black cross. We also annotated the $\delta_F/\sigma_{\delta_F}$ value from each map (at the candidate center) under the image. In the top row, we show a successful case, where the candidate forms a massive cluster, and the protocluster is found in all of the maps. This case was not very common when we included the large $6\,h^{-1}$Mpc separation maps, but it was usually the case for the most massive protoclusters that created a significant ($> 5\sigma$) flux decrement.

The first failure case, dropout, is the most common scenario for a missed protocluster identification. The protocluster creates a clear flux decrement in the ideal map and small separation survey maps, but the signal drops out in the large separation survey maps. An example is shown in the second row of Figure 4.9. In the example shown, the protocluster is successfully identified in the ideal and small $\langle d_\perp \rangle$ maps, but as the sightline separation increases, the region is less well-sampled and the flux values in the region never drop below the threshold. We also found plenty of cases where the candidate is missed in the $\langle d_\perp \rangle = 3$ or $4\,h^{-1}$Mpc maps, but found again in the larger separation maps, just due to how the sightlines and the protocluster line up in a given random survey realization.

The second failure case, bad merge, is another scenario that results in missing a protocluster, and is due to a weakness in our method for merging points during the identification procedure. We found a few cases where two protoclusters were linked by a dense filament, so that the two regions that should have been separate candidates were mistakenly merged. The grouped points were usually similar shapes in the different maps, but the flux minimum could end up in either protocluster depending on the reconstruction noise. If these candidates were correctly partitioned during the merging step, there would be another successful

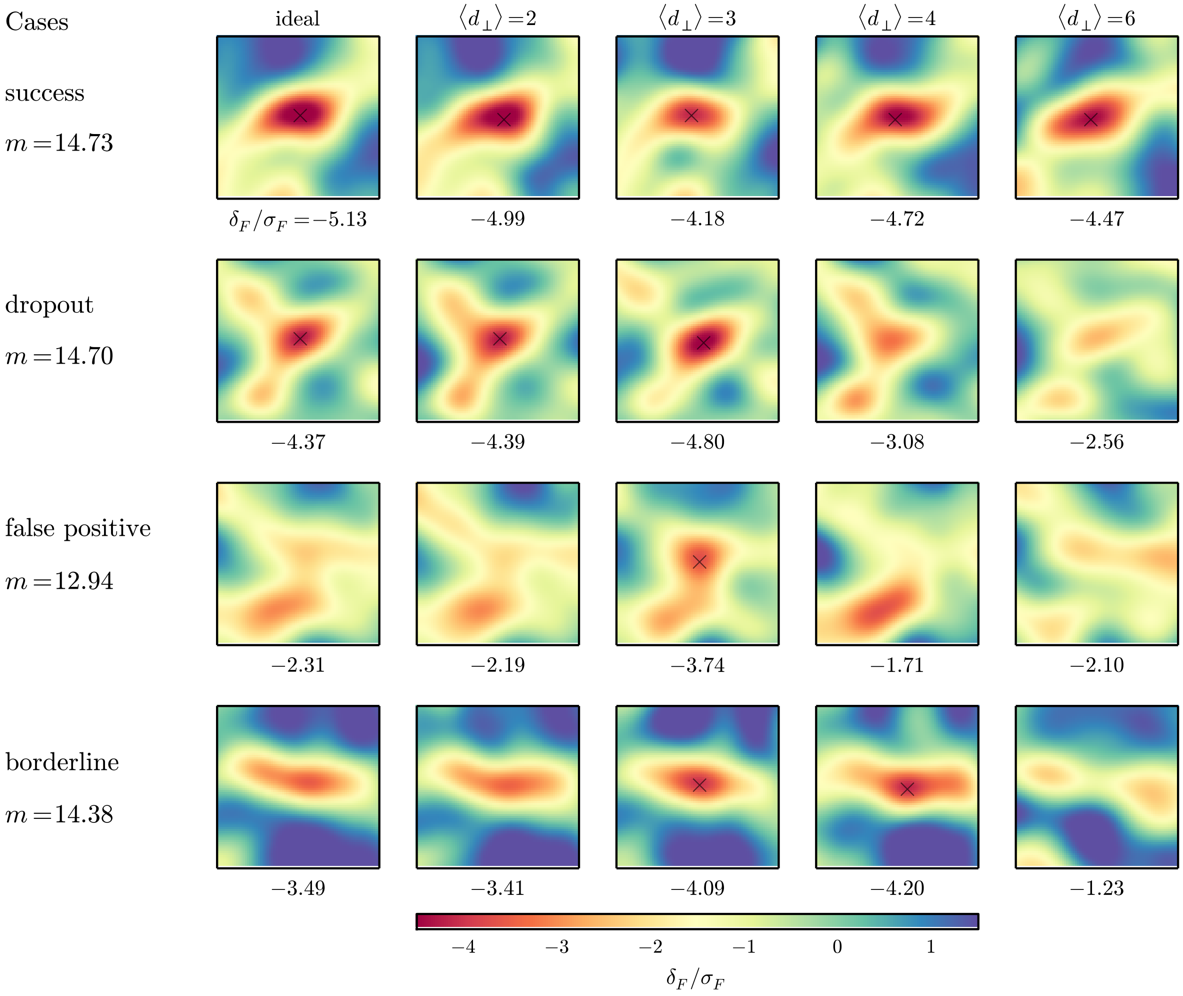

Figure 4.9: Slices from the ideal and mock maps centered on four candidates scenarios. The images show the $\delta_F/\sigma_{\delta_F}$ values in slices that are $40\,h^{-1}$Mpc across and $2\,h^{-1}$Mpc thick. From left to right, we show the ideal field and the mock reconstructions with increasing $\langle d_\perp \rangle$, and we mark successful identifications with a black cross. We also annotate the $\delta_F/\sigma_{\delta_F}$ value in the center of the map under each image. For each row, we annotate the case name and candidate mass $m = \log_{10}[M/(h^{-1}\mathrm{M}_\odot)]$ on the left. Top row: A success case where the protocluster is identified in all maps. Second row: As the average sightline separation increases, the sightlines do not sample the low flux region well enough and the candidate 'drops out'. Third row: A moderately low flux region with additional noise can create a false positive. Bottom row: A borderline protocluster where the reconstruction noise scatters the candidate over the threshold value. There is another failure case (bad merge) not shown here, but explained in the text.

identification in each map.

In regions of moderately low flux, it is possible for the reconstruction noise to scatter low, and create false positives. This scenario is origin of the increasing contamination (failure fraction) with increasing sightline separation. The third row of Figure 4.9 shows an example of this, where the candidate is not a protocluster, but is mistakenly identified in one of the maps. Of course, sometimes this reconstruction noise also scatters low mass protoclusters in the right direction. This is the origin of the final scenario we called borderline protoclusters. In this case, the protocluster creates a flux decrement just under the threshold, so that it is not identified in the noiseless map, but noise can scatter it over the threshold, so that it is identified in the mock reconstruction. An example is shown in the bottom row of Figure 4.9, where the flux decrement in the noiseless map is just under the threshold. In the $\langle d_\perp \rangle = 3$ and $4\,h^{-1}$Mpc reconstructions, the noise scatters the flux decrement to over 4σ, which makes it a successful candidate.

4.6 Conclusions

In this chapter, we characterized the signature of protoclusters at $z \simeq 2$–3, and demonstrated the success of a simple method for finding these protoclusters from the associated Lyα forest flux decrement. The tomographic reconstruction of the 3D Lyα forest transmitted flux field from individual sightlines is the crucial step to this method. In order to handle datasets with large numbers of pixels, we implemented a new fast Wiener Filter code, which we are making publicly available. This code will make it possible to run reconstructions on the scale of $N_{\rm pix} \gtrsim 10^6$, larger than the expected size of the ongoing CLAMATO survey.

We identified protoclusters at $z = 2.5$ using a large cosmological N-body simulation with sufficient resolution to capture individual absorption systems comprising the Lyα forest and covering enough volume to contain a respectable cluster sample. We constructed FoF halo catalogs for each simulation snapshot and defined clusters at $z = 0$ with a mass cut of $M \geq 10^{14}\,h^{-1}{\rm M}_\odot$. We then identified protoclusters by tracking cluster member particles from $z = 0$ back to $z = 2.5$ (by particle ID) and characterized the protocluster regions. The key signature of protoclusters is that they are outliers in density and flux on large scales. We found that protocluster centers are above the 95th percentile of the density and flux decrement and that the half-mass radius of typical protoclusters at this redshift is $4\,h^{-1}$Mpc. The density and flux profiles of protocluster regions are well fit by a Gaussian with a scale of $5\,h^{-1}$Mpc, suggesting that maps with several Mpc resolution should easily resolve these structures. We also found that the flux decrement and radius of a protocluster increases with its $z = 0$ mass, so that it is easiest to find the protoclusters that form the more massive clusters.

We reviewed our tomographic reconstruction method (a Wiener Filter) and some specifics of our application. Specifically, we assume a certain form of the signal covariance and that the noise covariance is diagonal. These assumptions significantly reduce the space complexity of our algorithm, so that we can easily fit the calculations on a single node, avoiding significant communication costs and taking advantage of shared-memory parallelism. Additionally, this design will easily take advantage of upcoming compute architectures, where the number of cores per node is expected to increase in the near future, and could easily be extended to run on GPUs.

We designed a procedure to identify protocluster candidates in the flux maps. To choose candidates we smooth the map, apply a threshold, and group the remaining points into candidates. We ran the procedure on noiseless maps and compared to the simulation halo catalog, finding that we can achieve 90 per cent candidate purity with this simple method. We also confirmed that the method tends to find protoclusters that form the most massive clusters ($> 3 \times 10^{14}\, h^{-1}\mathrm{M}_{\odot}$). The most massive halos in the identified protoclusters have masses of about $10^{13}\, h^{-1}\mathrm{M}_{\odot}$ – still very difficult to find at these redshifts using alternative methods.

Finally, we created realistic mock surveys (similar to the recent observations of Lee et al. (2014b)) and reconstructed the flux maps with our code. We found that surveys with an average sightline spacing $\langle d_{\perp} \rangle = 2.5\, h^{-1}\mathrm{Mpc}$ perform essentially the same as the ideal, noiseless map. Such surveys should identify protoclusters with a 90 per cent success rate, and find $70 - 80$ per cent of the protoclusters that form clusters with masses $> 3 \times 10^{14}\, h^{-1}\mathrm{M}_{\odot}$). Using the same conservative threshold, we would identify 5 protoclusters in the planned CLAMATO volume. However, the volume should contain about 30 protoclusters including those that form lower mass clusters.

Finding protoclusters at $z \simeq 2 - 3$ remains an observationally challenging problem. With relatively simple methods, we have demonstrated a promising new technique for finding protoclusters at these redshifts. As shown in Lee et al. (2014a) IGM tomography offers a novel method for mapping large volumes with high efficiency using existing facilities. The method can return large samples of protoclusters and does not suffer from projection effects (or redshift errors). The Lyα forest also has the advantage of only probing mildly nonlinear densities, allowing for *ab initio* calculation of the density-observable relation (i.e. the bias) via numerical simulations. Future work can easily extend this to reconstruct density maps, include redshift-space distortions, and incorporate more advanced models of protoclusters.

Acknowledgments

This work was published as Stark et al. (2014). We thank Andreu Font-Ribera, Zarija Lukić, and Peter Nugent for useful discussions. The simulation, mock surveys, and reconstructions discussed in this work were performed on the Edison Cray XC30 system at the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231.

Chapter 5

Finding cosmic voids using Lyα forest tomographic flux maps

The majority of the cosmic web, by volume, is made up of large, almost empty regions known as voids which are surrounded by walls, filaments and clusters (see e.g. van de Weygaert & Platen 2011, for a review). Within this paradigm, voids are regions which are practically devoid of galaxies. They are slightly prolate in shape and occur on a wide range of sizes from Mpc to tens of Mpc (Vogeley et al. 1994; Ceccarelli et al. 2006; Lavaux & Wandelt 2012).

The study of cosmic voids has received renewed theoretical attention recently. Voids are intrinsically interesting as a major constituent of the Universe (by volume) and one of the most visually striking features in galaxy maps. They form an interesting environment for the study of galaxy evolution. They may present an excellent laboratory for studying material which clusters most weakly (e.g. dark energy or massive neutrinos), and for testing modified gravity models. Future surveys are expected to find large samples of voids at a range of redshifts, enhancing the potential of void science.

The pristine environments of voids present an interesting setting for the study of early galaxy formation. Galaxies in low-redshift voids generally have smaller stellar masses, appear bluer, have a later morphological type, and have higher specific star formation rates than galaxies in average density environments (van de Weygaert & Platen 2011; Beygu et al. 2015), although the latter properties might be solely due to their lower stellar mass (Hoyle et al. 2005; Kreckel et al. 2011). Extending similar studies to higher redshifts to see whether similar trends hold is a pressing observational challenge.

Ryden (1995) was the first to discuss using voids as probes of cosmology. Park & Lee (2007) anticipated using void ellipticity as a cosmological probe and Lee & Park (2009); Bos et al. (2012) discussed constraining dark energy using voids. Lavaux & Wandelt (2012) investigated the potential for using stacked voids as a probe of geometrical distortions (the AP test; Alcock & Paczynski 1979). Chan et al. (2014) have studied the clustering of voids and Hamaus et al. (2014d) describe constraining cosmology with void-galaxy cross-correlations. Hellwing et al. (2010); Li (2011) have investigated studying the nature of dark matter using the properties of voids and Li et al. (2012); Clampitt et al. (2013) have suggested that void properties may provide a strong test of some modified gravity theories.

Observationally, studies of voids date back over three decades (Gregory & Thompson 1978; Longair & Einasto 1978; Kirshner et al. 1981). Recent redshift surveys have identified large samples of voids (e.g. 2dF: Ceccarelli et al. 2006; SDSS: Sutter et al. 2012, 2014a; VIPERS: Micheletti et al. 2014) and a measurement of the AP effect from voids in the local Universe has recently been reported by Sutter et al. (2014b); Hamaus et al. (2014a). Being underdense in both galaxies and dark matter, voids act like objects with an effectively negative mass, bending light rays away from them. This effect has been recently detected at high significance by Clampitt & Jain (2014).

In the absence of large-scale dynamical and environmental influences, voids would become increasingly isotropic as they evolve (Icke 1984). However, in modern theories of structure formation the frequent encounters with surrounding structures and the influence of large-scale tidal fields serve to reverse the simple trend expected for isolated voids (van de Weygaert & Platen 2011). As matter in the center of voids streams outwards faster than matter towards the boundary, the interior evolves into an almost uniform low density region surrounded by 'ridges' marking the void edge: often referred to as a 'bucket-shaped' density profile (see Ceccarelli et al. 2006; Hamaus et al. 2014c, for recent fits). The density in the center has a characteristic value of $\delta \approx -0.8$.

Historically, surveys of voids over large volumes have come from large, galaxy redshift surveys. However, finding voids in this manner requires a significant investment in telescope time due to the necessity of a high spatial sampling of tracer galaxies. For example, the void catalog presented in Sutter et al. (2012) found voids in the distribution of SDSS DR7 galaxies. Their 'bright' cut found voids with radii larger than $7\,h^{-1}$Mpc with galaxies separated by $8\,h^{-1}$Mpc. To find comparable galaxy separations at $z = 0.5$, 1.0, and 2.0 will require obtaining complete galaxy redshift samples for apparent limiting magnitudes of $I = 22.5$, 24.2, and 25.7, respectively (assuming galaxy luminosity functions from Dahlen et al. 2005 at $z \leq 1$, and Reddy et al. 2008 at $z = 2$). So while such galaxy number densities are just achievable up to $z \approx 1$ with existing telescopes, it becomes increasingly challenging at higher redshifts.

In light of the aforementioned challenges, it is understandable that little attention has been given to studying voids at $z > 1$ (although see D'Aloisio & Furlanetto 2007; Viel et al. 2008). However, recently it has been noted that given sufficient sightlines, the Lyα forest observed in a dense grid of faint background galaxies and quasars can be used to create three-dimensional maps of large-scale structure and that the observational requirements to map out cosmological volumes ($V \approx 10^6\,h^{-3}\text{Mpc}^3$) are within reach of existing facilities (Lee et al. 2014a). Indeed, the first pilot map on a small field has already been made using just a few hours of data from Keck telescope (Lee et al. 2014b). This method is ideally suited to finding extended structures at high redshift. In Stark et al. (2014), we investigated the possibility of finding protoclusters in tomographic Lyα maps. Here, we study the signature of cosmological voids in the Lyα forest.

5.1 Simulations and void finding

5.1.1 N-body simulations

In order to study the signal of voids in the Lyα forest, we make use of cosmological N-body simulations. We require a simulation which simultaneously covers a large cosmological volume while having a sufficiently small inter-particle spacing to model transmission in the IGM. The requirements are sufficiently demanding that we have used a pure N-body simulation, augmented with the fluctuating Gunn-Peterson approximation (FGPA; Petitjean et al. 1995; Croft et al. 1998; Meiksin & White 2001; Meiksin 2009). This same simulation was also used in Lee et al. (2014a) and Stark et al. (2014) so we only review the salient features here.

