Statistical Astrophysics

From Extrasolar Planets to the Large-scale Structure of the Universe

By

Collin A. Politsch

Department of Statistics & Data Science

Machine Learning Department

Carnegie Mellon University

A dissertation submitted to Carnegie Mellon University in accordance with the requirements of the joint Doctor of Philosophy degree in the Department of Statistics & Data Science and the Machine Learning Department.

June 2020

© Copyright by Collin A. Politsch, 2020.

All rights reserved.

Department of Statistics & Data Science, Dietrich College
Machine Learning Department, School of Computer Science
McWilliams Center for Cosmology, Mellon College of Science
Carnegie Mellon University

collinpolitsch@gmail.com

This research was conducted under the supervision of Dr. Larry Wasserman, Dr. Jessica Cisewski-Kehe, and Dr. Rupert A. C. Croft from January 2015 to June 2020.

First release, June 2020

Approved by:

Larry Wasserman
Advisor

Jessica Cisewski-Kehe
Co-Advisor

Rupert A. C. Croft
Co-Advisor

Ryan J. Tibshirani
Committee Member

Peter E. Freeman
Committee Member

Richard Scheines
Dean, Dietrich College of Humanities & Social Sciences

Tom M. Mitchell
Dean, School of Computer Science

Declaration of Authorship

I, Collin A. Politsch, declare that this dissertation titled, *Statistical Astrophysics: From Extrasolar Planets to the Large-scale Structure of the Universe* and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a research degree at this University.
- Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- Where the dissertation is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

For my dear mother, Carol.

1962 - 2015

"If we really saw the Universe, maybe we would understand it."

Jorge Luis Borges

Carnegie Mellon University

Statistical Astrophysics: From Extrasolar Planets to the Large-scale Structure of the Universe

by Collin A. Politsch

Abstract

Modern astronomical surveys compile massive catalogs of images, light curves, and spectra that allow us to study the Universe from the relatively local (e.g. stars and extrasolar planets) to the extremely remote (e.g. quasars and the cosmic microwave background). As the magnitude of these catalogs continues to grow exponentially, the presence of statisticians and computer scientists working at the interface of astronomy and astrophysics becomes increasingly essential to the advancement of the field. In this dissertation, we study a variety of astrophysical problems of a statistical nature.

In Chapters 2 and 3, we introduce trend filtering (Tibshirani, 2014) into the astronomical literature and demonstrate its broad utility by discussing how it can contribute to a variety of spectroscopic and time-domain studies. The astronomical observations we discuss are (1) the Lyman-α forest absorptions in the spectra of high redshift quasars; (2) the broader spectroscopic signatures of quasars, galaxies, and stars; (3) stellar light curves with planetary transits; (4) light curves of eclipsing binary star systems; and (5) supernova light curves. We study the Lyman-α forest in the greatest detail — using trend filtering to map the large-scale structure of the intergalactic medium along one-dimensional quasar-observer sightlines. The remaining studies share broad themes of: (1) estimating observable parameters of light curves and spectra; and (2) constructing observational spectral/light-curve templates.

In Chapters 4 and 5, we continue our Lyman-α absorption spectroscopy analysis of the intergalactic medium by utilizing the full redshift $z \gtrsim 2.1$ quasar catalog compiled by the Baryon Oscillation Spectroscopic Survey (Dawson et al., 2013) to reconstruct a 47 h^{-3} Gpc^3 three-dimensional large-scale structure map of the high redshift intergalactic medium — the largest volume map of the Universe to date — from the dense collection of one-dimensional quasar sightlines. We accompany the map with rigorous statistical error quantification and compile an extensive census of candidates for never-before-seen galaxy protoclusters and cosmic voids. The statistical reconstruction requires minimal assumptions on the underlying matter density field and is specifically optimized to recover three-dimensional structures lying between the one-dimensional sightlines backlit by quasars.

Acknowledgements

First, I would like to thank my primary dissertation advisor, Larry Wasserman, for his mentorship throughout the entirety of my time at Carnegie Mellon. So much of my growth during these years is directly attributable to his teachings and keen insights in supervising my research, his inspiration in the classroom during my early years of the Ph.D., and our many semesters spent working as a professor / head TA tandem. And most importantly, his compassionate nature helped make Carnegie Mellon a home away from home for me during a period in which I was navigating many difficult events in my personal life.

Many thanks to my dissertation co-advisors, Jessi Cisewski-Kehe and Rupert Croft, who helped me discover my love for astrophysics and played primary roles in my development into the interdisciplinary scientist I am today. I am incredibly fortunate to have had their guidance and mentorship, and I look forward to a long future of collaborations.

Thank you to the Ph.D. students of the Department of Statistics & Data Science — my friends, classmates, and colleagues — for truly making the department feel like a family. And an especially warm thanks to my Stats & DS Ph.D. cohort of 2014, a group of outstanding human beings and exceptionally talented individuals who navigated every single challenge the Ph.D. years brought in the same manner — together.

Thank you to my undergraduate advisor, Tyrone Duncan, who guided me through my early years of academic research at the University of Kansas and prepared me for a successful transition to the rigorous world of graduate school. Thank you to Ryan Tibshirani for his friendship, inspiration, and generous feedback on many aspects of the trend filtering work in this dissertation.

Finally, thank you to my family: Kent, Doris, Karen, Allison, Beth, and Isaac. Your unconditional love and support guides me in everything I do, and was absolutely essential during this period of time in which we endured the loss of my mother, Carol. I love you and I miss you dearly, Mom. This was all for you.

Contents

List of Figures

List of Tables

Abbreviations

BJD	**B**arycentric **J**ulian **D**ate
BOSS	**B**aryon **O**scillation **S**pectroscopic **S**urvey
Dec	**Dec**lination
DESI	**D**ark **E**nergy **S**pectroscopic **I**nstrument
DGP	**D**ata **G**enerating **P**rocess
DR	**D**ata **R**elease
EB	**E**clipsing **B**inary
eBOSS	**E**xtended **B**aryon **O**scillation **S**pectroscopic **S**urvey
Gpc	**G**iga**p**arse**c**
GRF	**G**aussian **R**andom **F**ield
GRR	**G**eneralized **R**idge **R**egression
IGM	**I**nter**g**alactic **M**edium
kpc	**k**ilo**p**arse**c**
LC	**L**ight **C**urve
LOS	**L**ine **o**f **S**ight
Lyα	**Ly**man-**α**
MJD	**M**odified **J**ulian **D**ate
Mpc	**M**ega**p**arse**c**
NGC	**N**orthern **G**alactic **C**ap
RA	**R**ight **A**scension
RBF	**R**adial **B**asis **F**unction
SDSS	**S**loan **D**igital **S**ky **S**urvey
SKRR	**S**patial **K**ernel **R**idge **R**egression
SGC	**S**outhern **G**alactic **C**ap
SN	**S**uper**n**ova
SURE	**S**tein's **U**nbiased **R**isk **E**stimate

Symbols

Symbol	Name	Units
λ	wavelength	Å
t	time	various
$f(\cdot)$	observed flux (noisy)	ergs/s/cm^2/Å
$f_0(\cdot)$	flux signal	ergs/s/cm^2/Å
$C(\cdot)$	continuum flux	ergs/s/cm^2/Å
$F(\cdot)$	transmitted flux fraction	none
$\overline{F}(\cdot)$	mean transmitted flux fraction	none
$\delta_F(\cdot)$	flux contrast	none
z	redshift	none

Chapter 1

Introduction

"We live on a hunk of rock and metal that circles a humdrum star that is one of 400 billion other stars that make up the Milky Way Galaxy, which is one of billions of other galaxies which make up a universe..."

Carl Sagan

The title of this dissertation, *Statistical Astrophysics: From Extrasolar Planets to the Large-scale Structure of the Universe*, serves to illustrate the full scale of astrostatistical problems we study in this dissertation, with extrasolar planets being the smallest scale phenomenon and the large-scale structure of the Universe (as revealed by the intergalactic medium) being the grandest cosmic stage. The astrophysical phenomena discussed in this dissertation that tacitly lie between these two titular extremes are: stars, eclipsing binary star systems, supernovae, galaxies, and quasars. We give a brief introduction below regarding the specific statistical problem we address for each phenomenon and we provide an outline for how our work is organized throughout this document.

Starting on the smallest scale (relatively speaking), extrasolar planets in our neighborhood of the Milky Way Galaxy can be detected by studying the apparent brightness of their host stars over time (i.e. the stellar light curve). In fortuitous circumstances, the orbit of an extrasolar planet may happen to pass directly in front of the line of sight between its host star and an observer on Earth. In such circumstances, the planet can be detected via a periodic dip in the apparent brightness of the star caused by the planetary eclipse event. This approach to detecting extrasolar planets — called the *transit method* — is currently the most effective method for detecting planets outside the Solar System [3]. And while the probability that a planet eclipses its host star from the vantage point of Earth is very small for any given planetary system, the massive catalogs of stellar light curves compiled by the *CoRoT* [4], *Kepler* [5, 6], and *TESS* [7] missions have enabled the detection of over 4,000 confirmed extrasolar planets to date, with several thousand more unconfirmed candidates. Furthermore, by studying the observable parameters of these stellar light curves with planetary transits (in particular, the planetary orbital period, the transit depth, the transit duration, and the ingress/egress duration), physical parameters of the system such as the radius of the planet, the semi-major axis, and the eccentricity and inclination of the orbit can be calculated. In this dissertation, we propose a novel method for nonparametrically estimating the transit depth and the transit duration given a phase-folded light curve of a planetary transit event. Our method may also be able to produce faithful estimates of the ingress/egress duration but further vetting is required in order to assert that claim definitively.

Just above extrasolar planets on the cosmic scale come stars. The diversity of stars observed within our galaxy is significant. Stars are categorized into one of seven main spectral classes — ranging from the coolest (M stars) to the hottest (O stars) — and then are further subcategorized based on their luminosity. Given a pair of stars with the same temperature, the luminosity subclass then differentiates between their sizes (e.g. subdwarfs, main-sequence stars, giants, supergiants). On the next level of the cosmic hierarchy, galaxies and quasars similarly exhibit rich intra-class

physical diversity — for example, differing morphologies and relative chemical compositions. Given an observed non-transient light source in the sky, the most powerful observational tool for determining its first order class — e.g. star, galaxy, or quasar — as well as its redshift (a monotonic function of its radial distance) and all subsequent hierarchical subclassifications is the coadded electromagnetic spectrum of the object, which encodes the intensities of the observed light across a continuous spectrum of wavelengths. The spectrum serves as a signature of the object from which we can decode these classifications, subclassifications, and redshifts from the relative strengths, widths, and shifts of known emission and absorption lines. In the modern age of massive astronomical surveys, it is essential to be able to carry out this process in a fully automated fashion. In order to do so we require large libraries of spectral templates that span the physical diversity of each object class so we can provide our statistical machine learning algorithms with a frame of reference from which they can learn how to classify celestial objects and their redshifts from their observational spectra. In this dissertation, we provide a flexible and efficient approach for compiling such spectral template libraries from observational spectra.

Stars may also be observed in pairs that orbit one another, which is referred to as a *binary star system* (for example, recall in *Star Wars* the Tatoo system that hosts Luke's home planet of Tatooine). Similar to the transit method for detecting and characterizing extrasolar planets, binary star systems can be identified by periodic dips in their light curves if we are fortunate enough for the stars to eclipse one another from our vantage point on Earth. In such cases, the system is referred to as an *eclipsing binary* (EB) star system. An EB light curve is characterized by periodic dips in the observed brightness that correspond to the eclipse events along the line of sight to an observer. In particular, there are two eclipses per orbital period — a primary and a secondary eclipse. The primary eclipse occurs when the hotter star is eclipsed by the cooler star and produces a comparatively deep dip in the observed brightness of the system. Conversely, the secondary eclipse occurs when the cooler star is eclipsed by the hotter star and produces a

comparatively shallow dip in the observed brightness. Depending on the effective temperature ratio and orbital period of the EB, the dips may range from very narrow and abrupt to very wide and smooth. Analogous to the transit method for extrasolar planets, physical parameters of an EB system (e.g. the temperature ratio, sum of fractional radii, photometric mass ratio, radial and tangential components of the eccentricity, fillout factor, and inclination) can be learned from the observable parameters of the light curve (e.g. the eclipse widths, depths, and separations). The current state-of-the-art procedure for learning these physical parameters consists of the following steps: (1) determine the orbital period of the binary system via a Fourier analysis of the observed light curve; (2) phase-fold the light curve with respect to the estimated orbital period; (3) denoise the phase-folded light curve and evaluate the denoised estimate on a regular grid in the phase space; (4) input the denoised phase-folded EB light curve into an artificial neural network (ANN) trained on a rigorous physical model to produce estimates for the physical parameters of the EB system [8]. In this dissertation, we propose a novel approach for denoising the phase-folded light curves of EB systems (the third step in the procedure above) that significantly improves upon the method currently utilized by the *Kepler* EB pipeline [9]. Our approach significantly reduces the statistical bias of the denoised phase-folded light curve estimate, which in turn reduces the systematic bias in the data that is provided to the ANN. If implemented in modern EB pipelines, we are confident our approach will significantly improve the accuracy of the estimated physical parameters of observed EB star systems.

A supernova (SN) is the death of a star that manifests as a catastrophic explosion producing a transient luminosity that, at maximum, can be comparable to an entire galaxy. SNe show light-curve variations on various time scales, with the initial core collapse occurring in a matter of seconds, the ascension to maximum light occurring over weeks or months, and the subsequent slow decay occurring over years. Type classifications and subclassifications of SNe are made based on both their light curves and their spectra. Analogous to our above discussion regarding

coadded spectral templates, fully automated SN classifications are enabled by compiling light-curve/spectral template libraries that span the full diversity of SNe within each class/subclass, as well as studying the observable parameters of the SN light curve (e.g. the maximum apparent brightness, the time of maximum, and the decline rate). In this dissertation, we provide a novel approach for constructing these light-curve/spectral templates from observational data and nonparametrically estimating the observable parameters of SN light curves. The improvement yielded by our approach primarily corresponds to cases where the observed SN has an especially high peak brightness and a fast decline rate. This behavior is particularly characteristic of the Type Ia SN class [10].

Finally, we devote the majority of this dissertation to the study of the largest scale astrophysical phenomenon — the intergalactic medium. The intergalactic medium (IGM) is a highly dilute gaseous medium that permeates the overwhelming volume of intergalactic space and contains a majority of the baryonic matter in the Universe. This ubiquitous gas is too diffuse to be directly observed in emission, but its presence is traced by absorptions in the light of luminous background sources — most notably, quasars, whose extreme luminosities enable studies of the IGM at vast radial distances. As light travels from a distant quasar along its path to Earth, the IGM leaves an absorption signature in the light, marking the atomic elements that are present in the intergalactic gas the light passes through [11]. This signature collectively reveals the presence of diffuse primordial hydrogen and helium residue in intergalactic space left over from the Big Bang, as well as a variety of metals occasionally ejected from galaxies by particularly forceful supernovae explosions [12]. However, the bulk of the IGM is composed of electrically neutral hydrogen (H I) gas, which marks its presence by absorbing light at the Lyman-α wavelength (1215.67 Å). Due to the constant doppler shifting of extragalactic radiation caused by the expansion of the Universe, Lyman-α absorptions are stretched over a series of wavelengths called the *Lyman-α forest*, which effectively provides a full one-dimensional map of the intergalactic H I

density distribution along the quasar-observer line of sight. For many years the one-dimensional analysis of Lyman-α forest spectra has enabled spatial analyses of the cosmological matter density distribution [e.g. 13–16], and more recently the denser sampling of observed quasar sightlines has enabled coherent measurements of large-scale structure across sightlines [e.g. 17–20]. The fact that structure can be measured across sightlines means that spatial statistical methods can be used to reconstruct full three-dimensional maps of the matter density distribution from closely sampled sets of quasar sightlines — a problem commonly known as *Lyman-α forest tomography* or *intergalactic medium tomography*. However, without the development of scalable statistical methods for handling the massive catalogs of quasar spectra collected by modern sky surveys, three-dimensional mapping via the Lyman-α forest has to this point been limited to simulated data sets and very small observational volumes [21–25]. In this dissertation, we present a 47 h^{-3} Gpc^3 Lyman-α absorption large-scale structure map of the intergalactic medium — the largest volume map of the Universe to date — which we reconstructed using a sample of approximately 160,000 quasar sightlines collected by the SDSS-III Baryon Oscillation Spectroscopic Survey [2, 26, 27]. In additional to the optimized map itself, we provide rigorous statistical inference that allows us to provide an extensive census of high significance candidates for never-before-seen galaxy protoclusters and cosmic voids embedded in the reconstructed absorption field. In total, this spatial analysis of intergalactic medium required approximately 3 million CPU hours of memory-heavy computations, for which we gratefully thank Yale University for the use of their YCRC high performance computing infrastructure.

The layout of this dissertation is as follows. In Chapter 2, we provide a detailed introduction of trend filtering [1, 28] into the astronomical literature. In Chapter 3, we detail our novel contributions to the various areas of one-dimensional time-domain astronomy and astronomical spectroscopy itemized above, which are powered by the application of trend filtering to astronomical observations. In Chapter 4, we describe our early investigations toward reconstructing a

large-scale map of the intergalactic medium via the Lyman-α forest. Finally, in Chapter 5, we detail our final Lyman-α absorption large-scale structure map of the intergalactic medium.

Chapter 2

Trend Filtering: A Modern Statistical Tool for Time-Domain Astronomy and Astronomical Spectroscopy

This chapter is based on our paper *Trend Filtering – I: A Modern Statistical Tool for Time-Domain Astronomy and Astronomical Spectroscopy* [29], which was published in *Monthly Notices of the Royal Astronomical Society* and was the first to introduce trend filtering [1, 28] into the astronomical literature.

The problem of denoising a one-dimensional signal possessing varying degrees of smoothness is ubiquitous in time-domain astronomy and astronomical spectroscopy. For example, in the time domain, an astronomical object may exhibit a smoothly varying intensity that is occasionally interrupted by abrupt dips or spikes. Likewise, in the spectroscopic setting, a noiseless spectrum typically contains intervals of relative smoothness mixed with localized higher frequency components such as emission peaks and absorption lines. In this work, we present trend filtering [1], a modern nonparametric statistical tool that yields significant improvements in this broad problem space of denoising *spatially heterogeneous* signals. When the underlying signal is spatially

heterogeneous, trend filtering is superior to any statistical estimator that is a linear combination of the observed data — including kernel smoothers, LOESS, smoothing splines, Gaussian process regression, and many other popular methods. Furthermore, the trend filtering estimate can be computed with practical and scalable efficiency via a specialized convex optimization algorithm [30], e.g. handling sample sizes of $n \gtrsim 10^7$ within a few minutes. In a companion paper [31] (Chapter 3), we explicitly demonstrate the broad utility of trend filtering to observational astronomy by carrying out a diverse set of spectroscopic and time-domain analyses.

2.1 Introduction

Many astronomical observations produce one-dimensional data with unknown or varying degrees of smoothness. These include data from time-domain astronomy, where transient events such as supernovae can show light-curve variations on timescales ranging from seconds to years [e.g., 32, 33]. Similarly, in astronomical spectroscopy, with wavelength (or frequency) as the independent variable, sharp absorption or emission-line features can be present alongside smoothly varying black-body or other continuum radiation [see, e.g., 11]. In each of these general settings, we observe a signal plus noise and would like to denoise the signal as accurately as possible. Indeed the set of statistical tools available for addressing this general problem is quite vast. Commonly used nonparametric regression methods include kernel smoothers [e.g., 34, 35], local polynomial regression [LOESS; e.g., 36, 37], splines [e.g., 38–40], Gaussian process regression [e.g., 41–43], and wavelet decompositions [e.g., 44–46]. A rich and elegant statistical literature exists on the theoretical and practical achievements of these methods [see, e.g., 47–49, for general references]. However, when the underlying signal is *spatially heterogeneous*, i.e. exhibits varying degrees of smoothness, the power of classical statistical literature is quite limited. Kernels, LOESS, smoothing splines, and Gaussian process regression belong to a broad family of nonparametric

methods called *linear smoothers*, which has been shown to be uniformly suboptimal for estimating spatially heterogeneous signals [50–52]. The common limitation of these methods is that they are not locally adaptive; i.e., by construction, they do not adapt to local degrees of smoothness in a signal. In particular, continuing with the example of a smoothly varying signal with occasional sharp features, a linear smoother will tend to oversmooth the sharp features and/or overfit the smooth regions in its effort to optimally balance statistical bias and variance. Considerable effort has been made to address this problem by locally varying the hyperparameter(s) of a linear smoother. For example, locally varying the kernel bandwidth [e.g., 53–57] irregularly varying spline knot locations [e.g., 58–60], and constructing non-stationary covariance functions for Gaussian process regression [e.g., 61–63]. However, since hyperparameters typically need to be estimated from the data, such exponential increases in the hyperparameter complexity severely limit the practicality of choosing the hyperparameters in a fully data-driven, generalizable, and computationally efficient fashion. Wavelet decompositions offer an elegant solution to the problem of estimating spatially heterogeneous signals, providing both statistical optimality [e.g., 52, 64] and only requiring data-driven tuning of a single (scalar) hyperparameter. Wavelets, however, possess the practical limitation of requiring a stringent analysis setting, e.g. equally-spaced inputs and sample size equal to a power of two, among other provisions; and when these conditions are violated, the optimality guarantees are void. So, seemingly at an impasse, the motivating question for this work is *can we have the best of both worlds?* More precisely, is there a statistical tool that simultaneously possesses the following properties:

1. Statistical optimality for estimating spatially heterogeneous signals

2. Practical analysis assumptions; for example, not limited to equally-spaced inputs

3. Practical and scalable computational speed

4. A one-dimensional hyperparameter space, with automatic data-driven methods for selection

In this chapter we introduce trend filtering [1], a statistical method that is new to the astronomical literature and provides a strong affirmative answer to this question.

The layout of this chapter is as follows. In Section 2.2 we provide both theoretical and empirical evidence of the superiority of trend filtering for estimating spatially heterogeneous signals compared to classical statistical methods. In Section 2.3 we introduce trend filtering, including a general overview of the estimator's machinery, its connection to spline methods, automatic methods for choosing the hyperparameter, uncertainty quantification, generalizations, and recommended software implementations in various programming languages. In Chapter 3, we directly illustrate the broad utility of trend filtering to astronomy by carrying out a diverse set of spectroscopic and time-domain analyses.

2.2 Classical statistical methods and their limitations

We begin this section by providing background and motivation for the nonparametric approach to estimating (or denoising) signals. We then discuss statistical optimality for estimating spatially heterogeneous signals, with an emphasis on providing evidence for the claim that trend filtering is broadly superior to classical statistical methods in this setting. Finally, we end this section by illustrating this superiority with a direct empirical comparison of trend filtering and several popular classical methods on simulated observations of a spatially heterogeneous signal.

2.2.1 Nonparametric regression

Suppose we observe noisy measurements of a response variable of interest (e.g., flux, magnitude, photon counts) according to the data generating process (DGP)

$$f(t_i) = f_0(t_i) + \epsilon_i, \qquad i = 1, \ldots, n \tag{2.1}$$

where $f_0(t_i)$ is the signal at input t_i (e.g., a time or wavelength) and ϵ_i is the noise at t_i that contaminates the signal, giving rise to the observation $f(t_i)$. Let $t_1, \ldots, t_n \in (a, b)$ denote the observed input interval and $\mathbb{E}[\epsilon_i] = 0$ (where we use $\mathbb{E}[\cdot]$ to denote mathematical expectation). Here, the general statistical problem is to estimate (or *denoise*) the underlying signal f_0 from the observations as accurately as possible. In the nonparametric setting, we refrain from making strong *a priori* assumptions about f_0 that could lead to significant modeling bias, e.g. assuming a power law or a light-curve/spectral template fit. Mathematically, a nonparametric approach is defined through the deliberately weak assumption $f_0 \in \mathcal{F}$ (i.e. the signal belongs to the function class $\mathcal{F}$) where $\mathcal{F}$ is *infinite-dimensional*. In other words, the assumed class of all possible signals $\mathcal{F}$ cannot be spanned by a finite number of parameters. Contrast this to the assumption that the signal follows a pth degree power law, i.e. $f_0 \in \mathcal{F}_{\text{PL}}$ where

$$\mathcal{F}_{\text{PL}} = \left\{ f_0 : f_0(t) = \beta_0 + \sum_{j=1}^{p} \beta_j t^j \right\}, \tag{2.2}$$

a class that is spanned by $p + 1$ parameters. Similarly, given a set of p spectral/light-curve templates $b_1(t), \ldots, b_p(t)$, the usual template-fitting assumption is that $f_0 \in \mathcal{F}_{\text{TEMP}}$ where

$$\mathcal{F}_{\text{TEMP}} = \left\{ f_0 : f_0(t) = \beta_0 + \sum_{j=1}^{p} \beta_j b_j((t - s)/v) \right\} \tag{2.3}$$

and s and v are horizontal shift and stretch hyperparameters, respectively. Both equations (2.2) and (2.3) represent very stringent assumptions about the underlying signal f_0. If the signal is anything other than exactly a power law in t — a highly unlikely occurrence — nontrivial statistical bias will arise by modeling it as such. Likewise, if a class of signals has a rich physical diversity [e.g., Type Ia supernova light curves, 10] that is not sufficiently spanned by the library of templates used in modeling, then statistical biases will arise. Depending on the size of the imbalance between class diversity and the completeness of the template basis, the biases could be significant. Moreover, these biases are rarely tracked by uncertainty quantification. To be clear, this is not a uniform criticism of template-fitting. For example, templates are exceptionally powerful tools for object classification and redshift estimation [e.g., 65, 66]. Furthermore, much of our discussion in Chapter 3 centers around utilizing the flexible nonparametric nature of trend filtering to construct more complete spectral/light-curve template libraries for various observational objects and transient events.

