Title

Constraining Reionization with the High-z Lyman-α Forest

Author

Wolfson, Molly Ann

Publication Date

2024

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

Constraining Reionization with the High-z Lyman-α Forest

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Physics

by

Molly Wolfson

Committee in charge:

Professor Joseph Hennawi, Chair
Professor Crystal Martin
Professor Peng Oh

June 2024

The Dissertation of Molly Wolfson is approved.

Professor Crystal Martin

Professor Peng Oh

Professor Joseph Hennawi, Committee Chair

May 2024

Constraining Reionization with the High-z Lyman-α Forest

by

Molly Wolfson

Acknowledgements

This dissertation is the result of six years of my life. It would be impossible to adequately thank everyone who has helped shape me into the person and researcher I am today, but I hope these following words will suffice.

First, I would like to thank my mentors in research. My advisor, Joseph Hennawi, without whom none of this would be possible. He has taught me how to think critically about science and ask questions that have led me to a better understanding of the world. The other members of my committee, Crystal Martin and Peng Oh, have provided invaluable guidance throughout my PhD. The many international scientists who have greatly impacted my work, including Frederick B. Davies, Sarah E. I. Bosman, Jose O norbe, Zarija Luki'c, and Martin G. Haehnelt. And finally, the entirety of the far-reaching ENIGMA group, which has been a fantastic environment in which to mature as a scientist, including Riccardo, Suk Sien, Caitlin, Shane, Elia, Silvia, Timo, Diego, and Linda.

Next, I want to thank the army of people who I have called my friends and family during this journey. My parents, brother, and sister-in-law, who have supported me moving all across the country in pursuit of this dream. They are my safety net that allows me to reach for these stars. Greta and Rachel, with whom I had the honor of living for two years. This dissertation was driven by their support and our shared love of matcha lattes. Craig and Sean, who spent hours in the gym with me. They kept me sane for six years. From playing video games to talking about anything, I always knew I could count on them to be there for me. Jenny, with whom I had the honor of being confused all the time. She is an inspiration and a great friend. Caitlin, with whom I (basically) share a birthday. Her compassion and dedication have greatly enriched my life. Beaks, Sara, Dmitriy, and Jordan. I can't imagine the last year without them, and

I am so grateful for the late nights and crazy plans. The many others who have made Santa Barbara an exciting adventure, including Aaron, Demi, Kaili, Minghan, and Potts. Clara, Colin, Jerry, Maddy, and Shirley, my college roommates. Each and every one of them inspires me every day, and I am so grateful to know that I have such a widespread and strong support system, wherever life takes me. I promise to always look forward to our reunions. Sam, my absolute rock in California, and Asya, my bestie in DC. I am so proud of you both for everything you have achieved in your personal and professional lives since we graduated six years ago. I am excited for all of our adventures in the future. Brett, Jenny, and Rachel, who I grew up with. They have cheered me on for well over a decade. I don't know who I would be without you all, and I hope I never need to find out. Emily and Michael, thank you for visiting me in Santa Barbara and for answering every phone call. To all these people, and many more, I wouldn't be who I am today, capable of doing all of this, without you.

Finally, I need to thank my dog, Pippin, for being the best boy.

Molly Wolfson
https://mollywolfson.github.io/
http://enigma.physics.ucsb.edu/

Department of Physics, UC Santa Barbara
email: mawolfson@ucsb.edu
phone: (518) 728-1896
citizenship: USA

Education

University of California, Santa Barbara (UCSB) Santa Barbara, CA
- Physics M.A. 2021; Physics Ph.D. expected June 2024

The University of Chicago Chicago, IL
- Physics B.A. with honors; Mathematics B.S. 2018

Research Experience

2018–present: **Graduate Student Researcher**, UCSB, Santa Barbara, CA
PI: Professor Joseph Hennawi

2017: **NSF Mathematics REU Participant**, UChicago, Chicago, IL
http://math.uchicago.edu/~may/REU2017/REUPapers/Wolfson.pdf

2017–2018: **Research Assistant**, Enrico Fermi Institute, Chicago, IL
PI: Professor Yau Wah

2016: **DHS Summer Research Intern**, NSTec, Las Vegas, NV
Supervisor: Dr. Eric Wagner

2015-2017: **Research Assistant**, James Franck Institute, Chicago, IL
PI: Professor Stuart A. Rice and Dr. Binhua Lin

Main Publications

1. **Wolfson, M.**, et al. "Measurements of the $z > 5$ Lyman-α forest flux auto-correlation functions from the extended XQR-30 data set" 2024, MNRAS, 531 (3), 3069-3087
2. **Wolfson, M.**, Hennawi, J. F., Davies, F. B., Oñorbe, J. "Forecasting constraints on the high-z IGM thermal state from the Lyman-α forest flux auto-correlation function" 2023, arXiv:2309.05647
3. **Wolfson, M.**, Hennawi, J. F., Davies, F. B., Oñorbe, J. "Forecasting constraints on the mean free path of ionizing photons at $z \geq 5.4$ from the Lyman-α forest flux auto-correlation function" 2023, MNRAS, 521 (3), 4056-4073
4. **Wolfson, M.**, Hennawi, J. F., Davies, F. B., Oñorbe, J., Hiss, H., Lukić, Z. "Improving IGM temperature constraints using wavelet analysis on high-redshift quasars" 2021, MNRAS, 508 (4), 5493-5513
5. **Wolfson, M.**, Liepold, C., Lin, B., and Rice, S. A. "A comment on the position dependent diffusion coefficient representation of structural heterogeneity" 2018, J. Chem. Phys., 148 (19), 194901

Work in Collaboration

1. Zhu, Y., Becker, G. D., Bosman, S. E. I., Cain, C., Keating, L. C., Nasir, F., D'Odorico, V., Bañados, E., Bian, F., Bischetti, M., Bolton, J. S., Chen, H., D'Aloisio, A., Davies, F. B., Davies, R. L., Eilers, A.-C., Fan, X., Gaikwad, P., Greig, B., Haehnelt, M. G., Kulkarni, G., Lai, S., Puchwein, E., Qin, Y., Ryan-Weber, E. V., Satyavolu, S., Spina, B., Walter, F., Wang, F., **Wolfson, M.**, Yang, J., "Damping Wing-Like Features in the Stacked Lyα Forest: Potential Neutral Hydrogen Islands at $z < 6$" 2024, arXiv:2405.12275

2. D’Odorico, V., Bañados, E., Becker, G. D., Bischetti, M., Bosman, S. E. I., Cupani, G., Davies, R., Farina, E. P., Ferrara, A., Feruglio, C., Mazzucchelli, C., Ryan-Weber, E., Schindler, J.-T., Sodini, A., Venemans, B. P., Walter, F., Chen, H., Lai, S., Zhu, Y., Bian, F., Campo, S., Carniani, S., Cristiani, S., Davies, F., Decarli, R., Drake, A., Eilers, A.-C., Fan, X., Gaikwad, P., Gallerani, S., Greig, B., Haehnelt, M. G., Hennawi, J. F., Keating, L., Kulkarni, G., Mesinger, A., Meyer, R. A., Neeleman, M., Onoue, M., Pallottini, A., Qin, Y., Rojas-Ruiz, S., Satyavolu, S., Sebastian, A., Tripodi, R., Wang, F., **Wolfson, M.**, Yang, J., Zanchettin, M. V., “XQR-30: the ultimate XSHOOTER quasar sample at the reionization epoch” 2023, MNRAS, 523 (1), 1399-1420

Awards and Honors

Worster Summer Research Fellowship	2023
Doctoral Student Travel Grant	2023
Mananya Tantiwiwat Fellowship Award	2022
UCSB Department of Physics, Department Service Award	2018-2020
University of Chicago Dean’s List	2014-2018
Enrico Fermi Institute Undergraduate Summer Research Grant	2017
Honorable Mention Poster at the Chicago Area Undergraduate Research Symposium	2017
UCISTEM Summer Research Grant	2015

Research Talks

2024 *Constraining Reionization with the high-z Lyman-α forest*
243th Meeting of the American Astronomical Society, January 7-11 2024, New Orleans, LA

2023 *Constraining Reionization with the high-z Lyman-α forest*
UC Berkeley Cosmology Seminar, October 31, 2023, Berkeley, CA

2023 *Constraining reionization with the Lyman-α forest flux auto-correlation function*
MIT Monday Afternoon Talk, July 10, 2023, Cambridge, MA

2023 *Constraining Reionization with the Lyman-α forest flux auto-correlation function*
Reionization in the Summer, June 26 - 29, 2023, Heidelberg, Germany

2023 *Constraining Reionization with the Lyman-α forest flux auto-correlation function*
UCSB Astro Lunch Seminar, May 31, 2023, Santa Barbara, CA

2023 *The Lyman-α forest flux auto-correlation function as a source of information on the $z > 5$ universe*
Future Cosmology, April 23 - 29, 2023, IESC Cargese, France

2022 *Constraining the mean free path of ionizing photons at $z > 5$ from the Lyman-α forest flux auto-correlation function*
UC Berkeley Cosmology Seminar, October 18, 2022, Berkeley, CA

2022 *Forecasting constraints on the high-z mean free path of ionizing photons from the Lyman-α forest auto-correlation function*
240th Meeting of the American Astronomical Society, June 12-16 2022, Pasadena, CA

2022 *Using the Lyman-α forest auto-correlation function to constrain the mean free path of ionizing photons at $z \geq 5.4$*
Reionization and Cosmic Dawn: Looking Forward to the Past, March 21 - 23, 2022, BCCP Berkeley, CA

Teaching and Supervision

2022–present: **Research Mentor** University of California, Santa Barbara
Supervised Linda Zhenyu Jin (undergrad) build an emulator with machine learning

2018–2019: **Teaching Assistant** UCSB Department of Physics
PHYS 3L (now 20AL), PHYS 4L (now 20BL) - introductory labs for physics majors

2016–2018: **Research Mentor** James Franck Institute
Trained and supervised Linsey Nowack (undergrad) on running diffusion experiments

2016–2018: **Physics Core Tutor** The University of Chicago Harper Library
Covered introductory physics course material and beyond

Synergistic Activities

1. **Collaborations:**
   (i) Member, the XQR-30 team, `https://xqr30.inaf.it/`

2. **Public Talks:**
   (i) “Cosmic Lighthouses: Navigating the Early Universe with Quasars", UCSB Grad Slam, March 12, 2023
   (ii) “The History of the Universe with High-Redshift Quasars", UCSB Lunch & Learn, June 2, 2023

3. **UCSB Service:**
   (i) Mentor, Graduate Scholars Program, 2020 - present
   (ii) Organizer, “Astro Lunch” a UCSB, KITP, and LCO talk series, 2019 - present
   (iii) Member, Women and Gender Minorities in Physics, 2018 - present
   (iv) President, Women and Gender Minorities in Physics, 2019 - 2022
   (v) Mentoring Chair, GradLife, 2019 - 2021
   (vi) Co-Author, APS Bridge Partnership Institution Application, 2020 - 2021
   (vii) LOC Member, “APS Conference for Undergraduate Women in Physics” 2018 - 2019
   (viii) Mentor, Women in Science and Engineer Mentoring Program, 2018 - 2019
   (ix) Finance Co-Chair, “Beyond Academia” industry conference, 2020 - 2021

4. **Invited Panels:**
   (i) “Being a Woman in Physics” UCSB SPS, 2021
   (ii) “Applying to Graduate School and Fellowships” APS CUWiP, 2019
   (iii) “Exploring Undergraduate Research Opportunities” UCSB Dept. of Physics, 2018

Abstract

Constraining Reionization with the High-z Lyman-α Forest

by

Molly Wolfson

Understanding the reionization of the intergalactic medium (IGM) by the first luminous sources remains an important open problem in cosmology. During Reionization ionization fronts propagate through the IGM, heating the reionized gas. This heat injection can be observed over a redshift interval of $\Delta z \sim 1$ due to the long cooling times in the low-density IGM. Simultaneously, the mean free path of ionizing photons (λ_{mfp}) describing the ultraviolet background (UVB) rapidly evolves as bubbles of reionized gas, where the UVB is stronger, merge. Thus, constraining the thermal state of the IGM and the evolution of λ_{mfp} can, in turn, be used to constrain reionization. Transmission in the Lyman-α (Lyα) forest, the ubiquitous Lyα absorption lines produced by residual neutral hydrogen in the IGM along quasar sightlines, offers a powerful tool to investigate these phenomena.

Here I present studies that look into utilizing the $z > 5$ Lyα forest to constrain Reionization via the thermal state of the IGM and λ_{mfp}. First, I use wavelet analysis as an initial clustering statistic and shows improvement on simulated temperature constraints when compared to measurements from the 1D Lyα forest power spectrum. This work also notes the importance of careful accounting of the correlations between datasets to ensure observational constraints are not over-confident. The next two studies use the Lyα forest auto-correlation function at $z \geq 5.4$ as the clustering statistic. The second study demonstrates the potential in constraining λ_{mfp} from simulated data while the third does the same for the temperature. Both of these do careful statistical inference tests to ensure the results are not over-confident and describe a method to guarantee statistical robustness of measurements. The final study makes measurements of the Lyα forest auto-correlation functions at $z > 5$ from the XQR-30 data set. Preliminary comparisons of these measurements to the models from the second study are done, showing agreement with previous measurements of λ_{mfp}. Each of these contribute to our understanding of the high-z Lyα forest and how it can be used to constrain Reionization.

Contents

Chapter 1

Introduction

1.1 A Brief History of the Universe

The Universe began 13.7 billion years ago with the Big Bang, which spawned everything in a hot, dense plasma where light couldn't escape. Since then the Universe has been expanding and cooling. Around 370 000 years later, the primordial plasma condensed into the first neutral atoms allowing light to escape for the first time. This light can be observed today as the cosmic microwave background (CMB). From then on the lack of sources of light caused the Universe to be totally dark. This cosmic 'dark age' lasted until the formation of the first galaxies, stars, and black holes which began to emit their own light. This light then reheated and reionized the neutral hydrogen in the intergalactic medium (IGM), meaning that the neutral hydrogen will lose its electron to become a negative ion. The time period where the neutral hydrogen is reionized is known as Reionization (or the epoch of hydrogen reionization). From here the Universe continued to evolve under the influence of gravity leading to the formation of modern galaxies like the one we live in at present day. An illustration of this timeline of the Universe and the evolution of the IGM can be found in Figure 1.1.

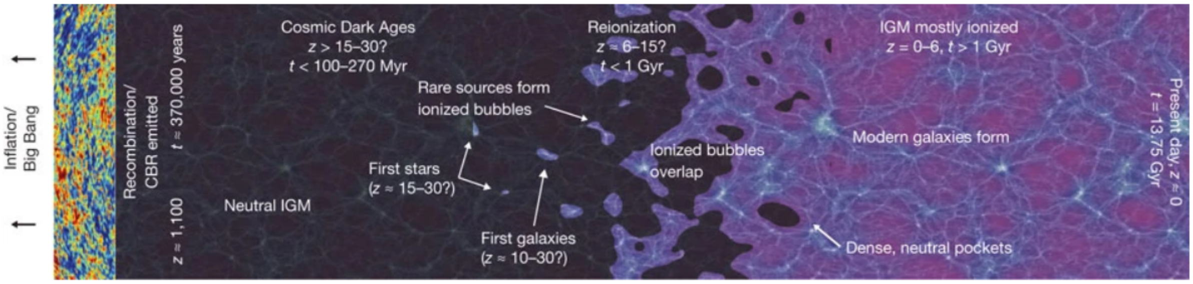

Figure 1.1: Diagram of the timeline of the Universe. This highlights the transition from a neutral IGM on the left to the ionized IGM of today on the right. This highlights several notable events including the Big Bang, Recombination, and Reionization. Reproduced from Robertson et al. (2010) with permission.

Understanding Reionization has been the main objective of my PhD over the past six years and this dissertation is dedicated to describing my efforts. To begin, in Chapter 1.2 I go over some of the definitions and physics used in my field of research. Then in Chapter 1.3 I describe some modern methods of studying reionization and in Chapter 1.4 I outline the topics in the rest of this dissertation.

1.2 Relevant (Astro)physics

1.2.1 Some Cosmology

As stated above, the Universe is expanding, which means that the distances between all galaxies, for example, is getting bigger solely from this expansion. Cosmologists have introduced the scale factor, $a(t)$, to describe this growth. The scale factor is a function of time because the Universe is not expanding at a constant rate, and in fact this rate of expansion depends on the energy density of the Universe, $\rho(t)$. To consider how rapidly the scale factor changes with time, we consider the Hubble rate, $H(t)$, where:

$$H(t) \equiv \frac{da/dt}{a}. \tag{1.1}$$

Note that generally cosmologists refer to the Hubble rate today as H_0. To look at the Hubble rate in terms of the energy density, we consider the Friedmann equation:

$$H^2(t) = \frac{8\pi G}{3}\left[\rho(t) + \frac{\rho_{\rm cr} - \rho_0}{a^2(t)}\right] \tag{1.2}$$

where the critical energy density, $\rho_{\rm cr}$, is:

$$\rho_{\rm cr} \equiv \frac{3H_0^2}{8\pi G}. \tag{1.3}$$

In terms of astronomical observations, since galaxies are getting physically farther apart with the expansion of the Universe they will appear like they are moving away from us here on Earth. This will cause the observed light to shift when compared to the light that was initially emitted. This can also be thought of as the light losing energy against the expansion of the Universe. We define this shift as cosmological redshift, z:

$$1 + z \equiv \frac{\lambda_{\rm obs}}{\lambda_{\rm emit}} = \frac{1}{a}. \tag{1.4}$$

The further away some object is, the greater the speed at which it moves away, which means the redshift will be greater. Thus, this redshift corresponds to the distance between us and a galaxy. And since light travels at a finite speed, c, the further away light is the older it is. If fact, cosmologists will refer to the time of events by its redshift, as can be seen in Figure 1.1. The higher redshift, z, the further back in time. Note that Reionization in Figure 1.1 is labeled as "$z = 6 - 15?$".

1.2.2 The Intergalactic Medium

The IGM is the diffuse gas between galaxies. The fraction of baryons in the IGM at $z \sim 6$ is thought to be $\sim 95\%$ (McQuinn, 2016) where the estimate today is closer to $\sim 50\%$. The IGM is of interest to both cosmologists and astronomers because it can be used to test structure formation and dark matter models, it impacts measurements of the CMB (as this light passes through the IGM before we observe it), it is the environment in which galaxies form and feed from, and more. It also is a unique location to study cosmology because the state of the IGM is largely determined by the cosmological initial conditions and evolution processes which we largely understand. The uncertain astrophysics such as feedback come into play mostly at later redshifts (closer to today) and low densities. Thus the IGM is a relatively clean theoretical location to study cosmological questions. The main physical properties of the IGM that I will discuss are the ionization state and the thermal state.

For the ionization state there are a few processes to consider. Electrons are ionized from the neutral hydrogen in the IGM by photons with energies greater than its ionization potential at the rate, Γ_{HI}:

$$\Gamma_{\mathrm{HI}} = c \int_{\nu_{\mathrm{T}}}^{\infty} d\nu \frac{u_\nu}{h_{\mathrm{P}}\nu} a_\nu^{\mathrm{HI}} \tag{1.5}$$

where a_ν^{HI} is the hydrogen photoelectric cross section, ν_{T} is the threshold (or Lyman limit) frequency required to ionize hydrogen, and u_ν is the specific energy density of the radiation field. These free electrons are then captured by other protons in the IGM with a rate $n_e \alpha_{\mathrm{HII}}(T)$ where α_{HII} is the total rate coefficient for radiative capture summed over recombinations to all energy levels (or Case A radiative recombination coefficient). The total hydrogen number density, n_{H}, can be written in terms of the neutral hydrogen number density, n_{HI}, and the ionized hydrogen number density, n_{HII}, as $n_{\mathrm{H}} = n_{\mathrm{HI}} + n_{\mathrm{HII}}$. The fraction of neutral hydrogen is $x_{\mathrm{HI}} = n_{\mathrm{HI}}/n_{\mathrm{H}}$ and the fraction of ionized hydrogen in $x_{\mathrm{HII}} = n_{\mathrm{HII}}/n_{\mathrm{H}}$. Then I can write down the rate of change in the fraction of neutral hydrogen in terms with a loss from photons and an increase from recombinations as

$$\frac{dx_{\mathrm{HI}}}{dt} = -x_{\mathrm{HI}}\Gamma_{\mathrm{HI}} + x_{\mathrm{HII}} n_e \alpha_{\mathrm{HII}}(T). \tag{1.6}$$

From here it is straightforward to derive the equilibrium fraction of

$$x_{\mathrm{HI}}^{\mathrm{eq}} = \frac{n_e \alpha_{\mathrm{HII}}(T)}{\Gamma_{\mathrm{HI}} + n_e \alpha_{\mathrm{HII}}(T)}. \tag{1.7}$$

For the thermal state, there are several processes to consider that contribute to heating and cooling. First is heating due to excess energy when photoionizing electrons from neutral hydrogen. For cooling the main processes relevant to consider in the high-z IGM are the adiabatic expansion of the Universe (as previously mentioned) and inverse Compton scattering off CMB photons. The equations for the rates here are not as relevant for the thesis so I will omit them here but see Meiksin (2009) for details.

1.2.3 The Lyman-α Forest

To begin, the Lyman series are the transitions as an electron goes from any excited state ($n \geq 2$) to the ground state in hydrogen. The Lyman-α (Lyα) transition is specifically the transition between the first excited state ($n = 2$) and the ground state ($n = 1$). This will happen both in emission (the $n = 2$ electron releases 1216Å light to become $n = 1$) and absorption (the $n = 1$ electron will absorb 1216Å light to become $n = 2$). Thus if 1216Å light encounters neutral hydrogen (such as neutral hydrogen in the IGM) it will be absorbed. Another transition to note is the Lyman-β (Lyβ), which happens between the second excited state ($n = 3$) to the ground state ($n = 1$) and has characteristic wavelength of 1026Å.

For this dissertation, I will be using this Lyα absorption in quasar spectra, in a region known as the Lyα forest. Quasars are supermassive black holes at the centers of galaxies, surrounded by gas. As this gas spirals into the black hole, it releases an extraordinary amount of energy in the form of light. To note later, this process makes quasars some of the brightest objects in the Universe which we can see very far away. Quasars emit light in a characteristic spectra (which includes Lyα emission). Now as this light travels from a quasar through the IGM to us on Earth, the light will redshift due to the expansion of the Universe, as was described in equation (1.4). If at a given location in the IGM there is neutral hydrogen, then the light that has redshifted to 1216Å could be absorbed.

I want to consider the specific region in quasars where we only see Lyα absorption so where the emitted light has 1026Å $< \lambda_{\rm emit} <$ 1216Å (between the rest frame Lyα and Lyβ emission lines). This region will have a collection of Lyα absorption lines due to neutral hydrogen in the IGM. The same logic for Lyα absorption applies to Lyβ absorption as well at higher energies which correspond to shorter wavelengths (and in fact this is true for all Lyman series lines). So emitted light with $\lambda_{\rm emit} <$ 1026Å will contain absorption for both Lyα and Lyβ where the Lyβ absorption will be closer to the quasar than the Lyα absorption in this region. These regions are illustrated in three quasar sightlines shown in Figure 1.2.

The three panels go from lowest z at the top to the highest z in the bottom panel. The top and middle panels roughly look like the quasar continuum emission with varying amounts of Lyα absorption lines. By the bottom panel the Lyα absorption is great enough that the resulting spectra instead looks like transmission spikes.

Through this cosmological redshift, the Lyα forest gives us spatial information about the neutral hydrogen in the IGM in front of quasars. The transmitted flux, F, (that we observe) is defined in terms of the the Lyα optical depth, $\tau_{\rm Ly\alpha}$, where $F = e^{-\tau}$, so the greater the optical depth the less light that makes it through. The optical depth itself is defined as:

$$\tau_{\rm Ly\alpha}(z) = 1.3\Delta_b \left(\frac{x_{\rm HI}}{10^{-5}}\right)\left(\frac{1+z}{4}\right)^{3/2}\left(\frac{dv/dx}{H(z)/(1+z)}\right)^{-1}, \tag{1.8}$$

where $\Delta_b = \rho/\langle\rho\rangle$ is the baryon density in units of the cosmic mean, $x_{\rm HI}$ is the fraction of hydrogen that is neutral, and dv/dx is the line-of-sight velocity gradient. With this definition (from McQuinn, 2016) it is easy to note that the Lyα forest is sensitive to low neutral fractions of $x_{\rm HI} \sim 10^{-5}$ at $z = 3$, which corresponds to a very small number density of $n_{\rm HI} \sim 10^{-10}$ cm^{-3}.

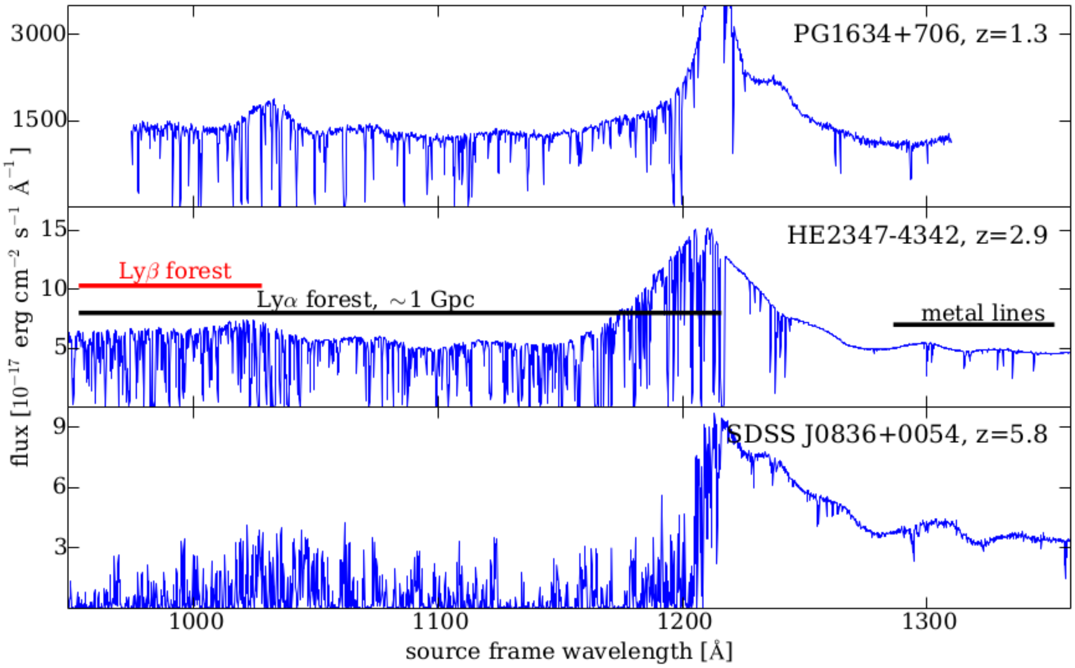

Figure 1.2: Lyα forest spectral region for three quasars chosen to span a large range in redshift. Image credit to McQuinn (2016).

Looking closer at some of the relevant physical parameters that affect the optical depth, we see:

$$\tau_{\mathrm{Ly}\alpha}(z) \propto x_{\mathrm{HI}} n_{\mathrm{H}} \propto \frac{n_{\mathrm{H}}^2 T^{-0.7}}{\Gamma_{\mathrm{HI}}}, \tag{1.9}$$

where T is the temperature of the gas and Γ_{HI} is the photoionization rate (see Rauch, 1998, for more information). From this, the Lyα forest contains additional information on the temperature of the IGM and the photoionization rate of the ultraviolet background (UVB). These parameters can be crucial to constraining Reionization.

1.3 Constraining Reionzation

Understanding the process of Reionization is one of the most important open questions in cosmology. CMB measurements thus far have only been able to constrain the midpoint of Reionization, placing it at $z_{\mathrm{reion}} = 7.7 \pm 0.7$ (Planck Collaboration et al., 2020). On the other hand, measurements of the Lyα optical depth and its scatter can only constrain the end of Reionization, since at earlier times the Lyα transition saturates to total absorption, as can be seen through equation (1.8). These measurements give qualitative evidence that Reionization is not complete until $z \sim 5 - 6$ (Yang et al., 2020; Bosman et al., 2022). However, a detailed timeline of Reionization and its completion is not yet known.

Alternative approaches to understanding Reionization arise from studying the thermal state of the IGM and the evolution of the UVB near the end of Reionization. During Reionization ionization fronts traverse the IGM, heating the reionized gas. This heat injection can be observed over a redshift interval of $\Delta z \sim 1$ due to the long cooling times in the low-density IGM (Boera et al., 2019). Simultaneously, near the end of Reionization the mean and topology of the UVB rapidly evolve as bubbles of reionized gas, where the UVB is stronger, merge (Zhu et al., 2023). Thus constraints on the evolution of the thermal state of the IGM and the UVB near the end of Reionization ($5 \lesssim z \lesssim 6$) can allow for constraints on and a better understanding of Reionization overall. Note that Lyα forest is sensitive to both the thermal state of the IGM and the UVB (through $\Gamma_{\rm HI}$) through equation 1.9.

To date, a lack of high-z, high-resolution measurements of the IGM and shortfalls in current modeling have not allowed statistical measurements of the Lyα forest at $z > 5$ to provide quantitative constraints on Reionization. This leads nicely into my dissertation, where I have worked on advancing the statistical methods and measurements for clustering of the high-z Lyα forest to allow for precise constraints on Reionization.

1.4 Overview of Dissertation

My thesis focuses on constraining Reionization via two separate measurements at high-z: the thermal state of the IGM and the mean free path of ionizing photons, $\lambda_{\rm mfp}$, which describes fluctuations in the UVB. Chapters 2 and 4 focus on applying statistical methods (wavelet analysis and the auto-correlation function of the Lyα forest, respectively) to the Lyα forest from cosmological simulating with the goal of improving the measurements of the thermal state of the IGM at $z \geq 5$. Chapter 3 uses the auto-correlation function from simulated data to investigate constraining λ_{mfp}. These three chapters focus on correct statistical analysis of simulation data such that we could be confident of our results when we apply them to observational data. Finally, Chapter 5 measures the auto-correlation function from observational quasar spectra, using the same statistical methods as in the previous chapters. These are the first measurements of the auto-correlation at $z > 5$. Moreover, they are pioneering measurements of Lyα forest clustering during the epoch of reionization, reaching $z = 6$ and surpassing the previous redshift limit of $z = 5.4$. Each of these chapters combine to push forward our understanding of the $z > 5$ IGM and the information it contains on Reionization.

Chapter 2

Improving IGM temperature constraints using wavelet analysis on high-redshift quasars

This chapter was reproduced from Wolfson et al. (2021) with only minor changes to fit the formatting of this dissertation. I'd like to thank my coauthors, without whom this work would not have been possible: Joseph F. Hennawi, Frederick B. Davies, Jose Oñorbe, Hector Hiss, and Zarija Lukić.

2.1 Introduction

The epoch of reionization, when the first luminous sources reionized the neutral hydrogen in the intergalactic medium (IGM), is one of the most dramatic periods of evolution in the young universe. During this time, ionization fronts impulsively heated reionized gas in the IGM to $\sim 10^4$ K. The exact amount of heat injected into the IGM depends on the proprieties of the luminous sources as well as the timing and duration of reionization

(McQuinn, 2012; Davies et al., 2016; D'Aloisio et al., 2019). After reionization, the IGM expands and cools through the adiabatic expansion of the universe and inverse Compton scattering off CMB photons. The combination of photoionization heating, Compton cooling, and cooling due to the expansion of the universe result in a tight power-law temperature-density relation for most of the IGM gas:

$$T = T_0 \Delta^{\gamma - 1} \tag{2.1}$$

for overdensity $\Delta = \rho/\bar{\rho}$, the mean density of the Universe $\bar{\rho}$, temperature at mean density T_0, and an expected slope γ (Hui & Gnedin, 1997; Puchwein et al., 2015; McQuinn & Upton Sanderbeck, 2016). However, the low-density IGM has long cooling times, so the thermal memory of reionization can persist for hundreds of Myr such that the thermal state of the IGM just after reionization ends contains important information on the state of the universe during reionization (Miralda-Escudé & Rees, 1994; Hui & Gnedin, 1997; Haehnelt & Steinmetz, 1998; Theuns et al., 2002a; Hui & Haiman, 2003; Lidz & Malloy, 2014; Oñorbe et al., 2017a,b). Describing the thermal state of the IGM (T_0 and γ) just after reionization, $z \sim 5 - 6$, is key to understand the evolution of the universe during reionization.

The premier probe of the IGM is Lyα absorption along sightlines to bright quasars at high redshift, known as the Lyα forest (Gunn & Peterson, 1965; Lynds, 1971). The properties of these absorption features are sensitive to the thermal state of the IGM from two effects: Doppler broadening due to thermal motions and Jeans (pressure) smoothing of the underlying baryon distribution. The rate at which pressure forces erase gravitational fluctuations is set by the local sound speed, and at IGM densities the pressure scale sound crossing time is approximately the Hubble time. Therefore, the pressure smoothing scale provides an integrated record of the thermal history of the IGM (Gnedin & Hui, 1998; Kulkarni et al., 2015; Nasir et al., 2016; Oñorbe et al., 2017a,b; Rorai et al., 2017). Both of these effects reduce the small-scale structure of the Lyα forest.

Several statistics have been used to measure the thermal state of the IGM, including the flux probability density (Becker et al., 2007; Bolton et al., 2008; Viel et al., 2009; Calura et al., 2012; Lee et al., 2015), the curvature (Becker et al., 2011; Boera et al., 2014; Gaikwad et al., 2021), the Doppler parameter distribution (Schaye et al., 1999, 2000; Ricotti et al., 2000; Bryan & Machacek, 2000; McDonald et al., 2001; Rudie et al., 2012; Bolton et al., 2010, 2012, 2014; Rorai et al., 2018; Gaikwad et al., 2021), and the joint distribution of the Doppler parameters with the Hydrogen Column Density (Hiss et al., 2018). One of the most commonly used statistics used to measure the structure of the Lyα forest is the 1D flux power spectrum ($P_{\mathrm{F}}(k)$) (Theuns et al., 2000; Zaldarriaga et al., 2001; Yèche et al., 2017; Walther et al., 2018; Boera et al., 2019; Gaikwad et al., 2021). The reduction in small-scale structure in the Lyα forest leads to a cut-off in the power at high k values. However, with measurements of higher-redshift quasars, closer to reionization, the optical depth and its scatter for Lyα photons increase (Fan et al., 2006; Becker et al., 2015), leading to more absorption and Gunn-Peterson troughs in the Lyα forest. Calculating the 1D flux power spectrum at these high redshifts thus mixes high signal-to-noise ratio transmission spikes with noisy absorption troughs, potentially leading to a loss of information.

Wavelet analysis provides an alternative statistical method to measure the structure of the Lyα forest over a range of characteristic scales (Lidz et al., 2010; Garzilli et al., 2012; Gaikwad et al., 2021) (though see also Theuns & Zaroubi (2000); Theuns et al. (2002a); Zaldarriaga (2002); Meiksin (2000)). Wavelets are localized in both frequency and real space, which allows them to encode Fourier information while remaining in configuration space. Therefore, wavelet analysis has the benefit of keeping the absorption troughs distinct from the transmission spikes because it produces a full decomposition of wavelet amplitudes along the spectrum. The ultimate statistic used in wavelet analysis is the full wavelet amplitude probability density function (PDF). The PDF potentially contains more information than the average, which is effectively encoded in the power spectrum. However, these wavelet amplitude PDFs are complicated owing to the large correlations between bins in one wavelet amplitude PDF as well as between different wavelet amplitude PDFs.

Our work builds off and improves upon the previous implementation of wavelet analysis done by Lidz et al. (2010) and Gaikwad et al. (2021). The work done in Lidz et al. (2010) used one of the two characteristic wavelet scales explored to constrain the thermal state of the IGM. Each wavelet scale picks out a frequency in the flux so, to compare to the constraints on the thermal state of the IGM from $P_{\mathrm{F}}(k)$, the number of smoothing scales used in wavelet analysis should be comparable to the number of band powers in $P_{\mathrm{F}}(k)$. Only using one scale will reduce the constraining power of the wavelet amplitude PDFs because it is missing information in other Fourier modes. Lidz et al. (2010) also ignored correlations between the bins in the wavelet amplitude PDFs, potentially significantly affecting the resulting error bars. Gaikwad et al. (2021) used eight wavelet scales in their analysis and included the correlations between the bins within each wavelet amplitude PDF. Their method still ignored the correlations between the bins for wavelet amplitude PDFs of different scales, again potentially effecting the resulting error bars. They also combined their $P_{\mathrm{F}}(k)$ measurements with their wavelet PDF measurements that were calculated from the same data set (along with the Doppler parameter distribution and the curvature statistic), ignoring correlations between all these statistics, to get a more precise measurement.

Our work quantifies the precision of parameter inference using wavelet amplitude PDFs and $P_{\mathrm{F}}(k)$. We show that measuring T_0 from our simple thermal model from the wavelet amplitude PDFs results in a 7% reduction of the 1σ errors when compared to the measurement from $P_{\mathrm{F}}(k)$ on the same mock data set. This confirms the potential that wavelet amplitude PDFs have to improve upon existing constraints on the thermal state of the IGM. Our wavelet analysis method uses more scales than previous works and spans the full range of scales probed by $P_{\mathrm{F}}(k)$. For the first time, we calculate and present the full correlation matrices between the bins of the wavelet amplitude PDFs as well as the cross-correlations between $P_{\mathrm{F}}(k)$ and the wavelet amplitude PDFs. We also combined the wavelet amplitude PDFs with $P_{\mathrm{F}}(k)$ while taking the cross-correlations into account and found that this did not further improve the measurement. Finally, we characterized the effects of ignoring cross-correlations for the wavelet amplitude PDFs and the combination of the two statistics.

In addition to the thermal state of the IGM, the small-scale structure of the Lyα forest is also sensitive to departures from cold dark matter (CDM), including models of warm dark matter (WDM). For WDM, the linear power spectrum is exponentially suppressed when compared to CDM on scales smaller than the free-streaming length of the WDM particle (Narayanan et al., 2000). The mass of the WDM particle, m_{WDM}, can then be constrained by requiring the initial conditions to have sufficient small-scale power to reproduce the properties of the Lyα forest (Viel et al., 2013; Iršič et al., 2017; Garzilli et al., 2017). Wavelet analysis thus also has the potential to improve constraints on the mass of a WDM particle from the small-scale structure in the Lyα forest.

The structure of this paper is as follows. We describe our procedure for generating simulated Lyα forest sightlines in Section 2.2. We then introduce and explore the properties of our wavelet analysis in Section 2.3. Our method for statistical inference is laid out in Section 2.4. Our results comparing the measurements from the wavelet analysis and power spectrum is in Section 2.5. We summarize in Section 2.6.

2.2 Simulating Lyman-alpha Forest Spectra

2.2.1 Hydro Simulations

For this work we use one simulation run that uses the Nyx code. Nyx is a cosmological hydrodynamical simulation code designed for simulating the Lyα forest. For more details on the numerical methods, scaling, and the heating and cooling rates see Almgren et al. (2013) and Lukić et al. (2015). We use a standard ΛCDM cosmological model consistent with the constraints from Planck Collaboration et al. (2020): $\Omega_b = 0.04964$, $\Omega_m = 0.3192$, $\Omega_\Lambda = 0.6808$, $h = 0.67038$, $\sigma_8 = 0.826$, and $n_s = 0.9655$. The simulation we used has a box size of length, $L_{\text{box}} = 20\,\text{Mpc}\,\text{h}^{-1}$ and 1024^3 resolution elements. To simulate reionization, we use the *flash* model from Oñorbe et al. (2019) which reionizes at $z_{\text{reion}} = 7.75$, and uses $\Delta T = 2 \times 10^4$ to parameterize the instantaneous heat injection from reionization. In this framework every cell in the simulation will be ionized at $z = 7.75$ and heated to ΔT, unless the cell was previously ionized by a different process (i.e. collisional ionization). We consider two snapshots from this simulation at $z = 5$ and $z = 6$. We output 10,000 skewers of the Lyα forest from each snapshot to use in our analysis, which is equivalent of a total pathlength of $200\,\text{Gpc}\,\text{h}^{-1}$. The pixel scale of the simulation snapshot is $\Delta v = 2.7\,\text{km}\,\text{s}^{-1}$ at $z = 5$ and is $\Delta v = 2.9\,\text{km}\,\text{s}^{-1}$ at $z = 6$. Since there is a larger dataset available at $z = 5$, we focus our work at this redshift. Figures shown in the main text will be at $z = 5$ unless otherwise specified while figures for $z = 6$ are available in Appendix 2.9.

2.2.2 Thermal Models

In the *flash* reionization model, the majority of the IGM follows the tight temperature-density relation of equation (2.1) after reionization, see Oñorbe et al. (2019) for more details. In order to create simulation Lyα absorption sightlines with different values of T_0, we adopt a semi-numerical approach to 'paint' on the temperature. We do this to each simulation cell using the density output from the simulation and setting the temperature according to equation (2.1) with our desired T_0. This is done for all densities with no cutoff. This is a simplistic model that does not take into account the full evolution of the thermal state of the IGM. However, the purpose of this paper is to present our statistical method and demonstrate its accuracy and precision on simulated data so a simple temperature model for the IGM thermal state is sufficient to achieve these aims. We use $\gamma = 1.35$, which was calculated by fitting the initial simulation snapshot to a power law. Our thermal grid consists of 81 values of T_0 from $\log(T_0) = 3.4$ to $\log(T_0) = 4.4$ with $\Delta \log(T_0) = 0.0125$. The Ly$\alpha$ opacity, $\tau_{\mathrm{Ly}\alpha}$ is related to the temperature via $\tau_{\mathrm{Ly}\alpha} = n_{\mathrm{HI}}\sigma_{\mathrm{Ly}\alpha} \propto T^{-0.7}/\Gamma_{\mathrm{HI}}$, see Rauch (1998). Because UV background photoionization, Γ_{HI}, is sourced by complex galaxy physics, it is not uniquely determined by the simulation. We therefore follow standard practice and adjust each model to have the same mean flux by rescaling τ such that $\langle e^{-\tau} \rangle = \langle F \rangle = 0.16$ at $z = 5$, which is within 1σ of the measurement presented in Boera et al. (2019). At $z = 6$ we use $\langle F \rangle = 0.011$ which is also consistent with recent measurements (Becker et al., 2015; D'Aloisio et al., 2018).

2.2.3 Forward Modeling Real Observations

To mimic realistic observational data from echelle spectrographs, (e.g. from Keck/HIRES, VLT/UVES, and Magellan/MIKE) we forward model a resolution of $R = 30,000$ and a signal to noise ratio per pixel (SNR) of the unabsorbed continuum of 10 (35) at $z = 5$ ($z = 6$). The resolution smooths the flux by a Gaussian filter with FWHM $= 10\,\mathrm{km\,s^{-1}}$ which means our simulations have ~ 4 pixels per FWHM of this resolution filter. For sim-

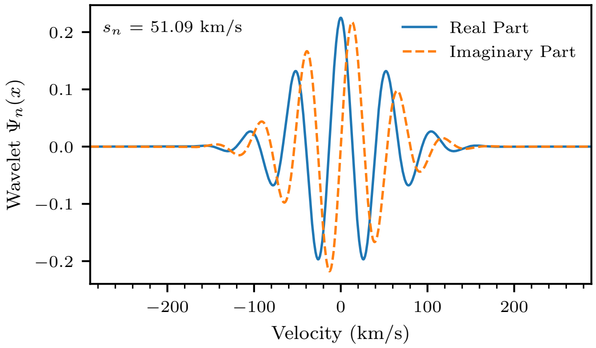

Figure 2.1: A complex Morlet Wavelet filter in real space with $s_n = 51.09\,\mathrm{km\,s^{-1}}$. The solid line shows the real part of the wavelet while the dashed line shows the imaginary part. The width of the oscillations are set by the smoothing scale.

plicity, we add flux-independent noise in the following way. We generate one 10,000 skewer x 1024 length realization of random noise all drawn from a Gaussian with $\sigma_N = 1/\mathrm{SNR}$ and add this noise realization to every temperature model. An example skewer of our initial and forward-modeled data is shown in the top panels of Figures 2.2 and 2.4 respectively. Using the same noise realizations over the different models ensures that different noise realizations will not adversely affect the inference on the T_0 for mock data.

We assume a fiducial data set size of 8 quasar spectra at both $z = 5$ and $z = 6$ that probe a redshift interval of $\Delta z = 0.2$ per quasar for a total pathlength of $\Delta z = 1.6$ (equivalent to 29 skewers). In the discussion going forward, each mock data set consists of a random selection of 29 skewers without replacement.

2.3 Wavelet Analysis

2.3.1 Formalism

Wavelets are localized in frequency and configuration space which allows wavelet amplitudes provide a breakdown of Fourier information at all locations along a quasar sightline. Following Lidz et al. (2010), we calculate wavelet amplitudes from a “complex

Morlet wavelet", which is shown in Figure 2.1 and has the functional form:

$$\Psi_n(x) = A \exp(ik_0 x) \exp\left[-\frac{x^2}{2s_n^2}\right]. \tag{2.2}$$

The normalization, A, is set by requiring that $|\Psi_n(k)| = 1$. With this normalization, the Fourier transform of a complex Morlet wavelet is

$$\Psi_n(k) = \pi^{-1/4}\sqrt{\frac{s_n}{\Delta u}} \exp\left[-\frac{(k-k_0)^2 s_n^2}{2}\right]. \tag{2.3}$$

This is a Gaussian in configuration space centered on k_0 with width $\sigma_k = \sqrt{2}/s_n$. We also require that $k_0 s_n = 6$ to ensure these filters have a close to zero mean.

To begin the analysis on our simulated spectra, we first compute the flux contrast of the Lyα forest, δ_F:

$$\delta_F = \frac{F - \bar{F}}{\bar{F}}. \tag{2.4}$$

Then we convolve this flux contrast field with a wavelet filter of smoothing scale s_n resulting in a filtered spectrum, a_n:

$$a_n(x) = \int dx' \Psi_n(x - x') \delta_F(x') \tag{2.5}$$

The filtered spectrum is a complex number, the modulus of which is called the "wavelet amplitude" $A_n(x) = |a_n(x)|^2$. We define the power spectrum as

$$\langle \delta_F(k) \delta_F(k') \rangle = 2\pi P_F(k) \delta_D(k - k') \tag{2.6}$$

where δ_D is the Dirac Delta function. With this definition of the power, the average wavelet amplitude is

$$\langle A_n(x) \rangle = \int_{-\infty}^{\infty} dk' 2\pi [\Psi_n(k')]^2 P_F(k'). \tag{2.7}$$

In words, this means that the average wavelet amplitude is the power spectrum averaged over a Gaussian centered on wave-number $k_0 = 6/s_n$ with standard deviation $\sqrt{2}/s_n$. Therefore, this average wavelet amplitude is effectively a band-power.

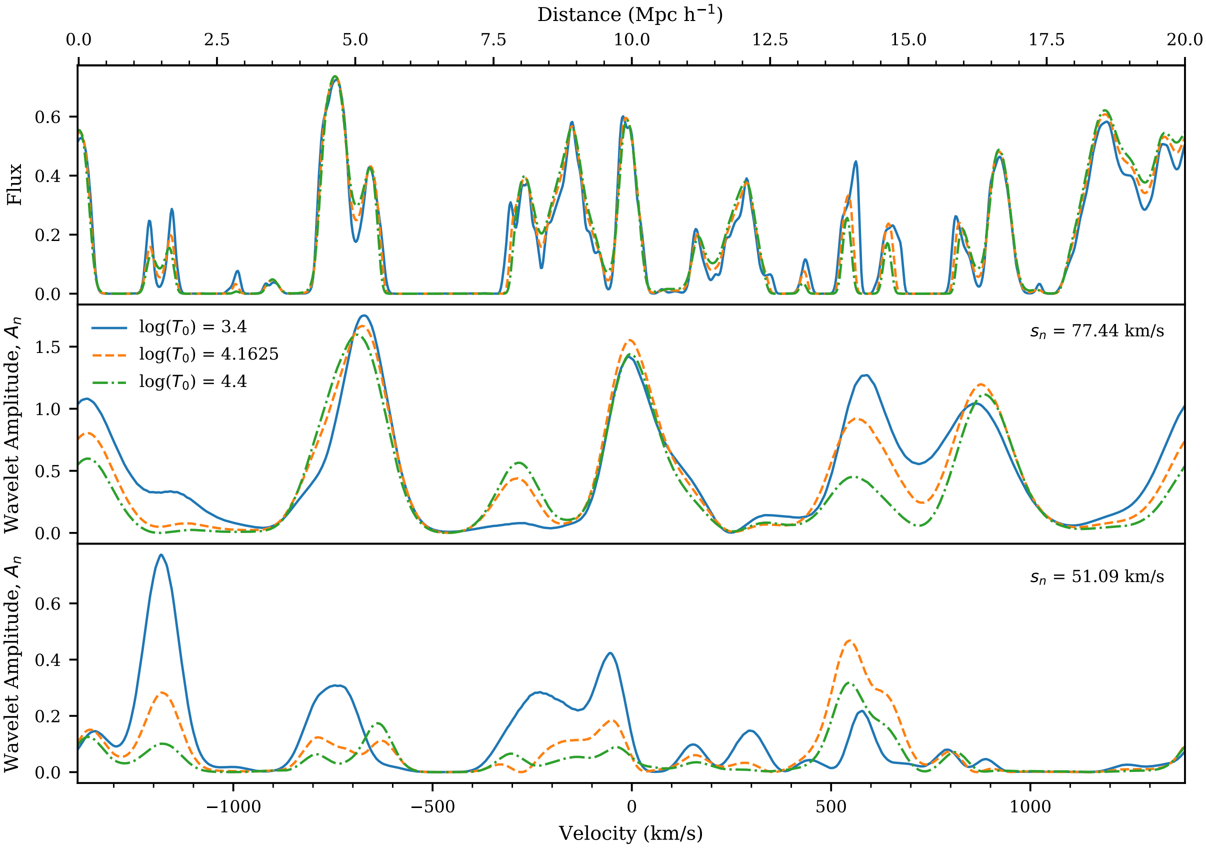

Figure 2.2: The top panel shows the flux from one simulation skewer at $z = 5$ for the three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). The middle panel shows the wavelet amplitude spectra for $s_n = 77.44\,\mathrm{km\,s^{-1}}$ and the bottom show the wavelet amplitude spectra for $s_n = 51.09\,\mathrm{km\,s^{-1}}$, both with the same T_0 values as the top panel. This shows that the largest values of wavelet amplitudes correspond to peaks in the flux that are roughly the same width as the oscillations set by the wavelet smoothing s_n scale.

Two wavelet amplitude spectra for an ideal simulated skewer at $z = 5$ are shown in the bottom two panels of Figure 2.2. For illustrative purposes in this section, we will mainly show wavelet amplitudes for $s_n = 51.09\,\mathrm{km\,s^{-1}}$ though Figure 2.2 also shows $s_n = 77.44\,\mathrm{km\,s^{-1}}$ for a comparison. Ultimately in our analysis at $z = 5$, we will use fifteen logarithmically spaced values of $2200\,\mathrm{km\,s^{-1}} > s_n > 5\,\mathrm{km\,s^{-1}}$ as described in Section 2.4.2. For $z = 6$ we still use fifteen logarithmically spaced values of s_n with slightly shifted values due to the redshift dependence of the simulation resolution and box size.

The purpose of Figure 2.2 is to show the relationship between the flux and wavelet amplitudes for different smoothing scales. The top panel shows the flux for the three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). The middle panel shows the wavelet amplitude spectra for $s_n = 77.44\,\mathrm{km\,s^{-1}}$ and the bottom show the wavelet amplitude spectra for $s_n = 51.09\,\mathrm{km\,s^{-1}}$, both with the same values of T_0 as the top panel. The smoothing scale sets the size of the features in the flux that are picked out, when the smoothing scale and the feature size in the flux align the resulting wavelet amplitude is greater. The middle panel has a greater value of s_n than the bottom panel, so it is going to pick out wider features in the flux. Consider the peak in the flux at $\sim 550\,\mathrm{km\,s^{-1}}$, which is smoother (and smaller) for $\log(T_0) = 4.1625$ (orange) flux than for $\log(T_0) = 3.4$ (blue). The corresponding wavelet amplitudes in the middle panel are greatest for $\log(T_0) = 3.4$ (blue) while the bottom panel are greatest for $\log(T_0) = 4.1625$ (orange), showing that the smaller feature in the flux agreed better with the smaller smoothing scale, as expected. The flux at $\log(T_0) = 4.4$ (green) is even smoother than the flux at $\log(T_0) = 4.1625$ (orange) but it does not have greater wavelet amplitude values in the bottom panel, this is because this peak corresponds to an ever smaller smoothing scale.

Figure 2.2 shows that the largest values of wavelet amplitudes correspond to peaks in the flux that are roughly the same width as the oscillations set by the wavelet smoothing s_n scale (an example of these oscillations can be seen in Figure 2.1). There can be offsets between features in flux and the corresponding features in the wavelet amplitude spectra, since the wavelets pick out features with the specific width set by s_n and, at larger scales, the wavelet will combine multiple features in the flux spectrum. This figure also demonstrates how wavelet analysis presents Fourier information in configuration space since the different wavelet amplitudes values convey frequency information along the quasar sightline.

In order to compare the spatial correlations between different values of s_n, consider Figure 2.3. The top panel of this figure shows a color plot of the wavelet amplitudes for different values of s_n along one line of sight; this is known as a “periodogram”. The bottom panel of the plot is the flux used the calculate the wavelet amplitudes, which is the same as in Figure 2.2 for $\log(T_0) = 4.1625$. The large trough in the flux at $-500\,\mathrm{km\,s^{-1}}$ is seen in the wavelet amplitudes for scales up to $s_n \sim 40\,\mathrm{km\,s^{-1}}$. The other troughs in the flux, such as the one near $750\,\mathrm{km\,s^{-1}}$, are also seen in the wavelet amplitudes across multiple smoothing scales, most prominently at the smaller values of s_n. The overall decline in the average wavelet amplitude value for smaller values of s_n follows from the cutoff in the power spectrum, as is expected from equation (2.7).

As discussed in Section 2.2.3, we forward modeled our simulation skewers to mimic real data by including the effects of the resolution and noise. We illustrate the change in the flux as well as the wavelet amplitudes for one example skewer with $\log(T_0) = 4.1625$ in Figure 2.4. Note that the “clean” flux and wavelet amplitudes in this figure matches the model in Figure 2.2 from the same temperature model. From the Figure, we see that adding noise to the flux is able to shift features, add additional features, and change the amplitude of features in the wavelet amplitude spectrum. These effects are more prominent on smaller scales, as seen in the large differences between the models in the bottom panel of Figure 2.4, since the noise power becomes comparable or greater than the flux power at these scales. Overall, this panel has greater values and more high-valued wavelet amplitudes for the forward modeled skewers than those without noise.

2.3.2 Wavelet Probability Density Function

As illustrated by Figures 2.2, 2.3, and 2.4 wavelet analysis converts the flux of the Lyα forest into wavelet amplitude spectra parameterized by s_n. The average value of

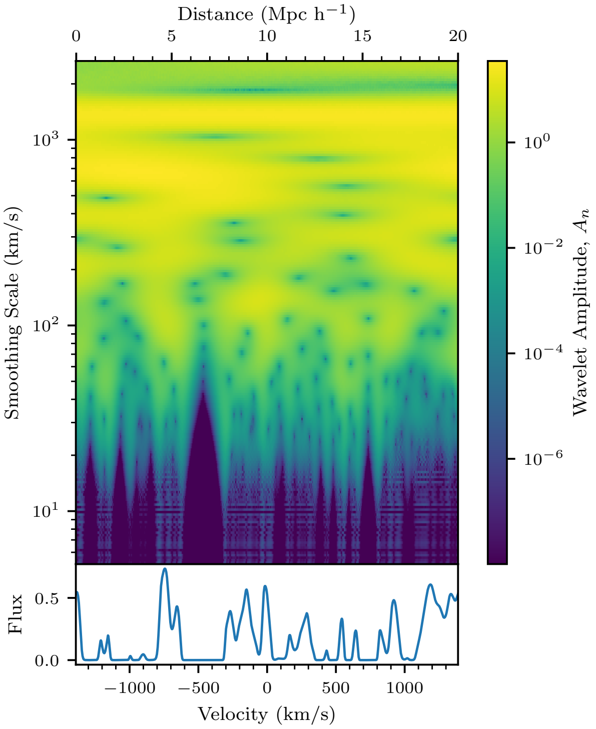

Figure 2.3: The bottom panel shows the flux from one simulation skewer with $\log(T_0) = 4.1625$. The top panel shows a periodogram of the wavelet amplitudes along the skewer for different smoothing scales. The different values of s_n will set different widths of oscillations that they pick out from the spectrum. This plot compares the location of the peaks and troughs in the wavelet amplitudes for different smoothing scales. It shows correlations between troughs at different values of s_n, for example the trough in the flux at $-500\,\mathrm{km\,s^{-1}}$ can clearly be seen in the wavelet amplitudes for scales up to $s_n \sim 40\,\mathrm{km\,s^{-1}}$. Note that the minimum wavelet amplitude shown on the plot is fixed at 10^{-8} for visual purposes.

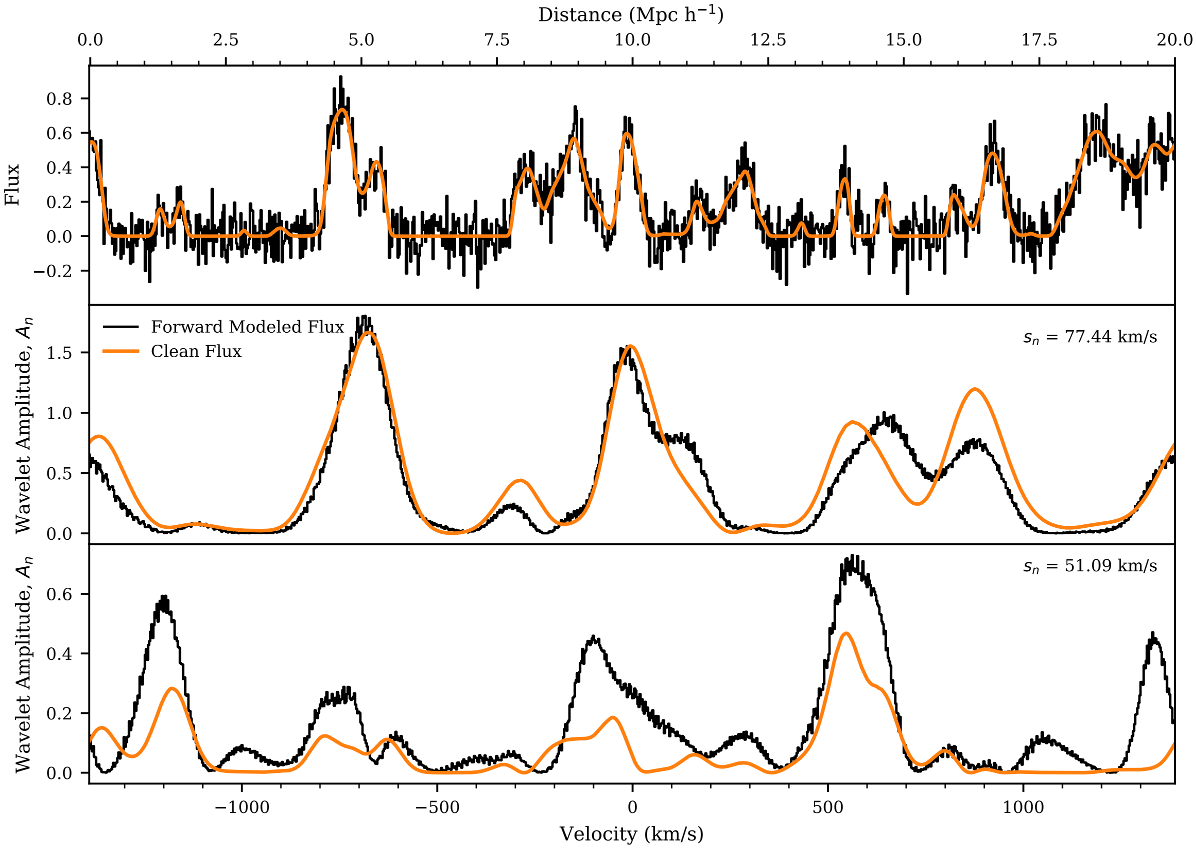

Figure 2.4: The top panel shows the flux for the $\log(T_0) = 4.1625$ model from the simulation (orange line) as well as forward modeled with resolution and noise (black histogram). The middle panel shows the wavelet amplitude spectra for $s_n = 77.44\,\mathrm{km\,s^{-1}}$ and the bottom show the wavelet amplitude spectra for $s_n = 51.09\,\mathrm{km\,s^{-1}}$, both with the same T_0 value as the top panel. The simulation skewer is the same as that shown in Figure 2.2. This shows the effect noise has on the flux and the resulting wavelet amplitudes for one skewer.

the wavelet amplitude spectra corresponds to $P_{\mathrm{F}}(k)$ via equation (2.7). The statistic we measure in our analysis is the wavelet amplitude probability density function (PDF), since this contains information on the full distribution of the wavelet amplitude values, rather than only the average.

PDFs for $s_n = 51.09\,\mathrm{km\,s^{-1}}$ are shown in Figure 2.5 for three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). The top panel shows the PDFs calculated from the ideal simulation with clean flux. The bottom panel shows the PDFs after forward-modeling the simulation output with resolution and noise to mimic real data, as discussed in Section 2.2.3. In both the top and the bottom panel, the black dotted line shows the same PDF for pure noise draws with SNR= 10 and our pixel resolution.

In the top panel, the ideal PDFs are skewed to the left, with lower IGM temperatures corresponding to a higher mean value, as is expected from $P_{\mathrm{F}}(k)$ and equation (2.7). The main effect of forward-modeling is the shift of the PDF from small values to larger values as was also seen in the bottom panel of Figure 2.4. This causes the suppression of wavelet amplitude values below $\sim 10^{-3}$. Initially the $\log(T_0) = 4.4$ (green) PDF had the largest tail below 10^{-3}, so the shift from small values to large values causes this model to change most dramatically from the top to the bottom panel. The PDF both shifts to the right and greatly increases the value of the PDF at the peak. The PDFs on the bottom panel are much more similar to the PDF for pure noise than in the top panel, which shows how the noise PDF is able to dominate over the signal. As the smoothing scale decreases and the overall PDF values decrease with it, as is inferred from Figure 2.3 and equation (2.7), the PDFs will become more dominated by the noise contribution.