The simulation employed 2560^3 equal mass ($8.6 \times 10^7\,h^{-1}\mathrm{M}_\odot$) particles in a $256\,h^{-1}$Mpc periodic, cubical box. This provides sufficient mass resolution to model the large-scale features in the IGM at $z = 2$–3 using the FGPA (Meiksin & White 2001; Rorai et al. 2013) and sufficient volume to find large voids. The assumed cosmology was of the flat ΛCDM family, with $\Omega_\mathrm{m} \approx 0.31$, $\Omega_\mathrm{b}h^2 \approx 0.022$, $h = 0.6777$, $n_s = 0.9611$, and $\sigma_8 = 0.83$, in agreement with Planck Collaboration et al. (2013a). The initial conditions were generated using second-order Lagrangian perturbation theory at $z_\mathrm{ic} = 150$, when the rms particle displacement was 40 per cent of the mean inter-particle spacing. The particle positions and velocities were evolved using the TreePM code of White (2002). Throughout the text, we will use the particle positions and velocities from the output at $z = 2.5$. Using the particle positions and velocities at $z = 2.5$, we generated mock Lyα forest spectra on a 2560^3 grid with the FGPA as described in Stark et al. (2014). In all, we generated 2560^3 grids with the matter density and Lyα forest flux in real- and redshift-space. For many purposes in this work, we did not need the high resolution provided by the 2560^3 grids and found it much easier to work with smaller grids. For this reason, we also downsampled the fields to 256^3 by simply averaging neighboring grid points. In the remainder of the chapter, when we refer to flux, we mean the Lyα forest transmitted flux fraction perturbation

$$\delta_F = F/\langle F \rangle - 1\ . \tag{5.1}$$

5.1.2 Void finding

There are a variety of methods and tools used to find voids in large-scale structure (Kauffmann & Fairall 1991; Platen et al. 2007; Neyrinck 2008; Sutter et al. 2015). We use a simple spherical underdensity method on the low resolution, gridded densities to construct our $z = 2.5$ void catalog. This technique is similar to spherical overdensity (SO) halo finding, but instead applied to underdensities. We identify voids by taking the 256^3 density grid and selecting points under some threshold value, then growing spheres around the points until the average density enclosed reaches a target value. We handle overlapping voids by only saving the void with the largest radius, and we also discard any remaining voids with a

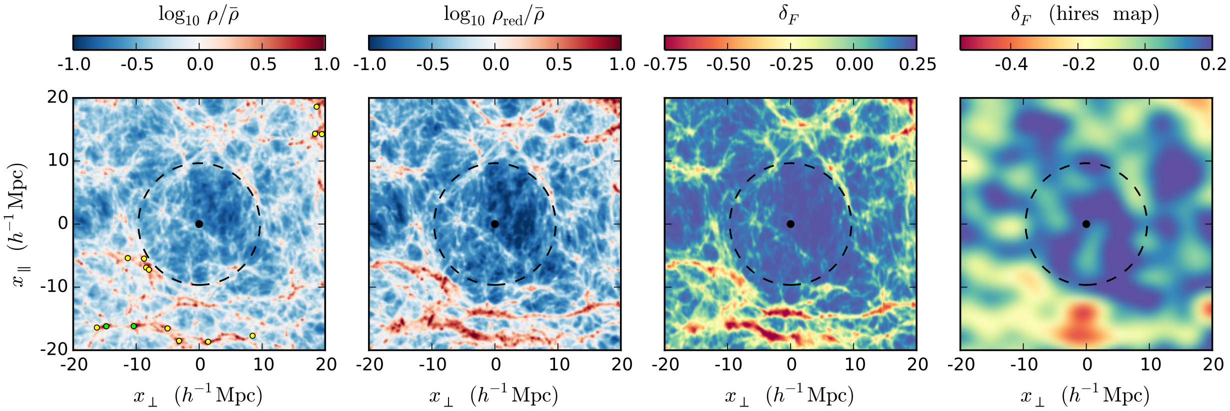

Figure 5.1: A slice through our simulation and a mock reconstructed flux map centered on a large void, with a radius of $9.7\,h^{-1}$Mpc. The slice is $40\,h^{-1}$Mpc across and $6\,h^{-1}$Mpc into the page. The vertical axis, $x_\parallel$, is the redshift direction, along the line of sight, and the horizontal axis, $x_\perp$, is one of the directions transverse to the line of sight (the other transverse direction is into the page). (Left) The matter density, in real space. The void center and radius are indicated with a black dot and dashed circle. Positions of halos with $M \geq 10^{12}\,h^{-1}\mathrm{M}_\odot$ are plotted as green dots and positions of halos with $M \geq 3\times 10^{11}\,h^{-1}\mathrm{M}_\odot$ are plotted as yellow dots. These halos would host galaxies at the spectroscopic limits for existing instrumentation. (Middle left) The matter density, in redshift space. (Middle right) The 'true' flux field over the same volume, in redshift space. (Right) A reconstructed flux field from a mock survey (also in redshift space) with average sightline spacing, $\langle d_\perp\rangle = 2.5\,h^{-1}$Mpc (see text).

radius $r < 2\,h^{-1}$Mpc.

The free parameters in this void-finding method are the threshold value $\rho_{\rm thresh}$ and enclosed average target value $\bar{\rho}_{\rm enc}$. For the density field, there is a well-motivated threshold value of $\rho_{\rm thresh} = 0.2\,\bar{\rho}$. This threshold value is a canonical density for a void core, corresponding to the central density at shell-crossing for spherical void models (van de Weygaert & Platen 2011). This critical density value is a standard choice in other void-finding codes (e.g. Neyrinck 2008). The choice of the average target value, however, is somewhat arbitrary. We first tested an average target value of $\bar{\rho}_{\rm enc} = 0.2\,\bar{\rho}$, but found that it produced voids that were far too small – the spheres never reached the apparent 'edge' surrounding the low-density core. We experimented with several other average target values and found that a value of $\bar{\rho}_{\rm enc} = 0.4\,\bar{\rho}$ resulted in good agreement between the sphere sizes and the apparent void edges.

In principle, there is nothing special about the specific threshold and average density values we chose, and these parameters should depend on the redshift. That is, as voids continue to evacuate, the core and average densities of the void will decrease. In practice, we found the final void catalog is not very sensitive to these settings, although the void radii clearly scale with the average target density setting. Since most large voids have central densities $< 0.2\,\bar{\rho}$ already, the exact value of the threshold mostly makes a difference in terms of how many points we must search over, and less of a difference in the void centers. We did find that a very small threshold (say $< 0.1\,\bar{\rho}$ for this redshift) forces voids to grow from positions that often look off-center by eye. This is due to the hierarchical nature of voids, in that the lowest-density point in a large void is typically the center of a smaller subvoid, sometimes referred to as the void-in-void scenario (c.f. fig. 6 of Neyrinck 2008 and Sheth & van de Weygaert 2004). Using the SO parameter values of the threshold $\rho_{\rm thresh} = 0.2\,\bar{\rho}$ and average target $\bar{\rho}_{\rm enc} = 0.4\,\bar{\rho}$, we found 16,167 voids which cover 15 per cent of the simulation volume.

Subsection 5.1.2 shows a slice through our simulation, centered on one of the largest voids in our catalog with $r = 9.7\,h^{-1}$Mpc. The slice is $40\,h^{-1}$Mpc across and $6\,h^{-1}$Mpc projected into the page. The four panels show the void in real-space density, redshift-space density, Lyα forest flux, and a tomographic flux map constructed from a mock survey. In each image, we marked the void center and radius with a black dot and dashed line, respectively. In the first panel, we also overplotted the positions of halos found in the same slice. Green dots mark the positions of halos with mass $M \geq 10^{12}\,h^{-1}{\rm M}_\odot$ (roughly an $L_\star$ halo at this redshift), while yellow dots mark the positions of halos with $3 \times 10^{11}\,h^{-1}{\rm M}_\odot \leq M < 10^{12}\,h^{-1}{\rm M}_\odot$. Based on simple abundance matching (see Figure 5.1.2), these halos should host galaxies with apparent magnitudes $\mathcal{R} < 24.7$ and $24.7 \leq \mathcal{R} < 25.6$, just bright enough for redshift determination with existing facilities. The second panel shows that in redshift space the void has a larger density contrast and extent in the line-of-sight direction due to the outflow of matter from the void. Such a large structure is easily visible in the redshift-space density and flux. Although the tomographic flux map is a blurred version of the true flux, the void structure is so large that it can still easily be picked out by eye. For reference, the tomographic map is one of the realizations from Stark et al. (2014) with an average sightline spacing of $\langle d_\perp \rangle = 2.5\,h^{-1}$Mpc, similar to the ongoing survey of Lee et al. (2014b). At the same time, the void is not captured by the galaxy positions even if we assume a complete galaxy sample. The relative sparsity of such halos highlights the difficulty in finding voids, even large ones, at high redshift using galaxies as tracers.

We compared our void catalog with that produced by a watershed void finder, similar to the ZOBOV code. The watershed method finds the set of connected elements all under some threshold. In ZOBOV, the elements are the Voronoi cells in the tessellation of the dark matter particle positions (where the density is estimated from the volume of the Voronoi cell), but in this case, we use the density values on the 2560^3 grid for simplicity. The watershed algorithm on a uniform grid is straightforward. We find the set of points under the given threshold, and keep a list of the under-threshold points that have not been assigned to a specific watershed. Starting with the minimum value point, we search grid neighbors to see if they are also under the threshold and add them to the current watershed if so. The search stops when there are no remaining neighbors under the threshold. These points are then removed from the list of unassigned points and we move on to the next watershed. After we assign all points under the threshold, we discard watersheds with an effective radius $r_{\rm eff} = (3V_{\rm shed}/4\pi)^{1/3}$ less than $2\,h^{-1}$Mpc, as we did with the spherical underdensity voids. Using this method with the same threshold of $\rho < 0.2\,\bar{\rho}$, we found 6,364 voids, covering 5 per cent of the simulation volume.

The sets of large voids in the spherical underdensity (or SO) catalog and the watershed catalog agree very well. We visually inspected the 100 largest voids in the SO catalog, and found matches in the watershed catalog. In most cases the watershed void effective radius was slightly smaller (by 1–2 h^{-1}Mpc), which explains the total count and volume difference, and the watershed voids typically have complex morphologies. The watershed voids often have an ellipsoidal core, with fingers stretching out into smaller low-density regions. We compared the SO void centers to the watershed void value-weighted centroids $\mathbf{x}_{\rm shed} = \sum_i \mathbf{x}_i \rho_i^{-1} / \sum_i \rho_i^{-1}$, where the sums are over all the points in the shed, and we weight by the inverse of the density so that the centering is driven by lower-density points. Unfortunately, the non-spherical geometries of the watershed voids tend to drive the centroid away from the center found with the SO method and the centers in the two catalogs tend to disagree by several Mpc.

In Figure 5.2, we show a slice centered on a large void. The left and right panels show the same void structure from two angles (the xz-plane and yz-plane). The top panels show the density field in this region, while the bottom panels show the void shape in the different catalogs. The black dot and circle are the center and radius of the void found with the SO method. The blue dot and circle show the value-weighted centroid and effective radius of the void found with the watershed method, and the blue triangle shows the ‘core’ point (the minimum value point within the shed). We also show the points in the void watershed in the bottom panels with the blue colorscale, where the color scales with the number of points in the projection. In order to damp out some of the complex structure of the original watershed void, we also tried running the watershed finder on a $2\,h^{-1}$Mpc smoothed density field (with an adjusted threshold of $0.45\,\bar{\rho}$). The watershed void found in the smoothed field is shown

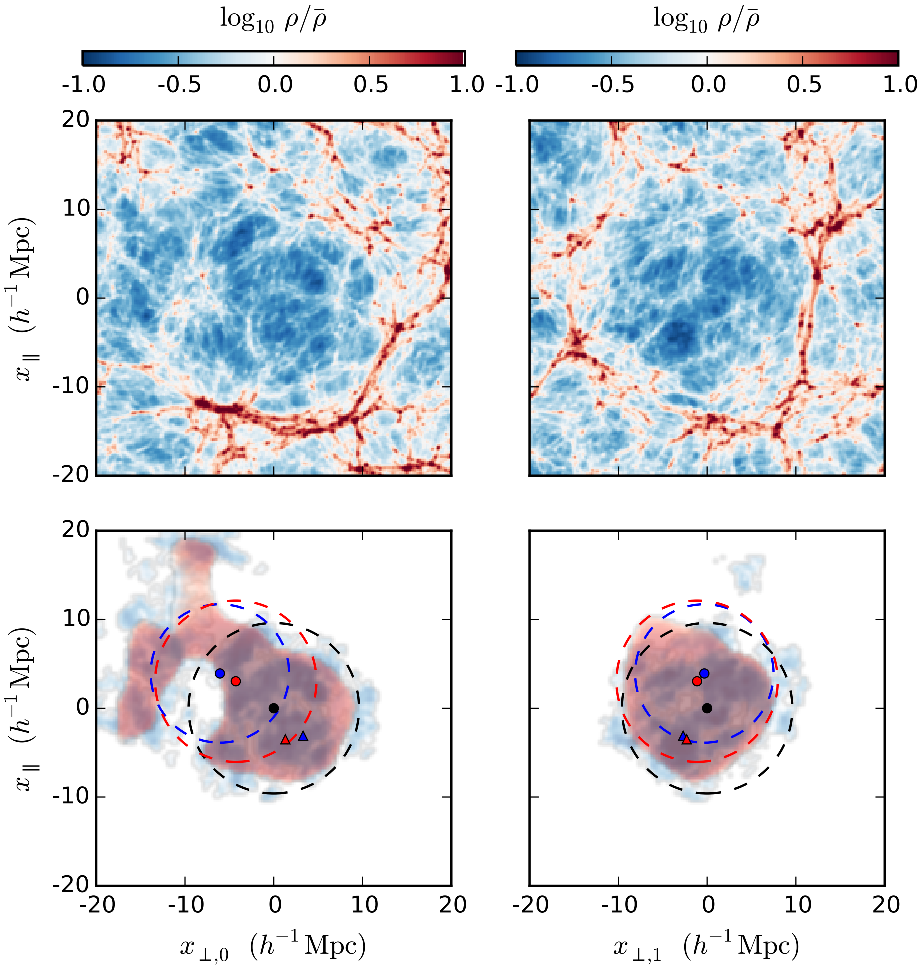

Figure 5.2: A density slice centered on a large void with a radius of $9.6\,h^{-1}$Mpc. Just as in Subsection 5.1.2, the slice is $40\,h^{-1}$Mpc across and $6\,h^{-1}$Mpc into the page. The panels on the left show the view along the simulation y axis, while the panels on the right show the view along the simultion x axis. (Top) The matter density. (Bottom) The corresponding voids from the SO catalog (black), the watershed catalog (blue), and the smoothed density watershed catalog (red). The dots and dashed line circles show the center of the voids and the SO radius or watershed effective radius. The triangles show the minimum value (core) points of the watersheds. The blue and red colorscales show the projection of the points in the watershed (the darker the color, the more points into the page). See the text for more details.

in red.

The SO and watershed voids have reasonably similar shapes as seen in the yz-plane (right panel). The extent of the watershed points (blue region) overlaps almost entirely with the SO circle (black), besides the small wayward blue blob at $(y, z) = (5, 15)\,h^{-1}$Mpc. However, seen in the xz-plane, the voids have very different shapes indeed. The slightly overdense region at $(x, z) = (-10, 0)\,h^{-1}$Mpc limits the growth of the SO void, but the watershed region reaches around this structure to the underdense region on the other side. This extension from the main underdense region is also seen in the smoothed version of the watershed void. The SO radius is $9.6\,h^{-1}$Mpc and the watershed effective radius is $7.8\,h^{-1}$Mpc, and $9.1\,h^{-1}$Mpc in the smoothed version. Although the radii are all fairly similar, it's amazing to see just how different the extents and centers differ. The SO center is $7.2\,h^{-1}$Mpc away from the watershed centroid and $5.3\,h^{-1}$Mpc away from the watershed core. At the same time, the watershed core and centroid are separated by a whopping $11.9\,h^{-1}$Mpc, far more than the effective radius. Overall, the watershed void spans 39, 26, and $32\,h^{-1}$Mpc in the x, y, and z directions respectively, meaning the small finger-like voids extending from the central underdensity are very long.

It is reassuring that these two methods for finding voids in the density field qualitatively agree well, but we decided to use the SO void catalog for the remainder of this work due to its simplicity. Overall, the centers and simple shapes of the SO voids provide cleaner radial profiles and should be easier to find in the tomographic maps later. We were also concerned that the non-trivial noise we expect in the maps from tomographic reconstruction might artificially combine or split watershed regions, whereas the spherical average in the SO method will be less affected by such noise. Since our tomographic maps come with a noise estimate, one could imagine a more sophisticated algorithm (e.g. a matched filter or likelihood-based method) for finding voids could be implemented. We leave such investigations to future work.

5.2 Voids at $z = 2.5$

In Figure 5.4, we show the cumulative number density of voids as a function of void radius. We plot the distribution of our $z = 2.5$ voids in black, and show the distribution of low-redshift voids with a green dashed line ($z \approx 0.5$), computed from eq. 21 of Lavaux & Wandelt (2012). As expected, there are many more small voids and voids are generally smaller at $z = 2.5$ than at $z \simeq 0$ (c.f. fig. 1 of Ceccarelli et al. 2006 or fig. 7 of Lavaux & Wandelt 2012). While voids with radii of $7\,h^{-1}$Mpc are common for low-redshift studies, we have only 126 voids with $r \geq 7\,h^{-1}$Mpc, which cover two per cent of the simulation volume. We note, however, that it is difficult to compare void sizes across works using different void-finding methods and working at different redshifts. For instance, we could increase the number of $r \geq 7\,h^{-1}$Mpc voids by simply increasing the average target value in our SO void finder. For the most part, this does not change which large voids are identified, but does shift centers and increase the cumulative number density at a particular value. See Colberg

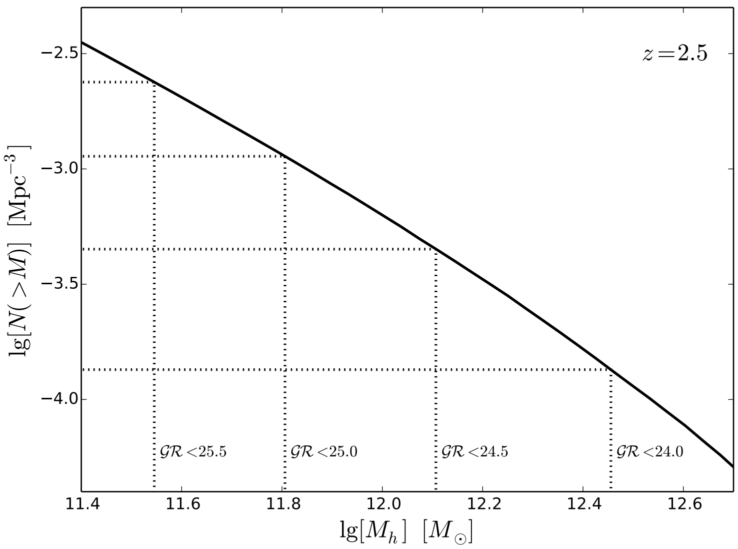

Figure 5.3: The cumulative, halo mass function in the simulation at $z = 2.5$, as a function of (FoF) halo mass. The dotted lines indicate the abundances of galaxies brighter than the listed apparent $\mathcal{R}$-magnitude limits, derived from the luminosity function of Reddy et al. (2008). Since at this redshift and these masses satellites make up a small fraction of galaxies by number, this plot allows us to approximately equate our mass-limited halo catalogs into flux-limited galaxy catalogs. Note we have used volumes and masses without factors of h in this figure to match the scalings adopted in Reddy et al. (2008).

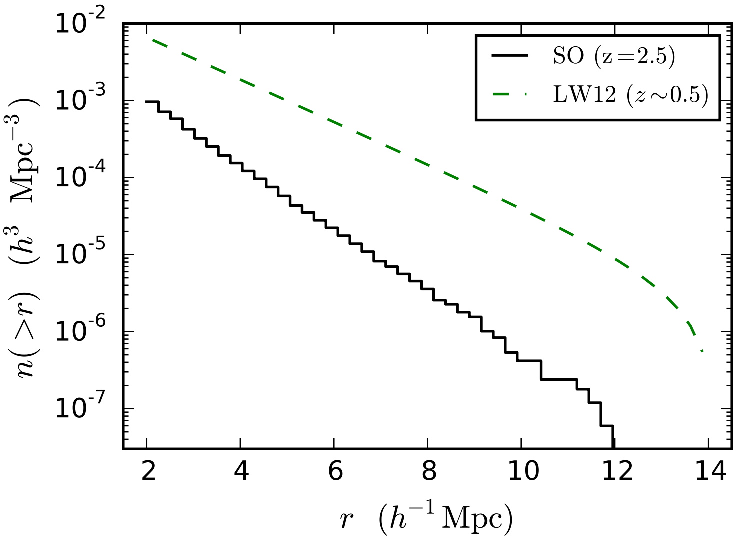

Figure 5.4: The cumulative, comoving number density of voids with radii greater than r vs. r. In black, we show the distribution of our SO density voids at $z = 2.5$. The green dashed line is the distribution of the voids found in Lavaux & Wandelt (2012), which are generally larger, as expected.

et al. (2008) for more detail about the difficulties of defining voids and differing results from various void finders.