Let $\widehat{f}_0$ be any statistical estimator for the signal f_0, derived from the noisy observations in equation (2.1). Further, let $p_t(t)$ denote the probability density function (pdf) that specifies the sampling distribution of the inputs on the interval (a, b), and let $\sigma^2(t) = \text{Var}(\epsilon(t))$ denote the noise level at input t. In order to assess the accuracy of the estimator it is common to consider the mean-squared prediction error (MSPE):

$$R(\widehat{f}_0) = \mathbb{E}\big[(\widehat{f}_0 - f)^2\big] \tag{2.4}$$

$$= \mathbb{E}\big[(\widehat{f}_0 - f_0)^2\big] + \overline{\sigma}^2 \tag{2.5}$$

$$= \int_a^b \Big(\text{Bias}^2(\widehat{f}_0(t)) + \text{Var}(\widehat{f}_0(t))\Big) \cdot p_t(t)dt + \overline{\sigma}^2, \tag{2.6}$$

where

$$\text{Bias}(\widehat{f}_0(t)) = \mathbb{E}[\widehat{f}_0(t)] - f_0(t) \tag{2.7}$$

$$\text{Var}(\widehat{f}_0(t)) = \mathbb{E}\Big(\widehat{f}_0(t) - \mathbb{E}[\widehat{f}_0(t)]\Big)^2 \tag{2.8}$$

$$\overline{\sigma}^2 = \int_a^b \sigma^2(t) \cdot p_t(t) dt. \tag{2.9}$$

The equality in equation (2.6) is commonly referred to as the bias-variance decomposition. The first term is the squared bias of the estimator $\widehat{f}_0$ (integrated over the input interval) and intuitively measures how appropriate the chosen statistical estimator is for modeling the observed phenomenon. The second term is the variance of the estimator that measures how stable or sensitive the estimator is to the observed data. And the third term is the irreducible error — the minimum prediction error we cannot hope to improve upon. The bias-variance decomposition therefore illustrates that an optimal estimator is one that combines appropriate modeling assumptions (low bias) with high stability (low variance).

2.2.1.1 Statistical optimality (minimax theory)

In this section, we briefly discuss a mathematical framework for evaluating the performance of statistical methods over nonparametric signal classes in order to demonstrate that the superiority of trend filtering is a highly general result. Ignoring the irreducible error, the problem of minimizing the MSPE of a statistical estimator can be equivalently stated as a minimization of the first term in equation (2.5) — the mean-squared estimation error (MSEE). In practice, low bias is attained by only making very weak assumptions about what the underlying signal may look like, e.g. f_0 has k continuous derivatives. An ideal statistical estimator for estimating

signals in such a class (call it $\mathcal{F}$) may then be defined as

$$\inf_{\widehat{f}_0} \Big(\sup_{f_0 \in \mathcal{F}} \mathbb{E}\big[(\widehat{f}_0 - f_0)^2\big] \Big). \tag{2.10}$$

That is, we would like our statistical estimator to be the minimizer (infimum) of the worst-case (supremum) MSEE over the signal class $\mathcal{F}$. This is rarely a mathematically tractable problem for any practical signal class $\mathcal{F}$. A more tractable approach is to consider how the worst-case MSEE behaves as a function of the sample size n. A reasonable baseline metric for a statistical estimator is to require that it satisfies

$$\sup_{f_0 \in \mathcal{F}} \mathbb{E}\big[(\widehat{f}_0 - f_0)^2\big] \to 0 \tag{2.11}$$

as $n \to \infty$. That is, for any signal $f_0 \in \mathcal{F}$, when a large amount of data is available, $\widehat{f}_0$ gets arbitrarily close to the true signal. In any practical situation, this is not true for parametric models because the bias component of the decomposition never vanishes. This, however, is a widely-held — perhaps, defining — property of nonparametric methods. Therefore, in order to distinguish optimality among nonparametric estimators, we require a stronger metric. In particular, we study *how quickly* the worst-case error goes to zero as more data is observed. This is the core idea of a rich area of statistical literature called *minimax theory* [see, e.g., 48, 67, 68]. For many infinite-dimensional classes of signals, theoretical lower-bounds exist on the rate at which the MSEE of *any* statistical estimator can approach zero. Therefore, if a statistical estimator is shown to achieve that rate, it can be considered optimal for estimating that class of signals. Formally, letting $g(n)$ be the rate at which the MSEE of the theoretically optimal estimator in equation (2.10) goes to zero (a monotonically decreasing function in n), we would like our estimator $\widehat{f}_0$ to satisfy

$$\sup_{f_0 \in \mathcal{F}} \mathbb{E}\big[(\widehat{f}_0 - f_0)^2\big] = \mathcal{O}(g(n)), \tag{2.12}$$

where we use $\mathcal{O}(\cdot)$ to denote big O notation. If this is shown to be true, we say the estimator *achieves the minimax rate over the signal class* $\mathcal{F}$. Loosely speaking, we are stating that a *minimax optimal* estimator is an estimator that learns the signal from the data just as quickly as the theoretical gold standard estimator in equation (2.10).

2.2.1.2 Spatially heterogeneous signals

Thus far we have only specified that the signal underlying most one-dimensional astronomical observations should be assumed to belong to a class $\mathcal{F}$ that is infinite-dimensional (i.e. non-parametric). Further, in Section 2.2.1.1 we introduced the standard metric used to measure the performance of a statistical estimator over an infinite-dimensional class of signals. Recalling the discussion in the abstract and Section 2.1, trend filtering provides significant advances for estimating signals that exhibit varying degrees of smoothness across the input domain. We restate this definition below.

Definition. A *spatially heterogeneous* signal is a signal that exhibits varying degrees of smoothness in different regions of its input domain.

Example. A smooth light curve with abrupt transient events.

Example. An electromagnetic spectrum with smooth continuum radiation and sharp absorption/emission-line features.

To complement the above definition we may also loosely define a *spatially homogeneous signal* as a signal that is *either* smooth *or* wiggly[1] across its input domain, but not both. As "smoothness" can be quantified in various ways these definitions are intentionally mathematically imprecise. A

[1] This is, in fact, a technical term used in the statistical literature.

class that is commonly considered in the statistical literature is the L_2 Sobolev class:

$$\mathcal{F}_{2,k}(C_1) := \left\{ f_0 : \int_a^b f_0^{(k)}(t)^2 dt < C_1 \right\}, \qquad C_1 > 0,\ k \in \mathbb{N}. \tag{2.13}$$

That is, an L_2 Sobolev class is a class of all signals such that the integral of the square (the "L_2 norm") of the kth derivative of each signal is less than some constant C_1. Statistical optimality in the sense of Section 2.2.1.1 for estimating signals in these classes (and some other closely related ones) is widely held among statistical methods in the classical toolkit; for example, kernel smoothers [69, 70], LOESS [71, 72], and smoothing splines [73]. However, a seminal result by [50, 51] showed that a statistical estimator can be minimax optimal over signal classes of the form (2.13) and still perform quite poorly on other signals. In particular, the authors showed that, when considering the broader L_1 Sobolev class

$$\mathcal{F}_{1,k}(C_2) := \left\{ f_0 : \int_a^b \left| f_0^{(k)}(t) \right| dt < C_2 \right\}, \qquad C_2 > 0,\ k \in \mathbb{N}, \tag{2.14}$$

all linear smoothers[2] — including kernels, LOESS, smoothing splines, Gaussian process regression, and many other methods — are strictly suboptimal. The key difference between these two types of classes is that L_2 Sobolev classes are rich in spatially homogeneous signals but not spatially heterogeneous signals, while L_1 Sobolev classes[3] are rich in both [see, e.g., 52].

The intuition of this result is that linear smoothers cannot optimally recover signals that exhibit varying degrees of smoothness across their input domain because they operate as if the signal possesses a fixed degree of smoothness. For example, this intuition is perhaps most clear when considering a kernel smoother with a fixed bandwidth. The result of [50, 51] therefore implies

[2] A *linear smoother* is a statistical estimator that is a linear combination of the observed data. Many popular statistical estimators, although often motivated from seemingly disparate premises, can be shown to fall under this definition. See, e.g., [48] for more details.

[3] The L_1 Sobolev class is often generalized to a nearly equivalent but slightly larger class — namely, signals with derivatives of bounded total variation. See [1] for the generalized definition.

that, in order to achieve statistical optimality for estimating spatially heterogeneous signals, a statistical estimator must be nonlinear (more specifically, it must be locally adaptive). [1] showed that trend filtering is minimax optimal for estimating signals in L_1 Sobolev classes. Since L_2 Sobolev classes are contained within L_1 Sobolev classes, this result also guarantees that trend filtering is also minimax optimal for estimating signals in L_2 Sobolev classes. Wavelets share this property, but require restrictive assumptions on the sampling of the data [64].

How large is this performance gap? The collective results of [1, 50, 51] reveal that the performance gap between trend filtering and linear smoothers when estimating spatially heterogeneous signals is significant. For example, when $k = 0$, the minimax rate over L_1 Sobolev classes (which trend filtering achieves) is $n^{-2/3}$, but linear smoothers cannot achieve better than $n^{-1/2}$. To put this in perspective, this result says that the trend filtering estimator, training on n data points, learns these signals with varying smoothness as quickly as a linear smoother training on $n^{4/3}$ data points. As we demonstrate in the next section, this gap in theoretical optimality has clear practical consequences.

In order to minimize the pervasion of technical statistical jargon, henceforth we simply refer to a statistical estimator that achieves the minimax rate over L_2 Sobolev classes as *statistically optimal for estimating spatially homogeneous signals*, and we refer to a statistical estimator that achieves the minimax rate over L_1 Sobolev classes as *statistically optimal for estimating spatially heterogeneous signals*. As previously mentioned, the latter implies the former, but not vice versa.

2.2.2 Empirical comparison

In this section we analyze noisy observations of a simulated spatially heterogeneous signal in order to compare the empirical performance of trend filtering and several classical statistical methods — namely, LOESS, smoothing splines, and Gaussian process regression. The mock

observations are simulated on an unequally-spaced grid $t_1, \ldots, t_n \sim \text{Unif}(0, 1)$ according to the data generating process

$$f(t_i) = f_0(t_i) + \epsilon_i \tag{2.15}$$

with

$$f_0(t_i) = 6 \sum_{k=1}^{3} (t_i - 0.5)^k + 2.5 \sum_{j=1}^{4} (-1)^j \phi_j(t_i), \tag{2.16}$$

where $\phi_j(t)$, $j = 1, \ldots, 4$ are compactly-supported radial basis functions distributed throughout the input space and $\epsilon_i \sim N(0, 0.125^2)$. We therefore construct the signal f_0 to have a smoothly varying global trend with four sharp localized features — two dips and two spikes. The signal and noisy observations are shown in the top panel of Figure 2.1.

In order to facilitate the comparison of methods we utilize a metric for the total statistical complexity (i.e. total wiggliness) of an estimator known as the *effective degrees of freedom* [see, e.g., 74]. Formally, the effective degrees of freedom of an estimator $\widehat{f}_0$ is defined as

$$\text{df}(\widehat{f}_0) = \overline{\sigma}^{-2} \sum_{i=1}^{n} \text{Cov}(\widehat{f}_0(t_i), f(t_i)) \tag{2.17}$$

where $\overline{\sigma}^2$ is defined in equation (2.9). In Figure 2.1 we fix all estimators to have 55 effective degrees of freedom. This exercise provides insight into how each estimator relatively distributes its complexity across the input domain. In the second panel we see that the trend filtering estimate has sufficiently recovered the underlying signal, including both the smoothness of the global trend and the abruptness of the localized features. All three of the linear smoothers, on the other hand, severely oversmooth the localized peaks and dips. Gaussian process regression also exhibits some undesirable oscillatory features that do not correspond to any real trend in the signal. In order to better recover the localized features the linear smoothers require a more complex fit, i.e. smaller LOESS kernel bandwidth, smaller smoothing spline penalization, and

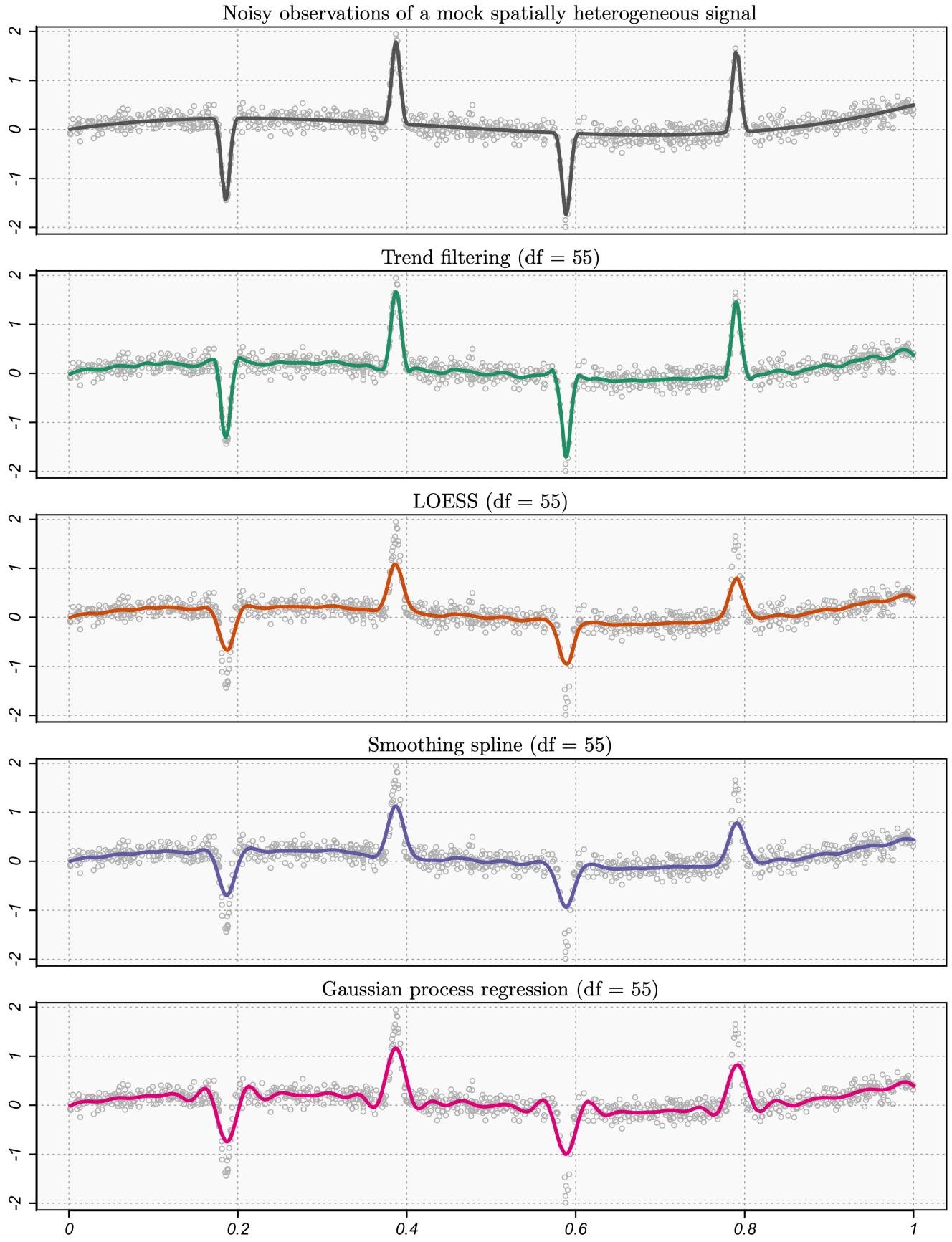


FIGURE 2.1: Comparison of statistical methods on data simulated from a spatially heterogeneous signal. Each statistical estimator is fixed to have 55 effective degrees of freedom in order to facilitate a direct comparison. The trend filtering estimator is able to sufficiently distribute its effective degrees of freedom such that it simultaneously recovers the smoothness of the global trend, as well as the abrupt localized features. The LOESS, smoothing spline, and Gaussian process regression each estimates the smooth global trend reasonably well here, but significantly oversmooths the sharp peaks and dips. Here, we utilize quadratic trend filtering (see Section 2.3.2).

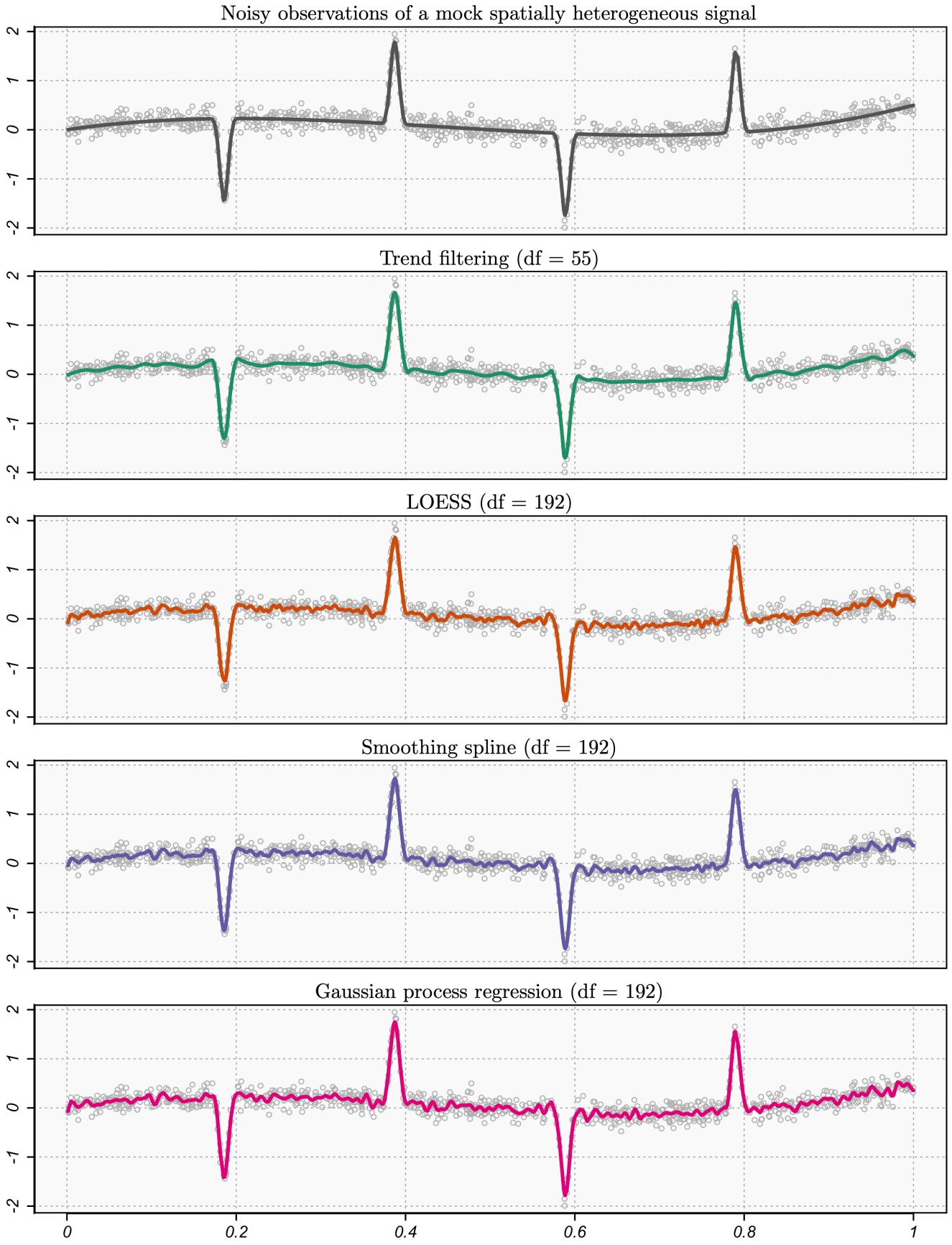


FIGURE 2.2: (Continued): Comparison of statistical methods on data simulated from a spatially heterogeneous signal. Here, each of the linear smoothers (i.e. the LOESS, smoothing spline, and Gaussian process regression) is fixed at 192 effective degrees of freedom — the complexity necessary for each estimator to recover the sharp localized features approximately as well as the trend filtering estimator with 55 effective degrees of freedom. While the linear smoothers now estimate the four abrupt features well, each severely overfits the data in the other regions of the input domain.

smaller Gaussian process noise-signal variance. In Figure 2.2 we show the same comparison, but we grant the linear smoothers more complexity. Specifically, in order to recover the sharp features comparably with the trend filtering estimator with 55 effective degrees of freedom, the linear smoothers require 192 effective degrees of freedom — approximately 3.5 times the complexity. As a result, although they now adequately recover the peaks and dips, each linear smoother severely overfits the data in the other regions of the input domain, resulting in many spurious fluctuations.

As discussed in Section 2.2.1.2, the suboptimality of LOESS, smoothing splines, and Gaussian process regression illustrated in this example is an inherent limitation of the broad *linear smoother* family of statistical estimators. Linear smoothers are adequate tools for estimating signals that exhibit approximately the same degree of smoothness throughout their input domain. However, when a signal is expected to exhibit varying degrees of smoothness across its domain, a locally-adaptive statistical estimator is needed.

2.3 Trend filtering

Trend filtering, in its original form, was independently proposed in the computer vision literature [75] and the applied mathematics literature [76], and has recently been further developed in the statistical and machine learning literature, most notably with [1, 28, 30, 77]. This work is in no way related to the work of [78], which goes by a similar name. At a high level, trend filtering is closely related to two familiar nonparametric regression methods: variable-knot regression splines and smoothing splines. We elaborate on these relationships below.

2.3.1 Closely-related methods

Splines have long played a central role in estimating complex signals [see, e.g., 79, 80, for general references]. Formally, a kth order spline is a piecewise polynomial (i.e. piecewise power law) of degree k that is continuous and has $k-1$ continuous derivatives at the knots. As their names suggest, variable-knot regression splines and smoothing splines center around fitting splines to observational data. Recall from equation (2.1) the observational data generating process (DGP)

$$f(t_i) = f_0(t_i) + \epsilon_i, \qquad t_1, \ldots, t_n \in (a, b), \tag{2.18}$$

where $f(t_i)$ is a noisy measurement of the signal $f_0(t_i)$, and $\mathbb{E}[\epsilon_i] = 0$. Given a set of knots $\kappa_1, \ldots, \kappa_p \in (a, b)$, the space of all kth order splines on the interval (a, b) with knots at $\kappa_1, \ldots, \kappa_p$ can be parametrized via a basis representation

$$m(t) = \sum_j \beta_j \eta_j(t), \tag{2.19}$$

where $\{\eta_j\}$ is typically the truncated power basis or B-spline basis. A suitable estimator for the signal f_0 may then be

$$\widehat{f}_0(t) = \sum_j \widehat{\beta}_j \eta_j(t), \tag{2.20}$$

where the $\widehat{\beta}_j$ are the ordinary least-squares (OLS) estimates of the basis coefficients. This is called a *regression spline*. The question of course remains where to place the knots.

2.3.1.1 Variable-knot regression splines

The variable-knot (or free-knot) regression spline approach is to consider all regression spline estimators with knots at a subset of the observed inputs, i.e. $\{\kappa_1, \ldots, \kappa_p\} \subset \{t_1, \ldots, t_n\}$ for all possible p. Formally, the variable-knot regression spline estimator is the solution to the following

constrained least-squares minimization problem:

$$\begin{aligned} \min_{\{\beta_j\}} \quad & \sum_{i=1}^{n} \left(f(t_i) - \sum_j \beta_j \eta_j(t_i) \right)^2 \\ \text{s.t.} \quad & \sum_{j \geq k+2} \mathbb{1}\{\beta_j \neq 0\} = p \\ & p \geq 0 \end{aligned} \tag{2.21}$$

where $p \geq 0$ is the number of knots in the spline and $\mathbb{1}(\cdot)$ is the indicator function satisfying

$$\mathbb{1}\{\beta_j \neq 0\} = \begin{cases} 1 & \beta_j \neq 0, \\ 0 & \beta_j = 0. \end{cases} \tag{2.22}$$

Furthermore, note that the equality constraint on the basis coefficients excludes those of the "first" $k+1$ basis functions that span the space of global polynomials and only counts the number of active basis functions that produce knots. The variable-knot regression spline optimization is therefore a problem of finding the *best subset* of knots for the regression spline estimator. Due to the sparsity of the coefficient constraint, the variable-knot regression spline estimator allows for highly locally-adaptive behavior for estimating signals that exhibit varying degrees of smoothness. However, the problem itself cannot be solved in polynomial time, requiring an exhaustive combinatorial search over all $\sim 2^n$ feasible models. It is common to utilize stepwise procedures based on iterative addition and deletion of knots in the active set, but these partial searches over the feasible set inherently provide no guarantee of finding the optimal global solution to the problem in equation (2.21).

In order to make the connection to trend filtering more explicit it is helpful to reformulate the constrained minimization problem in equation (2.21) into the following penalized unconstrained

minimization problem:

$$\min_{\{\beta_j\}} \quad \sum_{i=1}^{n} \left(f(t_i) - \sum_j \beta_j \eta_j(t_i) \right)^2 + \gamma \sum_{j \geq k+2} \mathbb{1}(\beta_j \neq 0), \tag{2.23}$$

where $\gamma > 0$ is a hyperparameter that determines the number of knots in the spline and the sum of indicator functions serves as a smoothness "penalty" on the ordinary least-squares minimization. Penalized regression is a popular area of statistical methodology [see, e.g., 49], in which the cost functional (i.e. the quantity to be minimized) quantifies a tradeoff between the training error of the estimator (here, the squared residuals) and the statistical complexity of the estimator (here, the number of knots in the spline). In particular, (2.23) is known as an ℓ_0-penalized least-squares regression because of the penalty's connection to the mathematical ℓ_0 vector quasi-norm.

2.3.1.2 Smoothing splines

Smoothing splines counteract the computational issue faced by variable-knot regression splines by simply placing knots at all of the observed inputs $t_1, \ldots, t_n$ and regularizing the smoothness of the fitted spline. For example, letting $\mathcal{G}$ be the space of all cubic natural splines with knots at $t_1, \ldots, t_n$, the cubic smoothing spline is the solution to the optimization problem

$$\min_{m \in \mathcal{G}} \quad \sum_{i=1}^{n} \big(f(t_i) - m(t_i) \big)^2 + \gamma \int_a^b \big(m''(t) \big)^2 dt, \tag{2.24}$$

where m'' is the second derivative of m and $\gamma > 0$ tunes the amount of regularization. Letting $\eta_1, \ldots, \eta_n$ be a basis for cubic natural splines with knots at the observed inputs, the optimization in equation (2.24) can be equivalently stated as a minimization over the basis coefficients:

$$\min_{\{\beta_j\}} \quad \sum_{i=1}^{n} \left(f(t_i) - \sum_j \beta_j \eta_j(t_i) \right)^2 + \gamma \sum_{j,k=1}^{n} \beta_j \beta_k \omega_{jk} \tag{2.25}$$

where

$$\omega_{jk} = \int_a^b \eta_j''(t)\eta_k''(t)dt. \tag{2.26}$$

The cost functional in equation (2.25) is differentiable and leads to a linear system with a special sparse structure (i.e. bandedness), which yields a solution that can both be found in closed-form and computed very quickly — in $\mathcal{O}(n)$ elementary operations. This particular choice of cost functional, however, produces an estimator that is a linear combination of the observations — a *linear smoother*. Therefore, as discussed and demonstrated in Section 2.2, smoothing splines are suboptimal for estimating spatially heterogeneous signals. Equation (2.25) is known as an ℓ_2-penalized least-squares regression because of the penalty's connection to the mathematical ℓ_2 vector-norm.