Figure 2.5 demonstrates the ability of the wavelet amplitude PDF to differentiate T_0 models both with and without forward modeling. This confirms that the wavelet amplitude PDFs are promising statistics to measure the thermal state of the IGM. In

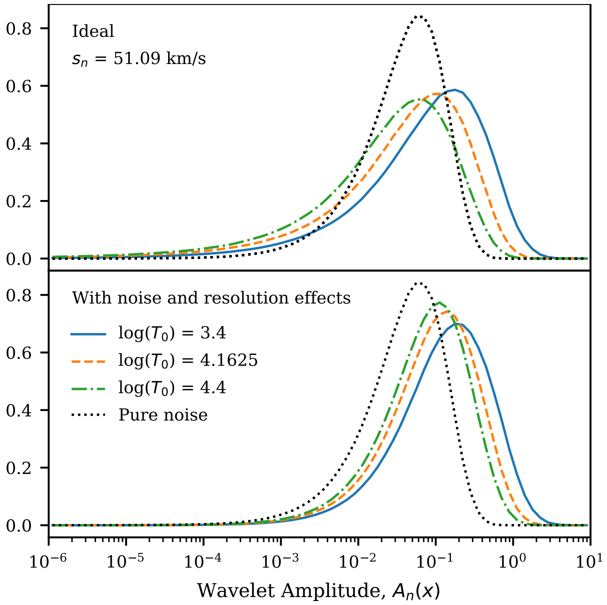

Figure 2.5: This figure shows the PDFs for $s_n = 51.09\,\mathrm{km\,s^{-1}}$ for three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). The top panel shows the PDFs calculated from 10,000 ideal simulation skewers (equivalent to a pathlength of $200\,\mathrm{Gpc\,h^{-1}}$). The bottom panel shows the PDFs after forward-modeling with resolution and noise to mimic real data. In both the top and the bottom panel, the black dotted line shows the same PDF for pure noise draws with SNR= 10 and our pixel resolution. The difference in the mean values of these PDFs in each panel is expected from $P_\mathrm{F}(k)$ and equation (2.7). The main effect of forward-modeling is the suppression of all wavelet amplitude values below $\sim 10^{-3}$, where the data is beginning to be dominated by the noise.

addition, it illustrates how the PDF quantifies the full distribution of wavelet amplitudes for multiple sightlines, rather than the values along one sightline or the average value which is encoded in the the power spectrum.

2.4 Statistical Methods

The goal of this paper is to calculate the statistical precision with which a realistic quasar data set can constrain the parameters governing the small-scale structure of the IGM, here limited to T_0, using wavelet analysis, specifically wavelet amplitude PDFs. The precision from this method can then be directly compared to the canonical approach using $P_\mathrm{F}(k)$. We will also consider the precision achieved when combining the wavelet amplitude PDFs and power spectrum as has recently been attempted in the literature (Gaikwad et al., 2021).

To calculate the statistical precision, we will use Bayes' Theorem:

$$P(T_0|\text{data}) = \frac{P(\text{data}|T_0)P(T_0)}{P(\text{data})}. \tag{2.8}$$

Here the "data" vector depends on the statistical method for which we are calculating the precision. For the power spectrum, the "data" are the band-powers comprising $P_\text{F}(k)$. For the wavelet analysis, we will use multiple values of s_n and thus have multiple wavelet amplitude PDFs we must consider. In this case, the "data" will be the wavelet amplitude PDFs concatenated one after the other from largest to smallest s_n (which corresponds to smallest to largest k). Finally, when combining the wavelet and power spectrum analysis, the "data" vector will be the concatenated PDFs vector from the wavelet case with $P_\text{F}(k)$ added onto the end of it.

We assume a flat prior, $P(T_0)$, over the range of T_0 values we have simulation data for and will normalize the posterior, $P(T_0|\text{data})$, to unity so we don't need to explicitly calculate $P(\text{data})$. In order to calculate the likelihood, $P(\text{data}|T_0) = \mathcal{L}$, we assume a multivariate Gaussian distribution. This likelihood has the form:

$$\mathcal{L} = \frac{1}{\sqrt{\det(\Sigma)(2\pi)^n}} \exp\left(-\frac{1}{2}(\text{data} - \text{model})^\text{T}\Sigma^{-1}(\text{data} - \text{model})\right) \tag{2.9}$$

where $\Sigma = \Sigma(T_0)$ is the model dependent covariance matrix, n is the number of points in the data vector. Both the data and model vectors depend on the statistic we are using and will be discussed in their respective sections. The choice of a multivariate Gaussian distribution for the likelihood has been used in previous wavelet studies (Lidz et al., 2010; Gaikwad et al., 2021) as well as for studies using the Lyα forest flux PDF (Lidz et al., 2006; Eilers et al., 2017). The base assumption is that each band power for $P_F(k)$ or each bin of the wavelet amplitude PDFs are Gaussian distributed. We show that this assumption is valid for our data in Appendix 2.7.

In our analysis, we estimate the covariance matrix from mock draws of the data by

$$\Sigma(T_0) = \frac{1}{N_{\text{mocks}}} \sum_{i=1}^{N_{\text{mocks}}} (\text{mock}_i - \text{model})(\text{mock}_i - \text{model})^{\text{T}} \tag{2.10}$$

where N_{mocks} is the number of forward-modeled mock draws used. This method estimates a model dependent covariance, not the covariance of the data itself, since we are using many draws in our calculation. For the power spectrum calculation we use $N_{\text{mocks}} = 5,000$. We increase the number of mocks to $N_{\text{mocks}} = 1,000,000$ for the wavelet amplitude PDFs and the combination of the power spectrum and the wavelet amplitude PDFs, since these matrices are much larger with more values close to zero. Note that mocks are a random combination of 29 skewers without replacement. In theory, there are $(10,000!)/(29! \times 9,971!) \approx 10^{85}$ unique sets of 29 skewers from 10,000 skewers. This means that mock data sets will be correlated since they will contain skewers that are also in other mock data sets. However, we do not approach the the total possible number of combinations for these calculations and expect this effect to be negligible.

To visualize the covariance matrix for each method, we define the correlation matrix, C. The correlation matrix is the covariance matrix with the diagonal normalized to 1. This is done to the ith, jth element by

$$C_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}. \tag{2.11}$$

2.4.1 Power Spectrum Likelihood

The resolution modifies $P_{\text{F}}(k)$ in a known way that can be corrected via the Fourier transform of the Gaussian resolution filter. There is also a white noise contribution to $P_{\text{F}}(k)$ due to the spectral noise which can be subtracted off. Therefore, the well known estimator for the true power is:

$$P_{\text{true}}(k) = \left\langle \frac{P_{\text{raw}}(k) - P_{\text{noise}}(k)}{W_R^2(k, \sigma_R, \Delta v)} \right\rangle \tag{2.12}$$

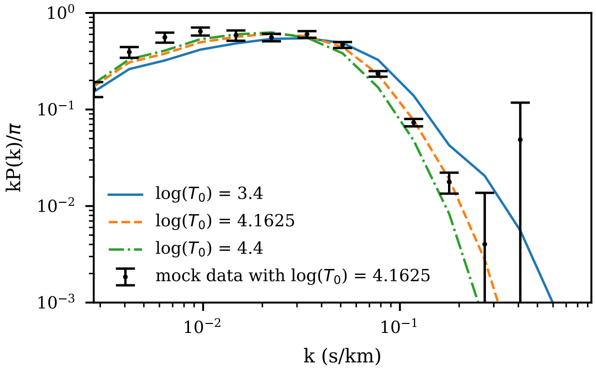

Figure 2.6: The power spectrum measurement, $P_{\rm F}(k)$, for one mock data set with $\log(T_0) = 4.1625$ (black points). The 1σ error bars are calculated from the square root of the diagonal of the covariance matrix. Also shown are model values of the power spectra for three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green).

where $W_R(k, \sigma_R, \Delta v)$ is the Window function

$$W_R(k, \sigma_R, \Delta v) = \exp\left(-\frac{1}{2}(k\sigma_R)^2\right)\frac{\sin(k\Delta v/2)}{(k\Delta v/2)}. \tag{2.13}$$

We have Gaussian white noise with SNR $=10$ added to the flux contrast, adding an extra factor of $\bar{F}$. The noise power is flat and has a value of

$$P_{\rm noise}(k) = \Delta v\left(\frac{1}{{\rm SNR}\cdot\bar{F}}\right)^2 \tag{2.14}$$

where the factor of Δv is our velocity pixel grid spacing which is $\Delta v = 2.7\,{\rm km\,s^{-1}}$ at $z = 5$.

For $R = 30,000$, $\exp\left(-\frac{1}{2}(k\sigma_R)^2\right) < 0.24$ when $k \geq 0.4$. This implies that $W_R^2(k, \sigma_R, \Delta v) <$ .06 when $k \geq 0.4$, so correcting these band-powers by this window function means dividing a noisy quantity, $P_F(k)$, by a very small number. When correcting by the window function, these band-powers blow up and the model covariance matrices we calculate via equation (2.10) are singular and ill-posed for inversion. We therefore choose to not correct by $W_R(k)$ in the "model" and "mock" data to ensure well-behaved covariance matrices. However, for visualization purposes we always show the resolution corrected power of equation (2.12) in the figures. The "model" is the power calculated from 10,000 flux skewers forward modeled with the resolution but not the noise, since there is no need to add additional noise when computing the mean. 10,000 skewers is equivalent to a total pathlength of 200 $\mathrm{Gpc\,h^{-1}}$. We calculate the "mock" data by computing the average $P_{\mathrm{F}}(k)$ for 29 fully forward modeled skewers and then subtracting off the noise power, equation (2.14). This data set size is equivalent to an 8 quasar data set as discussed in Section 2.2.3.

To choose the k values for our mock, we used 15 logarithmic band-powers spanning from $2\pi/l_{\mathrm{skewer}} = 0.0023\,\mathrm{s\,km^{-1}}$ to $\pi/\Delta v = 1.2\,\mathrm{s\,km^{-1}}$ at $z = 5$. The centers of the band-powers are listed in the first column of Table 2.1. We chose 15 band-powers in order to fully sample the shape of the power spectra while ensuring the low k (large scales) band-powers were populated by the discrete Fourier transform of the data. Figure 2.6 shows the power spectrum measurement, $P_{\mathrm{F}}(k)$, for one mock data set with $\log(T_0) = 4.1625$ (black points) at $z = 6$. An equivalent figure for $z = 6$ can be found in Appendix 2.20. The 1σ error bars are calculated from the square root of the diagonal of the covariance matrix. Also shown are three models of the power spectra for three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). This mock data set visually seems to best agree with the model for $\log(T_0) = 4.1625$ (orange) for $k > 0.5$, which is the true T_0 of the model.

The model correlation matrix (see equation (2.11)) for the power spectrum at $\log(T_0) = 4.1625$ is shown in Figure 2.7. There are positive correlations (red) between the band-powers where $4 \times 10^{-2}\mathrm{s\,km^{-1}} \lesssim k \lesssim 0.2\,\mathrm{s\,km^{-1}}$. The correlations between band-powers with $4 \times 10^{-2}\mathrm{s\,km^{-1}} \lesssim k \lesssim 0.2\,\mathrm{s\,km^{-1}}$ and $k < 4 \times 10^{-2}\mathrm{s\,km^{-1}}$ the correlations are negative (blue). This behavior arises from the underlying spatial correlations of the Lyα forest and is consistent with what has been seen for real data (Walther et al., 2018; Boera et al., 2019). At values of $k > 0.2\mathrm{s\,km^{-1}}$ (the smallest scales) noise dominates over the power spectrum, since the Lyα forest power spectrum exhibits a thermal cut-off at high-

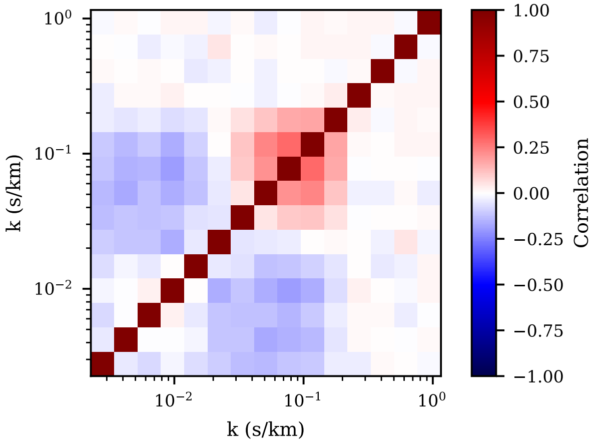

Figure 2.7: The correlation matrix for the power spectrum at $\log(T_0) = 4.1625$. The positive (red) and negative (blue) correlations on scales $k < 0.2\,\mathrm{s\,km^{-1}}$ arise from underlying spatial correlations of the Lyα forest. The very weak correlations seen in the regions where $k > 0.2\,\mathrm{s\,km^{-1}}$ are due to uncorrelated random Gaussian noise which dominates the signal on small-scales (high k).

k, whereas the noise power spectrum is flat. Random Gaussian noise is uncorrelated, making it hard to recover the signal from the Lyα forest and results in the very weak correlations shown in the correlation matrix in the regions where $k > 0.2\,\mathrm{s\,km^{-1}}$. We looked into the correlation matrix for SNR = 50 and SNR = 100 and found the values of the correlation matrix in the column above $k = 0.178\,\mathrm{s\,km^{-1}}$ were stronger. This agrees with our interpretation of the the weak correlations in Figure 2.7 since with higher SNR the noise power is smaller and will not dominate until higher k.

2.4.2 Wavelet Amplitude PDF Likelihood

In previous work, Lidz et al. (2010) measured the thermal state of the IGM with wavelets by assuming a Gaussian likelihood and ignoring correlations between PDF bins as well as between PDFs from different smoothing scales. Gaikwad et al. (2021) measured the thermal state with wavelets assuming a Gaussian likelihood including correlations between bins of the same PDF but not between the PDFs for different smoothing scales.

Table 2.1: The first column contains the band-powers for the power spectrum. Next are the corresponding smoothing scales (s_n) used to calculate the wavelet amplitudes. The last two columns contain the minimum and maximum values used for the PDF estimation for each smoothing scale. These were chosen as the 1.5th and 98.5th percentiles of the data for the whole thermal grid at these smoothing scales.

k ($\mathrm{s\,km^{-1}}$)	$s_n = 6/k$ ($\mathrm{km\,s^{-1}}$)	min $\log(A_n)$	max $\log(A_n)$
0.00278	2157.35	-0.671	1.683
0.00422	1423.32	-0.499	1.826
0.00639	939.04	-0.789	1.643
0.00968	619.54	-0.876	1.574
0.0147	408.74	-1.026	1.474
0.0222	269.67	-1.241	1.344
0.0337	177.91	-1.508	1.173
0.0511	117.38	-1.863	0.931
0.0775	77.44	-2.256	0.564
0.117	51.09	-2.651	0.039
0.178	33.71	-2.943	-0.430
0.270	22.24	-3.032	-0.585
0.409	14.67	-3.073	-0.627
0.620	9.68	-3.129	-0.688
0.939	6.39	-3.280	-0.830

Here, we improve upon these previous work and present the likelihood calculation taking into consideration all correlations, both between PDF bins and between PDFs of different smoothing scales.

For wavelet amplitude PDFs, there is no analytic way to correct for the window function and subtract off the full noise PDF. Instead, we choose to use 10,000 forward-modeled skewers (with resolution and noise) to calculate the "model" wavelet amplitude PDFs, which is equivalent to calculating the wavelet amplitude PDFs for a total path-length of $200\,\mathrm{Gpc\,h^{-1}}$. The "mock" data is calculated from the same forward-modeled skewers as the "model", though "mock" data is the average of 29 skewers (equivalent to an 8 quasar data set, see Section 2.2.3).

As was mentioned in Section 2.3.1, we use fifteen values of s_n to get fifteen wavelet amplitude PDFs. These fifteen s_n correspond to the centers of the power spectrum band-powers, k, that were discussed in Section 2.4.1 and are listed in the second column of Table 2.1. Qualitatively, this should ensure that the wavelet amplitude PDFs contain at least as much information as the power spectrum due to equation (2.7), allowing us to compare the resulting precision on an equal footing. To estimate the wavelet amplitude PDFs for each smoothing scale, we calculate histograms. This introduces three histogram parameters into our analysis: the maximum wavelet amplitude considered, the minimum wavelet amplitude considered, and the number of bins in the histogram.

When selecting the minimum and maximum wavelet amplitude considered for our PDF estimation, we want to ensure that all bins will be populated for the whole thermal grid so that the covariance matrix is well posed for inverting. We also want to make sure the maximum and minimum values span a large enough range to capture the most significant differences in the shape of the PDF. For these reasons, we chose the maximum and minimum values for the PDFs by calculating the 1.5th and 98.5th percentile of the "model" wavelet amplitudes calculated for every thermal model in our grid. The maximum and minimum values of the wavelet amplitudes considered for each smoothing scale s_n are listed in Table 2.1. We also need to select a number of bins that will sufficiently sample the shape of the PDF without making the data vector too long and the covariance matrix ill-suited for inversion. We found that 10 bins was a reasonable choice to achieve these aims.

Figure 2.8 shows the PDFs from one mock data set for each s_n with $\log(T_0) = 4.1265$ (black points) and $z = 5$. An equivalent figure for $z = 6$ can be found in Appendix 2.20. The 1σ error bars are calculated from the square root of the diagonal of the covariance matrix. Each panel also shows the "model" values of the PDFs for three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). This figure qualitatively illustrates the ability of the wavelet PDF to differentiate between different T_0 values, which we formally quantify with Bayesian inference as discussed in Section 2.4. The "model" PDFs for $s_n < 22.24\,\mathrm{km\,s^{-1}}$ all overlap because the noise dominates the signal on these scales and all three PDFs are equivalent to the pure noise PDF.

In order to understand the correlations present between the bins of a single wavelet PDF, we first calculate the model covariance matrix for $s_n = 51.09\,\mathrm{km\,s^{-1}}$ and $\log(T_0) = 4.1265$ and then plot the correlation matrix in Figure 2.9. There are positive correlations (red) between the bins at small wavelet amplitudes $A_n < 5 \times 10^{-2}$ with the other small values. For larger values, there are negative correlations (blue) between the larger wavelet amplitudes $A_n > 0.1$ and all other wavelet amplitude values. These effects are due to the shape of the PDF as well as the constraint that the PDF must integrate to 1. Increasing the counts for any wavelet amplitude value will cause the counts in the peak of the PDF (around $A_n \sim 0.1$ as seen in Figure 2.8 for $s_n = 51.09\,\mathrm{km\,s^{-1}}$) to decrease due to the integral constraint on the PDF. Meanwhile, the shape of the PDF means that when one bin along the tail ($A_n < 4 \times 10^{-2}$ as seen in Figure 2.8 $s_n = 51.09\,\mathrm{km\,s^{-1}}$) increases in

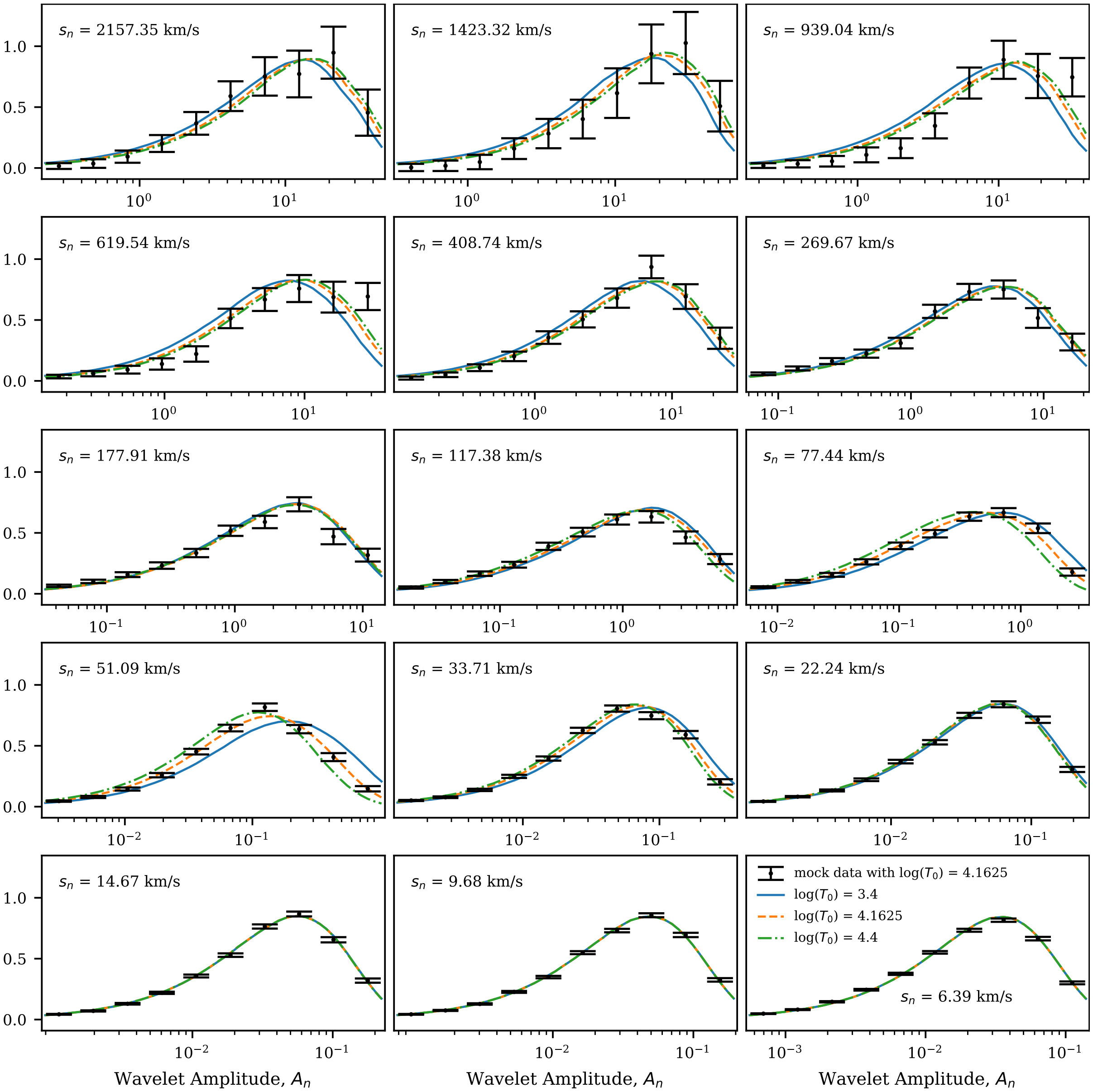

Figure 2.8: The black points show the PDFs from one mock data set for each s_n with $\log(T_0) = 4.1265$. The 1σ error bars are calculated from the square root of the diagonal of the covariance matrix. Each panel also shows the "model" values of the PDFs from the stated smoothing scale for three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). This figure qualitatively illustrates the ability of the wavelet PDF to differentiate between different T_0 values, which we formally quantify with Bayesian inference as discussed in Section 2.4.

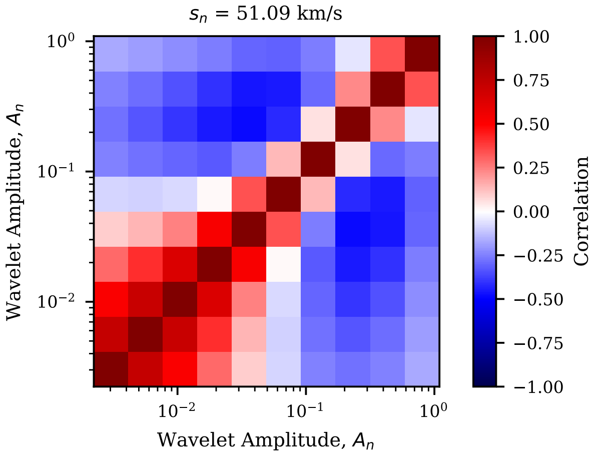

Figure 2.9: The correlation for the wavelet amplitude PDF for $\log(T_0) = 4.1625$ for $s_n = 51.09\,\mathrm{km\,s^{-1}}$. There are positive correlations (red) between the bins at small wavelet amplitudes $A_n < 5 \times 10^{-2}$ with the other small values. For larger values, there are negative correlations (blue) between the wavelet amplitudes $A_n > 0.1$ and all other wavelet amplitude values. These effects are due to the shape of the PDF as well as the constraint that the PDF must integrate to 1.

counts, the other tail bins will increase as well.

Ultimately, we will combine fifteen wavelet amplitude PDFs, each with a different value of s_n, in our measurement. Our measurement will include the correlations between the different PDFs, unlike the measurements from both Lidz et al. (2010) and Gaikwad et al. (2021) which ignore these correlations. We include these correlations by using non-zero off diagonal terms in each covariance matrix, $\Sigma(T_0)$, when computing the likelihood in equation (2.9). The correlations between different wavelet amplitude PDFs have never been considered in the previous literature on the Lyα forest. Our data vector is a concatenation of each wavelet amplitude PDF starting with the largest value of s_n going down to the smallest value, making it $n_{\mathrm{bins}} \times n_{s_n} = 10 \times 15 = 150$ points long. We expect that these correlations between different s_n values will be non-negligible due to the spatial correlations shown in the periodogram (Figure 2.3) as well as in the power spectrum (Figure 2.7).

In this case, the correlation matrix has dimension 150 × 150 and is shown in Figure

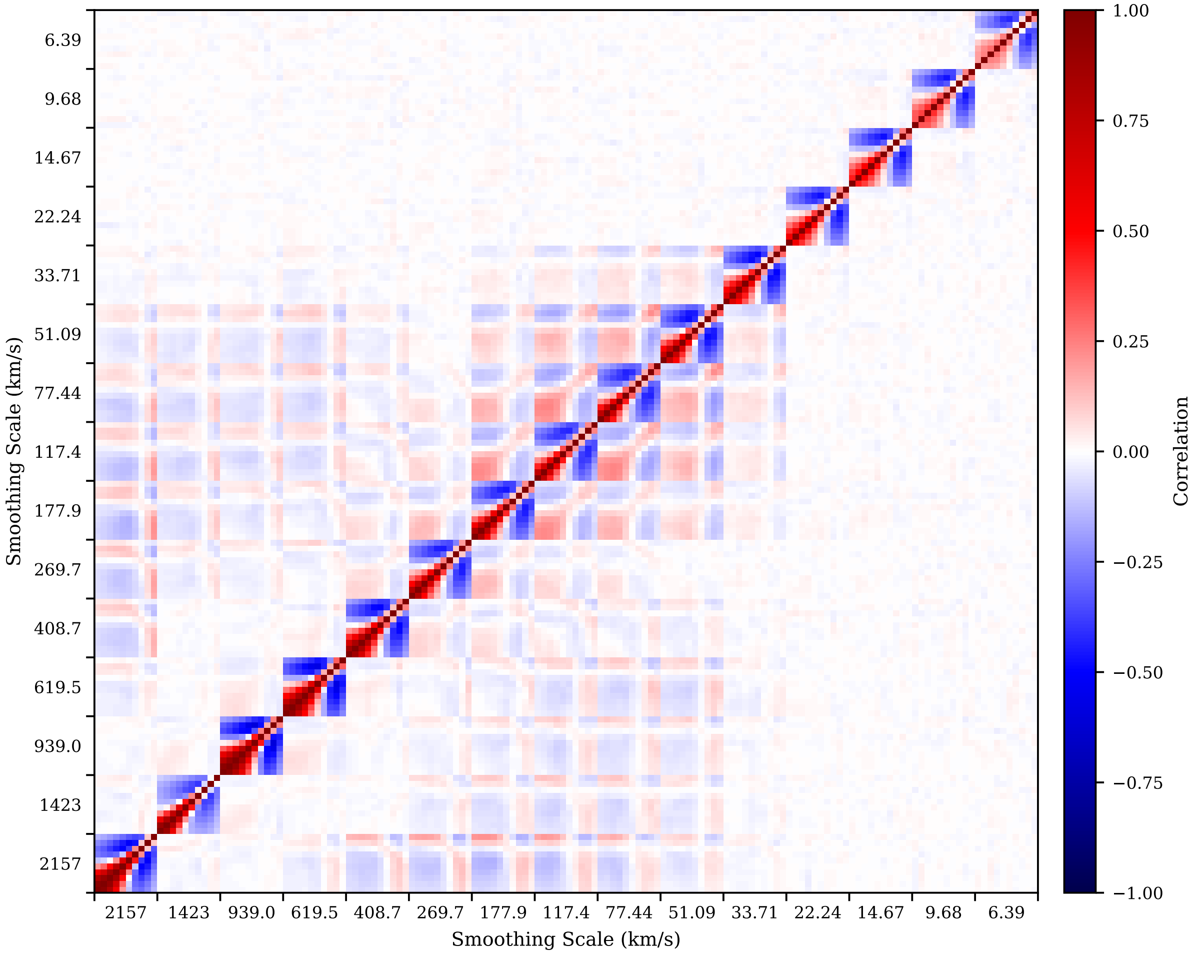

Figure 2.10: The correlation matrix for fifteen wavelet amplitude PDFs at $\log(T_0) = 4.1625$. The wavelet amplitude PDFs for large smoothing scales, $2157\,\mathrm{km\,s^{-1}} \geq s_n \geq 33.7\,\mathrm{km\,s^{-1}}$, have significant correlations off the diagonal. The correlations between the PDFs with $177.9\,\mathrm{km\,s^{-1}} > s_n > 33.71\,\mathrm{km\,s^{-1}}$ have the same pattern as the diagonal blocks modified by a small positive number (appearing mostly red). The correlations between the PDFs for $s_n > 408.7\,\mathrm{km\,s^{-1}}$ and $177.9\,\mathrm{km\,s^{-1}} > s_n > 33.71\,\mathrm{km\,s^{-1}}$ have the same pattern as the diagonal blocks modified by a small negative number (appearing mostly blue). For $s_n \leq 22.2\,\mathrm{km\,s^{-1}}$, the wavelet amplitudes begin to be dominated by noise, so the correlations between the PDFs for different values of s_n become very small. This pattern mimics that seen in the power spectrum correlation shown in Figure 2.7.

2.10. For visual purposes, the axes are labeled by the smoothing scale used to calculate the wavelet amplitude PDFs, but the correlations shown are between the wavelet amplitude bins (such as the labels in Figure 2.9). Each 10×10 block along the diagonal is the correlation matrix for a single s_n value. These diagonal blocks all appear very similar to the example shown for $s_n = 51.09\,\mathrm{km\,s^{-1}}$ in Figure 2.9, as expected from the similar shaped PDFs.

The wavelet amplitude PDFs for large smoothing scales, $2157\,\mathrm{km\,s^{-1}} \geq s_n \geq 33.7\,\mathrm{km\,s^{-1}}$, have significant correlations off the diagonal. The correlations between the PDFs with $177.9\,\mathrm{km\,s^{-1}} > s_n > 33.71\,\mathrm{km\,s^{-1}}$ have the same pattern as the diagonal blocks modified by a small positive number (appearing mostly red). The correlations between the PDFs for $s_n > 408.7\,\mathrm{km\,s^{-1}}$ and $177.9\,\mathrm{km\,s^{-1}} > s_n > 33.71\,\mathrm{km\,s^{-1}}$ have the same pattern as the diagonal blocks modified by a small negative number (appearing mostly blue). These modifications follow the same pattern as that in the correlation from the power spectrum shown in Figure 2.7 where there are positive correlations (red) between $4 \times 10^{-2}\mathrm{s\,km^{-1}} \lesssim k \lesssim 0.2\,\mathrm{s\,km^{-1}}$ and negative correlations between $4 \times 10^{-2}\mathrm{s\,km^{-1}} \lesssim k \lesssim 0.2\,\mathrm{s\,km^{-1}}$ and $k < 4 \times 10^{-2}\mathrm{s\,km^{-1}}$. This pattern arises from the underlying spatial correlations of the Lyα forest as was discussed for the power spectrum. For $s_n \leq 22.2\,\mathrm{km\,s^{-1}}$, the wavelet amplitudes begin to be dominated by noise, so the correlations between the PDFs of different smoothing scales become very small. This again mimics the behavior seen for $k > 0.2\mathrm{s\,km^{-1}}$ in the power spectrum correlation matrix shown in Figure 2.7.

Many of these off diagonal covariance elements are very small, and it is challenging to measure small correlations from finite noisy data sets. For the full set of fifteen wavelet amplitude PDFs, the covariance matrix is 100 times larger than the power spectrum covariance matrix, making the calculation even more time consuming and difficult. This noise in the wavelet amplitude PDFs covariance matrix becomes quite noticeable in the posterior measurement on T_0. We reduce this noise by smoothing the covariance matrices over multiple thermal grid with one spline per matrix element. A detailed discussion of this smoothing can be found in Appendix 2.8.

2.4.3 Joint Wavelet-Power Likelihood

Gaikwad et al. (2021) combined the wavelet amplitude PDFs with the power spectrum as well as the Doppler parameter distribution and curvature statistics to improve upon each of the individual measurements of the thermal state of the IGM. They did this by ignoring the correlations between PDFs for different smoothing scales as well as between the PDFs and the power spectrum despite the fact that these statistics were all measured from the same data set, and are thus surely correlated. This application has motivated us to combine the power spectrum and wavelet amplitude PDFs while paying careful attention to correlations to see if this improves the precision of our mock measurement. We expect there to be non-negligible correlations between the wavelet amplitude PDFs and the power spectrum from equation (2.7), since this says the mean wavelet amplitude, i.e. the first moment of the wavelet PDF, contains the same information as a band power.

When combining the wavelet amplitude PDFs and the power spectrum, the data vector is the 150 element wavelet amplitude PDFs, i.e. 10 PDF bins × 15 smoothing scales discussed in Section 2.4.2, with the addition of the 15 band-powers of $P_{\mathrm{F}}(k)$ added to the end. This makes the full data vector 165 points long and the correlation a complicated 165×165 matrix. To build intuition, we will first consider a subset of the full correlation matrix that consists of a single wavelet amplitude PDF with the power spectrum. The correlation matrix in this situation will be only 25×25, i.e. 10 wavelet PDFs values and 15 band-powers.

The correlation matrix for the wavelet amplitude PDF from $s_n = 51.09\,\mathrm{km\,s^{-1}}$ and the power spectrum is shown in Figure 2.11. The top panel has the full correlation matrix with the axes labeled by either the smoothing scale, $s_n = 51.09\,\mathrm{km\,s^{-1}}$, which was used to calculate the A_n or “Power” representing the different values of k. The top right 15×15 diagonal block is identical to the correlation matrix for the power spectrum shown in Figure 2.7 and the bottom left 10×10 diagonal block is identical to the correlation matrix for one $s_n = 51.09\,\mathrm{km\,s^{-1}}$ shown in Figure 2.9.

The off diagonal blocks show the correlations between the wavelet amplitude PDFs and the power spectrum. The bottom right rectangle of the correlation matrix is blown up in the bottom panel of the figure with the axes appropriately labeled by the wavelet amplitude A_n from the PDF and the k from the power spectrum. The strongest correlations (both positive and negative) between the power and wavelet amplitude PDF are found in the column at $k = 6/s_n = 0.12\,\mathrm{s\,km^{-1}}$. This k value corresponds to the same scales probed by $s_n = 51.09\,\mathrm{km\,s^{-1}}$. As the value of this k bin increases, we expect the wavelet amplitude PDF to shift to higher values so that the average wavelet amplitude in the PDF increases, as required by equation (2.7). This shift causes the larger values of A_n ($A_n > 0.2\,\mathrm{km\,s^{-1}}$) to be more common, resulting in a positive correlation (red) with larger PDF bins, while the smaller values of A_n ($A_n < 0.1\,\mathrm{km\,s^{-1}}$) are less common, resulting in a negative correlation (blue).

The behavior seen in the $k = 0.12\,\mathrm{s\,km^{-1}}$ column of the bottom panel is replicated for the columns above $3 \times 10^{-2}\mathrm{s\,km^{-1}} \lesssim k \lesssim 0.2\mathrm{s\,km^{-1}}$ modified by a small positive number. The columns where $k < 4 \times 10^{-2}\mathrm{s\,km^{-1}}$ show the same behavior modified by a small negative number. These modifications mimic the pattern in the correlations for the power spectrum, as shown in the upper right quadrant of the top panel and Figure 2.7. In particular, it replicates the positive (red) correlations for $3 \times 10^{-2}\mathrm{s\,km^{-1}} \lesssim k \lesssim 0.2\mathrm{s\,km^{-1}}$ with $k = 0.12\,\mathrm{s\,km^{-1}}$ and the negative (blue) correlations for $k < 4 \times 10^{-2}\mathrm{s\,km^{-1}}$ with $k = 0.12\,\mathrm{s\,km^{-1}}$.

As discussed in the beginning of this section, the total data vector for the combination

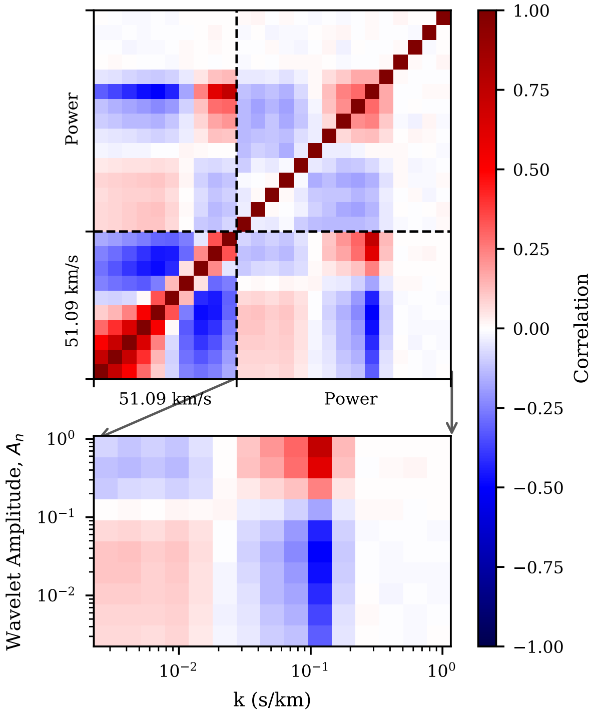

Figure 2.11: The top panel shows the correlation matrix for wavelet amplitude PDF with $s_n = 51.09\,\mathrm{km\,s^{-1}}$ and the power spectrum for $\log(T_0) = 4.1625$. The dotted lines separate the part of the matrix that corresponds to the wavelet amplitude PDF (labeled by $s_n = 51.09\,\mathrm{km\,s^{-1}}$) and the power spectrum (labeled "Power"). The bottom left diagonal block is identical to the correlation matrix shown in Figure 2.9 while the top right diagonal is identical to the correlation matrix shown in Figure 2.7. The bottom right rectangle of the correlation matrix is blown up in the bottom panel, which shows the correlations between the power spectrum and the wavelet amplitude PDFs with the appropriate labels of A_n and k. The strongest correlations (both positive and negative) between the power and wavelet amplitude PDF are found in the column at $k = 6/s_n = 0.12\,\mathrm{s\,km^{-1}}$. This k value corresponds to the same scales probed by $s_n = 51.09\,\mathrm{km\,s^{-1}}$. The behavior seen in the $k = 0.12\,\mathrm{s\,km^{-1}}$ column of the bottom panel is replicated for the columns above $3 \times 10^{-2}\mathrm{s\,km^{-1}} \lesssim k \lesssim 0.2\mathrm{s\,km^{-1}}$ modified by a small positive number. The columns where $k < 4 \times 10^{-2}\mathrm{s\,km^{-1}}$ show the same behavior modified by a small negative number. These modifications mimic the pattern in the correlations for the power spectrum, in particular the positive (red) correlations for $3 \times 10^{-2}\mathrm{s\,km^{-1}} \lesssim k \lesssim 0.2\mathrm{s\,km^{-1}}$ with $k = 0.12\,\mathrm{s\,km^{-1}}$ and the negative (blue) correlations for $k < 4 \times 10^{-2}\mathrm{s\,km^{-1}}$ with $k = 0.12\,\mathrm{s\,km^{-1}}$.

of the wavelet amplitude PDFs and the power spectrum is 165 points long. The full 165 × 165 correlation matrix is shown in Figure 2.12. The axes are labeled by either the smoothing scale used to calculate A_n or "Power" representing the k bands. The bottom left 150 × 150 diagonal block is identical to the correlation matrix for the wavelet amplitude PDFs shown in Figure 2.10 while the top right 15 × 15 diagonal block is identical to the correlation matrix for the power spectrum shown in Figure 2.7. The right most column above "Power" shows similar behavior for each smoothing scale as was discussed for the bottom panel of Figure 2.11. The column with the strongest correlations for each smoothing scale always corresponds to $k = 6/s_n$ and the behavior in other columns above "Power" follow the strongest bin modified by the correlations between the power bins.

This data vector is larger than the data vector of the wavelet amplitude PDFs, which was discussed in Section 2.4.2. Similarly, the noise in this covariance matrix is non-negligible and so we smooth the covariance matrix over the thermal grid with a spline in order to calculate the posteriors. This is discussed in more detail in Appendix 2.8.

2.5 Results

2.5.1 T_0 Measurements

We can calculate the posterior probability of T_0 given a mock data set, $P(T_0|\text{data})$, from equation (2.8). In Figure 2.13 we compare the posterior distribution of T_0 from one mock data set at $z = 5$ from three different methods: the power spectrum (blue triangles), the wavelet amplitude PDFs (orange circles), and both the power spectrum and wavelet amplitude PDFs (green triangles). This mock data set is the same one shown in Figures 2.6 and 2.8. Visually, the wavelet amplitude PDFs provide a more precise posterior for T_0 than the power spectrum does while. The measurement of T_0 for

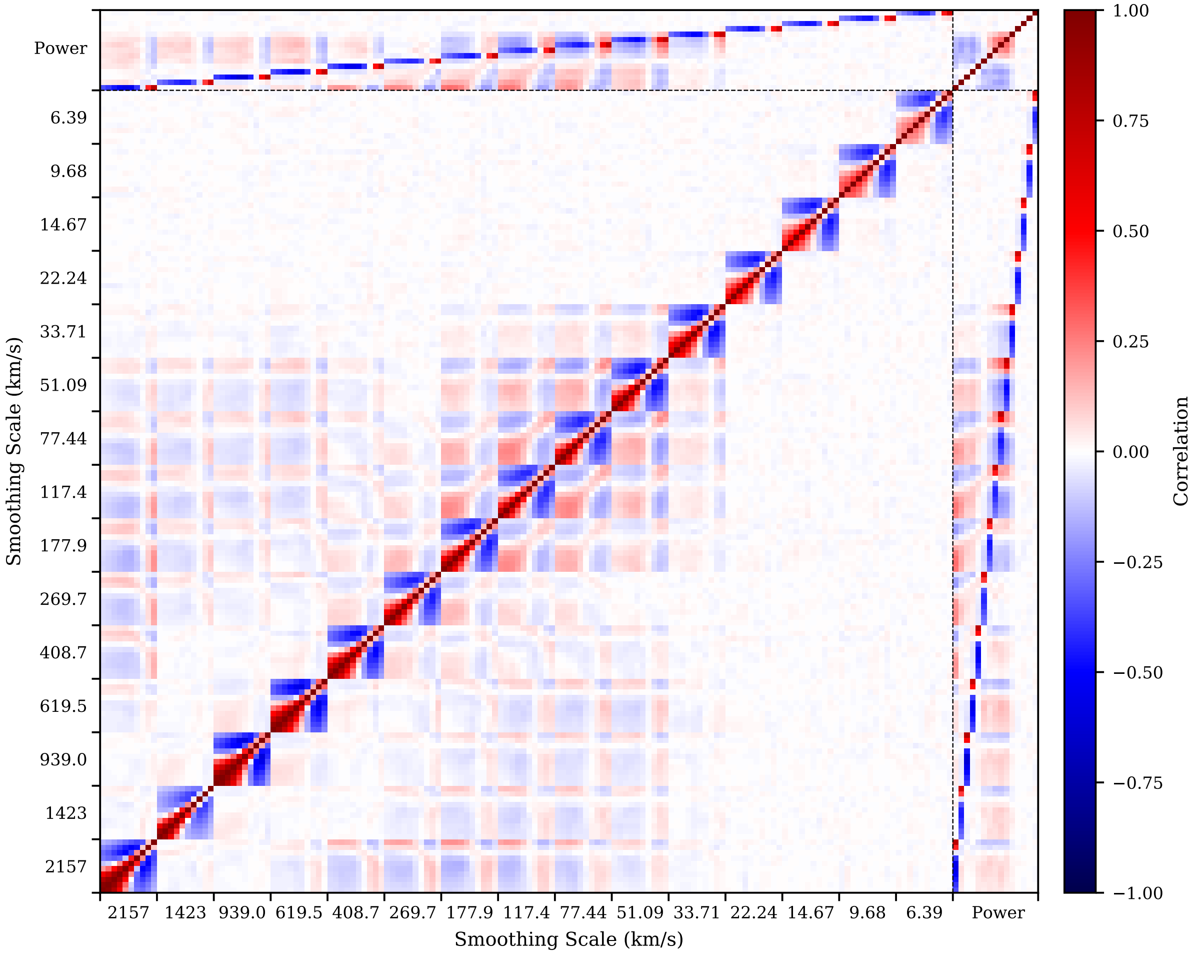

Figure 2.12: The correlation for all fifteen wavelet amplitude PDFs and the power spectrum combined for $\log(T_0) = 4.1625$. The axes are labeled by either the smoothing scale used to calculate A_n or "Power" representing the k bands. The bottom left 150×150 diagonal block is identical to the correlation matrix for the wavelet amplitude PDFs shown in Figure 2.10 while the top right 15×15 diagonal block is identical to the correlation matrix for the power spectrum shown in Figure 2.7. The right most column above "Power" shows similar behavior for each smoothing scale as was discussed for the bottom panel of Figure 2.11. The column with the strongest correlations for each smoothing scale always corresponds to $k = 6/s_n$ and the behavior in other columns above "Power" follow the strongest bin modified by the correlations between the power bins.

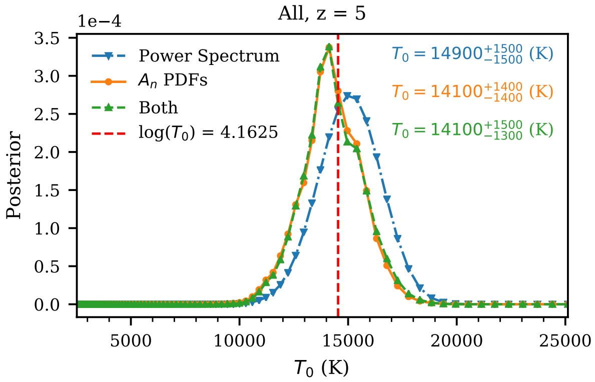

Figure 2.13: The posterior on T_0 for one mock data at $z = 5$ set from three different methods: the power spectrum (blue triangles), the wavelet amplitude PDFs (orange circles), and the combination of the power spectrum and the wavelet amplitude PDFs (green triangles). The mock data set used to calculate these posteriors is shown for the power in Figure 2.6 and for the wavelet amplitude PDFs in Figure 2.8. The vertical dotted red line shows the true value of T_0 for the mock data set. Qualitatively, the posterior from the power spectrum is less precise than the posterior from the wavelet amplitude PDFs and the combination of the power spectrum and wavelet amplitude PDFs do not improve the precision of the posterior over the wavelet amplitude PDFs alone. The text in the corner is a quantitative measurement of T_0, giving the median value with the equivalent 1σ errors for each method according to their colors in the same order as the legend.

these two methods are $T_0 = 14,900^{+1500}_{-1500}$ K (power spectrum) and $T_0 = 14,100^{+1400}_{-1400}$ K (wavelet amplitude PDFs). These errors are calculated by interpolating the cumulative distribution function (CDF) of the posterior onto the 15.9th and 84.1th percentiles, which correspond to the 1σ percentiles for a normal distribution. This region between these percentiles will be referred to as the 1σ region and the errors calculated from it as the equivalent 1σ errors throughout the end of this paper. The 1σ region is 7% smaller for the wavelet amplitude PDFs posterior than the power spectrum posterior, showing that the wavelet amplitude PDF is more sensitive to T_0 than the power for this mock data set. Combining the power spectrum and the wavelet PDFs has a negligible effect on the posterior distribution resulting in $T_0 = 14,100^{+1500}_{-1300}$ K. This 1σ region has the same width as the one for the wavelet amplitude PDFs alone and so the combination does not improve the measurement's precision.

In Figure 2.14 we compare the posterior distribution of T_0 from one mock data set at

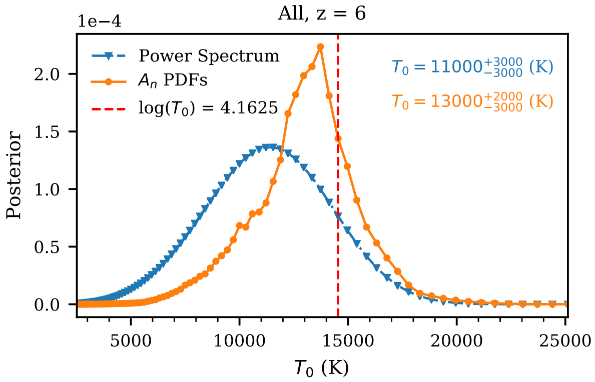

Figure 2.14: The posterior on T_0 for one mock data set at $z = 6$ from two different methods: the power spectrum (blue triangles) and the wavelet amplitude PDFs (orange circles). The mock data set used to calculate these posteriors is shown for the power in Figure 2.20 and for the wavelet amplitude PDFs in Figure 2.22. The vertical dotted red line shows the true value of T_0 for the mock data set. Qualitatively, the posterior from the power spectrum is less precise than the posterior from the wavelet amplitude PDFs. The text in the corner is a quantitative measurement of T_0, giving the median value with the equivalent 1σ errors for each method according to their colors in the same order as the legend.

$z = 6$ from two different methods: the power spectrum (blue triangles) and the wavelet amplitude PDFs (orange circles). This mock data set is the same one shown in Figures 2.20 and 2.22. The measurement of T_0 for these two methods are $T_0 = 11,000^{+3000}_{-3000}$ K (power spectrum) and $T_0 = 13,000^{+2000}_{-3000}$ K (wavelet amplitude PDFs) where the errors are calculated in the same way as they were for $z = 5$. At this redshift, the wavelet amplitude PDF measurements resulted in a 20% improvement of the 1σ errors when compared to the results from power spectrum measurement from the same mock data, almost three times the improvement seen at $z = 5$, though here the errors only have one significant figure so improvement is a coarser measurement.

To further quantify the difference in the precision of these posteriors, we calculated the equivalent 1σ errors for 1,000 mock data sets at $\log(T_0) = 4.1265$ for both $z = 5$ and $z = 6$. These resulting mean and variance of these values are listed in Table 2.2. On average, the posteriors for the wavelet amplitude PDFs at $z = 5$ are 7% smaller than those from the power spectrum, again showing that the wavelet amplitude PDFs

Table 2.2: This table shows the mean and variance of the 1σ values from 1,000 mock data sets at $\log(T_0) = 4.1265$ at both $z = 5$ and $z = 6$. The 1σ errors for the wavelet amplitudes PDFs are on average 7% smaller at $z = 5$ and 12% smaller at $z = 6$ than those for the power spectrum, though they do have a higher variance. The error calculated from combining the power spectrum and the wavelet analysis PDFs at $z = 5$ does not improve the errors on average over the wavelet amplitude PDFs alone.

z	Method	$\overline{\sigma}_+$	$\overline{\sigma}_-$
5	Power Spectrum	1490 ± 50	-1520 ± 50
	Wavelet Amplitude PDFs	1400 ± 200	-1400 ± 200
	Both	1400 ± 200	-1400 ± 200
6	Power Spectrum	3030 ± 190	-3140 ± 200
	Wavelet Amplitude PDFs	2700 ± 600	-2700 ± 600

are more sensitive on average than the power spectrum. However, the variance on these power spectrum errors are 75% smaller, meaning the power spectrum posteriors are more consistently large while the wavelet amplitude PDFs vary more in size. The average errors on the posteriors from combining both the wavelet amplitude PDFs and the power spectrum show no improvement over the errors from the wavelet amplitude PDFs alone again showing this combination does not improve the measurement.

For $z = 6$, the posteriors for the wavelet amplitude PDFs are 12% smaller than those from the power spectrum while the variance on the power spectrum errors are 67% smaller. This means that the wavelet amplitude PDFs are again more sensitive on average but the errors vary more than the power spectrum. This agrees with the results at $z = 5$, though again we find that the wavelets lead to an even greater improvement on the average sensitivity by a factor of two. Physically, at this higher redshift, more of the spectra consists of absorption troughs, giving the wavelet amplitude PDFs even greater potential to improve on the power spectrum measurements since they maintain spatial information.

2.5.2 Inference Testing

In order to test the fidelity of our statistical inference procedure and results, we perform an inference test. The goal of this test is to check that this calculated posterior behaves as a posterior probability should: if the true value of T_0 for the mock data falls into the equivalent of the 1σ region of the posterior $\sim 68\%$ of the time (and the 2σ region 95% of the time). We again calculate these equivalent 1σ and 2σ regions for our posteriors in the same way as discussed in Section 2.5.1. We integrate the posterior to get the CDF onto the 15.9th and 84.1th percentiles for 1σ and onto the 2.3rd and 97.7th percentile for 2σ. These percentiles correspond to the 1σ and 2σ percentiles for a normal distribution. From here, we count the number of times the true value of T_0 fell into these regions region. Ideally, the true value of T_0 should fall into the 1σ region 68.3% of the time and it should fall into the 2σ region 95.4% of the time.

We did this for 1,000 mock data sets at three different values of T_0 for $z = 5$ and one value of T_0 at $z = 6$. We chose to only look at one value of T_0 at the higher redshift because the posteriors are broader and we are more likely to run into edge effects at the other T_0 values. The errors are calculated by $\sqrt{N}/10000$ where N is the number of times the true value fell into the desired region and 1,000 is the total number of mocks used. The results, shown in Table 2.3, are consistent with the expected values of 68.3% and 95.4% within the calculated errors and we pass this inference test.

2.5.3 Ignoring Correlations

We further investigated the posterior distributions from the wavelet amplitude PDFs and the combined wavelet amplitude PDFs and power spectrum measurements at $z = 5$ while ignoring certain correlations. To begin, we considered the posterior from the wavelet amplitude PDFs alone. We constructed three distinct covariance matrices from our initial

Table 2.3: This table shows the results of our inference test for three values of T_0 and three statistical methods (power spectrum, wavelet amplitude PDFs, and the combination of the two) at $z = 5$. We have also included our results for one value of T_0 and two statistical methods (power spectrum and wavelet amplitude PDFs at $z = 6$. We calculated the equivalent 1σ and 2σ regions from the CDF and then determined the frequency with which the true T_0 values fell into these regions. These results are presented for 1,000 mock data sets and are consistent with a true distribution function with our expected errors.

z	Method	$\log(T_0)$	% in 1σ	% in 2σ
5	Power Spectrum	3.9	70.0 ± 2.6	94.6 ± 3.1
		4.1265	68.9 ± 2.6	96.0 ± 3.1
		4.2875	68.7 ± 2.6	96.2 ± 3.1
	Wavelet Amplitude PDFs	3.9	63.5 ± 2.5	94.0 ± 3.1
		4.1265	67.6 ± 2.6	95.0 ± 3.1
		4.2875	69.3 ± 2.6	95.9 ± 3.1
	Both	3.9	63.7 ± 2.5	93.1 ± 3.1
		4.1265	67.1 ± 2.6	94.1 ± 3.1
		4.2875	68.5 ± 2.6	95.3 ± 3.1
6	Power Spectrum	4.1265	68.1 ± 2.6	95.6 ± 3.1
	Wavelet Amplitude PDFs	4.1265	65.7 ± 2.6	93.6 ± 3.1

full calculation, which is shown in Figure 2.10. The first covariance matrix considered is made up of the same values along the diagonal and zeros for all off-diagonal elements. This is similar to the covariance considered in Lidz et al. (2010) and is referred to as the "no correlations" model in Figure 2.15 and Table 2.4. Next we construct a covariance matrix that includes the correlations between bins of the individual wavelet amplitude PDFs but ignores the correlations between different values of s_n. The resulting correlation matrix would have the same values for the fifteen 10×10 diagonal blocks in Figure 2.10 and zeros at all other locations. This is the similar to the covariance model considered in Gaikwad et al. (2021) and is referred to as the "PDF bin correlations" model in Figure 2.15 and Table 2.4. Finally we considered the full covariance matrix presented in this work in Figure 2.10, which we have labeled as "all correlations" in Figure 2.15 and Table 2.4. The resulting posteriors from these three models is shown in Figure 2.15. For the mock dataset shown in this figure (which is the same one shown throughout the rest of the paper) the "no correlations" model reduces the width of the posterior while the "PDF bin correlations" remains a similar width when compared to "all correlations". The median value does not agree for any of these posteriors though the whole distribution of the posteriors have significant overlap. We also performed the same inference test on the posteriors for these models which will be discussed at the end of this section with results in Table 2.4.

Next, we again constructed three different covariance matrices for the combination of the wavelet amplitude PDFs and the power spectrum where the initial covariance matrix is shown in Figure 2.12. First we considered only the correlations between PDF bins for the wavelet amplitude and the full correlations for the power spectrum. The resulting covariance matrix has fifteen 10×10 diagonal blocks followed by one 15×15 diagonal block and zeros at all other locations. This is similar to how the combination of different wavelet scales and different statistics were done in Gaikwad et al. (2021). We refer to this

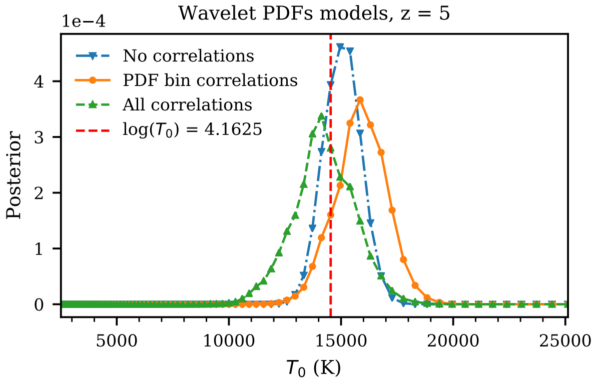

Figure 2.15: The posterior on T_0 for one mock data set for the wavelet amplitude PDFs using three different covariance matrices. The three matrices are described in more details in Section 2.5.3. They are: a diagonal-only covariance matrix which includes no correlations (blue triangles), a diagonal-block matrix that only contains correlations between PDF bins for the same wavelet scale (orange circles), and the full covariance matrix with all correlations (green triangles). The posterior from the diagonal matrix which has no correlations is much more narrow than the other two posteriors which are roughly the same width and height as each other.

model as the "PDF bin correlations" model in Figure 2.16 and Table 2.4. The subset of this matrix for the wavelet PDFs matches that used in Figure 2.15 with the same name. Next we consider the full wavelet correlations for the PDF bins as well as the different wavelet scales combined with the full power correlations but ignoring all cross correlations. The resulting matrix would have two diagonal blocks: one for the wavelet amplitude PDFs that has dimensions 150 × 150 and is shown in Figure 2.10, and one for the power spectrum that has dimensions 15 × 15 and is shown in Figure 2.7. We refer to this model as the "wavelet correlations" model in Figure 2.16 and Table 2.4. We also again considered the full covariance matrix presented in this work in Figure 2.12, which we have labeled as "All correlations" in Figure 2.16 and Table 2.4. The resulting posteriors from these three models is shown in Figure 2.16. For the mock dataset shown in this figure (which is the same one shown throughout the rest of the paper) the "PDF bin correlations" has the most narrow posterior while the "wavelet correlations" is more narrow than "all correlations" but broader than "PDF bin correlations". Again, each

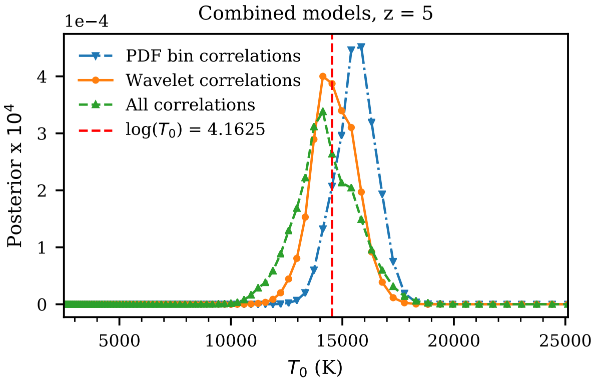

Figure 2.16: The posterior on T_0 for one mock data set for the combination of wavelet amplitude PDFs and the power spectrum using three different covariance matrices. The three matrices are described in more details in Section 2.5.3. They are: a diagonal-block matrix with 16 distinct blocks for the wavelet PDF bin correlations and power correlations separately (blue triangles), a diagonal-block matrix with 2 distinct diagonal blocks that contains the full wavelet correlations and the power correlations separately (orange circles), and the full covariance matrix with all correlations including the cross-correlations between the wavelet and the power spectrum (green triangles). Adding additional correlations caused the posterior distribution to broaden each time.

posterior has a shifted median value compared to the others, though all the posteriors have significant overlap with each other.

We repeat the inference test described in section 2.5.2 for these models and have presented the results in Table 2.4. For the wavelet amplitude PDF models, only the “all correltions” model passed our inference test with the true value of T_0 falling in our 1σ region for 67.6% of mock posteriors and the true value of T_0 falling in our 2σ region for 95.0% of mock posteriors. In comparison, the “no correlations” model only had the true value of T_0 falling in our 1σ region for 35.8% of mock posteriors and the true value of T_0 falling in our 2σ region for 62.7% of mock posteriors. We can roughly estimate that, since $68/36 \sim 1.9$, we would need to widen the posterior for the “no correlations” model by a factor of 1.9 to pass the inference test. If we take the “all correlations” model as the true information contained the in wavelet amplitude PDFs then the “no correlations” posterior needs to be both shifted and broadened to a lesser extent to match this posterior. A similar calculation can be done for the “PDF bin correlations” model which is more similar to the "all correlations" model initially.

The results for the inference test on the different models of the combined wavelet amplitude PDF and power spectrum correlations are also shown in Table 2.4. Here, again, only the "all correlations" model passed our inference test with the true value of T_0 falling in our 1σ region for 67.1% of mock posteriors and the true value of T_0 falling in our 2σ region for 94.1% of mock posteriors. In comparison, the "PDF bin correlations" model only had the true value of T_0 falling in our 1σ region for 47.7% of mock posteriors and the true value of T_0 falling in our 2σ region for 79.4% of mock posteriors. We can roughly estimate that, since $68/48 \sim 1.4$, we would need to widen the posterior for the "PDF bin correlations" model by a factor of 1.4 to pass the inference test. If we take the "all correlations" model as the true information contained the in combination of the wavelet amplitude PDFs and the power spectrum then the "PDF bin correlations" posterior needs to be both shifted and broadened to a lesser extent to match this posterior. A similar calculation can be done for the "wavelet correlations" model which is more similar to the "all correlations" model to start.