Figure 5.5 shows the radial profile of voids in density, radial peculiar velocity, and real-space flux, stacked by radius. The gray lines show the profile for individual voids, while the thick black lines show the mean. We chose to stack the voids in our catalog with radii in the range $5 \le r < 6\,h^{-1}$Mpc, of which there are 511. We note that the central density is about $0.2\,\bar{\rho}$ in the dark matter distribution, as a result of our choice of void finder. The average profile then rises almost continuously to the void edge, though individual voids show substructure within them (also visible in Subsection 5.1.2). The slope of our profile contrasts with the profile of more evolved voids at lower redshifts, which exhibit a 'bucket' profile [1]. We used the $z = 0$ simulation output to create a low-redshift SO void catalog (with adjusted threshold and average target values) and found that these voids do exhibit such a 'bucket' profile. The voids we are studying at $z = 2.5$, however, have not yet evolved to such a state and are still in the process of evacuating. The real-space flux profiles illustrate just how well the flux profile mirrors the density profile. The center of the voids have $F \simeq 1$, which translates to $\delta_F \simeq 0.25$ for our setting of $\langle F \rangle = 0.8$. In both the density and flux value, the stack profile almost reaches the mean value by $r = 10\,h^{-1}$Mpc.

The middle panel of Figure 5.5 shows the radial velocity profiles of the voids. The profiles are linearly increasing up to the void radius, where they peak around $120\,\mathrm{km\,s^{-1}}$

---

[1] For example Ceccarelli et al. (2006) propose $\rho(r)/\bar{\rho} = A_0 + A_3\,(r/R_V)^3$ while Hamaus et al. (2014b) proposes a 4-parameter model with a similar shape.

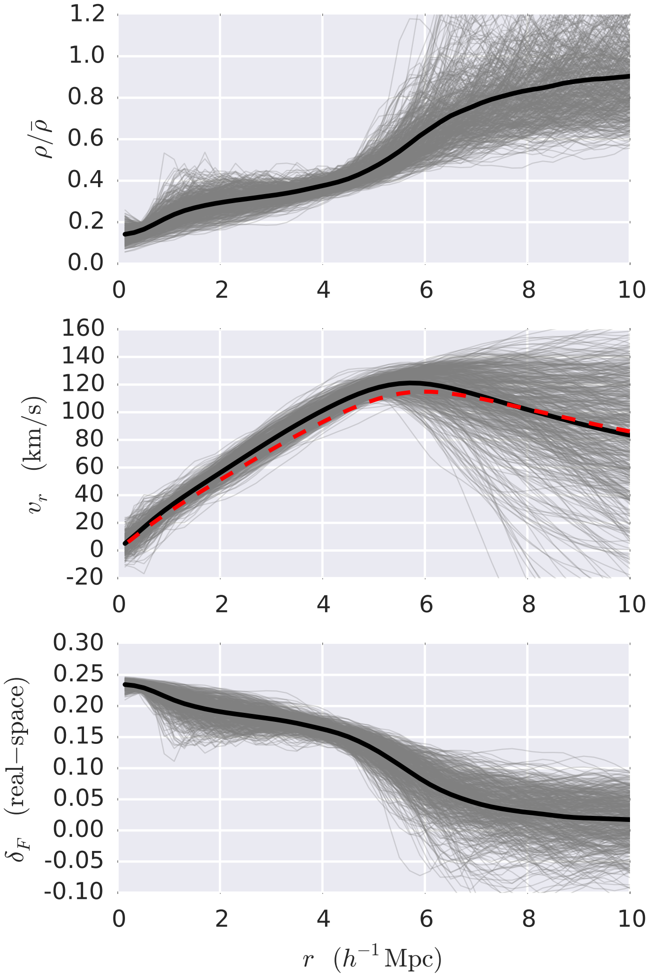

Figure 5.5: Radial profiles centered on voids with $5 \leq r < 6\,h^{-1}$Mpc. Gray lines are individual profiles, and the thick black line shows the mean. (Top) The matter density profile. (Middle) The radial peculiar velocity profile. The dashed red line shows the linear theory prediction using the density stack profile above. (Bottom) The real-space flux profile.

before turning over. The radial velocity profiles also have a fairly small scatter – at the average radius of $5.4\,h^{-1}$Mpc, the mean velocity is $119.7\,\mathrm{km\,s^{-1}}$ with a standard deviation of $5.4\,\mathrm{km\,s^{-1}}$ or about 5 per cent. Within the context of linear theory $\mathbf{v}(\mathbf{r}) \propto \nabla\nabla^{-2}\delta(\mathbf{r})$. If we assume a spherical mass distribution, this can be solved to yield $v_r \propto \delta(< r)/r^2$, where $\delta(< r)$ is the 'overdensity enclosed within r' in analogy with the Newtonian gravitational acceleration due to a spherical mass distribution. The radial velocity profile around spherical, or averaged, voids then becomes (Peebles 1993; Hamaus et al. 2014b) [2]

$$v(r) = -afH\frac{1}{r^2}\int_0^r \left(\frac{\rho(x)}{\bar{\rho}} - 1\right) x^2\,dx \tag{5.2}$$

where $f \simeq \Omega_\mathrm{m}^{0.55}(z)$ is the growth factor, which at $z = 2.5$ is close to 1. This form was shown in fig. 1 of Hamaus et al. (2014b) to fit the velocity profile of stacked voids in N-body simulations well at $z \simeq 0$. The dashed red line in Figure 5.5 shows this linear theory approximation, which we see compares favorably to the profile measured in our $z = 2.5$ voids (within 10 per cent over the range plotted). It is somewhat surprising that the linear theory prediction matches our simulated radial velocity profile result down to Mpc scales and for $|\delta| = 0.8$. The fact that this prediction also matched void radial velocity profiles at $z = 0$, with voids from a different finder method is impressive (Hamaus et al. 2014b).

Figure 5.6 shows the two dimensional profiles (in mass and flux) of stacks of voids with radii $5 \le r < 6\,h^{-1}$Mpc in both real and redshift space. Apart from some noise near the line-of-sight axes, the contours in Figure 5.6 are isotropic in the real-space panels but show extended, anisotropic profiles in redshift-space. This is an indication of the effect of peculiar velocities, which appear visually to be larger in our case than at lower redshifts when voids are traced by galaxies. We note that the profiles are much better measured at small radii where there is less scatter. Beyond the radius of the stack $r \approx 5.5\,h^{-1}$Mpc, the scatter in the individual profiles increases significantly. We believe this is the source of the extended orange contour in the bottom right panel, for instance.

Since we expect the stacked voids to be isotropic in real space, by symmetry, any observed anisotropy offers an opportunity to study such peculiar velocities. This could be particularly interesting for constraining models with modified gravity. For example, Clampitt et al. (2013) find that, driven by the outward-pointing fifth force, individual voids in chameleon models expand faster and grow larger than in a ΛCDM universe. Such effects would modify the profile of the stacked voids in a potentially observable manner, allowing observations of voids in the Lyα forest to test such models. Based on the radial velocity profiles shown in Figure 5.5 and the measured standard deviation, one would need only about 20 voids with a radial velocity measurement to reach one per cent standard error (assuming Poisson errors). With accurate enough radial velocity measurements from void anisotropies, it should be possible to detect deviations at the 10 per cent level with relatively small samples. Conversely, the

---

[2] Note, we have an additional factor of a in this expression compared to eq. 2.2 of Hamaus et al. (2014b), perhaps due to a difference in proper vs. comoving quantities. Again, we always use comoving scales and densities, and peculiar velocities.

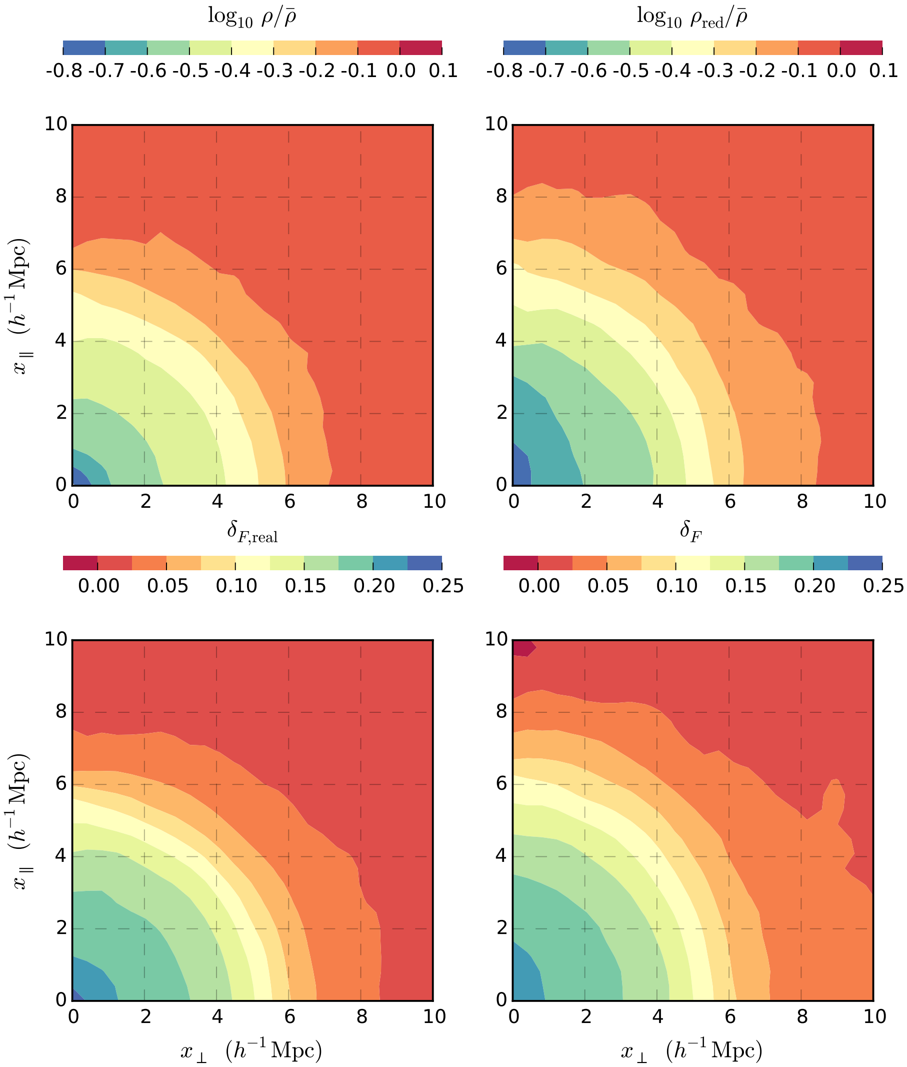

Figure 5.6: Contour plots of a stack of the voids with $5 \leq r < 6\,h^{-1}$Mpc, showing the impact of redshift-space distortions. The four panels show the density and flux in real and redshift space, binned in the distances parallel ($x_\parallel$) and perpendicular ($x_\perp$) to the line of sight. In the redshift-space fields, the radial profiles are clearly extended in the line-of-sight direction.

Table 5.1: Void catalogs

Field	SO Thresh.	SO Avg.	Count	Vol. Frac.
ρ	$0.20\,\bar{\rho}$	$0.4\,\bar{\rho}$	16,167	0.152
$\rho_{\rm red}$	$0.15\,\bar{\rho}$	$0.3\,\bar{\rho}$	16,338	0.151
δ_F	0.224	0.167	16,296	0.150
hires map	0.224	0.167	16,586	0.203
midres map	0.224	0.167	8,724	0.181
lores map	0.224	0.167	5,565	0.153

The spherical under/overdensity void catalogs used for comparison, found in our $V \approx 1.7 \times 10^7\,h^{-3}{\rm Mpc}^3$ simulation. We use the original ρ catalog as our 'truth' and varied the SO parameters for the $\rho_{\rm red}$ and δ_F catalogs to qualitatively match. The hires map uses a $d = 2.5\,h^{-1}$Mpc sightline spacing mock survey, while the midres map comes from a $d = 4.0\,h^{-1}$Mpc configuration, and the lores map from a $d = 6.0\,h^{-1}$Mpc configuration. The noise and smoothing inherent in the tomographic reconstruction process create larger differences in the catalog properties.

larger impact of redshift-space distortions in the Lyα flux field means they must be modeled in order to make a measurement of the Alcock-Paczynski effect (Alcock & Paczynski 1979) from stacked voids in the flux field (see Subsection 5.4.2).

5.3 Finding voids in flux

Underdense regions show up as high transmission regions in the 3D Lyα forest flux for $z = 2 - 3$, as shown in Subsection 5.1.2 and Figure 5.5. This is not necessarily the case at lower redshifts, since the characteristic density probed by the forest increases with time (Becker et al. 2011; Lukić et al. 2015). At lower redshifts, it is difficult to see differences in transmission passing through an underdense region vs. a moderately overdense region, since it takes a significant overdensity to create an observable absorption feature. Fortunately, there is a large overlap between the redshift range of the forest accessible from the ground and the redshift range where mean density structures scatter an observable fraction of the light. Given this, finding voids in flux at $z = 2$–3 is a matter of finding coherent high-transmission regions. In Stark et al. (2014), we outlined a simple method using tomographic flux maps to find protoclusters (coherent low-transmission regions) at these redshifts. We now adapt these methods to find coherent high-transmission regions, corresponding to high-redshift voids.

The void catalogs used in this section are listed in Table 5.1, including voids found in the redshift-space density, the flux, and three tomographic flux maps. In each case, we modified the SO threshold and average target parameters to create a void catalog with roughly the same void count and radius distribution as the real-space density catalog. For reference, we use three of the tomographic flux maps created in Stark et al. (2014). These maps were constructed by mocking up a realistic survey covering the simulation volume with signal-to-noise distributions similar to the pilot observations of Lee et al. (2014b), and several other settings of the average sightline spacing and minimum spectral signal-to-noise ratio. The tomographic maps were then generated by running our reconstruction code on the mock spectra. The three maps we use are generated from mock surveys with average sightline spacings of $\langle d_\perp \rangle = 2.5$, 4, and $6\,h^{-1}$Mpc. The smallest sightline separation configuration is similar to the ongoing CLAMATO survey, and the larger separation configurations are similar to what we expect from large-area surveys on 8–10 m telescopes like the Subaru Prime Focus Spectrograph (PFS; Takada et al. 2014) or the Maunakea Spectroscopic Explorer (Simons et al. 2014). We refer to these tomographic maps as the hires, midres, and lores flux maps. We discuss the characteristics of the individual catalogs in the following subsections and focus on our method to compare catalogs for now. Comparing across void catalogs, we want to confirm that there are nearby pairs of voids with similar radii. We expect to find the same set of voids in density and flux, as we have demonstrated how well-matched the fields are in previous sections. This is mostly a matter of determining the best SO parameters for the flux. The tomographic flux maps, however, are contaminated by and spectral noise in individual mock spectra and shot noise due to sparse sampling of the field, and this noise will certainly affect our capability of finding voids.

We use two metrics to compare the catalogs of voids found in different fields. The first metric is essentially the sum of the difference in the center positions and radii, which we call the match error. If we are comparing voids in catalog A with voids in catalog B, for each A-B pair, we compute the match error

$$\epsilon = \frac{\sqrt{(r_A - r_B)^2 + |\mathbf{x}_A - \mathbf{x}_B|^2/3^2}}{r_A}$$

where r is the radius and $\mathbf{x}$ is the center position. We chose this form of the error for several reasons. First, this form of the error also allows for trading off differences in radii and centers. We want to consider the differences relative to the size of the void, which will allow for larger center and radii differences for larger voids. Note that this form of the error assumes the radius of void A is the reference. Finally, we compare 1/3 of the center difference to the radius difference just due to the dimensionality (and empirically we found that the mean center difference is about 3 times the mean radius difference). Later in this section, we will show that a match error $\epsilon < 0.3$ qualifies as a good match for a void between two catalogs, and we will use this cut to count which voids are 'matched'. The second metric we use is the total volume overlap between the voids in both catalogs. Clearly, this metric is less useful for telling if a catalog A void is well-matched by a single catalog B void. However, it is a useful measure of how well-matched the catalogs are overall and does not depend on a specific form of the error nor a specific value to cut at. It is also useful in cases where noise in the tomographic maps artificially combines or splits voids – although the centers and radii might not match across catalogs, there will still be a sizable volume overlap. We can use these metrics to get a sense of void completeness and purity of each of the flux catalogs with respect to the density void catalog. The number of 'matched' density voids compared to the total number of density voids (the match fraction) is a measure of completeness. We

Table 5.2: Catalog comparison for $r \geq 6\,h^{-1}$Mpc voids

	Density	Flux	Hires map	Midres map	Lores map	Random
Density	–	0.994 / 0.937	0.660 / 0.631	0.478 / 0.568	0.269 / 0.517	0.0194 / 0.152
Flux	0.988 / 0.933	–	0.683 / 0.637	0.471 / 0.569	0.292 / 0.514	0.0188 / 0.153
Hires map	0.581 / 0.567	0.576 / 0.567	–	0.356 / 0.484	0.238 / 0.430	0.0261 / 0.153
Midres map	0.284 / 0.409	0.282 / 0.408	0.258 / 0.425	–	0.171 / 0.344	0.0243 / 0.152
Lores map	0.186 / 0.349	0.184 / 0.347	0.191 / 0.377	0.192 / 0.359	–	0.0204 / 0.153
Random	0.0193 / 0.151	0.0191 / 0.150	0.0368 / 0.204	0.0480 / 0.182	0.0398 / 0.149	–

The catalog void match fraction and volume overlap fractions (separated by a slash in each cell), for $r \geq 6\,h^{-1}$Mpc voids. For each row, we compute the fraction of voids with a match error $\epsilon < 0.3$ and the fraction of the total volume overlapped by voids in the catalog of that column.

also measure completeness by comparing the overlapping volume between two catalogs to the total volume in density voids (the overlap fraction). The purity of the flux catalogs can be measured by matching in the other direction (the fraction of matched flux voids) and by comparing the overlap to the total volume in flux voids.

Using these metrics, we first found that redshift-space distortions can create large differences in the centers and, to a lesser extent, the radii of the voids. In order to more easily compare voids found in density and flux, we created a void catalog using the redshift-space density. Voids found in redshift-space density matched those found in real-space density best (in terms of detecting the same voids with similar radii) when we used a threshold of $0.15\,\bar{\rho}$ and an average target of $0.3\,\bar{\rho}$. These densities are lower than the real-space values since outflows from voids drive densities lower.

Before applying these metrics to the void catalogs derived from the various flux maps, we also compared the redshift-space density void catalog to random void catalogs, mainly to get a sense of the worst-case performance. We created ten catalogs of 16,338 voids (the same number as the redshift-space density catalog), with centers uniformly distributed in the simulation domain, and with radii randomly drawn from the same distribution as that in the redshift-space density catalog. We compared each random void catalog against the density catalog, computing the fraction of density voids with a match error $\epsilon < 0.3$ and the fraction of the total volume overlap to the total volume in density voids. Overall, 2.7 per cent of the density voids were matched by voids in the random catalogs on average. It is reassuring to see that a small fraction of the density voids are matched by random voids which tells us that our cut of $\epsilon < 0.3$ is stringent enough. We also noticed that for the largest voids ($r \geq 8\,h^{-1}$Mpc), the average match fraction drops to 1.3 per cent. This is due to the fact that both the density and random catalogs contain just a few very large voids and it is even less likely that they will overlap enough to meet the match error cut. The average volume overlap fraction between density voids and voids in the random catalogs was 15 per cent, and did not change with the radius considered. This is not surprising since the voids cover roughly 15 per cent of the total volume, so random points will overlap about that often.