2.3.2 Definition

Trend filtering can be viewed as a blending of the strengths of variable-knot regression splines (local adaptivity and interpretability) and the strengths of smoothing splines (simplicity and speed). Mathematically, this is achieved by choosing an appropriate set of basis functions and penalizing the least-squares problem with an ℓ_1 norm on the basis coefficients (sum of absolute values), instead of the ℓ_0 norm of variable-knot regression splines (sum of indicator functions) or the ℓ_2 norm of smoothing splines (sum of squares).

This section is primarily summarized from [1] and [81]. Let the inputs be ordered with respect to the index, i.e. $t_1 < \cdots < t_n$. For the sake of simplicity, we consider the case when the inputs $t_1, \ldots, t_n \in (a, b)$ are equally spaced with $\Delta t = t_{i+1} - t_i$. See the aforementioned papers for the generalized definition of trend filtering to unequally spaced inputs.

For any given integer $k \geq 0$, the kth order trend filtering estimate is a piecewise polynomial of degree k with knots *automatically selected* at a sparse subset of the observed inputs $t_1, \ldots, t_n$.

In Figure 2.3, we provide an example of a trend-filtered data set for orders $k = 0, 1, 2$, and 3. Specifically, the panels of the figure respectively display piecewise constant, piecewise linear, piecewise quadratic, and piecewise cubic fits to the data with the automatically-selected knots indicated by the tick marks on the horizontal axes. Constant trend filtering is equivalent to total variation denoising [82], as well as special forms of the fused lasso of [83] and the variable fusion estimator of [84]. Linear trend filtering was independently proposed by [75] and [76]. Higher-order polynomial trend filtering ($k \geq 2$) was developed by [28] and [1]. In the Figure 2.3 example, the quadratic and cubic trend filtering estimates are nearly visually indistinguishable, and this is true in general. Although, as we see here, trend filtering estimates of different orders typically select different sets of knots.

Like the spline methods discussed in Section 2.3.1, for any order $k \geq 0$, the trend filtering estimator has a basis representation

$$m(t) = \sum_{j=1}^{n} \beta_j h_j(t), \tag{2.27}$$

but, here, the trend filtering basis $\{h_1, \ldots, h_n\}$ is the *falling factorial* basis, which is defined as

$$h_j(t) = \begin{cases} \prod_{i=1}^{j-1}(t - t_i) & j \leq k+1, \\ \prod_{i=1}^{j-1}(t - t_{j-k-1+i}) \cdot \mathbb{1}\{t \geq t_{j-1}\} & j \geq k+2. \end{cases} \tag{2.28}$$

Like the truncated power basis, the first $k+1$ basis functions span the space of global kth order polynomials and the rest of the basis adds the piecewise polynomial structure. However, the knot-producing basis functions of the falling factorial basis h_j, $j \geq k+2$ have small discontinuities in their jth order derivatives at the knots for all $j = 1, \ldots, k-1$, and therefore for orders $k \geq 2$ the trend filtering estimate is *close to*, but not quite a spline. The discontinuities are small enough, however, that the trend filtering estimate defined through the falling factorial basis

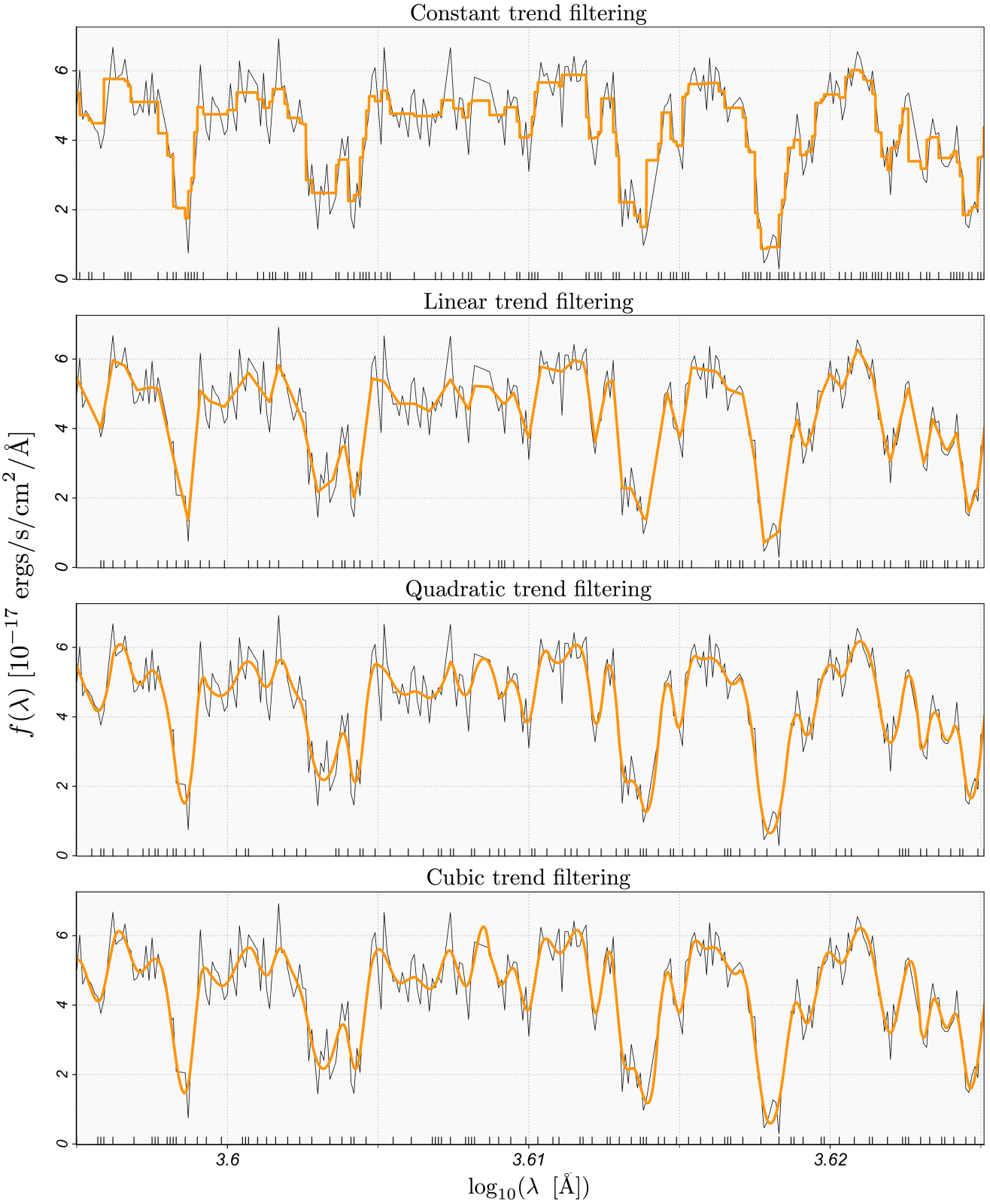


FIGURE 2.3: Piecewise polynomials with adaptively-chosen knots produced by trend filtering. From top to bottom, we show trend filtering estimates of orders $k = 0, 1, 2$ and 3, which take the form of piecewise constant, piecewise linear, piecewise quadratic, and piecewise cubic polynomials, respectively. The adaptively-chosen knots of each piecewise polynomial are indicated by the tick marks along the horizontal axes. The constant trend filtering estimate is discontinuous at the knots, but we interpolate here for visual purposes. The data set is taken from the Lyman-α forest of a mock quasar spectrum [85], sampled in logarithmic-angstrom space. We study this phenomenon in detail in Chapter 3.

representation is visually indistinguishable from the analogous spline produced by the truncated power basis [see 1, 81]. The advantage of utilizing the falling factorial basis in this context instead of the truncated power basis (or the B-spline basis) comes in the form of significant computational speedups, as we detail below.

Analogous to the continuous smoothing spline problem (2.24), we let $\mathcal{H}_k$ be the space of all functions spanned by the kth order falling factorial basis, and pose the trend filtering problem as a least-squares minimization with a derivative-based penalty on the fitted function. In particular, the kth order trend filtering estimator is the solution to the problem

$$\min_{m \in \mathcal{H}_k} \quad \sum_{i=1}^{n} \big(f(t_i) - m(t_i)\big)^2 + \gamma \cdot \mathrm{TV}(m^{(k)}), \tag{2.29}$$

where $m^{(k)}$ is the kth derivative of m, $\mathrm{TV}(m^{(k)})$ is the *total variation* of $m^{(k)}$, and $\gamma > 0$ is the model hyperparameter that controls the smoothness of the fit. When $m^{(k)}$ is differentiable everywhere in its domain, the penalty term simplifies to

$$\mathrm{TV}(m^{(k)}) = \int_a^b |m^{(k+1)}(t)| dt. \tag{2.30}$$

Avoiding the technical generalized definition of total variation [see, e.g., 1], we can simply think of $\mathrm{TV}(\cdot)$ as a generalized L_1 norm[4] for our piecewise polynomials that possess small discontinuities in the derivatives. Again referring back to the smoothing spline problem in equation (2.24), definitions (2.29) and (2.30) reveal that trend filtering can be thought of as an L_1 analog of the (L_2-penalized) smoothing spline problem. Moreover, note that unlike smoothing splines, trend filtering can produce piecewise polynomials of all orders $k \geq 0$.

[4] We use the upper-case notation L_p, $p = 1, 2$ for the p-norm of a continuous function, and ℓ_p, $p = 0, 1, 2$ for the p-norm of a vector.

Replacing m with its basis representation, i.e. $m(t) = \sum_j \beta_j h_j(t)$, yields the equivalent finite-dimensional minimization problem[5]:

$$\min_{\{\beta_j\}} \quad \sum_{i=1}^{n} \left(f(t_i) - \sum_{j=1}^{n} \beta_j h_j(t_i) \right)^2 + \gamma \cdot k! \cdot \Delta t^k \sum_{j=k+2}^{n} |\beta_j|. \tag{2.31}$$

The terms $k!$ and Δt^k are constants and can therefore be ignored by absorbing them into the hyperparameter γ. Visual inspection of equation (2.31) reveals that trend filtering is also analogous to the variable-knot regression spline problem (2.21) — namely, by replacing the ℓ_0 norm on the basis coefficients with an ℓ_1 norm. The advantage here is that the problem is now strictly convex and can be efficiently solved by various convex optimization algorithms. Furthermore, the ℓ_1 penalty still yields a sparse solution (i.e. many $\beta_j = 0$), which provides the automatic knot-selection property. Letting $\widehat{\beta}_1, \ldots, \widehat{\beta}_n$ denote the solution to (2.31) for a particular choice of $\gamma > 0$, the trend filtering estimate is then given by

$$\widehat{f}_0(t; \gamma) = \sum_{j=1}^{n} \widehat{\beta}_j h_j(t), \tag{2.32}$$

with the automatically-selected knots corresponding to the basis functions with $\widehat{\beta}_j \neq 0$, $j \geq k+1$.

The advantage of utilizing the falling factorial basis is found by reparametrizing the problem (2.31) into an optimization over the fitted values $m(t_1), \ldots, m(t_n)$. The problem then reduces to

$$\min_{\{m(t_i)\}} \quad \sum_{i=1}^{n} \big(f(t_i) - m(t_i) \big)^2 + \gamma \sum_{i=1}^{n-k-1} |\Delta^{(k+1)} m(t_i)| \cdot \Delta t \tag{2.33}$$

where $\Delta^{(k+1)} m(t_i)$ can be viewed as a discrete approximation of the $(k+1)$st derivative of m at t_i. For $k = 0$ the discrete derivatives are

$$\Delta^{(1)} m(t_i) = \frac{m(t_{i+1}) - m(t_i)}{\Delta t}, \tag{2.34}$$

[5] This may be recognized as a lasso regression [86], with the features being the falling factorial basis functions.

and then can be defined recursively for $k \geq 1$:

$$\Delta^{(k+1)} m(t_i) = \frac{\Delta^{(k)} m(t_{i+1}) - \Delta^{(k)} m(t_i)}{\Delta t}. \tag{2.35}$$

The penalty term in equation (2.33) can be viewed as a Riemann-like discrete approximation of the integral in equation (2.30). Because of the choice of basis, the problem has reduced to a simple generalized lasso problem [28, 87] with an identity predictor matrix and a banded[6] penalty matrix. This special structure allows the solution to be found in nearly linear time, i.e. $\mathcal{O}(n)$ elementary operations. In this work we utilize the specialized alternating direction method of multipliers (ADMM) algorithm of [30]. This algorithm has a linear complexity per iteration, so the overall complexity is $\mathcal{O}(nr)$ where r is the number of iterations necessary to converge to the solution. In the worst case scenario $r \sim n^{1/2}$, so the worst-case overall complexity is $\mathcal{O}(n^{1.5})$. The practical computational speed further illustrates the value of trend filtering to astronomy, as it is readily compatible with the large-scale analysis of one-dimensional data sets that has become increasingly ubiquitous in large sky surveys. We show a comparison in Table 2.1 of the computational costs associated with trend filtering and other popular one-dimensional nonparametric methods.

Given the trend filtering fitted values obtained by the optimization in equation (2.33) the full continuous-time representation of the trend filtering estimate follows by inverting the parametrization back to the basis function coefficients and plugging them into the basis representation in equation (2.32).

[6] A banded matrix only contains nonzero elements in the main diagonal and zero or more diagonals on either side.

	Method	Computational Complexity	Hyperparameters to estimate
Locally-adaptive	Wavelets	$\mathcal{O}(n)$	1
	Trend filtering	$\boldsymbol{\mathcal{O}(n^{1.5})}$	**1**
	Variable-knot regression splines	$\mathcal{O}(n \cdot \binom{n}{p})$	1
Non-adaptive	Uniform-knot regression splines	$\mathcal{O}(n)$	1
	Smoothing splines	$\mathcal{O}(n)$	1
	Kernel smoothers	$\mathcal{O}(n^2)$	1
	LOESS	$\mathcal{O}(n^2)$	1
	Gaussian process regression	$\mathcal{O}(n^3)$	3+

TABLE 2.1: Comparison of computational costs associated with popular one-dimensional nonparametric regression methods. The computational complexity column states the dependence on the sample size n of the number of elementary operations necessary to obtain the fitted values of each estimator (i.e. the estimator evaluated at the observed inputs). For trend filtering, the $\mathcal{O}(n^{1.5})$ complexity represents the worst-case complexity of the [30] convex optimization algorithm. In most practical settings the actual complexity of this algorithm is close to $\mathcal{O}(n)$. Variable-knot regression splines require a (nonconvex) exhaustive combinatorial search over the set of possible knots and the complexity therefore includes a binomial coefficient term $\binom{n}{p} = n!/(n!(n-p)!)$, where p is the number of knots in the spline. The remaining methods are explicitly solvable and the stated complexity represents the cost of an exact calculation. The $\mathcal{O}(n)$ complexity of wavelets relies on restrictive sampling assumptions (e.g., equally-spaced inputs, sample size equal to a power of two). The stated computational complexity of all methods represents the cost of a single model fit and does not include the cost of hyperparameter tuning. Gaussian process regression suffers from the most additional overhead in this regard because of the (often) large number of hyperparameters used to parametrize the covariance function (e.g., shape, range, marginal variance, noise variance). Each of the non-adaptive methods (linear smoothers) can be made to be locally adaptive (e.g., by locally varying the hyperparameters of the model), but at the expense of greatly increasing the dimensionality of the hyperparameter space to be searched over.

2.3.3 Extension to heteroskedastic weighting

Thus far we have considered the simple case where the observations are treated as equally-weighted in the cost functional (2.33). Recall from equation (2.18) the observational data generating process and define $\sigma_i^2 = \text{Var}(\epsilon_i)$ to be the noise level — the (typically heteroskedastic) uncertainty in the measurements that arises from instrumental errors and removal of systematic effects. When estimates for σ_i^2, $i = 1, \ldots, n$ accompany the observations, as they often do, they can be used to weight the observations to yield a more efficient statistical estimator (i.e. smaller mean-squared error). The error-weighted trend filtering estimator is the solution to the following minimization

Language	Recommended implementation
R	`github.com/glmgen`
C	`github.com/glmgen`
Python	`cvxpy.org`
Matlab	`http://stanford.edu/~boyd/l1_tf`
Julia	`github.com/JuliaStats/Lasso.jl`

TABLE 2.2: Recommended implementations for trend filtering in various programming languages. See Section 2.3.4 for details. We provide supplementary R code at `github.com/capolitsch/trendfilteringSupp` for selecting the hyperparameter via minimization of Stein's unbiased risk estimate (see Section 2.3.5) and various bootstrap methods for uncertainty quantification (see Section 2.3.6). Our implementations are built on top of the `glmgen` R package of [88].

problem:

$$\min_{\{m(t_i)\}} \quad \sum_{i=1}^{n} \big(f(t_i) - m(t_i)\big)^2 w_i + \gamma \sum_{i=1}^{n-k-1} |\Delta^{(k+1)} m(t_i)| \cdot \Delta t, \tag{2.36}$$

where the optimal choice of weights is $w_i = \sigma_i^{-2}$, $i = 1, \ldots, n$. Much of the publically available software for trend filtering allows for a heteroskedastic weighting scheme (see Section 2.3.4).

2.3.4 Software

Trend filtering software is available online across various platforms. For the specialized ADMM algorithm of [30] that we utilize in this work, implementations are available in R and C [88], as well as Julia [89]. Matlab and Python implementations are available for the primal-dual interior point method of [76], but only for equally-weighted linear trend filtering [90, 91]. We provide links to our recommended implementations in Table 2.2. Note that in all software implementations the trend filtering hyperparameter is called λ instead of γ, which we use here to avoid ambiguity with the notation for wavelength in our spectroscopic analyses throughout this thesis.

2.3.5 Choosing the hyperparameter

The choice of the piecewise polynomial order k generally has minimal effect on the performance of the trend filtering estimator in terms of mean-squared error and therefore can be treated as an *a priori* aesthetic choice based on how much smoothness is desired or believed to be present. For example, we use $k = 2$ (quadratic trend filtering) throughout our analyses in Chapter 3 so that the fitted curves are smooth, i.e. differentiable everywhere.

Given the choice of k, the hyperparameter $\gamma > 0$ is used to tune the complexity (i.e. the wiggliness) of the trend filtering estimate by weighting the tradeoff between the complexity of the estimate and the size of the squared residuals. Obtaining an accurate estimate is therefore intrinsically tied to finding an optimal choice of γ. The selection of γ is typically done by minimizing an estimate of the mean-squared prediction error (MSPE) of the trend filtering estimator. Here, there are two different notions of error to consider, namely, *fixed-input* error and *random-input* error. As the names suggest, the distinction between which type of error to consider is made based on how the inputs are sampled. As a general rule-of-thumb, we recommend optimizing with respect to fixed-input error when the inputs are regularly-sampled and optimizing with respect to random-input error on irregularly-sampled data.

Recall the DGP stated in equation (2.18) and let it be denoted by Q so that $\mathbb{E}_Q[\cdot]$ is the mathematical expectation with respect to the randomness of the DGP. Further, let $\sigma_i^2 = \mathrm{Var}(\epsilon_i)$. The fixed-input MSPE is given by

$$R(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_Q\Big[\big(f(t_i) - \widehat{f}_0(t_i;\gamma)\big)^2 \;\Big|\; t_1,\ldots,t_n\Big] \tag{2.37}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \Big(\mathbb{E}_Q\Big[\big(f_0(t_i) - \widehat{f}_0(t_i;\gamma)\big)^2 \;\Big|\; t_1,\ldots,t_n\Big] + \sigma_i^2\Big) \tag{2.38}$$

and the random-input MSPE is given by

$$\widetilde{R}(\gamma) = \mathbb{E}_Q\Big[\big(f(t) - \widehat{f}_0(t;\gamma)\big)^2\Big], \tag{2.39}$$

where, in the latter, t is considered to be a random component of the DGP with a marginal probability density $p_t(t)$ supported on the observed input interval. In each case, the theoretically optimal choice of γ is defined as the minimizer of the respective choice of error. Empirically, we estimate the theoretically optimal choice of γ by minimizing an estimate of (2.37) or (2.39). For fixed-input error we recommend Stein's unbiased risk estimate [SURE; 92, 93] and for random-input error we recommend K-fold cross validation with $K = 10$. We elaborate on SURE here and refer the reader to [94] for K-fold cross validation.

The SURE formula provides an unbiased estimate of the fixed-input MSPE of a statistical estimator:

$$\widehat{R}_0(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\big(f(t_i) - \widehat{f}_0(t_i;\gamma)\big)^2 + \frac{2\overline{\sigma}^2 \mathrm{df}(\widehat{f}_0)}{n}, \tag{2.40}$$

where $\overline{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\sigma_i^2$ and $\mathrm{df}(\widehat{f}_0)$ is defined in equation (2.17). A formula for the effective degrees of freedom of the trend filtering estimator is available via the generalized lasso results of [95]; namely,

$$\mathrm{df}(\widehat{f}_0) = \mathbb{E}[\text{number of knots in } \widehat{f}_0] + k + 1. \tag{2.41}$$

We then obtain our hyperparameter estimate $\widehat{\gamma}$ by minimizing the following plug-in estimate for (2.40):

$$\widehat{R}(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\big(f(t_i) - \widehat{f}_0(t_i;\gamma)\big)^2 + \frac{2\widehat{\overline{\sigma}}^2 \widehat{\mathrm{df}}(\widehat{f}_0)}{n}, \tag{2.42}$$

where $\widehat{\mathrm{df}}$ is the estimate for the effective degrees of freedom that is obtained by replacing the

expectation in equation (2.41) with the observed number of knots, and $\widehat{\overline{\sigma}}^2$ is an estimate of $\overline{\sigma}^2$. If a reliable estimate of $\overline{\sigma}^2$ is not available *a priori*, a data-driven estimate can be constructed [see, e.g., 48]. We provide a supplementary R package on the corresponding author's GitHub page[7] for implementing SURE with trend filtering. The package is built on top of the `glmgen` R package of [88], which already includes an implementation of K-fold cross validation.

Because of the existence of the degrees of freedom expression (2.41), trend filtering is also compatible with reduced chi-squared model assessment and comparison procedures under a Gaussian noise assumption [96, 97].

2.3.6 Uncertainty quantification

2.3.6.1 Frequentist

Frequentist uncertainty quantification for trend filtering follows by studying the sampling distribution of the estimator that arises from the randomness of the observational data generating process (DGP). In particular, most of the uncertainty in the estimates is captured by studying the variability of the estimator with respect to the DGP. We advise three different bootstrap methods [98] for estimating the variability of the trend filtering estimator, with each method corresponding to a distinct analysis setting. Here, we emphasize the terminology *variability* — as opposed to the variance of the trend filtering estimator — since, by construction, as a nonlinear function of the observed data, the trend filtering estimator has a non-Gaussian sampling distribution even when the observational noise is Gaussian. For that reason, each of our recommended bootstrap approaches is based on computing sample quantiles (instead of pairing standard errors with Gaussian quantiles).

[7] `https://github.com/capolitsch/trendfilteringSupp`

We restate the assumed DGP here for clarity:

$$f(t_i) = f_0(t_i) + \epsilon_i, \qquad t_1, \ldots, t_n \in (a, b) \tag{2.43}$$

where $\mathbb{E}[\epsilon_i] = 0$. We make the further assumption that the errors $\epsilon_1, \ldots, \epsilon_n$ are independent[8]. The three distinct settings we consider are:

S1. The inputs are irregularly sampled

S2. The inputs are regularly sampled and the noise distribution is known

S3. The inputs are regularly sampled and the noise distribution is unknown

The corresponding bootstrap methods are detailed in Algorithm 1 [nonparametric bootstrap; 98], Algorithm 2 [parametric bootstrap; 100], and Algorithm 3 [wild bootstrap; 101–103], respectively. We include implementations of each of these algorithms in the R package on our GitHub page.

Algorithm 1 Nonparametric bootstrap for random-input uncertainty quantification

Require: Training Data $(t_1, f(t_1)), \ldots, (t_n, f(t_n))$, hyperparameters γ and k, prediction input grid $t'_1, \ldots, t'_m$

1: **for all** b in $1 : B$ **do**

2: Define a bootstrap sample of size n by resampling the observed pairs with replacement:

$$(t_1^*, f_b^*(t_1^*)), \ldots, (t_n^*, f_b^*(t_n^*))$$

3: Let $\widehat{f}_b^*(t'_1), \ldots, \widehat{f}_b^*(t'_m)$ denote the trend filtering estimate fit on the bootstrap sample and evaluated on the prediction grid $t'_1, \ldots, t'_m$

4: **end for**

Output: The full trend filtering bootstrap ensemble $\{\widehat{f}_b^*(t'_i)\}_{\substack{i=1,\ldots,m \\ b=1,\ldots,B}}$

Given the full trend filtering bootstrap ensemble provided by the relevant bootstrap algorithm, for any $\alpha \in (0, 1)$, a $(1 - \alpha) \cdot 100\%$ quantile-based pointwise variability band is given by

$$V_{1-\alpha}(t'_i) = \left(\widehat{f}^*_{\alpha/2}(t'_i),\ \widehat{f}^*_{1-\alpha/2}(t'_i)\right), \qquad i = 1, \ldots, m \tag{2.44}$$

[8] If nontrivial autocorrelation exists in the noise then a block bootstrap [99] will yield a better approximation of the trend filtering variability than the bootstrap implementations we discuss.