2.5.4 Comparison to Previous Work

Lidz et al. (2010) made a measurement of the thermal state of the IGM with one wavelet amplitude PDF while ignoring the correlations between bins of the PDF. This measurement of T_0 is higher than those from most other statistics and also has larger error bars. This is shown in Walther et al. (2019), which used the same quasar data set as Lidz et al. (2010) when measuring the power spectrum but added additional data to roughly double the data set size. Figure 15 of Walther et al. (2019) shows the resulting thermal state constraints compared to Lidz et al. (2010) as well as other measurements. We investigated the effect of ignoring the correlations between PDF bins in Section 2.5.3

Table 2.4: This table shows the results of our inference test when ignoring correlations for either the wavelet amplitude PDFs or the combination of the wavelet amplitude PDFs and the power spectrum. The three models of correlations for each statistic is described in Section 2.5.3. The inference test was done for only one true value of T_0, $\log(T_0) = 4.1265$. We calculated the equivalent 1σ and 2σ regions from the CDF and then determined the frequency with which the true T_0 values fell into these regions for 1,000 mock data sets. Only the models that considered all the correlations as presented in this paper labeled as "all correlations" for each statistic passed the inference test. The other two models for both statistics did not recover the true value of T_0 the expected number of times.

Method	Correlations	% in 1σ	% in 2σ
Wavelet Amplitude PDFs	No correlations	35.8 ± 1.9	62.7 ± 2.5
	PDF bin correlations	55.2 ± 2.3	87.0 ± 2.9
	All Correlations	67.6 ± 2.6	95.0 ± 3.1
Both PDFs and Power	PDF bin correlations	47.7 ± 2.2	79.4 ± 2.8
	Wavelet correlations	55.3 ± 2.4	86.5 ± 2.9
	All Correlations	67.1 ± 2.6	94.1 ± 3.1

and found that this would result in underestimated errors. This would imply that the error bars from Lidz et al. (2010) would not reflect the true precision of the measurement. Additionally, Figure 2.15 shows that ignoring these correlations also shifts the peak of the posterior. However, we did not consider only using one smoothing scale as was done in Lidz et al. (2010), which we would expect to broaden the posterior of the measurement. Thus we can not precisely estimate how shifted or underestimated the measurement and errors from Lidz et al. (2010) would have been.

Gaikwad et al. (2021) made a measurement of the thermal state of the IGM using eight wavelet amplitude PDFs with $30\,\mathrm{km\,s^{-1}} \leq s_n \leq 100\,\mathrm{km\,s^{-1}}$, the power spectrum, and by combining these statistics along with the Doppler parameter distribution and curvature statistics. All of these statistics were calculated from the same data set and the correlations between PDF bins of different smoothing scale and the individual statistics were ignored. The main results are shown in Figure 14 of Gaikwad et al. (2021), where the different statistics have measurements that agree with each other but differ in error bar size and the joint constraints visually appear to have the smallest error bars most frequently. Our analysis has shown that combining the wavelet amplitude PDFs and the power spectrum from the same data set does not improve the measurement of T_0 on average. However, we did not further investigate the additional statistics that Gaikwad et al. (2021) considered nor did we consider the situation where the smoothing scales used to calculate the wavelet amplitudes did not span the range of k values considered in the power measurement. For these reasons, our work here varies considerably from the work done by Gaikwad et al. (2021). It is likely that combining the power spectrum and wavelet amplitude PDFs when there isn't a full correspondence between s_n and k would lead to an improvement of the measurement from either statistic alone. However, as long as some k and s_n values overlap we expect there to be non-negligible correlations that were been ignored in Gaikwad et al. (2021). The two additional statistics are also likely to improve the combined measurement beyond the combination of only power spectrum and wavelet amplitude PDFs. Adding these statistics to the work presented here would require calculating the correlations between the different statistics as we did between the power spectrum and wavelet amplitude PDFs. Calculating these statistics and exploring the relevant correlations is beyond the scope of this work. For these reasons we can not precisely estimate the correct size of the error bars from their combined measurement.

We did consider the effect of ignoring the correlations between PDFS from different smoothing scales on the posterior for the wavelet amplitude PDFs alone in Section 2.5.3 and Figure 2.15. We found that ignoring these correlations caused the posterior to shift and underestimate the errors (the orange and green lines in Figure 2.15). Using our inference test in Table 2.4, for the wavelet amplitude PDFs with "PDF bin correlations" only the true value of T_0 fell in the 1σ region 55%. This would imply the need to grow the 1σ region by a factor of $68/55 \sim 1.24$ or a 24% increase. It is therefore likely that the errors on the measurement from only the wavelet amplitudes in Gaikwad et al. (2021) are underestimated.

Overall, Figures 2.9, 2.10, 2.11, and 2.12 show that the off-diagonal terms in the covariance matrix are non-negligible and should be included in future analysis using wavelet amplitudes in order to achieve accurate error estimates. Figures 2.15 and 2.16 and Section 2.5.3 additionally demonstrate and discuss the effects of ignoring these correlations on the posteriors.

2.6 Conclusion

We have expanded upon the wavelet analysis methods used by Lidz et al. (2010) and Gaikwad et al. (2021) to study the thermal state of the IGM. Our method combines fifteen wavelet amplitude PDFs with smoothing scales that span the full range of scales probed by $P_{\rm F}(k)$ and, for the first time, provides a full accounting of the correlations between these PDFs. We also calculated $P_{\rm F}(k)$ from the same simulated data in order to compare the precision of measurements on T_0 from these statistics. In order to rigorously combine the wavelet amplitude PDFs and power spectrum, we calculated the cross-correlations between $P_{\rm F}(k)$ and the wavelet amplitude PDFs. We presented examples of each of these correlation matrices in Figures 2.9, 2.10, 2.11, and 2.12. Figures 2.10 and 2.12 showed the non-negligible off-diagonal correlations between the different smoothing scales and the different statistics. With our method at $z = 5$, the posterior of T_0 using the wavelet amplitude PDFs is on average 7% more precise than the power spectrum measurement on the same data. This means getting the same precision measurement with the power spectrum requires $\sim 15\%$ more data. Combining the power spectrum and wavelet amplitude PDFs did not significantly improve in precision of the posterior on T_0 over that from the wavelet amplitude PDFs alone, indicating that they contain the same information. At

$z = 6$ we found that the posterior of T_0 using the wavelet amplitude PDFs is on average 12% more precise which would require $\sim 15\%$ more data to achieve the same accuracy with the power spectrum. Additionally, we calculated posteriors on T_0 at $z = 5$ with covariance matrices that ignored the off-diagonal correlations between PDF bins, between smoothing scales, and between the different statistics in Figures 2.15 and 2.16. We were unable to pass an inference test with these posteriors (as reported in Table 2.4) which implies that the errors are underestimated in these cases. This further demonstrated the significance of the off-diagonal terms in the covariance matrices and that they must be computed for a robust statistical analysis.

Here we adopted a simple model of the thermal state of the IGM which depended on a single parameter T_0. For the more common and general case of multiple model parameters, the wavelet amplitude PDFs have even greater potential to better constrain model parameters when compared to the 1D flux power spectrum. To reiterate, the wavelet amplitude PDF characterizes the full wavelet distribution while the 1D flux power spectrum contains information on the mean of the wavelet amplitude PDF, see Equation (2.7). If we are only varying one model parameter and this parameter shifts the wavelet amplitude PDFs, the mean of the wavelet PDF may effectively contain all the information on the differences of the model. This is true in our thermal model with T_0, as can be seen in Figures 2.5 and 2.8. In more sophisticated models, like those of reionization (Oñorbe et al., 2019; Boera et al., 2019) and WDM (Viel et al., 2013; Iršič et al., 2017), we would want to vary multiple model parameters (such as γ, $\langle F \rangle$, and m_{WDM}). Multiple model parameters are likely to cause changes in the full distribution of wavelet amplitudes beyond shifts in the mean. Wavelet amplitude PDF are sensitive to these additional changes while the power spectrum is not, meaning the wavelet amplitude PDFs could better discriminate between models to an even greater extent than they do for only one model parameter when compared to the power spectrum. Investigation of models with multiple thermal parameters is beyond the scope of this paper but is a more realistic and promising area to explore wavelet analysis.

Wavelets are an independent statistic that can be used to probe the small-scale structure of the IGM through the Lyα forest. They can be used as a check on alternative statistics such as the power spectrum, Doppler parameter distribution, and curvature statistics since, in principle, each statistic may be sensitive to different systematics. In addition, the wavelet amplitude PDFs are higher precision than the power spectrum. Wavelets have an added benefit of providing Fourier information in configuration space which may be useful in other areas, such as in quasar proximity zones (Khrykin et al., 2016) or when looking for temperature fluctuations in the IGM (Theuns & Zaroubi, 2000; Theuns et al., 2002b; Zaldarriaga, 2002; Fang & White, 2004; Lai et al., 2006; McQuinn et al., 2011). We have not studied the effects of a late ending reionization with remaining temperature fluctuations or a varying UVB background on the shape of the wavelet PDFs but this an interesting area to explore in the future (Davies & Furlanetto, 2016a; Becker et al., 2018; Kulkarni et al., 2019; Keating et al., 2020b; Nasir & D'Aloisio, 2020; Gaikwad et al., 2020).

Implementing wavelet analysis can be challenging due to the large size of data vectors and covariance matrices involved. Often the cross correlations are ignored (Lidz et al., 2010; Gaikwad et al., 2021) which can lead to inaccurate parameter constraints. It is also impossible to remove the effect of both the resolution and the noise on the full wavelet amplitude PDFs, meaning you have to forward-model these effects, unlike in power spectrum analysis where the noise can be subtracted and the resolution effects removed by a window function correction. When studying data, we would calculate the covariance matrix from bootstrapping the data itself as has been done when using other statistics of the Lyα forest (see Boera et al. (2019) for example). This would reduce the computational time required since we would not be need to compute the covariance matrix for each model.

An interesting subject for future work will be to build an emulator using wavelet PDFs analogous to Lyα forest power spectrum emulators (see, e.g. Walther et al. (2019)). One issue of concern for wavelets is the large number of functions that need to be emulated (15 wavelets versus 1 power spectrum), although it could be possible to simply emulate the wavelet likelihood which is a single function. The emulation field has shifted towards iterative sampling, informed by the posterior probability distribution for a given observational dataset (Rogers et al., 2019; Takhtaganov et al., 2021), making emulating the likelihood consistent with current methods.

In the future, our method of wavelet analysis can be applied to quasar data and more sophisticated simulations to obtain precision constraints on the thermal state of the IGM. One can expand our analysis to constrain the timing of reionization as well as models of dark matter. As mentioned above, our approach can also be adapted to analyses in proximity zones or searches for IGM temperature fluctuations exploiting the space-preserving properties of wavelets.

2.7 Appendix A: Likelihood Choice

We chose to use a multivariate Gaussian distribution for the likelihood of our data. This assumption explicitly means that if we take multiple mock data sets and look at the distribution of a single point in our data vector (either a single bin from the wavelet amplitude histogram or a single k band from power spectra) it will be Gaussian. It also assumes that when looking at any two points from the data vector, we expect the resulting distribution to be a two-dimensional Gaussian (thus looking at many points will be a multi-dimensional Gaussian).

We can visually check the assumption that any two points from the data vector will

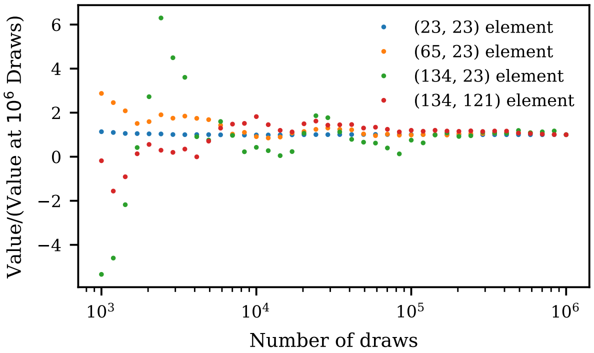

Figure 2.17: The distribution of mock draws for the power spectrum at $k = 0.12\,\mathrm{s\,km^{-1}}$ and the wavelet amplitude PDF at $A_n = 0.32$. The bottom left panel shows the 2D distribution of these bins where the red ellipse shows the 3σ region calculated from the covariance matrix for these two bins. The bottom right panel shows the distribution of values only for the power spectrum. The top left panel shows the distribution of values only for the wavelet amplitude PDF. All panels show good agreement with the assumption of a multi-variate Gaussian distribution.

result in a two-dimensional Gaussian distribution over many mocks. We will show this for one point in the power spectrum ($k = 0.12\,\mathrm{s\,km^{-1}}$) and one point in the wavelet amplitude PDF for $s_n = 51.09\,\mathrm{km\,s^{-1}}$ ($A_n = 0.32$). The distribution of these values for 1,000 mock draws is shown in Figure 2.17 along with the distributions of the individual Histogram and $P_{\mathrm{F}}(k)$ values. The red ellipse represents the 3σ contour from the 2×2 covariance matrix calculated for these two bins. The two dimensional distribution agrees very well with the ellipse by eye and the two histograms also visually appear Gaussian within the errors expected from counting statistics.

2.8 Appendix B: Covariance Matrix Calculation

As defined by equation (2.10), each model covariance matrix is calculated from mock draws of the data. This will inherently be a noisy calculation that will converge as $1/\sqrt{N}$ where N is the number of draws in the covariance matrix. As stated in the text, we used 1,000,000 mock draws when calculating the covariance matrix for the wavelet amplitude PDFs as well as the combination of the wavelet amplitude PDFs and the power

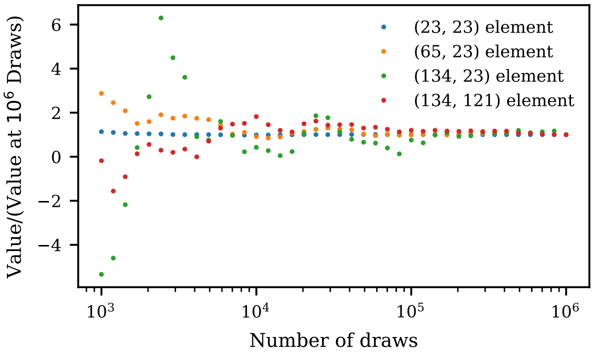

Figure 2.18: This figure shows the values of four distinct elements in the covariance matrix for the wavelet amplitude PDFs as we increased the number of mock draws used to calculate the covariance matrix. For simplicity, these distinct values are labeled by their index, rather than the smoothing scale and wavelet amplitude values associated for each bin in the PDFs. These points were chosen such that we had one on the diagonal, one off the diagonal where there are strong correlations, and two off the diagonal where there are weak correlations. As we approach 10^6 draws the values converge, showing that 10^6 is sufficient for our covariance matrix calculation.

spectrum. To check that 1,000,000 mock draws are sufficient to minimize the error from this calculation, we looked at the behavior of elements of the wavelet amplitude PDF covariance matrix in Figure 2.18. The values in the plot have been normalized such that at 10^6 draws they are 1. The four elements have been chosen such that there is one on the diagonal, one off the diagonal where there are strong correlations, and two off the diagonal where there are weak correlations. This plot shows that as we approach 10^6 draws the values vary significantly less than they do at lower values and thus the covariance elements are converging.

For both the wavelet amplitude PDFs and the combined power spectrum and wavelet amplitude PDFs covariance matrix, the data vector is long enough that there are many elements with very small cross correlations. These small values vary more (as seen in the (134, 23) and (134, 121) elements in Figure 2.18) such that they can still have non-negligible noise. This noise in the covariance leads to noise in the posterior measurement on T_0 (as discussed in Sections 2.4.2 and 2.4.3). At this point, for computational reasons, we decided not to increase the number of mock draws of data. Instead, we chose to

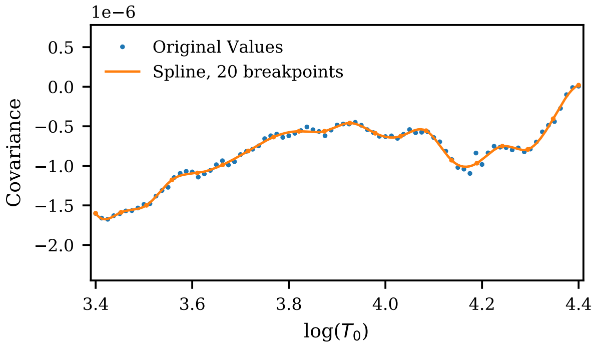

Figure 2.19: The covariance value for the same bin of the covariance matrix at different T_0. The bin chosen here corresponds to the histogram bin of $A_n = 0.027$ for $s_n = 51.09\,\mathrm{km\,s^{-1}}$ and the power spectrum band $k = 0.41\,\mathrm{s\,km^{-1}}$) and $A_n = 0.027$. The solid line shows the spline fit to these points with 20 equally spaced breakpoints.

smooth the covariance matrix across our thermal grid with a spline. We did this by fitting each individual element of the covariance matrix to spline with 20 equally spaced breakpoints. This results in 150×150 (165×165) splines for the wavelet amplitude PDFs (combined) covariance matrix. We chose 20 breakpoints to allow the spline to be flexible enough to find real patterns in the data while smoothing out noise.

We show one example of this spline in Figure 2.19. This is the spline for one of the elements of the covariance matrix shown in Figures 2.11 and 2.12. Specifically, this is for $k = 0.41\,\mathrm{s\,km^{-1}}$) and $A_n = 0.027$. This figure shows how the spline can replicate patterns in the calculated covariance matrix values while still reducing noise. For example, with the noisy values near $\log(T_0) = 4.2$.

The covariance matrices used in equation (2.9) for both the wavelet analysis and the combined wavelet and power spectrum analysis use the values of the spline at every T_0.

2.9 Appendix C: Redshift 6

Here we have included the figures for a mock data set at $z = 6$. As a reminder, at $z = 6$ we considered $\langle F \rangle = 0.011$, $R = 30,000$, and SNR $= 35$. We restricted our study

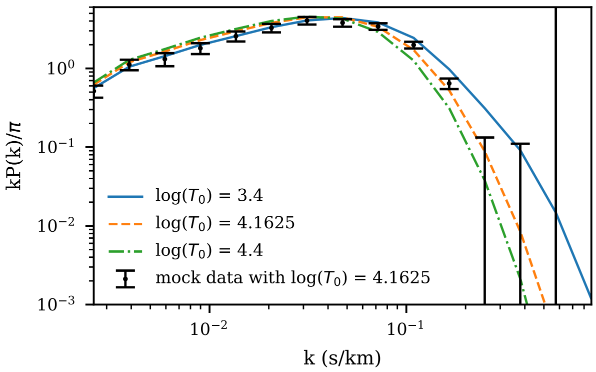

Figure 2.20: The power spectrum measurement, $P_{\mathrm{F}}(k)$, for one mock data set with $\log(T_0) = 4.1625$ (black points) at $z = 6$. The 1σ error bars are calculated from the square root of the diagonal of the covariance matrix. Also shown are model values of the power spectra for three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green).

to a higher SNR because the lower observed flux makes the noise power more dominant, see equation (2.14) for details. The mock data set still considers 8 quasars (equivalent to 29 simulation skewers).

First, in Figure 2.20 we show a mock power spectrum at $\log(T_0) = 4.1625$ with three models analogous to Figure 2.6 at $z = 5$. The error bars blow up at the largest values of k due to removing the noise power which is dominant at these small scales as well as correcting the window function. This Figure also shows three model power spectra at $z = 6$: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). The corresponding correlation plot for $\log(T_0) = 4.1625$ is shown in Figure 2.21. The structure here looks quite different than the one at $z = 5$ shown in Figure 2.7. The correlations come from underlying correlations in the high-z Lyα forest with the overall low mean flux. Again, the off-diagonal elements for high-k (small scales) are very small due to the dominance of the noise power over the signal at these scales.

Next, we look at the figures relevant for the wavelet amplitude PDFs at $z = 6$. The mock wavelet amplitude PDFs at all scales considered are shown in Figure 2.22. Note that the scales considered here are slightly different than those at $z = 5$ (as listed in

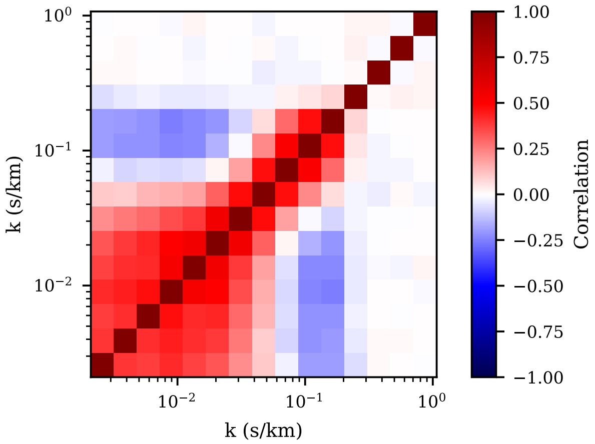

Figure 2.21: The correlation matrix for the power spectrum at $\log(T_0) = 4.1625$ and $z = 6$. The very weak correlations seen in the regions where $k > 0.2\,\mathrm{s\,km^{-1}}$ are due to uncorrelated random Gaussian noise which dominates the signal on small-scales (high k).

table 2.1 and Figure 2.8) because the size of the skewers and the nyquist frequency vary at these redshifts. At both redshifts we chose 15 logarithmically spaced values for s_n (and k). At this redshift and signal to noise ratio, we can see distinct bumps for the contribution of noise and wavelet amplitudes for $290.75\,\mathrm{km\,s^{-1}} \geq s_n \geq 55.09\,\mathrm{km\,s^{-1}}$. Again we have also shown three model wavelet amplitude PDFs: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). These wavelet amplitude PDFs all agree with each other for the smallest scales $s_n < 23.98\,\mathrm{km\,s^{-1}}$ where noise dominates the PDF which is the same for all temperature models. The correlation for all wavelet amplitude PDFs at $\log(T_0) = 4.1625$ is shown in Figure 2.23. The 10×10 diagonal blocks for the largest scales ($s_n > 440.7\,\mathrm{km\,s^{-1}}$) and the smallest scales ($36.34\,\mathrm{km\,s^{-1}} > s_n$) look similar to that shown for $z = 5$ in Figure 2.10 since the PDFs also have the same shape. The 10×10 diagonal blocks for the mid-range values of s_n are different from the others due to the combined PDF shape as seen in Figure 2.22. Again the off-diagonal blocks show a similar pattern to the diagonal blocks modified by positive or negative numbers, mimicking the off-diagonal correlations from the power at $z = 6$ seen in Figure 2.21. The off-diagonal correlation values between the $s_n \leq 23.98\,\mathrm{km\,s^{-1}}$ and all other values of

s_n show no strong correlations because the uncorrelated noise dominates at these small scales.

At this redshift we chose not to investigate the combination of the wavelet amplitude PDFs and the power spectrum due to the computational time required and the results at $z = 5$ which showed no significant improvement from combining these statistics when the same scales are covered in both.

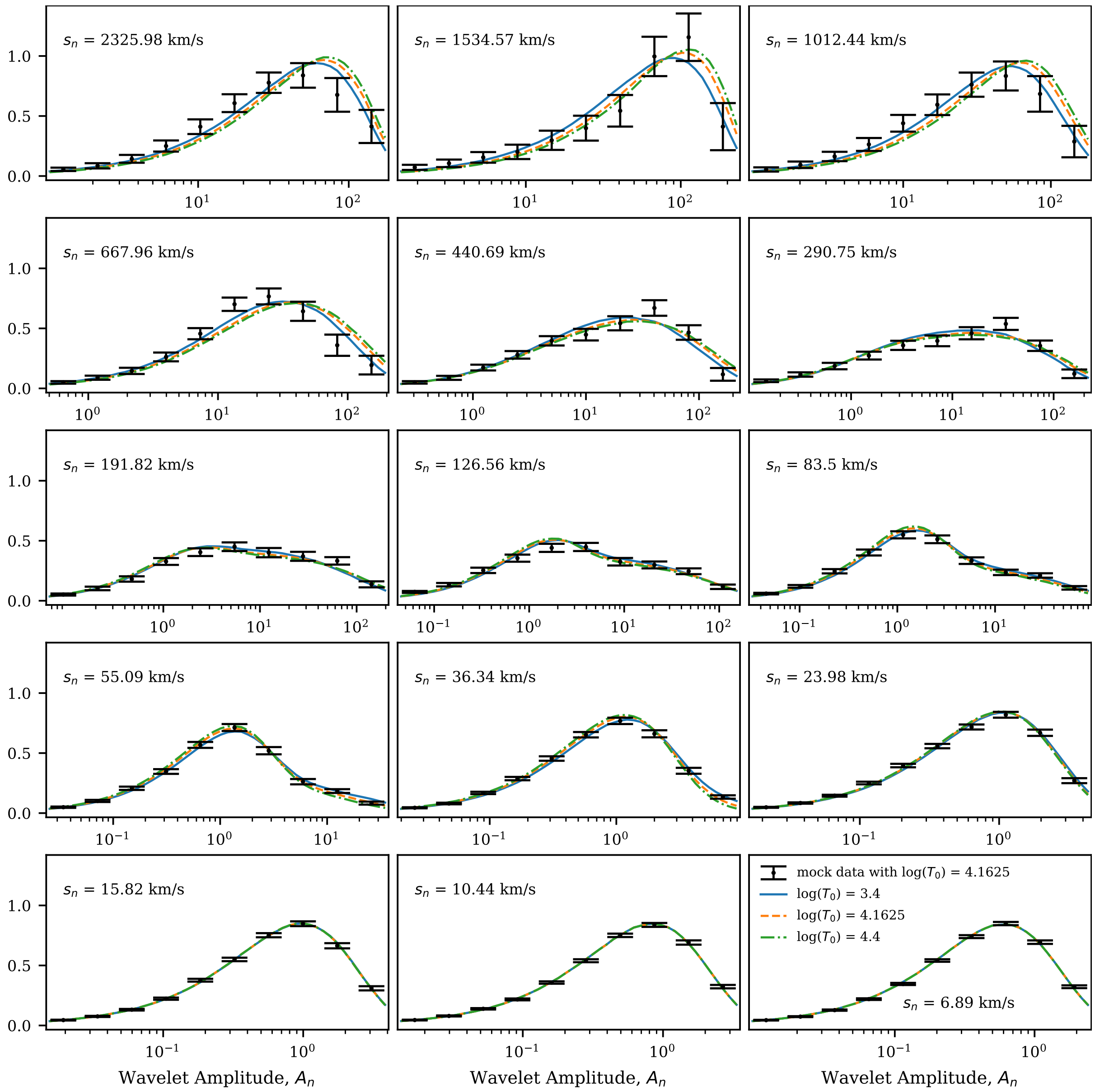

Figure 2.22: The black points show the PDFs from one mock data set for each s_n with $\log(T_0) = 4.1265$ and $z = 6$. The 1σ error bars are calculated from the square root of the diagonal of the covariance matrix. Each panel also shows the "model" values of the PDFs from the stated smoothing scale for three different values of T_0: $\log(T_0) = 3.4$ (blue), $\log(T_0) = 4.1625$ (orange), and $\log(T_0) = 4.4$ (green). This redshift shows broad PDFs with arguable two bumps in the mid-range values of s_n where the flux and noise power levels are comparable.

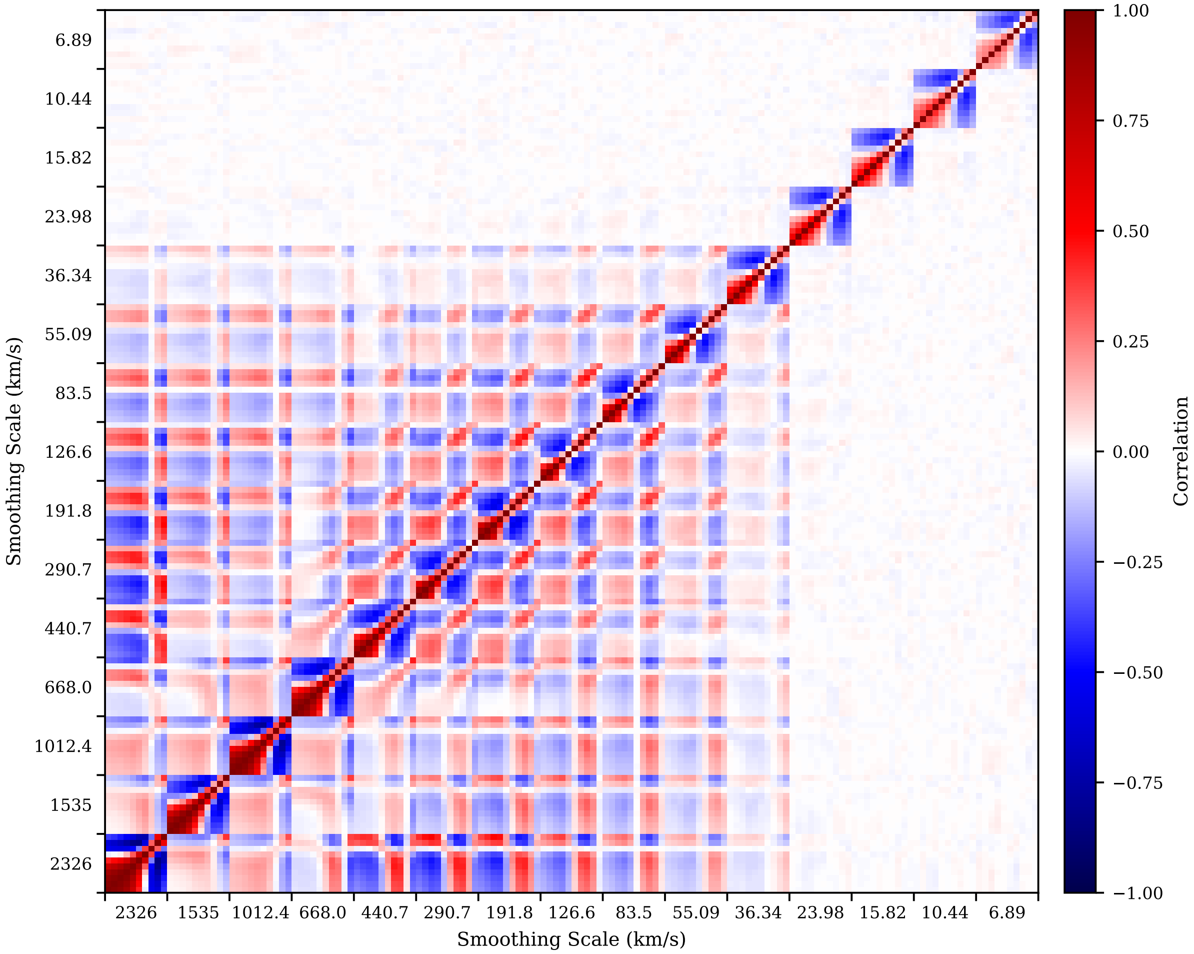

Figure 2.23: The correlation matrix for fifteen wavelet amplitude PDFs at $\log(T_0) = 4.1625$ and $z = 6$. The wavelet amplitude PDFs for large smoothing scales, $2326\,\mathrm{km\,s^{-1}} \geq s_n \geq 36.34\,\mathrm{km\,s^{-1}}$, have significant correlations off the diagonal. The off diagonal blocks show a similar repeating shape to those on the diagonal modified by numbers. These numbers are positive close to the diagonal and are negative further from the diagonal. For $s_n \leq 29.98\,\mathrm{km\,s^{-1}}$, the wavelet amplitudes begin to be dominated by noise, so the correlations between PDFs for different values of s_n become very small. This pattern mimics that seen in the power spectrum correlation shown in Figure 2.21. The pattern of the diagonal blocks are the most different for the mid range values of $290.7\,\mathrm{km\,s^{-1}} \geq s_n \geq 83.5\,\mathrm{km\,s^{-1}}$, which is where the PDF shapes are the most different from the typical shape, as seen in Figure 2.22.

Chapter 3

Forecasting constraints on the mean free path of ionizing photons at $z \geq 5.4$ from the Lyman-α forest flux auto-correlation function

This chapter was reproduced from Wolfson et al. (2023b) with only minor changes to fit the formatting of this dissertation. I'd like to thank my coauthors, without whom this work would not have been possible: Joseph F. Hennawi, Frederick B. Davies, and Jose Oñorbe.

3.1 Introduction

The neutral hydrogen in the intergalactic medium (IGM) was reionized by the first luminous sources during the epoch of reionization. This period was one of the most dramatic changes in the history of the universe. Current Planck constraints from the cosmic microwave background put the midpoint of reionization at $z_{\rm re} = 7.7 \pm 0.7$ (Planck Collaboration et al., 2020). There have also been multiple measurements that suggest reionization was not completed until after $z \leq 6$ (Fan et al., 2006; Becker et al., 2015, 2018; Bosman et al., 2018, 2022; Eilers et al., 2018; Boera et al., 2019; Yang et al., 2020; Jung et al., 2020; Kashino et al., 2020; Morales et al., 2021). However, much is still unknown about this process such as the exact timing, the impact on the thermal state of the IGM, the driving sources, and the number of photons that must be produced to complete reionization.

Characterizing the IGM both during and immediately after reionization will give vital information to answer these remaining questions. Of particular interest is the average distance that the ionizing photons travel through the IGM before interacting with its neutral hydrogen – also known as the mean free path of ionizing photons, $\lambda_{\rm mfp}$. The end of reionization results in a rapid increase in $\lambda_{\rm mfp}$ as the initially isolated regions of ionized hydrogen overlap to form a mostly ionized universe (Gnedin, 2000; Gnedin & Fan, 2006; Wyithe et al., 2008; D'Aloisio et al., 2018; Kulkarni et al., 2019; Keating et al., 2020b,a; Nasir & D'Aloisio, 2020; Cain et al., 2021; Gnedin & Madau, 2022). Detecting this rapid increase is therefore a clear signal of the end of reionization.

Direct measurements of $\lambda_{\rm mfp}$ at $z \leq 5.2$ have been achieved from stacked quasar spectra (Prochaska et al., 2009; Fumagalli et al., 2013; O'Meara et al., 2013; Worseck et al., 2014). Using a similar method, Becker et al. (2021) recently reported measurements of $\lambda_{\rm mfp} = 9.09^{+1.62}_{-1.28}$ proper Mpc at $z = 5.1$ and $\lambda_{\rm mfp} = 0.75^{+0.65}_{-0.45}$ proper Mpc at $z = 6$. This value at $z = 6$ is significantly smaller than extrapolations from previous lower z measurements (Worseck et al., 2014), causes tension with measurements of the ionizing output from galaxies (Cain et al., 2021; Davies et al., 2021), and also suggests a roughly 12-fold increase in $\lambda_{\rm mfp}$ between $z = 6$ and $z = 5.1$, potentially signalling the end of reionization. An alternative method presented in Bosman (2021) used lower limits on individual free paths towards high-z sources to place a 2σ limit of $\lambda_{\rm mfp} > 0.31$ proper Mpc at $z = 6.0$. This Bosman (2021) method is similar to other measurements using individual free paths (Songaila & Cowie, 2010; Rudie et al., 2013; Romano et al., 2019). Additional independent methods of measuring $\lambda_{\rm mfp}$ are necessary to verify these measurements. Of particular interest are methods that can be used at several redshift bins at $z > 5$ in order to study the evolution of $\lambda_{\rm mfp}$ in finer detail.

In this paper we investigate using the auto-correlation function of Lyα forest flux in high-z quasar sightlines to constrain $\lambda_{\rm mfp}$. The Lyα opacity, $\tau_{\rm Ly\alpha}$, is related to $\lambda_{\rm mfp}$ via $\tau_{\rm Ly\alpha} = n_{\rm HI}\sigma_{\rm Ly\alpha} \propto 1/\Gamma_{\rm HI} \propto 1/\lambda_{\rm mfp}^{\alpha}$ where α is typically between 3/2 and 2 (see e.g. Rauch (1998); Haardt & Madau (2012)). Additionally, during reionization the existence of significant neutral hydrogen in the IGM will cause a short mean free path value to also result in large spatial fluctuations in the ultraviolet background (UVB). This is because, during reionization ionizing photons are produced from the first sources and then quickly absorbed by the remaining neutral hydrogen. Thus there are large values of the UVB where the photons are produced and very small values where neutral hydrogen remains. If the mean free path is large, photons will travel further and effectively smooth the UVB (Mesinger & Furlanetto, 2009). The positive fluctuations in the UVB on small scales that accompany a short mean free path would then boost the flux of the Lyα forest on small scales, which could then be detected in the auto-correlation function. Various previous studies have investigated the effect of large scale variations in the UVB on the auto-correlation function and power spectrum of the Lyα forest (Zuo, 1992a,b; Croft, 2004; Meiksin & White, 2004; McDonald et al., 2005; Gontcho A Gontcho et al., 2014; Pontzen, 2014; Pontzen et al., 2014; D'Aloisio et al., 2018; Meiksin & McQuinn, 2019; Oñorbe et al., 2019). Our work is focused on determining if the effect of the fluctuating UVB on the auto-correlation function can lead to a constraint on $\lambda_{\rm mfp}$.

While the power spectrum has been a more popular statistic used on the high-z

Lyα forest to date (Boera et al., 2019; Walther et al., 2019; Gaikwad et al., 2021), the auto-correlation function has a few characteristics that make it easier to work with than the power spectrum. The two most obvious are the effect of noise and masking on the auto-correlation function when compared to the power spectrum. Astronomical spectrograph noise is expected to be white or uncorrelated. Uncorrelated noise only impacts the auto-correlation function at zero lag, since at all other lags the uncorrelated noise will average to zero. Therefore, by not measuring the auto-correlation at zero lag we have fully removed the effect of white noise. On the other hand, white noise is a constant positive value at all scales for the power spectrum. Thus the unknown noise level must be calculated and subtracted from power spectrum measurements which will add additional uncertainty to the final measurement. Additionally, real data often has regions of spectra that need to be removed from the quasar spectrum (e.g. for metal lines). Masking out these and other regions introduces a complicated window function to the power spectrum that must be corrected for (see e.g. Walther et al. (2019)) and will again increasing the uncertainty in the measurement. The auto-correlation function does not require a similar correction since masking only result in fewer points in bins for certain lags.

The structure of this paper is as follows. We discuss our simulation data in Section 3.2. The auto-correlation function and our other statistical methods are described in Section 3.3. We then discuss our results in Section 3.4 and summarize in Section 3.5. Here we also touch on how additional work on modeling $\lambda_{\rm mfp}$ in simulations as well as better statistical methods will improve these constraints.

3.2 Simulation Data

3.2.1 Models

In this work we use a simulation box run with `Nyx` code (Almgren et al., 2013). `Nyx` is a hydrodynamical simulation code that was designed for simulating the Lyα forest with updated physical rates from Lukić et al. (2015). The `Nyx` box has a size of $L_{\rm box} = 100$ cMpc h^{-1} with 4096^3 dark matter particles and 4096^3 baryon grid cells. This box is reionized by a Haardt & Madau (2012) uniform UVB that is switched on at $z \sim 15$. We have two snapshots of this simulation at $z = 5.5$ and $z = 6$. In this work we want to consider these models at seven redshifts: $5.4 \leq z \leq 6$ with $\Delta z = 0.1$. In order to consider the redshifts for which we do not have a simulation output, we select the nearest snapshot and use the desired redshift when calculating the proper size of the box and the mean density. This means we use the density fluctuations, temperature, and velocities directly from the nearest `Nyx` simulation output. We additionally used the $z = 6.0$ simulation snapshot to generate low-resolution skewers at $z = 5.7$ and found no significant change in our finally results, confirming that using the nearest simulation snapshot in this way is sufficient.

We also have separate boxes of fluctuating $\Gamma_{\rm HI}$ values generated with the semi-numerical method of Davies & Furlanetto (2016b). These boxes have a size $L_{\rm box} = 512$ cMpc and 128^3 pixels. We have one snapshot of these $\Gamma_{\rm HI}$ boxes at $z = 5.5$. To get the flux skewers used in this work, we combine random skewers of $\Gamma_{\rm HI}$ from these UVB boxes with the skewers from the `Nyx` box. The UVB boxes have a different resolution than the `Nyx` box, to generate a skewer of $\Gamma_{\rm HI}$ values we randomly selected a starting location and direction in the UVB box then linearly interpolated the $\log(\Gamma_{\rm HI})$ values onto the same length and resolution as the `Nyx` skewers.

The method of Davies & Furlanetto (2016b) allows for a spatially varying mean free

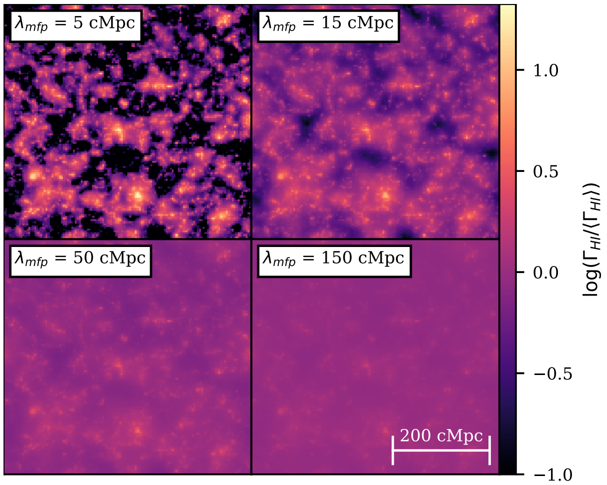

Figure 3.1: Each quadrant of this figure shows a slice through the box of the $z = 5.5$ UVB model used for four example values of $\lambda_{\rm mfp}$ (5, 15, 50, and 150 cMpc). The colorbar is cut off at $\log(\Gamma_{\rm HI}/10^{-12}{\rm s}^{-1}) = -1$ in order to better visualize the differences between the models. The models with smaller $\lambda_{\rm mfp}$ values show greater variation in the UVB than those with larger $\lambda_{\rm mfp}$, as expected.

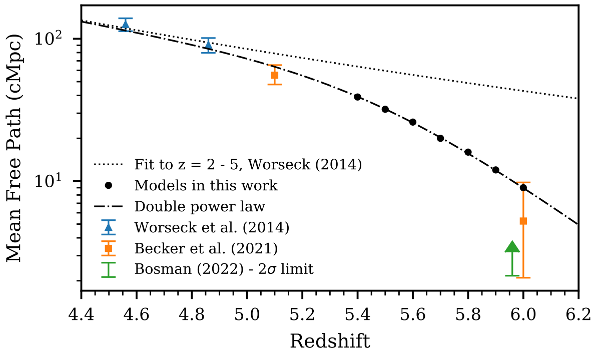

Figure 3.2: The blue triangles and orange squares show previous measurements of $\lambda_{\rm mfp}$ at high-z from Becker et al. (2021) and Worseck et al. (2014) respectively. The green limit is from Bosman (2021). Additionally, the dotted line shows the results of the power law fit to data from $z = 2 - 5$ from Worseck et al. (2014). For this work, we modified this power law fit into a double power law using the same low-z scaling by eye in order to agree with the Becker et al. (2021) points. This new scaling is shown by the dot-dashed line. We used this double power law as an example redshift evolution of $\lambda_{\rm mfp}$, where the values we modeled are shown as black circles.

path generated from fluctuations in the density of the sources of ionizing radiation with $\lambda \propto \Gamma_{\rm HI}^{2/3}\Delta^{-1}$, for λ, the local mean free path, and Δ, the local matter density. These simulations are scaled such that the mean value $\langle\lambda\rangle = \lambda_{\rm mfp}$ as desired. A brief summary of the Davies & Furlanetto (2016b) method is as follows. Cosmological initial conditions, independent of those from the 100 cMpc h^{-1} Nyx boxes, were generated for the 512 cMpc box and evolved to z = 5.5 via the Zel'dovich approximation (Zel'dovich, 1970). Halos were created via the approach of Mesinger & Furlanetto (2007) down to a minimum halo mass of $M_{\rm min} = 2 \times 10^9 M_\odot$. The ionizing luminosity of galaxies corresponding to each halo were determined following two steps: first the UV luminosities of galaxies were assigned by abundance matching to the Bouwens et al. (2015) UV luminosity function and then the ionizing luminosity of each galaxy was assumed to be proportional to its UV luminosity where the constant of proportionality is left as a free parameter. The ionizing background radiation intensity, J_ν, is then computed by a radiative transfer algorithm. The photoionization rate, $\Gamma_{\rm HI}$, is finally calculated by integrating over J_ν. For more details on the method see Davies & Furlanetto (2016b), Davies et al. (2018b) or Davies et al. 2022 in prep. where they also use this stitching procedure. Note that this method of generating UVB fluctuations ignores the effect of correlations between the baryon density in the Nyx boxes and the UVB. This is sufficient for the aims of this work but see Section 3.3.2 for a discussion on the effects of ignoring these correlations on the resulting auto-correlation function and therefore future measurements of $\lambda_{\rm mfp}$ from real data.

Example slices through the UVB boxes for four values of $\lambda_{\rm mfp}$ are shown in Figure 3.1 with a lower cutoff of $\log(\Gamma_{\rm HI}/\langle\Gamma_{\rm HI}\rangle) = -1$ for visual purposes. The top left box shows a slice of $\Gamma_{\rm HI}$ for the UVB simulation with the shortest $\lambda_{\rm mfp} = 5$ cMpc and has the greatest fluctuations. The bottom right box shows a slice of $\Gamma_{\rm HI}$ for the UVB simulation with the longest $\lambda_{\rm mfp} = 150$ cMpc and has the weakest fluctuations. This follows since overall longer λ values means that photons travel further and effectively smooth the UVB over these large scales.

We ran UVB boxes for 14 values of $\lambda_{\rm mfp}$ (in cMpc): 5, 6, 8, 10, 15, 20, 25, 30, 40, 50, 60, 80, 100, and 150. To generate UVB boxes for additional values of $\lambda_{\rm mfp}$ we linearly interpolated the $\log(\Gamma_{\rm HI})$ values at each location in the box between the two UVB boxes with the nearest $\lambda_{\rm mfp}$ values. This was done for three linearly spaced values between each existing $\lambda_{\rm mfp}$ values, resulting in a total of 53 UVB boxes.

To model a hypothetical evolution of $\lambda_{\rm mfp}$ as a function of redshift we used the double power law shown as the dot dashed line shown in Figure 3.2. This double power law was fit by eye with the following two considerations. We fixed the low z behavior to the power law fit from Worseck et al. (2014) for $z < 5$: $\lambda_{\rm mfp}(z) = (37 \pm 2) h_{70}^{-1} [(1+z)/5]^{-5.4 \pm 0.5}$ Mpc (proper). We also required consistency with the new measurements at higher z from Becker et al. (2021). The resulting double power law is:

$$\lambda_{\rm mfp}(z) = \frac{37 h_{70}^{-1} \left(\frac{5}{6.55}\right)^{5.4}}{\left(\frac{1+z}{6.55}\right)^{5.4} + \left(\frac{1+z}{6.55}\right)^{25.5}} \text{ Mpc (proper)}. \tag{3.1}$$

We then evaluated equation (3.1) at center of the seven redshift bins we considered and rounded to the nearest integer. The resulting true model $\lambda_{\rm mfp}$ values are listed in Table 3.1 and are plot as the black circles in Figure 3.2. If these values were already in our set of 53 models then nothing else was done. If not, we linearly interpolated the value of $\log(\Gamma_{\rm HI})$ at each point in the UVB simulation box between the two UVB boxes with the closest values of $\lambda_{\rm mfp}$ to get the final desired UVB box. This ultimately caused some redshifts to have 53 models of $\lambda_{\rm mfp}$ while others have 54. To generate the final flux skewers, we calculated the optical depths assuming a constant UVB then rescaled $\tau_{\rm mfp} = \tau_{\rm const.}/(\Gamma_{\rm HI}/\langle\Gamma_{\rm HI}\rangle)$. The $z = 5.5$ values of $\Gamma_{\rm HI}$ are used when generating flux skewers at all redshifts. This is justified because the value of $\lambda_{\rm mfp}$ is more important than the redshift evolution of the bias of the source population between $5 \leq z \leq 6$

Table 3.1: This table lists several relevant parameters for our simulations and mock data set. The second column lists the "true" values of the redshift-dependent $\lambda_{\rm mfp}$ calculated from equation (3.1). The third column gives the true values of $\langle F \rangle$ at each z from Bosman et al. (2022). These $\langle F \rangle$ values are the central value for the grid of values considered. The final column contains the number of quasar sightlines we modeled for one mock data set, which is the data set size in Bosman et al. (2022). These sightlines each have a length of $\Delta z = 0.1$.

z	$\lambda_{\rm mfp}$ (cMpc)	$\langle F \rangle$	# QSOs
5.4	39	0.0801	64
5.5	32	0.0591	64
5.6	26	0.0447	59
5.7	20	0.0256	51
5.8	16	0.0172	45
5.9	12	0.0114	28
6.0	9	0.0089	19

(Furlanetto et al., 2017).

The overall average of $\Gamma_{\rm HI}$ calculated in the UVB fluctuation simulations is not uniquely determined since this originates from complicated galaxy physics. Thus, we force the average mean flux, $\langle F \rangle$, to be the same for each model where the average is taken over all flux skewers considered. This is achieved by calculating a constant, a, such that $\langle e^{-a\tau} \rangle = \langle F \rangle$. Additionally, we want to consider how changes in $\langle F \rangle$ would affect the auto-correlation function and determine if there is a degeneracy with $\lambda_{\rm mfp}$. Therefore we create a grid of 9 values of $\langle F \rangle$ at each redshift. We chose the central value of $\langle F \rangle$ for a grid from Bosman et al. (2022) and chose the range of values to keep $\langle F \rangle > 0$ while not running into boundary issues during our inference.

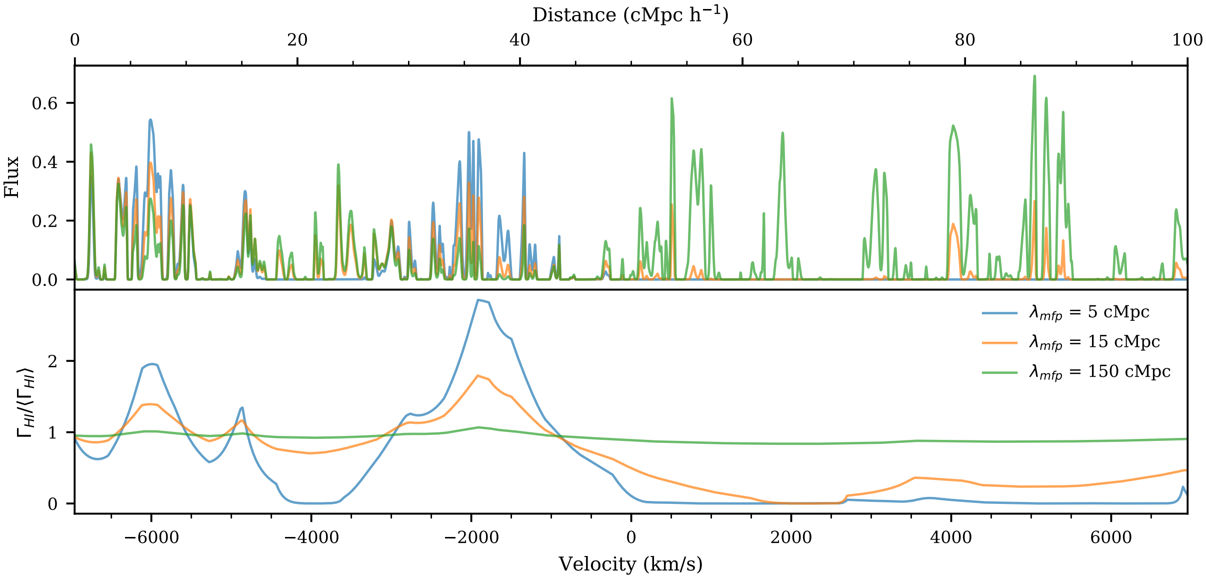

Figure 3.3: This figure shows the flux for one skewer of our simulation at at $z = 5.4$ with different values of λ_{mfp} all normalized to $\langle F \rangle = 0.0801$ in the top panel. The bottom panel shows the corresponding UVB skewer used to calculate the flux. Smaller λ_{mfp} values (such as $\lambda_{\mathrm{mfp}} = 5$ cMpc in blue) has greater variations in Γ_{HI} while the larger λ_{mfp} values (such as $\lambda_{\mathrm{mfp}} = 150$ cMpc in green) are more uniform. Larger values of Γ_{HI} leads to increased flux in that region which can be seen when comparing the two panels. Consider $\Delta v = -2000\,\mathrm{km\,s^{-1}}$, here $\lambda_{\mathrm{mfp}} = 5$ cMpc model (blue) has a peak in Γ_{HI}. The corresponding flux is boosted when compared to the other models. Additionally, for $\lambda_{\mathrm{mfp}} = 5$ cMpc (blue) the Γ_{HI} values are very small for $\Delta v \geq 0\,\mathrm{km\,s^{-1}}$ resulting in $F \sim 0$.

3.2.2 Comparison of Flux Skewers

We drew 1000 skewers from the Nyx simulation and 1000 independent skewers of Γ_{HI} from the UVB boxes to use in this work. One example flux skewer, which combines the Nyx simulation skewer and the Γ_{HI} values from the UVB boxes, at $z = 5.4$ is shown in Figure 3.3 for three different values of λ_{mfp} all normalized to $\langle F \rangle = 0.0801$. The bottom panel of this figure shows the corresponding UVB skewers that were used to calculate the flux. 2D slices of the UVB boxes these skewers came from are shown in Figure 3.1. The model shown with the shortest λ_{mfp}, 5 cMpc (blue), results in the greatest variation of $\Gamma_{\mathrm{HI}}/\langle\Gamma_{\mathrm{HI}}\rangle$. In particular, note that at $\Delta v = -2000\,\mathrm{km\,s^{-1}}$, the $\lambda_{\mathrm{mfp}} = 5$ cMpc model (blue) has a peak in Γ_{HI} and the corresponding flux is boosted when compared to the other models. Additionally, for $\lambda_{\mathrm{mfp}} = 5$ cMpc (blue) the Γ_{HI} values are very small for $\Delta v \geq 0\,\mathrm{km\,s^{-1}}$ resulting in $F \sim 0$. The model with the largest λ_{mfp}, 150 cMpc (green), shows a mostly uniform Γ_{HI} skewer throughout the whole velocity range leading to more consistent flux levels.

3.2.3 Forward Modeling

For this work we aim to model the resolution, noise, and size properties of a realistic data set. We first chose to model a simplified version of the XQR-30 (main and extended) data set[1]. The main XQR-30 data set consists of 30 spectra of the brightest $z > 5.8$ quasars observed with VLT/X-shooter (Vernet et al., 2011a). These spectra are supplemented with an extended data set consisting of 12 archival X-shooter spectra with comparable signal-to-noise ratio. See D'Odorico in prep. for additional information on these data. For this work we specifically model properties similar to the data set of Bosman et al. (2022) which consists of the etended XQR-30 data supplemented with

[1] https://xqr30.inaf.it/

additional archival X-Shooter data and archival Keck/ESI spectra which have a lower resolution than the X-shooter spectra.

For our simplified modeling, we use the resolving power of X-shooter for visible light with a 0.9" slit, so $R = 8800$. We also use a typical signal to noise ratio per $10\,\mathrm{km\,s^{-1}}$ pixel (SNR_{10}) of $\mathrm{SNR}_{10} = 35.9$, which is the median of all the data presented in Bosman et al. (2022). Additionally, we investigate how higher resolution data with access to smaller scales in the Lyα forest would impact measurements of λ_{mfp} from the auto-correlation function. To achieve this we consider a "high-resolution" data set with the same SNR_{10} and size properties as the "low-resolution" ($R = 8800$) data set but with $R = 30000$. This resolution is achievable with instruments such as Keck/HIRES, VLT/UVES, and Magellan/MIKE though the number of sightlines and noise properties used here do not represent a high-resolution data set currently in existence.

We model the resolution by smoothing the flux by a Gaussian filter then after smoothing we re-sampled such that there are 4 pixels per resolution element, where the resolution element is the FWHM. This means, for the low-resolution data set we smoothed by a Gaussian filter with FWHM $\approx 34\,\mathrm{km\,s^{-1}}$ then re-sampled so the pixel size was $\Delta v = 8.53\,\mathrm{km\,s^{-1}}$. For the high-resolution data set we smoothed by a Gaussian filter with FWHM $= 10\,\mathrm{km\,s^{-1}}$ then re-sampled so the pixel size was $\Delta v = 2.5\,\mathrm{km\,s^{-1}}$.

As stated above, we modeled a $\mathrm{SNR}_{10} = 35.9$. Using $\mathrm{SNR}_{\Delta v} = \mathrm{SNR}_{10}\sqrt{\Delta v / 10\,\mathrm{km\,s^{-1}}}$ this corresponds to a signal to noise ratio of 33.2 per 8.53 km/s low-resolution pixel and a signal to noise ratio of 18.0 per 2.5 km/s high-resolution pixel. For simplicity, we add flux-independent noise in the following way. We generate one realization of random noise drawn from a Gaussian with $\sigma_N = 1/\mathrm{SNR}_{\Delta v}$ for each SNR value and add this noise realization to every model at every redshift. The size of each noise realization is the number of skewers created (1000) by the number of pixels in the re-sampled flux skewers (1705 pixels for low-resolution and 5814 pixels for high-resolution). Using the same noise

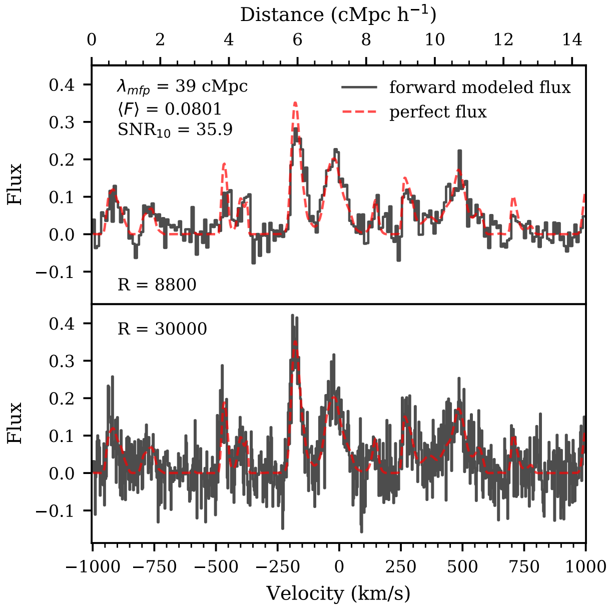

Figure 3.4: Both panels show initial and forward-modeled flux from a skewer with $\lambda_{\rm mfp} = 39$ cMpc and $\langle F \rangle = 0.0801$ at $z = 5.4$. The initial flux is the same in both panels (red dashed line) while the forward modeled flux (black histogram) varies. The top panel shows the low-resolution flux with $R = 8800$, which represents XQR-30 data. The bottom panel shows the high-resolution flux with $R = 30000$. Both of these resolutions have $\text{SNR}_{10} = 35.9$ which leads to differing $\text{SNR}_{\Delta v}$ as can be seen when comparing the two panels.

realization over the different models prevents stochasticity from different realizations of the noise from causing a noisy likelihood, which means the likelihood will be smooth as a function of model parameter. Thus the noise modeling will not unduly, adversely effect the parameter inference.

A section of one skewer for both the initial and forward-modeled flux is shown in Figure 3.4. Both panels shows a skewer at $z = 5.4$ with $\lambda_{\rm mfp} = 39$ cMpc and $\langle F \rangle = 0.0801$, our assumed true parameter values at this redshift. The initial flux in both panels is the same and is shown as a red dashed line. The top panel shows the low-resolution forward-modeled flux (black histogram) with $R = 8800$. The bottom panel shows the high-resolution forward-modeled flux (black histogram) with $R = 30000$. Again both of these panels have the same $\text{SNR}_{10} = 35.9$ which results in different noise levels per pixel, as can be seen when comparing the two panels.

We assume a fiducial data set size that matches the number of sightlines reported in Table 4 of Bosman et al. (2022) each with a length of $\Delta z = 0.1$. The number of sightlines are reported in the last column of Table 3.1 where to total pathlength considered is equal to these values multiplied by $\Delta z = 0.1$. Redshift bins of $\Delta z = 0.1$ correspond to distances of 33 to 29 cMpc h^{-1} when centered at $z = 5.4$ to $z = 6.0$. However, the `Nyx` simulation box is 100 cMpc h^{-1} long, much longer than these redshift bins. If we were to use the full 100 cMpc h^{-1} skewers in our calculation we would be averaging over fewer skewers to get the same total Δz path. We wanted to use a greater number independent skewers with more accurate lengths when compared with observed Lyα forest regions. For simplicity, we split all our skewers into two 40 cMpc h^{-1} regions which we treated as independent, giving us an effective number of 2000 independent skewers.

Note that unless otherwise specified the plots in this work mainly show results from the low resolution, $R = 8800$ data, since it represents existing XQR-30 data.

3.3 Methods

3.3.1 Auto-correlation function

The auto-correlation function of the flux ($\xi_F(\Delta v)$) is defined as

$$\xi_F(\Delta v) = \langle F(v)F(v + \Delta v)\rangle \tag{3.2}$$

where $F(v)$ is the flux of the Lyα forest and the average is performed over all pairs of pixels at the same velocity lag (Δv). The auto-correlation function is related to the power spectral density ($P_F(k)$) as

$$P_F(k) = \langle F\rangle^{-2} \int_{-\infty}^{\infty} \xi_F(\Delta v) e^{-ik\Delta v} d(\Delta v). \tag{3.3}$$

Note that this implies that the auto-correlation function should be sensitive to the same physical parameters as the power spectrum. Additionally, the auto-correlation function has nice properties with respect to white noise and spectral masks that make it a promising statistic to measure. Conventionally, the flux contrast field, $(F - \langle F \rangle)/\langle F \rangle$, is used when measuring statistics of the Lyα forest. Here, we chose to use the flux since $\langle F \rangle$ is small and has large uncertainties at high-z where we are most interested in this measurement. Using the flux thus prevents us from dividing by a small number which would come from an independent measurement and could potentially blow up the value of the flux contrast. This leads to the factor of $\langle F \rangle^{-2}$ in Equation (3.3).