5.3.1 Ideal flux

We ran our SO void finder on the 'true' flux grid (256^3) using a threshold of $\delta_F \geq 0.224$ and an average target of $\delta_F = 0.167$. We experimented with several values of the threshold and average target fluxes and found that these values resulted in a number of voids and radius distribution similar to the catalog of voids found in density. The mapping from flux to density evolves quickly with redshift, so these SO parameters would have to be adjusted for other redshifts and UV background prescriptions.

The flux void catalog matches the redshift-space density catalog very well. For all voids ($r \geq 2\,h^{-1}$Mpc), 84 per cent of the density voids and the flux voids are matched. The volume overlap fraction is also very high, at 86 per cent of the density void volume and also 86 per cent of the flux void volume. For larger voids ($r \geq 6\,h^{-1}$Mpc), the catalogs are even better matched. In this case, 99.4 per cent of the density voids and 98.8 per cent of the flux voids are matched, and the volume overlap fraction is 93.7 per cent of the density voids and 93.3 per cent of the flux voids. For reference, the density catalog contains 335 of these large voids, while the flux catalog contains 325.

5.3.2 Tomographic flux maps

We constructed the map void catalogs by running the SO void finder on the maps with the same SO parameters we used for the ideal flux field. We tried several other SO parameter settings on the maps, but found that the default value catalogs performed best in comparison to the density and flux catalogs. Small changes to the SO parameters resulted in slightly better performance, but changes larger than about $\Delta\delta_F = 0.01$, resulted in similar or worse performance, so we did not bother optimizing these parameter choices further. Unfortunately, our simple void-finding method does not consider noise in the map which can contaminate the set of thresholded points and the spherical averages. The noise in the map acts to scatter points below or above the SO threshold, creating false negatives and positives respectively. In the same way, the noise can affect the spherical averages used in the SO finder, resulting in inaccurate radii. However, this should be less of an issue for coherent structures spanning several map resolution scales, which is apparent in our results for small vs. large voids. The effects of the noise are apparent in the radius distribution of the map catalogs. In the hires map catalog, the number of very small voids ($r < 2.5\,h^{-1}$Mpc) is 3,902, about half of the number found in the density catalog (6,157). This is likely due to shot noise where sightlines did not sample these smaller structures well enough. The number of medium voids ($3\,h^{-1}\text{Mpc} \leq r < 6\,h^{-1}$Mpc) is about double that in the density catalog, and the number of large voids ($r \geq 7\,h^{-1}$Mpc) about the same (147 vs. 121). This explains why there is a similar total number of voids in the density and hires map catalogs, but more total volume in the map catalog (see Table 5.1). The radius distributions of the midres and lores map catalogs are more distorted by the noise. The midres map catalog contains about a half the number of small voids ($r < 5\,h^{-1}$Mpc) compared to the density catalog and the lores map catalog contain about a quarter. There are approximately double the number of large

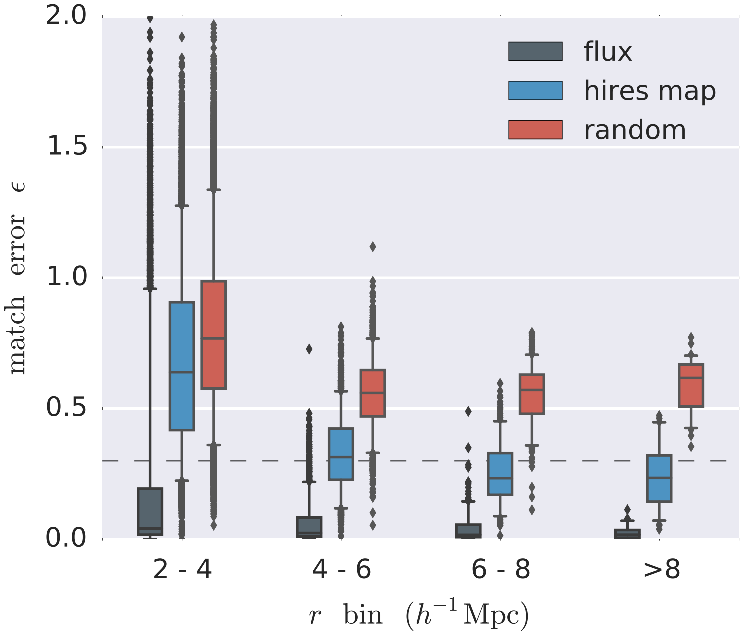

Figure 5.7: A box and whisker plot of the distributions of match errors vs. radius. The match error is computed between the redshift-space density voids and the voids in the flux catalog (gray), the hires map catalog (blue), and in one of the random catalogs (red). The line in the middle of the box is the median, the box extends from the 25th to the 75th percentile, the whiskers extend down to the 5th and up to the 95th percentiles, and points outside of this range are plotted individually. The match error threshold of $\epsilon = 0.3$ is marked with a dashed line. The flux catalog matches the density catalog exceptionally well. The hires map catalog is essentially random for small voids, but performs much better for large voids. The random catalog match error is fairly flat across radius bins.

voids ($r \geq 7\,h^{-1}$Mpc) in both map catalogs compared to the density catalog.

In Figure 5.7, we plot distributions of the match errors between the voids in the redshift-space density catalog and voids in the flux, hires map, and one of the random catalogs vs. radius. The line in the middle of the box shows the median, the box extends from the 25th to the 75th percentiles, the whiskers extend from the 5th to the 95th percentile, and samples outside this are plotted individually. The match error cut value of $\epsilon = 0.3$ is marked with a dashed line. The match errors against the random catalog tend to fall around $\epsilon = 0.6$, and there are few points under our cut of $\epsilon = 0.3$, again showing that this is a safe choice. It is also reassuring that the random errors are relatively flat over radius bins because we defined the match error relative to the original void radius. The gray distributions show just how well-matched the flux and density catalogs are and that the flux match errors overlap very little with the random errors. Overall, the hires map catalog misses a significant fraction of the small voids in the density catalog, but performs well for larger voids. For all voids

($r \geq 2\,h^{-1}$Mpc), the hires map catalog matches only 16.3 per cent of the density voids and the volume overlap fraction is 49.4 per cent. This can also be seen in the smallest radius bin in Figure 5.7, where the hires map is just a bit lower than the random distribution. However, considering larger voids ($r \geq 6\,h^{-1}$Mpc), the hires map catalog performs much better matching against 66 per cent of the density voids and overlapping with 63 per cent of the volume. In Figure 5.7, there is a clear trend that the hires map match errors decrease with radius, separating from the random distribution. The match fractions in the other direction (purity of the hires map voids) are similar at 17.8 per cent for all voids and 58.0 per cent for large voids. The lower match fractions for large voids in this case is driven by the hires map catalog having more large voids.

The midres map void catalog performs worse for all voids, but still matches a considerable fraction of the density voids. Overall, the midres map catalog matches only 6.2 per cent of the density voids, although it still overlaps with 40 per cent of the density void volume. If we consider larger voids ($r \geq 6\,h^{-1}$Mpc), the midres map catalog matches 48 per cent of the density voids, and matches 60 per cent of even larger voids ($r \geq 8\,h^{-1}$Mpc). The lores map catalog performs worse than this, but is still useful for finding large voids. The lores map catalog matches only 2.5 per cent of the density voids – consistent with the random catalogs – although it overlaps with 30 per cent of the volume. This indicates that the void finder is still able to find regions containing voids from the map, but does not recover an accurate center or radius. If we consider some of the largest voids ($r \geq 8\,h^{-1}$Mpc), the match fraction increases to 48.9 per cent and the volume overlap fraction is 61.9 per cent, again confirming that the larger the void, the better the maps perform.

Our results are also summarized in Table 5.2. In this table, we give the match and overlap fractions between the redshift-space density, flux, hires map, midres map, lores map, and random catalogs for the voids with $r > 6\,h^{-1}$Mpc. We note again that the first five catalogs are single catalogs while the random results are averages over the ten random catalog realizations. The trends between the catalogs are the same as described above: the random catalogs match 2 – 3 per cent and overlap about 15 per cent, and the correspondence between the density and flux catalogs is very high. Comparing the density (or flux) catalogs to the map catalogs, the match fraction drops to 60 – 70 per cent for the hires case, to 40 – 50 per cent for the midres case, and down to 20 – 30 per cent for the lores case. However, the volume overlap fraction remains relatively high for all of the maps indicating that the poor matching is more the fault of our simple void finding method than the maps truly missing the voids.

Overall, the maps perform decently matching voids with radii larger than the map resolution, but it is surprising that the maps still do not perform better for the largest voids. We visually inspected many of the large voids to see why the flux map void catalogs sometimes miss these large voids. We show two example slices of large voids in the flux, hires map, and lores map in Figure 5.8. The top row shows a successful void match in both maps, while bottom row shows a failure case. In the flux panels, we overplot the original void with a black dot and circle. In the map panels, we overplot the best match void and annotate the match error. In the top panels, the matching hires map void is a bit smaller and offset just a bit to

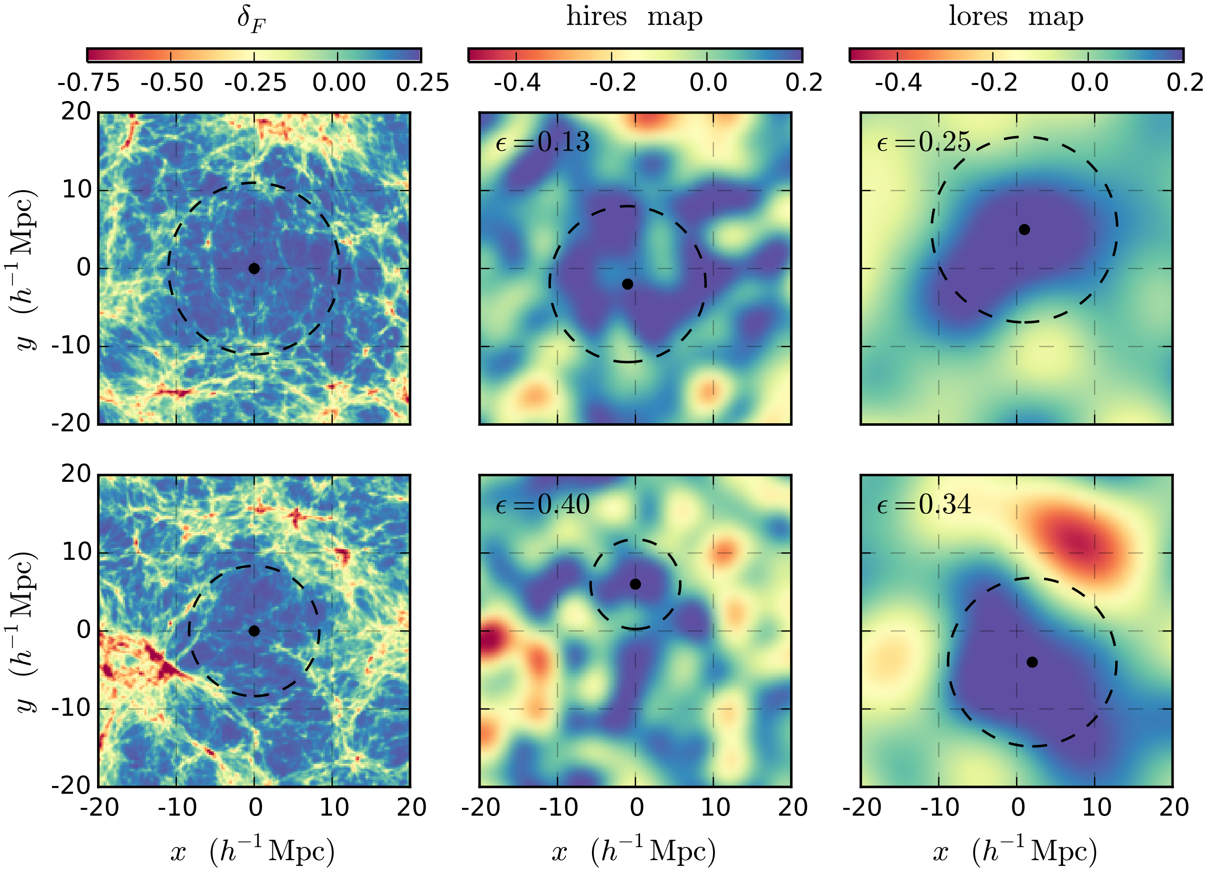

Figure 5.8: Slices of two large voids showing the flux, hires map, and lores map (from left to right). A void that is well-matched in the map void catalogs is shown on top, and a poor match case is shown on the bottom. The top void has a radius of $10.9\,h^{-1}$Mpc and the bottom void has a radius of $8.3\,h^{-1}$Mpc. Just as in Fig. 5.1.2, the slice is $40\,h^{-1}$Mpc across and $6\,h^{-1}$Mpc into the page, although in this case, the redshift direction is into the page. The black dot and circle in each panel show the void as found in each catalog. The flux panels show the original void (very close to the one found in density), while the map panels show the map voids with the lowest match error. We also annotate the match error of the map voids in the top-left corner of the images.

the bottom left. In the lores map, a noise feature around $(x, y) = (5, -5)\,h^{-1}$Mpc pushes the matching void center up farther, but with a similar radius, resulting in a sufficiently small match error. In the bottom hires map panel, there is significant noise around the center of the original void which pushes the center of the void up and restricts the growth of the void radius to a much smaller size. In the bottom lores map panel, the filamentary structure around $(x, y) = (5, -5)\,h^{-1}$Mpc is missing, which allows the void radius to grow much larger and results in a poor match. Interestingly, in the bottom row, the overall structure of the lores map matches the structure of the ideal flux better than the hires map by eye. However, we have not considered estimate of the noise in the map. Using the tomographic reconstruction method outlined in Stark et al. (2014), it is possible to compute the covariance of the map or to run Monte Carlo error estimates. Considering the amount of noise apparent in Figure 5.8, much could be gained by incorporating a noise estimate into a void finding procedure. We believe future work can make significant gains in void finding performance by considering the structure of voids beyond simple spheres and taking the map noise into account.

5.4 Discussion

5.4.1 Survey prospects

For the cosmology of our simulation, the comoving radial distance to $z = 2.5$ is $4050\,h^{-1}$Mpc, thus one degree subtends $70\,h^{-1}$Mpc. Assuming a $250\,h^{-1}$Mpc depth (e.g. $2.2 < z < 2.5$), each square degree of survey area translates into a volume of $1.2 \times 10^6\,h^{-3}\text{Mpc}^{-3}$. Given the number densities in Figure 5.4, we see that surveys like CLAMATO with $V \simeq 10^6\,h^{-3}\text{Mpc}^3$ would encompass about 150 voids larger than $5\,h^{-1}$Mpc in radius. If we assume a conservative void finding efficiency of 60 per cent, our simple method would recover ~ 90 voids. This would be the first detection of a significant population of high-redshift voids. Of course, this is a lower bound on the efficiency of identifying voids with a map of this resolution due to our conservative choice of what constitutes a match, and that there is still room for improvement in the method. Using the PFS, it is possible to double the target density, covering a larger redshift range at the cost of sightline density. In Lee et al. (2014a), we discussed piggybacking on the planned galaxy evolution survey described in Takada et al. (2014). Such a survey would provide a map of roughly $16\,\text{deg}^2$ or $8 \times 10^4\,h^{-2}\text{Mpc}^2$ area and $700\,h^{-1}$Mpc depth ($2.3 < z < 3.2$) for a total volume of $6 \times 10^7\,h^{-3}\text{Mpc}^3$, although at a coarser resolution of about $5\,h^{-1}$Mpc. This much larger volume would encompass ~ 3000 voids with $r \geq 5\,h^{-1}$Mpc, and would detect voids with an efficiency better than 30 per cent, providing a sample of around ~ 1000 voids. With an extended program on PFS of 100 nights, it is possible to construct a tomographic map covering $\sim 200\,\text{deg}^2$ with the same redshift coverage and resolution, providing a tenfold increase in volume, and therefore, the number of voids ($\sim 10^4$).

For comparison, similar volumes have been explored to find voids in low-redshift galaxy

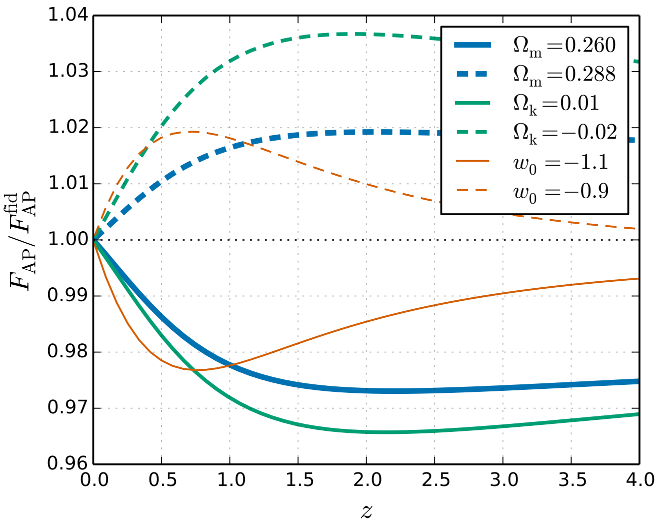

Figure 5.9: $F_{\rm AP}$ as a function of redshift for different cosmological models, divided by the prediction from our fiducial model. We show 5 per cent variations in $\Omega_{\rm m}$, 1 or 2 per cent variations in $\Omega_{\rm k}$ per cent, and 10 per cent variations in w_0. For all models, the value of the Hubble parameter h has been adjusted to keep the angular scale of the CMB oscillations $\theta_{\rm ls} = d_{\rm ls}/r_s$ fixed. State-of-the-art measurements of $F_{\rm AP}$ have 5 per cent uncertainties, measured at $z = 0.5$ up to $z = 2.4$ (Blake et al. 2011; Aubourg et al. 2014; Beutler et al. 2014; Samushia et al. 2014).

positions, although for somewhat larger voids. Pan et al. (2012) searched for $r > 10\,h^{-1}$Mpc voids in the Sloan Digital Sky Survey Data Release 7 main galaxy sample (out to $z = 0.1$), corresponding to a volume $V \approx 10^7\,h^{-1}{\rm Mpc}^3$, finding $\sim$ 1000 voids. Sutter et al. (2012) also found a similar number of voids in the SDSS DR7 main galaxy sample (out to $z = 0.2$) and the luminous red galaxy sample (out to $z = 0.44$). In total, the galaxy samples were split into 6 samples covering volumes from 10^6 to almost $10^9\,h^{-1}$Mpc (see their table 2). However, the larger volume samples were covered by brighter, more massive galaxies, with larger separations. By $z = 0.1$, the average galaxy separation in DR7 is already larger than $5\,h^{-1}$Mpc, making it difficult to find statistically significant small galaxy voids. Sutter et al. (2014a) provided an update to this analysis using the Baryon Oscillation Spectroscopic Survey Data Release 9 CMASS sample, split into 6 samples, each covering about $5 \times 10^8\,h^{-3}{\rm Mpc}^3$. This work found $\sim$ 1500 voids with large radii ($> 20\,h^{-1}$Mpc).