Algorithm 2 Parametric bootstrap for fixed-input uncertainty quantification (when noise distribution $\epsilon_i \sim Q_i$ is known *a priori*)

Require: Training Data $(t_1, f(t_1)), \ldots, (t_n, f(t_n))$, hyperparameters γ and k, assumed noise distribution $\epsilon_i \sim Q_i$, prediction input grid $t'_1, \ldots, t'_m$

1: Compute the trend filtering point estimate at the observed inputs:

$$(t_1, \widehat{f}_0(t_1)), \ldots, (t_n, \widehat{f}_0(t_n))$$

2: **for all** b in $1 : B$ **do**

3: Define a bootstrap sample by sampling from the assumed noise distribution:

$$f_b^*(t_i) = \widehat{f}_0(t_i) + \epsilon_i^* \qquad \text{where } \epsilon_i^* \sim Q_i, \quad i = 1, \ldots, n$$

4: Let $f_b^*(t'_1), \ldots, f_b^*(t'_m)$ denote the trend filtering estimate fit on the bootstrap sample and evaluated on the prediction grid $t'_1, \ldots, t'_m$

5: **end for**

Output: The full trend filtering bootstrap ensemble $\{\widehat{f}_b^*(t'_i)\}_{\substack{i=1,\ldots,m \\ b=1,\ldots,B}}$

Algorithm 3 Wild bootstrap for fixed-input uncertainty quantification (when noise distribution is not known *a priori*)

Require: Training Data $(t_1, f(t_1)), \ldots, (t_n, f(t_n))$, hyperparameters γ and k, prediction input grid $t'_1, \ldots, t'_m$

1: Compute the trend filtering point estimate at the observed inputs:

$$(t_1, \widehat{f}_0(t_1)), \ldots, (t_n, \widehat{f}_0(t_n))$$

2: Let $\widehat{\epsilon}_i = f(t_i) - \widehat{f}_0(t_i)$, $i = 1, \ldots, n$ denote the residuals

3: **for all** i **do**

4: Define a bootstrap sample by sampling from the following distribution:

$$f_b^*(t_i) = \widehat{f}_0(t_i) + u_i^* \qquad i = 1, \ldots, n$$

where

$$u_i^* = \begin{cases} \widehat{\epsilon}_i(1+\sqrt{5})/2 & \text{with probability } (1+\sqrt{5})/(2\sqrt{5}) \\ \widehat{\epsilon}_i(1-\sqrt{5})/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}) \end{cases}$$

5: Let $f_b^*(t'_1), \ldots, f_b^*(t'_m)$ denote the trend filtering estimate fit on the bootstrap sample and evaluated on the prediction grid $t'_1, \ldots, t'_m$

6: **end for**

Output: The full trend filtering bootstrap ensemble $\{\widehat{f}_b^*(t'_i)\}_{\substack{i=1,\ldots,m \\ b=1,\ldots,B}}$

where

$$\widehat{f}^*_\beta(t'_i) = \inf_g \left\{ g : \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{\widehat{f}^*_b(t'_i) \leq g\} \geq \beta \right\}, \qquad \beta \in (0,1). \tag{2.45}$$

Analogously, bootstrap sampling distributions and variability intervals for observable parameters of the signal may be studied by deriving a bootstrap parameter estimate from each trend filtering estimate within the bootstrap ensemble. For example, in Chapter 3 we examine the bootstrap sampling distributions of several observable light-curve parameters of exoplanet transits and supernovae.

2.3.6.2 Bayesian

There is a well-studied connection between ℓ_1-penalized least-squares regression and a Bayesian framework [see, e.g., 86, 104, 105]. A discussion specific to trend filtering can be found in [106].

2.3.7 Relaxed trend filtering

We are indebted to Ryan Tibshirani for a private conversation that motivated the discussion in this section. Trend filtering can be generalized to allow for greater flexibility through a technique that we call *relaxed trend filtering*[9]. Although the traditional trend filtering estimator is already highly flexible, there are certain settings in which the relaxed trend filtering estimator provides nontrivial improvements. In our experience, these typically correspond to settings where the optimally-tuned trend filtering estimator selects very few knots. For example, we use relaxed trend filtering in Chapter 3 to model the detrended, phase-folded light curve of a *Kepler* star with a planetary transit event [5, 6, 108].

The relaxed trend filtering estimate is defined through a two-stage sequential procedure in which the first stage amounts to computing the traditional trend filtering estimate discussed in Section

[9] We choose this term because the generalization of trend filtering to relaxed trend filtering is analogous to the generalization of the lasso [86] to the relaxed lasso [107].

2.3.2. Recall the trend filtering minimization problem in equation (2.31). For any given order $k \in \{0, 1, 2, \dots\}$ and hyperparameter $\gamma > 0$, let us amend our notation so that

$$\widehat{f}_0^{TF}(t) = \sum_{j=1}^{n} \widehat{\beta}_j^{TF} h_j(t) \tag{2.46}$$

denotes the basis representation of the traditional trend filtering estimate. Further, define the index set

$$\mathcal{K}_\gamma = \left\{ 1 \leq j \leq n \mid \widehat{\beta}_j^{TF} \neq 0 \right\} \tag{2.47}$$

that includes the indices of the non-zero falling factorial basis coefficients for the given choice of γ. Now let $\widehat{\beta}_j^{OLS}$, $j \in \mathcal{K}_\gamma$, denote the solution to the ordinary least-squares (OLS) minimization problem

$$\min_{\{\beta_j\}} \quad \sum_{i=1}^{n} \left(f(t_i) - \sum_{j \in \mathcal{K}_\gamma} \beta_j h_j(t_i) \right)^2, \tag{2.48}$$

and define the corresponding OLS estimate as

$$\widehat{f}_0^{OLS}(t) = \sum_{j \in \mathcal{K}_\gamma} \widehat{\beta}_j^{OLS} h_j(t). \tag{2.49}$$

That is, the OLS estimate (2.49) uses trend filtering to find the knots in the piecewise polynomial, but then uses ordinary least-squares to estimate the reduced set of basis coefficients. The relaxed trend filtering estimate is then defined as a weighted average of the traditional trend filtering estimate and the corresponding OLS estimate:

$$\widehat{f}_0^{RTF}(t) = \phi \widehat{f}_0^{TF}(t) + (1 - \phi) \widehat{f}_0^{OLS}(t), \tag{2.50}$$

for some choice of relaxation hyperparameter $\phi \in [0, 1]$. Relaxed trend filtering is therefore a generalization of trend filtering in the sense that the case $\phi = 1$ returns the traditional trend

filtering estimate.

In principle, it is preferable to jointly optimize the trend filtering hyperparameter γ and the relaxation hyperparameter ϕ, e.g. via cross validation. However, it often suffices to choose γ and ϕ sequentially, which in turn adds minimal computational cost on top of the traditional trend filtering procedure. Because of the trivial proximity of the falling factorial basis to the truncated power basis [established in 1, 81], it is sufficient to let $\widehat{f}_0^{OLS}$ be the kth order regression spline with knots at the input locations selected by the trend filtering estimator. In heteroskedastic settings, as discussed in Section 2.3.3, a piecewise polynomial or regression spline fit by weighted least-squares should be used in place of the OLS estimate (2.49).

2.4 Concluding remarks

The analysis of one-dimensional data arising from signals possessing varying degrees of smoothness is central to a wide variety of problems in time-domain astronomy and astronomical spectroscopy. Trend filtering is a modern statistical tool that provides a unique combination of (1) statistical optimality for estimating spatially heterogeneous signals; (2) natural flexibility for handling practical analysis settings (general sampling designs, heteroskedastic noise distributions, etc.); (3) practical computational speed that scales to massive data sets; and (4) a single model hyperparameter that can be chosen via automatic data-driven methods.

Software for trend filtering is freely available online across various platforms and we provide links to our recommendations in Table 2.3.4. Additionally, we make supplementary R code available on the corresponding author's GitHub page[10] for: (1) selecting the trend filtering hyperparameter by minimizing Stein's unbiased risk estimate (see Section 2.3.5); and (2) various bootstrap methods for trend filtering uncertainty quantification (see Section 2.3.6).

[10] `https://github.com/capolitsch/trendfilteringSupp`

In Chapter 3 we explicitly demonstrate the broad utility of trend filtering to astronomy by carrying out a diverse set of spectroscopic and time-domain analyses.

Chapter 3

Trend Filtering: Denoising Astronomical Signals with Varying Degrees of Smoothness

This chapter is based on our paper *Trend Filtering – II: Denoising Astronomical Signals with Varying Degrees of Smoothness* [31], which was published in *Monthly Notices of the Royal Astronomical Society*.

Trend filtering — first introduced into the astronomical literature in our companion paper [29] (Chapter 2) — is a state-of-the-art statistical tool for denoising one-dimensional signals that possess varying degrees of smoothness. In this chapter, we demonstrate the broad utility of trend filtering to observational astronomy by discussing how it can contribute to a variety of spectroscopic and time-domain studies. The observations we discuss are (1) the Lyman-α forest of quasar spectra; (2) more general spectroscopy of quasars, galaxies, and stars; (3) stellar light curves with planetary transits; (4) eclipsing binary light curves; and (5) supernova light curves. We study the Lyman-α forest in the greatest detail — using trend filtering to map the large-scale structure of the intergalactic medium along quasar-observer lines of sight. The remaining studies share broad themes of: (1) estimating observable parameters of light curves and spectra; and (2)

constructing observational spectral/light-curve templates. We also briefly discuss the utility of trend filtering as a tool for one-dimensional data reduction and compression.

3.1 Introduction

Many astronomical analyses can be described by the following problem setup. Suppose we collect noisy measurements of an observable quantity (e.g., flux, magnitude, photon counts) according to the data generating process

$$f(t_i) = f_0(t_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3.1}$$

where $t_1, \ldots, t_n$ is an arbitrarily-spaced grid of one-dimensional inputs (e.g., times or wavelengths), ϵ_i is mean zero noise, and f_0 is a signal that may exhibit varying degrees of smoothness across the input domain (e.g., a smooth signal with abrupt dips/spikes). Given the observed data sample, we then attempt to estimate (or denoise) the underlying signal f_0 from the observations by applying appropriate statistical methods. In [29] (Chapter 2), we introduced trend filtering [1, 28] into the astronomical literature. When the underlying signal is spatially heterogeneous, i.e. possesses varying degrees of smoothness, trend filtering is superior to popular statistical methods such as Gaussian process regression, smoothing splines, kernels, LOESS, and many others [1, 50, 51]. Furthermore, the trend filtering estimate can be computed via a highly efficient and scalable convex optimization algorithm [30] and only requires data-driven selection of a single scalar hyperparameter. In this chapter, we directly demonstrate the broad utility of trend filtering to observational astronomy by using it to carry out a diverse set of spectroscopic and time-domain analyses.

The outline of this chapter is as follows. In Section 3.2, we use trend filtering to study the Lyman-α forest of quasar spectra — a series of absorption features that can be used as a tracer of the matter density distribution of the intergalactic medium along quasar-observer lines of sight. We choose to study this application in depth and then illustrate the breadth of trend filtering's utility through our discussions in Section 3.3. The applications we discuss in Section 3.3 can be grouped into two broad (and often intertwined) categories: (1) deriving estimates of observable parameters from trend-filtered observations; and (2) using trend filtering to construct spectral/light-curve templates of astronomical objects/events. In Section 3.3.1, we discuss constructing spectral template libraries for astronomical objects by trend filtering coadded spectroscopic observations. We illustrate our approach on quasar, galaxy, and stellar spectra from the Sloan Digital Sky Survey [27, 109]. Emission-line parameters can also be robustly estimated by fitting radial basis functions (e.g., Gaussians) to trend-filtered estimates near emission lines. In Section 3.3.2, we use relaxed trend filtering to model the detrended, phase-folded light curve of a *Kepler* stellar system with a transiting exoplanet. We derive estimates and full uncertainty distributions for the transit depth and total transit duration. In Section 3.3.3, we use trend filtering to denoise a detrended, phase-folded *Kepler* light curve of an eclipsing binary (EB) system. We illustrate that trend filtering provides significant improvements upon the popular `polyfit` method of [8] that is used to model *Kepler* EB light curves and derive observable parameters. In Section 3.3.4, we discuss using trend filtering to construct light-curve templates of supernova (SN) explosions. We illustrate this approach on a SN light curve obtained from the Open Supernova Catalog [110]. Furthermore, we derive estimates and full uncertainty distributions for a set of observable parameters — namely, the maximum apparent magnitude, the time of maximum, and the decline rate. Finally, in Section 3.3.5, we briefly discuss a different, non-data-analysis application of trend filtering. Specifically, we discuss the use of trend filtering as a tool for fast and flexible one-dimensional data reduction and compression. The flexibility of the trend filtering estimator,

paired with its efficient speed and storage capabilities, make it a potentially powerful tool to include in large-scale (one-dimensional) astronomical data reduction and storage pipelines.

We utilize the [88] `glmgen` R implementation of the [30] trend filtering optimization algorithm in this work. See Chapter 2 for implementations in other programming languages.

3.2 Main Application: Quasar Lyman-α Forest

The Lyman-α (Lyα) forest is the name given to the absorption features seen in quasar spectra which are caused by neutral hydrogen (H I) in the intergalactic medium between a quasar and an observer. When emitted from an accretion disk close to the central black hole, the light from the quasar has a relatively smooth spectrum — a continuum — caused by the summed black-body emission of gas with different temperatures at different disk radii [111]. Emission lines are also seen, and their intensities and line ratios supply information on the physical conditions in the line emitting gas. At least twenty broad emission lines, broadened by high velocities and temperatures can be measured in a single active galactic nucleus, along with a similar number of narrow lines from colder gas [112]. The emitted spectrum therefore already consists of a superposition of components with varying degrees of smoothness. The Lyα forest arises when this spectrum is further processed with the addition of absorption lines. Light moving towards the observer is redshifted into resonance with the Lyα transition of H I, and the strength of absorption features is dictated by the densities of intergalactic material along the line of sight [113]. The smoothness of the absorption lines varies depending on the gas pressure, and thermal doppler broadening [114, 115]. Sharper absorption features, metal lines, are also caused by other intergalactic species, such as C IV, O VI , and Mg II [116, 117]. The usefulness of the Lyα forest as a cosmological probe [e.g., 118] stems from its relationship to the matter density field in the Universe, effectively mapping out structure along each quasar-observer line of sight [e.g., 23, 119]. In order to extract

this information from noisy spectra and separate it from other components, it is useful to have a method that can deal with the complexities outlined above, i.e. one that can naturally adapt to varying degrees of smoothness without extensive tuning.

The relative fluctuations in the Lyα forest transmitted flux fraction are of primary interest since they possess a monotonic relationship with the relative distribution of the absorbing H I. We utilize trend filtering to first denoise the spatially heterogeneous flux signal in an observed Lyα forest. Estimates for the fluctuations in transmitted flux due to absorbing H I are then typically produced by coupling the denoised Lyα forest with estimates for the quasar continuum and the cosmic mean transmitted flux in the Lyα forest. We take an alternative approach: directly estimating the mean flux level — defined as the product of the continuum and cosmic mean transmitted flux — as in [35]. The mean flux level is a very smooth, spatially homogeneous function within the truncated Lyα forest restframe. It is therefore appropriate to use a linear smoother (see Section 2.2.1) for this stage of estimation. Specifically, we use local polynomial regression [LOESS 120–122]. In this section, we illustrate these methods on a mock quasar Lyα forest from [85] and a real quasar Lyα forest from the Baryon Oscillation Spectroscopic Survey Data Release 12 [BOSS DR12; 27] of the Sloan Digital Sky Survey III [SDSS-III; 26, 123].

Historically, Lyα forest analyses have typically utilized kernel smoothers [e.g., 35, 124], wavelets [e.g., 45], or Gaussian processes [e.g., 125].

3.2.1 Notation

Suppose we observe a quasar located at redshift $z = z_0$. Ignoring systematic effects such as sky contamination and interstellar extinction for the moment, the observational DGP of the Lyα

forest can be assumed to follow the model

$$f(\lambda) = f_0(\lambda) + \epsilon(\lambda), \qquad \lambda \in \Lambda(z_0), \tag{3.2}$$

$$= \overline{F}(\lambda) \cdot C(\lambda) \cdot \big(1 + \delta_F(\lambda)\big) + \epsilon(\lambda), \tag{3.3}$$

where $f(\lambda)$ is the observed flux at wavelength λ, $f_0(\lambda)$ is the flux signal, $\epsilon(\lambda)$ is zero mean white Gaussian noise, $\Lambda(z_0) = (\lambda_{\text{Ly}\beta},\ \lambda_{\text{Ly}\alpha}) \cdot (1 + z_0)$ is the redshifted Lyα forest, $C(\lambda)$ is the flux of the unabsorbed quasar continuum, $F(\lambda) = f_0(\lambda)/C(\lambda)$ is the transmitted flux fraction, $\overline{F}(\lambda) = \mathbb{E}[F(\lambda)]$ is the mean transmitted flux fraction (over the sky) in the Lyα forest at redshift $z = \lambda/\lambda_{\text{Ly}\alpha} - 1$, and

$$\delta_F(\lambda) = F(\lambda)/\overline{F}(\lambda) - 1 \tag{3.4}$$

is the fluctuation about the mean Lyα transmitted flux at redshift $z = \lambda/\lambda_{\text{Ly}\alpha} - 1$. Here, δ_F is the quantity we are primarily interested in estimating since $\delta_F \propto \delta_{\text{H\,I}}^{-1}$ at each fixed redshift, where $\delta_{\text{H\,I}}$ is the density of H I (a latent variable). The estimation of the flux signal f_0 is viewed as an ancillary step.

Although, in principle, it is preferable to study the full spectral range $\Lambda(z_0)$ we have found that, in the nonparametric setting, estimating the quasar continuum near the localized Lyα and Lyβ emission peaks at the boundaries of the Lyα forest reduces the estimation accuracy in the interior of $\Lambda(z_0)$. Therefore, in this work we limit our analysis to the truncated Lyα forest range

$$\overline{\Lambda}(z_0) = (1045 \text{ Å}\ , 1195 \text{ Å}) \cdot (1 + z_0). \tag{3.5}$$

We simplify notation in this work by changing the input space of the functions introduced above by merely altering the input variable. For example, with respect to δ_F, we maintain the notation $\delta_F(\cdot)$ for all inputs λ, ν, z, ζ, while it is understood that a proper change of input spaces has

Input	Definition	Range (quasar at $z = z_0$)
λ	Observed wavelength	$\overline{\Lambda}(z_0)$
ν	Rest wavelength	$\overline{\Lambda}_{\text{rest}}(z_0) = \overline{\Lambda}(z_0)/(1+z_0)$
z	Redshift	$\Pi(z_0) = \overline{\Lambda}(z_0)/\lambda_{\text{Ly}\alpha} - 1$
ζ	Log-wavelength (scaled)	$Z(z_0) = 10^4 \cdot \log_{10}(\overline{\Lambda}(z_0))$

TABLE 3.1: Various input spaces utilized for the Lyα forest analysis. Notation of functions is held constant, e.g. $\delta_F(\cdot)$, and an alteration of the input variable implicitly indicates a change of input spaces. Logarithmic wavelengths are scaled for numerical stability of the trend filtering optimization algorithm.

taken place. The various input spaces are defined in Table 3.1.

3.2.2 Trend filtering the observed flux

We use quadratic trend filtering [1, 28] to estimate the flux signal f_0 of the observational model (5.1). In both BOSS DR12 and the [85] mock catalog, the quasar spectra are sampled on equally-spaced grids in logarithmic wavelength space with $\Delta \log_{10}(\lambda_i) = 10^{-4}$ dex (in logarithmic angstroms). Furthermore, flux measurement variances are provided by the BOSS pipeline [66], accounting for the statistical uncertainty introduced by photon noise, CCD read noise, and sky-subtraction error. We correct the BOSS spectrum for interstellar extinction with the [126] extinction law and the [127] dust map.

We fit the trend filtering estimator on the equally-spaced logarithmic grid and tune the complexity by minimizing Stein's unbiased risk estimate (SURE) of the fixed-input mean-squared error (see Chapter 2). More precisely, we fit the trend filtering estimator in the input space $Z(z_0) = 10^4 \cdot \log_{10}(\overline{\Lambda}(z_0))$, as defined in Table 3.1, where we add the scaling to unit spacing for numerical stability of the trend filtering convex optimization.

3.2.3 Nonparametric continuum estimation

We utilize a modified [35] approach to propagate the trend filtering estimate for the flux signal f_0 from Section 3.2.2 into an estimate for the fluctuation field δ_F along a line of sight to an observed quasar. Namely, given the trend filtering estimate $\widehat{f}_0$, we directly estimate the smooth mean flux level defined as the product $m = \overline{F} \cdot C$ and then define the δ_F estimates via the transformation $\widehat{\delta}_F := \widehat{f}_0/\widehat{m} - 1$. We carry out the estimation of m via a wide-kernel LOESS smooth of the trend filtering estimate, with the specific bandwidth of the kernel selected by optimizing over a large sample of mock spectra (detailed in Section 3.2.4). We find that regressing on the fitted values of the trend filtering estimate — instead of the observational DGP (5.1) — significantly improves the accuracy and robustness of the δ_F estimates. We carry out the LOESS estimation in the Lyα restframe $\overline{\Lambda}_{\text{rest}}(z_0)$ (see Table 3.1) in order to remove the effect of redshifting on the smoothness of m. The LOESS estimation of m is a fully nonparametric procedure and provides a reduction in bias over popular parametric approaches such as low-order power laws [128, 129] and principal components analyses [66, 130–134]. The sole assumption of our LOESS approach is that, in the restframe, the mean flux level m always has a fixed degree of smoothness, defined by an optimal fixed kernel bandwidth. There is of course a bias-variance tradeoff here; namely, the decreased bias comes with a modest increase in variance compared to a parametric power law or low-dimensional principal components model. Our decision in favor of the LOESS approach directly reflects our stance that low bias is preferable to low variance in this context since statistical uncertainty due to estimator variability is tracked by our uncertainty quantification (Section 3.2.6), while uncertainty due to modeling bias is not easily quantifiable. Therefore, we can be more confident that significant fluctuations in the estimated δ_F field are in fact real, and not due to statistical bias in the quasar continuum estimate.

To be explicit, the LOESS estimator for m is a regression on the data set

$$\{(\nu_i,\ \widehat{f}_0(\nu_i;\widehat{\gamma})\}_{i=1}^n, \qquad \nu_i \in \overline{\Lambda}_{\text{rest}}(z_0), \tag{3.6}$$

which can be viewed as arising from the DGP

$$\widehat{f}_0(\nu_i;\widehat{\gamma}) = m(\nu_i) + \rho_i, \tag{3.7}$$

where $\widehat{f}_0$ is the trend filtering estimate fixed at the minimum SURE hyperparameter $\widehat{\gamma}$, $e_i = \widehat{f}_0(\nu_i;\widehat{\gamma}) - f_0(\nu_i)$ are the errors of the trend filtering estimate, and $\rho_i = m(\nu_i) \cdot \delta_F(\nu_i) + e_i$ are autocorrelated fluctuations about zero. The LOESS estimator is the natural extension of kernel regression [135, 136] to higher-order local polynomials. Given a kernel function $K(\cdot)$ with bandwidth $h > 0$, for each $i = 1, \ldots, n$, the LOESS estimator is obtained by minimizing

$$\sum_{j=1}^n \Big(\widehat{f}_0(\nu_j;\widehat{\gamma}) - \phi_{\nu_i}(\nu_j;\beta_0,\ldots,\beta_d)\Big)^2 K\left(\frac{|\nu_j - \nu_i|}{h}\right), \tag{3.8}$$

and letting $\widehat{m}(\nu_i) = \widehat{\beta}_0$, where $\phi_{\nu_i}(\cdot\ ;\beta_0,\ldots,\beta_d)$ is a dth order polynomial centered at ν_i. Specifically, we utilize the local linear regression estimator (LLR; $d = 1$) and the Epanechnikov kernel [137]

$$K(t) = \frac{3}{4}(1-t^2)\mathbb{1}\{|t| < 1\}. \tag{3.9}$$

The LLR estimator is described in full detail by Algorithm 4. Given the trend filtering estimate $\widehat{f}_0$ and the LLR estimate $\widehat{m}$, the δ_F estimates are then defined as

$$\widehat{\delta}_F(z_i;\widehat{\gamma}) = \frac{\widehat{f}_0(z_i;\widehat{\gamma})}{\widehat{m}(z_i)} - 1, \qquad z_1,\ldots,z_n \in \Pi(z_0), \tag{3.10}$$

where we deliberately express $\widehat{m}$ as "hyperparameter-less" since γ has already been fixed at

the minimum SURE value $\widehat{\gamma}$ and we provide the optimal LOESS bandwidth value $h_0 = 74$ Å—optimized over the large mock sample. Here, we have also done a change of variables to redshift space — our desired input domain for studying the H I density fluctuations in the IGM.

Algorithm 4 LOESS (local linear) estimator for mean flux level

Require: Training Data $\{(\nu_i, \widehat{f}_0(\nu_i;\widehat{\gamma}))\}_{i=1}^n$, Bandwidth $h_0 = 74$ Å

1: **for all** i **do**

2: Let $\widehat{\beta}_0$, $\widehat{\beta}_1$ minimize the locally weighted sum of squares

$$\sum_{j=1}^{n} \Big(\widehat{f}_0(\nu_j;\widehat{\gamma}) - \beta_0 - \beta_1(\nu_j - \nu_i)\Big)^2 \cdot K\left(\frac{|\nu_j - \nu_i|}{h_0}\right).$$

3: Let $\widehat{m}(\nu_i) = \widehat{\beta}_0(\nu_i)$.