For each resolution and model we compute the auto-correlation function with a bin size of one FWHM of the resolution (either $34\,\mathrm{km\,s^{-1}}$ or $10\,\mathrm{km\,s^{-1}}$) starting from this resolution size out to 20 cMpc h^{-1} (half the length of the skewer) which corresponds to $\sim 2900\,\mathrm{km\,s^{-1}}$ at $z = 5.4$. The model value of the auto-correlation function was determined by taking the average of the auto-correlation function over all 2000 forward-modeled skewers. Each mock data set of the auto-correlation were calculated by taking an average over the appropriate number of random skewers for the number of quasars at that redshift from the initial 2000 forward-modeled skewers. The value of the auto-correlation function for small-scale bins is affected by the finite resolution. This effect is left in both the models and the mock data. We determine the errors on the models via the following estimate of the covariance matrix from mock draws of the data:

$$\Sigma(\boldsymbol{\xi}_{\mathrm{model}}) = \frac{1}{N_{\mathrm{mocks}}} \sum_{i=1}^{N_{\mathrm{mocks}}} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_{\mathrm{model}})(\boldsymbol{\xi}_i - \boldsymbol{\xi}_{\mathrm{model}})^{\mathrm{T}} \tag{3.4}$$

where $\boldsymbol{\xi}_i$ is the auto-correlation function calculated for the i-th mock data set, $\boldsymbol{\xi}_{\mathrm{model}}$ is the average value of the auto-correlation function over all 2000 skewers, and N_{mocks} is the number of forward-modeled mock data sets used. Both the mock data sets and the overall average have the same values of λ_{mfp} and $\langle F \rangle$ in this calculation, so we end up

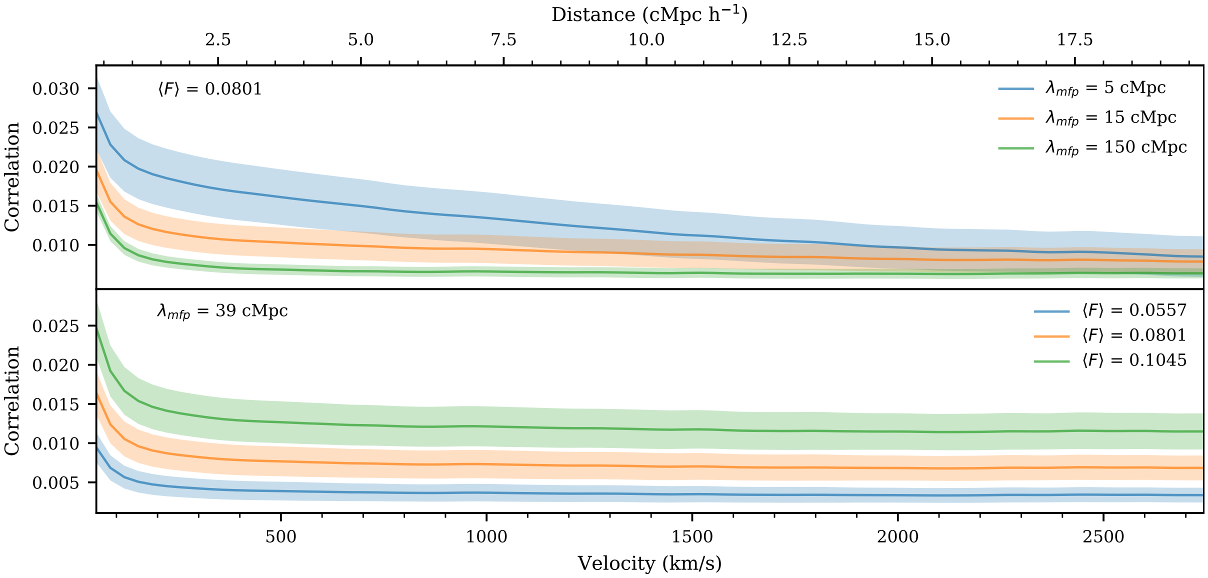

Figure 3.5: This figure demonstrates the effects of varying λ_{mfp} and $\langle F \rangle$ on the model values of the auto-correlation function at $z = 5.4$ and $R = 8800$. The solid lines show the model values calculated by averaging the auto-correlation function from all forward modeled skewers available while the shaded regions show the errors from the covariance matrix as estimated in equation (3.4). The top panel varies λ_{mfp} with a constant $\langle F \rangle$ labeled in the top left corner while the bottom panel does the opposite. Both λ_{mfp} and $\langle F \rangle$ change the auto-correlation function on all scales shown, though λ_{mfp} appears to effect small scales more than large scales. In the top panel, the model value of the auto-correlation function are further apart for $\lambda_{\mathrm{mfp}} = 5$ cMpc (blue) and $\lambda_{\mathrm{mfp}} = 15$ cMpc (orange) than for $\lambda_{\mathrm{mfp}} = 15$ cMpc (orange) and $\lambda_{\mathrm{mfp}} = 150$ cMpc (green), which is a greater difference in λ_{mfp} value. This means the auto-correlation function is more sensitive to small λ_{mfp} values than large λ_{mfp} values. Comparatively, in the bottom panel, the differences in the mean auto-correlation function appear roughly linear with varying $\langle F \rangle$ which should result in similar sensitivity for all $\langle F \rangle$ values.

with a covariance matrix at each parameter grid point. We use $N_{\mathrm{mocks}} = 500000$ for all models and redshifts in this work, see Appendix 3.6 for a discussion on the convergence of the covariance matrix.

Figure 3.5 shows the model value of the auto-correlation function with different parameter values at $z = 5.4$. The top panel shows models with a changing λ_{mfp} and constant $\langle F \rangle = 0.0801$. The solid lines show the model values calculated by averaging the auto-correlation function from all forward modeled skewers while the shaded regions show the errors from the diagonal elements of the covariance matrix as estimated in equation (3.4). Smaller λ_{mfp} values (such as $\lambda_{\mathrm{mfp}} = 5$ cMpc - blue) result in a greater correlation function at all scales, though mainly at small scales, and larger error bars than large $\lambda_{\rm mfp}$ values (such as $\lambda_{\rm mfp} = 150$ cMpc - green). These models are non-linearly spaced with greater differences between the models at small $\lambda_{\rm mfp}$ (blue vs orange) than large $\lambda_{\rm mfp}$ (orange vs green) which will result in variable sensitivity to $\lambda_{\rm mfp}$ from the auto-correlation function at different $\lambda_{\rm mfp}$ values. The bottom panel shows models with varying $\langle F \rangle$ and constant $\lambda_{\rm mfp} = 39$ cMpc. $\langle F \rangle$ sets the overall amplitude of the auto-correlation function. Here the differences between models are linear where larger $\langle F \rangle$ leads to larger auto-correlation values. This scaling is roughly $\propto \langle F \rangle^2$, which follows from the definition of the auto-correlation function.

To visualize the covariance matrix, we define the correlation matrix, C. The correlation matrix is the covariance matrix with the diagonal normalized to 1. This is done to the jth, kth element by

$$C_{jk} = \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}}. \tag{3.5}$$

One example correlation matrix is shown in Figure 3.6 for $z = 5.4$, $\lambda_{\rm mfp} = 39$ cMpc, and $\langle F \rangle = 0.0801$. All bins of the auto-correlation function are very-highly correlated which is due to the fact that each pixel in the Lyα forest contribute to multiple (in fact almost all) bins in the auto-correlation function.

3.3.2 Effect of Model Limits on the Auto-correlation Function

As stated in Section 3.2, the semi-numerical method to generate the fluctuating UVB with various $\lambda_{\rm mfp}$ ignores the correlation between the density and $\Gamma_{\rm HI}$. This is a result of the current limitations on available simulation boxes. We require that the UVB boxes are large enough to avoid suppressing UVB fluctuations and we require that the underlying hydrodrynamical simulation boxes of the IGM have a grid that is fine enough to resolve the small structures in the Lyα forest. Lukić et al. (2015) found that this grid needs

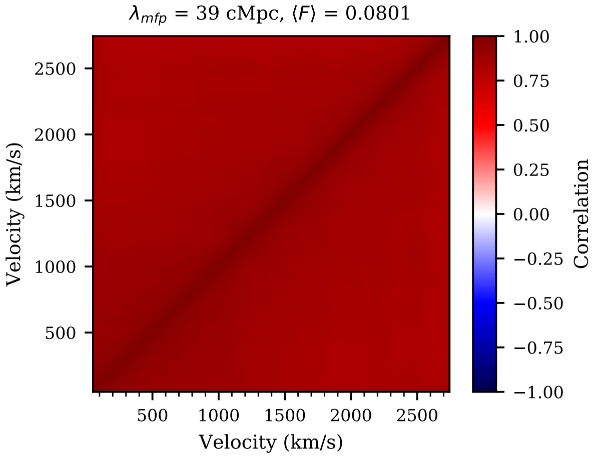

Figure 3.6: This figure shows the correlation matrix calculated with equation (3.5) with $N_{\text{mocks}} = 500000$ for the model at $z = 5.4$ with $\lambda_{\text{mfp}} = 39$ cMpc, $\langle F \rangle = 0.0801$, and $R = 8800$. The color bar is fixed to span from -1 to 1, which is all possible values of the correlation matrix. Here it is clear that all bins in the auto-correlation function are highly correlated with each other.

to have a grid resolution of 20 h^{-1} kpc to produce 1% convergence of Lyα forest flux statistics. Davies & Furlanetto (2016b) found that, with their 400 Mpc box of Γ_{HI} values, the tail of their optical depth distribution was impacted by cosmic variance, highlighting the need to go to even larger boxes. Having both a large box with a fine grid, which would be required to correlate the UVB and simulation box density, is currently too computationally expensive to be feasible.

In general, there is a positive correlation between density and Γ_{HI} and a negative correlation between density and transmitted flux. This means that in areas with high Γ_{HI} there should also be higher density which would in turn decrease the transmitted flux, therefore reducing the extra signal from the short λ_{mfp}. To quantitatively explore this, we used a `Nyx` simulation box with a size of $L_{\text{box}} = 40$ cMpc h^{-1} at $z = 5.8$. This box size has associated UVB values for $\lambda_{\text{mfp}} = 15$ cMpc generated with the same method of Davies & Furlanetto (2016b) as described in Section 3.2.1. For these UVB boxes the local matter density matches that of the `Nyx` simulations of the IGM. We selected skewers from the UVB boxes in two ways: from the same location as the `Nyx` skewers or from a

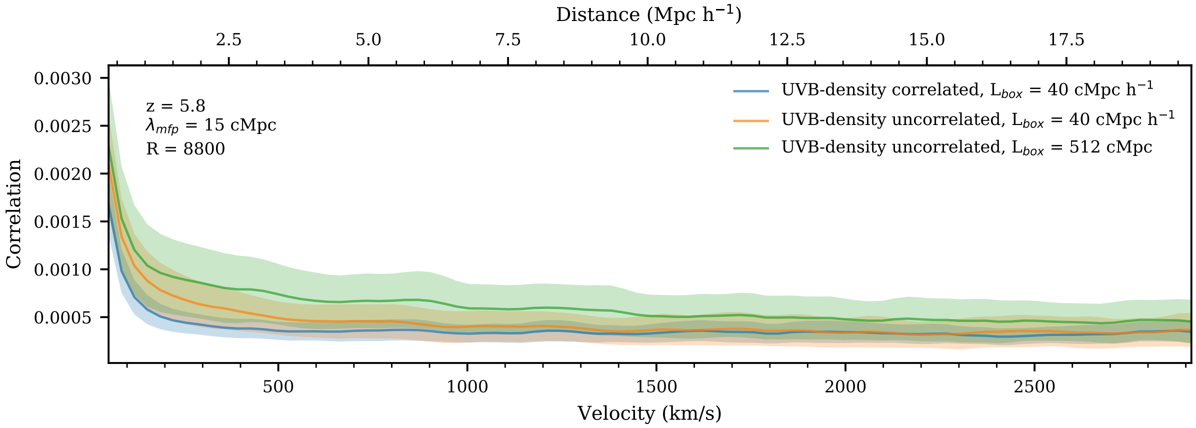

Figure 3.7: This figure demonstrates the effect of ignoring density correlations as well as using a small box size when generating $\Gamma_{\rm HI}$. The blue line shows the auto-correlation function when using a $\Gamma_{\rm HI}$ calculated with the appropriate density field and a box size of $L_{\rm box} = 40$ cMpc h^{-1}. The orange line shows the same for a $\Gamma_{\rm HI}$ calculated with the a random density field and a box size of $L_{\rm box} = 40$ cMpc h^{-1}, isolating the effect of density correlations when compared to blue. The green line shows the same for a $\Gamma_{\rm HI}$ calculated with the a random density field and a box size of $L_{\rm box} = 512$ cMpc, isolating the effect of the box size when compared with orange. Here we see that the correct density field will cause the signal on small scales to be reduced and that using a larger box size will increase the signal for all scales.

random location in the box. When the UVB skewers come from the same location as the Nyx skewers the density field and UVB field are correlated. When the UVB skewers come from a random location these two fields will not be correlated, which is analogous to the uncorrelated modeling adopted in the main text. The resulting auto-correlation function models are shown in Figure 3.7 as the blue and orange lines. The blue line is the model with UVB skewers from the 40 h^{-1} cMpc box that were derived from the same density field as the Nyx simulations. The orange line is the model with UVB skewers from the 40 h^{-1} cMpc box that were derived from a random density field. Comparing these two lines isolates the effects of ignoring the UVB-density correlations. Here we see that the density correlations reduce the auto-correlation signal at small scales while leaving the large scale signal unchanged. When correlated, $\Gamma_{\rm HI}$ is proportional to the local density field so the regions of high $\Gamma_{\rm HI}$ values will also be regions of higher density. Since the optical depth scales as a power of the local density field, the boosted signal on small scales from regions of high Γ_{HI} in the orange model will be reduced by the corresponding increased local density leading to the reduction in small scales in the blue model. Since the reduction is happening on small scales, this mimics the effect of instead having a model with a larger λ_{mfp}.

Additionally, we investigated the effect of the box size used to generate the UVB on the amount of fluctuations in Γ_{HI} seen at a fixed λ_{mfp}. Using a smaller box size, such as the 40 h^{-1} cMpc box considered in Oñorbe et al. (2019), can suppress fluctuations in the local λ value since there is a smaller volume that must average to λ_{mfp}. For this comparison, we use randomly selected UVB skewers from the 40 h^{-1} cMpc box as well as randomly selected skewers from our 512 cMpc UVB box with $\lambda_{\text{mfp}} = 15$ cMpc from the main text of this work as described in Section 3.2.1. The UVB skewers chosen with both of these methods are uncorrelated with the density field, so we isolate the effect of only the box size. The two resulting auto-correlation function models are also shown in Figure 3.7. Again, the orange line is the model with UVB skewers from the 40 h^{-1} cMpc box that were derived from a random density field. The green line shows the model with UVB skewers from the 512 cMpc box that has a random density field compared to the `Nyx` simulation. Comparing this green line to the orange line thus isolates the effect of the small box size where again the large box size is required for UVB fluctuations to converge for a given λ_{mfp}. Here we see that the green model has a greater signal than the orange at all scales. Therefore, both the blue and orange models with UVB skewers generated in a 40 h^{-1} cMpc box are likely underestimating the auto-correlation function on all scales. This makes it difficult to quantify the level of excess signal in the auto-correlation function that we get from ignoring correlations between the UVB and the local density field since the signal is underestimated on all scales when using the smaller UVB box. For this reason, we choose not to correct the mock data to account for the effect of using an uncorrelated UVB.

The mock data and models of the auto-correlation function from this study are self-consistently generated since both ignore the correlations between the UVB and the local density field. Therefore, the excess signal on small scales from modeling with an uncorrelated UVB will not bias the constraints obtained in this work. However, this excess on small scales needs to be accounted for when using the models of the auto-correlation function from the main text to constrain $\lambda_{\rm mfp}$ with observational data. We would expect measurements made by comparing data to models generated without UVB-density correlations to be biased towards larger values of $\lambda_{\rm mfp}$, since the reduced signal on small scales from real density correlations would look like a larger $\lambda_{\rm mfp}$ in our models. We can not quantify this potential bias with these simulations because, again, the small box size of 40 h^{-1} cMpc reduces the auto-correlation function signal on all scales. Modeling the UVB consistently with Lyα forest simulations in larger boxes is necessary to conclusively study the limitations of the model used in this work. We therefore leave this to future work.

3.3.3 Parameter Estimation

To quantify the precision with which $\lambda_{\rm mfp}$ can be measured we use Bayesian inference with a multi-variate Gaussian likelihood and a flat prior over the parameters of interest. This likelihood ($\mathcal{L} = p(\boldsymbol{\xi}|\lambda_{\rm mfp}, \langle F \rangle)$) has the form:

$$\mathcal{L} = \frac{1}{\sqrt{\det(\Sigma)(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}_{\rm model})^{\rm T}\Sigma^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}_{\rm model})\right) \tag{3.6}$$

where $\boldsymbol{\xi}$ is the auto-correlation function from our mock data, $\Sigma = \Sigma(\boldsymbol{\xi}_{\rm model})$ is the model dependent covariance matrix estimated by equation (3.4), and n is the number of points in the auto-correlation function. We discuss the assumption of using a multivariate Gaussian likelihood in Appendix 3.7.

Our models are defined by two parameters: $\lambda_{\rm mfp}$ and $\langle F \rangle$. We compute the posteriors for these parameters using Markov Chain Monte Carlo (MCMC) with the `EMCEE` package (Foreman-Mackey et al., 2013). We linearly interpolate the model values and covariance matrix elements onto a finer 2D grid of $\lambda_{\rm mfp}$ and $\langle F \rangle$ then use the nearest model during the MCMC. This fine grid has 137 values of $\lambda_{\rm mfp}$ and 37 values of $\langle F \rangle$. Our MCMC was run with 16 walkers taking 5000 steps each and skipping the first 500 steps of each walker as a burn-in.

Figure 3.8 shows the result of our inference procedure for one mock data set at $z = 5.4$. The top panel shows the mock data set with various lines relating to the inference procedure as follows. The green dotted line and accompanying text presents the true model that the mock data was drawn from. The mock data set is plot as the black point with error bars that come from the diagonal elements of the covariance matrix of the model that is nearest to the inferred model. The inferred model is the model that comes from the median of each parameter determined independently via the 50th percentile of the MCMC chains. The red lines and accompanying text shows the inferred model from MCMC. The errors on the inferred model written in the text are the distance between the 16th, 50th, and 84th percentiles of the MCMC chains. The blue lines show the models corresponding to 100 random draws from the MCMC chain to visually demonstrate variety of models that come from the resulting posterior. The bottom left panel shows a corner plot of the posteriors for both $\lambda_{\rm mfp}$ and $\langle F \rangle$. Here we see evidence of an extended tail out towards larger $\lambda_{\rm mfp}$ which is quantified in the asymmetric errors reported on the inferred value of $\lambda_{\rm mfp}$. This asymmetry comes from the non-linear spacing of the auto-correlation function models as discussed in Section 3.3.1.

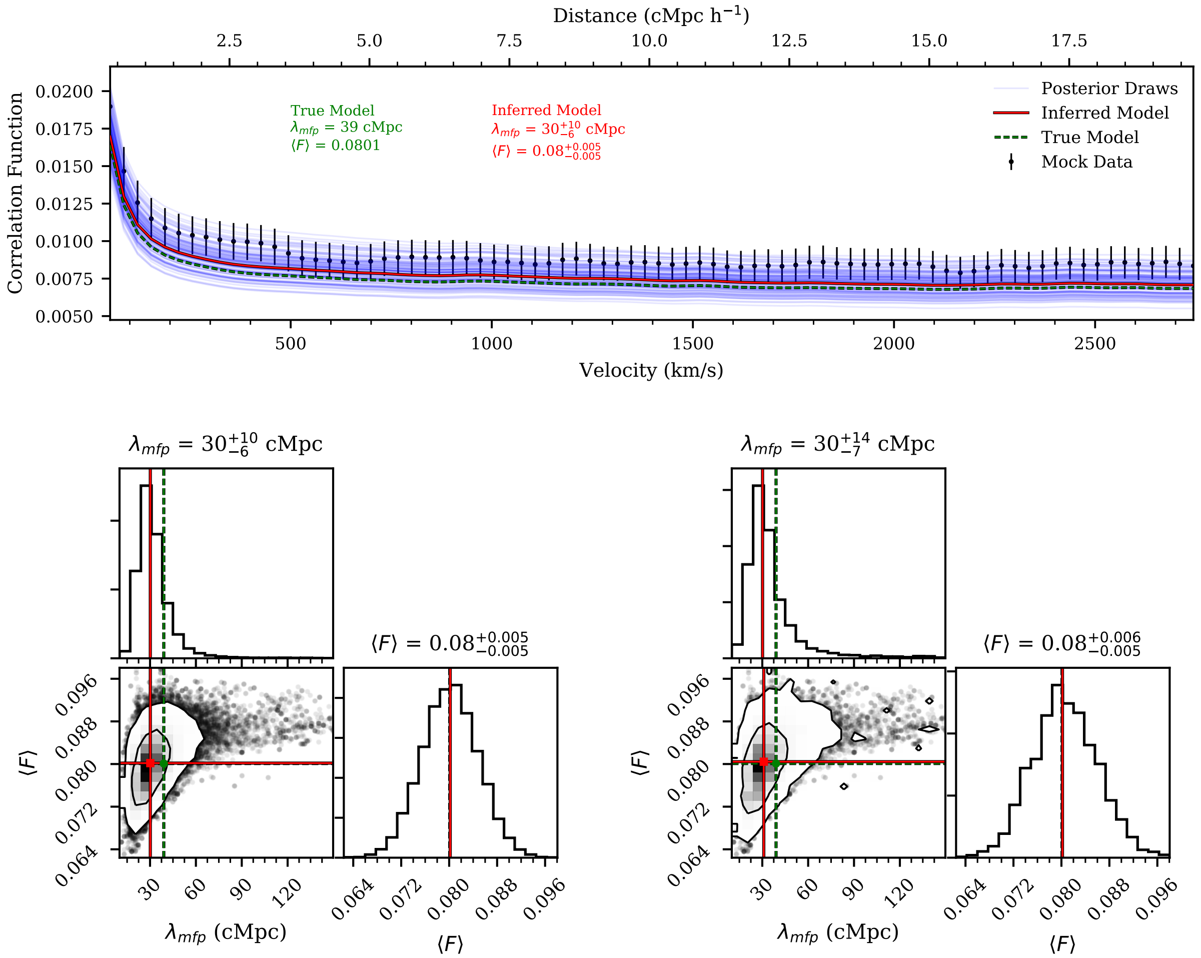

Figure 3.8: This figure illustrated the results of our inference procedure applied to one mock data set at $z = 5.4$. The top panel shows the resulting models from our inference procedure without re-weighting while the bottom panel has two corner plots that show the resulting parameters, the left without re-weighting and the right with re-weighting. In the top panel, the black points with error bars are the mock data with error bars from the inferred model. The inferred model was calculated by the median (50th percentile) of the MCMC chains of each parameter independently. The inferred model is shown as a red line while the accompanying red text reports errors calculated from the 16th and 84th percentiles of each parameter. In comparison, the true model the data was drawn from is the green dotted line and accompanying text. To demonstrate the width of the posterior, multiple faint blue lines are shown which are the models corresponding to the parameters from 100 random draws of the MCMC chain. The bottom left panel shows a corner plot of the values of λ_{mfp} and $\langle F \rangle$ that immediately result from our inference procedure. The bottom right panel shows the corner plot of the values of λ_{mfp} and $\langle F \rangle$ from our inference procedure using the re-weighting approach. This means the corner plot has been made with the weights calculated from our inference test as described in Section 3.3.4

3.3.4 Inference Test and Re-weighting

We perform a test to check the fidelity of our inference procedure in order to ensure that our resulting posteriors act in a statistically valid way. This will ensure any assumptions we make during our inference are justified. For example, in this work we used an approximate likelihood in the form of a multivariate Gaussian likelihood. The Lyα forest is known to be a non-Gaussian random field. By adopting a multivariate Gaussian likelihood here, we are tacitly assuming that averaging over all pixel pairs when calculating the auto-correlation function will Gaussianize the resulting distribution of the values of the auto-correlation function, as is expected from the central limit theorem. We discuss the distribution of these values for our mock data in detail in Appendix 3.7. If this assumption is not valid our reported errors may be either underestimated or overestimated.

The general idea of our inference test is to compare the true probability contour levels with the "coverage" probability. The coverage probability is the percent of time the probability of the true parameters of a mock data set fall above a given probability level over many mock data sets. In our case, we compute this over 500 mock data sets where the true parameters considered are samples from our priors. Ideally, this coverage probability should be equal to the chosen probability contour level. This calculation can be done at many chosen probabilities resulting in multiple corresponding coverage probabilities. Existing work that explore this coverage probability include Prangle et al. (2013); Ziegel & Gneiting (2013); Morrison & Simon (2017); Sellentin & Starck (2019).

When considering multiple chosen probabilities, P_{true}, and resulting coverage probabilities, P_{inf}, the results can be plot against each other. The results of our inference test at $z = 5.4$ from 500 posteriors with true parameters randomly drawn from our priors are shown in the left panel of Figure 3.9. The grey shaded regions around our resulting

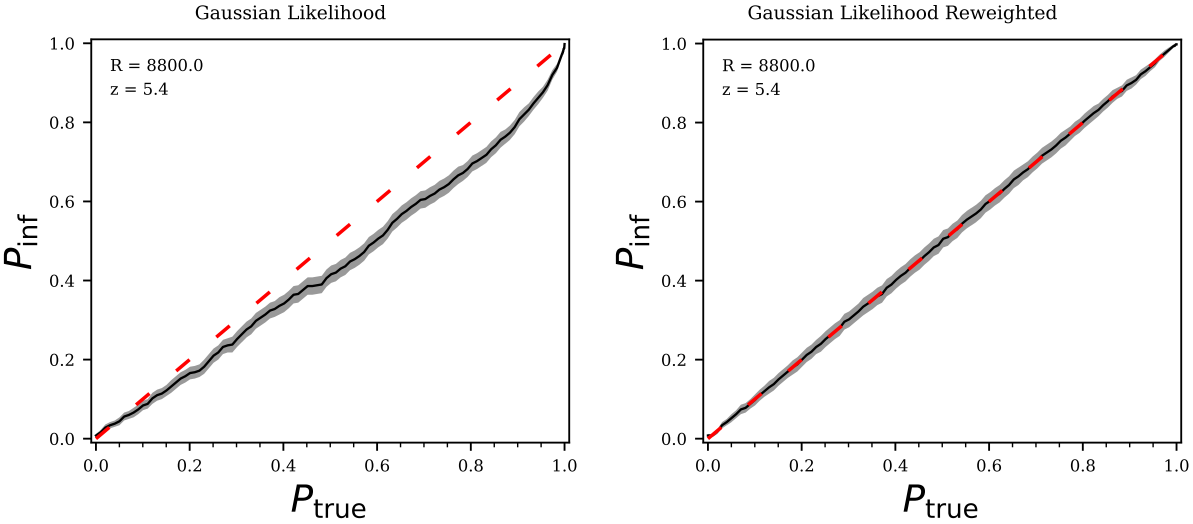

Figure 3.9: The left panel of this figure shows the coverage resulting from the inference test from 500 models at $z = 5.4$ and $R = 8800$ drawn from our priors on $\lambda_{\rm mfp}$ and $\langle F \rangle$. Here we see that the true parameters for the models fall above the 60th percentile in the MCMC chain $\sim$ 50% of the time, for example. The right panel of this figure shows the coverage resulting from the inference test with the use of one set of weights to re-weight the posteriors. With these weights we are able to pass the inference test.

line show the Poisson errors for our results. Again we expect $P_{\rm true} = P_{\rm inf}$ which would give the red dashed line in this figure. To interpret this plot, first consider one point, for example $P_{\rm true} \approx 0.6$. This represents the 60th percentile contour, which was calculated by the 60th percentile of the probabilities from the draws of the MCMC chain for each mock data set. Here, the true parameters fall within the 60th percentile contour only $\approx$ 50% of the time. This implies that our posteriors are too narrow and should be wider such that the true model parameters will fall in the 60th percentile contour more often, so we are in fact underestimating our errors. We run this inference test at all z considered in this work and found the deviation from the 1-1 line is worse at higher redshifts. See Appendix 3.8 for a discussion of the inference test at $z = 6$. We additionally run the inference test for mock data generated from a multi-variate Gaussian distribution in Appendix 3.9.

There has also been past work trying to correct posteriors that do not pass this coverage probability test (Prangle et al., 2013; Grünwald & van Ommen, 2014; Sellentin & Starck, 2019). In this work, we are using the method of Hennawi et al. in prep. where we can calculate one set of weights for the MCMC draws that broaden the posteriors in a mathematically rigorous way.

A brief description of the reweighting method from Hennawi et al. in prep. is as follows. Consider one data set which gives a corresponding posterior PDF. Initially we have:

$$\int d\hat{x} p(\hat{x}) H(p(\hat{x}) > p_0) = \alpha_0 \tag{3.7}$$

where $p(\hat{x})$ is the PDF of the posterior of some parameters $\hat{x}$, p_0 is a chosen posterior probability, H is the Heaviside function – causing the integrand to be 0 when the probability is less than the given contour p_0, and α_0 is the corresponding volume of the PDF where the probability of $\hat{x}$ is greater than p_0. This means that $\alpha_0 = 1 - C(p_0)$ where C is the cumulative distribution function.

If we instead consider our MCMC chain used to estimate the posterior with N_{samples} points each with probability $1/N_{\text{samples}}$ in the chain we get:

$$\frac{1}{N_{\text{samples}}} \sum_{i}^{N_{\text{samples}}} H(p_i > p_0) = \frac{\#\text{ of samples with } p > p_0}{N_{\text{samples}}} = \alpha_0 \tag{3.8}$$

where the last equality comes from the fact that this sum is a Monte Carlo integral.

Consider the corresponding percentile, P_{true}, of this probability contour. By definition, $C(P_{\text{true}}) = 1 - P_{\text{true}}$ (because the greatest values of the probability correspond to the smallest percentile contours). Thus we have:

$$\frac{1}{N_{\text{samples}}} \sum_{i}^{N_{\text{samples}}} H(p_i > p_0) = P_{\text{true}} \tag{3.9}$$

However, as discussed above, after running an inference test what was thought of as the P_{true} percentile contour is in reality the inferred percentile, P_{inf}, contour. Previous works Sellentin & Starck (2019) suggested relabeling the P_{true} contour as the P_{inf} contour.

However, another method to broaden (or condense) the probability contour is by using a set of weights. Consider re-writing equation (3.9) using weights, w:

$$\frac{1}{N_{\text{samples}}} \sum_{i}^{N_{\text{samples}}} w(x_i) H(p_i > p_0) \approx f(P_{\text{true}}) \tag{3.10}$$

You can then consider multiple values of P_{true} and absorb the factor of $\frac{1}{N_{\text{samples}}}$ into the weights:

$$A\boldsymbol{w} = f(\boldsymbol{P}_{\text{true}}) \tag{3.11}$$

where A is matrix of only 1s and 0s from the Heaviside function, $\boldsymbol{w}$ is the vector of weights, and $\boldsymbol{P}_{\text{true}}$ is the vector of probability contours considered. In fact, we can order the samples from the smallest probability value to the largest probability value such that A is an upper triangular matrix. To guarantee the new weighted probability contours behave as they should statistically (i.e. the true value falls in the P-th percentile contour P% of the time), we set $f(P_{\text{true}}) = P_{\text{inf}}$. This works because P_{inf} is the measured statistically correct percentile contour for this P_{true} value from the previous inference test. Therefore, we can compute weights by:

$$\boldsymbol{w} = A^{-1}\boldsymbol{P}_{\text{inf}}. \tag{3.12}$$

Note that this equation implies that we must run the inference test for the number of probability contours equal to the number of MCMC probability samples we have for each posterior. However, in practice we compute much fewer P_{inf} values during the inference test and then interpolate this vector onto one with the number of MCMC samples we have.

Thus we are able to calculate one set of $(5000 - 500) \times 16 = 72000$ weights that would be used for all 500 mock data sets to broaden the posteriors and pass this inference test. The weights calculated by this method, for a given set of MCMC chains, are unique. The line resulting from the inference test after calculating and using a set of weights is shown in the right panel of Figure 3.9. This line clearly falls along the 1-1 line as expected so our calculated weights allow us to re-weight our posteriors into a statistically correct form. See Appendix 3.8 for a discussion of the re-weighting at $z = 6$.

We show the re-weighted posteriors on $\lambda_{\rm mfp}$ and $\langle F \rangle$ in the bottom right part of Figure 3.8. The weights give greater importance to larger values of $\lambda_{\rm mfp}$ in the tail of values to the right, effectively broadening the posteriors and increasing the errors on the fit. For the mock data set in Figure 3.8 the re-weighted marginalized posterior for $\lambda_{\rm mfp}$ gives $\lambda_{\rm mfp} = 30^{+14}_{-7}$ cMpc whereas before the inferred value was 30^{+10}_{-6} cMpc, so the new errors are $\sim 30\%$ larger. When looking at the 2D distribution in this panel we see that the weights do introduce an additional source of noise to the posterior distribution.

This whole inference procedure is not the optimal and will not give the best constraints on $\lambda_{\rm mfp}$ possible from this statistic. The need to use re-weighting, or some method to correct our posteriors to pass an inference test, comes from our incorrect (though frequently used) assumption of a multivariate Gaussian likelihood. The values of the auto-correlation function at these high z do not sufficiently follow a multivariate Gaussian distribution to justify this assumption, which should be a warning for other studies of the Lyα forest at these z. Using a more correct form of the likelihood (such as a skewed distribution) or likelihood-free inference (such as approximate Bayesian computation as used in Davies et al. (2018b) or other machine learning methods) would lead to more optimal posteriors that better reflect the information in the auto-correlation function. Therefore, future work on this inference procedure will improve the constraints on $\lambda_{\rm mfp}$.

3.4 Results

In order to consider the range of observational constraints possible from one set of $\lambda_{\rm mfp}$ and $\langle F \rangle$ values because of cosmic variance, we study the distribution of measurements

z	Model $\lambda_{\rm mfp}$ (cMpc)	Measured $\lambda_{\rm mfp}$ (cMpc)	
		$R = 8800$	$R = 30000$
5.4	39	40^{+27}_{-9}	32^{+7}_{-5}
5.5	32	35^{+12}_{-6}	33^{+6}_{-4}
5.6	26	28^{+7}_{-4}	27^{+5}_{-3}
5.7	20	22^{+7}_{-4}	20^{+4}_{-3}
5.8	16	18^{+6}_{-4}	16^{+3}_{-3}
5.9	12	14^{+5}_{-4}	13^{+3}_{-3}
6.0	9	12^{+6}_{-3}	11^{+4}_{-2}

Table 3.2: This table contains the results of analyzing the $\lambda_{\rm mfp}$ posteriors for the model value of the auto-correlation function. The mock data at each z has the same value of $\lambda_{\rm mfp}$ as recorded in Table 3.1. The first column contains the modeled value of $\lambda_{\rm mfp}$ at each z that was used in this measurement. The next column contains the resulting measurements at each z for $R = 8800$ data while the last column has the resulting measurements for $R = 30000$ data. In general the trend of the errors is to initially decrease with increasing redshift and then stay about flat beyond $z = 5.7$. This trend follows from the evolution in the true value of $\lambda_{\rm mfp}$ and the data set size at each z.

for 100 mock data sets. For each z we use the $\lambda_{\rm mfp}$ and $\langle F \rangle$ values reported in Table 3.1. Each mock data set is chosen by randomly selecting the appropriate number of skewers given the data set size at each redshift, and averaging the auto-correlation function for each individual skewer. For each mock data set, we perform MCMC as described in Section 3.3.3 and then re-weight the resulting posteriors following Section 3.3.4. Once we have the weights and the chains resulting from our inference procedure we can calculate the marginalized posterior for $\lambda_{\rm mfp}$.

We calculate the marginalized re-weighted posteriors for 100 mock data sets at each z and R. All 100 marginalized re-weighted posteriors are shown as the faint blue lines in Figure 3.10 at $z = 5.4$ for $R = 8800$ (top panel) and $R = 30000$ (bottom panel). In addition to the randomly selected mock data sets, we computed the re-weighted posterior using the model value of the auto-correlation. This is shown as the blue histogram in Figure 3.10. Using the model value as mock data is the ideal case and removes the luck of the draw from affecting the precision of this posterior. The measurement resulting from the model data is written in the blue text of this figure and the values at each z and R are reported in Table 3.2.

Figure 3.10 shows the results from all 100 mock data sets (blue lines) at $z = 5.4$ in order to visualize the various shapes of the resulting re-weighted posteriors. These data sets all have the true values of $\lambda_{\rm mfp} = 39$ cMpc, $\langle F \rangle = 0.0801$, and a 64 quasar data set. The top panel shows the low-resolution $R = 8800$ results and the bottom panel shows the high-resolution $R = 30000$ results. The re-weighted histograms in Figure 3.10 are noisy, much like is seen in the bottom right panel of Figure 3.8. This is a direct consequence of our re-weighting procedure and will be improved with further work on likelihood-free inference. There are roughly two behaviors of posteriors shown here: those that have a large peak at low values and those that are lower limits starting at low value and staying non-zero at our upper boundary of 150 cMpc. For both the lower resolution and higher resolution data, the model values of the auto-correlation function give posteriors with typical widths when compared to the mock data. Both model posteriors also contain the true value of $\lambda_{\rm mfp}$ within their 1σ error bars. Overall, the higher resolution data does produce tighter, more precise posteriors for both the model value and the mock data.

Table 3.2 reports the measurements that result from using the model values of the auto-correlation function as our data. This is an ideal scenario that removes luck of the draw from the resulting measurement. The first column contains the modeled value of $\lambda_{\rm mfp}$ at each z that was used in this measurement, which also appear in Table 3.1. The next column has the resulting measurements for $R = 8800$ data while the last column has the resulting measurement for the $R = 30000$ data. The errors initially decrease with increasing redshift and then stay about the same beyond $z = 5.7$. There are two important factors to consider when looking at this trend. First is the trend of the true value of $\lambda_{\rm mfp}$ with z where $\lambda_{\rm mfp}$ decreases with increasing z. The auto-correlation function is more sensitive to smaller values of $\lambda_{\rm mfp}$ as discussed in Section 3.3.1. Briefly this is due to the fact that smaller $\lambda_{\rm mfp}$ produce greater fluctuations resulting in a larger signal. This means we would expect the results to get more precise and thus have smaller errors at higher z. The other factor is the size of the data set, which is greatest at the lowest z. We would expect the measurements to be less precise and thus have larger errors for the smaller data sets at high z. These effects combine resulting in the trend we see. When comparing the $R = 8800$ and $R = 30000$ measurements, we find that the $R = 30000$ values are on average 40% more precise. Note that it also appears that the measured values of $\lambda_{\rm mfp}$ are always biased high. However, these posterior distributions have tails to greater values of $\lambda_{\rm mfp}$ which causes the reported median measured value of $\lambda_{\rm mfp}$ to be greater than the most likely value of $\lambda_{\rm mfp}$.

In order to visualize the differences between measurements at different redshifts, we plot the results for five random mock data sets with $R = 8800$ in Figure 3.11. Each violin

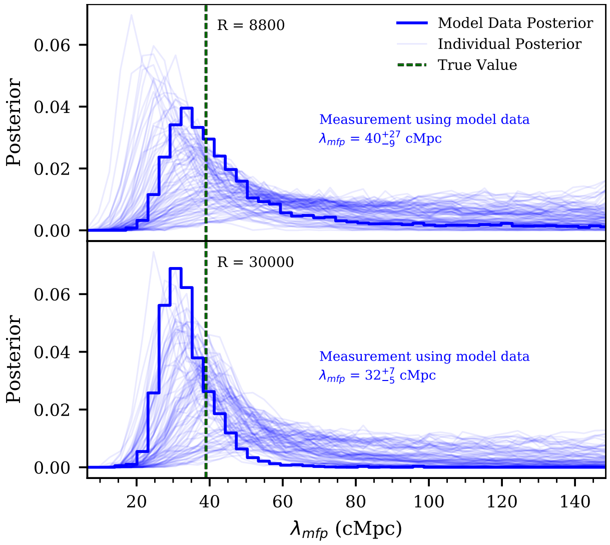

Figure 3.10: This figure shows 100 re-weighted posteriors of λ_{mfp} at $z = 5.4$ with true $\lambda_{\text{mfp}} = 39$ cMpc and $\langle F \rangle = 0.0801$ (blue faint lines). The top panel shows the low-resolution $R = 8800$ results and the bottom panel shows the high-resolution $R = 30000$ results. It also displays the re-weighted posterior (thick blue histogram) from the model's value of the auto-correlaiton function with the measurement of this average posteriors written in blue text. This demonstrates the different possible behaviors the posterior can have from our method. Overall, the higher resolution data does produce more precise posteriors, including the average posterior which is seen in the higher peak and smaller reported errors.

is the re-weighted marginalized posterior for one randomly selected mock data set at the corresponding redshift. The light blue shaded region demarcates the 2.5th and 97.5th percentiles (2σ) of the MCMC draws while the darker blue shaded region demarcates the 16th and 84th percentiles (1σ) of the MCMC draws. The dot dashed line is the double power law, equation (3.1), which we used to determine the true λ_{mfp} evolution as shown in Figure 3.2.

Looking at the posteriors for a given redshift (one column in the figure), the only difference between the posteriors is the random mock data set drawn. This still produces different precision results as seen in Figure 3.10 for $z = 5.4$. There are then three differences between mock data sets shown for a given panel. First is again the mock data is chosen at random so there will be fluctuations in the precision with the luck of the draw. The mock data at different redshift also have different true λ_{mfp} values, shown in the dot-dashed black line, where the smallest λ_{mfp} value is at the highest z. The auto-correlation function is most precise at small inferred λ_{mfp} values which are more likely at the highest z. Additionally, the redshifts each have different data set sizes, as reported in Table 3.1. The highest redshifts have the smallest data set sizes, leading to greater scatter in the precision of the posteriors. Again, the individual posteriors are noisy, resulting from the re-weighting procedure as described in Section 3.3.4.

Here all the mock data sets are at our lower-resolution, $R = 8800$, which was chosen to mimic the existing XQR-30 data. In Appendix 3.10 we discuss the same plot (Figure 3.17) but with the higher resolution, $R = 30000$, data. The only difference between the data used in Figure 3.11 and Figure 3.17 is the resolution of the mock data. The randomly chosen mock data sets, the data set sizes, and the true values are the same.

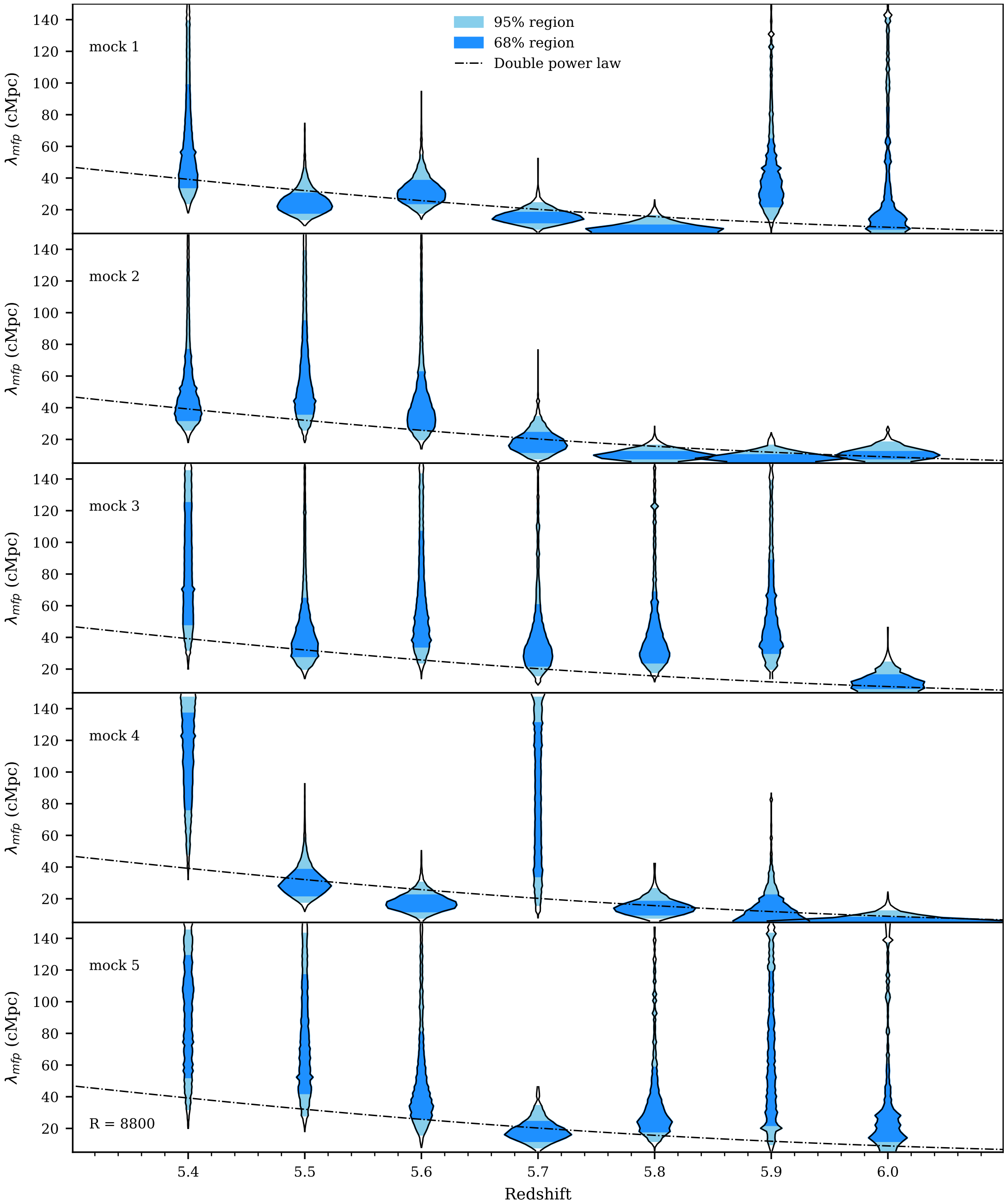

Figure 3.11: Each panel of this figure shows one posterior for a different randomly selected low-resolution ($R = 8800$) mock data set at each z. For each posterior, the light blue shaded region demarcates the 2.5th and 97.5th percentile of the MCMC draws while the darker blue shaded region demarcates the 17th and 83rd percentile of the MCMC draws. The black dot dashed line shows the double power law from equation (3.1) and Figure 3.2. The behavior of each posterior at the different z is determined by the luck of the draw when selecting the mock data, the true $\lambda_{\rm mfp}$ value at each z, and the data set size at each z. The true $\lambda_{\rm mfp}$ values and data set sizes are reported in Table 3.1.

3.5 Conclusions

In this work we have investigated to what precision $\lambda_{\rm mfp}$ can be constrained using the auto-correlation function of Lyα forest flux in quasar sightlines. Overall, we found that the auto-correlation function is sensitive to the value of $\lambda_{\rm mfp}$ across multiple redshift bins and for realistic mock data with both high and low resolution. We computed the marginalized re-weighted posterior for $\lambda_{\rm mfp}$ for 100 mock data sets with properties similar to the XQR-30 extended data set at $5.4 \leq z \leq 6.0$ We additionally considered 100 mock data sets with $R = 30000$, over three times greater than XQR-30 data resolution. The re-weighted posterior showed a variety of behaviors based on the luck of the draw of the mock data chosen, the true value of $\lambda_{\rm mfp}$ for the mock data, and the data set size at each z.

We considered an ideal data set which had the model value of the auto-correlation function, effectively removing the luck of the draw from our measurement. The error on these measurements for both the high resolution and low resolution data initially got smaller (more precise) with increasing redshift then stayed about the same beyond $z = 5.7$. This followed from the changing true value of $\lambda_{\rm mfp}$ and the size of the data set at each z. Small values of $\lambda_{\rm mfp}$ lead to greater fluctuations in the UVB and thus produce an increased signal in the auto-correlation function. This makes the auto-correlation function more sensitive to smaller values of $\lambda_{\rm mfp}$ than larger values of $\lambda_{\rm mfp}$ where the fluctuations are smaller. This work has opened up the possibility for future measurements of $\lambda_{\rm mfp}$ with the auto-correlation function by quantifying the sensitivity of this method.

Of particular interest is the measurement at $z = 6.0$, where recent measurements imply a rapid evolution of $\lambda_{\rm mfp}$. For our ideal model data at $z = 6.0$ with $R = 8800$, we get $\lambda_{\rm mfp} = 12^{+6}_{-3}$ cMpc where the true value we modeled was $\lambda_{\rm mfp} = 9$ cMpc. In comparison, the measurement from Becker et al. (2021) at $z = 6.0$ is $0.75^{+0.65}_{-0.45}$ proper

Mpc (or $5.25^{+4.25}_{-3.15}$ cMpc). Thus our ideal measurement with this new statistical method has comparable error bars as those from Becker et al. (2021). We therefore expect that a measurement using this technique on real data will provide a competitive, secondary check on the value of $\lambda_{\rm mfp}$ at $z = 6.0$. Additionally, we have shown that our method can be applied to multiple fine redshift bins from $5.4 \leq z \leq 6.0$ to precisely constrain the evolution of $\lambda_{\rm mfp}$.

Note that our procedure uses a multi-variate Gaussian likelihood, MCMC, and a set of weights for the MCMC chains that ensures our posteriors pass an inference test. The original failure of our procedure to pass an inference test is likely due to the incorrect assumption that the auto-correlation function follows a multi-variate Gaussian distribution, as discussed in Appendix 3.7. This result should caution against using a multi-variate Gaussian likelihood with other statistics, such as the power spectrum, when making measurements at $z > 5$ as the same issue of non-Gaussian data likely applies. This is especially concerning if the low value of $\lambda_{\rm mfp}$ with high corresponding fluctuations in the UVB at high-z holds true. In the future, better likelihoods or likelihood-free inference will allow for a more optimal inference procedure (see e.g. Davies et al. (2018b) or Alsing et al. (2019)). This will lead to tighter constraints on $\lambda_{\rm mfp}$ from the auto-correlation function.

For this work, we used the method of Davies & Furlanetto (2016b) to generate the UVB boxes as described in section 3.2.1. This assumes a fixed source model which could potentially prove to be incorrect. For example, if fainter galaxies had higher escape fractions it would reduce the strength of UVB fluctuations at fixed $\lambda_{\rm mfp}$, also reducing the auto-correlation signal. This would bias $\lambda_{\rm mfp}$ measurements through this method from real data compared to these models (though it is consistent for our mock data generated from our models). We leave a detailed consideration of the effect of other source models to future work.

Our work also discussed the effect of the current limitations in modeling the UVB and Lyα forest on the auto-correlation function. Namely, our UVB boxes are not correlated with the density of our `Nyx` simulation box, where in reality these quantities are physically correlated. We considered the effect of a correlated UVB in Section 3.3.2. We found that the correlation between high density areas with increased UVB values would reduce the auto-correlation signal for a fixed $\lambda_{\rm mfp}$ on small scales, since higher density leads to reduced transmission. This would again bias a measurement from real data, where these correlations would exist, because the true signal for a given $\lambda_{\rm mfp}$ should be lower than it is in our models, which mimics a model with a larger $\lambda_{\rm mfp}$ value. However, this comparison was done in a small box (40 cMpc h^{-1}) which suppresses UVB fluctuations on all scales as is also discussed in Section 3.3.2. Suppressing fluctuations in the UVB causes the auto-correlation signal to be lower in these boxes. Thus in this comparison the signal is smaller from the density correlations but the UVB fluctuations are also under-estimated due to the box size. The existence of both of these effects means that we were not able to quantify any potential bias from the uncorrelated UVB boxes. The mock data used in this work is generated in the same ways as the models they are compared to, so the measurements here are self-consistent. However, any attempts to compare these models with actual data will need to take into account the effect of using an uncorrelated UVB in the modeling. Thus, future work on UVB models will be necessary before observational $\lambda_{\rm mfp}$ constraints can be produced.

Another potential physical impact on the auto-correlation signal is fluctuation in the temperature of the IGM. Oñorbe et al. (2019) showed that fluctuations in the temperature of the IGM impacted the largest scales of the power spectrum at $z > 5$. We therefore would conclude these fluctuations would also impact the auto-correlation function, which is the Fourier transform of the power spectrum. However Oñorbe et al. (2019) also considered a fluctuating UVB and found that this effectively cancelled out the impact of the thermal fluctuations on the largest scales of the power spectrum. We leave further work on the impact of temperature fluctuations along with UVB fluctuations to future work.

Continuum errors will effect the measurement of the auto-correlation on larger scales which are less important than the small scales when considering λ_{mfp}. The reconstruction done in Bosman et al. (2022) is shown to reconstruct the continuum within 8%. Additionally, Eilers et al. (2017) showed that continuum errors had minimal effect on the shape of the normalized flux PDF at $z = 5$ where transmission is low. We have left a detailed exploration of the effect of continuum errors on the auto-correlation function for future work.

We also note that there is additional $z > 5$ Lyα forest data in telescope archives with lower SNR that could be used in our analysis. Here we limited the consideration to mock XQR-30 data (and a high-resolution analog) but will consider the impact of adding noisier data in future work.

The value of λ_{mfp} and its evolution at high z is important for understanding reionization. Measuring λ_{mfp} at high z is a difficult task that so far has been restricted to two redshift bins at $z > 5$. This work has shown that the auto-correlation function of the Lyα forest flux provides a new, competitive way to constrain λ_{mfp} in multiple redshift bins at $z \geq 5.4$.

3.6 Appendix A: Convergence of the Covariance Matrices

We calculate the covariance matrices for our models with mock draws, as defined in equation (3.4). Using mock draws is inherently noisy and it should converge as $1/\sqrt{N}$

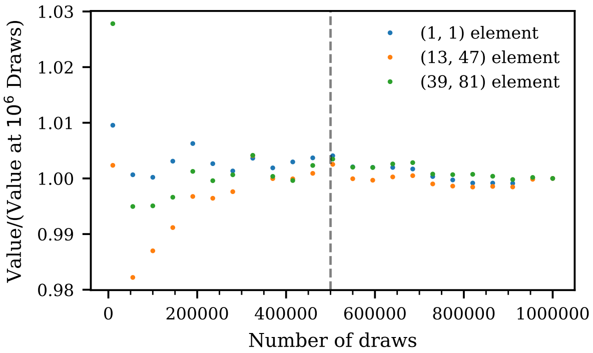

Figure 3.12: This figure shows the behavior of three elements of the model covariance matrix ($z = 6$, $R = 8800$, $\lambda_{\rm mfp} = 9$ cMpc, and $\langle F \rangle = 0.0089$) for different numbers of mock draws. At all values of the number of mocks considered, the covariance elements fall within 3% of their final value. By around ~ 200000 draws, all of the values fall within 1% of the final value. For this reason, we believe using 500000 mock draws is sufficient to generate the covariance matrices used in this study. 500000 mock draws is represented by the vertical dashed black line.

where N is the number of draws used. As stated in the text, we used 500000 mock draws. To check that this number is sufficient to minimize the error in our calculation, we looked at the behavior of elements of one covariance matrix in Figure 3.12. This covariance matrix is for the model with $z = 6$, $R = 8800$, $\lambda_{\rm mfp} = 9$ cMpc, and $\langle F \rangle = 0.0089$, chosen because $z = 6$ has the lowest "true" $\lambda_{\rm mfp}$ value which would lead to the largest fluctuations in the UVB. The values in the plot have been normalized to 1 at 10^6 draws. The three elements have been chosen such that there is one diagonal value and two off-diagonal values in different regions of the matrix. Looking at the correlation matrix in Figure 3.6 (which is for a different model but the qualitative behavior is the same for this model) we see that at all values both on and off the diagonal of the correlation matrix are high and positive, so we expect the convergence for all elements to be roughly the same. At all values of the number of mock draws considered, the covariance elements fall within 3% of their final value. By around ~ 200000 draws, the values fall within 1% of the final value. For this reason, we believe using 500000 mock draws is sufficient to generate the covariance matrices used in this study. In Figure 3.12, 500000 mock draws is represented by the vertical dashed black line.

3.7 Appendix B: Non-Gaussian distribution of the values of the auto-correlation function

For our inference, we used the multi-variate Gaussian likelihood defined in equation (3.6). This functional form assumes that the distribution of mock draws of the auto-correlation function is Gaussian distributed about the mean for each bin. In order to visually check this we will look at the distribution of mock draws from two bins of the auto-correlation function for two different models.

Both Figures 3.13 and 3.14 show the distribution of 1000 mock data sets from the velocity bins of the auto-correlation function with $\Delta v = 85.0\,\mathrm{km\,s^{-1}}$ and $\Delta v = 289.0\,\mathrm{km\,s^{-1}}$. The bottom left panels show the 2D distribution of the auto-correlation values from these bins. The blue (green) ellipses represents the theoretical 68% (95%) percentile contour calculated from the covariance matrix calculated for each model from equation (3.4). The red crosses shows the calculated mean. The top panels show the distribution of only the $v = 289.0\,\mathrm{km\,s^{-1}}$ bins while the right panels show the distribution of only the $v = 85.0\,\mathrm{km\,s^{-1}}$ bins.

Figure 3.13 shows mock values of two bins of the auto-correlation function for the model at $z = 5.4$ with $R = 8800$, $\lambda_{\mathrm{mfp}} = 39$ cMpc and $\langle F \rangle = 0.0801$. These mock data sets consist of 64 quasar sightlines of length $\Delta z = 0.1$. Both the 1D and 2D distributions seem relatively well described by Gaussian distributions by eye though they do show some evidence of non-Gaussian tails to larger values. The number of points falling in each contour both fall within 2% of the expected values. In the bottom left panel with the 2D distribution there are more mock values falling outside the 95% contour to the top right (higher values) than in any other direction. For this reason the distribution is not exactly Gaussian but a Gaussian visually appears as an acceptable approximation.

Figure 3.14 shows mock values of two bins of the auto-correlation function for the model at $z = 6$ with $R = 8800$, $\lambda_{\mathrm{mfp}} = 9$ cMpc, and $\langle F \rangle = 0.0089$. These mock data sets consist of 19 quasar sightlines of length $\Delta z = 0.1$. In both the top and right panels, which show the distribution of values for one bin of the auto-correlation function, it is clear that the distribution of mock draws is skewed and a Gaussian is not a good approximation for the distributions. This is quantified by the percent of points in the two ellipses from the bottom left panel labeled in the top right with 79.0% of the mock draws falling within the 68% contour and 92.2% of the mock draws falling within the 95% contour. The points outside of the contours are highly skewered towards the top right (higher values). It is only possible for the auto-correlation function to be negative due to noise, which generally averages to very small values approaching zero at the non-zero lags of the auto-correlation function. However real fluctuations in the UVB cause the positive fluctuations, making them much more likely and cause the resulting skewed distribution at high z where the overall signal is closer to zero.

Figures 3.13 and 3.14 show the changing distribution of the auto-correlation value with λ_{mfp}, $\langle F \rangle$, and mock data set size. There is a greater deviation from a multi-variate Gaussian distribution at higher z. It is possible that adding additional sightlines will cause the auto-correlation function to better follow a multi-variate Gaussian distribution due to the central limit theorem, though investigating this in detail is beyond the scope of the paper. The incorrect assumption of the multi-variate Gaussian likelihood thus contributes to the failure of our method to pass an inference test as discussed in Section 3.3.4 for $z = 5.4$ and Appendix 3.8 for $z = 6$. For our final constraints, we calculated weights for our MCMC chains such that the resulting posteriors do pass our inference test, as discussed in Section 3.3.4. The whole method of assuming a multi-variate Gaussian then re-weighting the posteriors in non-optimal and future work using a more correct likelihood or likelihood-free inference will improve our results.

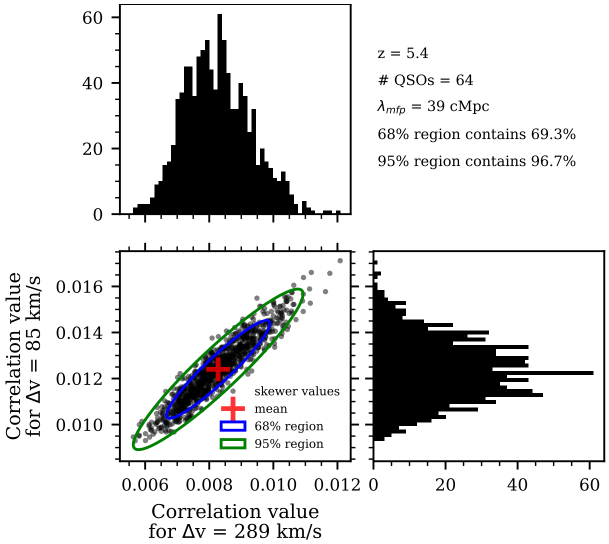

Figure 3.13: This figure shows the distribution 1000 mock draws from two bins of the auto-correlation function ($\Delta v = 85.0\,\mathrm{km\,s^{-1}}$ and $\Delta v = 289.0\,\mathrm{km\,s^{-1}}$) for one model ($z = 5.4$, $R = 8800$, $\lambda_{\mathrm{mfp}} = 39$ cMpc, and $\langle F \rangle = 0.0801$). The top panel shows the distribution of only the $\Delta v = 289.0\,\mathrm{km\,s^{-1}}$ bin while the right panel shows the distribution of only the $\Delta v = 85.0\,\mathrm{km\,s^{-1}}$ bin. The blue (green) circle represents the 68% (95%) ellipse calculated from the covariance matrix calculated for this model from equation (3.4). The red plus shows the calculated mean. Additionally the percent of mock draws that fall within each of these contours is written in the top right. Both the 1D and 2D distributions seem relatively well described by a Gaussian distribution. In the 2D plot, there are more points outside the 95% contour to the top right than on any other side but it is not extreme.