5.4.2 High-redshift void cosmology

Much of the recent discussion of voids as cosmological probes has focused on them as a means to measuring the Alcock-Paczynski (AP) parameter,

$$F_{\rm AP} = \frac{1+z}{c} D_{\rm A} H \ , \tag{5.3}$$

where D_{A} is the angular diameter distance and the Hubble parameter, H, encodes distortions in the line-of-sight (LOS) direction. Note that this measurement measures $H(z)$ directly, rather than an integral as measured by e.g. Type Ia supernovae. In Figure 5.9, we show how specific variations in cosmological parameters affect the AP parameter, giving a qualitative idea of how accurate these measurements must be. Specifically, we show changes in the AP parameter with 5 per cent variations in Ω_{m}, 1 or 2 per cent variations in Ω_{k} per cent, and 10 per cent variations in w_0.

Lavaux & Wandelt (2012) have proposed measuring the AP parameter using the anisotropy of stacked voids, in the context of voids identified in galaxy surveys. Even though voids should be spherically symmetric when averaged, each void will have a certain random asymmetry that will add noise to the global stack. Lavaux & Wandelt (2012) showed that the uncertainty due to this intrinsic scatter when averaging N voids can be approximated by (their eq. 35):

$$\frac{\sigma_{AP}}{F_{\mathrm{AP}}} \approx \frac{1}{\sqrt{N}} \ . \tag{5.4}$$

In terms of cosmological parameters, Figure 5.9 shows that it requires substantial changes in cosmological parameter values, by today's standards, to produce one or two per cent changes in F_{AP}. Therefore, in order to be competitive with other cosmological probes, the stack should be done using several thousand voids.

In order to accurately estimate the uncertainty when stacking N voids identified in the flux field, we would have to study the intrinsic scatter in the asymmetry of the void flux profiles, as a function of redshift and void size. We would also have to take into account the effect of potential systematics like errors in centering and measuring radii, as well as different sources of contamination in the Lyα flux. But assuming that we would also need ten thousand voids to have a one per cent measurement, we can use the discussion above to estimate that we could achieve this uncertainty with a CLAMATO-like survey over 100 square degrees, or with a PFS-like survey over 200 square degrees.

Moreover, F_{AP} can also be robustly constrained from anisotropic measurements of the Baryon Acoustic Oscillation (BAO) scale (Eisenstein & Hu 1998; Seo & Eisenstein 2003). BAO measurements typically report ratios of separations with respect to a fiducial model along the line of sight ($\alpha_\parallel \pm \sigma_\parallel$) and transverse ($\alpha_\perp \pm \sigma_\perp$) directions, as well as their correlation coefficient (r). One can translate these values into a ratio of F_{AP} with respect to F_{AP} in the fiducial model:

$$f_{\mathrm{AP}} = \frac{F_{\mathrm{AP}}}{F_{\mathrm{AP}}^{\mathrm{fid}}} = \frac{\alpha_\perp}{\alpha_\parallel} \ , \tag{5.5}$$

with an uncertainty given by

$$\frac{\sigma_f^2}{f_{\mathrm{AP}}^2} = \frac{\sigma_\parallel^2}{\alpha_\parallel^2} + \frac{\sigma_\perp^2}{\alpha_\perp^2} - 2\frac{r\sigma_\parallel\sigma_\perp}{\alpha_\parallel\alpha_\perp} \ . \tag{5.6}$$

For instance, recent BAO measurements from the BOSS collaboration (Aubourg et al. 2014; Samushia et al. 2014) can be translated into ~ 5 per cent measurements of f_{AP} both at

$z = 0.57$ (from the galaxy survey) and at $z = 2.4$ (from the Lyα survey), raising the bar for measurements from voids.

5.5 Conclusions

In this chapter, we characterized the signal of cosmological voids in the high-redshift matter density field and demonstrated how we can use Lyα forest tomographic maps to find high-redshift voids. We used a simple spherical over/underdensity approach to identifying voids in a large cosmological simulation (with a box size of $256\,h^{-1}$Mpc or a volume of $1.7 \times 10^7\,h^{-3}$Mpc^3) at $z = 2.5$, resulting in a catalog of $\sim 16,000$ voids with radii of 2–$12\,h^{-1}$Mpc. We also tested finding voids with a watershed approach and found that the resulting catalog was similar to that produced by the spherical overdensity method, but with more complex geometries that changed the void centroid non-trivially. For simplicity, we used the spherical overdensity void finding method throughout. This makes our results somewhat conservative, i.e. it is likely that more sophisticated void-finding methods will have improved performance.

Overall, the signature of high-redshift voids in flux is similar to what has been found for low-redshift voids in density. The radial density profile of voids is low ($\rho/\bar{\rho} = 0.2$–0.4) and rises more steeply closer to the radius of the void. One difference we noticed is that the high-redshift voids are typically less evacuated than their low-redshift analogues, giving them a steeper inner profile and less pronounced rise at the edge. The shape of the density profile is clearly mirrored in flux with high transmission inside the radius ($\delta_F = 0.25$–0.15), and dropping down to the mean flux beyond the radius. Interestingly, the radial velocity profiles show very little scatter and the mean radial velocity profile matches up to the linear theory prediction very well. This could be a promising testbed for any (modified gravity) theory predicting differences in void outflow velocities.

Using our void finding method, we identified voids in an ideal flux field and in three tomographic flux maps generated from mock surveys with spatial samplings of $\langle d_\perp \rangle = 2.5$, 4, and $6\,h^{-1}$ (hires, midres, and lores maps). We compared the flux void catalogs to the density void catalogs by considering how well 'matched' pairs of voids are in terms of their centers and radii. We found excellent agreement between the density and ideal flux void catalogs, where 99 per cent of the large voids ($r > 6\,h^{-1}$Mpc) are well-matched. The noise in the tomographic maps clearly impacts the efficiency of finding voids, reducing the fraction of well-matched large voids down to 66, 48, and 27 per cent in the hires, midres, and lores maps, respectively. However, when we inspected individual cases of poorly matched voids, we found that many of these are due to noise in the maps artificially breaking up or merging high-transmission regions. It is clear that a more sophisticated void finder, especially one that models a noise component, would perform much better on the tomographic maps. Implementing such a method is beyond the scope of the current work.

Using these matching results, we can provide a conservative forecast for the number of voids that can be found in dense Lyα surveys. Our hires map has a signal-to-noise ratio distribution and sightline spacing similar to the ongoing CLAMATO survey. With a sky coverage of one deg^2, the CLAMATO data would produce a tomographic map covering $V \approx 10^6\,h^{-1}\mathrm{Mpc}^3$, and our proposed void-finding method would identify about 100 voids with $r > 5\,h^{-1}\mathrm{Mpc}$ in such a volume. With a 16 deg^2 survey on the PFS, we would identify about 1000 voids with $r > 5\,h^{-1}\mathrm{Mpc}$, although at a degraded purity. A 100-night dedicated Lyα forest survey across 200 deg^2 on the PFS would increase this number by a further order of magnitude to $> 10^4$ voids.

These populations of high-redshift voids could be useful for many purposes, including tests of modified gravity, as an AP test and for studying high-redshift void galaxies. Previous works have considered voids as a clean environment for studying galaxy evolution, where galaxies are very isolated and their evolution is not complicated by environmental effects (e.g. see van de Weygaert & Platen 2011, sec. 5). However, existing studies of void galaxies are concentrated at low redshift, where such objects are much easier to find (van de Weygaert et al. 2011). At low redshifts, the evidence points to the different properties of void galaxies being caused by their low stellar mass, independent of other influence from their void environment (Hoyle et al. 2005; Kreckel et al. 2011; Tinker et al. 2008). It would be very interesting to see whether similar behavior is seen at higher redshifts, where we expect the processes of galaxy formation could be different. Current galaxy redshift surveys can probe only down to $L \sim L_\star$ in galaxy luminosity at these redshifts, and we would naively expect high-redshift voids identified through Lyα forest tomography to also be void of such bright galaxies. However, the James Webb Space Telescope and its NIRSPEC spectrograph [3] will have the ability to target $L \sim 0.3L_\star$ galaxies within voids identified through CLAMATO and PFS.

With dense Lyα forest surveys covering larger volumes, such as a dedicated program on the PFS covering 200 deg^2, it is possible to identify a population of 10^4 voids. Such a large number of voids would naively translate to a one per cent AP measurement, although this is just a statistical estimate and it is possible that there would be larger systematic errors in such a measurement.

The data used in this project are available at:
`http://tinyurl.com/lya-tomography-sim-data`. We hope that making this data publicly available will reduce the barrier to future work on Lyα forest tomography and high-redshift voids. The data release includes gridded simulation quantities, the tomographic flux maps, a grid of hires flux skewers, FoF halo catalogs, void catalogs, and the protocluster catalog from (Stark et al. 2014). Due to space limitations, we downsampled the gridded quantities from the full 2560^3 grid to a 640^3 grid. Although this process erases some small-scale structure, the resolution is still more than enough for our purposes. The gridded quantities include the $z = 2.5$ density, redshift-space density, flux, real-space flux, and peculiar velocities and the $z = 0$ density and peculiar velocities. We also include example Python and C++ sources for reading the files.

---

[3] `http://www.stsci.edu/jwst/instruments/nirspec`

Acknowledgments

This work was published as Stark et al. (2015a). We thank Kathryn Kreckel for useful discussions.

Bibliography

Agertz, O., Moore, B., Stadel, J., et al. 2007, MNRAS, 380, 963

Alcock, C., & Paczynski, B. 1979, Nature, 281, 358

Aldrovandi, S. M. V., & Pequignot, D. 1973, A&A, 25, 137

Alexander, D. M., & Hickox, R. C. 2012, New A Rev., 56, 93

Almgren, A. S., Bell, J. B., Lijewski, M. J., Lukić, Z., & Van Andel, E. 2013, ApJ, 765, 39

Almgren, A. S., Beckner, V. E., Bell, J. B., et al. 2010, ApJ, 715, 1221

Angulo, R. E., Springel, V., White, S. D. M., et al. 2012, MNRAS, 425, 2722

Aubourg, É., Bailey, S., Bautista, J. E., et al. 2014, ArXiv e-prints, arXiv:1411.1074

Balogh, M. L., Pearce, F. R., Bower, R. G., & Kay, S. T. 2001, MNRAS, 326, 1228

Barnes, J., & Hut, P. 1986, Nature, 324, 446

Baugh, C. M. 2006, Reports on Progress in Physics, 69, 3101

Bechtold, J., Dobrzycki, A., Wilden, B., et al. 2002, ApJS, 140, 143

Becker, G. D., Bolton, J. S., Haehnelt, M. G., & Sargent, W. L. W. 2011, MNRAS, 410, 1096

Becker, G. D., Hewett, P. C., Worseck, G., & Prochaska, J. X. 2013, MNRAS, 430, 2067

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

Benson, A. J. 2012, Nature, 17, 175

Benson, A. J., & Bower, R. 2010, MNRAS, 405, 1573

Bernardi, M., Sheth, R. K., SubbaRao, M., et al. 2003, AJ, 125, 32

Beutler, F., Saito, S., Seo, H.-J., et al. 2014, MNRAS, 443, 1065

Beygu, B., Kreckel, K., van der Hulst, J. M., et al. 2015, ArXiv e-prints, arXiv:1501.02577

Bird, S., Vogelsberger, M., Sijacki, D., et al. 2013, MNRAS, 429, 3341

Black, J. H. 1981, MNRAS, 197, 553

Blake, C., Glazebrook, K., Davis, T. M., et al. 2011, MNRAS, 418, 1725

Blas, D., Lesgourgues, J., & Tram, T. 2011, J. Cosmology Astropart. Phys., 7, 34

Bleem, L. E., Stalder, B., Brodwin, M., et al. 2014, ArXiv e-prints, arXiv:1403.7186

Böhringer, H., Schuecker, P., Guzzo, L., et al. 2004, A&A, 425, 367

Bolton, J. S., & Becker, G. D. 2009, MNRAS, 398, L26

Bolton, J. S., Becker, G. D., Haehnelt, M. G., & Viel, M. 2014, MNRAS, 438, 2499

Bolton, J. S., Haehnelt, M. G., Viel, M., & Springel, V. 2005, MNRAS, 357, 1178

Bolton, J. S., Viel, M., Kim, T.-S., Haehnelt, M. G., & Carswell, R. F. 2008, MNRAS, 386, 1131

Bond, J. R., Kofman, L., & Pogosyan, D. 1996, Nature, 380, 603
Borde, A., Palanque-Delabrouille, N., Rossi, G., et al. 2014, J. Cosmology Astropart. Phys., 7, 5
Bos, E. G. P., van de Weygaert, R., Dolag, K., & Pettorino, V. 2012, MNRAS, 426, 440
Brown, P. N., Byrne, G. D., & Hindmarsh, A. C. 1989, SIAM J. Sci. Stat. Comput., 10, 1038
Bryan, G. L., Machacek, M., Anninos, P., & Norman, M. L. 1999, ApJ, 517, 13
Bryan, G. L., & Machacek, M. E. 2000, ApJ, 534, 57
Bryan, G. L., Norman, M. L., Stone, J. M., Cen, R., & Ostriker, J. P. 1995, Computer Physics Communications, 89, 149
Burenin, R. A., Vikhlinin, A., Hornstrup, A., et al. 2007, ApJS, 172, 561
Busca, N. G., Delubac, T., Rich, J., et al. 2013, A&A, 552, A96
Cai, Z., Fan, X., Bian, F., et al. 2014, in American Astronomical Society Meeting Abstracts, Vol. 223, American Astronomical Society Meeting Abstracts #223, 358.21
Caucci, S., Colombi, S., Pichon, C., et al. 2008, MNRAS, 386, 211
Ceccarelli, L., Padilla, N. D., Valotto, C., & Lambas, D. G. 2006, MNRAS, 373, 1440
Cen, R. 1992, ApJS, 78, 341
Cen, R., Miralda-Escudé, J., Ostriker, J. P., & Rauch, M. 1994, ApJ, 437, L9
Chan, K. C., Hamaus, N., & Desjacques, V. 2014, Phys. Rev. D., 90, 103521
Chiang, Y.-K., Overzier, R., & Gebhardt, K. 2013, ApJ, 779, 127
—. 2014, ApJ, 782, L3
Cisewski, J., Croft, R. A. C., Freeman, P. E., et al. 2014, MNRAS, 440, 2599
Clampitt, J., Cai, Y.-C., & Li, B. 2013, MNRAS, 431, 749
Clampitt, J., & Jain, B. 2014, ArXiv e-prints, arXiv:1404.1834
Coc, A., Uzan, J.-P., & Vangioni, E. 2013, ArXiv e-prints: 1307.6955, arXiv:1307.6955 [astro-ph.CO]
Colberg, J. M., Pearce, F., Foster, C., et al. 2008, MNRAS, 387, 933
Colella, P. 1990, Journal of Computational Physics, 87, 171
Colella, P., & Glaz, H. M. 1985, Journal of Computational Physics, 59, 264
Compostella, M., Cantalupo, S., & Porciani, C. 2013, MNRAS, 435, 3169
Conroy, C., Shapley, A. E., Tinker, J. L., Santos, M. R., & Lemson, G. 2008, ApJ, 679, 1192
Cooke, R. J., Pettini, M., Jorgenson, R. A., Murphy, M. T., & Steidel, C. C. 2014, ApJ, 781, 31
Courtois, H. M., Hoffman, Y., Tully, R. B., & Gottlöber, S. 2012, ApJ, 744, 43
Crighton, N. H. M., Bielby, R., Shanks, T., et al. 2011, MNRAS, 414, 28
Croft, R. A. C., Weinberg, D. H., Bolte, M., et al. 2002, ApJ, 581, 20
Croft, R. A. C., Weinberg, D. H., Katz, N., & Hernquist, L. 1998, ApJ, 495, 44
Croft, R. A. C., Weinberg, D. H., Pettini, M., Hernquist, L., & Katz, N. 1999, ApJ, 520, 1
Cucciati, O., Zamorani, G., Lemaux, B. C., et al. 2014, A&A, 570, A16
Dahlen, T., Mobasher, B., Somerville, R. S., et al. 2005, ApJ, 631, 126
D'Aloisio, A., & Furlanetto, S. R. 2007, MNRAS, 382, 860
Davis, M., Efstathiou, G., Frenk, C. S., & White, S. D. M. 1985, ApJ, 292, 371
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10

Delubac, T., Bautista, J. E., Busca, N. G., et al. 2014, ArXiv e-prints, arXiv:1404.1801

Diener, C., Lilly, S., Ledoux, C., et al. 2014, ArXiv e-prints, arXiv:1411.0649

Dobrzycki, A., Bechtold, J., Scott, J., & Morita, M. 2002, ApJ, 571, 654

Dodelson, S. 2003, Modern cosmology (Academic Press)

Dolag, K., Borgani, S., Murante, G., & Springel, V. 2009, MNRAS, 399, 497

Ebeling, H., Edge, A. C., Allen, S. W., et al. 2000, MNRAS, 318, 333

Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605

—. 1999, ApJ, 511, 5

Fabian, A. C. 1994, ARA&A, 32, 277

—. 2012, ARA&A, 50, 455

Fan, X., Strauss, M. A., Becker, R. H., et al. 2006, AJ, 132, 117

Faucher-Giguère, C.-A., Kereš, D., & Ma, C.-P. 2011, MNRAS, 417, 2982

Faucher-Giguère, C.-A., Lidz, A., Hernquist, L., & Zaldarriaga, M. 2008a, ApJ, 688, 85

Faucher-Giguère, C.-A., Lidz, A., Zaldarriaga, M., & Hernquist, L. 2009, ApJ, 703, 1416

Faucher-Giguère, C.-A., Prochaska, J. X., Lidz, A., Hernquist, L., & Zaldarriaga, M. 2008b, ApJ, 681, 831

Feretti, L., Giovannini, G., Govoni, F., & Murgia, M. 2012, A&A Rev., 20, 54

Ferland, G. J., Peterson, B. M., Horne, K., Welsh, W. F., & Nahar, S. N. 1992, ApJ, 387, 95

Finley, H., Petitjean, P., Noterdaeme, P., & Pâris, I. 2014, ArXiv e-prints, arXiv:1408.4799

Finoguenov, A., Guzzo, L., Hasinger, G., et al. 2007, ApJS, 172, 182

Font-Ribera, A., Miralda-Escudé, J., Arnau, E., et al. 2012, J. Cosmology Astropart. Phys., 11, 59

Font-Ribera, A., Arnau, E., Miralda-Escudé, J., et al. 2013, J. Cosmology Astropart. Phys., 5, 18

Font-Ribera, A., Kirkby, D., Busca, N., et al. 2014, J. Cosmology Astropart. Phys., 5, 27

Garzilli, A., Bolton, J. S., Kim, T.-S., Leach, S., & Viel, M. 2012, MNRAS, 424, 1723