4: **end for**

Output: $\{\widehat{m}(\nu_i)\}_{i=1}^n$

Although h_0 is chosen to directly optimize $\widehat{\delta}_F$ accuracy, an estimate for the quasar continuum $C(\cdot)$ arises intrinsically:

$$\widehat{C}(\nu_i) = \overline{F}(\nu_i)^{-1} \cdot \widehat{m}(\nu_i), \qquad \nu_i \in \overline{\Lambda}_{\text{rest}}(z_0), \tag{3.11}$$

where precise estimates of $\overline{F}$ follow from a rich literature [e.g., 14, 133, 138–141]. The δ_F estimates could then be equivalently restated as

$$\widehat{\delta}_F(z_i;\widehat{\gamma}) = \frac{\widehat{F}(z_i;\widehat{\gamma})}{\overline{F}(z_i)} - 1, \qquad z_1, \ldots, z_n \in \Pi(z_0), \tag{3.12}$$

where

$$\widehat{F}(z_i;\widehat{\gamma}) = \widehat{f}_0(z_i;\widehat{\gamma})/\widehat{C}(z_i). \tag{3.13}$$

3.2.4 Calibrating continuum smoothness

We utilize a sample of 124,709 mock quasar spectra from the [85] catalog to optimize the bandwidth of the LOESS estimator for the mean flux level that intrinsically removes the effect of

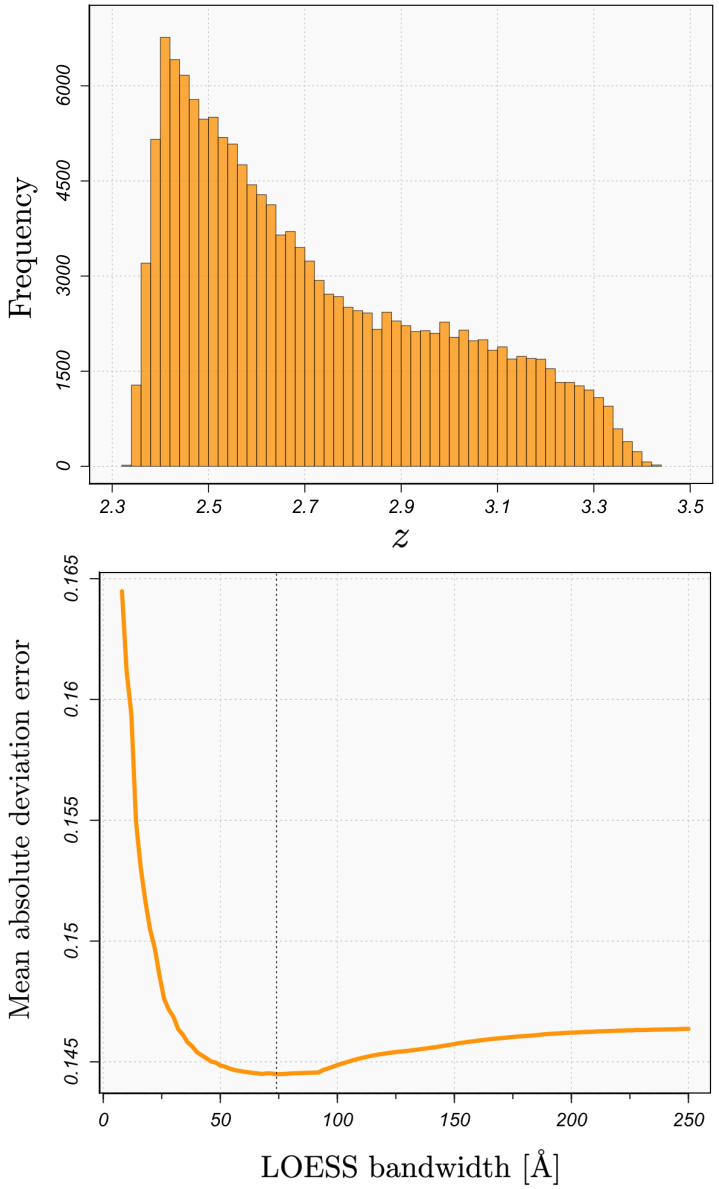


FIGURE 3.1: **Top**: Distribution of mock quasar redshifts (data reduction detailed in Section 3.2.5). We utilize this sample of 124,709 quasars to calibrate the optimal nonparametric continuum smoothness. **Bottom**: Mean absolute deviation error curve for selecting the optimal kernel bandwidth for the LOESS (local linear) estimator of the mean flux level, averaged over the 124,709 spectra in the mock sample. The optimal choice of bandwidth is $h_0 = 74$ Å.

the quasar continuum. Our mock data reduction is detailed in the Section 3.2.5 and the redshift distribution of the quasars is shown in the top panel of Figure 3.1.

For each mock quasar Lyα forest with DGP Q_j, $j = 1, \ldots,$ 124,709, we first compute the trend filtering hyperparameter value that minimizes the *true* fixed-input mean-squared prediction error

$$\gamma_0^j = \underset{\gamma>0}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Q_j} \Big[\big(f(\zeta_i) - \widehat{f}_0(\zeta_i; \gamma) \big)^2 \;\Big|\; \zeta_1, \ldots, \zeta_n \Big]. \tag{3.14}$$

Then, given the trend filtering restframe fitted values $\{(\nu_i,\ \widehat{f}_0(\nu_i; \gamma_0^j)\}_{i=1}^n$, we fit a LOESS estimator with bandwidth h and define the error (as a function of h) of the resulting δ_F estimator as the fixed-input mean absolute deviation (MAD) error

$$R_j(h) = \frac{1}{n} \sum_{i} \mathbb{E}_{Q_j} \Big[\big| \delta_F(\nu_i) - \widehat{\delta}_F(\nu_i; \gamma_0^j, h) \big| \;\Big|\; \nu_1, \ldots, \nu_n \Big], \tag{3.15}$$

where $\mathbb{E}_{Q_j}$ denotes the mathematical expectation over the randomness arising from the observational DGP Q_j. Because we can repeatedly sample from each mock quasar DGP, the expectations in equations (3.14) and (3.15) can be computed to an arbitrary precision. Here, we utilize 300 realizations of each DGP to approximate the mathematical expectations.

We then define the optimal choice of h as the minimizer of the summed error over the full sample of m =124,709 mock quasar spectra:

$$h_0 = \underset{h>0}{\operatorname{argmin}} \sum_{j=1}^{m} R_j(h). \tag{3.16}$$

The aggregate error curve is shown in the bottom panel of Figure 3.1, yielding an optimal value of $h_0 = 74$ Å. We find that defining $R(h)$ as the conditional MAD error — instead of the conditional MSE — provides an essential boost in robustness that keeps the error from being dominated by a very small proportion of worst-case estimates. More complex choices of bandwidth, e.g. variable bandwidths that depend on the S/N ratio of the trend filtering estimator, the restframe pixel spacing, and/or the redshift of the quasar, do not significantly improve upon the $h_0 = 74$

Å restframe fixed bandwidth.

3.2.5 Mock quasar Lyman-α forest reduction

The [85] mock quasar catalog is designed to mimic the observational data generating processes of the quasar spectra released in Data Release 11 of the Baryon Oscillation Spectroscopic Survey [27]. We pool the first three realizations of the mock catalog, i.e. `M3_0_3/000`, `M3_0_3/001`, and `M3_0_3/002` and remove all damped Lyα systems (DLAs), Lyman-limit systems (LLS), and broad absorption line quasars (BALs). We assume no metal absorption in the Lyα forest and correct estimation and subtraction of the sky. We mask all pixels with `and_mask` $\neq$ `0` or `or_mask` $\neq$ `0`. Finally, we retain only the spectra with ≥ 500 pixels in the truncated Lyα forest and those with a fixed-input-optimal trend filtering hyperparameter satisfying $\gamma_0 < 5.25$. Spectra with $\gamma_0 \geq 5.25$ correspond to the very lowest S/N ratio observations, where the trend filtering estimate typically reduces to a global power law fit (zero knots). The final mock sample contains 124,709 quasar spectra.

3.2.6 Uncertainty quantification

Given the assumed noise model $\epsilon_i \sim N(0, \sigma_i^2)$, $i = 1, \ldots, n$ provided by the BOSS pipeline [66], we can construct a pointwise variability band for $\widehat{\delta}_F$ via an augmentation of the parametric bootstrap outlined in Algorithm 2 of Chapter 2. Specifically, given the parametric bootstrap ensemble of trend filtering estimates provided by the parametric bootstrap algorithm, we fit the mean flux level of each with the LOESS estimator detailed in Algorithm 4, and then define the $\widehat{\delta}_F$ bootstrap ensemble

$$\widehat{\delta}^*_{F,b}(z_i) = \frac{\widehat{f}^*_b(z_i)}{\widehat{m}^*_b(z_i)} - 1, \qquad i = 1, \ldots, n, \;\; b = 1, \ldots, B, \tag{3.17}$$

where, for each $b = 1, \ldots, B$, $\widehat{m}_b^*$ is the LOESS estimate fit to the data set $\big\{(\nu_i, \widehat{f}_b^*(\nu_i))\big\}_{i=1}^n$. Note that refitting the LOESS estimator on each bootstrap realization allows us to track the extra variability introduced into the δ_F estimates by the uncertainty in the continuum estimate. A $(1-\alpha)\cdot 100\%$ quantile-based pointwise variability band for $\widehat{\delta}_F$ is then given by

$$V_{1-\alpha}(z_i) = \Big(\widehat{\delta}^*_{F,\alpha/2}(z_i),\ \widehat{\delta}^*_{F,1-\alpha/2}(z_i)\Big), \qquad i = 1, \ldots, n. \tag{3.18}$$

where

$$\widehat{\delta}^*_{F,\beta}(z_i) = \inf_g \left\{ g : \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\big\{\widehat{\delta}^*_{F,b}(z_i) \le g\big\} \ge \beta \right\}, \tag{3.19}$$

for any $\beta \in (0,1)$.

In particular, given a quasar at redshift $z = z_0$ with wavelength inputs $\{\lambda_i\}_{i=1}^n$, $\lambda_i \in \overline{\Lambda}(z_0)$, let

$$V_{\widehat{f}_0(\lambda_i)}(\alpha) = \Big(\widehat{f}^*_{\lambda_i}(\alpha/2),\ \widehat{f}^*_{\lambda_i}(1-\alpha/2)\Big), \qquad i = 1, \ldots, n \tag{3.20}$$

denote the $1-\alpha$ pointwise variability band of $\widehat{f}_0$, with the bounds given by the $\alpha/2$ and $1-\alpha/2$ quantiles of the parametric bootstrap ensemble from Algorithm 2.

Although the bootstrap does not provide finite sample coverage guarantees, we are able to study the coverage properties of the bands in equations (3.20) and (3.18) empirically on the mock sample of 124,709 quasar spectra. Let Q_j, $j = 1, \ldots, 124{,}709$ denote the DGPs of the mock quasar Lyα forests. For any $\alpha \in (0,1)$, define the *proportion coverage* of the bands in equations (3.20) and (3.18) to be

$$\text{coverage}^j_{f_0}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_{Q_j}\Big(f_0(\lambda_i) \in\ V_{\widehat{f}_0(\lambda_i)}(\alpha)\Big), \tag{3.21}$$

and

$$\text{coverage}^{j}_{\delta_F}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}_{Q_j}\Big(\delta_F(z_i) \in V_{\widehat{\delta}_F(z_i)}(\alpha)\Big), \tag{3.22}$$

respectively, where $\mathbb{P}_{Q_j}$ denotes the probability over the randomness of the DGP Q_j. We utilize 300 realizations from each DGP to compute (3.21) and (3.22) and then define the average empirical coverages over the mock sample as

$$\text{coverage}_{f_0}(\alpha) = \frac{1}{m}\sum_{j=1}^{m} \text{coverage}^{j}_{f_0}(\alpha), \tag{3.23}$$

and

$$\text{coverage}_{\delta_F}(\alpha) = \frac{1}{m}\sum_{j=1}^{m} \text{coverage}^{j}_{\delta_F}(\alpha). \tag{3.24}$$

Table 3.2 displays the results for various values of α. In particular, the *bootstrap inner-percentile width* is given by $1 - \alpha$.

3.2.7 Results

A mock quasar spectrum from [85] and a real quasar spectrum from the Baryon Oscillation Spectroscopic Survey [BOSS; 27] are displayed in Figure 3.2, with the results of our Lyα forest analysis overlaid. Starting with the top panel, the quadratic trend filtering estimate is fit on the equally-spaced observations in logarithmic wavelength space and then transformed to the restframe wavelength space, where the LOESS estimate for the mean flux level — the product of the continuum and cosmic mean Lyα absorption — is then fit to the trend filtering fitted values. The δ_F estimates are then computed according to (3.10) and displayed in redshift space in the second panel, where they closely track the true δ_F defined by (3.4). Furthermore, the 95% bootstrap variability band defined in (3.18) can be seen to almost fully cover the true δ_F signal. The estimated δ_F can be interpreted as an inversely proportional proxy for the deviations

$\widehat{f}_0$		$\widehat{\delta}_F$	
Bootstrap percentile width	**Empirical coverage**	**Bootstrap percentile width**	**Empirical coverage**
0.50	0.426	0.50	0.346
0.55	0.471	0.55	0.384
0.60	0.517	0.60	0.424
0.65	0.564	0.65	0.466
0.70	0.612	0.70	0.510
0.75	0.663	0.75	0.556
0.80	0.715	0.80	0.607
0.85	0.771	0.85	0.664
0.90	0.831	0.90	0.732
0.91	0.844	0.91	0.747
0.92	0.857	0.92	0.764
0.93	0.871	0.93	0.782
0.94	0.884	0.94	0.800
0.95	0.898	0.95	0.819
0.96	0.912	0.96	0.838
0.97	0.926	0.97	0.858
0.98	0.942	0.98	0.880
0.99	0.960	0.99	0.908

TABLE 3.2: Empirical coverages of parametric bootstrap variability bands of various widths (Algorithm 2), averaged over the sample of 124,709 mock quasar Lyα forests. The left column displays the average empirical coverage of the variability bands of the flux signal estimator $\widehat{f}_0$ and the right columns displays the same for the δ_F estimator $\widehat{\delta}_F = \widehat{f}_0/\widehat{m} - 1$, where $\widehat{m}$ is the LOESS estimate for the mean flux level. See Section 3.2.6 for more details.

from the mean H I density in the intervening intergalactic medium at each redshift — with negative values of δ_F corresponding to (epoch-relative) over-densities of H I and positive values corresponding to under-densities.

The third and fourth panels are the analogous plots for a BOSS quasar spectrum Lyα forest (Data Release 12, Plate = 6487, MJD = 56362, Fiber = 647). The quasar is located in the Northern Galactic Cap at (RA, Dec, z) $\approx$ (196.680°, 31.078°, 2.560).

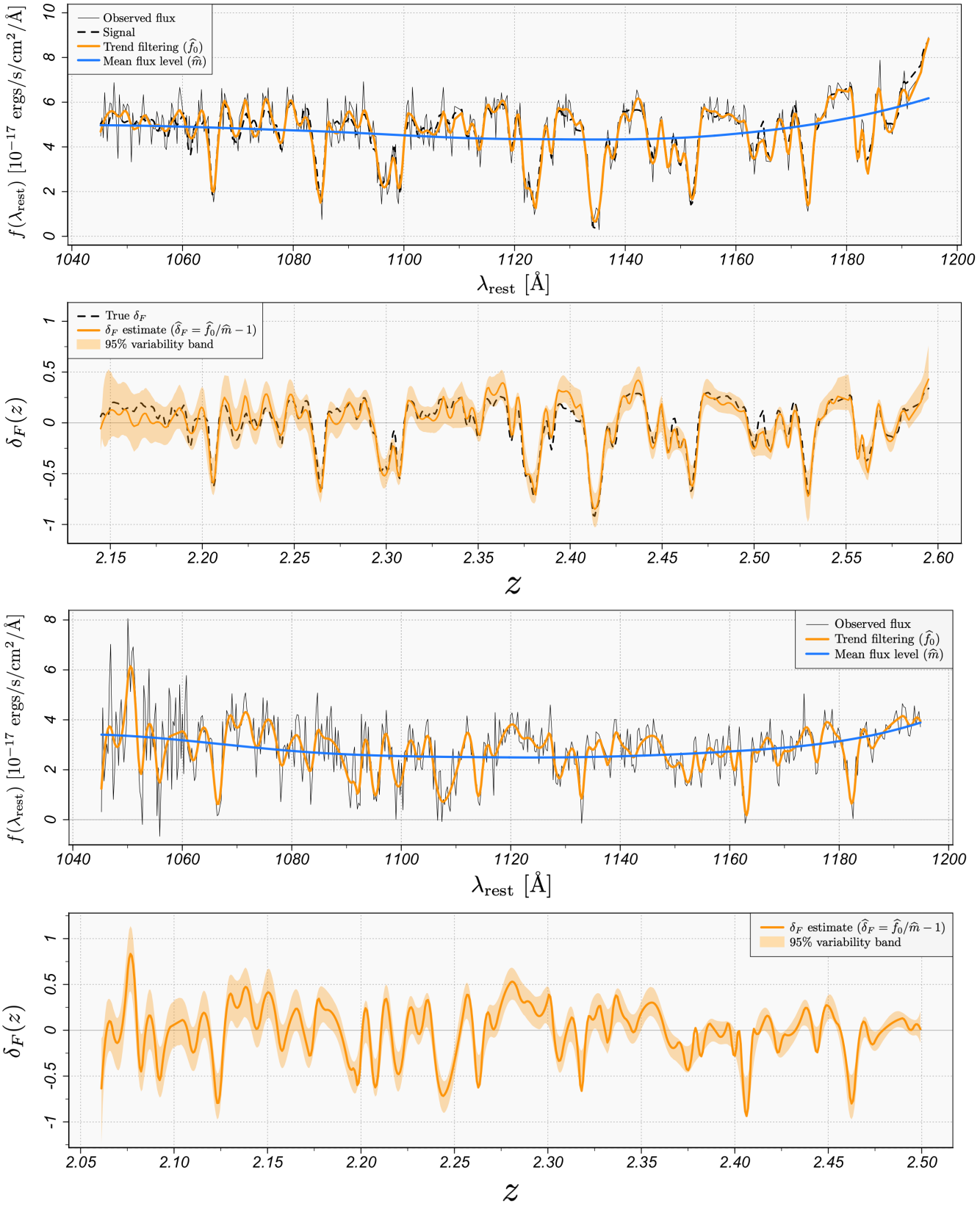


FIGURE 3.2: Results of Lyα forest analysis. **Top panel**: Lyα forest of a mock quasar spectrum [85] in the restframe, with the quadratic trend filtering estimate shown in orange and the LOESS (local linear) estimate for the mean flux level shown in blue. **Second panel**: The redshift-space fluctuations in the Lyα transmitted flux fraction, with our estimate superposed. The fluctuations inversely trace the relative under- and over-densities of H I in the intervening intergalactic medium between Earth and the quasar. **Third and Fourth panels**: Analogous plots for a real quasar Lyα forest from the twelfth data release of the Baryon Oscillation Spectroscopic Survey [27, ; Plate = 6487, MJD = 56362, Fiber = 647]. The quasar is located at (RA, Dec, z) $\approx$ (196.680°, 31.078°, 2.560).

3.3 Further Applications

The analysis of signals possessing varying degrees of smoothness permeates many areas of astronomy. In this section, we discuss a variety of further applications for which trend filtering may find use. For the sake of brevity, we discuss these applications in less detail than the Lyα forest analysis in Section 3.2. Naturally, we expect trend filtering may find many uses in astronomy beyond those we explicitly discuss.

3.3.1 Spectral template generation and estimation of emission-line parameters

Automated spectral classification and redshift estimation pipelines require rich template libraries that span the full space of physical objects in the targeted set in order to achieve high statistical accuracy. In this section, we discuss using trend filtering to construct spectral templates from observational spectra. We describe the procedure here for generating a single spectral template from a well-sampled observed spectrum. Suppose we observe coadded flux measurements of a targeted source at wavelengths $\lambda_1, \ldots, \lambda_n \in \Lambda$ according to the data generating process

$$f(\lambda_i) = f_0(\lambda_i) + \epsilon_i, \tag{3.25}$$

where the observations have been corrected for systematic effects (e.g., sky subtraction, interstellar extinction, etc.) and the ϵ_i are mean-zero errors that arise from instrumental uncertainty and systematic miscalibrations. After removal of potentially problematic pixels (e.g., near bright sky lines), let $\widehat{f}_0(\lambda),\ \lambda \in \Lambda$ denote the trend filtering estimate fit to the observations. Given a confident object classification and redshift estimate z_0 (e.g., determined by visual inspection), we

then define the restframe spectral template

$$b(\lambda_{\text{rest}}) = \widehat{f}_0(\lambda/(z_0+1)), \qquad \lambda_{\text{rest}} \in \Lambda/(z_0+1), \tag{3.26}$$

and store it in the respective object template library.

In Figure 3.3, we show three optical coadded spectra from the twelfth data release of the Baryon Oscillation Spectroscopic Survey [BOSS DR12; 27] of the Sloan Digital Sky Survey III [SDSS-III; 26, 123]. The figure is zoomed to a subinterval of the optical range for visual clarity and the spectra are easily identifiable as a quasar, a galaxy, and a star, respectively. We fit an error-weighted quadratic trend filtering estimate to each spectrum in the logarithmic-angstrom wavelength space in which the BOSS spectra are gridded, and tune the hyperparameter to minimize Stein's unbiased risk estimate (see Section 2.3.5). The trend filtering estimates give good results, adequately adapting to even the strongest emission and absorption features in each spectrum.

As in the BOSS spectroscopic pipeline [66], after a spectral classification and redshift measurement have been precisely determined, emission-line parameter estimates can then be obtained by fitting Gaussian radial basis functions to the emission lines of the spectrum "best fit" — nonlinearly optimizing the amplitudes, centroids, and widths. We propose that the trend filtering estimate of a spectrum be used as the "best fit" for this type of procedure, e.g. instead of the low-dimensional principal components models currently used by the BOSS pipeline. The relative magnitudes of the emission-line parameter estimates can then be used to determine object subclassifications.

3.3.2 Exoplanet transit modeling

In this section we discuss the use of trend filtering for modeling the photometric time series of an extrasolar planet transiting its host star. Given a phase-folded stellar light curve, corrected for

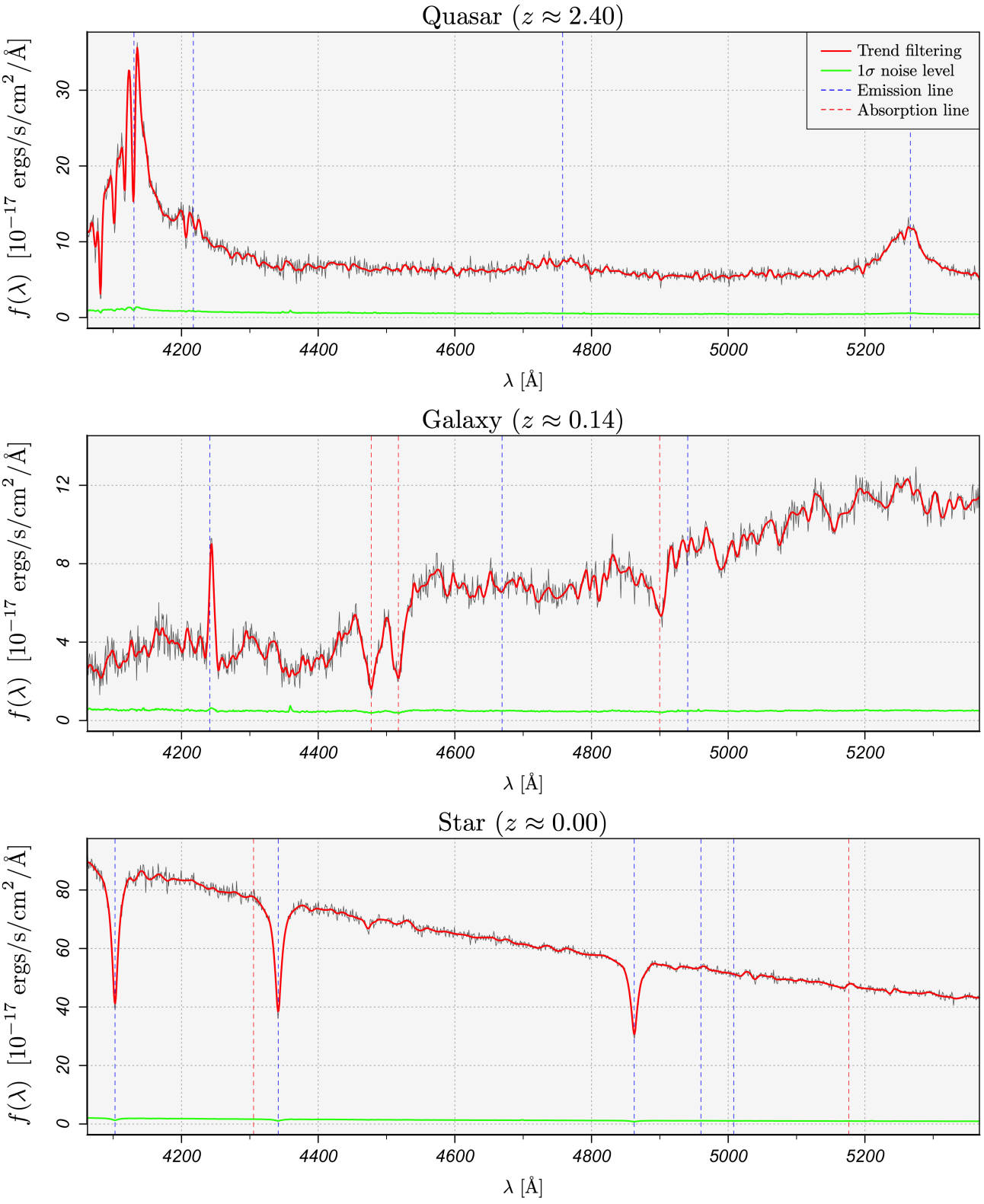


FIGURE 3.3: Optical coadded spectra collected by the Baryon Oscillation Spectroscopic Survey [27] of the Sloan Digital Sky Survey III [26, 123]. From top to bottom, a quasar (DR12, Plate = 7140, MJD = 56569, Fiber = 58) located at (RA, Dec, z) $\approx$ (349.737°, 33.414°, 2.399), a galaxy (DR12, Plate = 7140, MJD = 56569, Fiber = 68) located at (RA, Dec, z) $\approx$ (349.374°, 33.617°, 0.138), and a star (DR12, Plate = 4055, MJD = 55359, Fiber = 84) located at (RA, Dec, z) $\approx$ (236.834°, 0.680°, 0.000). We fit a quadratic trend filtering estimate to each spectrum in the logarithmic wavelength space in which the observations are equally spaced, and optimize the hyperparameter by minimizing Stein's unbiased risk estimate. Given confidently determined redshifts (e.g., determined by visual inspection), the trend filtering estimate for each object can be scaled to the restframe and stored as a spectral template. Furthermore, emission-line parameter estimates for a spectrum can be obtained by fitting Gaussian radial basis functions to the emission lines of the trend filtering estimate.

stellar variability and spacecraft motion systematics, trend filtering can be used to automatically produce fully nonparametric estimates and uncertainties for the transit depth and total transit duration.