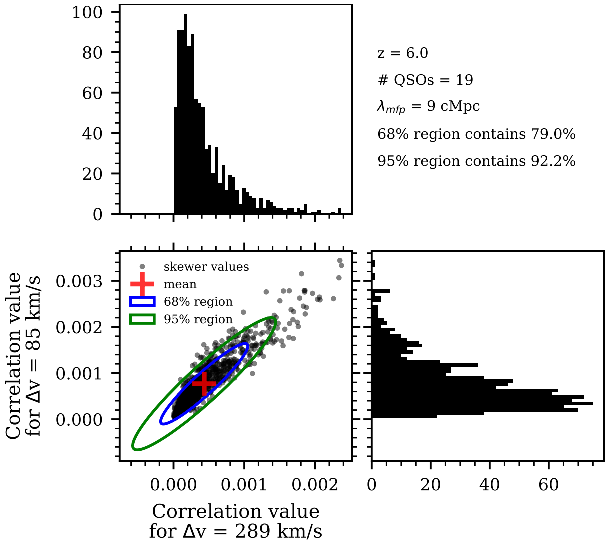

Figure 3.14: This figure shows the distribution 1000 mock draws from two bins of the auto-correlation function ($\Delta v = 85.0\,\mathrm{km\,s^{-1}}$ and $\Delta v = 289.0\,\mathrm{km\,s^{-1}}$) for one model ($z = 6$, $R = 8800$, $\lambda_{\mathrm{mfp}} = 9$ cMpc, and $\langle F \rangle = 0.0089$). The top panel shows the distribution of only the $\Delta v = 289.0\,\mathrm{km\,s^{-1}}$ bin while the right panel shows the distribution of only the $\Delta v = 85.0\,\mathrm{km\,s^{-1}}$ bin. The blue (green) circle represents the 68% (95%) ellipse calculated from the covariance matrix calculated for this model from equation (3.4). The red plus shows the calculated mean. Additionally the percent of mock draws that fall within each of these contours is written in the top right. Both the 1D and 2D distributions do not seem to be well described by a Gaussian with 79.0% of the mock draws falling within the 68% contour and 92.2% of the mock draws falling within the 95% contour.

3.8 Appendix C: Inference test at high redshift

Here we present the results of the inference test at $z = 6$. This calculation was done following the procedure described in Section 3.3.4. Figure 3.15 shows the results for $z = 6$ and can be compared to the $z = 5.4$ results in Figure 3.9. The left panel here shows the initial coverage plot which deviates greatly from the expected $P_{\text{inf}} = P_{\text{true}}$ line, much more so than the $z = 5.4$. This likely comes from a greater deviation from the assumption of a multi-variate Gaussian likelihood as described in Appendix 3.7. The $z = 6$ mock data show highly skewed distributions that are not well described by a Gaussian likelihood.

This initial coverage plot only ever reaches a value of $P_{\text{inf}} \approx 0.8$, which becomes an issue for the re-weighting. In the right panel of Figure 3.15 the re-weighted inference line thus still only able to reach $P_{\text{inf}} \approx 0.8$ creating a plateau in the line once it reaches this value. One way to reach higher values is to increase the number of steps in the MCMC chain. We tried to triple the number of steps but did not see much improvement in the inference test. For computational reasons we stick with the numbers used at other redshifts resulting in 72000 total steps as described in Section 3.3.4. This plateau at $P_{\text{true}} = 0.8$ means that our 1-σ (68th percentile) contours are robust but our 2-σ (95th percentile) contours are underestimated since we can only correct up to $\sim$ 80th percentile.

The inference lines at other redshifts are available upon request. For $5.4 \leq z \leq 5.8$ the coverage plots after re-weighting do not plateau, like the re-weighted coverage plot shown in Figure 3.9. Both the $z = 6.0$ and the $z = 5.9$ coverage plots plateau after re-weighting, like that in Figure 3.15. This means our re-weighted posteriors at $z = 5.9$ and $z = 6$ may still need additional work to further enlarge probability contours above the value of the plateau.

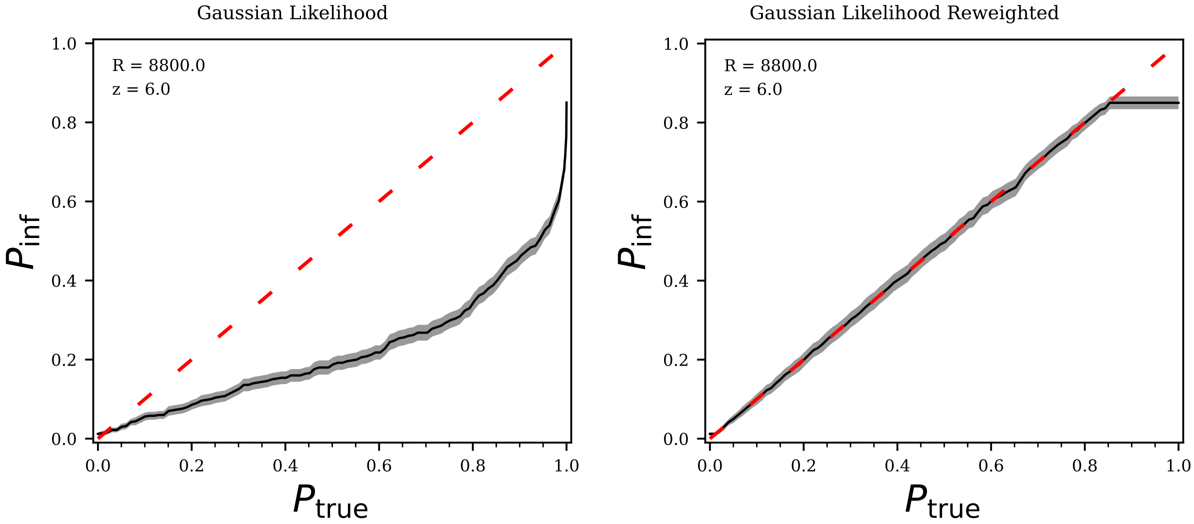

Figure 3.15: The left panel of this figure shows the coverage plot resulting from the inference test from 500 models at $z = 6$ and $R = 8800$ drawn from our priors on $\lambda_{\rm mfp}$ and $\langle F \rangle$. Here we see that the true parameters for the models fall above the 60th percentile in the MCMC chain $\sim 20\%$ of the time, for example. The right panel of this figure shows the coverage plot resulting from the inference test with the use of one set of weights to re-weight the posteriors. With these weights the true parameters for the models fall on the $P_{\rm inf} = P_{\rm true}$ line up to $P_{\rm true} \sim 0.8$. This is because the original coverage plot was only able to reach $P_{\rm inf} \sim 0.8$ so our re-weighting could only match up to this value.

3.9 Appendix D: Gaussian data inference test

As shown in Appendix 3.7, the distribution of mock values of the auto-correlation function is not exactly Gaussian distributed. In order to confirm the failure of our mock data to pass an inference test (as discussed in Section 3.3.4 and Appendix 3.8) comes from the use of a multi-variate Gaussian likelihood, we generate Gaussian distributed data and run inference tests. For one value of $\lambda_{\rm mfp}$ and $\langle F \rangle$, we randomly generate a mock data set from a multi-variate Gaussian with the given mean model and covariance matrix that we calculated for our mock data in Section 3.3.1. We can then continue with the inference test as described in Section 3.3.4. The results for this inference test for $z = 5.4$ and $z = 6.0$ (both with $R = 8800$) are shown in Figure 3.16. Here both redshifts inference lines fall along the 1-1 line that is expected for all probability contour, $P_{\rm true}$,

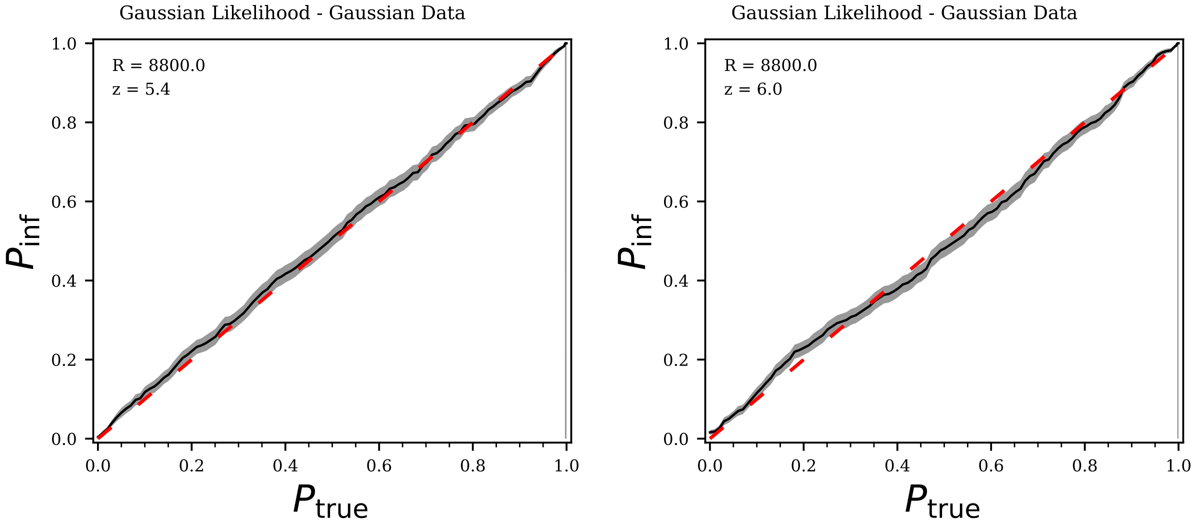

Figure 3.16: Both panels of this figure shows the coverage plot resulting from the inference test from 500 data sets generated by randomly drawing points from the mean model and covariance matrix. The the means and covariance matrices used come from $z = 5.4$ and $R = 8800$ in the left panel and $z = 6.0$ and $R = 8800$ in the right panel. The true parameter values for both panels were drawn from our priors on $\lambda_{\rm mfp}$ and $\langle F \rangle$. In both panels, the Gaussian mock data produced inference lines that fall on top of the 1-1 line within errors, as expected for the statistically correct posteriors.

values. This behavior is also seen at the other redshifts and $R = 30000$. The fact that perfectly Gaussian data passes an inference test with the same likelihood, priors, and method as was used on mock data confirms that the failure of our mock data to pass an inference test is due to the non-Gaussian distribution of the mock data.

3.10 Appendix E: High-resolution results

In section 3.4 we only show the posteriors for multiple mock data sets at different redshifts for $R = 8800$. Here we present the same results but for mock data with $R = 30000$. Each violin plot in Figure 3.17 is the re-weighted marginalized posterior for one randomly selected mock data set at the corresponding redshift. The light red shaded region demarcates the 2.5th and 97.5th percentiles (2σ) of the MCMC draws while the darker red shaded region demarcates the 16th and 84th percentiles (1σ) of the MCMC draws. Beneath the red violins are blue violins for the posteriors for the same data with $R = 8800$ as shown in Figure 3.11. The dot dashed line is the double power law, equation (3.1), which we used to determine the true λ_{mfp} evolution as shown in Figure 3.2. The random mock data selected for this figure matches exactly with the random mock data used to make Figure 3.11. The only difference between the data used in these two figures is the resolution. Generally, the posteriors from the $R = 30000$ data shown in Figure 3.17 are more precise than those from the $R = 8800$ data.

Again, looking at the posteriors for a given redshift (one column in the figure), the only difference between the posteriors is the random mock data set drawn. These results still have varying precision as is expected from luck of the draw with the mock data sets. There are then three differences between mock data sets shown for a given panel. First is the same as the difference between mocks at one redshift: the mock data is chosen at random so there is just the luck of the draw. The mock data at each redshift also vary with the true λ_{mfp} value, shown in the dot-dashed black line, where the smallest λ_{mfp} value is at the highest z. The auto-correlation function is most precise at small inferred λ_{mfp} values which are more likely at the highest z. Additionally, the redshifts each have different data set sizes, as reported in Table 3.1. The highest redshifts have the smallest data set sizes, leading to greater scatter in the precision of the posteriors. Again, the individual posteriors are noisy, resulting from the re-weighting procedure as described in Section 3.3.4.

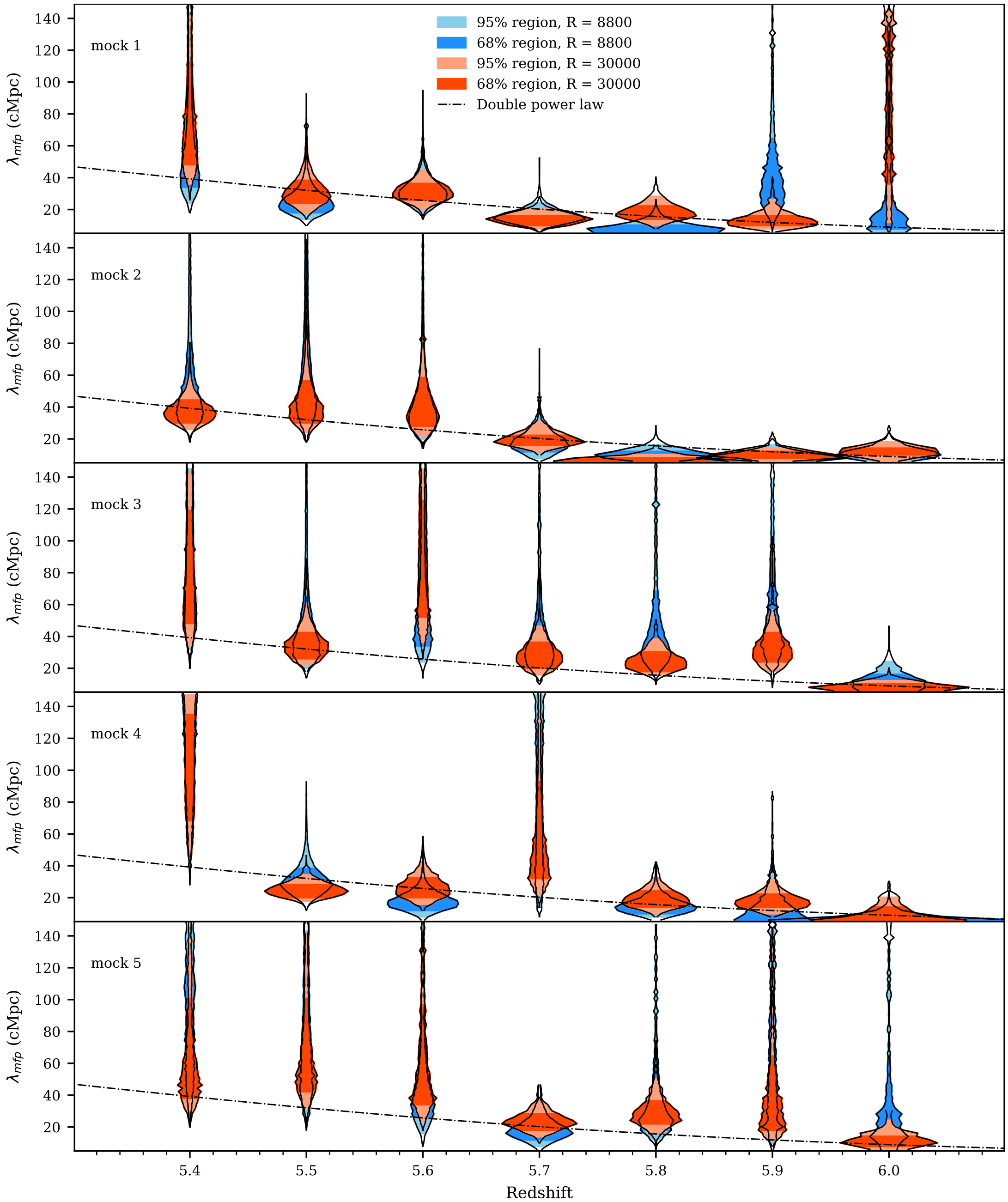

Figure 3.17: Each panel of this figure shows one posterior for a different randomly selected high-resolution ($R = 30000$) mock data set at each z in shades of red. Note that the low-resolution ($R = 8800$) mock data posteriors are plot below the high resolution posteriors in blue. For each posterior, the light red shaded region demarcates the 2.5th and 97.5th percentile of the MCMC draws while the darker red shaded region demarcates the 16th and 84th percentile of the MCMC draws. The black dot dashed line shows the double power law from equation (3.1) and Figure 3.2. The behavior of each posterior at the different z is determined by the luck of the draw when selecting the mock data, the true λ_{mfp} value at each z, and the data set size at each z. The true λ_{mfp} values and data set sizes are reported in Table 3.1.

Chapter 4

Forecasting constraints on the high-z IGM thermal state from the Lyman-α forest flux auto-correlation function

This chapter was reproduced from Wolfson et al. (2023a) with only minor changes to fit the formatting of this dissertation. I'd like to thank my coauthors, without whom this work would not have been possible: Joseph F. Hennawi, Frederick B. Davies, Zarija Lukić, and Jose Oñorbe.

4.1 Introduction

Understanding the epoch of reionization, the time period where the first luminous sources emitted photons that re-ionized the intergalactic medium (IGM), remains a major open problem for studies of the early universe. The midpoint of reionization has been constrained as $z_{\rm re} = 7.7 \pm 0.7$ from the cosmic microwave background (Planck Collaboration et al., 2020). Initial measurements of transmission in the Lyman-α (Lyα) forest (Gunn & Peterson, 1965; Lynds, 1971) of high redshift quasars suggested that reionization was complete by $z \sim 6$ (Fan et al., 2006; McGreer et al., 2011, 2015). Additional methods used to constrain reionization include observations of Lyα emission from high redshift galaxies (see e.g. Jung et al., 2020; Morales et al., 2021) and large Lyα absorption troughs (see e.g. Becker et al., 2018; Kashino et al., 2020). Measurements of the Lyα forest optical depths scatter on levels that suggest reionization is not actually complete until $z < 6$ (Fan et al., 2006; Becker et al., 2015; Bosman et al., 2018; Eilers et al., 2018; Yang et al., 2020; Bosman et al., 2022).

An alternative, indirect method to constrain reionization is by looking at the thermal history of the IGM at $z > 5$ (Boera et al., 2019; Walther et al., 2019; Gaikwad et al., 2021). During reionization, ionization fronts propagate through the IGM and impulsively heat the reionized gas in the IGM to $\sim 10^4$ K (McQuinn, 2012; Davies et al., 2016; D'Aloisio et al., 2019). The details of the driving sources, the timing, and duration of reionization will determine the precise amount of heat injected. After reionization, the IGM expands and cools through the adiabatic expansion of the universe and inverse Compton scattering off CMB photons. The combination of these physical processes will allow the IGM gas to relax into a state described by a tight power-law relation between the temperature and density:

$$T = T_0 \Delta^{\gamma-1}. \tag{4.1}$$

Where $\Delta = \rho/\bar{\rho}$ is the overdensity, $\bar{\rho}$ is the mean density of the Universe, T_0 is the temperature at mean density, and γ is the slope of the relationship (Hui & Gnedin, 1997; Puchwein et al., 2015; McQuinn & Upton Sanderbeck, 2016). The low-density IGM has long cooling times, so the thermal memory of reionization will persist for hundreds of Myr. This means that thermal state of the IGM at the end and after reionization, $z \sim 5-6$, can provide key insights into reionization (Miralda-Escudé & Rees, 1994; Hui & Gnedin, 1997; Haehnelt & Steinmetz, 1998; Theuns et al., 2002a; Hui & Haiman, 2003; Lidz & Malloy, 2014; Oñorbe et al., 2017a,b).

The Lyα optical depth, $\tau_{\mathrm{Ly}\alpha}$ is related to the temperature via

$$\tau_{\mathrm{Ly}\alpha} = n_{\mathrm{HI}}\sigma_{\mathrm{Ly}\alpha} \propto T^{-0.7}/\Gamma_{\mathrm{HI}}, \tag{4.2}$$

see Rauch (1998). Thus, several statistics have been used to measure the thermal state of the IGM from the Lyα forest, including the flux probability density (Becker et al., 2007; Bolton et al., 2008; Viel et al., 2009; Calura et al., 2012; Lee et al., 2015), the curvature (Becker et al., 2011; Boera et al., 2014; Gaikwad et al., 2021), the Doppler parameter distribution (Schaye et al., 1999, 2000; Ricotti et al., 2000; Bryan & Machacek, 2000; McDonald et al., 2001; Rudie et al., 2012; Bolton et al., 2010, 2012, 2014; Rorai et al., 2018; Gaikwad et al., 2021), the joint distribution of the Doppler parameters with the Hydrogen Column Density (Hiss et al., 2018), and wavelets (Lidz et al., 2010; Garzilli et al., 2012; Gaikwad et al., 2021). One of the most commonly used statistics for measuring the structure of the Lyα forest is the 1D flux power spectrum, $P_F(k)$ (Theuns et al., 2000; Zaldarriaga et al., 2001; Yèche et al., 2017; Walther et al., 2018; Boera et al., 2019; Gaikwad et al., 2021; Wolfson et al., 2021).

The thermal state of the IGM significantly influences the Lyα forest, primarily through two mechanisms: Doppler broadening, which is driven by thermal motions, and Jeans (pressure) smoothing, which affects the distribution of the underlying baryons. To understand Jeans smoothing, it's crucial to consider the role of pressure forces. Pressure forces, influenced by the thermal state, erase gravitational fluctuations at a rate determined by the local sound speed. At low densities, like those of the IGM, this sound-crossing time is approximately the Hubble time. Thus, the Jeans (pressure) smoothing scale serves as a record of the thermal history of the IGM over extensive timescales (Gnedin & Hui, 1998; Kulkarni et al., 2015; Nasir et al., 2016; Oñorbe et al., 2017a,b; Rorai et al., 2017). Both Doppler broadening and Jeans smoothing reduce the small-scale structure of the Lyα forest. These reductions in small-scale structure of the Lyα forest lead to a cut-off in $P_F(k)$ at high-k.

An alternative to the power spectrum is the Lyα forest flux auto-correlation function, which is the Fourier transform of the power spectrum. In this work we will explore the ability of the auto-correlation function to constrain the thermal state of the IGM at $z > 5$. The auto-correlation function of the Lyα forest has two statistical properties that make it easier to work with than the power spectrum. First is that uncorrelated noise (which is the expectation for astronomical spectrograph noise) will not impact non-zero lags of the auto-correlation function, as it will average to zero. Thus there is no need to account for uncorrelated noise with the auto-correlation function. For the power spectrum, uncorrelated noise is a constant positive value at all scales. Thus the unknown noise level must be calculated and subtracted from power spectrum measurements which will add additional uncertainty to the final measurement. Additionally, observational quasar spectra often have regions that need to be removed (e.g. for metal lines). Masking out these and other regions introduces a complicated window function to the power spectrum that must be corrected for (see e.g. Walther et al., 2019) and will again increase the uncertainty in the measurement. The auto-correlation function does not require a similar correction since masking will only change the number of pixel pairs used at a given velocity lag.

Many previous studies have measured the Lyα forest flux auto-correlation function at lower redshifts for a wide range of applications (McDonald et al., 2000; Rollinde et al., 2003; Becker et al., 2004; D'Odorico et al., 2006). In addition, the first measurement of the Lyα forest flux auto-correlation function at $z > 5$ was presented in Wolfson et al.

(2023c) for moderate resolution quasar spectra.

In this work we will investigate constraints on T_0 and γ that can be achieved from measurements of the Lyα forest flux auto-correlation function. We will do this by creating mock observational measurements of the auto-correlation function and comparing to model values of the auto-correlation function determined via semi-numerical methods applied to hydrodynamical simulations. By applying Bayesian statistics to this setup we will get mock posterior distributions for T_0 and γ.

Beyond a thermal state that follows a tight power law described by T_0 and γ, reionization can lead to significant fluctuations in the temperature of the IGM (D'Aloisio et al., 2015; Davies et al., 2018a). At the same time, fluctuations in the ultraviolet background (UVB) arise during reionization because the ionizing photons produced will be absorbed by the remaining neutral hydrogen at short distances from their initial sources (Davies & Furlanetto, 2016b; Gnedin et al., 2017; D'Aloisio et al., 2018). These distances are characterized by the mean free path of ionizing photons, $\lambda_{\rm mfp}$ (Mesinger & Furlanetto, 2009). Various previous studies have investigated the effect of large scale variations in the UVB on the auto-correlation function and power spectrum of the Lyα forest (Zuo, 1992a,b; Croft, 2004; Meiksin & White, 2004; McDonald et al., 2005; Gontcho A Gontcho et al., 2014; Pontzen, 2014; Pontzen et al., 2014; D'Aloisio et al., 2018; Meiksin & McQuinn, 2019; Oñorbe et al., 2019). In particular, Wolfson et al. (2023b) showed that the positive fluctuations in the UVB that accompany small $\lambda_{\rm mfp}$ values boost the flux of the Lyα forest on small scales, which can be detected in the auto-correlation function.

We will use an additional hydrodynamical simulation that models fluctuations in both the temperature and the UVB to determine the effect on the Lyα forest flux auto-correlation function. In addition to looking at the qualitative differences between these models, we will quantify the likelihood ratio between different models for mock data sets. This provides a quantitative way to discuss constraints on a discrete set of models.

The structure of this paper is as follows. We discuss our grid of simulations that vary T_0 and γ in Section 4.2. The auto-correlation function and our other statistical methods to constrain these parameters are described in Section 4.3 with our results being discussed in Section 4.3.4. We discuss our second set of simulations for models of the IGM with temperature and UVB fluctuations in Section 4.4 and use the auto-correlation function to quantitatively distinguish between these models in Section 4.4.3. Finally, we summarize in Section 4.5.

4.2 Simulation Data

4.2.1 Simulation box

In this work we use a simulation box of size $L_{\rm box} = 100$ comoving Mpc (cMpc) h^{-1} run with `Nyx` code (Almgren et al., 2013). `Nyx` is a hydrodynamical simulation code that was designed for simulating the Lyα forest with updated physical rates from Lukić et al. (2015). The simulation has with 4096^3 dark matter particles and 4096^3 baryon grid cells. It is reionized by a Haardt & Madau (2012) uniform UVB that is switched on at $z \sim 15$. We have two snapshots of this simulation at $z = 5.5$ and $z = 6.0$. We consider models at seven redshifts: $5.4 \leq z \leq 6.0$ with $\Delta z = 0.1$. In order to consider the redshifts for which we do not have a simulation output, we select the nearest snapshot and use the desired redshift when calculating the proper size of the box and the mean density. This means we use the density fluctuations and velocities directly from the nearest `Nyx` simulation output. As a test, we used the $z = 6.0$ simulation snapshot to generate skewers at $z = 5.7$ and found no significant change in our finally results, thus the nearest grid point interpolation between snapshot redshifts is sufficient.

We generate grids of thermal models through a semi-numerical method to set the temperature along the sightlines. For each value of T_0 and γ, we set the temperature of each cell following Equation (4.1) for all densities with no cutoff. Our method does not take into account the full evolution of the thermal state of the IGM, only the instantaneous temperature. This simple model is sufficient to achieve the aim of this paper, which is to see if the auto-correlation function is sensitive to the thermal state. To make our grid we use 15 values of T_0 and 9 values of γ resulting in 135 different combinations of these parameters at each z. The values of T_0 and γ in our grid of thermal models were chosen based on the current models and available data, as shown in Figure 4.1. We generate a model for the evolution of the thermal state of the IGM by a method similar to Upton Sanderbeck et al. (2016) with $z_{\rm reion} = 7.7$, $\Delta T = 20,000$ K, and $\alpha_{\rm UVB} = 1.5$. For more information on the calculation of the temperature field see Davies et al. (2018 a). We select central "true" T_0 and γ values at each redshift from this model, which are shown as black points in Figure 4.1 and listed in Table 4.1. At all z, we use the errors on the measurements reported in Gaikwad et al. (2021) at $z = 5.8$ ($\Delta T_0 = 2200\,\rm K$ and $\Delta\gamma = 0.22$) and modeled from $T_0 - 4\Delta T_0$ to $T_0 + 4\Delta T_0$ and $\gamma - 4\Delta\gamma$ to $\gamma + 4\Delta\gamma$ in linear bins.

Our simulations don't predict the overall average of the UVB, $\langle\Gamma_{\rm UVB}\rangle$, because this value originates from complicated galaxy physics that are not included in the simulations. In addition our method of post-processing different thermal states would affect the resulting $\langle\Gamma_{\rm UVB}\rangle$. Instead we choose to model a variety of potential $\langle\Gamma_{\rm UVB}\rangle$ values through the mean transmitted flux, $\langle F\rangle$. This is done be re-scaling the optical depths along the skewer, τ, such that $\langle e^{-\tau}\rangle = \langle F\rangle$ when averaging the transmitted flux over all skewers. These $\langle F\rangle$ model values are centered on the values presented in Bosman et al. (2022) for each redshift bin. We chose a range of models spanning $4\Delta\langle F\rangle$ where the $\Delta\langle F\rangle$ is the redshift dependent value reported in Bosman et al. (2022). These choices of $\langle F\rangle$ are listed in the last column of Table 4.1.

Table 4.1: This table lists the central "true" values of the redshift-dependent thermal state models used in this work. The last column states the central "true" value of $\langle F \rangle$ modeled in this work, which are the measurements from Bosman et al. (2022).

z	T_0 (K)	γ	$\langle F \rangle$
5.4	9,149	1.352	0.0801
5.5	9,354	1.338	0.0591
5.6	9,572	1.324	0.0447
5.7	9,804	1.309	0.0256
5.8	10,050	1.294	0.0172
5.9	10,320	1.278	0.0114
6.0	10,600	1.262	0.0089

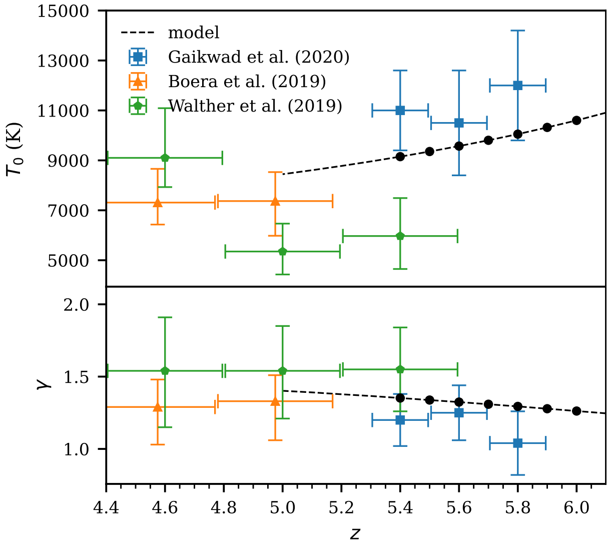

Figure 4.1: The blue squares, orange pentagons, and green triangles show previous measurements of T_0 and γ at high z from Gaikwad et al. (2021), Boera et al. (2019), and Walther et al. (2019) respectively. The dashed line shows the results for a thermal evolution model calculated with methods similar to Upton Sanderbeck et al. (2016) and Davies et al. (2018a) This model has $z_{\rm reion} = 7.7$, $\Delta T = 20,000$ K, and $\alpha_{\rm UVB} = 1.5$. We use this model as our "true" redshift evolution for T_0 and γ in this work. The chosen models are shown as black circles.

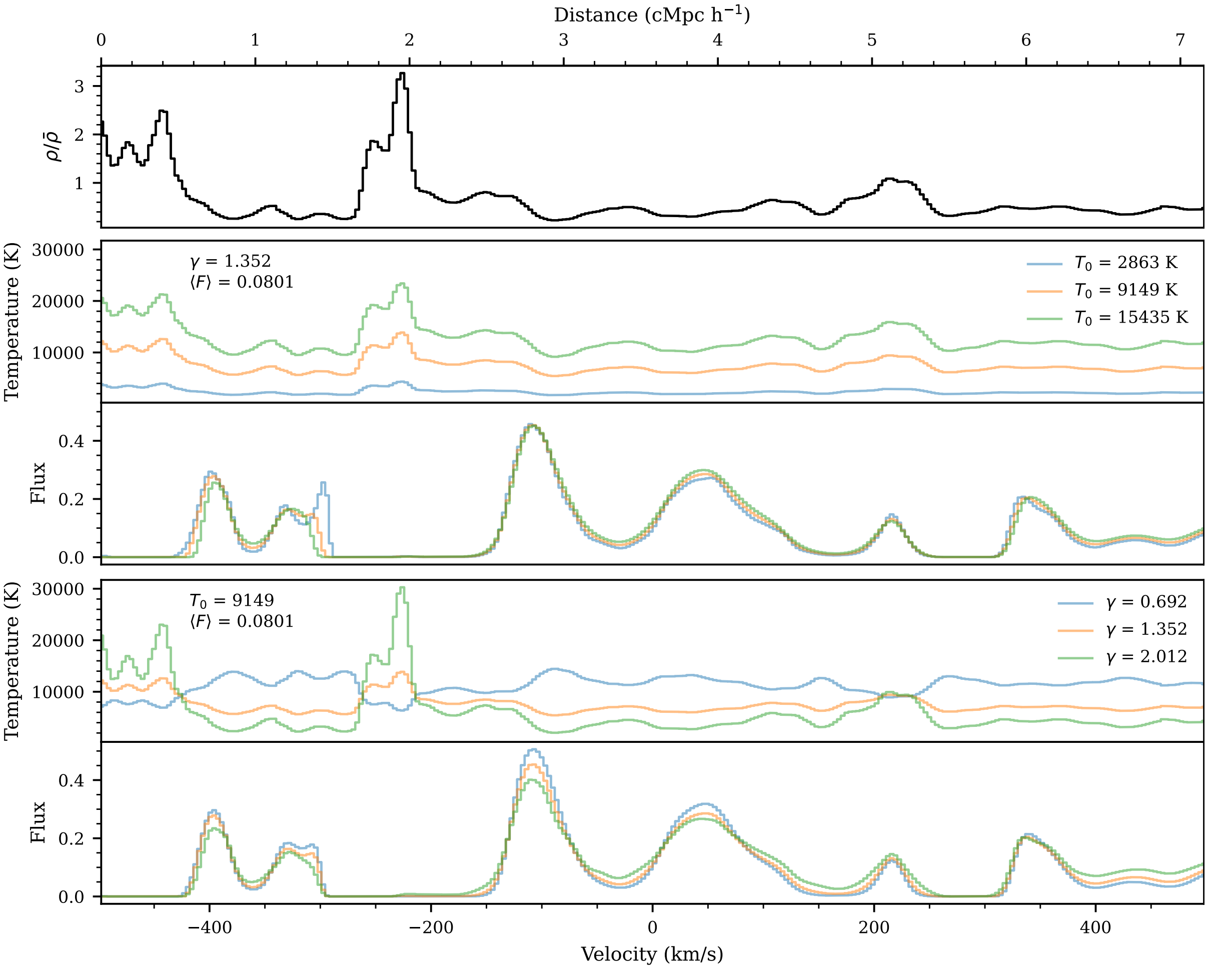

Figure 4.2: The top panel shows the density along a section of one skewer in black for $z = 5.4$. There are then two pairs of panels each depicting the temperature (top) and flux (bottom) along this skewer. The first pair varies T_0 with constant $\gamma = 1.352$ and $\langle F \rangle = 0.0801$. Shifting T_0 causes a corresponding shift in the temperature values along the skewer. Hotter temperatures (orange and green) smooths the flux, as seen clearly in the loss of a second transmission spike at $z \sim -300\,\mathrm{km\,s^{-1}}$. The second pair varies γ with constant $T_0 = 9148\,\mathrm{K}$ and $\langle F \rangle = 0.0801$. When $\gamma > 1$ (orange and green) the temperature is directly proportional to the density fluctuations while $\gamma < 1$ (blue) causes the temperature to be inversely proportional to the density fluctuations. When temperature is inversely proportional to density, lower densities have higher temperatures. Low densities and higher temperatures will locally increase the flux so the $\gamma < 1$ (blue) model will lead to transmission spikes with the greatest flux, as seen at $v \sim -100\,\mathrm{km\,s^{-1}}$.

We draw 1000 skewers from the simulation box. One example skewer at $z = 5.4$ for different T_0 and γ models is shown in Figure 4.2. The top panel shows the density of this skewer for all models in black. There are then two pairs of panels each depicting the temperature (top) and flux (bottom) along this skewer. The second and third panels vary T_0 with constant $\gamma = 1.352$ and $\langle F \rangle = 0.0801$. The coldest model, $T_0 = 2863\,\mathrm{K}$ (blue), has some of the sharpest features. This is seen at $v \sim -300\,\mathrm{km\,s^{-1}}$ where the low T_0 (blue) model has a secondary sharp peak in the flux. In comparison the hottest model, $T_0 = 15\,435\,\mathrm{K}$ (green), has one wider transmission spike. In addition, increasing T_0 decreases $\tau_{\mathrm{Ly}\alpha}$ as described in Equation (4.2), which in turn increases the transmitted flux. For this reason we get the greatest transmission from the $T_0 = 15435$ K (green) model, seen in the transmission spike at $v = 50\,\mathrm{km\,s^{-1}}$. With fixed $\langle F \rangle$ this leads to greater variation in the flux for higher T_0 models.

The fourth and fifth panels vary γ with constant $T_0 = 9149$ K and $\langle F \rangle = 0.0801$. When $\gamma > 1$ (orange and green) the temperature is directly proportional to the density fluctuations while $\gamma < 1$ (blue) causes the temperature to be inversely proportional to the density fluctuations. When temperature is inversely proportional to density, lower densities have higher temperatures. Low densities and higher temperatures will locally increase the flux so the $\gamma < 1$ (blue) model will lead to transmission spikes with the greatest flux, as seen at $v \sim -100\,\mathrm{km\,s^{-1}}$.

4.2.2 Forward Modeling

In order to mimic realistic high-resolution observational data from echelle spectrographs, (e.g. from Keck/HIRES, VLT/UVES, and Magellan/MIKE) we forward model our ideal simulation skewers to have imperfect resolution and flux levels. We consider a resolution of $R = 30000$ and a signal to noise ratio per $10\,\mathrm{km\,s^{-1}}$ pixel (SNR_{10}) of

$\mathrm{SNR}_{10} = 30$ at all redshifts.

We model this resolution by smoothing the flux by a Gaussian filter with FWHM $= 10\,\mathrm{km\,s^{-1}}$. After smoothing we re-sampled the new flux such that the new pixel size was $\Delta v = 2.5\,\mathrm{km\,s^{-1}}$. With this pixel scale, $\mathrm{SNR}_{10} = 30$ corresponds to a signal to noise ratio of the pixel size ($\mathrm{SNR}_{\Delta v}$) of 15. For simplicity, we add flux-independent noise in the following way. We generate one 1000 skewer × skewer length realization of random noise drawn from a Gaussian with $\sigma_N = 1/\mathrm{SNR}_{\Delta v}$ and add this noise realization to every model at every redshift. Using the same noise realization over the different models prevents stochasticity from different realizations of the noise from adding additional variations between the models. Thus the noise modeling will not unduly corrupt the parameter inference.

As discussed in Section 4.2.1 simulation skewers are 100 cMpc h^{-1} long, much longer than the $\Delta z = 0.1$ redshift bins we have chosen to analyze. Therefore, we split these skewers into two regions of length $\Delta z = 0.1$ and treating these two regions as independent, resulting in a total of 2000 skewers. Note that $\Delta z = 0.1$ corresponds to 33 cMpc h^{-1} at $z = 5.4$ and 29 cMpc h^{-1} at $z = 6.0$.

The initial and forward-modeled flux for one $z = 5.4$ skewer is shown in Figure 4.3. This skewer has $T_0 = 9149\,\mathrm{K}$, $\gamma = 1.352$, and $\langle F \rangle = 0.0801$ (our assumed "true" parameters at this redshift). The forward modeled skewer, as is always true, uses $R = 30000$ and $\mathrm{SNR}_{10} = 30$. The initial flux is plotted as the red dashed line while the forward modeled flux is plotted as the black histogram.

We assume a fiducial data set size of 20 quasar spectra that probe a redshift interval of $\Delta z = 0.1$ per quasar for a total pathlength of $\Delta z = 2.0$ at all redshifts. This is a reasonable number of high-z, high resolution quasar observations to consider for a future measurement.

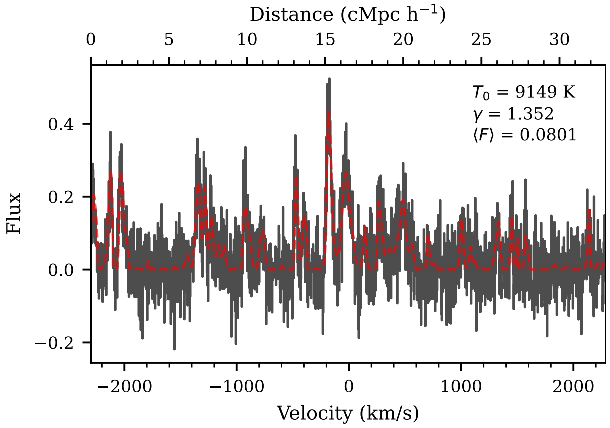

Figure 4.3: A forward-modeled skewer at $z = 5.4$ with $T_0 = 9149\,\mathrm{K}$, $\gamma = 1.352$, and $\langle F \rangle = 0.0801$ (our assumed "true" parameters at this redshift). This skewer, as is true for all skewers, is forward modeled with $R = 30000$ and $\mathrm{SNR}_{10} = 30$. The initial flux from the simulations is a red dashed line while the forward modeled flux is a black histogram.

4.3 Methods

4.3.1 Auto-correlation

The auto-correlation function of the flux, $\xi_F(\Delta v)$, is defined as

$$\xi_F(\Delta v) = \langle F(v) F(v + \Delta v) \rangle \tag{4.3}$$

where $F(v)$ is the flux of the Lyα forest and the average is performed over all pairs of pixels with the same velocity lag, Δv. Conventionally, the flux contrast field, $\delta_F = (F - \langle F \rangle)/\langle F \rangle$, is used when measuring the power spectrum of the Lyα forest. Here, we chose to use the flux since $\langle F \rangle$ is small and has large uncertainties at high-z where we are most interested in this measurement. Using the flux thus prevents us from dividing by a small number which comes from an independent measurement and could potentially blow up the value of the flux contrast. The auto-correlation function of the flux contrast can be written as

$$\xi_{\delta_f}(\Delta v) = \frac{\xi_F(\Delta v) - \langle F \rangle^2}{\langle F \rangle^2}. \tag{4.4}$$

ξ_{δ_f} can be computed via the Fourier transform of the dimensionless power spectrum of the Lyα forest flux contrast, $\Delta^2_{\delta_F}(k) = kP_{\delta_f}(k)/\pi$. In one dimension this can be written as:

$$\xi_{\delta_f}(\Delta v) = \int_0^\infty \Delta^2_{\delta_f}(k)\cos(k\Delta v) d\ln k \tag{4.5}$$

The dimensionless power, $\Delta^2_{\delta_f}(k)$, is a smoothly rising function that has a sharp cutoff set by the thermal state of the IGM. Higher temperature values lead to sharper cutoffs as the power at small scales in the Lyα forest is removed. Equation (4.5) can be particularly useful when building intuition for the trends seen in the auto-correlation function with changing T_0 and γ, which we will discuss later in this section.

We compute the auto-correlation function with the following consideration for the velocity bins. We set the left edge of the smallest bin to be the resolution length, $10\,\mathrm{km\,s^{-1}}$, and continue with linear bin sizes with a width of the resolution length, $10\,\mathrm{km\,s^{-1}}$, up to $300\,\mathrm{km\,s^{-1}}$. Then we switch to logarithmic bin widths where $\log(\Delta v) = 0.029$ out to a maximal distance of $2700\,\mathrm{km\,s^{-1}}$. This results in 59 velocity bins considered where the first 28 have linear spacing. The center of our smallest bin is $15\,\mathrm{km\,s^{-1}}$ and the center of our largest bin is $2295\,\mathrm{km\,s^{-1}}$. This largest bin corresponds to $\sim$ 16.5 cMpc h^{-1} at $z = 5.4$. We chose to use linear bins on the smallest scales because this is where the thermal state has the greatest effect on the Lyα forest flux. At larger scales we switch to logarithmic binning as this is only sensitive to $\langle F \rangle$ and not the thermal parameters. The main aim of this work is to constrain the thermal parameters so having fine binning at large scales is not as important. To check this we compared out results at $z = 5.4$ to those when using linear bins at all scales and found no significant change to the constraints on the parameters. However, using linear bins at all scales results in 268 total bins, which significantly slowed down our computations. Therefore we used the linear-logarithmic bins at all z throughout the rest of this work.

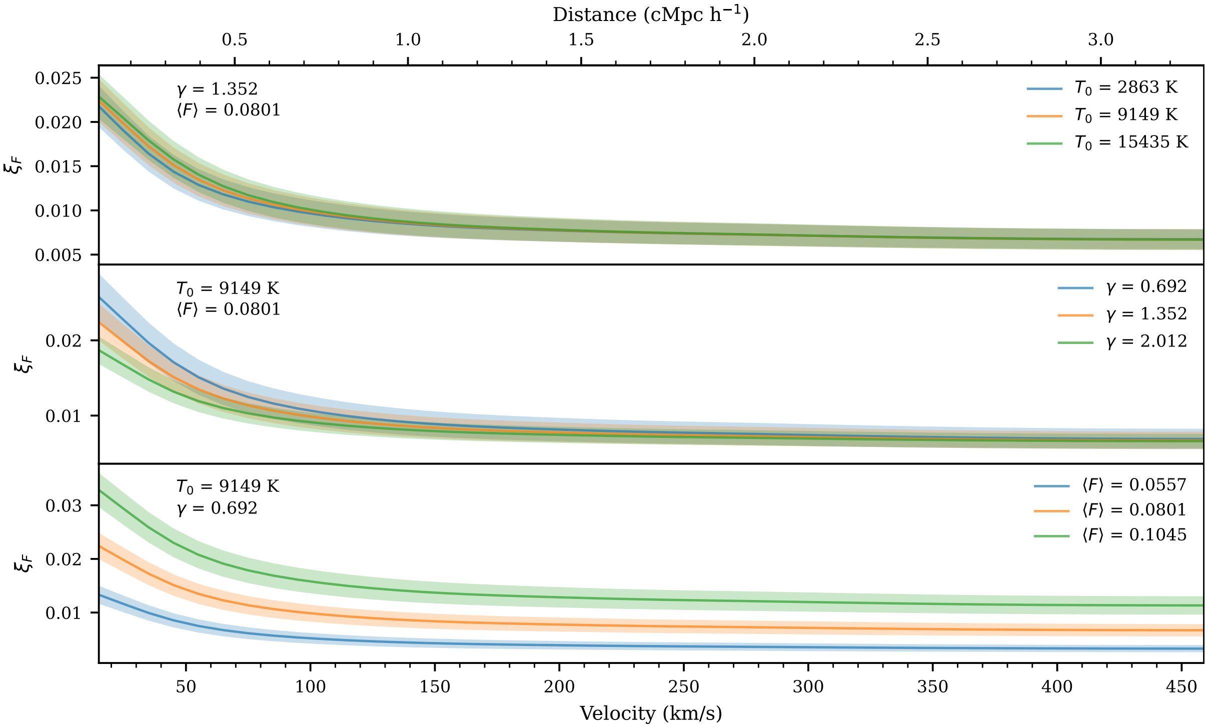

Figure 4.4: This figure demonstrates the effects of varying the parameters on the auto-correlation function from the simulations at $z = 5.4$. Each of the three panels varies one parameter from T_0, γ, and $\langle F \rangle$ while keeping the others constant. The constant parameter values are written in the top left of each panel. The solid lines show the model values and the shaded regions show errors estimated by the diagonals of the covariance matrices. T_0 (top panel) and γ (middle panel) affect the auto-correlation function on small scales. $\langle F \rangle$ (bottom panel) affects the auto-correlation function on all scales.

The model value of the auto-correlation function was determined by taking the average of the auto-correlation function over all 2000 forward-modeled skewers. Each mock data set of the auto-correlation was calculated by taking an average over 20 random skewers (representing 20 quasar sightlines) from the initial 2000 forward-modeled skewers. The value of the auto-correlation function at the smallest velocity lags is affected by the finite resolution. This effect is left in both the models and the mock data.

We show the correlation functions calculated for different thermal state parameters in Figure 4.4 at $z = 5.4$. The solid lines show the mean values while the shaded regions represent the errors estimated from the diagonal of the covariance matrices. We discuss the computation of these covariance matrices later in this section.

The top panel shows models that vary T_0 with constant γ and $\langle F\rangle$. Varying T_0 results in small changes for the smallest velocity lags, where the second bin centered on $25\,\mathrm{km\,s^{-1}}$ has the largest percent change in the models. The middle panel has models that vary γ with constant T_0 and $\langle F\rangle$ where the effect of changing γ is strongest on small scales. The bottom panel has models that vary $\langle F\rangle$ with constant T_0 and γ. $\langle F\rangle$ sets the amplitude of the auto-correlation function at all velocity lags. Here the differences between models are linear where larger $\langle F\rangle$ leads to larger auto-correlation values. This scaling is roughly $\propto \langle F\rangle^2$, which follows from the definition of the auto-correlation function.

For the thermal models, larger T_0 and smaller γ lead to larger correlation function values on small scales. Though these models do not seem to show large differences by eye, we will investigate what statistically rigorous measurements could look like in Section 4.3.4.

To build intuition for the behavior of the auto-correlation function with the thermal parameters we refer to Equation (4.5). In Appendix 4.6 we show the integrand from this equation for $\Delta v = 15\,\mathrm{km\,s^{-1}}$. As mentioned above $\Delta^2_{\delta_f}(k)$ has a sharp thermal cutoff which would naively lead to the belief that the auto-correlation function could show lower values at small-scales for hotter thermal states. However, as seen in Figure 4.4, greater T_0 values have greater values of the auto-correlation function at small scales. This behavior is explained by both the great variation in the flux in these models, as described earlier in the section, and by the behavior of $\Delta^2_{\delta_f}(k)$ at small k values with a linear y-scale, as seen in Figure 4.14.

We compute the covariance matrices for the models from mock draws of the data:

$$\Sigma(T_0, \gamma, \langle F\rangle) = \frac{1}{N_{\mathrm{mocks}}} \sum_{i=1}^{N_{\mathrm{mocks}}} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_{\mathrm{model}})(\boldsymbol{\xi}_i - \boldsymbol{\xi}_{\mathrm{model}})^{\mathrm{T}} \tag{4.6}$$

where $\boldsymbol{\xi}_i = \boldsymbol{\xi}_i(T_0, \gamma, \langle F\rangle)$ is the i-th mock auto-correlation function, $\boldsymbol{\xi}_{\mathrm{model}} = \boldsymbol{\xi}_{\mathrm{model}}(T_0, \gamma, \langle F\rangle)$

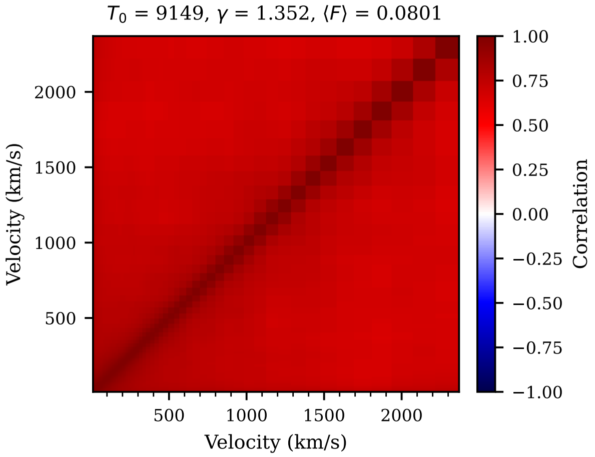

Figure 4.5: This figure shows the correlation matrix calculated with equation (4.7) for the model at $z = 5.4$ with $T_0 = 9149\,\mathrm{K}$, $\gamma = 1.352$, $\langle F \rangle = 0.0801$. The color bar is fixed to span from -1 to 1, which is all possible values of the correlation matrix. This illustrates that all bins in the auto-correlation function are highly correlated with each other.

is the model value of the auto-correlation function, and N_{mocks} is the number of forward-modeled mock data sets used. We use $N_{\mathrm{mocks}} = 500000$ for all models and redshifts in this work, see Appendix 4.7 for a discussion on the convergence of the covariance matrix. Note that $\boldsymbol{\xi}_i(T_0, \gamma, \langle F \rangle)$ and $\Sigma(T_0, \gamma, \langle F \rangle)$ are computed at each point on the grid of T_0, γ, and $\langle F \rangle$, resulting in 1215 separate computations.

To visualize the covariance matrix, we define the correlation matrix, C, which expresses the covariances between jth and kth bins in units of the the diagonal elements of the covariance matrix. This is done for the jth, kth element by

$$C_{jk} = \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}}. \tag{4.7}$$

One example correlation matrix is shown in Figure 4.5 for $z = 5.4$ with $T_0 = 9149\,\mathrm{K}$, $\gamma = 1.352$, $\langle F \rangle = 0.0801$. All bins of the auto-correlation function are highly correlated, which is caused by each pixel in the Lyα forest contributing to multiple (in fact almost all) bins in the auto-correlation function.

4.3.2 Parameter Estimation

To quantitatively constrain the parameters we modeled (T_0, γ, and $\langle F \rangle$), we use Bayesian inference with a multivariate Gaussian likelihood and a flat prior over the parameters. This likelihood, $\mathcal{L} = p(\boldsymbol{\xi}|T_0, \gamma, \langle F \rangle)$, has the form:

$$\mathcal{L} = \frac{1}{\sqrt{\det(\Sigma)(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}_{\text{model}})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}_{\text{model}})\right) \tag{4.8}$$

where $\boldsymbol{\xi}$ is the auto-correlation function from our mock data, $\boldsymbol{\xi}_{\text{model}} = \boldsymbol{\xi}_{\text{model}}(T_0, \gamma, \langle F \rangle)$ is the model value of the auto-correlation function, $\Sigma = \Sigma(T_0, \gamma, \langle F \rangle)$ is the model dependent covariance matrix estimated by Equation (4.6), and $n = 59$ is the number of points in the auto-correlation function. We discuss the assumption of using a multivariate Gaussian likelihood in Appendix 4.8. This discussion shows that our mock data does not exactly follow a Guassian distribution. This discrepancy may affect our parameter inference, we investigate the consequences of this assumption in a later section.

Our models are defined by three (T_0, γ, and $\langle F \rangle$) parameters. We compute the posteriors on these parameters using Markov Chain Monte Carlo (MCMC) with the `EMCEE` (Foreman-Mackey et al., 2013) package. We linearly interpolate the model values and covariance matrix elements onto a finer 3D grid of T_0, γ, and $\langle F \rangle$ then use the nearest model during the MCMC. This fine grid has 29 values of T_0, 33 values of γ, and 41 values of $\langle F \rangle$ which corresponds to adding 1, 3, and 4 points between the existing grid points respectively. Our MCMC was run with 16 walkers taking 3500 steps each and skipping the first 500 steps of each walker as a burn-in.

Figure 4.6 shows the result of our inference procedure for one mock data set at $z = 5.4$. The top panel shows the mock data set with various lines relating to the inference procedure as follows. The green dotted line and accompanying text present the model value for the simulation that the mock data was taken from. The mock data set is plot as the black points with error bars that come from the diagonal elements of the covariance matrix of the model that is nearest to the inferred model. The inferred model is the model that comes from the median of each parameter's samples determined independently via the 50th percentile of the MCMC chains. The red line and accompanying text present this inferred model. The errors on the inferred model written in the text are from the 16th and 84th percentiles of the MCMC chains. The blue lines show models corresponding to 100 random draws from the MCMC chain to visually demonstrate variety of models that come from the resulting posterior. The bottom left panel shows a corner plot of the posteriors for T_0, γ, and $\langle F \rangle$.

4.3.3 Inference Test and Re-weighting

We test to check the fidelity of our inference procedure. This test ensures that the behavior of our posteriors is statistically correct and checks the validity of any assumptions we make during our inference. For example, in this work we used an approximate likelihood in the form of a multivariate Gaussian likelihood. The Lyα forest is known to be a non-Gaussian random field. By adopting a multivariate Gaussian likelihood here, we are tacitly assuming that averaging over all pixel pairs when calculating the auto-correlation function will Gaussianize the resulting distribution of the values of the auto-correlation function, as is expected from the central limit theorem. We discuss the distribution of these values for our mock data in detail in Appendix 4.8. If this assumption is not valid our reported errors may be either underestimated or overestimated.

The general idea of our inference test is to compare the true probability contour levels with the "coverage" probability. The coverage probability is the percent of time (over many mock data sets) the true parameters of a mock data set fall within a given probability contour. In our case, we compute this over 300 mock data sets where the true parameters considered are sampled from our priors. Ideally, this coverage probability

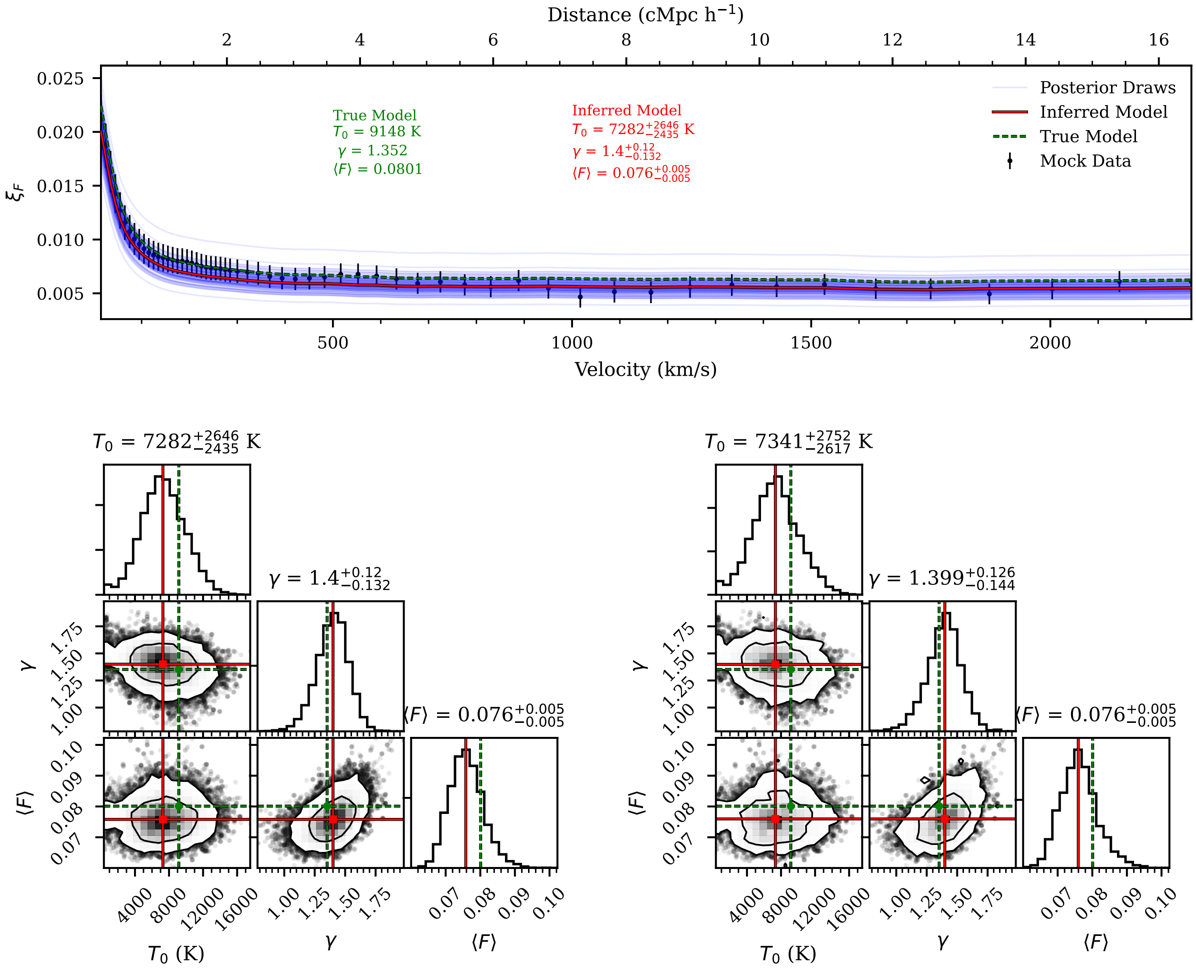

Figure 4.6: This figure illustrates the results of our inference procedure applied to one mock data set at $z = 5.4$. The top panel shows the data and models that resulted from our inference procedure, the bottom left has the corner plot resulting from the fit, and the bottom right has the same corner plot which has been re-weighted to pass our inference test. In the top panel, the black points are the mock data with error bars from the diagonals of the covariance matrix of the inferred model. The inferred model was calculated by the median (50th percentile) of the MCMC chains of each parameter independently. The inferred model is shown as a red line while the accompanying red text reports errors calculated from the 16th and 84th percentiles of each parameter. In comparison, the "true" model, which was used to generate the data, is shown as a green dotted line. The parameters for this model is written in the accompanying green text. To demonstrate the width of the posterior, multiple faint blue lines are shown which are the models corresponding to the parameters from 100 random draws of the MCMC chain. The bottom left panel shows a corner plot of the values of T_0, γ, and $\langle F \rangle$ that immediately result from our inference procedure. The bottom right panel shows the corner plot of the values of T_0, γ, and $\langle F \rangle$ from our inference procedure that has been re-weighted with the weights calculated from our inference test as described in Section 4.3.3. For this mock data set, the "true" model parameters fall within the 68th percentile contours.

should be equal to the chosen probability contour level. This calculation can be done at many chosen probabilities resulting in multiple corresponding coverage probabilities. Existing work that explore this coverage probability include Prangle et al. (2013); Ziegel & Gneiting (2013); Morrison & Simon (2017); Sellentin & Starck (2019).

We plot the results of our inference procedure (i.e. $P_{\rm inf}$ vs $P_{\rm true}$) at $z = 5.4$ from 300 posteriors in the left panel of Figure 4.7. The grey shaded regions around our resulting line show the Poisson errors for our results. Again we expect $P_{\rm true} = P_{\rm inf}$ which would give the red dashed line in this figure. To interpret this plot, first consider one point, for example $P_{\rm true} \approx 0.6$. This represents the 60th percentile contour, which was calculated by the 60th percentile of the probabilities from the draws of the MCMC chain for each mock data set. Here, the true parameters fall within the 60th percentile contour only $\approx 52\%$ of the time. This implies that our posteriors are too narrow and should be wider such that the true model parameters will fall in the 60th percentile contour more often, so we are in fact underestimating our errors. We run this inference test at all z considered in this work and found the deviation from the 1-1 line is worse at higher redshifts. See Appendix 4.9 for a discussion of the inference test at $z = 6$. We additionally run the inference test for mock data generated from a multi-variate Gaussian distribution in Appendix 4.10. The inference test using Gaussian mock data agrees with the 1-1 line, which indicates the reason for the inference test from forward-modeled data failing is that the distribution of the data is not perfectly Gaussian.