Genel, S., Vogelsberger, M., Springel, V., et al. 2014, MNRAS, 445, 175

Gladders, M. D., & Yee, H. K. C. 2005, ApJS, 157, 1

Gnedin, N. Y., & Hui, L. 1998, MNRAS, 296, 44

Gontcho, S. G. A., Miralda-Escudé, J., & Busca, N. G. 2014, ArXiv e-prints, arXiv:1404.7425

Gregory, S. A., & Thompson, L. A. 1978, ApJ, 222, 784

Greig, B., Bolton, J. S., & Wyithe, J. S. B. 2014, ArXiv e-prints, arXiv:1411.1687

Guth, A. H. 1981, Phys. Rev. D., 23, 347

Haardt, F., & Madau, P. 1996, ApJ, 461, 20

—. 2012, ApJ, 746, 125

Habib, S., Pope, A., Lukić, Z., et al. 2009, Journal of Physics Conference Series, 180, 012019

Haehnelt, M. G., & Tegmark, M. 1996, MNRAS, 279, 545

Hamaus, N., Sutter, P. M., Lavaux, G., & Wandelt, B. D. 2014a, J. Cosmology Astropart. Phys., 12, 13

Hamaus, N., Sutter, P. M., & Wandelt, B. D. 2014b, ArXiv e-prints, arXiv:1409.7621

—. 2014c, Physical Review Letters, 112, 251302

Hamaus, N., Wandelt, B. D., Sutter, P. M., Lavaux, G., & Warren, M. S. 2014d, Physical Review Letters, 112, 041304

Hao, J., McKay, T. A., Koester, B. P., et al. 2010, ApJS, 191, 254

Heitmann, K., Lawrence, E., Kwan, J., Habib, S., & Higdon, D. 2014, ApJ, 780, 111

Heitmann, K., Lukić, Z., Fasel, P., et al. 2008, Computational Science and Discovery, 1, 015003

Hellwing, W. A., Juszkiewicz, R., & van de Weygaert, R. 2010, Phys. Rev. D., 82, 103536

Hernquist, L., Katz, N., Weinberg, D. H., & Miralda-Escudé, J. 1996, ApJ, 457, L51

Hill, G. J., Gebhardt, K., Komatsu, E., & MacQueen, P. J. 2004, in American Institute of Physics Conference Series, Vol. 743, The New Cosmology: Conference on Strings and Cosmology, ed. R. E. Allen, D. V. Nanopoulos, & C. N. Pope, 224

Hockney, R. W., & Eastwood, J. W. 1988, Computer simulation using particles (CRC Press)

Hoyle, F., Rojas, R. R., Vogeley, M. S., & Brinkmann, J. 2005, ApJ, 620, 618

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

Icke, V. 1984, MNRAS, 206, 1P

Ikeuchi, S. 1986, Ap&SS, 118, 509

Jenkins, E. B., & Ostriker, J. P. 1991, ApJ, 376, 33

Kaiser, N. 1987, MNRAS, 227, 1

Katz, N., Weinberg, D. H., & Hernquist, L. 1996, ApJS, 105, 19

Kauffmann, G., & Fairall, A. P. 1991, MNRAS, 248, 313

Kim, T.-S., Cristiani, S., & D'Odorico, S. 2001, A&A, 373, 757

Kim, T.-S., Viel, M., Haehnelt, M. G., Carswell, R. F., & Cristiani, S. 2004, MNRAS, 347, 355

Kirshner, R. P., Oemler, Jr., A., Schechter, P. L., & Shectman, S. A. 1981, ApJ, 248, L57

Kitaura, F.-S., Gallerani, S., & Ferrara, A. 2012, MNRAS, 420, 61

Kitaura, F. S., Jasche, J., Li, C., et al. 2009, MNRAS, 400, 183

Knebe, A., Knollmann, S. R., Muldrew, S. I., et al. 2011, MNRAS, 415, 2293

Kochanek, C. S., White, M., Huchra, J., et al. 2003, ApJ, 585, 161

Koester, B. P., McKay, T. A., Annis, J., et al. 2007, ApJ, 660, 239

Kolb, E. W., & Turner, M. S. 1990, The early universe. (Westview Press)

Kollmeier, J. A., Weinberg, D. H., Oppenheimer, B. D., et al. 2014, ApJ, 789, L32

Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, ApJS, 192, 18

Kravtsov, A. V., & Borgani, S. 2012, ARA&A, 50, 353

Kreckel, K., Joung, M. R., & Cen, R. 2011, ApJ, 735, 132

Kulkarni, G., Hennawi, J. F., Oñorbe, J., Rorai, A., & Springel, V. 2015, ArXiv e-prints, arXiv:1504.00366

Lacey, C., & Cole, S. 1994, MNRAS, 271, 676

Lavaux, G., & Wandelt, B. D. 2012, ApJ, 754, 109

Lee, J., & Park, D. 2009, ApJ, 696, L10

Lee, K.-G., Hennawi, J. F., White, M., Croft, R. A. C., & Ozbek, M. 2014a, ApJ, 788, 49

Lee, K.-G., Hennawi, J. F., Stark, C., et al. 2014b, ApJ, 795, L12

Lee, K.-G., Hennawi, J. F., Spergel, D. N., et al. 2015, ApJ, 799, 196

Li, B. 2011, MNRAS, 411, 2615
Li, B., Zhao, G.-B., & Koyama, K. 2012, MNRAS, 421, 3481
Lidz, A., Faucher-Giguère, C.-A., Dall'Aglio, A., et al. 2010, ApJ, 718, 199
Linde, A. D. 1982, Physics Letters B, 108, 389
Loeb, A., & Furlanetto, S. R. 2013, The First Galaxies in the Universe (Princeton University Press)
Longair, M. S., & Einasto, J., eds. 1978, IAU Symposium, Vol. 79, The large scale structure of the universe; Proceedings of the Symposium, Tallin, Estonian SSR, September 12-16, 1977
Lukić, Z., Stark, C. W., Nugent, P., et al. 2015, MNRAS, 446, 3697
Madau, P., & Dickinson, M. 2014, ARA&A, 52, 415
Marriage, T. A., Acquaviva, V., Ade, P. A. R., et al. 2011, ApJ, 737, 61
Martizzi, D., Jimmy, Romain, T., & Moore, B. 2014, MNRAS, 443, 1500
McCarthy, I. G., Schaye, J., Ponman, T. J., et al. 2010, MNRAS, 406, 822
McDonald, P. 2003, ApJ, 585, 34
McDonald, P., Miralda-Escudé, J., Rauch, M., et al. 2000, ApJ, 543, 1
McDonald, P., Seljak, U., Cen, R., Bode, P., & Ostriker, J. P. 2005a, MNRAS, 360, 1471
McDonald, P., Seljak, U., Cen, R., et al. 2005b, ApJ, 635, 761
McDonald, P., Seljak, U., Burles, S., et al. 2006, ApJS, 163, 80
McQuinn, M., Hernquist, L., Lidz, A., & Zaldarriaga, M. 2011, MNRAS, 415, 977
McQuinn, M., Lidz, A., Zaldarriaga, M., et al. 2009, ApJ, 694, 842
McQuinn, M., & White, M. 2011, MNRAS, 415, 2257
McQuinn, M., & Worseck, G. 2014, MNRAS, 440, 2406
Meiksin, A. 2000, MNRAS, 314, 566
Meiksin, A., Bryan, G., & Machacek, M. 2001, MNRAS, 327, 296
Meiksin, A., & Tittley, E. R. 2012, MNRAS, 423, 7
Meiksin, A., Tittley, E. R., & Brown, C. K. 2010, MNRAS, 401, 77
Meiksin, A., & White, M. 2001, MNRAS, 324, 141
—. 2004, MNRAS, 350, 1107
Meiksin, A. A. 2009, Reviews of Modern Physics, 81, 1405
Micheletti, D., Iovino, A., Hawken, A. J., et al. 2014, A&A, 570, A106
Mitchell, N. L., McCarthy, I. G., Bower, R. G., Theuns, T., & Crain, R. A. 2009, MNRAS, 395, 180
Murphy, D. N. A., Geach, J. E., & Bower, R. G. 2012, MNRAS, 420, 1861
Muzzin, A., Wilson, G., Yee, H. K. C., et al. 2009, ApJ, 698, 1934
Neyrinck, M. C. 2008, MNRAS, 386, 2101
Palanque-Delabrouille, N., Yèche, C., Borde, A., et al. 2013, A&A, 559, A85
Pan, D. C., Vogeley, M. S., Hoyle, F., Choi, Y.-Y., & Park, C. 2012, MNRAS, 421, 926
Park, D., & Lee, J. 2007, Physical Review Letters, 98, 081301
Peebles, P. J. E. 1968, ApJ, 153, 1
—. 1993, Principles of Physical Cosmology (Princeton University Press)
Peterson, J. R., & Fabian, A. C. 2006, Phys. Rep., 427, 1

Petitjean, P., Mueket, J. P., & Kates, R. E. 1995, A&A, 295, L9

Pichon, C., Vergely, J. L., Rollinde, E., Colombi, S., & Petitjean, P. 2001, MNRAS, 326, 597

Pierre, M., Pacaud, F., Duc, P.-A., et al. 2006, MNRAS, 372, 591

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2013a, ArXiv e-prints, arXiv:1303.5076 [astro-ph.CO]

—. 2013b, ArXiv e-prints, arXiv:1303.5089

—. 2015, ArXiv e-prints, arXiv:1502.01589

Platen, E., van de Weygaert, R., & Jones, B. J. T. 2007, MNRAS, 380, 551

Pontzen, A. 2014, Phys. Rev. D., 89, 083010

Postman, M., Lubin, L. M., Gunn, J. E., et al. 1996, AJ, 111, 615

Press, W. H., Rybicki, G. B., & Schneider, D. P. 1993, ApJ, 414, 64

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical recipes in C. The art of scientific computing (Cambridge University Press)

Rauch, M., Miralda-Escudé, J., Sargent, W. L. W., et al. 1997, ApJ, 489, 7

Reddy, N. A., Steidel, C. C., Pettini, M., et al. 2008, ApJS, 175, 48

Rees, M. J. 1986, MNRAS, 218, 25P

Regan, J. A., Haehnelt, M. G., & Viel, M. 2007, MNRAS, 374, 196

Reichardt, C. L., Stalder, B., Bleem, L. E., et al. 2013, ApJ, 763, 127

Ricotti, M., Gnedin, N. Y., & Shull, J. M. 2000, ApJ, 534, 41

Rorai, A., Hennawi, J. F., & White, M. 2013, ApJ, 775, 81

Rudie, G. C., Steidel, C. C., Trainor, R. F., et al. 2012, ApJ, 750, 67

Ryden, B. S. 1995, ApJ, 452, 25

Rykoff, E. S., Rozo, E., Busha, M. T., et al. 2014, ApJ, 785, 104

Saad, Y. 2003, Iterative methods for sparse linear systems (Siam)

Samushia, L., Reid, B. A., White, M., et al. 2014, MNRAS, 439, 3504

Schaye, J., Theuns, T., Rauch, M., Efstathiou, G., & Sargent, W. L. W. 2000, MNRAS, 318, 817

Scholz, T. T., & Walters, H. R. J. 1991, ApJ, 380, 302

Seljak, U., Makarov, A., McDonald, P., & Trac, H. 2006, Physical Review Letters, 97, 191303

Seo, H.-J., & Eisenstein, D. J. 2003, ApJ, 598, 720

Shapiro, P. R., & Kang, H. 1987, ApJ, 318, 32

Sheth, R. K., & van de Weygaert, R. 2004, MNRAS, 350, 517

Shull, J. M., France, K., Danforth, C. W., Smith, B., & Tumlinson, J. 2010, ApJ, 722, 1312

Sijacki, D., & Springel, V. 2006, MNRAS, 366, 397

Simons, D. A., Crampton, D., Côté, P., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9145, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 15

Skory, S., Hallman, E., Burns, J. O., et al. 2013, ApJ, 763, 38

Slosar, A., Ho, S., White, M., & Louis, T. 2009, J. Cosmology Astropart. Phys., 10, 19

Slosar, A., Font-Ribera, A., Pieri, M. M., et al. 2011, J. Cosmology Astropart. Phys., 9, 1

Slosar, A., Iršič, V., Kirkby, D., et al. 2013, J. Cosmology Astropart. Phys., 4, 26

Springel, V. 2005, MNRAS, 364, 1105

Springel, V., & Hernquist, L. 2002, MNRAS, 333, 649

Stark, C. W., Font-Ribera, A., White, M., & Lee, K.-G. 2015a, ArXiv e-prints, arXiv:1504.03290

Stark, C. W., Khandai, N., Lukić, Z., et al. 2015b, in prep.

Stark, C. W., White, M., Lee, K.-G., & Hennawi, J. F. 2014, ArXiv e-prints, arXiv:1412.1507

Strang, G. 1968, SIAM J. Numerical Analysis, 5, 506

Sutter, P. M., Lavaux, G., Wandelt, B. D., & Weinberg, D. H. 2012, ApJ, 761, 44

Sutter, P. M., Lavaux, G., Wandelt, B. D., et al. 2014a, MNRAS, 442, 3127

Sutter, P. M., Pisani, A., Wandelt, B. D., & Weinberg, D. H. 2014b, MNRAS, 443, 2983

Sutter, P. M., Lavaux, G., Hamaus, N., et al. 2015, Astronomy and Computing, 9, 1

Syphers, D., & Shull, J. M. 2014, ApJ, 784, 42

Szabo, T., Pierpaoli, E., Dong, F., Pipino, A., & Gunn, J. 2011, ApJ, 736, 21

Takada, M., Ellis, R. S., Chiba, M., et al. 2014, PASJ, 66, 1

Theuns, T., Leonard, A., Efstathiou, G., Pearce, F. R., & Thomas, P. A. 1998, MNRAS, 301, 478

Theuns, T., Zaroubi, S., Kim, T.-S., Tzanavaris, P., & Carswell, R. F. 2002, MNRAS, 332, 367

Tinker, J. L., Conroy, C., Norberg, P., et al. 2008, ApJ, 686, 53

Tittley, E. R., & Meiksin, A. 2007, MNRAS, 380, 1369

Tytler, D., Paschos, P., Kirkman, D., Norman, M. L., & Jena, T. 2009, MNRAS, 393, 723

van de Weygaert, R., & Platen, E. 2011, International Journal of Modern Physics Conference Series, 1, 41

van de Weygaert, R., Kreckel, K., Platen, E., et al. 2011, The Void Galaxy Survey (Springer), 17

Vanden Berk, D. E., Richards, G. T., Bauer, A., et al. 2001, AJ, 122, 549

Verner, D. A., & Ferland, G. J. 1996, ApJS, 103, 467

Viel, M., Becker, G. D., Bolton, J. S., & Haehnelt, M. G. 2013a, Phys. Rev. D., 88, 043502

Viel, M., Colberg, J. M., & Kim, T.-S. 2008, MNRAS, 386, 1285

Viel, M., Haehnelt, M. G., & Springel, V. 2004, MNRAS, 354, 684

—. 2006, MNRAS, 367, 1655

Viel, M., Matarrese, S., Mo, H. J., Theuns, T., & Haehnelt, M. G. 2002, MNRAS, 336, 685

Viel, M., Schaye, J., & Booth, C. M. 2013b, MNRAS, 429, 1734

Vogeley, M. S., Geller, M. J., Park, C., & Huchra, J. P. 1994, AJ, 108, 745

Vogelsberger, M., Genel, S., Sijacki, D., et al. 2013, MNRAS, 436, 3031

Vogt, S. S., Allen, S. L., Bigelow, B. C., et al. 1994, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, Vol. 2198, Proc. SPIE Instrumentation in Astronomy VIII, David L. Crawford; Eric R. Craine; Eds., Volume 2198, p. 362, ed. D. L. Crawford & E. R. Craine, 362

Voronov, G. S. 1997, Atomic Data and Nuclear Data Tables, 65, 1

Weinberg, S. 1993, The first three minutes : a modern view of the origin of the universe (Basic Books)

White, M. 2001, A&A, 367, 27

—. 2002, ApJS, 143, 241
White, M., Pope, A., Carlson, J., et al. 2010, ApJ, 713, 383
Wiener, N. 1949, Extrapolation, interpolation, and smoothing of stationary time series, Vol. 2 (MIT press Cambridge, MA)
Willick, J. A. 2000, ArXiv Astrophysics e-prints, astro-ph/0003232
Wilson, G., Muzzin, A., Yee, H. K. C., et al. 2009, ApJ, 698, 1943
Worseck, G., Prochaska, J. X., Hennawi, J. F., & McQuinn, M. 2014, ArXiv e-prints: 1405.7405, arXiv:1405.7405
Yang, H.-Y. K., Sutter, P. M., & Ricker, P. M. 2012, MNRAS, 427, 1614
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579
Yuan, T., Nanayakkara, T., Kacprzak, G. G., et al. 2014, ApJ, 795, L20
Zaldarriaga, M. 2002, ApJ, 564, 153
Zel'dovich, Y. B. 1970, A&A, 5, 84
Zhang, Y., Anninos, P., & Norman, M. L. 1995, ApJ, 453, L57
Zhang, Y., Anninos, P., Norman, M. L., & Meiksin, A. 1997, ApJ, 485, 496
Zhang, Y., Meiksin, A., Anninos, P., & Norman, M. L. 1998, ApJ, 495, 63

Appendix A

Notation

In this thesis, I stick to standard community conventions and symbols. In order to make this explicit, I have attempted to make a complete list of notation that often goes unmentioned.

- The most common time coordinates in cosmology are the redshift z or the scale factor $a = \frac{1}{1+z}$. At the Big Bang, z is infinite and a is 0, and today, z is 0 and a is 1.
- The Hubble rate is $H = (da/dt)/a = H_0 E$, where H_0 is the Hubble rate today. The Hubble rate today is typically parameterized as $H_0 = 100\,h\,\mathrm{Mpc}^{-1}\,\mathrm{km\,s}^{-1}$.
- The (proper) time since the Big Bang is $t = \frac{1}{H_0}\int_0^a \frac{a' da'}{E(a')} = \frac{1}{H_0}\int_\infty^z \frac{dz'}{(1+z')E(z')}$. The evolution function $E(z)$ depends on the cosmological model adopted, based on which energy components are included and how their energy density dilutes with expansion.

For any numbers quoted in this dissertation that require an assumed cosmology, I take a fiducial flat ΛCDM model with:
$\Omega_\mathrm{b} = 0.05$, $\Omega_\mathrm{m} = 0.3$, $\Omega_\Lambda = 0.7$, $h = 0.7$, and $\sigma_8 = 0.8$.
In this model, the Universe includes radiation (relativistic components), baryons and dark matter (collectively referred to as matter), and a 'cosmological constant' type of dark energy. The symbols for the energy densities of these components are, respectively, ρ_r, ρ_b and ρ_dm ($\rho_\mathrm{m} = \rho_\mathrm{b} + \rho_\mathrm{dm}$), and ρ_Λ.