We demonstrate our approach on long-cadence photometric observations of Kepler-10 [KOI-72, KIC 11904151; 5]. Kepler-10 is confirmed to host at least two exoplanets: Kepler-10b [KOI-72 b, KIC 11904151 b; 142] and Kepler-10c [KOI-72 c, KIC 11904151 c; 108]. Each planet was first detected via the transit method — a measurable periodic dimming in the photometry caused by the planet crossing in front of the host star. Here, we use trend filtering to estimate a nonparametric transit model for Kepler-10c and derive depth and duration measurements. The results of our analysis are displayed in Figure 3.4. The top panel displays a sample of the Kepler-10 long-cadence (30-min. increment), quarter-stitched, median detrended, relative flux light curve processed by the *Kepler* pipeline [143], which we accessed from the NASA Exoplanet Archive [144]. The observed transit events of Kepler-10c are indicated by the vertical dashed lines. The middle panel displays the light curve (with 1σ error bars) after phase-folding with respect to the $\sim$45.29 day orbital period of Kepler-10c and zooming in on the transit event. We fit a relaxed quadratic trend filtering estimate (see Section 2.3.7) to the phased data, weighted by the measurement variances provided by the *Kepler* pipeline and tuned by sequential K-fold cross validation. The optimal relaxation hyperparameter is $\widehat{\phi} = 0.14$, indicating that relaxation significantly improves the traditional trend filtering estimate in this setting. In particular, we find that relaxation allows the estimate to faithfully capture the sharp transitions corresponding to the beginning of the ingress phase and the end of the egress phase. The relaxed trend filtering estimate is overlaid on the phase-folded light curve, along with 95% variability bands. Estimates for the transit depth and total transit duration follow immediately from the relaxed trend filtering estimate, as detailed below. The estimated inception and termination of the transit event are indicated by the vertical dashed lines in the middle panel. The nonparametric bootstrap sampling

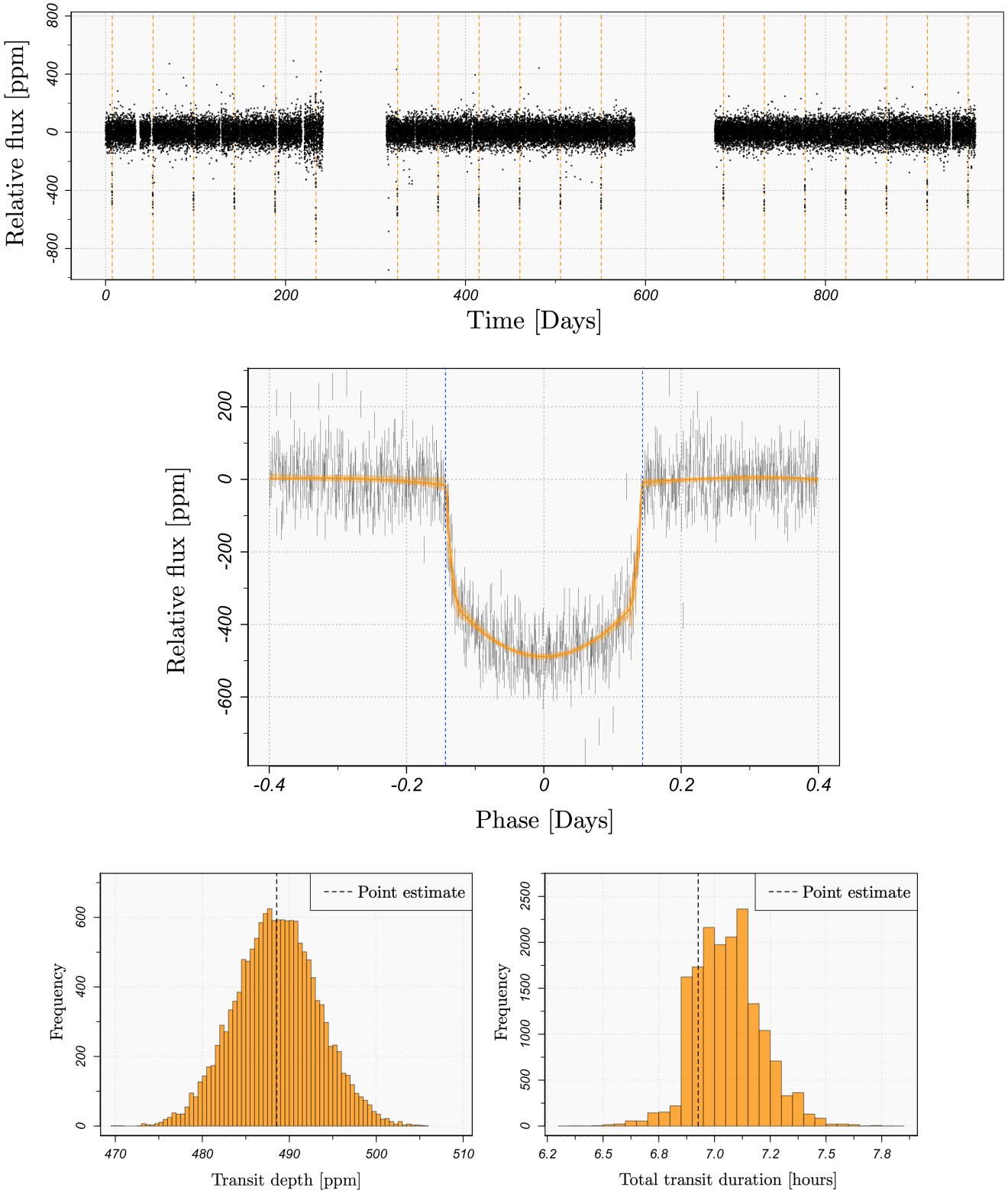


FIGURE 3.4: Kepler-10c transit light curve analysis. **Top**: Long-cadence (30-min.), quarter-stitched, median detrended, relative flux light curve (`LC_DETREND`) of the confirmed exoplanet host Kepler-10 [KOI-072, KIC 11904151; 142], processed by the *Kepler* pipeline [143] and obtained from the NASA Exoplanet Archive [144]. Vertical lines indicate the observed transit events of the system's second confirmed planet Kepler-10c [KOI-072 c, KIC 11904151 c; 108]. **Middle**: Phase-folded transit light curve for Kepler-10c (∼45.29 day orbital period) with 1σ error bars. The error-weighted relaxed trend filtering estimate, optimized by sequential K-fold cross validation, is superposed with 95% variability bands. The estimated inception and termination of the transit event are indicated by the vertical dashed lines. The estimated transit depth and total transit duration are $\widehat{\delta} = 488.292$ ppm and $\widehat{T} = 6.927$ hours, respectively. **Bottom**: Bootstrap sampling distributions of the transit depth and transit duration estimates.

distributions (see Algorithm 2) of the transit depth and total transit duration measurements are displayed in the bottom panel. Our point estimates for the transit depth and total transit duration for Kepler-10c are $\widehat{\delta} = 488.292$ ppm and $\widehat{T} = 6.927$ hours, respectively.

The knot-selection property of trend filtering is particularly appealing in this setting because it provides interpretation as to where the transit event begins and ends. Specifically, we define our estimate of the inception of the transit event $\widehat{T}_0$ as the leftmost knot selected by the trend filtering estimator and we define our estimate of the termination of the transit event $\widehat{T}_1$ as the rightmost knot[1]. The total transit duration estimate then follows as $\widehat{T} = \widehat{T}_1 - \widehat{T}_0$. Naturally, we define our transit depth estimate $\widehat{\delta}$ to be the minimum of the relaxed trend filtering estimate.

It is thus far unclear to us whether trend filtering can also reliably detect the end of the ingress and beginning of the egress phases — e.g. via knot selection or examining changes in the estimated derivatives — and therefore also provide nonparametric ingress/egress duration measurements. We recommend pairing trend filtering with the traditional analytical transit model search [e.g., 145] for these particular measurements. That is, given the transit depth and total transit duration measurements provided by trend filtering, an analytical planet model can be fit over a parameter space that is constrained by the trend filtering parameter measurements. Coupling the two methods in this way therefore also provides significant computational speedups over a traditional single-stage analytical model search since it greatly reduces the dimensionality of the parameter space to be searched over. Furthermore, the relaxed trend filtering estimate may provide a benchmark χ^2 statistic for the constrained analytical model comparison.

A dedicated paper on this application of trend filtering is forthcoming.

[1] Note that the boundary points of the input space are not considered knots.

3.3.3 Eclipsing binary modeling

In this section we discuss how trend filtering can further improve the work of the *Kepler* Eclipsing Binary working group, specifically in regard to the Eclipsing Binaries via Artificial Intelligence (EBAI) pipeline used to characterize *Kepler* eclipsing binary stars via their phase-folded light curves [8, 9, 146, 147]. The EBAI pipeline utilizes an artificial neural network (ANN) to estimate a set of physical parameters for each binary pair (the temperature ratio, sum of fractional radii, photometric mass ratio, radial and tangential components of the eccentricity, fillout factor, and inclination) from the observables of the phase-folded light curve (e.g., the eclipse widths, depths, and separations). [8] outline the EBAI light curve pre-processing algorithm, which they call `polyfit`, that provides the crucial step of taking a noisy, irregularly-spaced, phase-folded light curve (detrended for spacecraft motion and normalized by the median flux) and outputting a denoised and gridded phase-folded light curve, which is then fed to the ANN. We propose that trend filtering be used for this pre-processing step instead of the `polyfit` algorithm for the reasons detailed below.

An eclipsing binary (EB) light curve is characterized by periodic dips in the observed brightness that correspond to eclipse events along the line of sight to an observer. In particular, there are two eclipses per orbital period — a primary and a secondary eclipse. The primary eclipse occurs when the hotter star is eclipsed by the cooler star and produces a comparatively deep dip in observed brightness. Analogously, the secondary eclipse occurs when the cooler star is eclipsed by the hotter star and produces a comparatively shallow dip in the observed brightness. Depending on the effective temperature ratio and orbital period of the EB, the dips may range from very narrow and abrupt to very wide and smooth. In Figure 3.5, we display a detrended, median-normalized, long-cadence (30 min. increment) light curve of a *Kepler* EB [KIC 6048106; 5, 9], with the primary eclipses and secondary eclipses designated by dashed red and blue lines,

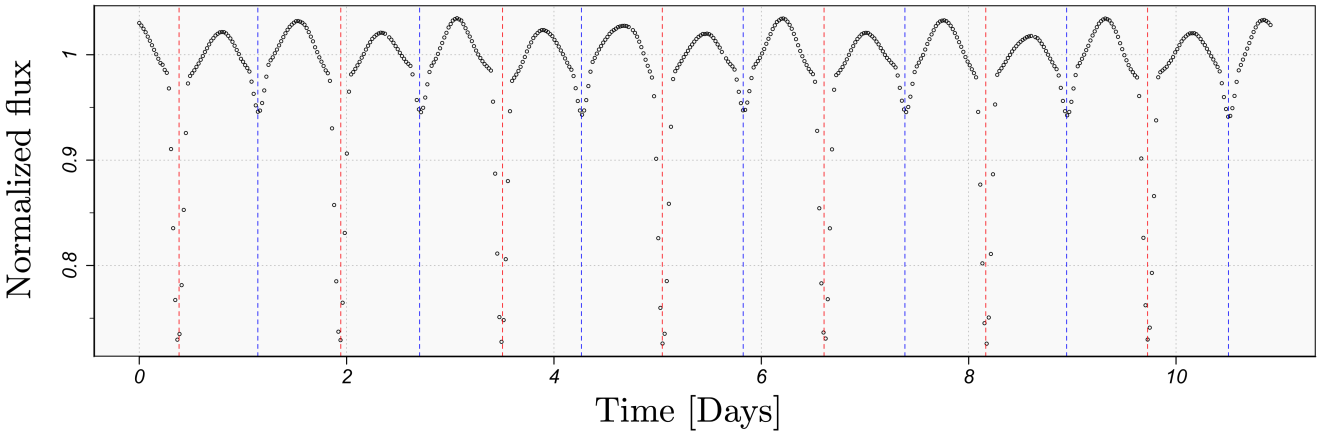


FIGURE 3.5: Long-cadence (30-min. increment), detrended, median-normalized light curve of a *Kepler* eclipsing binary system [KIC 6048106; 5, 9]. The vertical red lines mark the primary eclipses (the eclipses of the hotter star) and the vertical blue lines mark the secondary eclipses (the eclipses of the cooler star). KIC 6048106 has an orbital period of ∼1.559 days.

respectively. The orbital period of KIC 6048106 is ∼1.559 days.

After phase-folding with respect to the estimated EB orbital period and centering the primary eclipse at `Phase = 0`, the purpose of the EBAI light-curve pre-processing step is to faithfully extract the signal of the phase-folded light curve and evaluate it on a regular grid so that it can then be input into the EBAI ANN. We show a comparison of the `polyfit` algorithm of [8] and our trend filtering approach in Figure 3.6 on the phase-folded KIC 6048106 light curve. We choose a relatively high S/N light curve here in order to elucidate the significant statistical bias that underlies `polyfit`. The `polyfit` algorithm fits a piecewise quadratic polynomial by weighted least-squares with four knots selected by a randomized computational search over the phase space. The piecewise quadratic polynomial is forced to be continuous, but no smoothness constraints are placed on the derivatives of the estimate at the knots. Recalling our discussion in Section 2.3.1, this overly-stringent modeling assumption leads to significant statistical bias in the light-curve estimate. The bias is particularly apparent by examining the residuals of the `polyfit` estimate, which we display below the light curve. Moreover, recalling our discussion of variable-knot regression splines in Section 2.3.1.1, a randomized partial search over the space of feasible knots inherently provides no guarantee of finding a global solution — leaving the algorithm susceptible

to extreme failure scenarios. We display a quadratic trend filtering estimate of the KIC 6048106 phase-folded light curve in the bottom panel of Figure 3.6, with the hyperparameter chosen by K-fold cross validation. The trend filtering estimate accurately recovers the signal of the light curve (clearly apparent here by examining a high S/N ratio light curve) and produces a desirable random residual scatter about zero. Since the statistical bias introduced by the `polyfit` pre-processing stage propagates through to the EBAI ANN as systematic bias in the input data, we are confident that the use of trend filtering will in turn improve the error rate of the EBAI ANN output-parameter estimates.

3.3.4 Supernova light-curve template generation and estimation of observable parameters

In this section we demonstrate the use of trend filtering for generating light-curve templates of supernova (SN) events and estimating observable parameters. We illustrate our approach on SN 2016coi [ASASSN-16fp; 148] by constructing a B-band light-curve template from the well-sampled observations of [149] and [150] and deriving nonparametric estimates and full uncertainty distributions for the maximum apparent brightness, the time of maximum, and the decline rate parameter $\Delta m_{15}(B)$ introduced by [151]. The improvement yielded by trend filtering as a tool for SN light-curve template generation, compared to, for example, the `SNooPy` cubic smoothing spline approach of the Carnegie Supernova Project [CSP; 39, 152, 153] primarily corresponds to light curves with an especially bright peak magnitude and fast decline rate. In such cases, trend filtering is better able to recover the abruptness of the peak, the initial sharp decline, and the subsequent slow decay. This behavior is particularly characteristic of Type Ia SNe [e.g., 10]. In cases where the peak is not particularly prominent, trend filtering and cubic smoothing splines produce nearly identical estimates. Our procedure for generating the SN light-curve templates requires reasonably well-sampled observations (in particular, with the

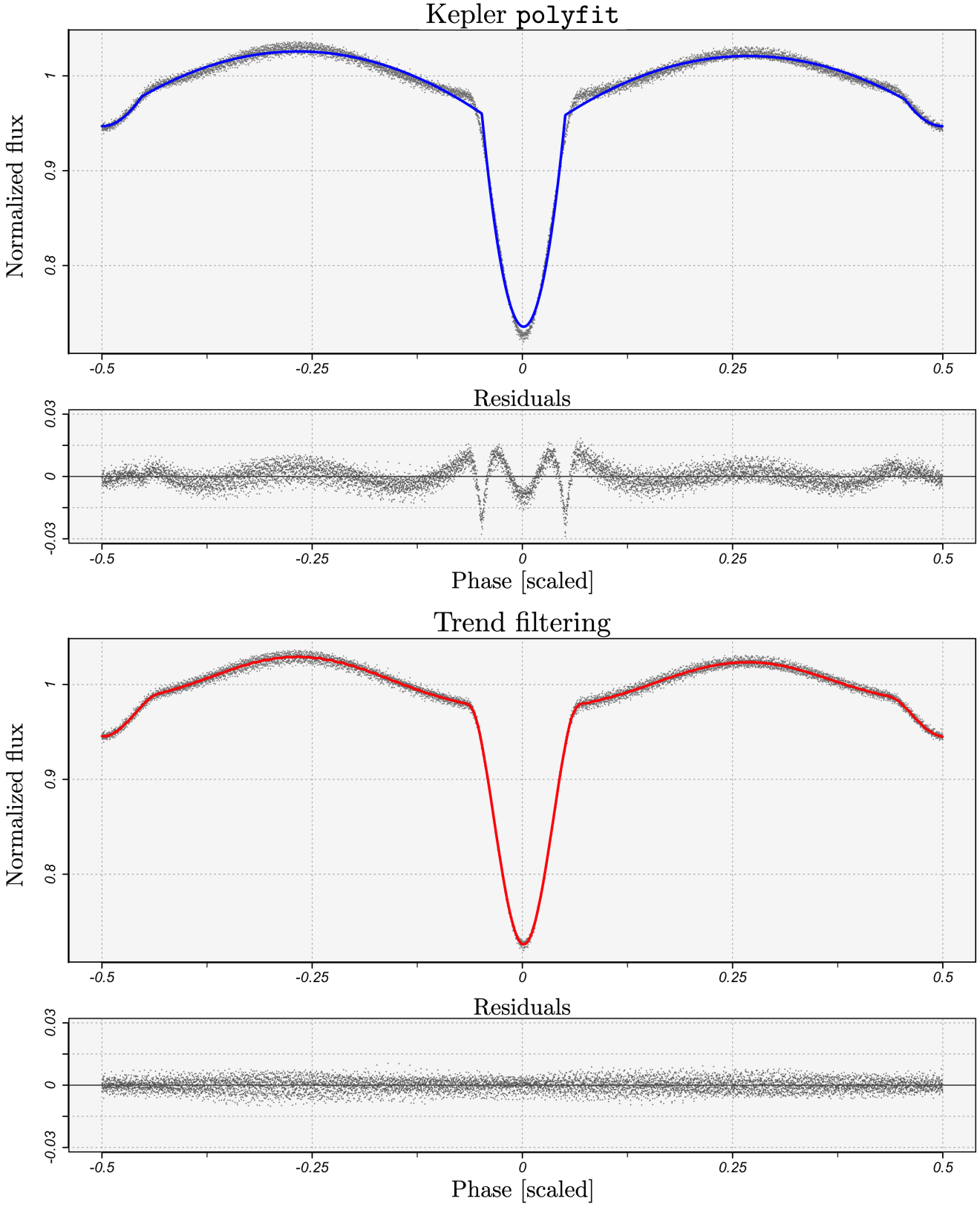


FIGURE 3.6: Comparison of the `polyfit` algorithm of [8] and our trend filtering approach for denoising phase-folded eclipsing binary light curves. The light curve shown in this example comes from the *Kepler* eclipsing binary system KIC 6048106 [5, 9]. **Top**: The `polyfit` algorithm fits a piecewise quadratic polynomial by weighted least-squares with four knots selected by a randomized search over the phase space. The estimate is constrained to be continuous but no constraints are enforced on the derivatives at the knots. The overly-stringent assumed model leads to significant statistical bias, which is readily apparent by examining the autocorrelation in the residuals. **Bottom**: Trend filtering is sufficiently flexible to accurately denoise the diverse set of signals observed in phase-folded eclipsing binary light curves. Here, the goodness-of-fit is clear by the random, mean-zero residual scatter.

initial observation occurring before maximum light). The resulting template libraries can then be used to classify SNe with partially-sampled light curves and derive parameter estimates [e.g., 154].

The SN light-curve template generation procedure is analogous to the spectral template generation procedure discussed in Section 3.3.1, so we discuss it in less formal detail here. Naturally, the same procedure can also be implemented for generating fixed-time spectral templates of SN events. Given a well-sampled light curve, corrected for systematic effects (e.g., K-corrections and interstellar reddening), we propose the use of quadratic trend filtering to generate a "best fit" to the observations. Given a confident type classification, the trend filtering estimate can then be stored as a light-curve template in the respective library.

We show a B-band light curve for SN 2016coi [ASASSN-16fp; 148] in the top panel of Figure 3.7. The light curve is an aggregation of observations collected by [149] and [150], which we accessed from the Open Supernova Catalog [110]. SN 2016coi was discovered on May 27, 2016 (UT 2016-05-27.55) by the All Sky Automated Survey for SuperNovae [ASAS-SN; 155] in the galaxy UGC 11868 at redshift $z \approx 0.0036$. The classification of SN 2016coi currently remains intermediate between Type Ib and Type Ic [150, 156–158]. We fit a quadratic trend filtering estimate to the observations, weighted by measurement uncertainties and with the hyperparameter selected by K-fold cross validation. The trend filtering estimate, along with 95% nonparametric bootstrap variability bands, is overlaid in the top panel of the figure. In the bottom panels we show the univariate nonparametric bootstrap sampling distributions for the estimates of the maximum apparent magnitude, the time of maximum, and the decline rate — defined as the relative change in the B-magnitude light curve from maximum light to the magnitude 15 days after the maximum [151]. We also show a bivariate bootstrap sampling distribution for maximum apparent magnitude versus decline rate. The bimodality in the bootstrap sampling distributions arises

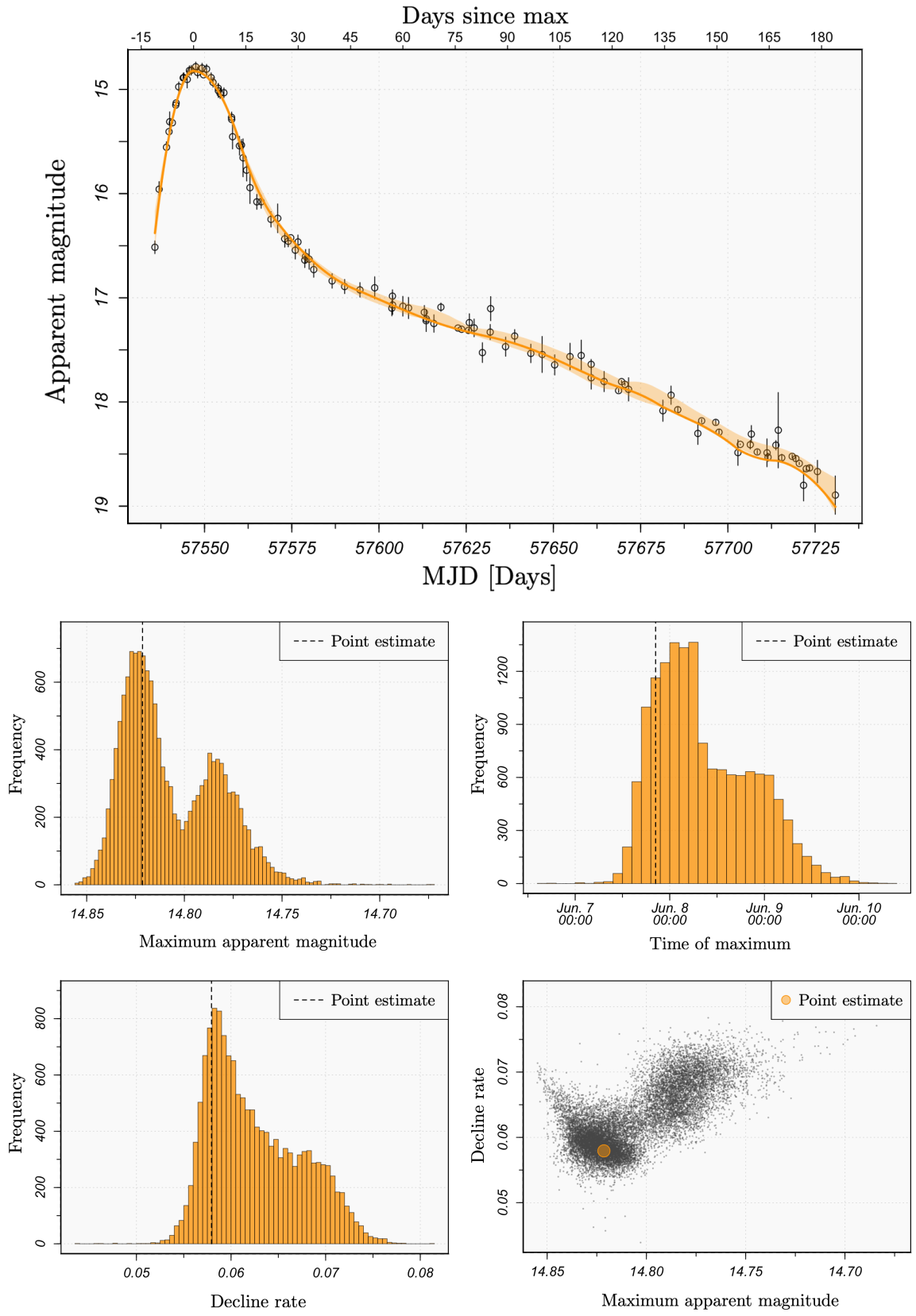


FIGURE 3.7: SN light-curve analysis (SN 2016coi). **Top**: B-band photometry of the supernova SN 2016coi [ASASSN-16 fp; 148] discovered on May 27, 2016 by the All Sky Automated Survey for SuperNovae [ASAS-SN; 155] in the galaxy UGC 11868 at redshift $z \approx 0.0036$. We fit a quadratic trend filtering estimate, tuned by K-fold cross validation, and overlay 95% nonparametric bootstrap variability bands. **Bottom**: Univariate/bivariate nonparametric bootstrap sampling distributions of the observable parameter estimates derived from the trend filtering light-curve estimate. The bimodality in the bootstrap parameter distributions arises from systematic discrepancies between the observations of the two separate observers [149, 150].

from systematic discrepancies between the two separate sets of B-band observations that form the light curve [149, 150].

3.3.5 Data reduction and compression

Although the primary focus of this chapter is the use of trend filtering as a tool for astronomical data analysis, it also possesses potential utility for large-scale reduction and compression of one-dimensional data sets. This owes to several factors: its speed, scalability, flexibility, and representation as a sum of basis functions with a sparse coefficient vector.

Given a one-dimensional set of n observations $(t_1, f(t_1)), \ldots, (t_n, f(t_n)) \in (a, b) \times \mathbb{R}$, trend filtering can quickly provide a flexible, lower-dimensional approximation of the data where the dimensionality is controlled by the choice of the hyperparameter γ. In this context γ is a subjective choice that specifies the amount of (lossy) compression desired—unrelated to the discussion in Section 2.3.5. For any given choice of γ, let p be the number of knots selected by the trend filtering estimator. The corresponding continuous-time representation of this lower-dimensional approximation is fully encoded by the knot locations and the sparse basis vector with $p + k + 1$ nonzero entries, which can be stored efficiently. The falling factorial basis then serves as the "dictionary" from which the continuous-time representation can be losslessly recovered. Gridded uncertainty measurements for the reduced observations can also be computed and stored, though not in a sparse format, via the methods discussed in Section 2.3.6.