There has been much recent work trying to correct posteriors that do not pass this coverage probability test (see e.g. Prangle et al., 2013; Grünwald & van Ommen, 2014; Sellentin & Starck, 2019). In this work, we are using the method of Hennawi et al. in prep. where we calculate one set of weights for the MCMC draws that broaden the posteriors in a mathematically rigorous way. This method has been described in some detail in Wolfson et al. (2023b) so we refer to that paper for details on computing this set

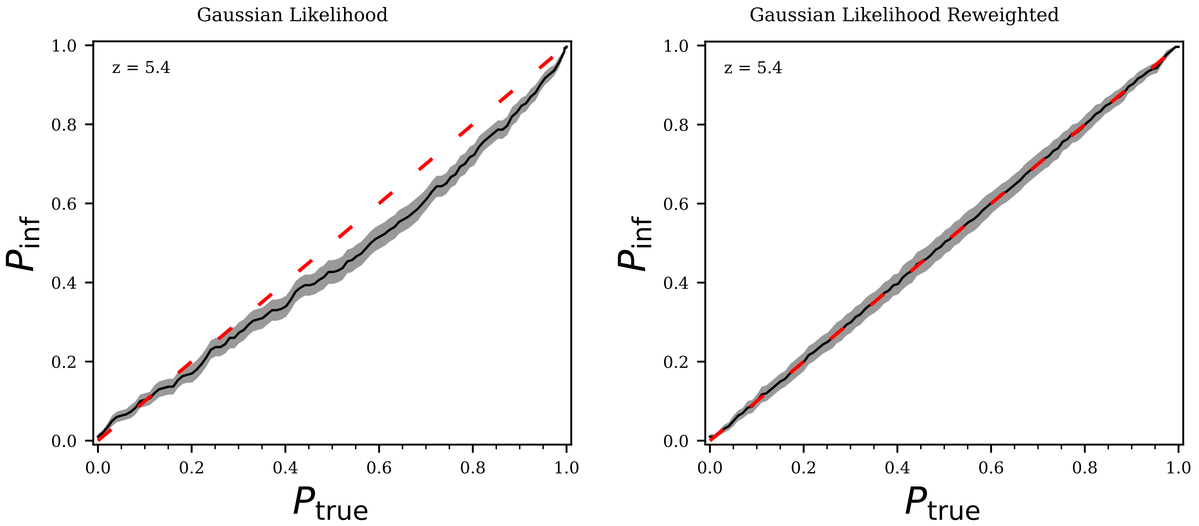

Figure 4.7: The left panel shows the coverage resulting from the inference test at $z = 5.4$ from 300 mock data sets with parameters drawn from our priors on T_0, γ, and $\langle F \rangle$. This shows that, for example, the true parameters fall above the 60th percentile in the MCMC chain $\sim 50\%$ of the time. The line falls below the 1-1 line, meaning that the posteriors are overconfident (too narrow). The right panel of this figure shows the coverage resulting from the inference test with the use of one set of weights to re-weight the posteriors, which passes.

of weights. For now, we will proceed with discussing the effect of adding these weights to the posteriors.

We show the re-weighted posteriors on T_0, γ, and $\langle F \rangle$ in the bottom right panel of Figure 4.6. The weights give greater importance to values of T_0, γ, and $\langle F \rangle$ that are outside of the 68% contour, effectively broadening the posteriors and increasing the errors on the fit. For the mock data set in Figure 4.6 the re-weighted marginalized posterior for T_0 gives $T_0 = 7341^{+2752}_{-2617}$ K whereas before the inferred value was 7282^{+2646}_{-2435} K, so the new errors are $\sim 6\%$ larger. The re-weighted posterior for γ gives $\gamma = 1.399^{+0.126}_{-0.144}$ whereas before the inferred value was $1.400^{+0.120}_{-0.132}$, so the new errors are $\sim 7\%$ larger. The error on $\langle F \rangle$ does not change. When looking at the 2D distributions in this corner plot, such as the (γ, $\langle F \rangle$) distribution in the middle panel of the bottom row, we see small regions outside of the main 95% contour that are important. This comes from weighting one draw quite highly, which demonstrates how the weights introduce an additional source of noise to the posterior distribution.

This whole inference procedure is not the optimal and may not give the best possible constraints on T_0 or γ from the auto-correlation function. The need to use re-weighting, or some method to correct our posteriors to pass an inference test, comes from our incorrect (though frequently used) assumption of a multivariate Gaussian likelihood. The values of the auto-correlation function at these high z do not sufficiently follow a multivariate Gaussian distribution to justify this assumption, which should be a warning for other studies of the Lyα forest at these z. Using a more correct form of the likelihood (such as a skewed distribution) or likelihood-free inference (such as approximate Bayesian computation as used in Davies et al. (2018b) or other machine learning methods) would lead to more optimal posteriors that better reflect the information content of the auto-correlation function.

4.3.4 Thermal state measurements

We study the distribution of measurements for 100 mock data sets with one "true" (T_0, γ, $\langle F \rangle$) model in order to account for cosmic variance. For each z we use the T_0, γ, and $\langle F \rangle$ values reported in Table 4.1. Each mock data set is chosen by randomly selecting and averaging the auto-correlation function over 20 skewers. For each mock data set, we perform MCMC as described in Section 4.3.2 and then re-weight the resulting posteriors following Section 4.3.3. Once we have the weights and the chains resulting from our inference procedure we can calculate the marginalized posterior for T_0 and γ.

At $z = 5.4$, all 100 marginalized re-weighted posteriors are shown as the faint blue lines in Figure 4.8 for T_0 (top panel) and γ (bottom panel). Attempting to fit the model value of the auto-correlation function removes the luck of the draw that exists in selecting mock data and gives the optimal precision of the posteriors. The resulting posteriors from fitting the model is shown as the thick blue histogram in the figure. The measurement resulting from this fit is written in the blue text of this figure and the values at each z are reported in Table 4.2.

The re-weighted histograms in Figure 4.8 are noisy, much like is seen in the bottom right panel of Figure 4.6. This is a direct consequence of our re-weighting procedure and will be improved with further work on likelihood-free inference. For T_0, the model value of the auto-correlation function gives a posterior with a width that is typical of those from the mock data. For γ, the posterior has a slightly narrower peak. Also for γ, posteriors that peak at lower γ values are broader than those that peak at higher γ values. Both model posteriors contain the true value of T_0 and γ within their 1σ error bars.

Table 4.2 reports the measurements that result from using the model values of the auto-correlation function as our data at all z. This is an ideal scenario that removes luck of the draw from the resulting measurement. The first (third) column contains the "true" modeled value of T_0 (γ) at each z that was used in this measurement. The second (fourth) column contains the measurements for T_0 (γ) calculated by the 16th, 50th, and 84th percentiles. In general the trend of the errors is to increase with increasing redshift. At $z = 5.4$, the measurement of the model constrains T_0 to 29% and γ to 9%.

In order to visualize the differences between measurements at different redshifts, we plot the results for two random mock data sets in Figure 4.9. The first and third panels show the marginalized posteriors for T_0 while the second and fourth panels shows the marginalized posteriors for γ. Each violin is the re-weighted marginalized posterior for one randomly selected mock data set at the corresponding redshift. The light blue shaded region demarcates the 2.5th and 97.5th percentiles (2σ) of the MCMC draws while the darker blue shaded region demarcates the 16th and 84th percentiles (1σ) of the MCMC draws. The dot dashed line is the true simulated model value evolution as shown in

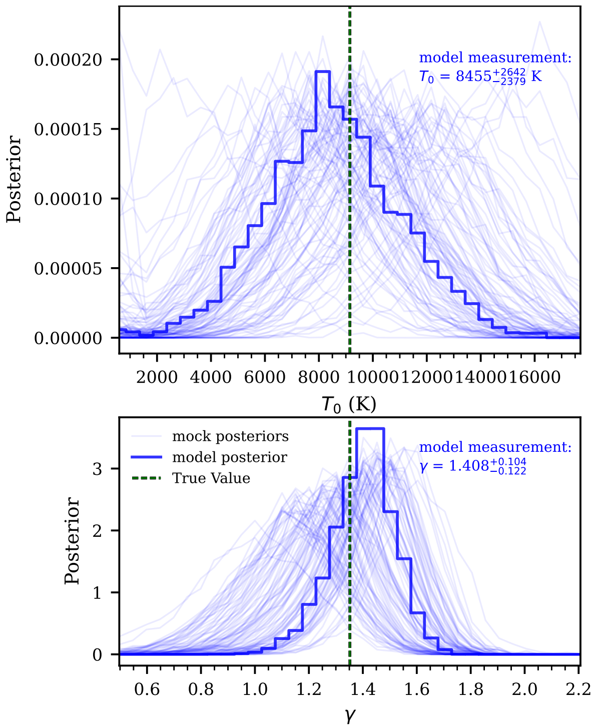

Figure 4.8: 100 re-weighted marginalized posteriors of T_0 and γ at $z = 5.4$ from mock data sets with true $T_0 = 9149$, $\gamma = 1.352$, and $\langle F \rangle = 0.0801$ (faint blue lines). The top panel shows the marginalized posteriors for T_0 and the bottom panel shows the marginalized posteriors for γ. Both panels also show the re-weighted posterior from the model value of the auto-correlation function (thick blue histograms). The measurement resulting from fitting the model are written in blue text. This demonstrates the different possible behaviors the posterior can have for different mock data sets with the same "true" T_0, γ, and $\langle F \rangle$ values.

z	Model T_0	Measured T_0	Model γ	Measured γ
5.4	9149	8455^{+2642}_{-2379}	1.352	$1.408^{+0.104}_{-0.122}$
5.5	9354	8643^{+3152}_{-3054}	1.338	$1.422^{+0.116}_{-0.141}$
5.6	9572	8480^{+3720}_{-3642}	1.324	$1.433^{+0.121}_{-0.151}$
5.7	9804	8222^{+5188}_{-4176}	1.309	$1.460^{+0.139}_{-0.166}$
5.8	10050	8346^{+4926}_{-4576}	1.294	$1.485^{+0.157}_{-0.204}$
5.9	10320	7892^{+6111}_{-4655}	1.278	$1.513^{+0.170}_{-0.223}$
6.0	10600	9574^{+6219}_{-5133}	1.262	$1.511^{+0.196}_{-0.256}$

Table 4.2: The results of fitting to the models of the auto-correlation function for given T_0 and γ values. The first (third) column contains the modeled value of T_0 (γ) at each z. The second (fourth) column contains the measurements for T_0 (γ) calculated by the 16th, 50th, and 84th percentiles. In general the trend of the errors is to increase with increasing redshift.

Figure 4.1 and reported in Table 4.1.

Looking at the posteriors for a given redshift (one column in the figure), the only difference between the posteriors is the random mock data set drawn. This still produces different precision results as seen in Figure 4.8 for $z = 5.4$. For the different posteriors within one section of this figure, the random mock data set differs as does the true values of T_0 and γ. Again, the individual posteriors are noisy, resulting from the re-weighting procedure as described in Section 4.3.3. The behavior here echos that found with the model measurements where the precision of the constraints on T_0 and γ decrease with increasing z. In the highest redshift bins, $z > 5.7$, the posteriors for the mock data sets have high values at the boundary of our prior much more often.

4.4 Inhomogeneous Reionization

So far in this work we have used a semi-numerical method to "paint on" different thermal states to our simulations for a tight temperature-density relationship. This is

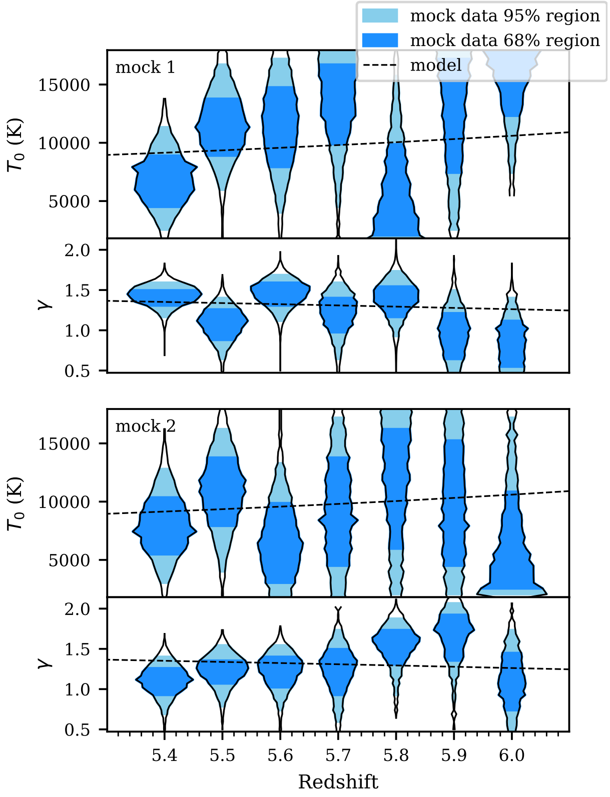

Figure 4.9: The marginalized posteriors for two random mock data sets at each z for T_0 and γ. The first and third panels show the marginalized posteriors for T_0 while the second and fourth panels show the same for γ. For each posterior, the light blue shaded region demarcates the 2.5th and 97.5th percentile of the weighted MCMC draws while the darker blue shaded region demarcates the 17th and 83rd percentile of the weighted MCMC draws. There are 14 total random mock data sets used to make this figure. For a given T_0 and γ posterior pair (in the first and second or third and fourth panels) the mock data set is the same. The behavior of each posterior is partially determined by the luck of the draw when selecting the mock data. The size of the data set is consistent across z but the true parameter values of the mock data varies as shown by the black dot dashed line. This black dot dashed line was also shown in Figure 4.1 and the values at each z are reported in Table 4.1.

sufficient to explore the sensitivity of the Lyα forest flux auto-correlation function to the thermal state of the IGM at high-redshifts. However, as previously discussed, recent measurements of the Lyα optical depth at $z > 5.5$ have shown scatter that can't be explained by density fluctuations alone (Fan et al., 2006; Becker et al., 2015; Bosman et al., 2018; Eilers et al., 2018; Bosman et al., 2022). It is possible that these fluctuations come from fluctuations in the temperature field (D'Aloisio et al., 2015; Davies et al., 2018 a) or fluctuations in the UVB (Davies & Furlanetto, 2016b; Gnedin et al., 2017; D'Aloisio et al., 2018). Fluctuations in either of these fields can arise if reionization is extended or patchy.

On top of the measurements of fluctuations in the Lyα forest optical depth at $z > 5.5$, recent measurements of the mean free path of ionizing photons at $z > 5$ suggest a UVB that cannot be well described by uniform fields (Becker et al., 2021; Bosman, 2021; Gaikwad et al., 2023; Zhu et al., 2023).

In order to explore the effect of temperature and UVB fluctuations on the Lyα forest flux auto-correlation function, we consider a set of four simulation models. These simulations have two different reionization models (one of which causes temperature fluctuations) and two UVB models (one of which has fluctuations). These simulations and their results will be described in detail in the following sections.

4.4.1 Simulation box

For these models we use an additional Nyx simulation box with a size of $L_{\text{box}} = 40$ cMpc h^{-1} and 2048^3 resolution elements at $z = 5.8$. A slice through the density field of this simulation is shown in the top left panel of Figure 4.10.

We consider two reionization models: an instantaneous model and an extended, inhomogeneous model (the "flash" and inhomogeneous methods described in Oñorbe et al.

(2019), respectively). The instantaneous model of reionization assigns all resolution elements the same redshift of reionization, $z_{\text{reion, HI}}$. For this work we use $z_{\text{reion, HI}} = 7.75$. A brief summary of the inhomogeneous model of reionization is as follows, each resolution element is assigned its own redshift of reionization such that reionization has a given midpoint, $z^{\text{median}}_{\text{reion, HI}}$, and duration, $\Delta z_{\text{reion, HI}}$. For this work we use $z^{\text{median}}_{\text{reion, HI}} = 7.75$ and $\Delta z_{\text{reion, HI}} = 4.82$. It is possible for cells to be ionized before the redshift of reionization through other processes such as collisional reionization. In both models, at the redshift of reionization for a given resolution element heat, ΔT, is injected. In both of our reionization models $\Delta T = 2 \times 10^4$ K. These two models result in two different temperature fields. We say that the instantaneous reionization model has “no temperature fluctuations” and the inhomogeneous reionization model has “temperature fluctuations”.

The bottom row of Figure 4.10 shows slices through the resulting temperature field from these two simulations: one with no temperature fluctuations on the left and one with temperature fluctuations on the right. From this figure we see that model with temperature fluctuations has a larger scatter in the temperature with the greater abundance of colder (darker blue) regions. These cold regions correspond to the regions of higher density in the top left panel. This follows from the model of reionization where the denser regions reionize (and are heated) first and thus have more time to cool to a lower temperature by $z = 5.8$.

In addition to a constant UVB model, we have a model with uvb fluctuations. This UVB model was generated by the same method presented in Oñorbe et al. (2019) with $\lambda_{\text{mfp}} = 15$ cMpc. The method follows the approach of Davies & Furlanetto (2016b) where we consider modulations in the ionization state of optically thick absorbers assuming that $\lambda_{\text{mfp}} \propto \Gamma^{2/3}_{\text{UVB}}/\Delta$ where Δ is the local matter density. For the fluctuating UVB, Γ_{HI} was calculated on a uniform grid of 64^3 at $z = 6$ and then linearly interpolated the $\log \Gamma_{\text{UVB}}$ field to match the hydrodynamical simulation with 2048^3. The top right panel of Figure

4.10 shows a slice through the UVB model with fluctuations. The largest UVB values are in the same location as the high density areas shown in the top left panel. These are the densest regions of the simulation which contain the majority of the sources of ionizing photons. We do not show the model without UVB fluctuations as this is a constant field.

Thus our four models of reionization are (1) no temp. fluctuations and no UVB fluctuations (2) no temp. fluctuations with UVB fluctuations (3) temp. fluctuations with no UVB fluctuations and (4) both temperature and UVB fluctuations. All four models are normalized to $\langle F \rangle = .0172$, which is the measured value at $z = 5.8$ from Bosman et al. (2022). We do not consider multiple values of $\langle F \rangle$ for these models since they represent four discrete models and we will not try to constrain any parameters.

We now consider the effect of these four simulation models on the transmitted flux. Figure 4.11 shows one skewer from each of the four different reionization models at $z = 5.8$. The top panel shows the resulting Lyα forest flux. The second panel shows the density field along the skewer. The third panel shows the temperature along the skewer. The bottom panel shows the UVB background values. Each panel has four lines representing models with no temperature and no UVB fluctuations (solid blue), no temperature fluctuations with UVB fluctuations (dashed blue), temperature fluctuations with no UVB fluctuations (solid red), and both temperature and UVB fluctuations (dashed red). Comparing the solid lines to each other isolates the effect of temperature fluctuations only. When comparing these two models, we see that a positive scatter in the temperature of the IGM leads to increased flux over $-1600\,\mathrm{km\,s^{-1}} < v < -1000\,\mathrm{km\,s^{-1}}$. Comparing the dashed lines to the solid lines of the same color isolates the effect of UVB fluctuations. For example consider $v > 1000\,\mathrm{km\,s^{-1}}$ where the models with UVB fluctuations (dashed) in the bottom panel are constantly greater than the models without UVB fluctuations (solid). In the top panel, these positive fluctuations in the UVB boost the flux in these dashed lines over the solid lines of the same color.

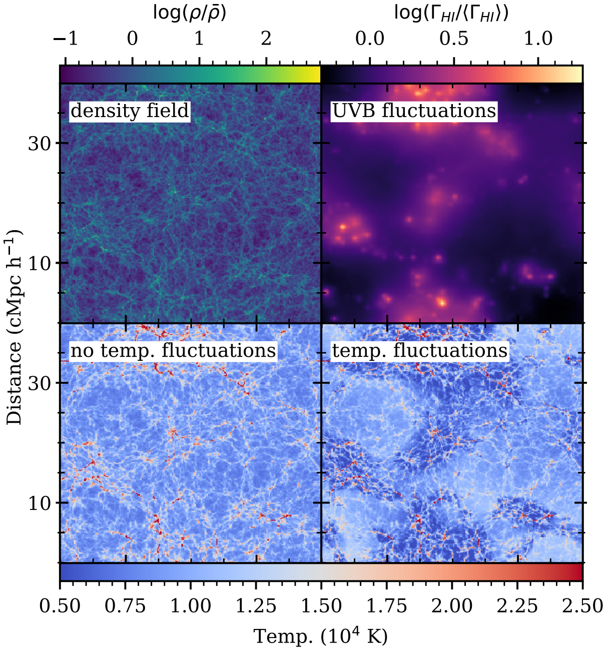

Figure 4.10: This figure shows slices of density field (top left), the temperature field (bottom row), and UVB (top right) for the `Nyx` simulation described in Section 4.4. The bottom left panel shows the temperature field without fluctuations. The bottom right panel right shows the temperature field with fluctuations. The model with temperature fluctuations has a greater scatter in the temperature field, as can be seen by the greater abundance of colder (darker blue) regions. These cold regions correspond to the regions of higher density in the top left panel. The top right panel shows a slice through the UVB field of the simulation with $\lambda_{\rm mfp} = 15$ cMpc, which gives a fluctuating UVB. The largest UVB values are in the same location as the high density areas shown in the top left panel. These are the densest regions of the simulation which contain the majority of the sources of ionizing photons.

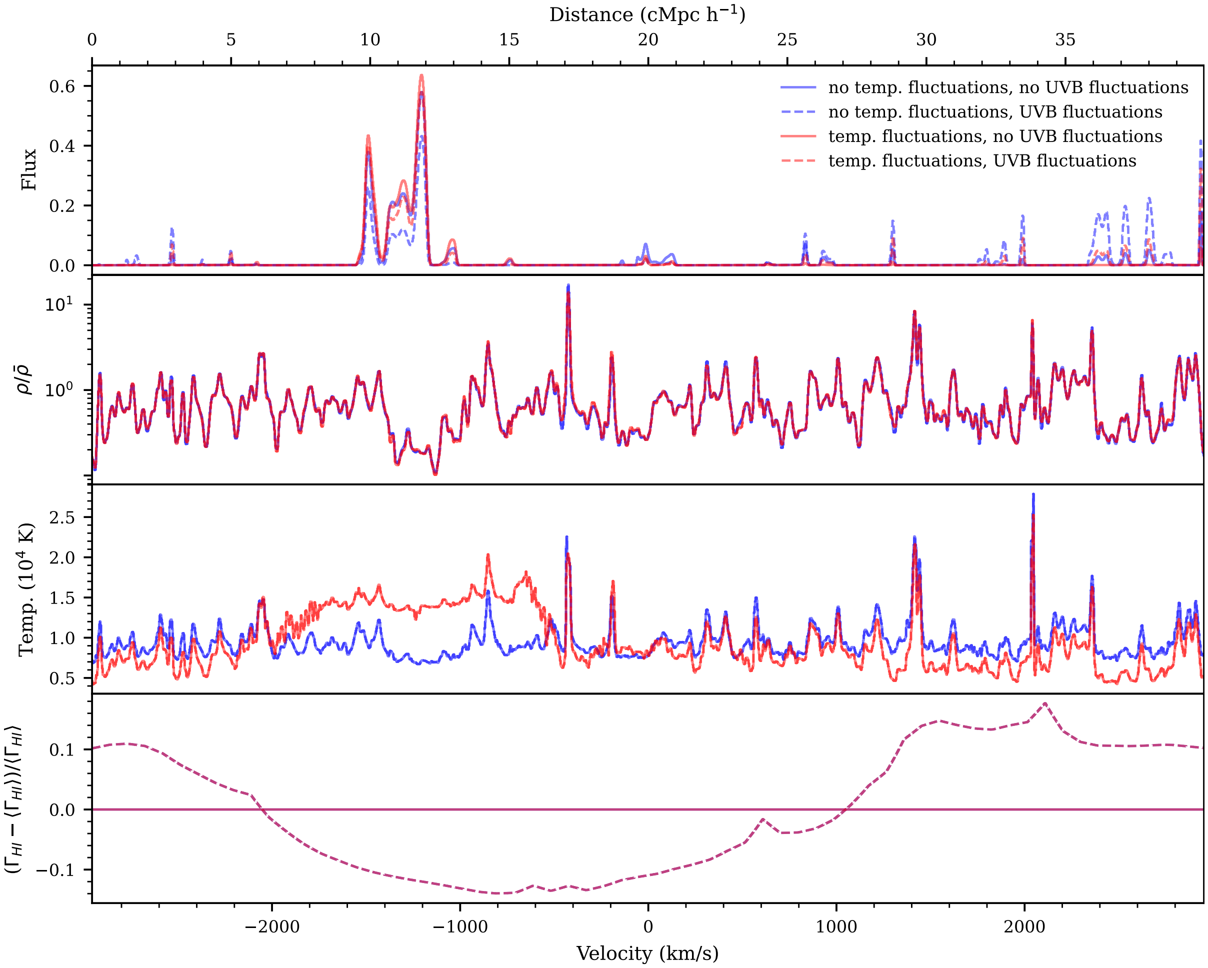

Figure 4.11: This figure shows one skewer from the four various reionization models at $z = 5.8$. The top panel shows the resulting Lyα forest flux. The second panel shows the density field along the skewer. The third panel shows the temperature along the skewer. The bottom panel shows the UVB background values. Each panel has four lines representing models with no temperature and no UVB fluctuations (solid blue), no temperature fluctuations with UVB fluctuations (dashed blue), temperature fluctuations with no UVB fluctuations (solid red), and both temperature and UVB fluctuations (dashed red). Comparing the solid lines to each other isolates the effect of temperature fluctuations only. When comparing these two models, we see that a positive scatter in the temperature of the IGM leads to increased flux over $-1600\,\mathrm{km\,s^{-1}} < v < -1000\,\mathrm{km\,s^{-1}}$. Comparing the dashed lines to the solid lines of the same color isolates the effect of UVB fluctuations. For example consider $v > 1000\,\mathrm{km\,s^{-1}}$ where the models with UVB fluctuations (dashed) in the bottom panel are constantly greater than the models without UVB fluctuations (solid). In the top panel, these positive fluctuations in the UVB boost the flux in these dashed lines over the solid lines of the same color.

In general, the UVB fluctuations are anti-correlated with the temperature fluctuations. This follows from the dense regions in the simulations causing negative temperature fluctuation and positive UVB simulation as discussed earlier. For example, consider the positive temperature fluctuation and negative UVB fluctuation at $-1600\,\mathrm{km\,s^{-1}} < v < -1000\,\mathrm{km\,s^{-1}}$. Overall this anti-correlation will result in the effects of these two fluctuating fields to cancel out, causing the model with both temperature and UVB fluctuations (dashed red) to look similar to the model with no temperature fluctuations and np UVB fluctuations (solid blue). This is indeed generally seen across the flux panel of Figure 4.11.

From here, we forward model the skewers in the same way as discussed in Section 4.2.2 with $R = 30000$ and $\mathrm{SNR}_{10} = 30$. The only difference is that we leave the skewers with the full 40 cMpc h^{-1} length and then use only 15 (where before we used 20) skewers when calculating mock data sets. The mock data sets here and in the previous section contain the same pathlength corresponding to 20 observed quasars with $\Delta z = 0.1$. We do not show an example of the forward modeled skewer here as they are very similar to that shown in Figure 4.3.

4.4.2 Auto-correlation

The auto-correlation functions are computed via Equation (4.3) and the covariance matrices are computed via Equation (4.6).

Figure 4.12 shows the correlation function for the four reionization models at $z = 5.8$ with a logarithmic y-axis. The inset shows the first $100\,\mathrm{km\,s^{-1}}$ of the auto-correlation functions with a linear y-axis to highlight the differences at small scales. The lines show the model value while the shaded regions are the error estimated from the diagonals of the covariance matrices. The colors and line styles here match those in Figure 4.11

with the model with no temperature fluctuations and no UVB fluctuations (solid blue), no temperature fluctuations with UVB fluctuations (dashed blue), temperature fluctuations with no UVB fluctuations (solid red), and both temperature and UVB fluctuations (dashed red). Comparing the red to the blue lines with the same style isolates the effect of temperature fluctuations while comparing the dashed to the solid line with the same color isolates the effect of UVB fluctuations. Note that the shaded regions are about the same size for all four models.

First compare the model with no temperature fluctuations and no UVB fluctuations (solid blue) and the model with temperature fluctuations with no UVB fluctuations (solid red), which isolates the effect of temperature fluctuations. The model values for these models show that adding temperature fluctuations boosts the value of the auto-correlation function for $\Delta v < 1800\,\mathrm{km\,s^{-1}}$. This follows from the additional variation added by the temperature fluctuations.

Now consider the model with no temperature fluctuations and no UVB fluctuations (solid blue) and the model with no temperature fluctuations with UVB fluctuations (dashed blue), which adds UVB fluctuations to a model without temperature fluctuations. Comparing these line in the inset shows that adding UVB fluctuations increases the value of the auto-correlation function on small scales. This result falls in line with that found in Wolfson et al. (2023b) which says that a shorter λ_{mfp} value leads to greater boosts on small scales of the auto-correlation function. At larger scales there is a slight boost in the model with no temperature fluctuations with UVB fluctuations (dashed blue) seen with the logarithmic scale.

Finally consider the model with temperature fluctuations with no UVB fluctuations (solid red) and the model with both temperature and UVB fluctuations (dashed red), which compares adding UVB fluctuations to a model with temperature fluctuations. In this case adding UVB fluctuations decreases the value of the auto-correlation function for $\Delta v < 1800\,\mathrm{km\,s^{-1}}$. This is the opposite effect as adding UVB fluctuations to a model without temperature fluctuations (seen in comparing the blue lines) and the results from Wolfson et al. (2023b). However, there is an anti-correlation between the UVB and temperature fluctuations resulting from the correlations with the density field. For a fluctuating UVB, the UVB is highest where the density is greatest, since this is where ionizing photon sources are located. For a fluctuating temperature model, the temperature is lowest where the density is greatest, which decreases the transmitted flux. This causes more constant flux levels and decreases the auto-correlation function values at these small scales, as seen in these lines. Ultimately, the correlations with density cause the model with both temperature and UVB fluctuations (dashed red) to be most similar to the model with no temperature fluctuations and no UVB fluctuations (solid blue). Note that on small scales there is still a boost in the model with both temperature and UVB fluctuations (dashed red) over the model without both fluctuations (solid blue), which comes from increased variation in the flux.

4.4.3 Ruling-out Reionization scenarios

For these four reionization models, there is no grid of parameters that can be constrained via MCMC. Instead, we will investigate how confidently other models can be ruled out given mock data from a single model. We will rule out models via the likelihood ratio, $\mathcal{R}$, which is defined as

$$\mathcal{R} = \frac{\mathcal{L}(\text{model})}{\mathcal{L}(\text{reference model})} \tag{4.9}$$

Again for this we assume the likelihood, $\mathcal{L}$, is the multivariate Gaussian likelihood from Equation (4.8).

Here we assume that the mock data comes from the model with both temperature and UVB fluctuations (red dashed lines in the Figures 4.11 and 4.12). Therefore, we will

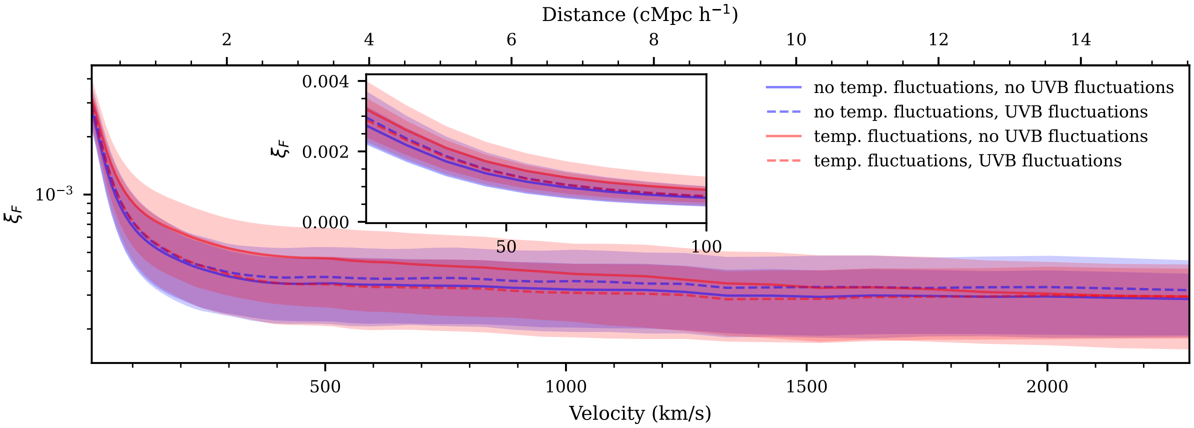

Figure 4.12: This figure shows the correlation function for the four reionization models at $z = 5.8$ with a logarithmic y-axis. The lines show the model values of the correlation function while the shaded region shows the errors estimated from the diagonal of the covariance matrices. The colors and line styles here match those in Figure 4.11 with the model with no temperature fluctuations and no UVB fluctuations (solid blue), no temperature fluctuations with UVB fluctuations (dashed blue), temperature fluctuations with no UVB fluctuations (solid red), and both temperature and UVB fluctuations (dashed red). Comparing the red to the blue lines with the same style isolates the effect of temperature fluctuations while comparing the dashed to the solid line with the same color isolates the effect of UVB fluctuations. Note that the shaded regions are about the same size for all four models. The inset shows the first 100 km s^{-1} of the auto-correlation functions with a linear y-axis to see the differences at small scales.

be looking at the value of the likelihood for the mock data sets using the other three reionization models divided by the likelihood for what we know is the true mock data model (with both temperature and UVB fluctuations). To investigate the distribution of potential likelihood ratio values, we use 1000 mock data sets.

The distribution of the 1000 likelihood ratio values for each of the alternative reionization models are shown in Figure 4.13. The violin plots show the full distribution where the light orange shaded region demarcates the 2.5th and 97.5th percentiles (2σ) of the ratio values while the darker orange shaded region demarcates the 16th and 84th percentiles (1σ) of the ratio values. The solid black line shows where the ratio is equal to 1, which is where both models are just as likely given the mock data. The dashed, dot-dashed, and dotted back lines show the value where the alternative models are ruled out at the 1, 2, and 3 σ levels respectively.

Overall, it is most difficult to rule out the model with no temperature fluctuations and no UVB fluctuations (solid blue lines in previous plots), as is seen in the left most violin in Figure 4.13. This distribution has 44.6% of the mock data sets that favor the incorrect, alternative reionization scenario than the true model with both temperature and UVB fluctuaions. Then only 40.4%, 17.4%, and 3.4% of mock data sets can be ruled out at the 1, 2, and 3 σ levels respectively. This follows from the auto-correlation values for these models seen in Figure 4.12 and the discussion there about how the temperature fluctuations and UVB fluctuations are anti-correlated and thus produce an auto-correlation function most similar to the model which lacks both of these fluctuations.

The next most difficult model to rule out is the model with no temperature fluctuations but with UVB fluctuations (dashed blue lines in the previous plot) as seen in the central violin in Figure 4.13. This distribution has 26.5% of the mock data sets that favor the incorrect, alternative reionization scenario than the true model with both temperature and UVB fluctuations. Then 60.6%, 23.8%, and only 0.3% of mock data sets can be ruled out at the 1, 2, and 3 σ levels respectively. Between this and the left plot there are fewer mock data sets here that can be ruled out at least at the 3σ level but over half of them can be ruled out at 1σ.

The easiest model to rule out is the model with temperature fluctuations but with no UVB fluctuations (solid red lines in the previous plots) as seen in the right most violin in Figure 4.13. This distribution has only 21.8% of the mock data sets that favor the incorrect, alternative reionization scenario than the true model with both temperature and UVB fluctuations. Then 73.9%, 54.0%, and 7.9% of mock data sets can be ruled out at the 1, 2, and 3 σ levels respectively, which is the greatest percentages out of the three alternative models. This also follows from the differences between these models in Figure 4.12. The model with temperature fluctuations but no UVB fluctuations has the greatest values of the auto-correlation function at most scales, making it the easiest to

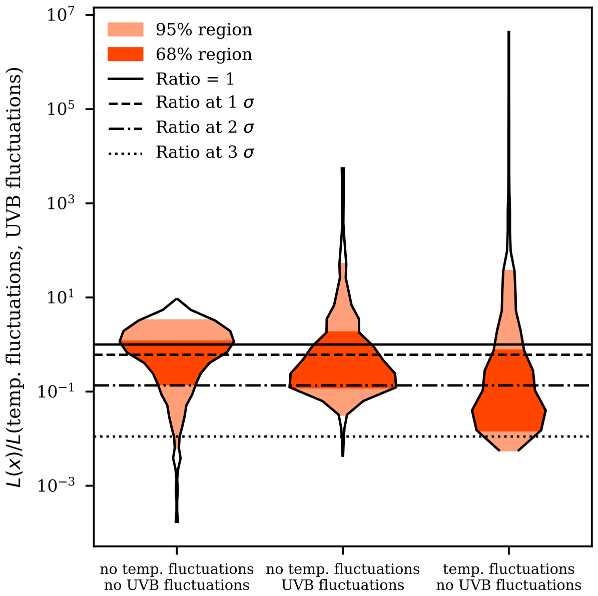

Figure 4.13: This figure shows the distribution of likelihood ratios from 1000 mock data sets where the mock data originates from the model with both temperature and UVB fluctuations. The violin plots show the full distribution where the light orange shaded region demarcates the 2.5th and 97.5th percentiles (2σ) of the ratio values while the darker orange shaded region demarcates the 16th and 84th percentiles (1σ) of the ratio values. The solid black line shows where the ratio is equal to 1, which is where both models are just as likely given the mock data. The dashed, dot-dashed, and dotted back lines show the value where the alternative models are ruled out at the 1, 2, and 3 σ levels respectively.

distinguish.

This is the distribution of the likelihood ratio for 1000 mock data sets. For a given observational data set, the luck of the draw would ultimately determine if it is possible to rule out each model. It is possible that the incorrect models is favored over the true model from which the mock data was drawn, though this was always true for less than half of the mock data sets.

4.5 Conclusions

In this work we have investigated the precision of possible constraints on the thermal state of the IGM from the auto-correlation function of Lyα forest flux in high resolution quasar observations. This came in two forms: constraining T_0 and γ when the IGM thermal state follows a tight power law of the form of Equation (4.1) and investigating the likelihood ration for models with temperature fluctuations from different reionization scenarios.

We discussed the results of constraints on T_0 and γ in Section 4.3. Overall, we found that the auto-correlation function is sensitive to T_0 and γ across multiple redshift bins for realistic mock data of 20 quasars with $R = 30000$. We computed the marginalized re-weighted posterior for T_0 and γ for 100 mock data sets at $5.4 \leq z \leq 6.0$. The re-weighted posterior showed a variety of behaviors based on the luck of the draw of the mock data chosen, the true value of T_0 and γ for the mock data, and the data set size at each z. We also considered an ideal data set which had the model value of the auto-correlation function, effectively removing the luck of the draw from our measurement. The error on these measurements for both the T_0 and γ increase with redshift, which may be from the low $\langle F \rangle$. At $z = 5.4$ we found that ideal data can constrain T_0 to 29% and γ to 9%.

Note that our procedure uses a multi-variate Gaussian likelihood, MCMC, and a set of weights for the MCMC chains that ensures our posteriors pass an inference test. The original failure of our procedure to pass an inference test is due to the incorrect assumption that the auto-correlation function follows a multi-variate Gaussian distribution, as discussed in Appendix 4.8. This result should caution against using a multi-variate Gaussian likelihood with other statistics, such as the power spectrum, when making measurements at $z > 5$ as the same issue of non-Gaussian data may appear. In the future, better likelihoods or likelihood-free inference will allow for a more optimal inference procedure (see e.g. Davies et al., 2018b; Alsing et al., 2019).

We discussed the likelihood ratios for four different reionization models in Section 4.4, assuming a Gaussian distribution of data. Looking at mock data from model which has temperature fluctuations and UVB fluctuations, we found that it is easiest to rule out a model with temperature fluctuations and no UVB fluctuations and it is most difficult to rule out a model with no temperature or UVB fluctuations. The actual ability to distinguish between models depends on the luck of the draw for the actual data that is measured. In this most difficult case, we found that 40.4% of mock data sets from the model with temperature and UVB fluctuations can rule out a model without temperature or UVB fluctuations at $> 1\sigma$ level.

Note that the UVB models used in this section were computed in in a small box (40 cMpc h^{-1}) which suppresses UVB fluctuations on all scales, as was discussed in Wolfson et al. (2023b). Suppressing fluctuations in the UVB causes the auto-correlation signal to be lower in these boxes. For this reason, it may be easier to distinguish between models with and without UVB fluctuations if they were generated in a larger box. Thus, future work on UVB models will be necessary to get the best possible constraints on reionization from these models.

Both the thermal state and the UVB fluctuations affect the Lyα forest flux auto-correlation function. Modeling both of these physical effects by varying multiple parameters in a larger box will allow the auto-correlation function to constrain the two simultaneously. This will allow us to put quantitative constraints on the thermal state of the IGM, the λ_{mfp} that describes the UVB, and ultimately reionization. We leave this exploration to future work.

This work assumed 20 high-resolution quasar observations in our forecasting. There are currently over 100 known quasars above a redshift of 6, a subset of which already have high resolution spectroscopic observations. Thus the 20 quasars used in this work is reasonable for a near-future observational constraint. In addition, the number of known quasars with high resolution observations is expected to continue to grow in the coming years which would only improve the prospects of this constraint.

Here we used the auto-correlation function of the Lyα forest flux. We chose to look at this clustering statistic for a couple statistical properties: namely that uncorrelated noise averages out and spectral masking is easy to implement. In comparison to the power spectra, the auto-correlation function has a covariance matrix with large off-diagonal values which makes it more difficult to intuitively look at the resulting fits to data (e.g. Figure 4.6). In addition to the intuition, these large off-diagonal values also make it more difficult to trouble-shoot the inference procedure using a Gaussian likelihood. Using a statistic with a more diagonal covariance matrix, like the power spectra, is easier to implement when fitting data.

Constraining the thermal state of the IGM, such as through constraining characteristic T_0 and γ values at high-z is an important method to constrain reionization. Measuring these parameters at $z > 5$ is a difficult task that has so far been done with few methods (Boera et al., 2019; Walther et al., 2019; Gaikwad et al., 2021). This work has shown that the auto-correlation function of the Lyα forest flux provides a new, competitive way to constrain T_0 and γ in multiple redshift bins at $z \geq 5.4$.

4.6 Appendix A: Power spectra models

As explained in Section 4.3.1, the dimensionless power, $\Delta^2_{\delta_f}(k)$, can be written as the Fourier transform of the auto-correlation function of the flux contrast, $\xi_{\delta_f}(\Delta v)$. $\xi_{\delta_f}(\Delta v)$ is explicitly written in terms of $\Delta^2_{\delta_f}(k)$ in Equation (4.5), which says $\xi_{\delta_f}(\Delta v)$ is the integral of $\Delta^2_{\delta_f}(k)\cos(k\Delta v)$ in logarithmic k bins. We refer to $\Delta^2_{\delta_f}(k)\cos(k\Delta v)$ as the integrand for the rest of this discussion. To build intuition for the auto-correlation function at small scales we show the integrand for $\Delta v = r = 15\,\mathrm{km\,s^{-1}}$ in Figure 4.14.

This Figure mimics the set up of Figure 4.4 for the auto-correlation function where each panels varies one parameter while keeping the others constant. For these panels the solid lines show the model values calculated by averaging Δ^2 from all forward modeled

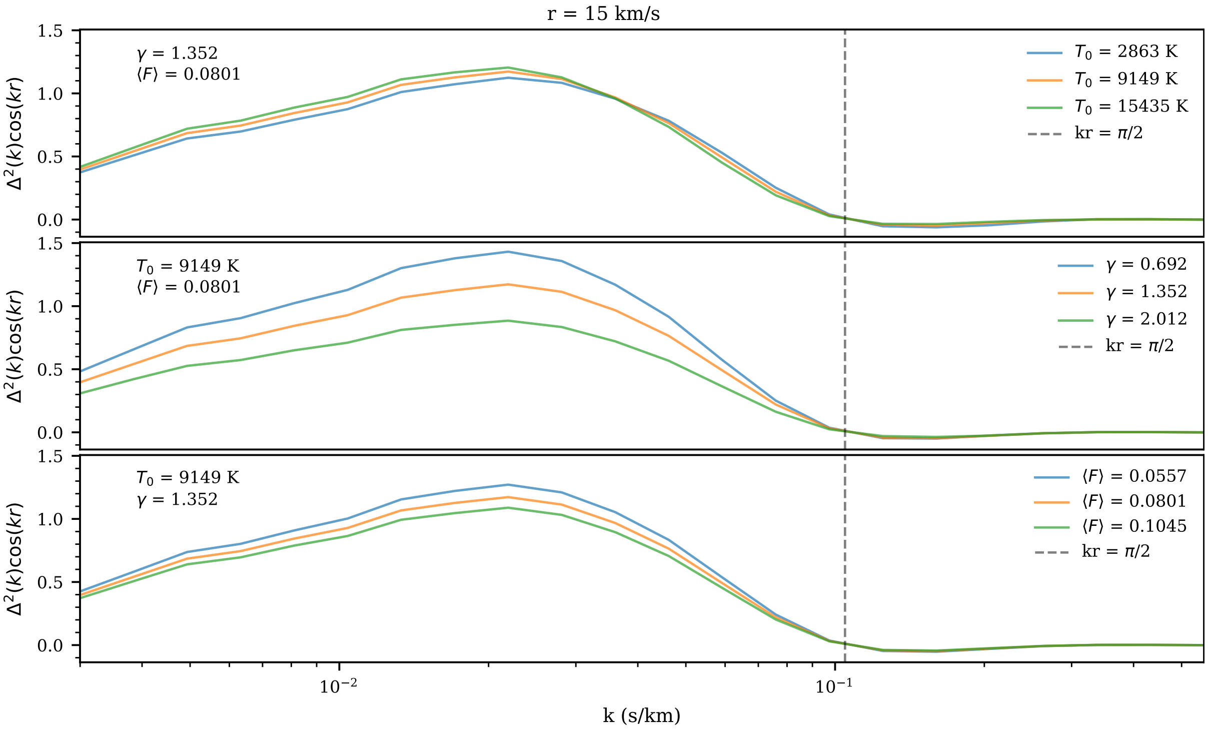

Figure 4.14: This figure shows the mean value of $\Delta^2 \cos(kr)$ where $r = 15\,\mathrm{km\,s^{-1}}$ for different sets of parameters. Each panels varies one parameter while keeping the others constant with T_0, γ, and $\langle F \rangle$ varying in the top, middle, and bottom panels respectively. For these panels the solid lines show the model values calculated by averaging Δ^2 from all forward modeled skewers available. This figure is meant to explain the behavior of the auto-correlation seen in Figure 4.4 at $\Delta v = r = 15\,\mathrm{km\,s^{-1}}$ due to the relation in Equation (4.5).

skewers available. The vertical grey dashed line shows where $\cos(kr) = 0$.

In the top panel T_0 varies while γ and $\langle F \rangle$ are constant. At small k the larger values of T_0 have larger values of the integrand while at small k there is thermal cutoff and smaller values of T_0 now have larger values of the integrand. When integrating over these logarithmic bins the greater T_0 values end up with more area and thus the auto-correlation functions are also greater.

4.7 Appendix B: Convergence of the Covariance Matrices

We calculate the covariance matrices for our models with mock draws, as defined in equation (4.6). Using mock draws is inherently noisy and it should converge as $1/\sqrt{N}$ where N is the number of draws used. As stated in the text, we used 500000 mock draws. To check that this number is sufficient to minimize the error in our calculation, we looked at the behavior of elements of one covariance matrix in Figure 4.15. This covariance matrix is for the model with $z = 5.4$, $T_0 = 9149$ K, $\gamma = 1.352$, and $\langle F \rangle = 0.0801$, which is the "true" model at this redshift. The correlation matrix for this model is also shown in Figure 4.5.

The values in the plot have been normalized to 1 at 10^6 draws. The four elements have been chosen such that there is one diagonal value and three off-diagonal values in different regions of the matrix. At all values of the number of mock draws considered, the covariance elements fall within 2% of their final value. By around ~ 100000 draws, the values fall within 0.5% of the final value. For this reason, using 500000 mock draws is sufficient to generate the covariance matrices used in this study. In Figure 4.15, 500000 mock draws is represented by the vertical dashed grey line.

4.8 Appendix C: Non-Gaussian distribution of the values of the auto-correlation function

For our inference, we used the multi-variate Gaussian likelihood defined in equation (4.8). This functional form assumes that the distribution of mock draws of the auto-correlation function is Gaussian distributed about the mean for each bin. In order to

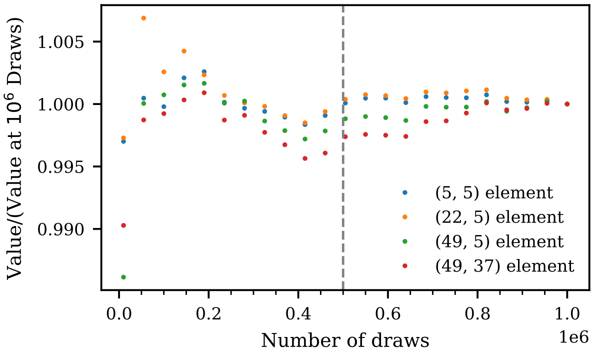

Figure 4.15: This figure shows the behavior of four elements of the model covariance matrix ($z = 5.4$, $T_0 = 9149$ K, $\gamma = 1.352$, and $\langle F \rangle = 0.0801$) for different numbers of mock draws. At all values of the number of mocks considered, the covariance elements fall within 2% of their final value. By around ~ 100000 draws, all of the values fall within 0.5% of the final value. For this reason, using 500000 mock draws is sufficient to generate the covariance matrices used in this study. 500000 mock draws is represented by the vertical dashed grey line.

visually check this we will look at the distribution of mock draws from two bins of the auto-correlation function for two different models.

Both Figures 4.16 and 4.17 show the distribution of 1000 mock data sets from the velocity bins of the auto-correlation function with $\Delta v = 25.0\,\mathrm{km\,s^{-1}}$ and $\Delta v = 65.0\,\mathrm{km\,s^{-1}}$. The bottom left panels show the 2D distribution of the auto-correlation values from these bins. The blue (green) ellipses represents the theoretical 68% (95%) percentile contour calculated from the covariance matrix calculated for each model from equation (4.6). The red crosses shows the calculated mean. The top panels show the distribution of only the $v = 65.0\,\mathrm{km\,s^{-1}}$ bins while the right panels show the distribution of only the $v = 25.0\,\mathrm{km\,s^{-1}}$ bins.

Figure 4.16 shows mock values of two bins of the auto-correlation function for the model at $z = 5.4$ with $T_0 = 9148$ K, $\gamma = 1.352$, and $\langle F \rangle = 0.0801$. Both the 1D and 2D distributions seem relatively well described by Gaussian distributions by eye though they do show some evidence of non-Gaussian tails to larger values. The number of points falling in each contour both fall within 1% of the expected values. In the bottom left panel with the 2D distribution there are more mock values falling outside the 95% contour to the top right (higher values) than in any other direction. For this reason the distribution is not exactly Gaussian but a Gaussian visually appears as an acceptable approximation.

Figure 4.17 shows mock values of two bins of the auto-correlation function for the model at $z = 6$ with $T_0 = 10600$ K, $\gamma = 1.262$, and $\langle F \rangle = 0.0089$. In both the top and right panels, which show the distribution of values for one bin of the auto-correlation function, the distribution of mock draws is skewed with tails to the right. This is quantified by the percent of points in the two ellipses from the bottom left panel labeled in the top right with 72.3% of the mock draws falling within the 68% contour and 94.0% of the mock draws falling within the 95% contour. The points outside of the contours are highly skewered towards the top right (higher values). It is only possible for the auto-correlation function to be negative due to noise, which generally averages to very small values approaching zero at the non-zero lags of the auto-correlation function. This can be seen in the black points and histogram do not go below 0, though the 95% ellipse shown in green in the bottom left panel does go negative for $\Delta v = 65\,\mathrm{km\,s^{-1}}$.

Figures 4.16 and 4.17 show the changing distribution of the auto-correlation value with z, T_0, γ, and $\langle F \rangle$. There is a greater deviation from a multi-variate Gaussian distribution at higher z. It is possible that adding additional sightlines will cause the auto-correlation function to better follow a multi-variate Gaussian distribution due to the central limit theorem, though investigating this in detail is beyond the scope of the paper. However, even with more sightlines $\langle F \rangle$ will be low at high-z so we still expect the distribution to be skewed as the values mainly will not fall below 0. The incorrect assumption of the multi-variate Gaussian likelihood thus contributes to the failure of our method to pass an inference test as discussed in Section 4.3.3 for $z = 5.4$ and Appendix 4.9 for $z = 6$. For our final constraints, we calculated weights for our MCMC chains such that the resulting posteriors do pass our inference test, as discussed in Section 4.3.3. The whole method of assuming a multi-variate Gaussian then re-weighting the posteriors in

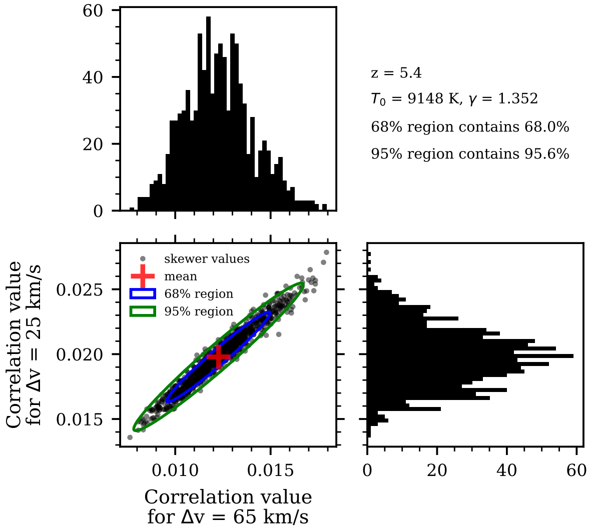

Figure 4.16: This figure shows the distribution 1000 mock draws from two bins of the auto-correlation function ($\Delta v = 25.0\,\mathrm{km\,s^{-1}}$ and $\Delta v = 65.0\,\mathrm{km\,s^{-1}}$) for one model ($z = 5.4$, $T_0 =$ 9148 K, $\gamma = 1.352$, and $\langle F \rangle = 0.0801$). The top panel shows the distribution of only the $\Delta v = 65.0\,\mathrm{km\,s^{-1}}$ bin while the right panel shows the distribution of only the $\Delta v = 25.0\,\mathrm{km\,s^{-1}}$ bin. The blue (green) circle represents the 68% (95%) ellipse calculated from the covariance matrix calculated for this model from equation (4.6). The red plus shows the calculated mean. Additionally the percent of mock draws that fall within each of these contours is written in the top right. Both the 1D and 2D distributions seem relatively well described by a Gaussian distribution. In the 2D plot, there are more points outside the 95% contour to the top right than on any other side but it is not extreme.

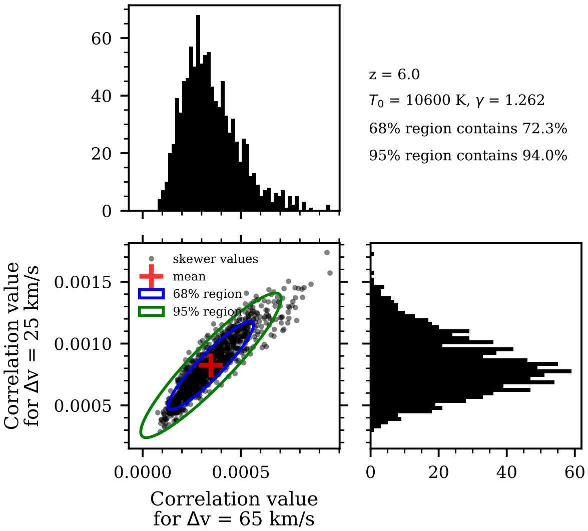

Figure 4.17: This figure shows the distribution 1000 mock draws from two bins of the auto-correlation function ($\Delta v = 25.0\,\mathrm{km\,s^{-1}}$ and $\Delta v = 65.0\,\mathrm{km\,s^{-1}}$) for one model ($z = 6$, $T_0 =$ 10600 K, $\gamma = 1.262$, and $\langle F \rangle = 0.0089$). The top panel shows the distribution of only the $\Delta v = 65.0\,\mathrm{km\,s^{-1}}$ bin while the right panel shows the distribution of only the $\Delta v = 25.0\,\mathrm{km\,s^{-1}}$ bin. The blue (green) circle represents the 68% (95%) ellipse calculated from the covariance matrix calculated for this model from equation (4.6). The red plus shows the calculated mean. Additionally the percent of mock draws that fall within each of these contours is written in the top right. Both the 1D and 2D distributions are not well described by a Gaussian with 72.3% of the mock draws falling within the 68% contour and 94.% of the mock draws falling within the 95% contour.

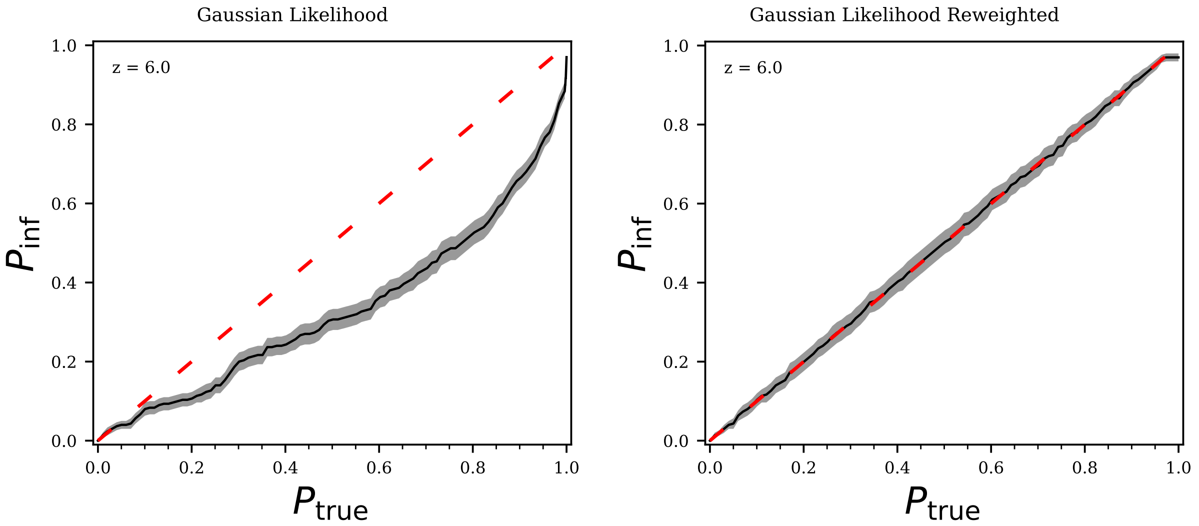

Figure 4.18: The left panel of this figure shows the coverage resulting from the inference test from 300 models at $z = 6.$ drawn from our priors on T_0, γ, and $\langle F \rangle$. Here we see that the true parameters for the models fall above the 60th percentile in the MCMC chain $\sim 35\%$ of the time, for example. The right panel of this figure shows the coverage resulting from the inference test with the use of one set of weights to re-weight the posteriors. With these weights we are able to pass the inference test.

non-optimal and future work using a more correct likelihood or likelihood-free inference will improve our results.

4.9 Appendix D: Inference test at high redshift

Here we present the results of the inference test at $z = 6$. This calculation was done following the procedure described in Section 4.3.3. Figure 4.18 shows the results for $z = 6$ and can be compared to the $z = 5.4$ results in Figure 4.7. The left panel here shows the initial coverage plot which deviates greatly from the expected $P_{\rm inf} = P_{\rm true}$ line, much more so than the $z = 5.4$. This likely comes from a greater deviation from the assumption of a multi-variate Gaussian likelihood as described in Appendix 4.8. The $z = 6$ mock data show highly skewed distributions that are not well described by a Gaussian likelihood. The inference lines at other redshifts are available upon request.

4.10 Appendix E: Gaussian data inference test

As shown in Appendix 4.8, the distribution of mock values of the auto-correlation function is not exactly Gaussian distributed. In order to confirm the failure of our mock data to pass an inference test (as discussed in Section 4.3.3 and Appendix 4.9) comes from the use of a multi-variate Gaussian likelihood, we generate Gaussian distributed data and run inference tests. For one value of T_0, γ, and $\langle F \rangle$, we randomly generate a mock data set from a multi-variate Gaussian with the given mean model and covariance matrix calculated from our mock data as described in Section 4.3.1. We can then continue with the inference test as described in Section 4.3.3. The results for this inference test for $z = 5.4$ and $z = 6.0$ are shown in Figure 4.19. Here both redshifts inference lines fall along the 1-1 line that is expected for all probability contour, $P_{\rm true}$, values. This behavior is also seen at the other redshifts. The fact that perfectly Gaussian data passes an inference test with the same likelihood, priors, and method as was used on mock data confirms that the failure of our mock data to pass an inference test is due to the non-Gaussian distribution of the mock data.

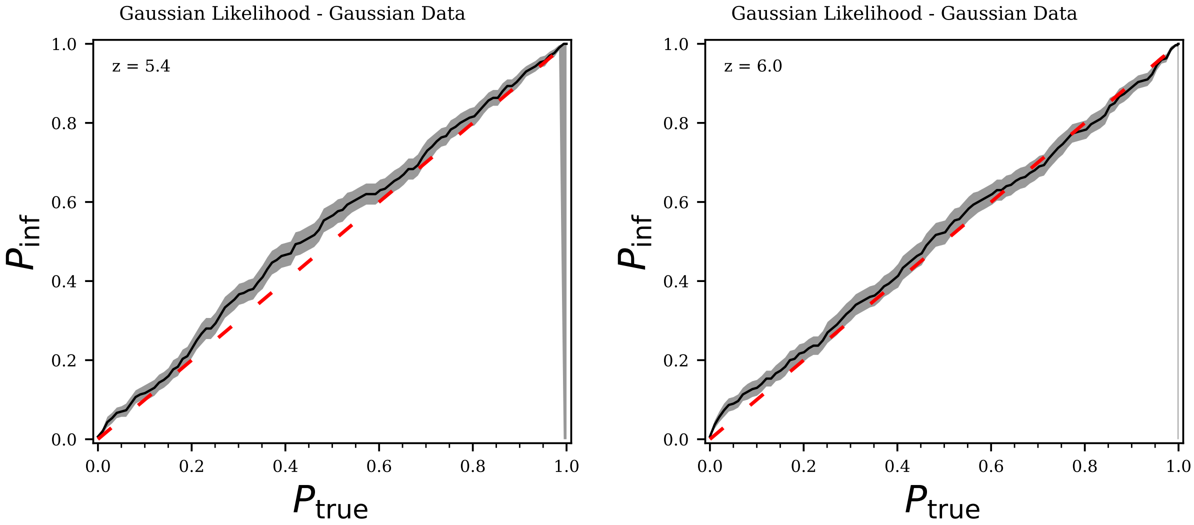

Figure 4.19: Both panels of this figure shows the coverage plot resulting from the inference test from 300 data sets generated by randomly drawing points from the mean model and covariance matrix. The the means and covariance matrices used come from $z = 5.4$ in the left panel and $z = 6.0$ in the right panel. The true parameter values for both panels were drawn from our priors on T_0, γ, and $\langle F \rangle$. In both panels, the Gaussian mock data produced inference lines that fall on top of the 1-1 line within errors, as expected for the statistically correct posteriors.