- The ratio of the mean energy density of a specific component X to the critical density is Ω_X, e.g. $\Omega_\mathrm{m} = \bar{\rho}_\mathrm{m}/\rho_\mathrm{crit}$.
- These values are typically evaluated at $z = 0$, sometimes indicated with a 0 subscript, e.g. $\Omega_\mathrm{m,0}$, although this is usually omitted for brevity or to annoy readers.
- Since the Universe is flat in this model, $\Omega r + \Omega_\mathrm{m} + \Omega_\Lambda = 1$ always.
- The evolution function is $E(a) = \sqrt{\Omega_\mathrm{r,0} a^{-4} + \Omega_\mathrm{m,0} a^{-3} + \Omega_{\Lambda,0}}$.

We can relate time, distance, and wavelength coordinates assuming traveling at the speed of light.

- The comoving scale x is related to the proper scale $r = ax$, and $dr = c\,dt = a\,dx = \frac{c}{aH}da = -\frac{c}{(1+z)H}dz$. Thus, the comoving distance x to some redshift is $x(z) = \frac{c}{H_0}\int_0^z \frac{dz'}{E(z')}$.
- The peculiar velocity $v = dr/dt - \dot{a}x = a\dot{x}$ is then related as $dr = \frac{1}{H}dv$.
- Finally, we can also use wavelength or frequency coordinates: if the observed wavelength is λ and the rest-frame wavelength is λ_0, $\lambda = (1+z)\lambda_0$, so that $dr = \frac{c}{\lambda_0 H}d\lambda$.

All distances and scales are comoving and are quoted in 'h units' such as h^{-1}Mpc. All densities are comoving and typically in mean units. Masses are typically in units of $h^{-1}\mathrm{M}_\odot$. Velocities are always the peculiar velocity, and are typically in units of $\mathrm{km\,s^{-1}}$.

A.1 Fourier convention and power spectra

The comoving mode k corresponding to the comoving scale x is $k = 2\pi/x$. The Fourier transform of a periodic field f, sampled over a volume V in n dimensions, is $\hat{f}$.

$$\begin{aligned} \hat{f}(\mathbf{k}) &= \int f(\mathbf{x}) e^{i\mathbf{k}\cdot\mathbf{x}} d^n x \\ f(\mathbf{x}) &= \frac{1}{(2\pi)^n}\int \hat{f}(\mathbf{k}) e^{-i\mathbf{k}\cdot\mathbf{x}} d^n k \end{aligned} \tag{A.1}$$

Note that if f is dimensionless, this convention forces $\hat{f}$ to have dimensions of volume.

The auto-correlation field of f is

$$\xi_{ff}(\mathbf{x}) = \langle f(\mathbf{x}') f(\mathbf{x}' + \mathbf{x}) \rangle \ , \tag{A.2}$$

where the brackets indicate a volume-weighted average over all $\mathbf{x}'$. The auto-power spectrum field of f, $P_{ff}(\mathbf{k})$ is defined as the Fourier transform of $\xi_{ff}(\mathbf{x})$. After some manipulation, we have

$$P_{ff}(\mathbf{k}) = V \hat{f}(\mathbf{k}) \hat{f}^*(\mathbf{k}) \ , \tag{A.3}$$

where the $*$ superscript means the complex conjugate. When we mention the power spectrum in cosmology, $P(k)$, we typically mean the spherically-averaged auto power spectrum field of some quantity. This is because if the field is isotropic, the spherically-averaged power spectrum contains all of the information of the power spectrum field. Thus,

$$P(k) = V \langle \hat{f}(\mathbf{k}) \hat{f}^*(\mathbf{k}) \rangle \ , \tag{A.4}$$

where the average is over shells of $k = |\mathbf{k}|$. Note that the power spectrum also has units of volume. Sometimes we consider the dimensionless power spectrum instead:

$$\Delta^2(k) = \frac{d\sigma^2}{d\log k} \ , \tag{A.5}$$

where $\sigma^2 = \frac{1}{(2\pi)^n} \int P(k) d^n k$ is the mass variance. So in 3D, $\Delta^2(k) = \frac{k^3 P(k)}{2\pi^2}$, and in 1D, $\Delta^2(k) = \frac{kP(k)}{\pi}$.

Appendix B

Atomic rates

Here we provide some details on the reaction rates implemented in the Nyx code. In order to provide an easy comparison to the Gadget code, we have also implemented the atomic rates from Katz et al. (1996). However, we also implement more accurate rates which will be used in future Nyx studies exploring cosmological effects in Lyα forest statistics.

We explicitly keep track of the net loss of thermal energy resulting from atomic collisional processes. Those rates are shown in Table B.2. In addition to the tabulated cooling rates, Nyx includes cooling from inverse-Compton scattering off CMB photons as in Peebles (1968):

$$L_C = \frac{4\sigma_T a k_B}{m_e c} n_e T_{\rm CMB}^4(z) \left[T - T_{\rm CMB}(z)\right] \ , \tag{B.1}$$

where σ_T is the Thomson cross section, a is the radiation density constant, k_B is the Boltzmann constant, m_e is the electron mass, c is the speed of light, and $T_{\rm CMB}$ is the temperature of the microwave background, which we take to be $T_{\rm CMB} = 2.725$.

The atomic rates are a compilation of observed laboratory data, and as such, the fitting functions are used to interpolate and extrapolate between and beyond these data points. In the literature many different fits to the atomic rates have been used in IGM simulations. To build intuition for the differences they make on the Lyα forest flux, we present the different hydrogen recombination rates found in several works, including the Enzo and Gadget codes used for most of the Lyα forest simulations in recent years. Since $\tau \propto n_{\rm H\,I} \propto \alpha_{r,{\rm H\,II}}$, an error in the hydrogen recombination rate directly propagates to the same error in τ. As Figure B.1 demonstrates, some of the fits are inaccurate by ~ 20 per cent at $T = 10^4\,{\rm K}$, even though it can be calculated from first principles, as done e.g. in Ferland et al. (1992). We show a large temperature range for completeness; at low temperatures three-body recombination is dominant at most densities, whereas at very high temperatures the neutral hydrogen fraction is vanishingly small.

Another essential ingredient in modeling the thermal and ionization history of the IGM is the ultra-violet background, and especially the hydrogen photoionization rate, Γ_{HI}. Due to its low surface brightness this is not a directly measurable quantity in observations, however, indirectly it can be seen via such astrophysical phenomena as the quasar proximity effect.

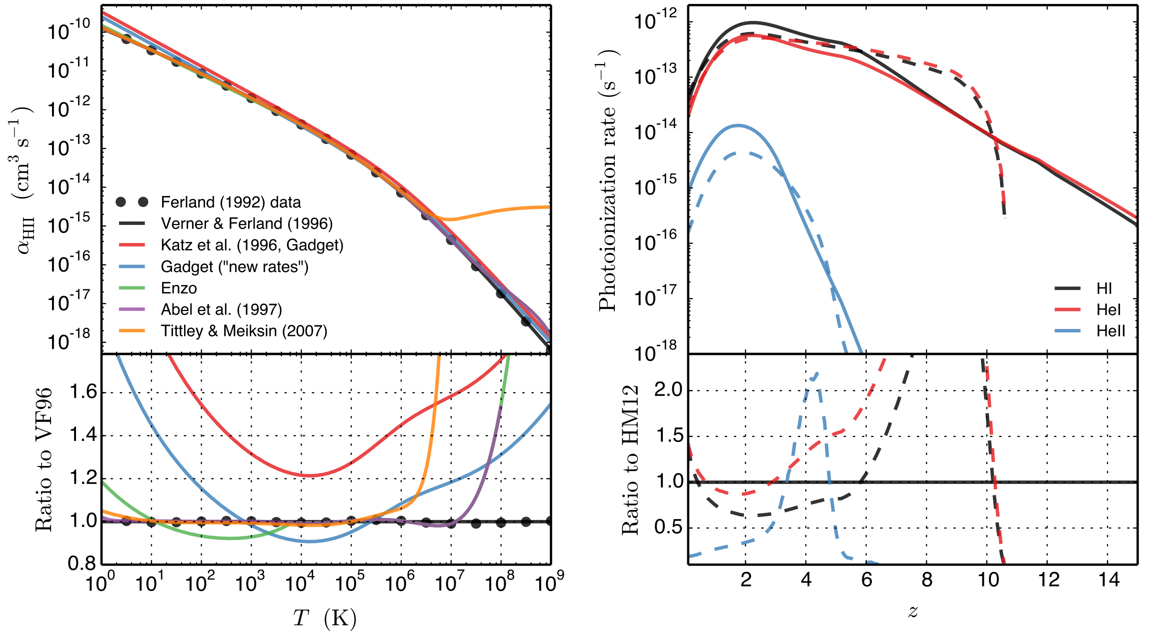

Figure B.1: *Left*: A comparison of different hydrogen recombination rates used in recent simulations in the literature and Verner & Ferland (1996) calculated data. *Right*: Comparison of photoionization rates published in recent works: solid lines are Haardt & Madau (2012), and dashed lines are Faucher-Giguère et al. (2009) (revised in Faucher-Giguère et al. 2011).

Table B.1: Atomic rates in Nyx

Coefficient	Fitting formula [cm^{3}s^{-1}]	Comment
	$a\left[\sqrt{T/T_0}\left(1+\sqrt{T/T_0}\right)^{1-b}\left(1+\sqrt{T/T_1}\right)^{1+b}\right]^{-1}$	Verner & Ferland (1996)
$\alpha_{\rm r,HII}$	$a = 7.982\times10^{-11}$, $b = 0.7480$, $T_0 = 3.148$, $T_1 = 7.036\times10^{5}$	
$\alpha_{\rm r,HeII}$	$a = 3.294\times10^{-11}$, $b = 0.6910$, $T_0 = 15.54$, $T_1 = 3.676\times10^{7}$	$(T \le 10^6)$
$\alpha_{\rm r,HeII}$	$a = 9.356\times10^{-10}$, $b = 0.7892$, $T_0 = 4.266\times10^{-2}$, $T_1 = 4.677\times10^{6}$	$(T > 10^6)$
$\alpha_{\rm r,HeIII}$	$a = 1.891\times10^{-10}$, $b = 0.7524$, $T_0 = 9.370$, $T_1 = 2.774\times10^{6}$	
$\alpha_{\rm d,HeII}$	$1.9\times10^{-3}\left(1+0.3e^{\frac{-9.4\times10^4}{T}}\right)e^{\frac{-4.7\times10^5}{T}}T^{-\frac{3}{2}}$	Aldrovandi & Pequignot (1973)
	$A\dfrac{(1+P\,U^{1/2})}{(X+U)}U^m e^{-U}$; $U = \dfrac{11604.5E}{T}$	Voronov (1997)
$\Gamma_{e,\rm HI}$	$A = 2.91\times10^{-8}$, $E = 13.6$, $P = 0$, $X = 0.232$, $m = 0.39$	
$\Gamma_{e,\rm HeI}$	$A = 1.75\times10^{-8}$, $E = 24.6$, $P = 0$, $X = 0.180$, $m = 0.35$	
$\Gamma_{e,\rm HeII}$	$A = 2.05\times10^{-9}$, $E = 54.4$, $P = 1$, $X = 0.265$, $m = 0.25$	

Recombination (α_i) and collisional ionization ($\Gamma_{e\,i}$) rates in the Nyx code. $\alpha_{\rm d,HeII}$ is the dielectronic recombination rate of singly-ionized helium. Temperatures are in K, and rates are tabulated in the code in the temperature range $1 \le T \le 10^9$ K.

Table B.2: Cooling rates in NYX

Type	Fitting formula [erg cm^{3}s^{-1}]	Comment
Bremsstrahlung	$1.426 \times 10^{-27} T^{\frac{1}{2}} Z_i^2 \langle g_{ff} \rangle$; $\langle g_{ff} \rangle = \begin{cases} 0.79464 + 0.1243 \log(T/Z^2); & T/Z^2 \leq 3.2 \times 10^5\,\mathrm{K} \\ 2.13164 - 0.1240 \log(T/Z^2); & T/Z^2 > 3.2 \times 10^5\,\mathrm{K} \end{cases}$	Shapiro & Kang (1987)
Neutral Hydrogen		Scholz & Walters (1991)
	$10^{-20}\exp(213.7913 - 113.9492y + 25.06062y^2 - 2.762755y^3 + 0.1515352y^4$ $-3.290382 \times 10^{-3}y^5 - 1.18415 \times 10^5 T^{-1})$	$2 \times 10^3 \leq T \leq 10^5$
	$10^{-20}\exp(271.25446 - 98.019455y + 14.00728y^2 - 0.9780842y^3 + 3.356289 \times 10^{-2}y^4 -$ $-4.553323 \times 10^{-4}y^5 - 1.18415 \times 10^5 T^{-1})$	$T > 10^5$
	$y \equiv \ln(T)$	
Helium		Black (1981)
HeI	$9.38 \times 10^{-22} T^{\frac{1}{2}} e^{-285335.4/T} \left(1 + \sqrt{\frac{T}{5 \times 10^7}}\right)^{-1}$	
HeII	$\left(5.54 \times 10^{-17} T^{-0.397} e^{-473638/T} + 4.85 \times 10^{-22} T^{\frac{1}{2}} e^{-631515/T}\right) \left(1 + \sqrt{\frac{T}{5 \times 10^7}}\right)^{-1}$	
Recombinations		Black (1981)
HII	$2.851 \times 10^{-27} T^{\frac{1}{2}} \left(5.914 - \frac{1}{2}\ln T + 0.01184 T^{\frac{1}{3}}\right)$	
HeII	$1.55 \times 10^{-26} T^{0.3647} + 1.24 \times 10^{-13} \left(1 + 0.3e^{\frac{-9.4 \times 10^4}{T}}\right) e^{\frac{-4.7 \times 10^5}{T}} T^{-\frac{3}{2}}$	
HeIII	$1.140 \times 10^{-26} T^{\frac{1}{2}} \left(6.607 - \frac{1}{2}\ln T + 7.459 \times 10^{-3} T^{\frac{1}{3}}\right)$	

The Cooling rates used in NYX. Note that the helium rates are from Black (1981) but were modified by a different temperature factor than in Cen (1992). In the Bremsstrahlung expression, $Z = 1$ for HII and HeII, and $Z = 2$ for HeIII. Temperature is in K, and the rates are tabulated in the code in temperature range $1 \leq T \leq 10^9$ K.

These kinds of measurements are quite uncertain, and instead one often tries to calculate the UVB intensity and spectral shape by combining all possible sources of ionizing flux (Haardt & Madau 2012). These calculations are also quite uncertain and have a large number of input assumptions.

It is beyond the scope of this work to examine all the potential physical processes and simulations used to create these different models of the UVB, as well as their accuracy; we refer interested readers a recent work by Kollmeier et al. (2014). Instead, in Figure B.1 we simply show the differences in photoionization rates of the most recent works on this topic by Haardt & Madau (2012) and Faucher-Giguère et al. (2009). Note that the latter rates were updated in 2011 (Faucher-Giguère et al. 2011). The right panel of Figure B.1 clearly demonstrates that differences between the two works – and therefore our understanding of the UVB – are rather significant. The effect of different rates on the temperature of the IGM we show in Figure 2.5, confirming that the 'feedback' onto the dynamical evolution of the gas is much smaller.

Appendix C

Reconstruction derivation and implementation

In this appendix, we briefly review the Wiener filter, to establish our notation, and describe our efficient numerical algorithm for map making.

C.1 Wiener filter

We assume our data is made up of the signal we are interested in and additive noise $\mathbf{d} = \mathbf{s}_p + \mathbf{n}$. In order to keep coordinates clear, we use a p subscript to indicate 'pixel' coordinates, and an m subscript to indicate 'map' coordinates. Note that some other texts characterize this difference with the instrumental response matrix $\mathbf{R}$ as $\mathbf{s}_p = \mathbf{R}\mathbf{s}_m$. We want to make a linear estimate of the signal $\hat{\mathbf{s}} = \mathbf{L}\mathbf{d}$, with minimal error $\epsilon = E[|\mathbf{s}_m - \hat{\mathbf{s}}|^2]$. We start by simplifying the error expression.

$$\epsilon = \mathrm{tr}\left(E[\mathbf{s}_m\mathbf{s}_m^T] - E[\mathbf{s}_m\hat{\mathbf{s}}^T] - E[\hat{\mathbf{s}}\mathbf{s}_m^T] + E[\hat{\mathbf{s}}\hat{\mathbf{s}}^T]\right)$$

The first term $E[\mathbf{s}_m\mathbf{s}_m^T]$ is just the signal covariance $\mathbf{S}_{mm}$. The second term is

$$\begin{aligned}
E[\mathbf{s}_m\hat{\mathbf{s}}^T] &= E[\mathbf{s}_m(\mathbf{L}\mathbf{d})^T] = E[\mathbf{s}_m\mathbf{d}^T\mathbf{L}^T] \\
&= E[\mathbf{s}_m(\mathbf{s}_p^T + \mathbf{n}^T)\mathbf{L}^T] \\
&= (E[\mathbf{s}_m\mathbf{s}_p^T] + E[\mathbf{s}_m\mathbf{n}^T])\mathbf{L}^T = \mathbf{S}_{mp}\mathbf{L}^T
\end{aligned}$$

since we assume $E[\mathbf{s}_m\mathbf{n}^T] = 0$. By a similar manipulation, the third term $E[\hat{\mathbf{s}}\mathbf{s}_m^T] = \mathbf{L}\mathbf{S}_{pm}$. The fourth term is

$$\begin{aligned}
E[\hat{\mathbf{s}}\hat{\mathbf{s}}^T] &= \mathbf{L}E[\mathbf{d}\mathbf{d}^T]\mathbf{L}^T \\
&= \mathbf{L}(E[\mathbf{s}_p\mathbf{s}_p^T] + E[\mathbf{s}_p\mathbf{n}^T] + E[\mathbf{n}\mathbf{s}_p^T] + E[\mathbf{n}\mathbf{n}^T])\mathbf{L}^T \\
&= \mathbf{L}(\mathbf{S}_{pp} + \mathbf{N})\mathbf{L}^T
\end{aligned}$$

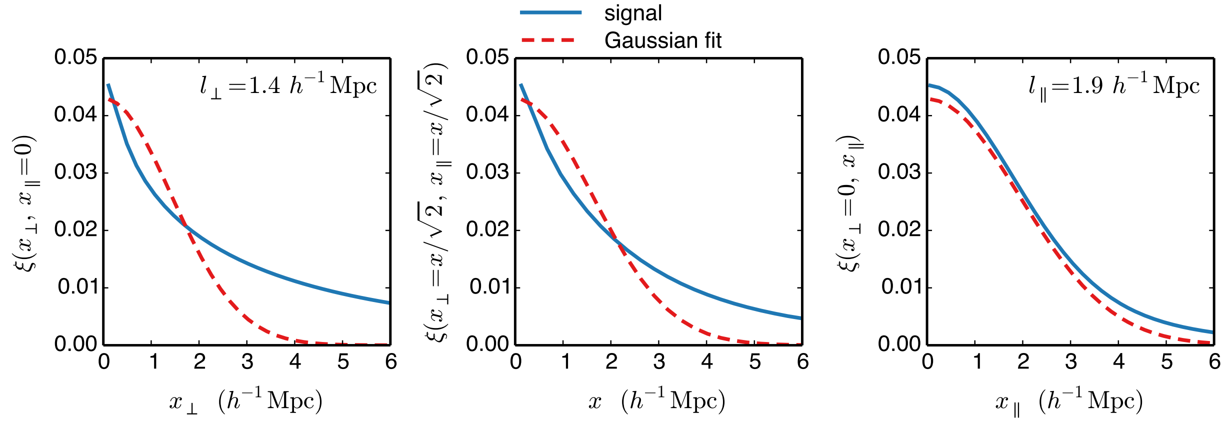

Figure C.1: The signal correlation function (solid line) compared to our assumed Gaussian form (dashed line), with amplitude, $l_\perp$, and $l_\parallel$ fit to the signal data.