3.4 Concluding remarks

In order to illustrate the broad utility of trend filtering to astronomy, we demonstrated its promising performance on a wide variety of problems across time-domain astronomy and astronomical spectroscopy. We studied the Lyman-α forest of quasar spectra with the most depth — using

trend filtering to map the relative distribution of neutral hydrogen in the high redshift intergalactic medium along quasar-observer lines of sight. Furthermore, we discussed how trend filtering can be used to (1) generate galaxy, quasar, and stellar spectral templates and estimate emission-line parameters; (2) produce nonparametric models for exoplanet transit events in phase-folded stellar light curves, providing estimates for the transit depth and total duration; (3) improve upon the `polyfit` algorithm utilized by the *Kepler* Eclipsing Binary via Artificial Intelligence (EBAI) pipeline for denoising phase-folded eclipsing binary light curves (as a preliminary step to estimating the physical parameters); (4) generate supernova light-curve templates and produce nonparametric estimates of the maximum apparent magnitude, the time of maximum, and the decline rate; (5) quickly and efficiently compress large collections of one-dimensional data sets. Naturally, we expect trend filtering will find uses in astronomy beyond those that we explicitly discussed.

Chapter 4

Three-dimensional cosmography of the high redshift Universe using intergalactic absorption: Early Investigations

This chapter is based on my Advanced Data Analysis (ADA) / Data Analysis Project (DAP) report titled *Exploring the Intergalactic Medium* and my thesis proposal titled *Multi-resolution Regression, Divide and Conquer Risk Estimation, and the Large-scale Universe*, which were submitted in Spring 2016 and Spring 2017, respectively, in accordance with the requirements of the joint Doctor of Philosophy degree in Statistics and Machine Learning at Carnegie Mellon University. This chapter also includes material from our article *Mapping the Large-Scale Universe through Intergalactic Silhouettes* [159], which was published in *CHANCE*.

In this chapter we discuss spatial statistical methods for reconstructing the three-dimensional matter density distribution of the intergalactic medium using the Lyman-α forest absorptions observed in closely-sampled sightlines of high redshift quasars — a problem that is commonly referred to as *Lyman-α forest tomography*. This work may therefore be viewed as an extension of the problem discussed in Chapter 3.2, where we detailed a novel approach for reconstructing the

matter density distribution of the intergalactic medium along one-dimensional quasar sightlines. This chapter also serves as a precursor to Chapter 5, which details our more recent work on this topic.

4.1 The Lyman-α forest

A majority of the baryonic matter in the Universe takes the form of a highly dilute gaseous medium that permeates the overwhelming volume of intergalactic space. Dubbed the *intergalactic medium* [IGM; 160], this ubiquitous gas is too diffuse to be directly observed in emission, but its presence is traced by absorptions in the light of luminous background sources — most notably, quasars. Quasars are distant, extremely luminous cosmological objects powered by supermassive black holes [161]. The immense gravity of these black holes and the friction of the matter falling inward result in an outpouring of thermal radiation from outside the event horizon that shines anywhere between 10 and 10,000 times brighter than the entire Milky Way. This thermal radiation spans the electromagnetic spectrum and allows cosmologists to study the IGM by analyzing the absorption patterns in the observed spectra left by the chemical elements present in intergalactic space. As light travels from a distant quasar along its path to Earth, the IGM leaves an absorption signature in the light, marking the atomic elements that are present in the intervening intergalactic gas at each point along the light's path [11]. This signature collectively reveals the presence of diffuse primordial hydrogen and helium residue in intergalactic space left over from the Big Bang, as well as a variety of metals occasionally ejected from galaxies by particularly forceful supernovae explosions [12]. However, the bulk of the IGM is composed of electrically neutral hydrogen (H I) gas, which marks its presence by absorbing a very specific wavelength of light: the so-called Lyman-α (Lyα) wavelength (1215.67 Å). A Lyα absorption occurs when a photon at the Lyα wavelength hits a H I atom in the gaseous IGM, is absorbed,

and sends the electron from the neutral ground state ($n = 1$) to the first excited state ($n = 2$), whence the photon is then reemitted in a random direction in three-dimensional space. This *scattering* of photons results in a decrease in the flux of the quasar continuum (the emitted radiation) at the Lyα spectral line, where flux is a measure of the amount of energy traversing a two-dimensional surface per unit time per unit area. Mathematically, the observed flux at wavelength λ is defined as

$$f(\lambda) = \frac{\partial^2 Q_e(\lambda)}{\partial t \partial A}, \tag{4.1}$$

where $Q_e(\lambda)$ is the radiant energy received at wavelength λ, t is time, and A is the two-dimensional surface area traversed by the radiation, e.g. the collecting area of a telescope. Because electrically neutral hydrogen constitutes the bulk of the baryonic composition of the IGM, the Lyα forest — the series of absorptions originating from the Lyα transition — is therefore the richest source of observational data for cosmologists to study the large-scale structure of the IGM [113, 162–164].

The original detection of the Lyα forest dates back to 1970 [163], but scientific studies of the forest did not mature until the early 1990s with the advent of high-resolution spectrometers — an instrument that connects to a telescope and splits the observed radiation into a continuous spectrum that shows the intensity at each wavelength. In particular, the commissioning of the High Resolution Echelle Spectrometer [HIRES; 165–167] on the Keck telescope in Mauna Kea, Hawaii marked the beginning of the golden age of Lyα forest cosmology.

Figure 4.1 shows a simulated electromagnetic spectrum of a quasar ∼10.9 billion lightyears from Earth [85]. The orange dashed curve shows the quasar continuum — the intensity of the light at each wavelength when it was emitted from the quasar — and the black (wiggly) curve shows the intensity of the light when observed from Earth. The decrease in the intensity of the quasar's light in this region of the electromagnetic spectrum is due to intergalactic H I gas partially absorbing and scattering the light passing through it. This phenomenon is analogous to observing a distant

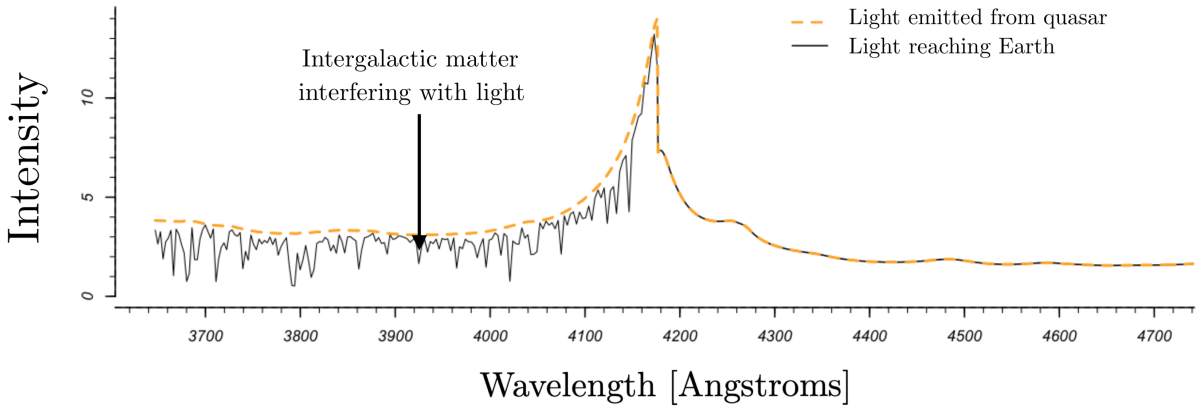


FIGURE 4.1: Simulated electromagnetic spectrum of a quasar ∼10.9 billion lightyears from Earth [85]. The orange dashed curve shows the intensity of light at each wavelength at the time that it was first emitted by the quasar and the solid black curve shows the spectrum as observed from Earth, after H I gas absorption in the intergalactic medium. The Lyα absorptions appear over a series of wavelengths because of the constant doppler-shifting of lightwaves traversing intergalactic space caused by the expansion of the Universe — leading to the so-called *Lyman-α forest*.

lighthouse through a patchy cloud of fog. That is, when seen through a dense patch, the light is dim. When seen through a relatively thin patch, the light is bright. The intergalactic "fog" in this case is too diffuse to be directly observed, but studying its matter density distribution is fortuitously made possible by analyzing the fraction of light that is absorbed along its line of sight to Earth. The fraction of absorbed light is nonlinearly related to the H I density — the latent (i.e. unobservable) quantity of interest, but the relation is monotonic which allows the absorption fraction to serve as a suitable proxy for the H I density. In practice, the intensity of the unabsorbed light originally emitted from the quasar (orange curve) is not known, and accurately estimating this curve represents a crucial step in any statistical analysis of the Lyα forest. The literature on this topic is extensive and includes a wide variety of approaches. These include principal components analyses over a set of candidate functional shapes (i.e. spectral templates) [66, 130–134], interpolation of observed regions deemed to have minimal absorption [168], methods based on low-order polynomials [128, 129], functional regression [169], and nonparametric smooths of the observed spectrum subsequently scaled to match the cosmic mean absorption fraction [31, 35] — with the latter being our own work discussed in Section 3.2.3.

As previously stated, H I gas absorbs at a fixed wavelength of 1215.67 Å. Figure 4.1 should therefore provoke a couple questions. Why do the absorptions appear at a series of wavelengths? And, why are those wavelengths significantly longer than 1215.67 Å? The answer to both questions is that the Universe is expanding. When the Universe began with the Big Bang ∼13.8 billion years ago, the fabric of space began an outward expansion and has continued to expand ever since that moment. This expansion leads to a fascinating phenomenon known as *redshift*. Namely, when light travels through intergalactic space the expansion of the space it is traveling through actually causes the light itself to be continuously doppler-shifted to longer wavelengths. For a (fixed) absorption line such as Lyα, the consequence is that a forest of absorptions becomes inscribed in the light's spectrum in a similar manner to how a seismograph operates. Although the pen of the seismograph is stationary, the paper continuously moves underneath it, recording the amplitude of the pen's oscillations as a one-dimensional curve. Here, redshift is the mechanism that moves the paper. And the resulting forest effectively provides a one-dimensional map of the H I density along the path from Earth to the quasar (provided one can produce an accurate estimate of the quasar continuum), with longer wavelengths corresponding to the H I density at more distant points along the path. Given a spectrum of an extragalactic object, the redshift of the object is defined as the relative scale factor by which the emitted light has been stretched by the moment of observation

$$z = \lambda_{\rm obs}/\lambda_{\rm emit} - 1, \tag{4.2}$$

and is proportional to the radial distance of the object (see Figure 4.2). In practice, the redshift of extragalactic objects can be inferred by identifying prominent emission lines in the observational spectrum. For example, if the Lyα emission peak of an observed quasar appears at wavelength $\overline{\lambda}_{\rm obs}$, the redshift of the quasar is given by

$$z = \overline{\lambda}_{\rm Ly\alpha}/\lambda_{\rm Ly\alpha} - 1, \tag{4.3}$$

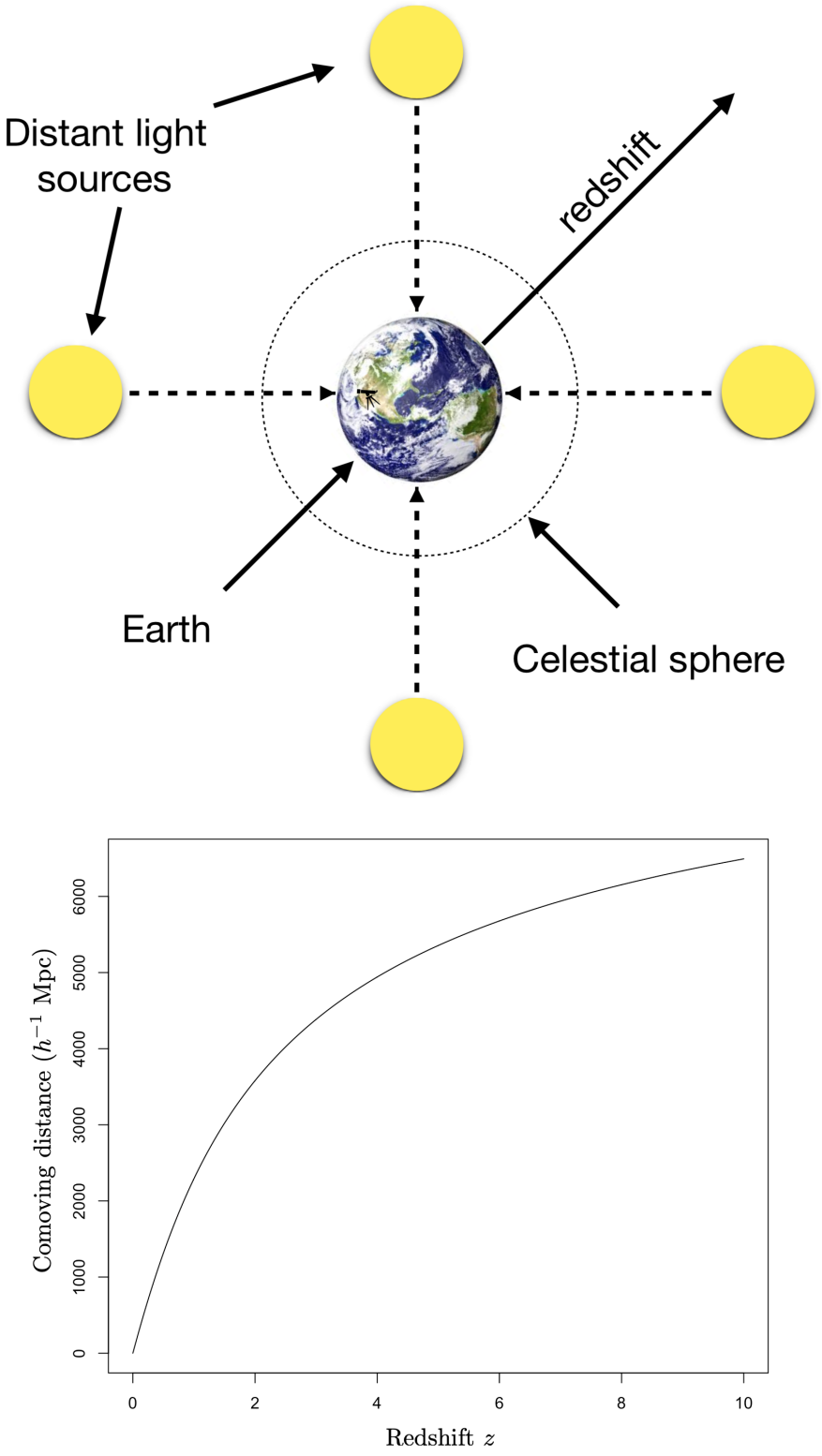


FIGURE 4.2: **Top:** Illustration of the celestial sphere and redshift as a radial coordinate. The redshift of an extragalactic source is nonlinearly, but monotonically related to the radial comoving distance of the source. Celestial coordinate systems are inherently spherical, with observations of redshift surveys recorded in some parametrization of the three-dimensional geometric space $\mathbb{S}^2 \times \mathbb{R}^+$. We utilize the equatorial coordinate system throughout this thesis. **Bottom:** Redshift-distance relation under the modern cosmological model (with *Planck* CMB parameters [170]). Here, comoving distance is the distance between two objects at the present cosmological time and is given in units of h^{-1} megaparsecs (Mpc) where $h = H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})$ with Hubble constant H_0. The effect of redshift increases at greater radial separations because of the increasing recession velocity of objects due to the expansion of the Universe.

where $\lambda_{\mathrm{Ly}\alpha} = 1215.67$ Å is the restframe Lyα wavelength.

Another important observational phenomenon arises due to the vast distances quasars can be observed at and the finite speed at which their light travels to us. Namely, the light we observe from quasars is actually extremely old, with its age being directly related to the distance it traversed on its path to Earth. For example, an observed spectrum of a quasar 10.9 billion lightyears from Earth — such as that displayed in Figure 4.1 — is in fact a signature of what that quasar looked like 10.9 billion years ago — 6.4 billion years before the Earth even formed. Mapping the matter distribution of the IGM via Lyα absorptions in the spectra of distant quasars is therefore akin to charting how our adolescent Universe evolved into its present-day form.

Large-scale cosmological hydrodynamical simulations provide illustrations of the Universe's large-scale structure [171–174], both via its point process matter distribution (galaxies and quasars) and its continuous intergalactic medium matter distribution. Both processes arise as a result of gravity and dark energy acting on the primordial matter fluctuations left over from the nearly perfect, scale-invariant Gaussianity of the Big Bang temperature fluctuations, as revealed by the Cosmic Microwave Background radiation [CMB; 170, 175–180]. The mutual gravitational forces of the matter have subsequently pulled the Universe's large-scale structure into a highly non-uniform weblike distribution — appropriately dubbed the *cosmic web*. Dark matter — matter that thus far only reveals its presence by way of its gravitational force — is also a central acting force in the modern Big Bang cosmological model [ΛCDM; 181, 182] and possesses a weblike distribution closely tracing the baryonic matter distribution [183, 184].

4.2 Lyman-α forest tomography

The past decade of Lyα forest observational cosmology has progressed to a point where it is now, in principle, possible to use the aggregate of currently available Lyα forest sightlines to

statistically reconstruct a continuous large-scale structure map of the IGM in all three dimensions — a problem commonly known as *Lyman-α forest tomography* [21–25]. Mapping the IGM in three dimensions can be viewed as an extension of charting the locations of the galaxies and quasars that the IGM envelopes, in the sense that the discrete distribution of galaxies and quasars and the continuous distribution of intergalactic gas together provide a complete picture of how baryonic matter is distributed throughout the Universe. Scientifically, a large-scale IGM map will allow for testing of cosmological models in an entirely new regime of both scale and age of the Universe, potentially leading to a more refined constraints on the parameters that define the ΛCDM cosmological model. Moreover, because the H I gas traces the total distribution of gravitating matter on large scales, a large-scale map of the IGM will allow us to locate many never-before-seen objects such as galaxy protoclusters and cosmic voids [185].

Certainly, the greatest challenge in this sort of spatial modeling — as is the case in the majority of statistical analyses — is not in producing the point estimate itself, but rather accompanying the point estimate with reliable statistical inference procedures. For any given predicted structure in the map — e.g. a galaxy protocluster or a cosmic void — what is the probability that the structure indeed exists? Is there enough signal in the map to detect significant cross-correlations with the anisotropies of the CMB [186], from which the IGM evolved? Does the established inference adequately account for the uncertainty introduced by the sequential observational pipeline of analyses the data have already undergone? All of these are exceedingly difficult questions that rely on the development of rigorous statistical inference. Our initial investigations in this chapter are light on statistical inference, but our more recent work on this topic — discussed in Chapter 5 — is accompanied with significantly more statistical rigor.

At the time this thesis work was carried out, the most prolific Lyα quasar catalog to date was compiled by the Baryon Oscillation Spectroscopic Survey [BOSS; 2], which operated from 2008 to 2014 within the third phase of the Sloan Digital Sky Survey [SDSS-III; 26, 109]. During that time

BOSS measured the spectra of 194,240 high redshift ($z \gtrsim 2.1$) quasars in two large contiguous regions in the Northern and Southern Galactic Caps. In both this chapter and the next, we utilize the BOSS Lyα quasar catalog (discussed further in the section below) to produce our large-scale structure map of the IGM. Beyond the groundbreaking work of the BOSS collaboration, the rapid influx of Lyα data is assured to continue into the foreseeable future. The next generation of the BOSS collaboration, the extended BOSS survey [eBOSS; 187, 188] began gathering data immediately following the completion of the SDSS-III phase, and is preparing its final data release at the time of this writing. eBOSS, however, primarily focused on adding low-z non-Lyα quasars to complement the catalog of Lyα quasars observed by its predecessor, and therefore would not significantly improve our high redshift intergalactic medium map. The Dark Energy Spectroscopic Instrument [DESI; 189] Survey conducted on the Mayall 4-meter telescope at Kitt Peak National Observatory in Arizona very briefly began its 5-year data collection in March of this year before being put on hiatus due to the COVID-19 pandemic. DESI will measure spectra for approximately 700,000 Lyα ($z \gtrsim 2.1$) quasars over a sky area of 14,000 square degrees (50 quasars per square degree). The COSMOS Lyman-α Mapping And Tomography Observations [CLAMATO; 24] survey using the Low Resolution Imaging Spectrograph (LRIS) on the Keck-I telescope at Mauna Kea, Hawaii is actively collecting spectra of bright star-forming galaxies over comparatively smaller regions than BOSS, eBOSS, and DESI, but with a much more densely populated target selection on the sky, allowing for higher resolution mappings. Moreover, though not a Lyα survey, the Square Kilometer Array [SKA; 190, 191] has begun construction on the world's largest radio telescope, with which it will pioneer the collection of observational data from the so-called Dark Ages, using a different region of the electromagnetic spectrum — the 21 cm radio emission. Mapping this completely uncharted epoch will utilize the same variety of statistical methods developed for Lyα forest tomography. See, for example [192, 193], for discussions on *21 cm tomography*.

4.3 The Baryon Oscillation Spectroscopic Survey

The Baryon Oscillation Spectroscopic Survey [BOSS; 2] operated from 2008 to 2014 within the third phase of the Sloan Digital Sky Survey [SDSS-III; 27] and was designed to measure the scale of baryon acoustic oscillations [BAOs; 194] — a feature imprinted on matter clustering by acoustic waves that propagate through the primordial plasma in the pre-recombination Universe — at a variety of redshifts, allowing for constraints on the ΛCDM model and a path to understanding dark energy [195]. At low redshifts ($z < 0.7$), BOSS utilized the point process distribution of luminous galaxies [galaxy clustering; 196–218] to measure the BAO scale and constrain cosmological parameters. Quasars can be observed at higher redshifts because of their extreme luminosity, which the BOSS collaboration exploited to study redshift $2.15 < z < 3.5$ cosmological matter distribution, measuring the BAO scale and constraining cosmological parameters through both the point process distribution of observed quasars [quasar clustering; 219, 220] and their Lyα forest absorption [13, 15, 18, 19, 221, 222], as well as cross-correlations of the various matter tracers [223–226].

The spatial distribution of the galaxies and quasars observed by BOSS is shown in Figure 4.3, with galaxies shown in black and quasars shown in red. In order for the ultraviolet Lyα forest absorptions in the quasar spectra to be observed by ground-based spectrographs such as those used by the BOSS, the Lyα wavelength must be redshifted by a factor of $\gtrsim 3$ since most ultraviolet light is absorbed in Earth's atmosphere. In particular, the spectra of $z \gtrsim 2.1$ quasars in the BOSS catalog provide useful observations for analyzing the intergalactic medium via the Lyα forest and we therefore refer to them as Lyα quasars. The consequence of this lower bound on obtaining Lyα forest observations from the ground-based spectrographs is that the $0 < z \lesssim 2$ intergalactic medium is currently inaccessible to us, and our three-dimensional reconstruction is therefore limited to redshifts $z \gtrsim 2$. An upper bound on three-dimensional reconstruction of

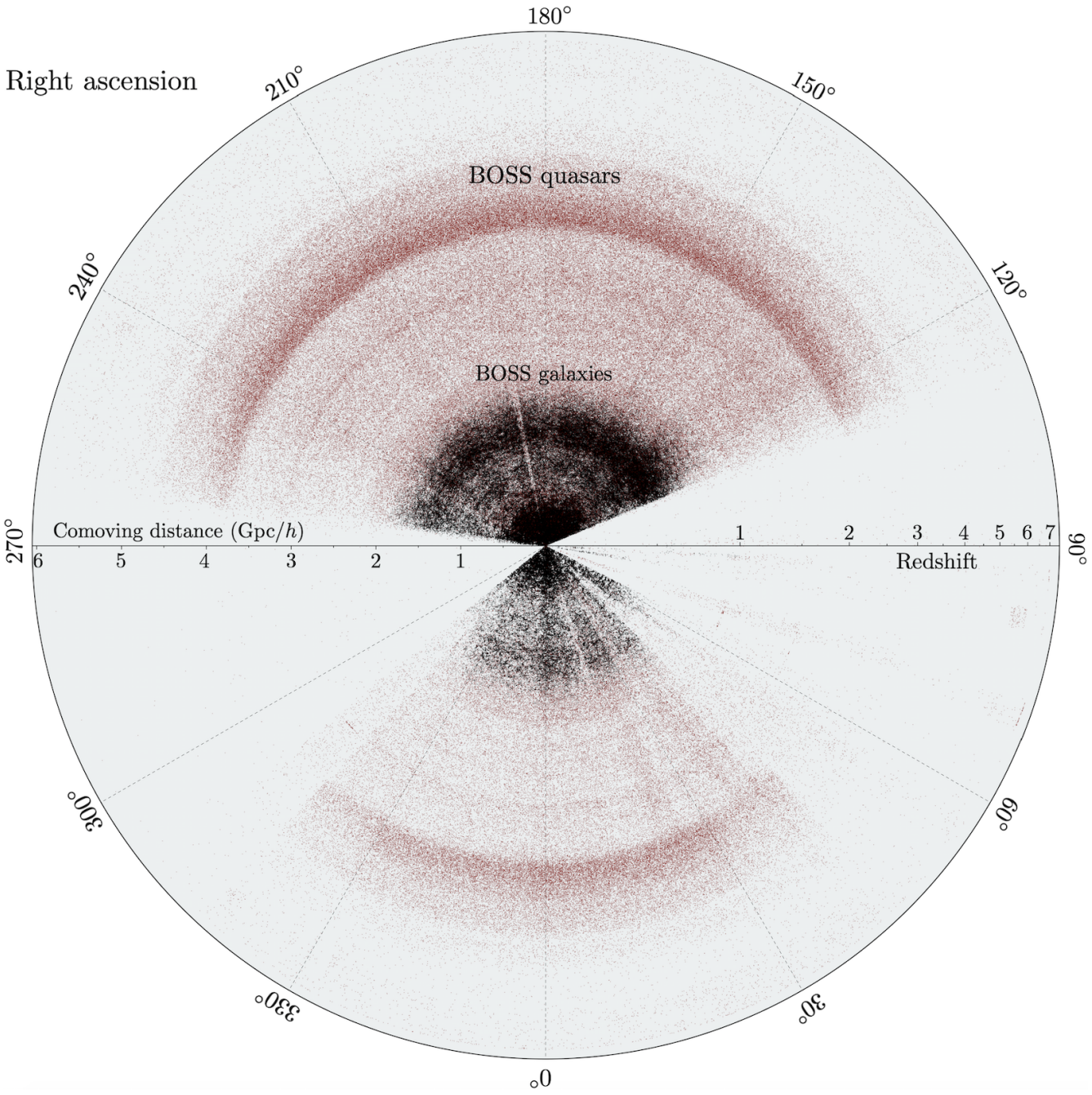


FIGURE 4.3: Earth-centric map of the galaxies (black) and quasars (red) observed by the Baryon Oscillation Spectroscopic Survey [BOSS; 27], in a right ascension-comoving distance projection. Only redshift $z \gtrsim 2.1$ quasars are suitable for Lyα forest analyses since the forest is still in the ultraviolet range at lower redshifts and therefore inaccessible to ground-based telescopes.

the intergalactic medium also arises at higher redshifts as quasars become increasingly faint and difficult to detect. In this chapter we pursue a reconstruction over the redshift range $1.95 \leq z \leq 3$. In Chapter 5, we revise this range to $1.98 \leq z \leq 3.15$.