Chapter 5

Measurements of the $z > 5$ Lyman-α forest flux auto-correlation functions from the extended XQR-30 data set

This chapter was reproduced from Wolfson et al. (2023c) with only minor changes to fit the formatting of this dissertation. I'd like to thank my coauthors, without whom this work would not have been possible: Joseph F. Hennawi, Sarah E. I. Bosman, Frederick B. Davies, Zarija Lukić, George D. Becker, Huanqing Chen, Guido Cupani, Valentina D'Odorico, Anna-Christina Eilers, Martin G. Haehnelt, Laura C. Keating, Girish Kulkarni, Samuel Lai, Andrei Mesinger, Fabian Walter, and Yongda Zhu.

5.1 Introduction

The reionization of the neutral hydrogen in the intergalactic medium (IGM) is one of the major phase changes in our Universe's history. Understanding the timing of this process has been the focus of many recent studies. Current Planck constraints put the midpoint of reionization at $z_{\rm re} = 7.7 \pm 0.7$ (Planck Collaboration et al., 2020) with mounting evidence that it was not completed until after $z \leq 6$ (Fan et al., 2006; Becker et al., 2015, 2018; Bosman et al., 2018, 2022; Eilers et al., 2018; Boera et al., 2019; Yang et al., 2020; Jung et al., 2020; Kashino et al., 2020; Morales et al., 2021).

Before the end of reionization, the mean free path of hydrogen-ionizing photons ($\lambda_{\rm mfp}$) is expected to be short due to the significant neutral hydrogen remaining in the IGM which will absorb these photons close to their sources. In some models, as reionization ends $\lambda_{\rm mfp}$ will rapidly increase due to the overlap of initially isolated ionized bubbles and the photo-evaporation of dense photon sinks (Gnedin, 2000; Shapiro et al., 2004; Furlanetto & Oh, 2005; Gnedin & Fan, 2006; Wyithe et al., 2008; Sobacchi & Mesinger, 2014; Park et al., 2016; Kulkarni et al., 2019; Keating et al., 2020b,a; Nasir & D'Aloisio, 2020; Cain et al., 2021; Gnedin & Madau, 2022). Thus detecting an increase in $\lambda_{\rm mfp}$ will provide insights into the end of reionization.

Becker et al. (2021) reported the first direct measurement of $\lambda_{\rm mfp}$ at $z \sim 6$ from stacked quasar spectra. Zhu et al. (2023) updated this measurement and added two additional redshift bins at $z = 5.31$ and $z = 5.65$. They found that $\lambda_{\rm mfp} = 9.33^{+2.06}_{-1.80}$, $5.40^{+1.47}_{-1.40}$, $3.31^{+2.74}_{-1.34}$, and $0.81^{+0.73}_{-0.48}$ pMpc at z =5.08, 5.31, 5.65, and 5.93, respectively. Becker et al. (2021) and Zhu et al. (2023) expanded on previous measurements of $\lambda_{\rm mfp}$ at $z \leq 5.1$ (Prochaska et al., 2009; Fumagalli et al., 2013; O'Meara et al., 2013; Worseck et al., 2014). The Zhu et al. (2023) measurement has $\lambda_{\rm mfp}$ rapidly increasing between $z = 6$ and $z = 5.1$, potentially signalling the end of reionization. The values at $z \geq 5.3$ are significantly smaller than extrapolations from previous lower z measurements (Worseck et al., 2014) based on a fully ionized IGM. In addition, the value at $z \sim 6$ may cause tension with measurements of the ionizing output from galaxies (Cain et al., 2021; Davies et al., 2021).

Alternative methods to constrain $\lambda_{\rm mfp}$ are needed to check the measurements discussed above and to constrain the timing of reionization in finer redshift bins. One such method from Bosman (2021) used lower limits on individual free paths (the distance ionizing radiation travels from an individual source) towards high-z sources to place a 2σ limit of $\lambda_{\mathrm{mfp}} > 0.31$ proper Mpc at $z = 6.0$. This Bosman (2021) method is similar to other measurements using individual free paths (Songaila & Cowie, 2010; Rudie et al., 2013; Romano et al., 2019). Additionally, Gaikwad et al. (2023) constrained λ_{mfp} for $4.9 < z < 6.0$ with $\Delta z = 0.1$ by comparing the observed probability distribution function of the Lyα optical depth to predictions from simulations with a fluctuating ultraviolet background (UVB) driven by a short λ_{mfp}. The measurement of λ_{mfp} at $z < 5.1$ in Gaikwad et al. (2023) shows a good agreement with the measurements from Worseck et al. (2014) and Becker et al. (2021). At $z = 6.0$ Gaikwad et al. (2023) measured $\lambda_{\mathrm{mfp}} = 8.318^{+7.531}_{-4.052}$ comoving Mpc (cMpc) h^{-1}, which agrees with the Zhu et al. (2023) measurement at the 1.2σ level and also falls above the lower limit found by Bosman (2021).

The level of fluctuations in the UVB are set by the distribution of ionizing photon sources and λ_{mfp}. For large values of λ_{mfp}, photons travel further from their sources and effectively creates a more uniform UVB (Mesinger & Furlanetto, 2009). Alternatively, small values of λ_{mfp} lead to greater fluctuations in the UVB, causing some regions to have very large Γ_{HI} values. These fluctuations then imprint themselves on the Lyα forest flux transmission in high-z quasar spectra via the Lyα opacity, $\tau_{\mathrm{Ly}\alpha}$ where $\tau_{\mathrm{Ly}\alpha} = n_{\mathrm{HI}}\sigma_{\mathrm{Ly}\alpha} \propto 1/\Gamma_{\mathrm{HI}} \propto 1/\lambda_{\mathrm{mfp}}^{\alpha}$ where $3/2 < \alpha < 2$ (see e.g. Rauch, 1998; Haardt & Madau, 2012). Many previous studies have investigated the effect of large scale variations in the UVB on the structure of the Lyα forest (Zuo, 1992a,b; Croft, 2004; Meiksin & White, 2004; McDonald et al., 2005; Gontcho A Gontcho et al., 2014; Pontzen, 2014; Pontzen et al., 2014; D'Aloisio et al., 2018; Meiksin & McQuinn, 2019; Oñorbe et al., 2019). This is similar to the argument explored by Gaikwad et al. (2023) in using the probability distribution function of the Lyα optical depth to constrain $\lambda_{\rm mfp}$. The probability distribution function of the Lyα optical depth does not consider the 2-point clustering, which can be quantified through the auto-correlation function and the power spectrum, which is the Fourier transform of the auto-correlation function, of the Lyα forest flux. Beyond the effect of UVB fluctuations, the power spectrum of the Lyα forest flux contrast has been measured at high z and used to constrain the thermal state of the IGM (Boera et al., 2019; Walther et al., 2019; Gaikwad et al., 2021) as well as warm dark matter particle mass (Viel et al., 2013; Iršič et al., 2017; Garzilli et al., 2017).

This work is specifically building on Wolfson et al. (2023b) which investigated the effect of a fluctuating UVB on small scales in Lyα forest transmission at $z \geq 5.4$. They found that the Lyα forest transmission on small scales will be boosted for small values of $\lambda_{\rm mfp}$ and that this can be quantified with the Lyα forest flux auto-correlation function. They used the auto-correlation function to recover $\lambda_{\rm mfp}$ from simulated mock data. The Lyα forest flux auto-correlation function has yet to be measured at $z \gtrsim 5.5$ for observational data. Many previous studies have measured the Lyα forest flux auto-correlation function at lower redshifts for a wide range of applications (McDonald et al., 2000; Rollinde et al., 2003; Becker et al., 2004; D'Odorico et al., 2006).

In this paper we use the XQR-30 extended data set to measure the Lyα forest flux auto-correlation function. We discuss this observational data in Section 5.2. The details on the data selection and measurement process with a full account of relevant errors are described in Section 5.3. We then discuss our resulting measurements in Section 5.4 and some preliminary comparisons to simulations in Section 5.5. We summarize our results in Section 5.6.

5.2 Data

The quasar spectra used in this work are a subset of those presented in Bosman et al. (2022). The data reduction was performed and discussed in detail there but will be summarized again in this work for the sake of completeness. Additionally, more information on the continuum reconstructions can be found in Bosman et al. (2021).

All of the observations used in this work comes from the XQR-30 program[1] (1103.A0817(A), D'Odorico et al., 2023), which consists of a sample of 30 very luminous quasars at $z \gtrsim 5.8$ observed with the X-Shooter instrument (Vernet et al., 2011b) on ESO's Very Large Telescope. We use 24 quasars from the XQR-30 sample which do not show strong broad absorption lines (BALs) that would create issues in the modelling of the intrinsic continuum (Bischetti et al., 2022) and could also possibly contaminate the Lyα forest region. Three additional spectra (PSO J231-20, ATLAS J2211-3206, and SDSS J2310+1855) were identified as hi-BALs so we exclude regions of the spectra where there is possible strong OVI contamination (7770Å $< \lambda_{\rm obs} <$ 7870Å, $\lambda_{\rm obs} <$ 7280Å, and $\lambda_{\rm obs} <$ 6700Å respectively). All XQR-30 spectra have signal-to-noise ratios (SNRs) larger than 20 per $10\,{\rm km\,s^{-1}}$ pixel measured over 1165Å $< \lambda_{\rm rest} <$ 1170Å (Table 5.1). In addition to the 24 XQR-30 quasars, we use 11 archival X-Shooter spectra that are from the extended XQR-30 sample (D'Odorico et al., 2023). These spectra have SNR ¿ 40 per $10\,{\rm km\,s^{-1}}$ pixel from the literature (Table 5.1, marked with *). The extended XQR-30 sample has a median effective resolving power over all 42 quasars of $R \simeq 11400$ and 9800 in the visible (5500Å $< \lambda_{\rm obs} <$ 10200Å) and infrared arm (10200Å $< \lambda_{\rm obs} <$ 24800Å) of X-Shooter, respectively (D'Odorico et al., 2023).

All quasars are reduced with the same procedure. Observations are first flat-fielded and sky-subtracted following the method of Kelson (2003). The spectra are extracted

[1] https://xqr30.inaf.it/

Table 5.1: The extended XQR-30 quasars included in this work. Those with a * represent the extended data set quasars which did not get new spectra in the XQR-30 program. References correspond to: discovery, redshift determination.

Quasar ID	$z_{\rm qso}$	SNR pix^{-1}	Refs.
PSO J323+12	6.5872	35.9	Mazzucchelli et al. (2017), Venemans et al. (2020)
PSO J231-20	6.5869	42.3	Mazzucchelli et al. (2017), Venemans et al. (2020)
VDES J0224-4711	6.5223	24.4	Reed et al. (2017), Wang et al. (2021)
PSO J036+03*	6.5405	61.4	Venemans et al. (2015), Venemans et al. (2020)
PSO J1212+0505	6.4386	55.8	Mazzucchelli et al. (2017), Decarli et al. (2018)
DELS J1535+1943	6.3932	22.6	Wang et al. (2019), Bosman et al. (2022)
ATLAS J2211-3206	6.3394	37.5	Chehade et al. (2018)/Farina et al. (2019), Decarli et al. (2018)
SDSS J0100+2802*	6.3269	560.5	Wu et al. (2015), Venemans et al. (2020)
ATLAS J025-33*	6.318	127.3	Carnall et al. (2015), Becker et al. (2019)
SDSS J1030+0524*	6.309	69.6	Fan et al. (2001), Jiang et al. (2007)
PSO J060+24	6.192	49.7	Bañados et al. (2016), Bosman et al. (2022)
PSO J065-26	6.1871	77.9	Bañados et al. (2016), Venemans et al. (2020)
PSO J359-06	6.1722	68.8	Wang et al. (2016), Eilers et al. (2021)
PSO J217-16	6.1498	73.0	Bañados et al. (2016), Decarli et al. (2018)
ULAS J1319+0950*	6.1347	81.7	Mortlock et al. (2009), Venemans et al. (2020)
CFHQS J1509-1749*	6.1225	43.0	Willott et al. (2007), Decarli et al. (2018)
PSO J239-07	6.1102	56.3	Bañados et al. (2016), Eilers et al. (2021)
SDSS J0842+1218	6.0754	83.2	De Rosa et al. (2011)/Jiang et al. (2015), Venemans et al. (2020)
ATLAS J158-14	6.0685	60.3	Chehade et al. (2018), Eilers et al. (2021)
VDES J0408-5632	6.0345	86.6	Reed et al. (2017), Reed et al. (2017)
SDSS J1306+0356*	6.033	65.3	Fan et al. (2001), Venemans et al. (2020)
ATLAS J029-36	6.021	57.1	Carnall et al. (2015), Becker et al. (2019)
SDSS J2310+1855	6.0031	113.4	Jiang et al. (2016), Wang et al. (2013)
PSO J007+04	6.0015	54.4	Jiang et al. (2015)/Bañados et al. (2014), Venemans et al. (2020)
ULAS J0148+0600*	5.998	152.0	Jiang et al. (2015), Becker et al. (2019)
SDSS J0818+1722*	5.997	132.1	Fan et al. (2006), Becker et al. (2019)
PSO J029-29	5.984	65.6	Bañados et al. (2016), Bañados et al. (2016)
PSO J108+08	5.9485	104.8	Bañados et al. (2016), Bañados et al. (2016)
PSO J183-12	5.917	61.8	Bañados et al. (2014), Bosman et al. (2022)
PSO J025-11	5.844	50.6	Bañados et al. (2016), Bosman et al. (2022)
PSO J242-12	5.837	22.9	Bañados et al. (2016), Bosman et al. (2022)
PSO J065+01	5.833	25.1	D'Odorico et al. (2023), Bosman et al. (2022)
SDSS J0836+0054*	5.804	73.8	Fan et al. (2001), Bosman et al. (2022)
PSO J308-27	5.7985	53.2	Bañados et al. (2016), D'Odorico et al. (2023)
SDSS J0927+2001*	5.7722	53.8	Fan et al. (2006), Wang et al. (2010)

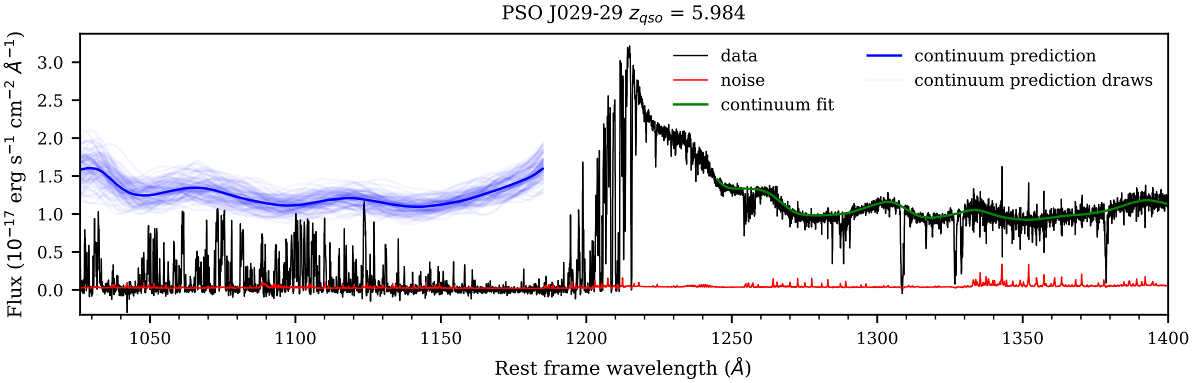

Figure 5.1: The X-Shooter spectrum of the Lyα transmission region for the quasar PSO J029-29 from the XQR-30 sample. The noise vector is shown in red and the PCA-reconstructed continuum is shown in blue. The light blue lines show draws of the continuum reconstruction with the appropriate scatter from the covariance matrix of the PCA reconstruction. The pixel scale is $10\,\mathrm{km\,s^{-1}}$ and the SNR of the Lyα region (reconstruction divided by uncertainty) is SNR = 50.6.

(Horne, 1986) separately for the visible and infrared arms of the instrument which are then stitched together over the $10110\text{Å} < \lambda_{\rm obs} < 10130\text{Å}$ spectral window. The infrared spectrum is re-scaled to match the observed mean flux in the optical arm. The spectrum is then interpolated over the overlap window in order to minimize the risk of creating an artificial step in the spectrum between the arms to which the continuum-fitting method may be non-linearly sensitive (see discussion in Bosman et al., 2022). The reduction routines are described in more detail in Becker et al. (2009). Further details are presented in D'Odorico et al. (2023).

An example spectrum from the program is shown in Figure 5.1 for PSO J029-29. The black line shows the reduced XQR-30 spectrum and the red line shows the noise vector. The intrinsic continuum reconstructed with the method described in Section 5.3.1 is shown by the solid blue line, while the continuum fit to the red side of the quasars emission is shown in green. The light blue lines show draws of the continuum reconstruction with the appropriate scatter from the covariance matrix of the PCA reconstruction. The sampling procedure for these draws are also discussed in Section 5.3.1.

5.3 Methods

5.3.1 Continuum reconstruction

For each quasar, the continuum, $F_{\text{cont}}(\lambda_{\text{rest}})$, was reconstructed using Principal Component Analysis (PCA). To do this, we consider both the red side ($\lambda_{\text{rest}} > 1280$Å) and the blue side ($\lambda_{\text{rest}} < 1220$Å) of the quasar continuum with respect to the Lyα emission. At low-z, both sides of the quasar continuum are transmitted through the IGM, as the IGM is mainly ionized. Thus we can use PCA to find the optimal linear decomposition of both the red side and the blue side of the low-z quasar continuum, then construct an optimal mapping between the the linear coefficients from the two decompositions. At high-z, the red side of quasar continua will be transmitted while the blue side is absorbed by remaining neutral hydrogen in the IGM, see e.g. Figure 5.1. We can thus get the PCA decomposition for the red side of the continuum then use the optimal mapping, determined from low-z quasars, to predict the blue side coefficients and thus the continuum (Francis et al., 1992; Yip et al., 2004). This method has been historically used to get the continuum for the Lyα forest in Suzuki et al. (2005) then it was further expanded, for example by: McDonald et al. (2005); Pâris et al. (2011); Davies et al. (2018 d,c); Ďurovčíková et al. (2020). Previously, Bosman et al. (2021) determined the most accurate PCA method and Bosman et al. (2022) further improved this method with the log-PCA approach of Davies et al. (2018d,c).

This work uses the same reconstructions that were generated for Bosman et al. (2022) using the log-PCA approach. The PCA consists of 15 red-side components and 10 blue-side components. The training set amounted to 4597 quasars from the SDSS-III Baryon Oscillation Spectroscopic Survey (BOSS, Dawson et al., 2013) and the SDSS-IV Extended BOSS (eBOSS, Dawson et al., 2016) at $2.7 < z < 3.5$ with SNR ¿ 7. Intrinsic continua were obtained automatically using a modified version of the method of Dall'Aglio et al.

(2008), originally based on the procedures outlined in Young et al. (1979) and Carswell et al. (1982). These continua are re-normalized so that they match the observed mean Lyα transmission at $z \sim 3$ that was measured from high-resolution spectra (Faucher-Giguère et al., 2008; Becker et al., 2013) to prevent bias from the low spectral resolution of the SDSS spectrograph (as described in Dall'Aglio et al., 2009). The reconstructions were tested with an independent set of 4597 quasars from eBOSS. As described in Bosman et al. (2022), this testing revealed that there is no bias in reconstructing the blue-side emission lines and that the method predicts the underlying continuum within 8%. The reconstruction error on this testing set gives us the mean, $\boldsymbol{\mu}_{\rm cont}$, and covariance, $\boldsymbol{\Sigma}_{\rm cont}$, of the PCA reconstruction as shown in Figure 2 of Bosman et al. (2022).

In the following steps, we always forward-model the full wavelength-dependent uncertainties from the reconstruction of $F_{\rm cont}(\lambda_{\rm rest})$ into all measurements and model comparisons. We do this by randomly drawing realizations of the continuum error, $\boldsymbol{E}_{\rm cont} \sim N(\boldsymbol{\mu}_{\rm cont}, \boldsymbol{\Sigma}_{\rm cont})$, where N is the normal distribution. We create a realization of the predicted continuum with this error, $\boldsymbol{C}_{\rm pred}$, from the fit quasar continuum, $\boldsymbol{C}_{\rm fit}$, via:

$$\boldsymbol{C}_{\rm pred} = \boldsymbol{C}_{\rm fit} \times \boldsymbol{E}_{\rm cont}. \tag{5.1}$$

We use 500 of these continuum draws to analyze each quasar's spectrum. When we performed bootstrap re-sampling as described in Section 5.4.3, each draw uses a random selection of these 500 continua. Figures showing all PCA fits and blue-side predictions for all extended XQR-30 quasars are shown in Zhu et al. (2021).

5.3.2 Pixel masking

We want to use flux from the quasar continuum that exclusively corresponds to Lyα forest absorption. To do this, we only use wavelengths larger than the Lyβ emission at the redshift of the quasar, or $\lambda_{\rm rest} > 1026$Å. Additionally, we want to exclude the quasars proximity zone, which is the region close to the quasar where the IGM has been ionized by the quasar's own emission and the transmission is enhanced. For this reason, we consider $\lambda_{\rm rest} < 1185$Å following Bosman et al. (2022) which corresponds to $\sim 7650\,{\rm km\,s^{-1}}$ from emission at $z \sim 6$. This is a conservative estimate based on Bosman et al. (2018), which found no effect on the Lyα transmission in spectral stacks over this wavelength.

The data reduction procedure should automatically reject outlier pixels. However, we check for and exclude anomalous pixels that meet either of the following conditions: the SNR at the unabsorbed continuum level is ¡ 2 per pixel or if pixels have negative flux at $> 3\sigma$ significance. This excludes 0% of pixels for the SNR cut at all redshifts and $0.07 - 0.47$% of pixels for the negative flux cut depending on redshift.

5.3.3 DLA exclusion

Damped Lyα absorption systems (DLAs) are intervening systems along quasar sightlines with hydrogen column densities $N_{\rm HI} \geq 10^{20.3}$ cm^{-2}. These systems result in significant damping wings in the Lyα absorption profile (Wolfe et al., 2005; Rafelski et al., 2012). DLAs in quasar spectra at $z \gtrsim 6$ can cause complete absorption of Lyα transmission over $\Delta v = 2000\,{\rm km\,s^{-1}}$ and additional suppression over $\Delta v \gtrsim 5000\,{\rm km\,s^{-1}}$ intervals (D'Odorico et al., 2018; Bañados et al., 2019; Davies, 2020). DLAs can arise in the circumgalactic medium (CGM) of galaxies which are not typically included in reionization simulations, including those discussed in Section 5.5. For this reason, we attempt to remove DLAs from our observations based on the presence of metals in the spectra. This does leave open the possibility that DLAs from pristine neutral patches of the IGM remain in our observations.

We remove DLAs by identifying and masking out their locations in our spectra. The detection of $z \gtrsim 5$ DLAs relies on the identification of associated low-ionization metal absorption lines, since the Lyα absorption from the DLA may not be distinguishable from the highly-opaque IGM. The typical transitions are CII, OI, SiII, and MgII. DLA metallicities at $z \gtrsim 5$ vary so even relatively weak metal absorption could indicate a DLA. The identification of intervening metal absorbers in the extended XQR-30 sample has been described in detail in Davies et al. (2023) and Sodini et al. (2024). Due to the high SNR of the X-Shooter spectra, we expect to be ¿ 90% complete to absorption corresponding to $\log N_{\rm MgII}/{\rm cm}^{-2} \gtrsim 13$.

We adopt the following criteria for our masks, following Bosman et al. (2022). We mask the central $\Delta v = 3000\,{\rm km\,s^{-1}}$ for systems with metal column densities $\log N_{\rm CII}/{\rm cm}^{-2} > 13$, $\log N_{\rm OI}/{\rm cm}^{-2} > 13$, or $\log N_{\rm SiII}/{\rm cm}^{-2} > 12.5$, measured through the $\lambda_{\rm rest} = 1334.53$Å, 1302.16Å, and 1526Å transitions, respectively. When none of these ions are accessible, we also exclude the central $\Delta v = 3000\,{\rm km\,s^{-1}}$ for systems with $\log N_{\rm MgII}/{\rm cm}^{-2} > 13$ based on the high rates of co-occurrence of the MgII 2796.35, 2803.53Å doublet (Cooper et al., 2019). We exclude a larger window of $\Delta v = 5000\,{\rm km\,s^{-1}}$ around intervening systems with $\log N_{\rm OI,\,CII,\,SiII,\,MgII}/{\rm cm}^{-2} > 14$ due to the likely presence of extended damping wings. We do not exclude systems based on the presence of highly ionized ions alone (e.g. C IV, Si IV) since the corresponding gas is likely highly ionized (Cooper et al., 2019).

We investigate the effect of this mask on the measurement of the auto-correlation function in Appendix 5.8.

5.3.4 Resulting normalized flux

After combining the masks of the bad pixels discussed in Section 5.3.2 and the DLAs discussed in Section 5.3.3, we only considered sightlines that maintain at least 10% of the pixels in a given redshift bin. Only using spectra that maintain at least 10% of pixels limits noisy contributions to the measurement from short spectra that may only consist

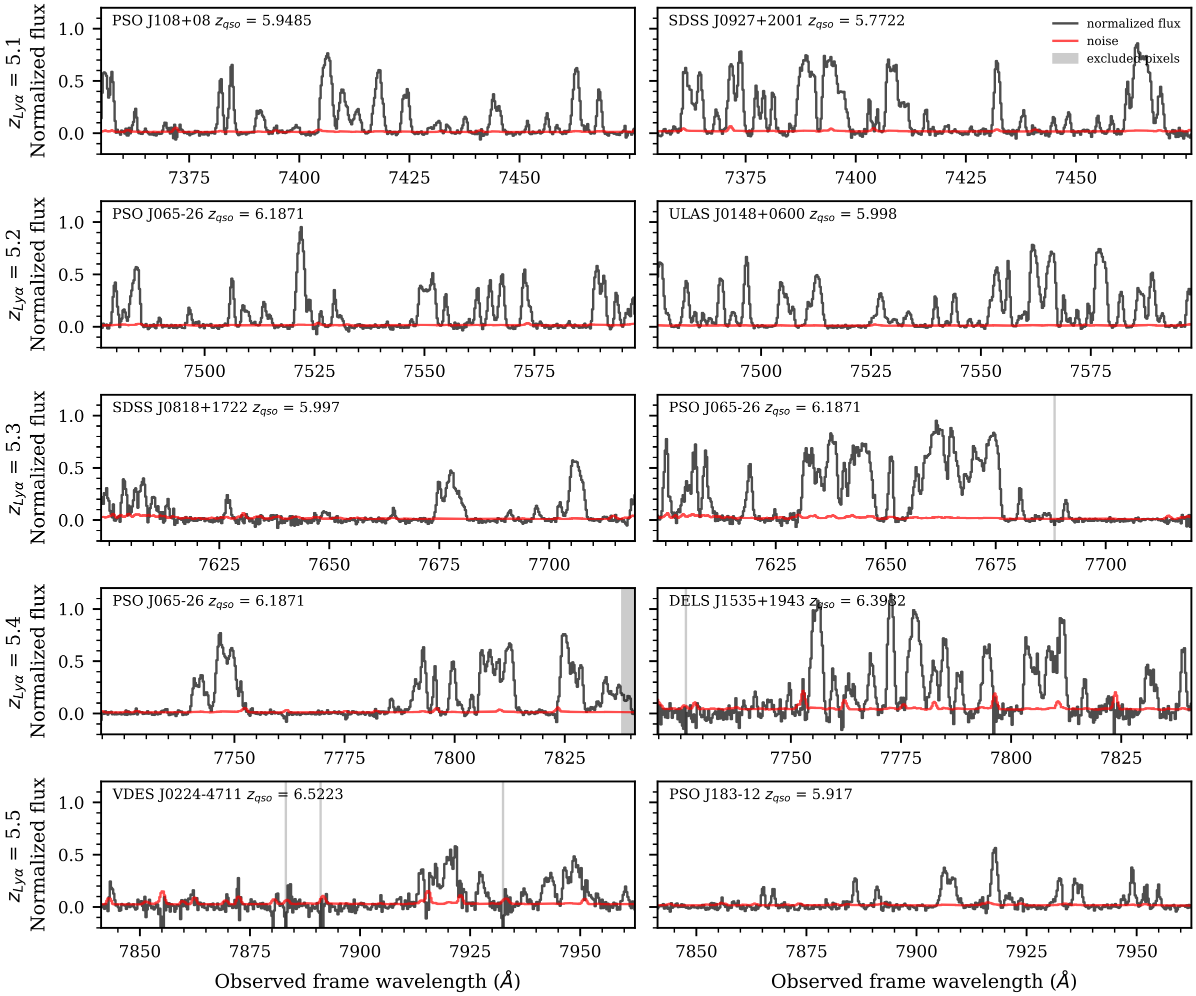

Figure 5.2: This figure shows the continuum normalized flux for two randomly selected quasars at five values of z of the Lyα forest from $5.1 \leq z \leq 5.5$. These sections are centered on the given z and span $\Delta z = 0.05$. The continuum normalized flux is shown in black with the continuum normalized uncertainty in red. The shaded regions indicate excluded pixels based on the masking procedure described in Section 5.3.2 and 5.3.3. Each row shares the same y-axis to demonstrate the decrease in $\langle F \rangle$ with increasing z (down the rows). Note that the normalized flux for all the quasars considered in each measurement can be found in the online supplementary material.

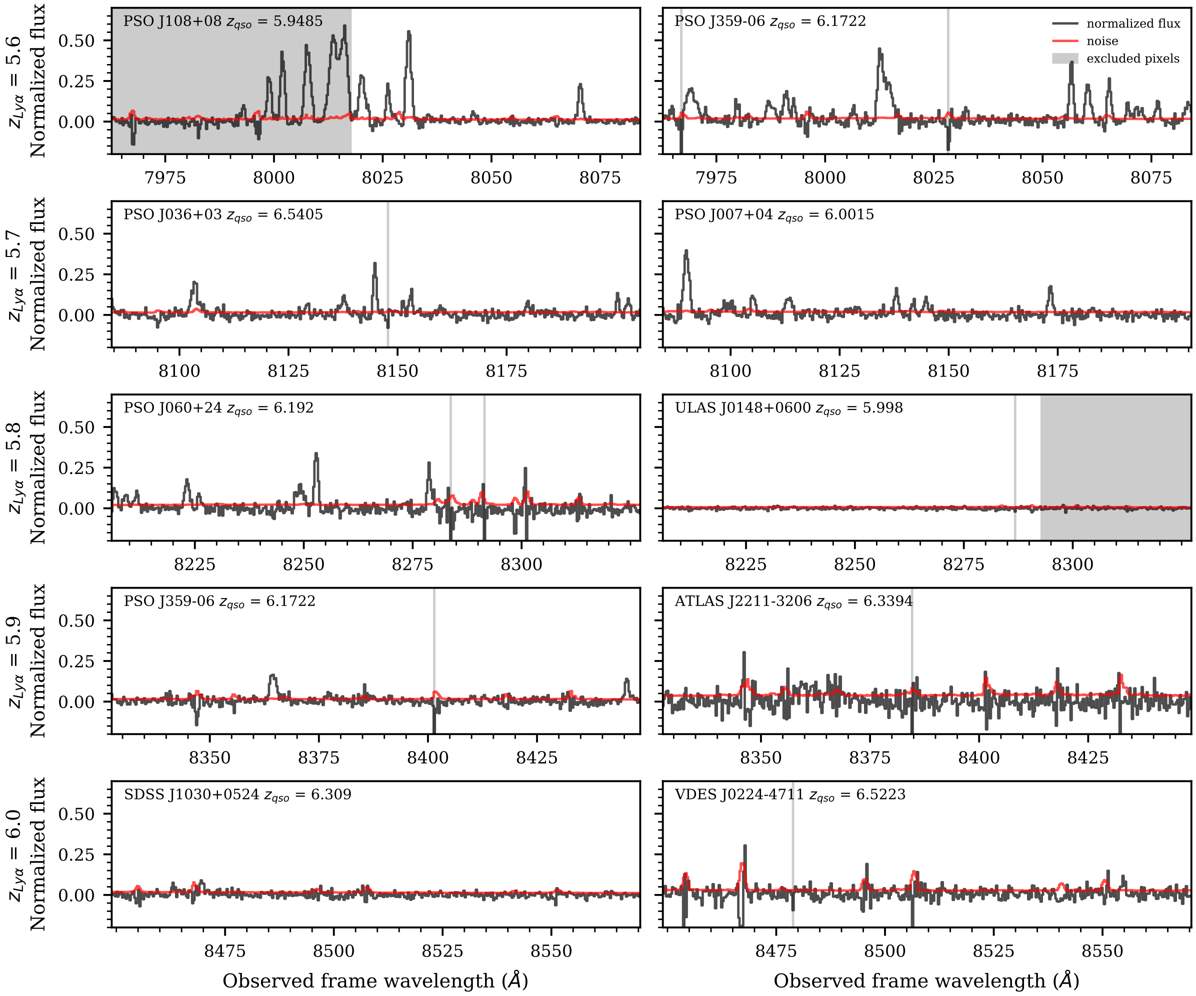

Figure 5.3: The same as Figure 5.2 except for $5.6 \leq z \leq 6.0$. The y-axis spans a smaller range than that in Figure 5.2.

of one transmission spike. Two random examples of the normalized flux from quasars in our sample at each redshift are shown in Figures 5.2 and 5.3. The normalized flux for all the sightlines used in each redshift bin can be found in the online supplementary material, which demonstrate the variance between the sightlines at a given redshift.

Figure 5.2 shows the normalized flux for two quasar sightlines for $5.1 \leq z \leq 5.5$ while Figure 5.3 has the same for $5.6 \leq z \leq 6.0$. Each row has the same z and each column shows a random quasar sightline. The value of z increases down the rows. The y-axis is fixed within Figures 5.2 and 5.3 though it varies between the two figures. The fixed y-axis illustrates the rough trend of decreasing $\langle F \rangle$ with increasing z. Both of the random sightlines shown at $z = 6$ have very limited transmission, which highlights the difficulty in making statistical measurements of the Lyα forest at high redshifts.

5.4 Results

5.4.1 Mean flux

The mean flux in this paper was calculated as the average of the normalized flux values for the non-excluded pixels as shown in Figures 5.2 and 5.3. The resulting values are reported in Table 5.2 and plotted as a function of redshift in Figure 5.4. The error on the $\langle F \rangle$ values were computed by bootstrap re-sampling the quasar sightlines considered at each z for 500,000 data set realizations and computing the variance on these values. See Section 5.4.3 for more information on how the bootstrap realizations were generated.

Figure 5.4 shows the $\langle F \rangle$ values computed in this work in red, the previous measurement of Bosman et al. (2022) in black, and the measurements of Becker et al. (2013), Bosman et al. (2018), and Eilers et al. (2018) in blue, orange, and green, respectively. Our measurement is in agreement with that from Bosman et al. (2022), as is expected

z	N_{los}	$\langle F \rangle$
5.1	24	0.1456 ± 0.0075
5.2	29	0.1314 ± 0.0072
5.3	29	0.1097 ± 0.0087
5.4	33	0.0830 ± 0.0086
5.5	34	0.0567 ± 0.0055
5.6	34	0.0474 ± 0.0053
5.7	29	0.0269 ± 0.0044
5.8	26	0.0181 ± 0.0035
5.9	15	0.0089 ± 0.0018
6.0	14	0.0090 ± 0.0023

Table 5.2: The second column lists the numbers of lines of sight at each z in our sample. The third column reports the mean flux, $\langle F \rangle$, value that was directly computed from this sample. The error on $\langle F \rangle$ comes from bootstrap re-sampling of the sightlines.

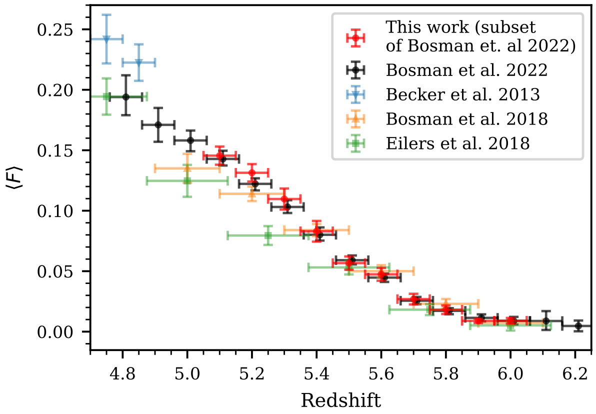

Figure 5.4: Recent measurements of the average Lyα transmission, $\langle F \rangle$, at high-z. The measured $\langle F \rangle$ from this work are shown in red. This is computed directly by taking the average of the non-excluded normalized flux values from the masks created as discussed in Sections 5.3.2 and 5.3.3. The errors come from bootstrap re-sampling the quasar sightlines. Note that the measurement shown in red comes from a subset of the quasar sightlines used in Bosman et al. (2022) which are plotted in black. Additional data points from previous works are shown in blue, orange, and green over the same z range (Becker et al., 2013; Bosman et al., 2018; Eilers et al., 2018).

since the data used here is a subset of that used in that work and our method is the same. In addition, we use the same continuum reconstruction and masking procedure as in Bosman et al. (2022). At $z = 5.1$ and $z = 5.2$ our measurement appears greater than that from Bosman et al. (2022), but the data set we considered is much smaller and the measurements are consistent within the error bars. A discussion of the agreement of $\langle F \rangle$ with previous work can be found in Bosman et al. (2022).

5.4.2 Auto-correlation Function

The auto-correlation function of the flux ($\xi_F(\Delta v)$) is defined as

$$\xi_F(\Delta v) = \langle F(v)F(v + \Delta v)\rangle \tag{5.2}$$

where $F(v)$ is the normalized flux of the Lyα forest and the average is performed over all pairs of pixels at the same velocity lag (Δv). The pixels that have been masked as discussed in Sections 5.3.2 and 5.3.3 are not used when computing the auto-correlation function for each quasar. See Appendix 5.8 for a discussion of the effect of the DLA exclusion on the measurement of the auto-correlation function. Note that different quasar sightlines will have a different number of pixel pairs contributing to the same velocity bin. Thus, when combining the different quasar sightlines, we weight each quasar's contribution by the numbers of pixel pairs in each bin. The number count of pixel pairs contributing to each auto-correlation function bin is output during the auto-correlation function computation.

We compute the auto-correlation function with the following consideration for the velocity bins. We start with the left edge of the smallest bin to be $40\,\mathrm{km\,s^{-1}}$ and use linear bins with a width of $40\,\mathrm{km\,s^{-1}}$ up to $280\,\mathrm{km\,s^{-1}}$. The choice of $40\,\mathrm{km\,s^{-1}}$ was done as it is roughly the size of a resolution element for these observations. Then we switch to logarithmic bin widths where $\log_{10}(\Delta v) = 0.058$ out to a maximal distance of $2700\,\mathrm{km\,s^{-1}}$. This results in 22 velocity bins considered where the first 6 have linear spacing. The center of our smallest bin was $60\,\mathrm{km\,s^{-1}}$ and our largest bin was $2223\,\mathrm{km\,s^{-1}}$, which corresponds to ~ 16 cMpc h^{-1} at $z = 5.5$. We chose to use linear bins on the smallest scales because the effect of λ_{mfp} is greatest on small scales and these scales already have access to the most pixel pairs which reduces noise. Larger scales are more sensitive to $\langle F \rangle$ than λ_{mfp} so having fewer bins here is not as important. In addition, there are fewer pixel pairs at large scales to begin with so using larger bins will increase the pixel pairs per bin and reduce noise.

Previously, Wolfson et al. (2023b) demonstrated the sensitivity of the auto-correlation function to λ_{mfp} for mock data at $z \geq 5.4$. Generally, they found that shorter λ_{mfp} values cause a greater boost in the auto-correlation function on the smallest scales. We compute the auto-correlation functions of the XQR-30 data set discussed in Section 5.2. The measured auto-correlation function from the extended XQR-30 data set can be seen in Figures 5.5 and 5.6. The errors on these plots come from bootstrap sampling of the quasar sightlines when computing the mean auto-correlation function and will be discussed in more detail in Section 5.4.3. The first few velocity bins of the final measurement with error from the diagonal of the covariance matrix estimated via bootstrap re-sampling are in Table 5.3. The full measurement, error bars, as well as the full bootstrap covariance matrices for each redshift are available to download online[2].

[2] https://github.com/mollywolfson/lya_autocorr/

Table 5.3: The table lists the auto-correlation function measurement for the first six bins of the auto-correlation function at all z with errors from the diagonal of the covariance matrix estimated from bootstrap re-sampling. The full measurement values of the auto-correlation function at all z can be found online.

z	**Central velocity (km s^{-1})**						
	60	100	140	180	220	260	...
5.1	0.0413 ± 0.0040	0.0331 ± 0.0035	0.0291 ± 0.0033	0.0271 ± 0.0032	0.0270 ± 0.0032	0.0262 ± 0.0031	...
5.2	0.0365 ± 0.0043	0.0290 ± 0.0037	0.0256 ± 0.0032	0.0235 ± 0.0028	0.0222 ± 0.0025	0.0210 ± 0.0024	...
5.3	0.0291 ± 0.0047	0.0227 ± 0.0042	0.0201 ± 0.0038	0.0191 ± 0.0037	0.0178 ± 0.0037	0.0168 ± 0.0036	...
5.4	0.0185 ± 0.0032	0.0143 ± 0.0026	0.0130 ± 0.0024	0.0122 ± 0.0023	0.0117 ± 0.0023	0.0113 ± 0.0022	...
5.5	0.0105 ± 0.0020	0.0076 ± 0.0016	0.0064 ± 0.0013	0.0054 ± 0.0012	0.0048 ± 0.0010	0.00436 ± 0.00081	...
5.6	0.0078 ± 0.0016	0.0057 ± 0.0012	0.0047 ± 0.0011	0.00406 ± 0.00093	0.00365 ± 0.00083	0.00361 ± 0.00084	...
5.7	0.00298 ± 0.00082	0.00206 ± 0.00062	0.00193 ± 0.00065	0.00182 ± 0.00060	0.00158 ± 0.00048	0.00141 ± 0.00044	...
5.8	0.00197 ± 0.00065	0.00120 ± 0.00040	0.00082 ± 0.00026	0.00072 ± 0.00022	0.00070 ± 0.00027	0.00083 ± 0.00032	...
5.9	0.00055 ± 0.00023	0.00030 ± 0.00013	0.000150 ± 0.000045	0.000124 ± 0.000085	0.00020 ± 0.00012	0.000189 ± 0.000095	...
6.0	0.00053 ± 0.00023	0.00027 ± 0.00014	0.000180 ± 0.000096	0.00023 ± 0.00014	0.00028 ± 0.00017	0.00019 ± 0.00012	...

Figure 5.5 has two panels that show the auto-correlation function of this data set at different z. The top panel shows $5.1 \leq z \leq 5.5$ while the bottom panel shows $5.6 \leq z \leq 6.0$. They are shown in two different panels in order to better accommodate the dynamic range of the auto-correlation function over our range of z. The overall amplitude of the auto-correlation function of the flux is set by $\langle F \rangle^2$, which decreases with increasing z.

In order to better visually demonstrate the differences in the shape of the auto-correlation function on small scales, we also plot the measured auto-correlation function normalized and shifted by $\langle F \rangle^2$ at each z in Figure 5.6. Note that the $\langle F \rangle$ value used is redshift dependent and is reported in Table 5.2. This is equivalent to the auto-correlation function of the flux density field. The color of the normalized auto-correlation function at each z matches those from Figure 5.5. This has been split into two panels for visual clarity to more easily see the behavior in each redshift bin. The top panel has $z = 5.1, 5.3, 5.5, 5.7, 5.9$ while the bottom panel has $5.2, 5.4, 5.6, 5.8, 6.0$. By looking at the smallest scales, $v < 500\,\mathrm{km\,s^{-1}}$ or $x < 4$ cMpc h^{-1} at $z = 5.5$, there is a trend of increasing small-scale values of the auto-correlation function with increasing redshift. For example, the lines for $5.8 \leq z \leq 6.0$ have the greatest auto-correlation value (in shades of purple). Note that these points have the largest error bars, likely caused by both the limited number of sightlines and the low transmission at these redshifts. Both $\langle F \rangle$ and λ_{mfp} affect the small scale boost in the auto-correlation function. Smaller $\langle F \rangle$ will lead to larger fluctuations in the flux contrast field and thus a boost on the small scales. Wolfson et al. (2023b) found that shorter λ_{mfp} values also cause a boost in the auto-correlation function on the smallest scales. These effects are not completely degenerate since the overall auto-correlation function shape differs as shown in the forecast measurements of Wolfson et al. (2023b).

We isolate the redshift evolution of the smallest velocity bin ($60\,\mathrm{km\,s^{-1}}$) of the normalized auto-correlation function in Figure 5.7. Again, the $\langle F \rangle$ value used is redshift

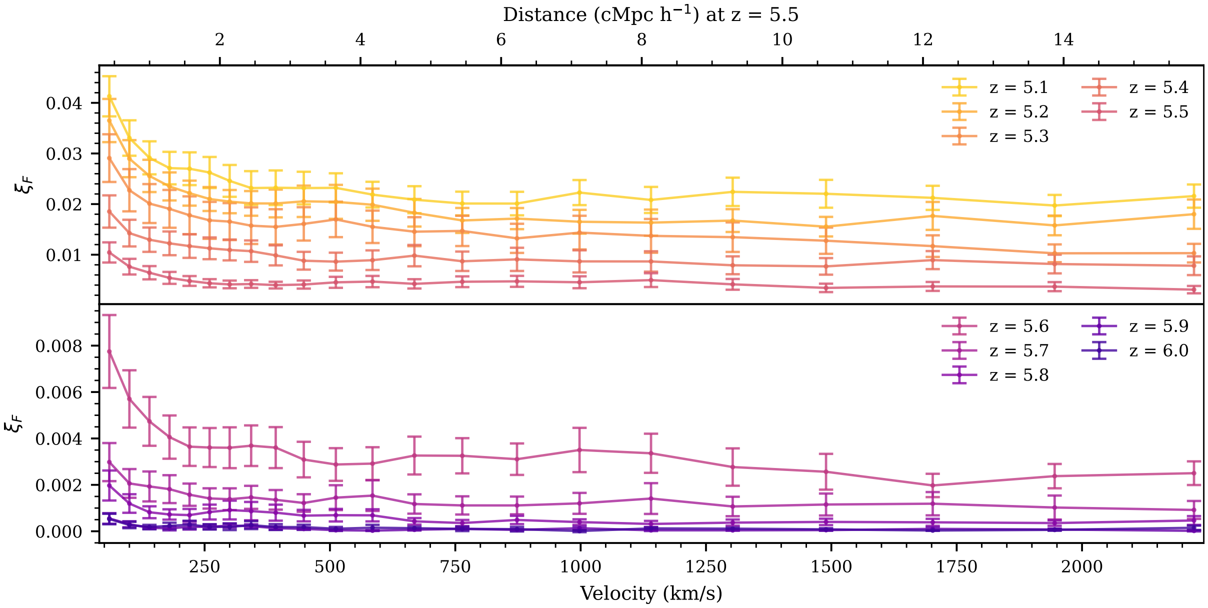

Figure 5.5: The auto-correlation function of Lyα transmission in ten redshift bins for XQR-30 data. The top panel shows the lower z bins, $5.1 \leq z \leq 5.5$ while the lower panel shows the higher z bins, $5.6 \leq z \leq 6.0$. The main trend seen in these plots is the evolution of $\langle F \rangle$ which is very small at high-z.

dependent and is reported in Table 5.2. The errors are computed by propagating the statistical uncertainty from bootstrap re-sampling both the auto-correlation function and $\langle F \rangle$. In general these values increase with redshift, which is expected from decreasing $\langle F \rangle$ as well as λ_{mfp}. However, the errors also increase with redshift and the values at the highest redshift are consistent with each other within errors.

5.4.3 Bootstrap Covariance Matrices

In order to calculate the error on $\langle F \rangle$ and the auto-correlation functions we used bootstrap re-sampling. To compute the values we performed averages over $N_{\mathrm{boot}} = 500000$ realizations of the data set. Each realization is a random selection of N_{los} quasars with replacement. In addition, each choice of quasar goes along with a choice of the 500 continuum realizations that were generated as described at the end of Section 5.3.1. The

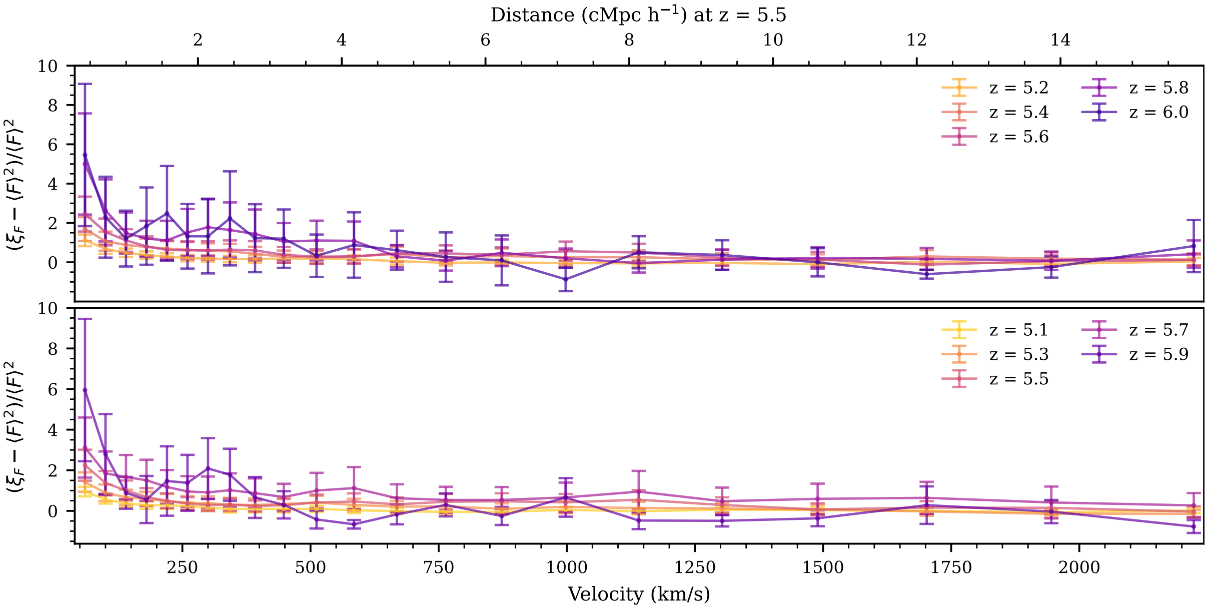

Figure 5.6: The auto-correlation function of Lyα transmission normalized and shifted by the mean transmission, $\langle F \rangle$, in ten redshift bins for XQR-30 data. This is equivalent to the auto-correlation function of the flux density field. The errors are computed by propagating the statistical uncertainty from bootstrap re-sampling both the auto-correlation function and $\langle F \rangle$. This is split into two panels for visual clarity, so as to not overcrowd the panels. The top panel has $z = 5.1, 5.3, 5.5, 5.7, 5.9$ while the bottom panel has $5.2, 5.4, 5.6, 5.8, 6.0$. This figure makes the trend of higher redshift bins having larger boosts of the auto-correlation function on small scales when dividing out the flux evolution more visible.

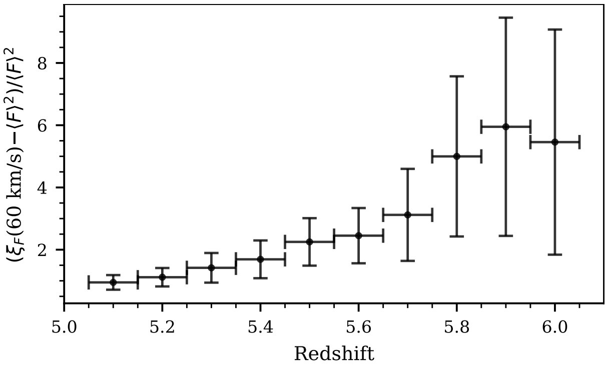

Figure 5.7: The value of the first bin of the auto-correlation function of Lyα transmission normalized and shifted by the mean transmission, $\langle F \rangle$, as a function of redshift. The errors are computed by propagating the statistical uncertainty from bootstrap re-sampling both the auto-correlation function and $\langle F \rangle$. These values are also shown in Figure 5.6. There is a general trend of increasing value with redshift, though the errors also increase. The highest redshift values are consistent with no evolution.

computed mean flux for the ith sample is thus $\langle F\rangle_i$ and the error on $\langle F\rangle$, σ_F is:

$$\sigma_F = \left(\frac{1}{N_{\text{boot}} - 1} \sum_{i=1}^{N_{\text{boot}}} \left(\langle F\rangle_i - \langle F\rangle \right)^2 \right)^{1/2}. \tag{5.3}$$

These errors are reported in Table 5.2 and shown in Figure 5.4.

For the auto-correlation function, ξ, we compute the entire bootstrap covariance matrix, not only the diagonal error. Again we chose N_{boot} realizations of the observed data set by randomly selecting N_{los} quasars with replacement each with their own random selection of the continuum realization. For any given bootstrap realization we computed the average of the auto-correlation function over the chosen sightlines to construct a realization of the average auto-correlation function, ξ_i. The covariance matrix was then computed by averaging over the ensemble of bootstrap realizations in the following way:

$$\mathbf{\Sigma}_{\text{boot}} = \frac{1}{N_{\text{boot}} - 1} \sum_{i=1}^{N_{\text{boot}}} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_{\text{data}})(\boldsymbol{\xi}_i - \boldsymbol{\xi}_{\text{data}})^{\text{T}}. \tag{5.4}$$

For visualization purposes, we use the diagonal of the bootstrap covariance matrices to estimate the error bars on the auto-correlation function shown in Figure 5.5. Specifically we define $\sigma_{\text{boot}} = \sqrt{(}\text{diag}(\mathbf{\Sigma}_{\text{boot}}))$. The diagonal of the covariance matrix is not a full description of the error since the bins of the auto-correlation function are highly correlated and should thus fluctuate in a correlated way, thus making the full covariance matrix necessary in any computations. The error bars in Figure 5.6, σ_Δ, come from combining the bootstrap estimate of the errors for ξ_F with bootstrap estimate of the errors on $\langle F\rangle$ via:

$$\sigma_\Delta = \frac{\xi_F - \langle F\rangle^2}{\langle F\rangle^2} \sqrt{\left(\frac{\sigma_{\text{boot}}}{\xi_F}\right)^2 + 2\left(\frac{\sigma_F}{\langle F\rangle}\right)^2} \tag{5.5}$$

Additionally we define the correlation matrix, C, which expresses the covariances between jth and kth bins in units of the the diagonal elements of the covariance matrix. This is done for the jth, kth element by

$$C_{jk} = \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}}. \tag{5.6}$$

The bootstrap correlation matrices for the measured auto-correlation functions at each z are shown in Figure 5.8. Based on the simulated correlation matrices from Wolfson et al. (2023b), we expect there to be significant off diagonal values of these bootstrap correlation matrices. This is because, generally, each pixel in the Lyα forest contributes to every bin of the auto-correlation function so the different velocity bins in the auto-correlation function are highly covariant. Large off-diagonal values are seen in the bootstrap correlation matrices in Figure 5.8 for $z < 5.8$. At the highest three redshifts, especially $z = 5.9$ and $z = 6.0$, the number of quasar lines of sight are quite small and the transmission is quite low, leading to large noise fluctuations and non-converged off-diagonal values. In particular, there are negative values off the diagonal for $z = 5.9$ and $z = 6.0$ which we do not see in our simulated covariance matrices. We expect noisy fluctuations in the off-diagonal covariance matrix values to go away with the addition of more quasar sightlines, though low transmission at the highest redshifts will still make this computation difficult.

5.5 Modeling the measurement

In order to interpret the physical implications of the measured auto-correlation function, we construct forward models with the properties of the observed quasars. Functions to convert any set of simulation skewers into auto-correlation function measurements are available online at `https://github.com/mollywolfson/lya_autocorr/`. In addition, there is a `Jupyter Notebook` that goes through an example of forward-modeling simulation skewers and then computing the auto-correlation function. The simulation method used here was introduced in Wolfson et al. (2023b) for a simplified mock data set. We have updated this method to include continuum uncertainty, noise vectors from observational data, and a $\Gamma_{\rm HI}$ box that matches the density field of the main simulation suite.

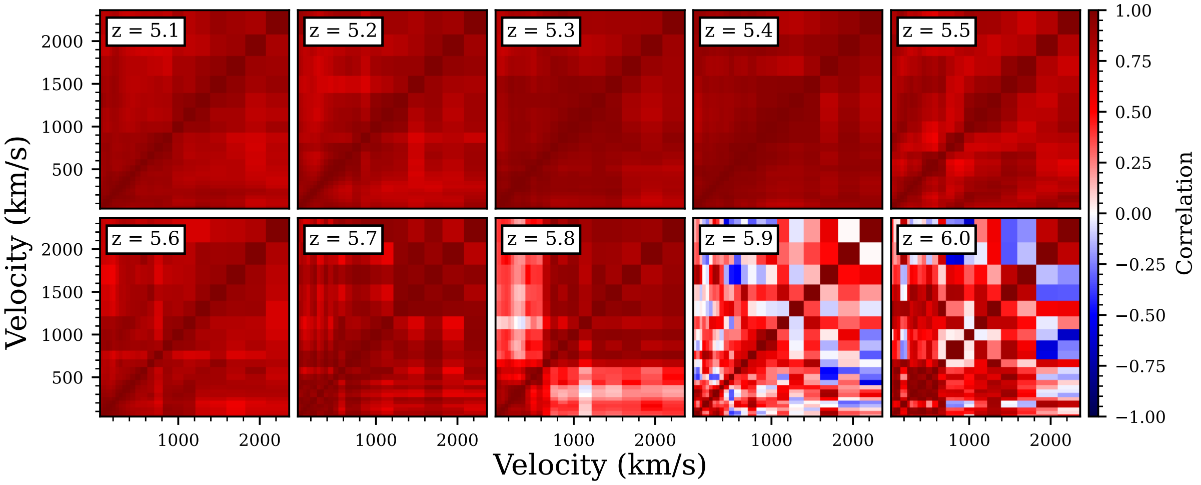

Figure 5.8: The correlation matrices from bootstrap re-sampling the auto-correlation function in the ten redshift bins considered in this work. For $z < 5.8$ we see very strong positive off-diagonal values of the correlation matrices. This behavior is expected since each pixel in the Lyα forest contribute to every bin of the auto-correlation function, making these bins highly correlated. The fluctuations in the correlation matrix values are caused by noise due to the limited sightlines available to bootstrap. At $z \geq 5.8$ the number of sightlines is small and the transmission is low, causing large noise fluctuations. For $z = 5.9$ and $z = 6.0$, the sightlines are so few and so non-transmissive that noise fluctuations lead to negative values in the correlation matrices. There is no physical explanation for these negative values. The numbers of sightlines used at each z are listed in Table 5.2.

We will briefly describe this updated method here, for more information see Wolfson et al. (2023b).

Note that this paper is using a simple model that only varies $\lambda_{\rm mfp}$ and $\langle F \rangle$. More sophisticated modeling that includes variation in the IGM thermal state, patchy reionization, and more robust UVB modeling is left for future work. For an initial investigation into the effect of the IGM thermal state and inhomogeneous reionization on the Lyα forest flux auto-correlation function see Wolfson et al. (2023a). They found that these mainly affect scales $v < 100\,{\rm km\,s^{-1}}$, which corresponds to only the smallest bin considered here. Thus, while additional simulation work is necessary to include all relevant parameters, the models presented here are sufficient for an initial comparison.

5.5.1 Simulation box

To begin, we use a `Nyx` simulation box (Almgren et al., 2013). `Nyx` is a hydrodynamical simulation code designed to simulate the Lyα forest with updated physical rates from Lukić et al. (2015). The `Nyx` box has a size of $L_{\rm box} = 100$ cMpc h^{-1} with 4096^3 dark matter particles and 4096^3 baryon grid cells. This box is reionized by a Haardt & Madau (2012) uniform UVB that is switched on at $z \sim 15$, which means these simulation boxes do not include the effects of a patchy, inhomogeneous reionization.

We have three snapshots of this simulation at $z = 5.0$, $z = 5.5$, and $z = 6$ and we want to model all ten redshifts $5.1 \leq z \leq 6.0$ with $\Delta z = 0.1$. In order to consider the redshifts for which we do not have a simulation output, we select the nearest snapshot and use the desired redshift when calculating the proper size of the box and the mean density. This means we use the density fluctuations, temperature, and velocities directly from the nearest `Nyx` simulation output. Previously, in Wolfson et al. (2023b) we tested this choice of simulation interpolation by using the $z = 6.0$ simulation snapshot to generate skewers at $z = 5.7$ and found no change in the results.

In addition, we have a grid of boxes of $\Gamma_{\rm HI}/\langle\Gamma_{\rm HI}\rangle$ values generated with the semi-numerical method of Davies & Furlanetto (2016b) corresponding to a fluctuating UVB for different $\lambda_{\rm mfp}$ values, all at $z = 5.5$. These boxes have a size of $L_{\rm box} = 100\ h^{-1}$ cMpc, 64^3 pixels, and are generated from the density field of the `Nyx` simulation box. The method of Davies & Furlanetto (2016b) uses Mesinger & Furlanetto (2007) and Bouwens et al. (2015) to create halos and assign UV luminosities from the density field. They then get the ionizing luminosity of each galaxy by assuming it to be proportional to its UV luminosity where the constant of proportionality is left as a free parameter. Finally the ionizing background radiation intensity, J_ν, is computed by a radiative transfer algorithm and $\Gamma_{\rm HI}$ is finally calculated by integrating over J_ν. For more information on this method of generating $\Gamma_{\rm HI}$ boxes see Davies & Furlanetto (2016b), Davies et al. (2018b), or Wolfson et al. (2023b). Note that this modeling assumes a number of relations, such as local $\lambda \propto \Gamma_{\rm HI}^{2/3}\Delta^{-1}$. Additional work looking into the effect on the UVB from varying these assumptions is necessary to get robust constraints on $\lambda_{\rm mfp}$ from these models. We leave this for future work and use this simple one parameter model for an initial, qualitative comparison with the data.