Altogether, the error is

$$\begin{aligned}\epsilon &= \mathrm{tr}\mathbf{S}_{mm} - \mathrm{tr}(\mathbf{S}_{pm}\mathbf{L}^T) - \mathrm{tr}(\mathbf{L}\mathbf{S}_{mp}) + \mathrm{tr}(\mathbf{L}(\mathbf{S}_{pp} + \mathbf{N})\mathbf{L}^T) \\ &= \mathrm{tr}\mathbf{S}_{mm} - 2\mathrm{tr}(\mathbf{L}\mathbf{S}_{mp}) + \mathrm{tr}(\mathbf{L}(\mathbf{S}_{pp} + \mathbf{N})\mathbf{L}^T)\end{aligned}$$

Taking the derivative of the error with respect to the operator, we have

$$\frac{\partial \epsilon}{\partial \mathbf{L}} = -2\mathbf{S}_{mp} + 2(\mathbf{S}_{pp} + \mathbf{N})^T\mathbf{L}^T$$

And then evaluating $\partial\epsilon/\partial\mathbf{L} = 0$ to find the minimum error, we have the optimal operator $\mathbf{L} = \mathbf{S}_{mp}(\mathbf{S}_{pp} + \mathbf{N})^{-1}$.

C.2 Signal covariance

The form we assume for the signal covariance is a product of two Gaussians, as shown in Equation 4.2. The flux correlation function should roughly have this form, but it is certainly not correct in detail. In this section, we consider the difference between the true signal covariance and our model (with appropriate $l_\perp$ and $l_\parallel$ values), and how this model inadequacy might affect our reconstruction results.

In Figure C.1, we compare the correlation function of δ_F from the simulation (labeled signal) and our model fit to the ideal signal (labeled Gaussian fit). We have smoothed the signal along the line of sight to match a typical spectrograph resolution ($R \sim 1100$). From left to right, the three panels show different slices through the $(x_\perp, x_\parallel)$ plane, first all perpendicular, for $x_\perp = x_\parallel$, and for all parallel. We also annotated the fit Gaussian scales $l_\perp$ and $l_\parallel$. The Gaussian product shape does well along the line of sight, due to the fact that we have mocked the instrumental smoothing with a Gaussian filter, and the unsmoothed

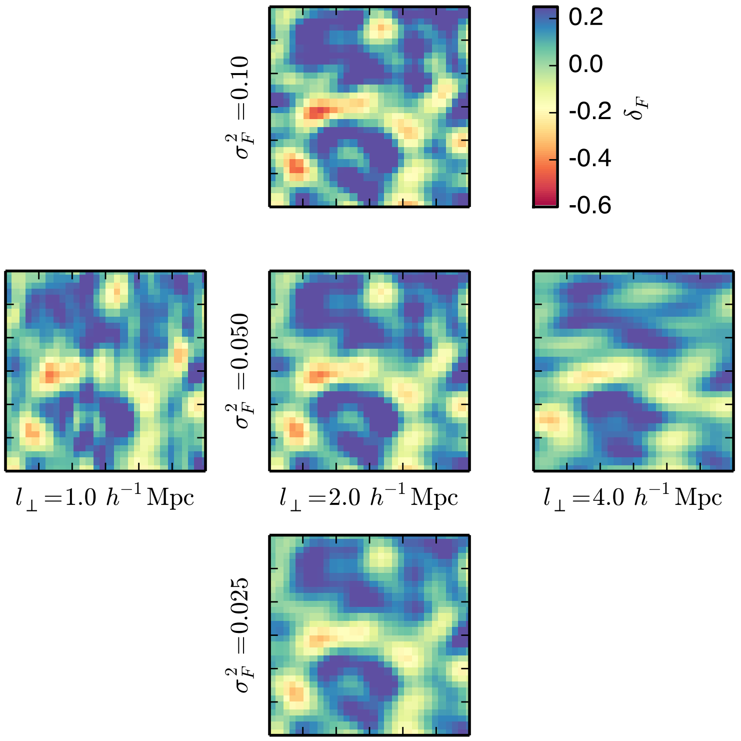

Figure C.2: Slices through reconstructions run on the same pixel data, varying the signal covariance parameters $l_\perp$ and σ_F^2. The slices are $2\,h^{-1}$Mpc thick, projected over the x-axis (perpendicular to line of sight). The vertical axis of the images is along the line of sight. From left to right, we double the $l_\perp$ value, from 1 to 4. From bottom to top, we double the σ_F^2 value, from 0.025 to 0.1. Any reasonable deviations from the best case parameter values do not affect the morphology of the resulting map.

flux correlation is small for scales larger than the filter scale. Across the line of sight, our model does much worse. In a future iteration, we will consider using a sum of Gaussians. Such a mismatch between the simulation and model might be worrying, but we argue that this is not a concern for our application. In the case of Wiener filtering, most elements of the operator $\mathbf{S}(\mathbf{S}+\mathbf{N})^{-1}$ are close to 0 or 1, and the shape of $\mathbf{S}$ only changes values in the intermediate regime (see e.g. Press et al. 1992, for discussion).

In order to test the effect of an inaccurate covariance assumption on the reconstruction, we ran several reconstructions of the same pixel data, varying the signal covariance parameters $l_\perp$ and σ_F^2. The same slice from each reconstruction is shown in Figure C.2. We vary $l_\perp$ from left to right and σ_F^2 from top to bottom. Overall, it appears that any reasonable changes to the parameter values (relative to the best-fit values) do not affect the morphology of the structures in the map. Increasing the flux variance increases the variance in the final map. This is due to the increase in all $\mathbf{S}$ elements relative to $\mathbf{N}$ so that pixels have larger weights in the reconstruction. That is, increasing the flux variance parameter should have the same effect as reducing all pixel noise estimates. Varying the correlation scale $l_\perp$ (or $l_\parallel$) has a more dramatic effect. With a fixed sightline sampling and a smaller correlation scale, the noise will obviously have a larger effect on the map, as the pixels are less correlated. As we increase the correlation scale, structures become increasing smoothed out. We found that changes in $l_\parallel$ behave the same as changes in $l_\perp$, so we did not add it to the plot.

C.3 Numerical implementation and scaling

Computationally, the map making process consists of two steps. First, there is the matrix inversion and matrix-vector multiply $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = (\mathbf{S}_{pp} + \mathbf{N})^{-1}\mathbf{d}$. The second step of the map process is just the multiplication $\mathbf{m} = \mathbf{S}_{mp}\mathbf{x}$. The matrix $\mathbf{A}$ is symmetric and positive definite, so there are several computationally efficient methods for obtaining the solution $\mathbf{x}$. Since our signal and noise matrices are both relatively sparse, we use the preconditioned conjugate gradient (PCG) method with a Jacobi pre-conditioner (Saad 2003) [1]. PCG is an iterative method which converges rapidly for sufficiently sparse matrices. For reasonable survey strategies, we do not expect a large number of pixels within a flux correlation scale (several Mpc), so methods that perform better for sparse matrices should be advantageous. We use the stopping condition that the residual is smaller than the norm of the data times a tolerance parameter, $|\mathbf{r}| = |\mathbf{b} - \mathbf{A}\mathbf{x}| < \mathrm{tol}|\mathbf{b}|$.

The real advantage of PCG for our problem, however, is that it never uses $\mathbf{A}$ directly, but only products of $\mathbf{A}$ and a vector. Since we know the functional form of $\mathbf{S}_{pp}$, and we assume $\mathbf{N}$ is diagonal, we do not have store the matrix $\mathbf{A}$, and instead compute elements when needed. This changes the space complexity of the algorithm from N_{pix}^2 to N_{pix}. For a typical problem where $N_{\mathrm{pix}} = 10^6$ the difference in storage is about 8 TB (for $\mathbf{A}$ stored in double precision), demanding several hundred nodes on modern systems, versus six vectors of length N_{pix}, requiring about 50 MB and easily fitting on a single node. Clearly, the performance of the PCG solve depends on how quickly we can compute elements of $\mathbf{A}$. We speed up the element lookup by using a small table of $\exp(x)$ for the Gaussian. This reduces each element lookup to 10 add/multiply operations.

Altogether, the cost of the reconstruction algorithm is $N_{\mathrm{lookup}}(N_{\mathrm{iter}}N_{\mathrm{pix}}^2 + N_{\mathrm{map}}N_{\mathrm{pix}})$, where N_{lookup} is the number of operations involved in computing elements of $\mathbf{A}$ and N_{iter} is the number of iterations before the PCG reaches the stop condition. We expect problem sizes of up to 10^6 pixels and 10^6 map points, so assuming 100 iterations, the calculation takes 10^{15} operations. This estimate indicates that we will likely not need to parallelize the code beyond shared memory, especially since the number of cores per node is expected to increase in coming years.

In order to choose a tolerance value for the PCG stop condition, we tested the PCG result against a direct Cholesky factorization for small problems. We generated a mock dataset with $N_{\mathrm{pix}} = 4000$, fixed the pixel positions and signal, and generated 10 noise realizations with SNR = 5. With multiple noise realizations, we can estimate the map variance due to noise compared to the error of the PCG solve. For each of the 10 noise realizations, we ran the reconstruction with the Cholesky solve and with the PCG solve with tol values of 1, 0.1, 0.01, and 10^{-3}. The Cholesky reconstruction took 18 seconds on average while the lowest tol PCG reconstruction took 0.9 seconds on average. The average number of PCG iterations to reach the various tol values were 5, 13, 29, and 37. We computed the standard deviation of the Cholesky map values over noise realizations σ_m to have a measure of the variance due to the noise at each map point. The average σ_m is 0.06 and the max is 0.11. We then computed the absolute difference of the PCG maps and Cholesky maps relative to the map noise std., $\epsilon = |\hat{\mathbf{s}}_{\text{PCG}} - \hat{\mathbf{s}}_{\text{Chol}}|/\sigma_m$. This error captures the fact that the PCG error must be smaller for map points with small noise variance. We found that the errors have an exponential distribution, with maximum values over all map points of 30, 4.5, 0.47, and 0.059 respectively for the 4 PCG tolerance settings. Since the max error of the tol = 0.01 PCG maps is less than unity and the error distribution is exponential, this tolerance setting is in the safe regime where the PCG residual error in the map is significantly smaller than the noise. In practice the PCG tol value should be adjusted for the problem at hand (if the SNR is very different), but this is a conservative choice for Lyα forest data in the near future.

---

[1] Also see `http://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf`

One practical issue with the expressions in Equation 4.1 is that it does not easily allow for masking bad pixels. If we have any pixels with $n = \inf$, the PCG routine will return nan's. The ability to mask data is critical for Lyα forest data, where we may run into sky lines that add significant noise, or any pixels that should be masked entirely. We can rewrite the map expression using the fact that the noise covariance may be formed as a product of a lower triangular matrix with its transpose. In the case of our noise covariance, this $N_{ij} = (n_i \delta_{ik})(n_j \delta_{jk})^T$. It follows that

$$\mathbf{m} = \mathbf{S}_{mp}\mathbf{w}(\mathbf{w}\mathbf{S}_{pp}\mathbf{w} + \mathbf{I})^{-1}\mathbf{w}\mathbf{d} \tag{C.1}$$

where $\mathbf{w} = \mathbf{n}^{-1}$. In this new expression, the matrix to be inverted is definite even for pixels with $w = 0$, and the PCG solves will work as expected. This expression requires more operations than the simpler Equation 4.1, but they add negligible overhead.

The reconstruction code implemented for this work consists of a static library and a few executables, written in C++, with no dependencies. The code can be compiled and run with no parallelism, but we recommend enabling OpenMP if available. The code is publicly available at `http://github.com/caseywstark/dachshund`, and includes some documentation and a test suite.

We performed scaling tests of our code to give an idea of what problem scale the code is able to handle within a reasonable wall time. We ran the test problems on the Edison machine at NERSC. Each Edison node has two 12-core Intel "Ivy Bridge" processors clocked at 2.4 GHz. We created mock surveys like the ones in Section 4.5 with an average sightline spacing of $\langle d_\perp \rangle = 2\,h^{-1}$Mpc and adjusted the volume to make problem sizes of $\log_2 N_{\text{pix}} = 15$, 16, 17, 18, and 19. We ran the $\log_2 N_{\text{pix}} = 15$, 16, and 17 problems with OMP_NUM_THREADS set to 1, 2, 4, 6, 8, 12, 16, 18, and 24. We set the number of threads to test the standard

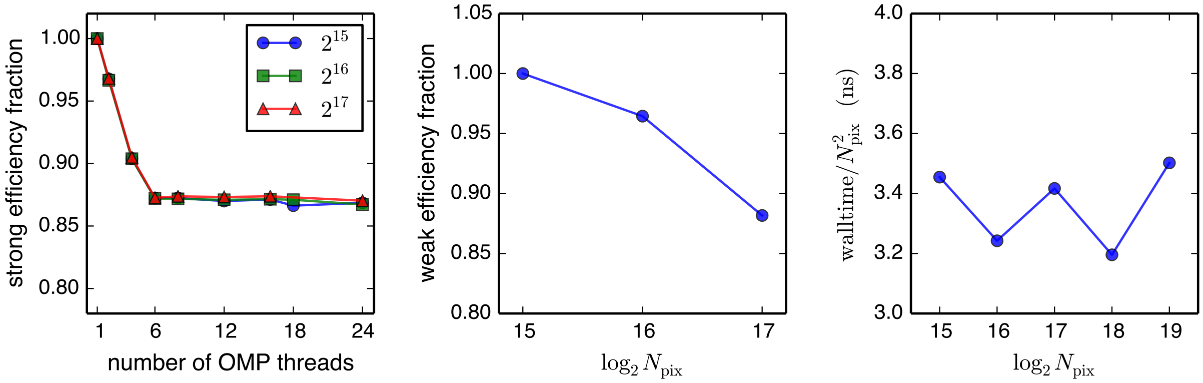

Figure C.3: Three scaling tests for our code. Left: 3 strong scaling tests, where we fix the problem size and increase the number of threads. As long as all threads are given enough work, the speedup should be linear. We show the efficiency $e_{\rm strong} = t_n/(nt_1)$ where n is the number of threads and the t is the walltime. Middle: a weak scaling test, where we fix the work per thread, increasing the problem size and number of threads. We show the efficiency $e_{\rm weak} = t_n/t_1$. Right: The element-wise time for several problem sizes.

powers of two, but also included multiples of 6 to test the Edison NUMA boundaries, which have a significant effect.

In the first panel of Figure C.3, we show the strong scaling efficiency (the walltime compared to what is expected for linear scaling) for these problems. We show the efficiency $e_{\rm strong} = t_n/(nt_1)$ where n is the number of threads and the t is the walltime for each run. The result is independent of the problem size. The relative speedup drops from 1.0 to 0.87 as the number of threads increases from 1 to 6, and then remains the same up to a full node. This suggests there is an increasing (but small) cost for threads to access memory until we hit the first NUMA barrier at 6 threads and is constant after that. In the middle panel, we show a weak scaling problem, increasing the number of threads from 1 to 4 to 16 as the problem size doubles (since the algorithm scales as $N^2_{\rm pix}$). We show the efficiency $e_{\rm weak} = t_n/t_1$. The decrease in efficiency is similar to the strong scaling case, where the 16 thread case is 0.88 of the max efficiency. Finally, in the third panel we show the walltime per $N^2_{\rm pix}$ element from problems all run with 24 threads, doubling in size. The up-down pattern in this panel is not due to random system behavior, but instead the number of PCG iterations. The bottom runs took 12 iterations while the top took 13 due to small differences in the noise realizations. This test confirms the expected $N^2_{\rm pix}$ scaling of the code and also demonstrates how fast the code is. If we consider the number of threads $n = 24$, the number of iterations $i = 12$, the clock speed $s = 2.4\,{\rm ns}^{-1}$, and the element-wise time $t = 3.2\,{\rm ns}$, the number of clock cycles taken per element per iteration is $nsti^{-1} = 15$. This is close to our estimate of 10 operations per lookup and multiply, even though the element-wise time measurement is an overestimate, including other operations like the $\mathbf{S}_{mp}\mathbf{x}$ multiply.

C.4 Error estimation

There are two possibilities for estimating the errors of the map values. First, we can compute the map covariance $\mathbf{M} = \mathbf{S}_{mp}(\mathbf{S}_{pp} + \mathbf{N})^{-1}\mathbf{S}_{pm}$ directly. This option is straightforward, but prohibitively expensive computationally. The inverse and product on the right of the map covariance is now a matrix instead of a vector, meaning we must run a PCG solve for each row of the solution matrix. One could also abandon an iterative method and perform a direct inverse. Either way, the computational complexity of the covariance calculation is a factor of N_{pix} greater than the map calculation. For any interesting problem, this is very expensive indeed.

Instead, we propose using Monte Carlo error estimation. We run n reconstructions on data with random noise realizations (consistent with the noise estimates), and estimate the map variance over the n results. We expect the required number of reconstructions n to be much smaller than N_{pix}, making this method much cheaper. For synthetic data sets, such as in this work, this method also allows us to test the effect of noise in the data and the effect of the sightline sampling independently.

C.5 Alternate smooth map construction

For our protocluster application, we are primarily interested in large-scale fluctuations. A simple way to pick out large-scale fluctuations is to smooth the field on the scale we are interested in, as we did earlier. This acts as a basic matched filter. However, instead of smoothing a high-resolution reconstruction, we could start with a different estimator that picks out large-scale fluctuations. We can think of our signal split into low and high-frequency components $\mathbf{s} = \mathbf{s}_l + \mathbf{s}_h$. The Wiener Filter estimate of the low-frequency signal is $\hat{\mathbf{s}}_l = (\mathbf{S}_{ll} + \mathbf{S}_{hl})(\mathbf{S} + \mathbf{N})^{-1}\mathbf{d}$. We can split the signal with a Gaussian filter $\mathbf{G}$ such that $\mathbf{s}_l = \mathbf{G}\mathbf{s}$ and $\mathbf{s}_h = \mathbf{s} - \mathbf{G}\mathbf{s}$. It follows that $\hat{\mathbf{s}}_l = \mathbf{S}\mathbf{G}(\mathbf{S}+\mathbf{N})^{-1}\mathbf{d}$. Compare this to the expression for a smoothed map, $\mathbf{G}\hat{\mathbf{s}} = \mathbf{G}\mathbf{S}(\mathbf{S} + \mathbf{N})^{-1}\mathbf{d}$. These expressions only differ by the position of the Gaussian filter, but it is an important distinction. In the case of the smoothed map, the filter acts on the map values, whereas in the case of the smooth signal reconstruction, the filter acts on the weighted pixel values. However, for any practical case where the filter scale is larger than the pixel and map spacing, these expressions will be very close to one another, and the distinction is no longer important.