To combine the `Nyx` box with the $\Gamma_{\rm HI}$ values generated via the Davies & Furlanetto (2016b) method, we linearly interpolated $\log(\Gamma_{\rm HI}/\langle\Gamma_{\rm HI}\rangle)$ onto the higher resolution grid of the `Nyx` simulation box. We then re-scale the optical depths from the `Nyx` box with a constant UVB, $\tau_{\rm const.}$, by these fluctuating $\Gamma_{\rm HI}$ values to get the optical depths for a fluctuating UVB, $\tau_{\rm mfp} = \tau_{\rm const.}/(\Gamma_{\rm HI}/\langle\Gamma_{\rm HI}\rangle)$. This implies that we need to know $\langle\Gamma_{\rm HI}\rangle$ to compute our final optical depths, which is not known a priori. We therefore determine this value by matching an overall mean flux $\langle F\rangle$, where we vary $\langle F\rangle$ over a range of models based off the measurement of Bosman et al. (2022). We look at the relationship between $\langle F\rangle$, $\lambda_{\rm mfp}$, and $\langle\Gamma_{\rm HI}\rangle$ in Appendix 5.9. We generate 1000 skewers from this simulation method for each $\lambda_{\rm mfp}$ and $\langle F \rangle$ at each z for $5.1 \leq z \leq 6.0$. These skewers come from the same location in the simulation box for all parameter values and z.

5.5.2 Forward modeling

Our simulations provide skewers of the optical depth of the Lyα forest for given $\lambda_{\rm mfp}$ and $\langle F \rangle$ values. In order to compare these (or any) simulated skewers with the results of our observational measurement, we forward model the telescope resolution, the noise properties of our observed sightlines, and the continuum uncertainty from the PCA continuum fit. This section will describe how each property is modeled for our simulation skewers, see the `lya-autocorr` git repository to follow along with an example simulation skewers being forward modeled.

To model the resolution of X-shooter for visible light with a 0.9" slit, we convolved the flux by a Gaussian line-spread function with FWHM $\approx 34\,\mathrm{km\,s^{-1}}$. This corresponds the nominal resolving power ($R \sim 8800$) of the X-Shooter setup used for the XQR-30 data. However, as noted in Section 5.2 the actual data has a higher median resolving power in the visible of $R = 11400$ (D'Odorico et al., 2023). Future work will use the measured resolving power for each quasar in the modeling but using the nominal value for all is sufficient for this initial comparison. After using this Gaussian filter we interpolated the line-spread-function convolved flux onto the exact velocity grid from the observation. This step also reduced the simulation skewers from the box size to the same length as our observations, as 100 cMpc h^{-1} corresponds to $\Delta z \sim 0.3$ at the relevant redshifts and our observations have $\Delta z = 0.1$.

We add noise to the interpolated, line-spread-function convolved flux, $\boldsymbol{F}_{\rm res}$, according to the noise vectors for each quasar sightline, $\boldsymbol{\sigma}_{\rm qso}$, with random normal distribution realization, $\boldsymbol{N}_{\rm qso} \sim N(0,1)$, via

$$\boldsymbol{F}_{\rm noise} = \boldsymbol{F}_{\rm res} + (\boldsymbol{N}_{\rm qso} \times \boldsymbol{\sigma}_{\rm qso})\,. \tag{5.7}$$

$\boldsymbol{F}_{\rm noise}$ is thus the flux modeled with both the resolution of the telescope and the noise properties of our observed sightlines. This modeling choice is valid because of the low flux in the Lyα forest at these redshifts such that we are background limited in the observations.

To model continuum error, we used the mean, $\boldsymbol{\mu}_{\rm cont}$, and covariance, $\boldsymbol{\Sigma}_{\rm cont}$, of the PCA reconstruction just as we do for the data as described in Section 5.3.1. We randomly draw realizations of the continuum error, $\boldsymbol{E}_{\rm cont} \sim N(\boldsymbol{\mu}_{\rm cont}, \boldsymbol{\Sigma}_{\rm cont})$, where N is the normal distribution. In our simulations we do not fit and normalize by the quasar continuum so we model continuum error by:

$$\boldsymbol{F}_{\rm cont} = \boldsymbol{F}_{\rm noise}/\boldsymbol{E}_{\rm cont} \tag{5.8}$$

where $\boldsymbol{F}_{\rm cont}$ is the final fully forward-modeled Lyα forest spectra. We investigate the effect of the continuum modeling on the resulting models of the auto-correlation function in Appendix 5.7.

Ultimately, we generate $N_{\rm skewer}$ forward-modeled copies of each of the $N_{\rm los}$ quasars in the sample, where $N_{\rm skewer} = 1000$ from the simulation and $N_{\rm los}$ is the number of quasar sightlines at each redshift as listed in Table 5.2. For example at $z = 5.1$ we have $N_{\rm skewers} \times N_{\rm los} = 1000 \times 24 = 24000$ total forward-modeled Lyα forest spectra.

Figure 5.9 shows the normalized flux of the $z = 5.6$ Lyα forest from PSO J029+29 with four examples of the normalized flux from our simulations that were forward modeled with this quasar's properties. The thick line in the middle is the flux from the quasar while the other four thinner lines are from the simulation. The visual similarities between the observed data and the forward modeled data highlights the ability of our forward

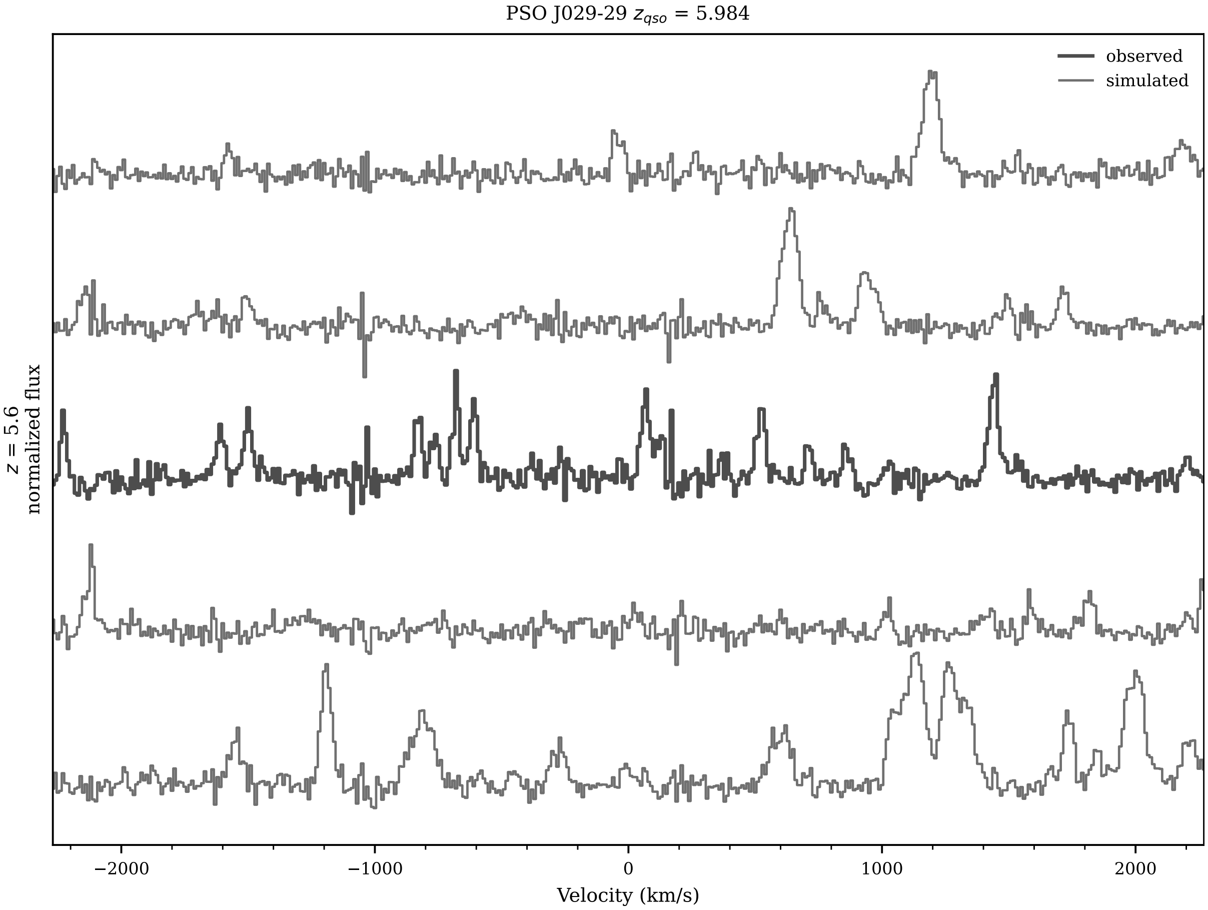

Figure 5.9: This figure compares the observed Lyα forest flux at $z = 5.6$ from PSO J029-29 with forward modeled simulation skewers modeled to have the same noise properties as this quasar. The thick line in the middle is the observed flux while the other four thinner lines are from the forward modeled simulations. The visual similarities between the observed and simulated Lyα forest flux shown here demonstrates the success of our forward-modeling procedure.

modeling methods to mimic realistic data. The remaining figures all show data and simulations at $z = 5.6$ because this redshift has the maximal observed sightlines with $N_{\text{los}} = 34$ and there is a nearby measurement of λ_{mfp} at $z = 5.6$ by Zhu et al. (2023). Note that N_{los} does not affect the convergence of our simulated models but it determines the convergence of the bootstrap covariance matrix estimate which we will compare to later in the section.

5.5.3 Modeled auto-correlation function

We then computed the auto-correlation function of these forward modeled skewers in the same way as the actual data, with equation (5.2), for each copy of the skewer. We used the same mask from the observed quasar when computing the auto-correlation function. This includes the DLA mask as described in Section 5.3.3. In the observations, the DLA mask corresponds to regions in the spectra where the transmission is low. However, for the simulations the DLA mask corresponds to random parts of the spectra. We choose to include this part of the mask for the simulation data in order to keep the number of pixel pairs used per quasar sightline the same between simulations and observations. A discussion on the effect of the DLA mask on the measured auto-correlation function can be found in Appendix 5.8.

To create a mock data set, we randomly selected $N_{\rm los}$ quasars from the 1000 forward modeled skewers without replacement. We then assigned each of the randomly selected skewers one of each of the $N_{\rm los}$ quasars, so each mock data set had exactly one skewer forward modeled with the properties of each quasar. The value of the auto-correlation function from the mock data set, ξ_i, is then the weighted average of the auto-correlation function from these $N_{\rm los}$ forward modeled skewers, where the weights are the number of pixels pairs in each bin of the auto-correlation function. We defined the model value of the auto-correlation function, $\boldsymbol{\xi}_{\rm model} = \boldsymbol{\xi}_{\rm model}(\lambda_{\rm mfp}, \langle F \rangle)$, to be the weighted average of the auto-correlation functions from all $N_{\rm los} \times N_{\rm skewers}$ skewers generated. The simulated covariance matrices, $\boldsymbol{\Sigma}_{\rm sim}$, are computed for each $\lambda_{\rm mfp}$ and $\langle F \rangle$ values from $N_{\rm mocks}$ mock data sets in the following way:

$$\boldsymbol{\Sigma}_{\rm sim}(\boldsymbol{\xi}_{\rm model}) = \frac{1}{N_{\rm mocks}} \sum_{i=1}^{N_{\rm mocks}} (\boldsymbol{\xi}_i - \boldsymbol{\xi}_{\rm model})(\boldsymbol{\xi}_i - \boldsymbol{\xi}_{\rm model})^{\rm T}. \tag{5.9}$$

Figure 5.10 shows nine measurements of the auto-correlation function from nine different mock data sets generated from the simulations at $z = 5.6$ (colored triangles). These

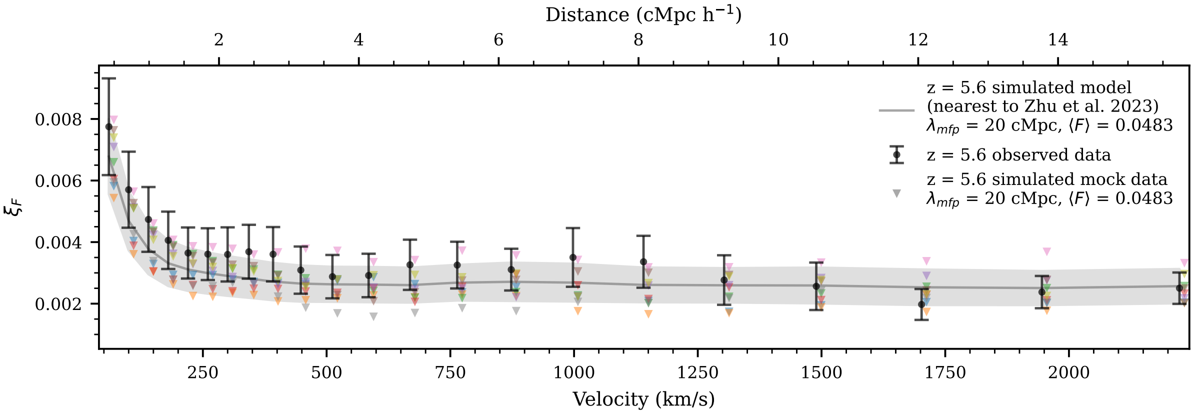

Figure 5.10: The black points show the observed auto-correlation function from the extended XQR-30 data discussed in this work at $z = 5.6$. The colored triangles show the auto-correlation value for 9 different simulated mock data sets. The mock data sets shown here were all modeled with $\lambda_{\rm mfp} = 20$ cMpc and $\langle F \rangle = 0.0483$, the closest $\lambda_{\rm mfp}$ value to the Zhu et al. (2023) measurement and the closest $\langle F \rangle$ value to our measurement listed in Table 5.2. The model value of the auto-correlation is shown as the grey line with the shaded region representing the diagonal elements from the corresponding simulated covariance matrix.

mock measurements were generated from the $\lambda_{\rm mfp} = 20$ cMpc and $\langle F \rangle = 0.0483$ simulation, the closest $\lambda_{\rm mfp}$ value to the Zhu et al. (2023) measurement and the closest $\langle F \rangle$ value to our measurement listed in Table 5.2. This model value of the auto-correlation function is shown as the grey line where the grey shaded region shows the error from the diagonal of the simulated covariance matrix. The black points show the measured auto-correlation function at $z = 5.6$ with error bars from the bootstrap covariance matrix. This plot demonstrates that our forward modeling procedure leads to mock correlation function measurements that are visually similar to our actual measurement. This plot also shows that our measured auto-correlation function and the model with the value from Zhu et al. (2023) agree within $1\sigma_{\rm boot}$ for nearly all the points, though again these errors come from the diagonal of the covariance matrix only and therefore do not include information on the strong off-diagonal covariance between auto-correlation function bins. We discuss the comparison of our measured auto-correlation function and the measurements of Zhu et al. (2023) and Gaikwad et al. (2023) in Section 5.5.5.

5.5.4 Model based covariance matrices

Figure 5.11 shows correlation matrices from the forward modeled data for six different parameter values at $z = 5.6$. The parameter values shown are $\lambda_{\rm mfp} = 5, 15, 150$ cMpc going down the rows and then $\langle F \rangle = 0.0303, 0.0591$ across the columns, both of which span the full range of parameter values available to us. Going from the left to the right column, we see that increasing the $\langle F \rangle$ weakly increases the off-diagonal values of the correlation matrices, however the effect going down the rows is much stronger. Going down the rows shows that an increase in $\lambda_{\rm mfp}$ decreases the off-diagonal values for the correlation matrix. This means that shorter $\lambda_{\rm mfp}$ models have more highly covariant bins in the auto-correlation function.

To compare a bootstrap covariance matrix from the data with the forward modeled covariance matrices, Figure 5.12 shows the bootstrap correlation matrix at $z = 5.6$ with the same color bar as Figure 5.8. Additionally, Figure 5.12 shows the simulated correlation matrix for the $\lambda_{\rm mfp} = 20$ cMpc and $\langle F \rangle = 0.0483$ model to directly compare to the bootstrapped matrix. Again, this is the model with the closest $\lambda_{\rm mfp}$ value to the Zhu et al. (2023) measurement and the closest $\langle F \rangle$ value to our measurement. The bootstrap covariance matrix is still quite noisy due to the limited data available so it is difficult to determine the best matching simulated covariance matrix. The bootstrap correlation matrix has regions of high off diagonal values, such as $1200\,{\rm km\,s^{-1}} < v < 2000\,{\rm km\,s^{-1}}$ as well as individual pixels with relatively small off-diagonal values, such as the combination of $v = 60\,{\rm km\,s^{-1}}$ and $v = 1702\,{\rm km\,s^{-1}}$. This potentially suggests additional structure in the bootstrap covariance matrix compared to the simulated covariance data, but these fluctuations appear consistent with the noise.

As can be seen in Figure 5.11, the correlation matrices, and therefore the covariance matrices, strongly depend on the model value of $\lambda_{\rm mfp}$. For this reason, when attempting

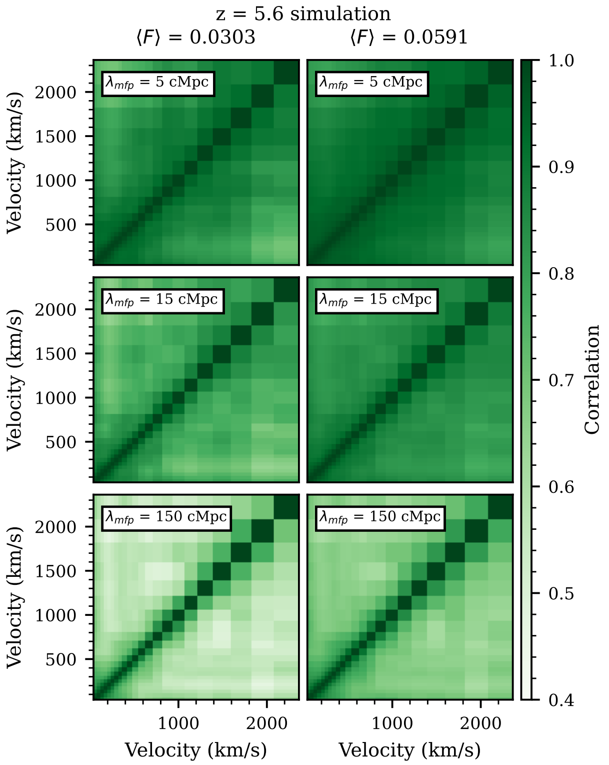

Figure 5.11: Correlation matrices for six different simulation model values at $z = 5.6$. These six covariance matrices come from the combination of $\lambda_{\rm mfp} = 5, 15, 150$ cMpc and $\langle F \rangle = 0.0303, 0.0591$ as labeled in the title of each subplot. These include the maximal and minimal $\lambda_{\rm mfp}$ and $\langle F \rangle$ values simulated at $z = 5.6$. This shows the model-dependence of the correlation (and thus covariance) matrices. Larger values of $\lambda_{\rm mfp}$ result in weaker off-diagonal correlation matrix values, as is seen going down the rows. Smaller $\langle F \rangle$ values also appear to cause weaker off-diagonal correlation matrix values (as seen when comparing the left and right columns) but this effect is weaker than the effect of $\lambda_{\rm mfp}$.

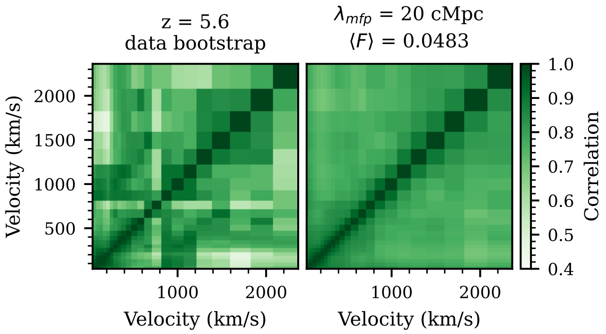

Figure 5.12: The correlation matrix computed via bootstrap re-sampling the data at $z = 5.6$ (left) and the simulated correlation matrix from the model with $\lambda_{\rm mfp} = 20$ cMpc and $\langle F \rangle = 0.0483$ (right). This model was chosen as the model with the closest $\lambda_{\rm mfp}$ value to the Zhu et al. (2023) measurement and the closest $\langle F \rangle$ value to our measurement. This bootstrap correlation matrix is also shown in Figure 5.8 with a different color bar and has been reproduced here with the color bar used in Figure 5.11, to more easily compare the values of the correlation matrix from data to the simulated examples. The bootstrap covariance matrix is noisy due to the limited data available, though this redshift was selected as the bin with the maximal value of $N_{\rm los} = 34$.

to fit this data to a model, we would be fitting both the measured auto-correlation function as well as the covariance structure between the bins. While the amplitude of the correlation function might favor one combination of model parameters, it is conceivable that the level of fluctuations between two correlated correlation function bins, which is quantified by the covariance matrix, could favor a different combination of parameters. For this reason, fitting these models to our measurements is quite challenging and we leave this discussion for future work.

5.5.5 Comparison to previous work

We model the auto-correlation function at any value of $\lambda_{\rm mfp}$ and $\langle F \rangle$ via nearest grid-point emulation from our initial grid of values. Therefore, we can compare our auto-correlation function measurement to the models with the $\lambda_{\rm mfp}$ values measured in Gaikwad et al. (2023) and Zhu et al. (2023) which updated the measurements of Becker et al. (2021). Since we need to specify both $\lambda_{\rm mfp}$ and $\langle F \rangle$ to get our models, we use the measured $\langle F \rangle$ from this work to get the models representing the $\lambda_{\rm mfp}$ values from the corresponding alternative measurements. Figure 5.13 has ten panels, each of which has one of our measured $\frac{\xi_F - \langle F \rangle^2}{\langle F \rangle^2}$ values (shown as the black points) at a given z. We have chosen to show $\frac{\xi_F - \langle F \rangle^2}{\langle F \rangle^2}$ instead of the regular auto-correlation function because we have to use the nearest grid point on a coarse $\langle F \rangle$ grid which could be quite far from the measured $\langle F \rangle$ value. This would have a large effect on the auto-correlation function value and a smaller effect on $\frac{\xi_F - \langle F \rangle^2}{\langle F \rangle^2}$.

Gaikwad et al. (2023) measured $\lambda_{\rm mfp}$ at each of these redshifts and so each panel has our model with their $\lambda_{\rm mfp}$ values (green lines). Zhu et al. (2023) has measured $\lambda_{\rm mfp}$ for $z = 5.08$, 5.31, 5.65, and 5.93. We show the models for the measured $\lambda_{\rm mfp}$ values from Zhu et al. (2023) in the $z = 5.1$, 5.3, (5.6 and 5.7), and 6.0 panels respectively (red lines). Finally, we also show the model for $\lambda_{\rm mfp} = 150$ cMpc, our most uniform UVB (blue line).

Making a quantitative comparison of these models with the measured auto-correlation function is difficult due to the expected large-off diagonal values of the covariance matrix as well as the noise in the bootstrap covariance matrices as shown in Figure 5.8. For this reason we leave detailed quantitative comparisons and fitting for future work. It is interesting to note that our measurements fall above the models from Zhu et al. (2023), Gaikwad et al. (2023), and $\lambda_{\rm mfp} = 150$ cMpc for $z < 5.8$. Also note that models from Zhu et al. (2023) and Gaikwad et al. (2023) show a small boost over the most uniform UVB model for $z < 5.8$.

5.6 Conclusions

In this work we have measured the auto-correlation function of the Lyα forest flux from the extended XQR-30 data set in 10 redshift bins, $5.1 \leq z \leq 6.0$. This is the first measurement of the auto-correlation function of the Lyα forest at these redshifts.

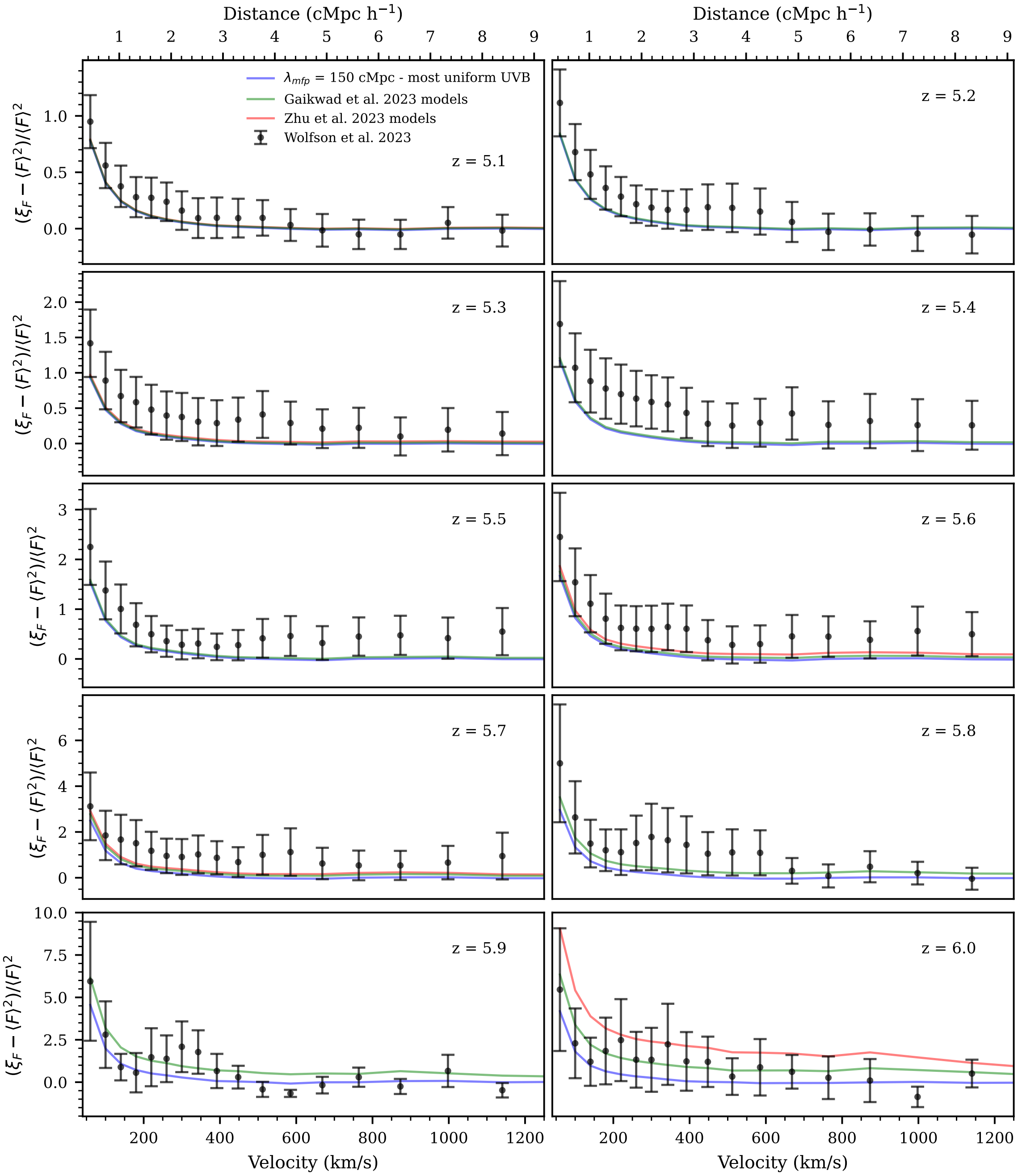

Figure 5.13: The auto-correlation function of Lyα transmission normalized and shifted by the mean transmission, $\langle F \rangle$, in ten redshift bins measured in this work. Gaikwad et al. (2023) measured $\lambda_{\rm mfp}$ at each of these redshifts and so each panel has our model with their $\lambda_{\rm mfp}$ values (green lines). Zhu et al. (2023) has measured $\lambda_{\rm mfp}$ for $z = 5.08$, 5.31, 5.65, and 5.93. We show the model models for the measured $\lambda_{\rm mfp}$ values from Zhu et al. (2023) in the $z = 5.1$, 5.3, (5.6 and 5.7), and 6.0 panels respectively (red lines). The model for a uniform UVB value (blue line) is also shown as a comparison.

Our final assembled data set includes 36 $z > 5.7$ quasars with SNR > 20 per spectral pixel. This data set was analyzed while fully accounting for the error from continuum reconstruction, instrumentation, and contamination from DLAs. We measured the average transmission, $\langle F \rangle$, from this data and found good agreement with previous work. We found that the boost in the auto-correlation function on the smallest scales increases when increasing z, which may suggest a decrease in $\lambda_{\rm mfp}$. We additionally measured covariance matrices of the auto-correlation function by bootstrap re-sampling the available data. The convergence of these matrices was hindered by noise from the limited number of sightlines and low transmission, especially for the highest redshift bins, $z \geq 5.8$. The auto-correlation function measurements as well as the bootstrap covariance matrices are available to download online at `https://github.com/mollywolfson/lya_autocorr/`. Note that this is the best available sample of quasars at these redshifts in terms of size, resolution, and SNR. Increasing the number of observations, especially at $z \gtrsim 6.5$, with the same quality would greatly improve these measurements.

In addition, we introduced Lyα forest simulations with a fluctuating UVB model described by $\lambda_{\rm mfp}$. This comparison indicates preliminary agreement between these models and our measurements. We found that the covariance matrices produced from the simulations had a strong dependence on $\lambda_{\rm mfp}$. In order to fit these models to our data, we would need to use an estimate of the covariance matrix for the bins of the auto-correlation function. In this work we have presented two options for this covariance matrix: the bootstrap estimate, $\mathbf{\Sigma}_{\rm boot}$, and the simulation covariance matrices, $\mathbf{\Sigma}_{\rm sim}$. Ideally we would like to use $\mathbf{\Sigma}_{\rm boot}$ when fitting, however as seen in Figure 5.8, these covariance matrices are quite noisy and non-converged. Therefore, we could hope to use $\mathbf{\Sigma}_{\rm sim}$, where the off-diagonal structure depends strongly on the value of $\lambda_{\rm mfp}$. This dependence of $\mathbf{\Sigma}_{\rm sim}$ on $\lambda_{\rm mfp}$ means that fitting the models to the data would require fitting both the mean line as well as this covariance structure, which is subtle. Thus, additional work is necessary to get robust measurements of $\lambda_{\rm mfp}$, which we leave to the future. We did show a preliminary comparison of our measured auto-correlation function to models with the $\lambda_{\rm mfp}$ values measured by Gaikwad et al. (2023) and Zhu et al. (2023), leaving a quantitative comparison of these results to future work.

With this work we have included a link to a Git repository with the code necessary to measure the auto-correlation function from any set of simulation skewers. This will allow other simulation groups to compare the auto-correlation function from their simulations to our measured auto-correlation function and thus foster more work on this statistic.

Future work to get a robust measurement of $\lambda_{\rm mfp}$ from the Lyα forest auto-correlation function include further considerations in the modeling methods. The Davies & Furlanetto (2016b) method to generate $\Gamma_{\rm HI}$ for various $\lambda_{\rm mfp}$ assumes a fixed source model. Other source model choices could impact the fluctuations in $\Gamma_{\rm HI}$ seen at a fixed $\lambda_{\rm mfp}$ value, and thus bias measurements from observation data when compared with these models. Additionally, rare bright sources could cause boosts in the auto-correlation function for individual sightlines that aren't modeled in our simulations. We leave a detailed investigation into these effects on the auto-correlation function models and covariance matrices to future work.

Note that in order to generate UVB fluctuations due to different $\lambda_{\rm mfp}$ values that matched the density field of our `Nyx` simulation, we also generated UVB fluctuations in a 100 cMpc h^{-1} box. Wolfson et al. (2023b) found that using a 40 cMpc h^{-1} box to generate UVB fluctuations significantly reduced the auto-correlation function on all scales when compared to a 512 cMpc box. Future work would be needed to understand the effect of the box size on any measured $\lambda_{\rm mfp}$ from the auto-correlation function with a 100 cMpc h^{-1} UVB box.

Additionally, this work ignored the effect of inhomogeneous reionization beyond a fluctuating UVB. It is expected that a patchy, inhomogeneous reionization process would have other physical effects, such as additional fluctuations in the thermal state of the IGM. We leave an exploration of the effect of the temperature of the IGM on the Lyα forest flux auto-correlation function, including the effect of temperature fluctuations, to a future work.

Overall, this first measurement of the $z > 5$ Lyα forest flux auto-correlation functions opens up an exciting new way to measure $\lambda_{\rm mfp}$ at the tail-end of reionization.

5.7 Appendix A: Continuum uncertainty modeling effect

Figure 5.14 quantifies the difference in the auto-correlation models calculated from forward-modeled skewers with or without continuum uncertainty multiplied in, as described in Section 5.5.2. The first and third panels show the auto-correlation function from the simulations with (solid line) and without (dashed line) modeling continuum uncertainty at $z = 5.1$ and 6. The different colors represent different parameter values of $\lambda_{\rm mfp}$ and $\langle F \rangle$ used. The second and fourth panels show the relative difference in percent, defined as:

$$\frac{\xi_{\rm cont} - \xi_{\rm no\ cont}}{\xi_{\rm no\ cont}}. \tag{5.10}$$

At $z = 5.1$ there is a $< 1\%$ of a difference between the auto-correlation models with and without the continuum error. At $z = 6.0$ there is a larger difference between the models where the difference is $< 8\%$ for all the parameter values. However, the effect is most noticeable when $\langle F \rangle$, and hence the auto-correlation function which goes as $\langle F \rangle^2$, is quite small. For the other $\langle F \rangle$ value at this redshift the error is $< 2\%$. These values are typically positive because of the bias in the continuum reconstruction as seen in Figure 2 of Bosman et al. (2022).

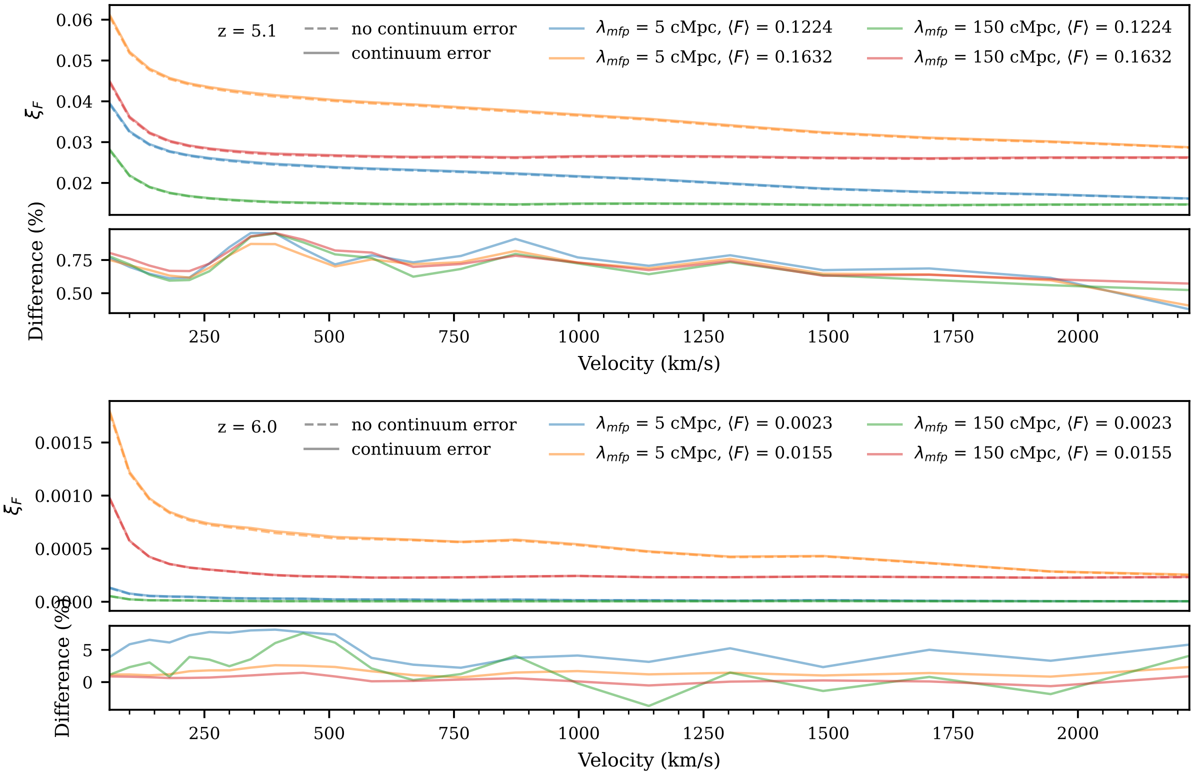

Figure 5.14: The first and third panels show the auto-correlation function from the simulations with (solid line) and without (dashed line) modeling continuum uncertainty at redshifts of 5.1 and 6. The different colors represent different parameter values of $\lambda_{\rm mfp}$ and $\langle F \rangle$ used. The second and fourth panels show the relative difference between these lines defined by Equation (5.10), in percent.

We computed the difference in our measured auto-correlation function at all z with and without continuum error. The difference in the measured data ranges from at most 0.4% to 1.8% with a stronger effect at the highest redshifts.

5.8 Appendix B: DLA modeling effect

In order to investigate how the DLA mask that was described in Section 5.3.3 we compute the measured auto-correlation functions without using this mask. This is shown for all redshifts in Figure 5.15 in red. The original measurement including this mask is shown in black. The measurement at $z = 5.2$ is not impacted at all by the DLA mask as no sightline has a detected DLA in this redshift range. Otherwise, for most scales at most redshifts ignoring the DLA mask reduces the auto-correlation function values. This follows as generally the regions masked in our procedure are regions with high absorption.

5.9 Appendix C: Mean UVB for fixed mean flux

We computed the $\langle \Gamma_{\mathrm{HI}} \rangle$ values that arose in our simulations for given values of $\langle F \rangle$ and λ_{mfp}. These are shown in Figure 5.16 for three fixed values of $\langle F \rangle$. Each fixed $\langle F \rangle$ value is shown in a different color. These lines demonstrate that increasing $\langle \Gamma_{\mathrm{HI}} \rangle$ is required in order to maintain a given $\langle F \rangle$ when decreasing λ_{mfp}. This follows from the effect of small λ_{mfp} on $\langle F \rangle$. Consider Figure 3 from Wolfson et al. (2023b) which shows the flux along the line of sight of a skewer for different λ_{mfp} values. Small λ_{mfp} causes there to be large regions of the Lyα forest with no transmitted flux. Therefore, higher flux values in regions where there is some transmitted flux are required to match the average to models with more areas of transmitted flux (in this case larger λ_{mfp} models), meaning a larger $\langle \Gamma_{\mathrm{HI}} \rangle$. This may seem in conflict to the assumption that $\lambda \propto \Gamma_{\mathrm{HI}}^{2/3}$,

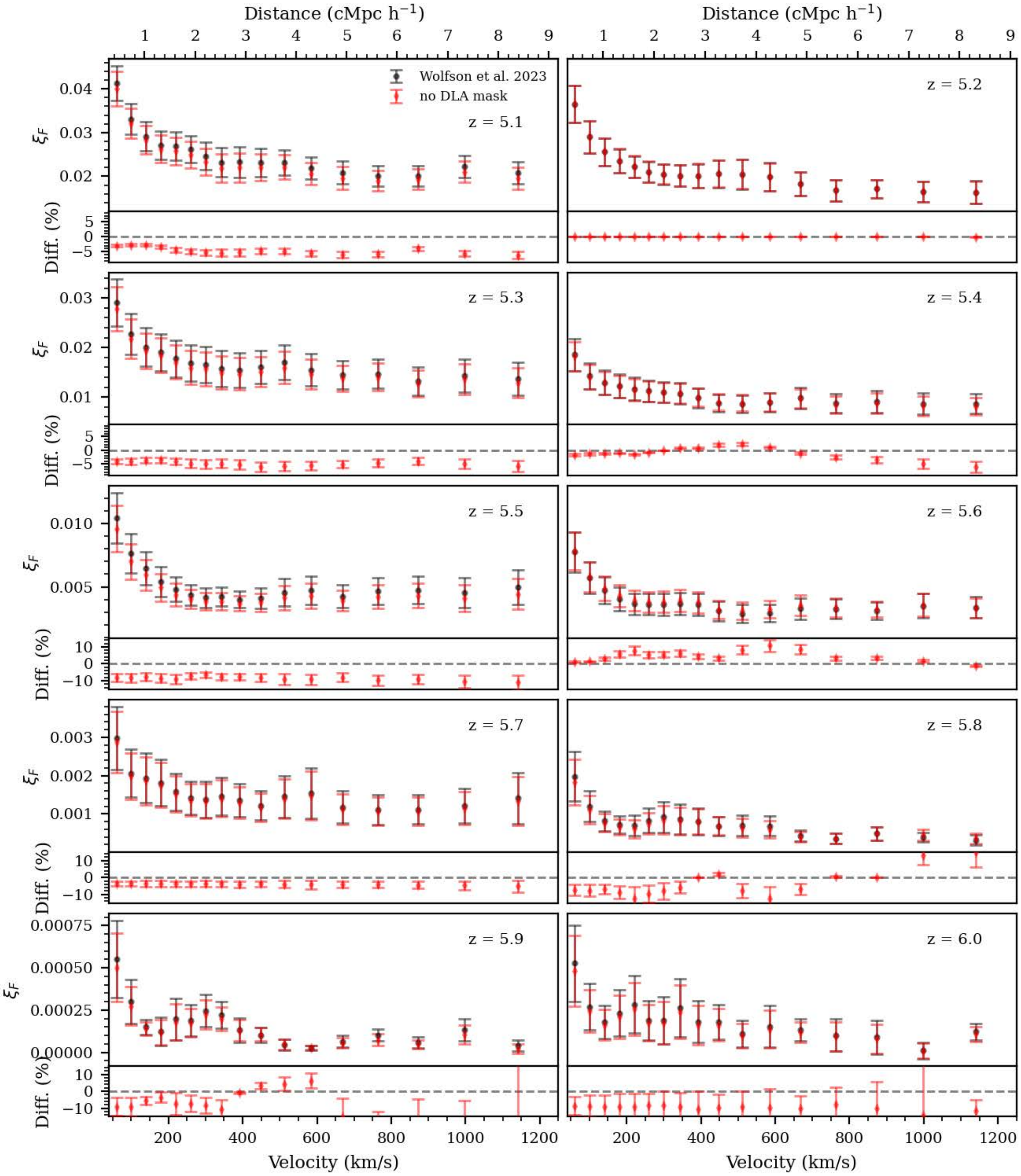

Figure 5.15: The black points show auto-correlation function of Lyα transmission in ten redshift bins measured in this work. The red points show the measured auto-correlation function of the Lyα transmission when ignoring the masks for the DLAs as described in Section 5.3.3. In general, ignoring the DLA mask decreases the auto-correlation function values.

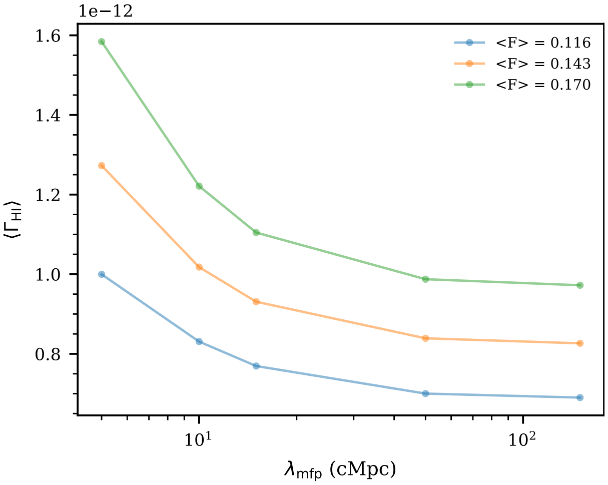

Figure 5.16: The $\langle \Gamma_{\rm HI} \rangle$ values as a function of $\lambda_{\rm mfp}$ for fixed $\langle F \rangle$ at $z = 5.1$. There are three values of $\langle F \rangle$ shown in three colors. This demonstrates that increasing $\langle \Gamma_{\rm HI} \rangle$ if required when decreasing $\lambda_{\rm mfp}$ in order to maintain a given $\langle F \rangle$.

however this is a local relation and the overall average has additional influences.

Bibliography

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

Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, MNRAS, 488, 4440

Bañados, E., Venemans, B. P., Morganson, E., et al. 2014, AJ, 148, 14

Bañados, E., Venemans, B. P., Decarli, R., et al. 2016, ApJS, 227, 11

Bañados, E., Rauch, M., Decarli, R., et al. 2019, ApJ, 885, 59

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

Becker, G. D., Bolton, J. S., Madau, P., et al. 2015, MNRAS, 447, 3402

Becker, G. D., D'Aloisio, A., Christenson, H. M., et al. 2021, MNRAS, 508, 1853

Becker, G. D., Davies, F. B., Furlanetto, S. R., et al. 2018, ApJ, 863, 92

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

—. 2009, ApJ, 698, 1010

Becker, G. D., Sargent, W. L. W., & Rauch, M. 2004, ApJ, 613, 61

Becker, G. D., Pettini, M., Rafelski, M., et al. 2019, ApJ, 883, 163

Bischetti, M., Feruglio, C., D'Odorico, V., et al. 2022, Nature, 605, 244

Boera, E., Becker, G. D., Bolton, J. S., & Nasir, F. 2019, ApJ, 872, 101

Boera, E., Murphy, M. T., Becker, G. D., & Bolton, J. S. 2014, MNRAS, 441, 1916

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

Bolton, J. S., Becker, G. D., Raskutti, S., et al. 2012, MNRAS, 419, 2880

Bolton, J. S., Becker, G. D., Wyithe, J. S. B., Haehnelt, M. G., & Sargent, W. L. W. 2010, MNRAS, 406, 612

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

Bosman, S. E. I. 2021, arXiv e-prints, , arXiv:2108.12446

Bosman, S. E. I., Fan, X., Jiang, L., et al. 2018, MNRAS, 479, 1055

Bosman, S. E. I., Ďurovčíková, D., Davies, F. B., & Eilers, A.-C. 2021, MNRAS, 503, 2077

Bosman, S. E. I., Davies, F. B., Becker, G. D., et al. 2022, MNRAS, 514, 55

Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2015, ApJ, 803, 34

Bryan, G. L., & Machacek, M. E. 2000, ApJ, 534, 57

Cain, C., D'Aloisio, A., Gangolli, N., & Becker, G. D. 2021, ApJL, 917, L37

Calura, F., Tescari, E., D'Odorico, V., et al. 2012, MNRAS, 422, 3019

Carnall, A. C., Shanks, T., Chehade, B., et al. 2015, MNRAS, 451, L16

Carswell, R. F., Whelan, J. A. J., Smith, M. G., Boksenberg, A., & Tytler, D. 1982, MNRAS, 198, 91

Chehade, B., Carnall, A. C., Shanks, T., et al. 2018, MNRAS, 478, 1649

Cooper, T. J., Simcoe, R. A., Cooksey, K. L., et al. 2019, ApJ, 882, 77

Croft, R. A. C. 2004, ApJ, 610, 642

Dall'Aglio, A., Wisotzki, L., & Worseck, G. 2008, A&A, 491, 465

—. 2009, arXiv e-prints, , arXiv:0906.1484

D'Aloisio, A., McQuinn, M., Davies, F. B., & Furlanetto, S. R. 2018, MNRAS, 473, 560

D'Aloisio, A., McQuinn, M., Maupin, O., et al. 2019, ApJ, 874, 154

D'Aloisio, A., McQuinn, M., & Trac, H. 2015, ApJL, 813, L38

Davies, F. B. 2020, MNRAS, 494, 2937

Davies, F. B., Becker, G. D., & Furlanetto, S. R. 2018a, ApJ, 860, 155

Davies, F. B., Bosman, S. E. I., Furlanetto, S. R., Becker, G. D., & D'Aloisio, A. 2021, ApJL, 918, L35

Davies, F. B., & Furlanetto, S. R. 2016a, MNRAS, 460, 1328

—. 2016b, MNRAS, 460, 1328

Davies, F. B., Furlanetto, S. R., & McQuinn, M. 2016, MNRAS, 457, 3006

Davies, F. B., Hennawi, J. F., Eilers, A.-C., & Lukić, Z. 2018b, ApJ, 855, 106

Davies, F. B., Hennawi, J. F., Bañados, E., et al. 2018c, ApJ, 864, 143

—. 2018d, ApJ, 864, 142

Davies, R. L., Ryan-Weber, E., D'Odorico, V., et al. 2023, MNRAS, 521, 289

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10

Dawson, K. S., Kneib, J.-P., Percival, W. J., et al. 2016, AJ, 151, 44

De Rosa, G., Decarli, R., Walter, F., et al. 2011, ApJ, 739, 56

Decarli, R., Walter, F., Venemans, B. P., et al. 2018, ApJ, 854, 97

D'Odorico, V., Viel, M., Saitta, F., et al. 2006, MNRAS, 372, 1333

D'Odorico, V., Feruglio, C., Ferrara, A., et al. 2018, ApJL, 863, L29

D'Odorico, V., Bañados, E., Becker, G. D., et al. 2023, MNRAS, 523, 1399

Eilers, A.-C., Davies, F. B., & Hennawi, J. F. 2018, ApJ, 864, 53

Eilers, A.-C., Hennawi, J. F., & Lee, K.-G. 2017, ApJ, 844, 136

Eilers, A.-C., Hennawi, J. F., Decarli, R., et al. 2021, ApJ, 914, 74

Fan, X., Narayanan, V. K., Lupton, R. H., et al. 2001, AJ, 122, 2833

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

Fang, T., & White, M. 2004, ApJL, 606, L9

Farina, E. P., Arrigoni-Battaia, F., Costa, T., et al. 2019, ApJ, 887, 196

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

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

Francis, P. J., Hewett, P. C., Foltz, C. B., & Chaffee, F. H. 1992, ApJ, 398, 476

Fumagalli, M., O'Meara, J. M., Prochaska, J. X., & Worseck, G. 2013, ApJ, 775, 78

Furlanetto, S. R., Mirocha, J., Mebane, R. H., & Sun, G. 2017, MNRAS, 472, 1576

Furlanetto, S. R., & Oh, S. P. 2005, MNRAS, 363, 1031

Gaikwad, P., Srianand, R., Haehnelt, M. G., & Choudhury, T. R. 2021, MNRAS, 506, 4389

Gaikwad, P., Rauch, M., Haehnelt, M. G., et al. 2020, MNRAS, 494, 5091

Gaikwad, P., Haehnelt, M. G., Davies, F. B., et al. 2023, arXiv e-prints, , arXiv:2304.02038

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

Garzilli, A., Boyarsky, A., & Ruchayskiy, O. 2017, Physics Letters B, 773, 258

Gnedin, N. Y. 2000, ApJ, 535, 530

Gnedin, N. Y., Becker, G. D., & Fan, X. 2017, ApJ, 841, 26

Gnedin, N. Y., & Fan, X. 2006, ApJ, 648, 1

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

Gnedin, N. Y., & Madau, P. 2022, Living Reviews in Computational Astrophysics, 8, 3

Gontcho A Gontcho, S., Miralda-Escudé, J., & Busca, N. G. 2014, MNRAS, 442, 187

Grünwald, P., & van Ommen, T. 2014, arXiv e-prints, , arXiv:1412.3730

Gunn, J. E., & Peterson, B. A. 1965, ApJ, 142, 1633

Haardt, F., & Madau, P. 2012, ApJ, 746, 125

Haehnelt, M. G., & Steinmetz, M. 1998, MNRAS, 298, L21

Hiss, H., Walther, M., Hennawi, J. F., et al. 2018, ApJ, 865, 42

Horne, K. 1986, PASP, 98, 609

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

Hui, L., & Haiman, Z. 2003, ApJ, 596, 9

Iršič, V., Viel, M., Haehnelt, M. G., et al. 2017, Phys. Rev. D, 96, 023522

Jiang, L., Fan, X., Vestergaard, M., et al. 2007, AJ, 134, 1150

Jiang, L., McGreer, I. D., Fan, X., et al. 2015, AJ, 149, 188

—. 2016, ApJ, 833, 222

Jung, I., Finkelstein, S. L., Dickinson, M., et al. 2020, ApJ, 904, 144

Kashino, D., Lilly, S. J., Shibuya, T., Ouchi, M., & Kashikawa, N. 2020, ApJ, 888, 6

Keating, L. C., Kulkarni, G., Haehnelt, M. G., Chardin, J., & Aubert, D. 2020a, MNRAS, 497, 906

Keating, L. C., Weinberger, L. H., Kulkarni, G., et al. 2020b, MNRAS, 491, 1736

Kelson, D. D. 2003, PASP, 115, 688

Khrykin, I. S., Hennawi, J. F., McQuinn, M., & Worseck, G. 2016, ApJ, 824, 133

Kulkarni, G., Hennawi, J. F., Oñorbe, J., Rorai, A., & Springel, V. 2015, ApJ, 812, 30

Kulkarni, G., Keating, L. C., Haehnelt, M. G., et al. 2019, MNRAS, 485, L24

Lai, K., Lidz, A., Hernquist, L., & Zaldarriaga, M. 2006, ApJ, 644, 61

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

Lidz, A., Faucher-Giguère, C.-A., Dall'Aglio, A., et al. 2010, ApJ, 718, 199

Lidz, A., Heitmann, K., Hui, L., et al. 2006, ApJ, 638, 27

Lidz, A., & Malloy, M. 2014, ApJ, 788, 175

Lukić, Z., Stark, C. W., Nugent, P., et al. 2015, MNRAS, 446, 3697

Lynds, R. 1971, ApJL, 164, L73

Mazzucchelli, C., Bañados, E., Venemans, B. P., et al. 2017, ApJ, 849, 91

McDonald, P., Miralda-Escudé, J., Rauch, M., et al. 2001, ApJ, 562, 52

—. 2000, ApJ, 543, 1

McDonald, P., Seljak, U., Cen, R., Bode, P., & Ostriker, J. P. 2005, MNRAS, 360, 1471

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

McGreer, I. D., Mesinger, A., & Fan, X. 2011, MNRAS, 415, 3237

McQuinn, M. 2012, MNRAS, 426, 1349

—. 2016, ARA&A, 54, 313

McQuinn, M., Hernquist, L., Lidz, A., & Zaldarriaga, M. 2011, MNRAS, 415, 977

McQuinn, M., & Upton Sanderbeck, P. R. 2016, MNRAS, 456, 47

Meiksin, A. 2000, MNRAS, 314, 566

Meiksin, A., & McQuinn, M. 2019, MNRAS, 482, 4777

Meiksin, A., & White, M. 2004, MNRAS, 350, 1107

Meiksin, A. A. 2009, Reviews of Modern Physics, 81, 1405

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

—. 2009, MNRAS, 400, 1461

Miralda-Escudé, J., & Rees, M. J. 1994, MNRAS, 266, 343

Morales, A. M., Mason, C. A., Bruton, S., et al. 2021, ApJ, 919, 120

Morrison, J., & Simon, N. 2017, arXiv e-prints, , arXiv:1702.06986

Mortlock, D. J., Patel, M., Warren, S. J., et al. 2009, A&A, 505, 97

Narayanan, V. K., Spergel, D. N., Davé, R., & Ma, C.-P. 2000, ApJL, 543, L103

Nasir, F., Bolton, J. S., & Becker, G. D. 2016, MNRAS, 463, 2335

Nasir, F., & D'Aloisio, A. 2020, MNRAS, 494, 3080

Oñorbe, J., Davies, F. B., Lukić, et al. 2019, MNRAS, 486, 4075

Oñorbe, J., Hennawi, J. F., & Lukić, Z. 2017a, ApJ, 837, 106

Oñorbe, J., Hennawi, J. F., Lukić, Z., & Walther, M. 2017b, ApJ, 847, 63

O'Meara, J. M., Prochaska, J. X., Worseck, G., Chen, H.-W., & Madau, P. 2013, ApJ, 765, 137

Pâris, I., Petitjean, P., Rollinde, E., et al. 2011, A&A, 530, A50

Park, H., Shapiro, P. R., Choi, J.-h., et al. 2016, ApJ, 831, 86

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6

Pontzen, A. 2014, PhRvD, 89, 083010

Pontzen, A., Bird, S., Peiris, H., & Verde, L. 2014, ApJL, 792, L34

Prangle, D., Blum, M. G. B., Popovic, G., & Sisson, S. A. 2013, arXiv e-prints, , arXiv:1301.3166

Prochaska, J. X., Worseck, G., & O'Meara, J. M. 2009, ApJL, 705, L113

Puchwein, E., Bolton, J. S., Haehnelt, M. G., et al. 2015, MNRAS, 450, 4081

Rafelski, M., Wolfe, A. M., Prochaska, J. X., Neeleman, M., & Mendez, A. J. 2012, ApJ, 755, 89

Rauch, M. 1998, ARA&A, 36, 267

Reed, S. L., McMahon, R. G., Martini, P., et al. 2017, MNRAS, 468, 4702

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

Robertson, B. E., Ellis, R. S., Dunlop, J. S., McLure, R. J., & Stark, D. P. 2010, Nature, 468, 49

Rogers, K. K., Peiris, H. V., Pontzen, A., et al. 2019, JCAP, 2019, 031

Rollinde, E., Petitjean, P., Pichon, C., et al. 2003, MNRAS, 341, 1279

Romano, M., Grazian, A., Giallongo, E., et al. 2019, A&A, 632, A45

Rorai, A., Carswell, R. F., Haehnelt, M. G., et al. 2018, MNRAS, 474, 2871

Rorai, A., Hennawi, J. F., Oñorbe, J., et al. 2017, Science, 356, 418

Rudie, G. C., Steidel, C. C., & Pettini, M. 2012, ApJL, 757, L30

Rudie, G. C., Steidel, C. C., Shapley, A. E., & Pettini, M. 2013, ApJ, 769, 146

Schaye, J., Theuns, T., Leonard, A., & Efstathiou, G. 1999, MNRAS, 310, 57

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

Sellentin, E., & Starck, J.-L. 2019, JCAP, 2019, 021

Shapiro, P. R., Iliev, I. T., & Raga, A. C. 2004, MNRAS, 348, 753

Sobacchi, E., & Mesinger, A. 2014, MNRAS, 440, 1662

Sodini, A., D'Odorico, V., Salvadori, S., et al. 2024, arXiv e-prints, , arXiv:2404.10722

Songaila, A., & Cowie, L. L. 2010, ApJ, 721, 1448

Suzuki, N., Tytler, D., Kirkman, D., O'Meara, J. M., & Lubin, D. 2005, ApJ, 618, 592

Takhtaganov, T., Lukić, Z., Müller, J., & Morozov, D. 2021, ApJ, 906, 74

Theuns, T., Schaye, J., & Haehnelt, M. G. 2000, MNRAS, 315, 600

Theuns, T., Schaye, J., Zaroubi, S., et al. 2002a, ApJL, 567, L103

Theuns, T., & Zaroubi, S. 2000, MNRAS, 317, 989

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

Upton Sanderbeck, P. R., D'Aloisio, A., & McQuinn, M. J. 2016, MNRAS, 460, 1885

Ďurovčíková, D., Katz, H., Bosman, S. E. I., et al. 2020, MNRAS, 493, 4256

Venemans, B. P., Verdoes Kleijn, G. A., Mwebaze, J., et al. 2015, MNRAS, 453, 2259

Venemans, B. P., Walter, F., Neeleman, M., et al. 2020, ApJ, 904, 130

Vernet, J., Dekker, H., D'Odorico, S., et al. 2011a, A&A, 536, A105

—. 2011b, A&A, 536, A105

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

Viel, M., Bolton, J. S., & Haehnelt, M. G. 2009, MNRAS, 399, L39

Walther, M., Hennawi, J. F., Hiss, H., et al. 2018, ApJ, 852, 22

Walther, M., Oñorbe, J., Hennawi, J. F., & Lukić, Z. 2019, ApJ, 872, 13

Wang, F., Wu, X.-B., Fan, X., et al. 2016, ApJ, 819, 24

Wang, F., Yang, J., Fan, X., et al. 2019, ApJ, 884, 30

Wang, F., Fan, X., Yang, J., et al. 2021, ApJ, 908, 53

Wang, R., Carilli, C. L., Neri, R., et al. 2010, ApJ, 714, 699

Wang, R., Wagg, J., Carilli, C. L., et al. 2013, ApJ, 773, 44

Willott, C. J., Delorme, P., Omont, A., et al. 2007, AJ, 134, 2435

Wolfe, A. M., Gawiser, E., & Prochaska, J. X. 2005, ARA&A, 43, 861

Wolfson, M., Hennawi, J. F., Davies, F. B., Lukić, Z., & Oñorbe, J. 2023a, arXiv e-prints, , arXiv:2309.05647

Wolfson, M., Hennawi, J. F., Davies, F. B., & Oñorbe, J. 2023b, MNRAS, 521, 4056

Wolfson, M., Hennawi, J. F., Davies, F. B., et al. 2021, MNRAS, 508, 5493

Wolfson, M., Hennawi, J. F., Bosman, S. E. I., et al. 2023c, arXiv e-prints, , arXiv:2309.03341

Worseck, G., Prochaska, J. X., O'Meara, J. M., et al. 2014, MNRAS, 445, 1745

Wu, X.-B., Wang, F., Fan, X., et al. 2015, Nature, 518, 512

Wyithe, J. S. B., Bolton, J. S., & Haehnelt, M. G. 2008, MNRAS, 383, 691

Yang, J., Wang, F., Fan, X., et al. 2020, ApJ, 904, 26

Yèche, C., Palanque-Delabrouille, N., Baur, J., & du Mas des Bourboux, H. 2017, JCAP, 2017, 047

Yip, C. W., Connolly, A. J., Vanden Berk, D. E., et al. 2004, AJ, 128, 2603

Young, P. J., Sargent, W. L. W., Boksenberg, A., Carswell, R. F., & Whelan, J. A. J. 1979, ApJ, 229, 891

Zaldarriaga, M. 2002, ApJ, 564, 153

Zaldarriaga, M., Hui, L., & Tegmark, M. 2001, ApJ, 557, 519

Zel'dovich, Y. B. 1970, A&A, 5, 84

Zhu, Y., Becker, G. D., Bosman, S. E. I., et al. 2021, ApJ, 923, 223

Zhu, Y., Becker, G. D., Christenson, H. M., et al. 2023, ApJ, 955, 115

Ziegel, J. F., & Gneiting, T. 2013, arXiv e-prints, , arXiv:1307.7650

Zuo, L. 1992a, MNRAS, 258, 36

—. 1992b, MNRAS, 258, 45