In Figure 4.4, we show the Earth-centric spatial distribution of BOSS Lyα quasars within a 5° declination window along the celestial equator (declination $\delta = 0°$) and in Figure 4.5 we show the full three-dimensional distribution of Lyα quasars observed by BOSS. The BOSS footprint (i.e. sky coverage) is primarily focused on two contiguous regions of the sky (shown in Figure 4.6), with the principal criterion for selecting these regions being to minimize attenuation due to interstellar dust (within the Milky Way). The BOSS catalog constitutes a coverage of 10,400 deg^2 (approximately 25% of the sky).

As discussed in Section 3.2, each quasar Lyα forest provides a dense series of observations lying along a one-dimensional sightline that is foreground to the quasar. In Figure 4.7, we show again the BOSS Lyα quasars along the celestial equator, now with their foreground sightlines superposed. In Figure 4.8 we show a three-dimensional rendering of a small sample of Lyα sightlines centered at equatorial coordinates $(\alpha, \delta) = (0°, 0°)$, with the Cartesian coordinate axes given in units of h^{-1} Mpc, where $h = H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})$ and H_0 is the Hubble constant. The goal of Lyα forest tomography is to construct a three-dimensional spatial model that takes in these line of sight observations and reconstructs the full three-dimensional absorption field. Prior to three-dimensional modeling the sightline observations must first be propagated to the flux contrast scale that directly traces the H I gas density, which we discuss in the section below.

4.4 Transforming spectra to flux contrast

Three-dimensional reconstruction of Lyα forest absorption fields relies crucially on the preliminary step of estimating the latent unabsorbed continuum of each coadded quasar Lyα forest and

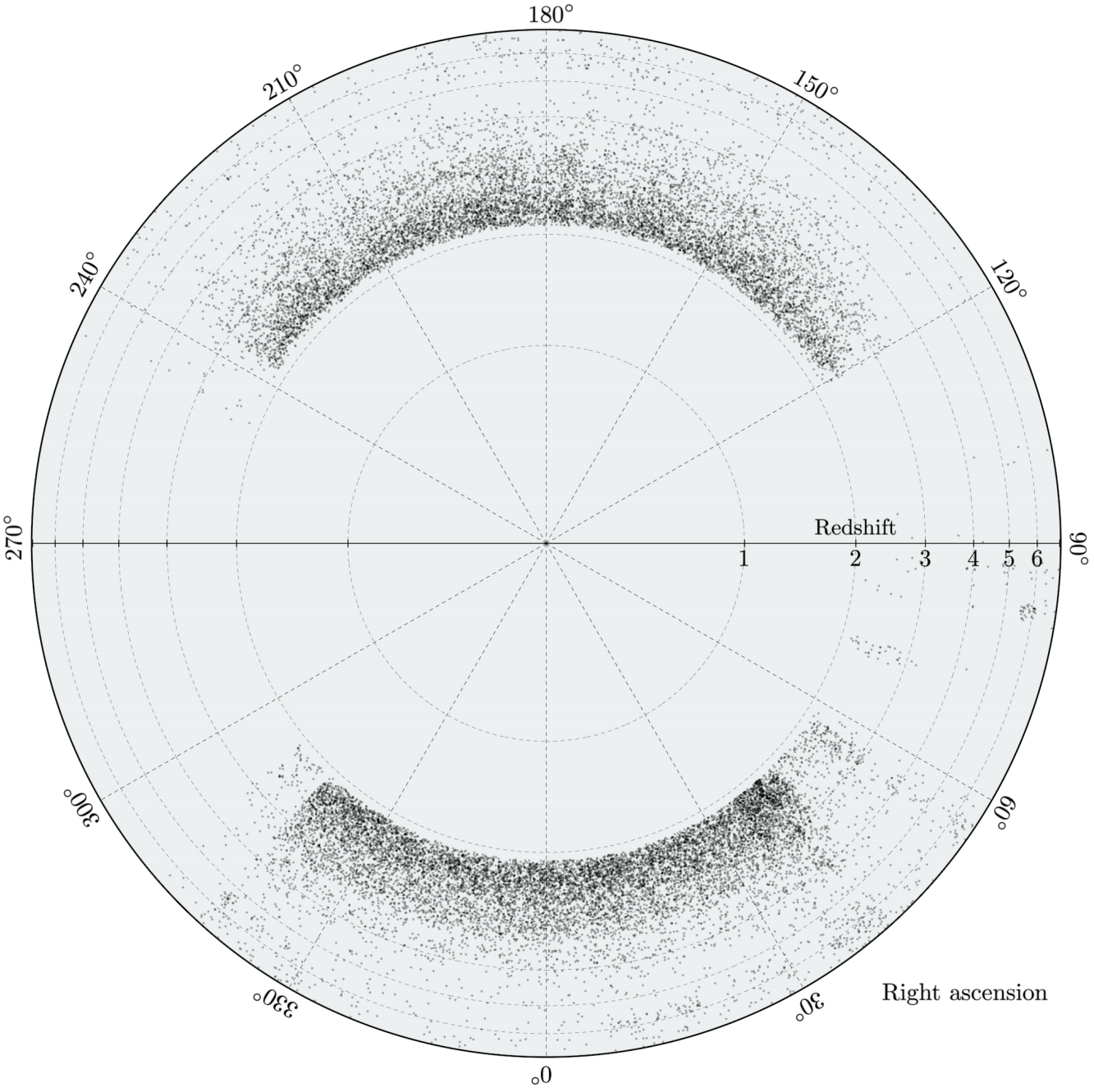


FIGURE 4.4: RA-redshift distribution of the BOSS quasars at sufficiently high redshifts for Lyα forest analysis — i.e. Lyα quasars. Here, we show all Lyα quasars along the celestial equator (declination $\delta = 0°$) plus or minus 2.5°, with the scale being linear in comoving distance. The quasars are primarily located in two contiguous regions on the celestial sphere — one in the Northern Galactic Cap (top) and one in the Southern Galactic Cap (bottom).

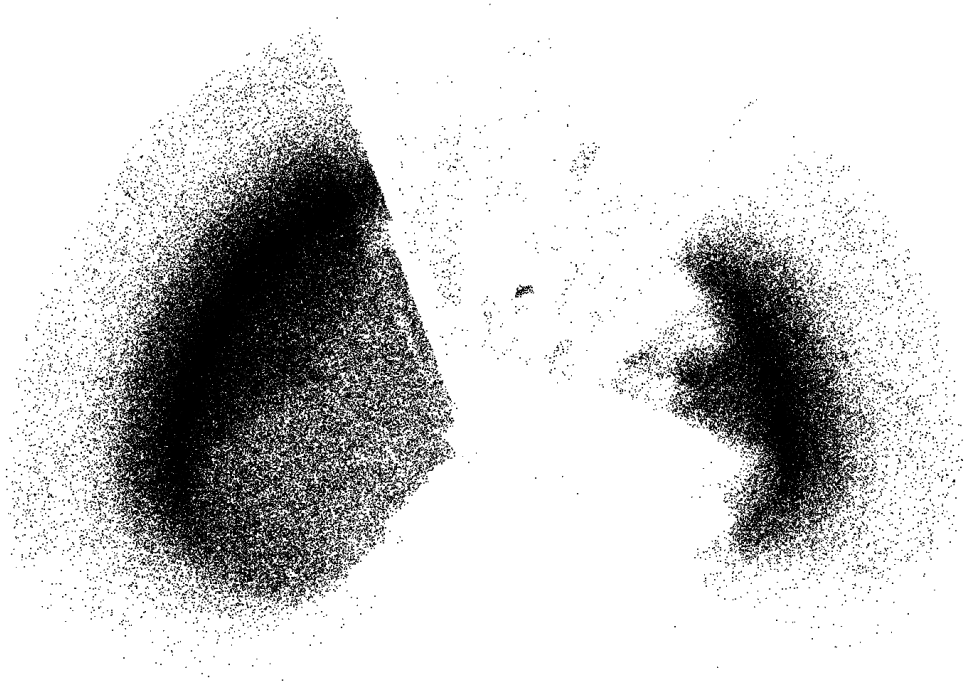


FIGURE 4.5: Three-dimensional distribution of BOSS Lyα quasars. The Lyα forest is not accessible to ground-based telescopes at redshifts $0 \leq z \lesssim 1.92$ due to atmospheric opacity to ultraviolet wavelengths, therefore producing a spherical void in the volume over which we can study the intergalactic medium. The density of observed quasars also degrades on the high redshift end as they become increasingly faint to observers on Earth.

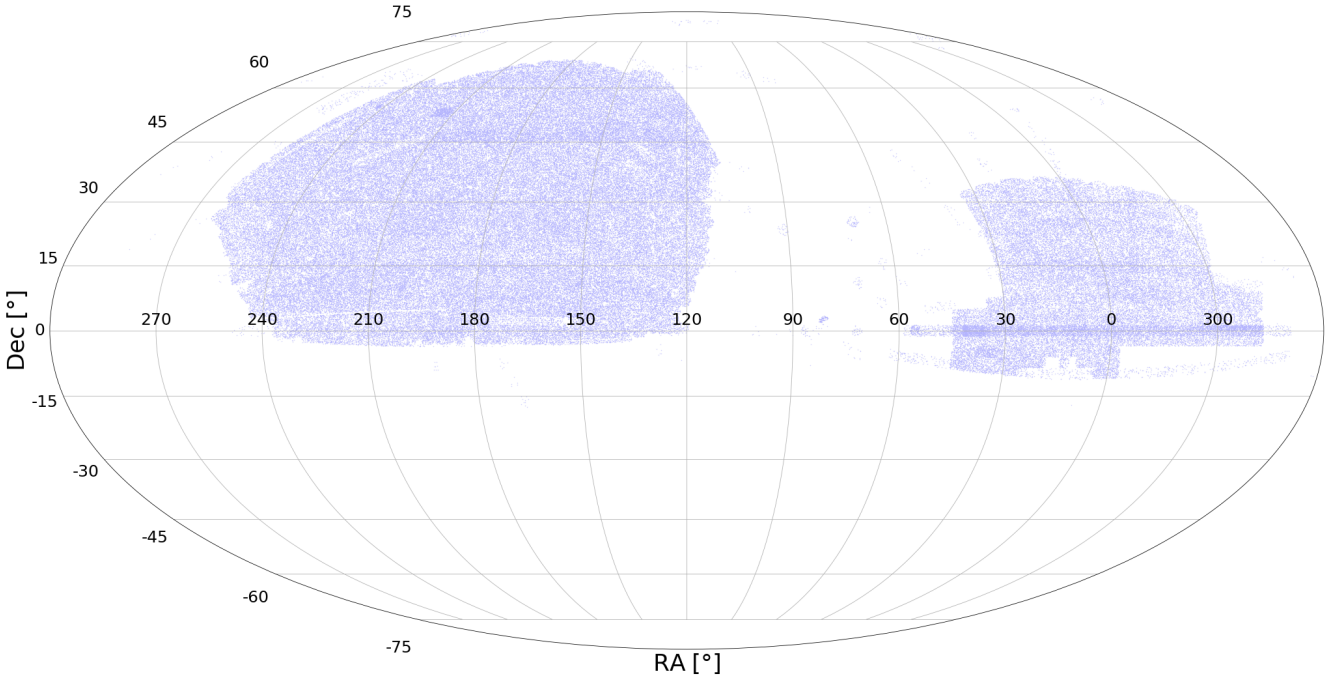


FIGURE 4.6: Sky distribution of BOSS Lyα quasars, shown in equatorial coordinates. The total footprint is 10,400 deg^2 ($\sim$25% sky coverage) and the total number of distinct $z \geq 2.1$ quasars is 208,360 ($\sim$20 per sq. degree).

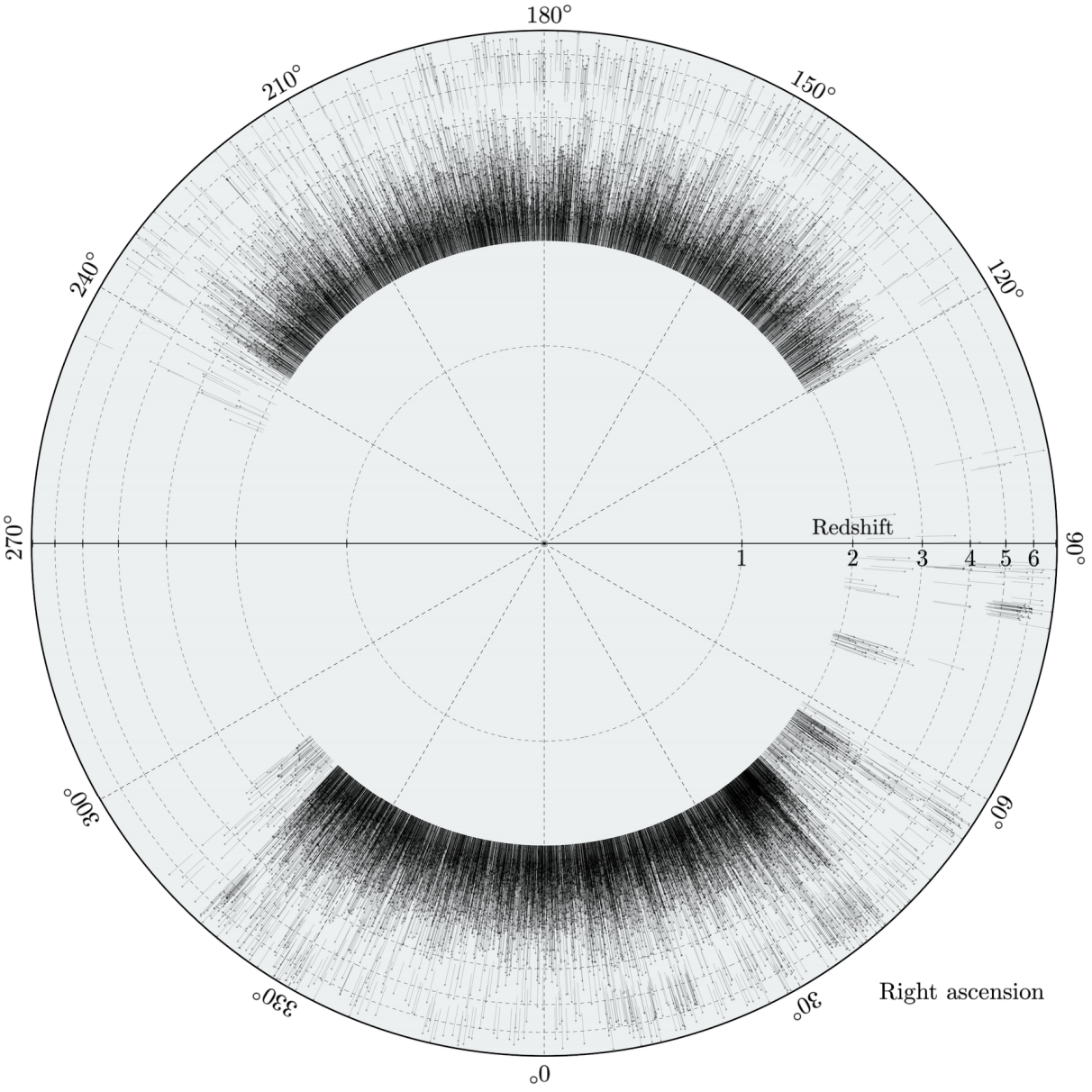


FIGURE 4.7: Collection of BOSS Lyα sightlines observed along the celestial equator. The sightlines terminate at $z \sim 1.92$ on the low redshift end due to the opacity of Earth's atmosphere to ultraviolet wavelengths. On the high redshift end the sightlines becomes too sparse to allow for three-dimensional reconstruction of the full absorption field.

the broader procedure for transforming the raw flux observations to the Lyα flux contrast scale that traces the density fluctuations of H I gas in the intergalactic medium. Our work on three-dimensional absorption field reconstruction in this chapter predated our published work on modeling one-dimensional Lyα forest absorption fields in Chapter 2, so the procedure we use here for transforming the spectra to the flux contrast scale is a more primitive approach. In Chapter 5, which details our most recent work on three-dimensional mapping the intergalactic

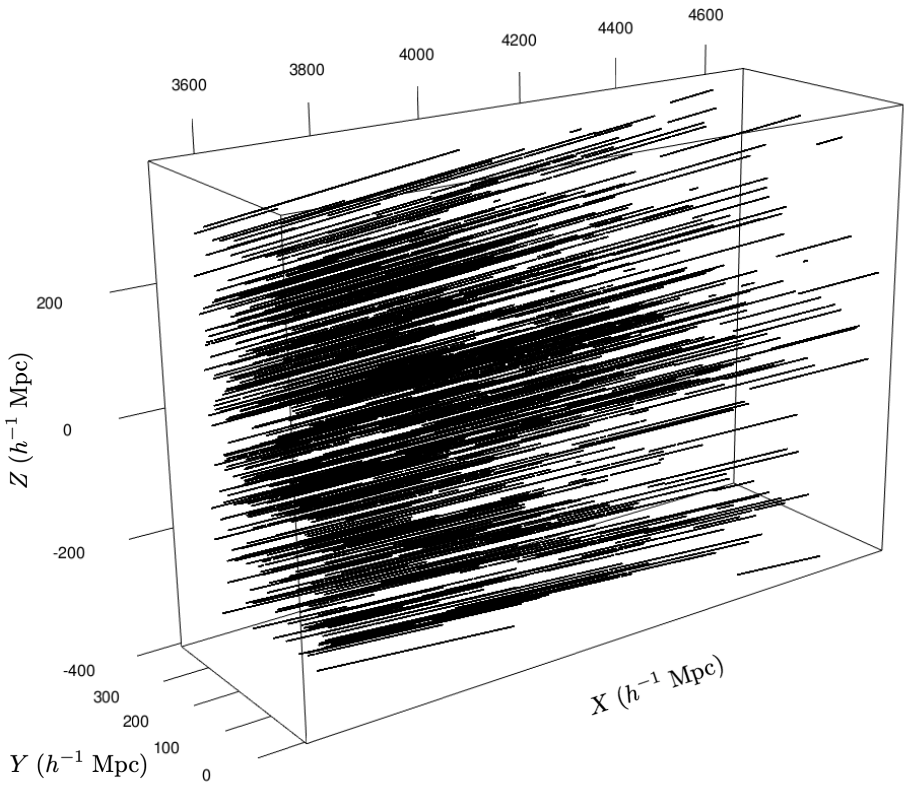


FIGURE 4.8: Sample of Lyα forest sightlines centered at equatorial coordinates $(\alpha, \delta) = (0°, 0°)$, with Cartesian axes (in comoving h^{-1} Mpc). The goal of Lyα forest tomography is to reconstruct the full three-dimensional Lyα absorption field by smoothing the sample of one-dimensional sightlines.

medium, we make use of the one-dimensional pipeline detailed in Chapter 3.

Suppose we observe a BOSS Lyα quasar located at redshift $z = z_0$. The observational data generating process (DGP) of the quasar flux in the Lyα forest can be assumed to follow the model

$$f(\lambda_i) = f_0(\lambda_i) + \epsilon_i, \qquad \lambda_i \in \Lambda(z_0), \tag{4.4}$$

$$= \overline{F}(\lambda_i) \cdot C(\lambda_i) \cdot \big(1 + \delta_F(\lambda_i)\big) + \epsilon_i, \tag{4.5}$$

where $f(\lambda_i)$ is the coadded flux at wavelength λ_i, $f_0(\cdot)$ is the flux signal, $\{\epsilon_i\}_{i=1}^n$ are uncorrelated Gaussian measurement errors attributable to photon noise, CCD readout noise, and sky-subtraction error (assumed $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma_i^2$), $\{\lambda_i\}_{i=1}^n$ form an equally spaced

wavelength grid in log-space with $\Delta \log_{10}(\lambda_i) = 10^{-4}$ dex log-Angstroms, z_0 is the redshift of the quasar, $C(\cdot)$ is the flux of the unabsorbed quasar continuum, $F(\cdot) = f_0(\cdot)/C(\cdot)$ is the transmitted flux fraction, $\overline{F}(\lambda_i) = \mathbb{E}[F(\lambda_i)]$ is the mean Lyα transmitted flux fraction at redshift $z_i = \lambda_i/\lambda_{\text{Ly}\alpha} - 1$, and $\delta_F(\lambda_i) = F(\lambda_i)/\overline{F}(\lambda_i) - 1$ is the transmitted flux contrast at redshift $z_i = \lambda_i/\lambda_{\text{Ly}\alpha} - 1$. The flux contrast δ_F inversely traces H I density fluctuations in the intervening intergalactic medium and, by definition, is mean zero across the sky at each fixed redshift, with negative contrasts corresponding to H I densities above the cosmic mean and positive contrasts corresponding to H I densities below the cosmic mean.

Here we estimate the mean flux level $m(\lambda) = \overline{F}(\lambda) \cdot C(\lambda)$ by fitting a smooth local polynomial regression [LOESS; 120–122] directly to the pixels in the observed wavelength frame, weighting the pixels by the inverse measurement variance provided by the BOSS spectroscopic pipeline [66]. For each Lyα sightline, we then define the flux contrast estimates to be

$$\widehat{\delta}_F(\lambda_i) = \frac{f(\lambda_i)}{m(\lambda_i)} - 1 - \text{bias}, \qquad i = 1, \ldots, n, \tag{4.6}$$

where $m(\cdot)$ is the LOESS estimate with bandwidth 300 Å fit to the flux observations $f(\lambda_1), \ldots, f(\lambda_n)$ and bias = 0.071291 is a scalar bias term that we utilize here so the aggregated set of flux contrast estimates empirically match the theoretical mean zero property of the defined flux contrast. Here, the 300 Å bandwidth is a purely heuristic choice we made based on what seemed to visually produce reasonable estimates of the mean flux level over a large sample of quasar sightlines. Recall we take a much more data-driven approach in Chapter 3. The LOESS estimation of the mean flux level in the Lyα forest is illustrated in Figure 4.9.

Unlike our more recent work, here we do not track the additional statistical uncertainty introduced into the flux contrast estimates by the stochastic transformation involving the mean flux estimator. Given the measurement uncertainties on the observational scale σ_i^2, we simply take the LOESS

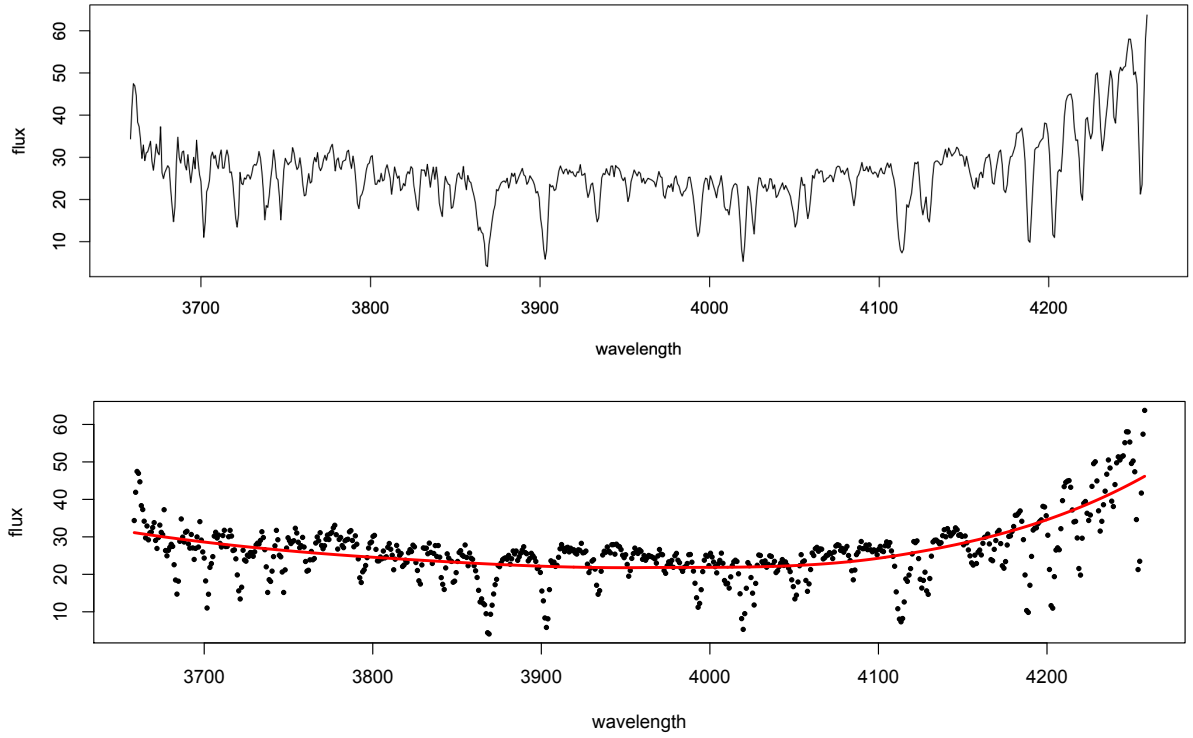


FIGURE 4.9: **Top:** Observed Lyα forest of a BOSS quasar located at RA = 12.02527, Dec = -1.05598, $z = 2.5338 \pm 0.00013$. **Bottom:** Pixelization of the Lyα forest with a low order LOESS estimate of the mean flux level — the product of the quasar continuum and the mean Lyman-α flux transmission at each redshift. The flux contrast estimates are then defined as $\widehat{\delta}_F(\lambda) := f(\lambda)/m(\lambda) - 1 - \text{bias}$, where $f(\cdot)$ is the observed flux, $m(\cdot)$ is the low order LOESS smooth, and bias = 0.071291 is a scalar bias term that scales the aggregated sample of flux contrast estimates to be mean zero.

estimate of the mean flux level to be fixed, yielding measurement variance estimates of

$$\widehat{\text{Var}}(\widehat{\delta}_F(\lambda_i)) = \sigma_i^2/\widehat{m}^2(\lambda_i), \qquad i = 1, \ldots, n, \tag{4.7}$$

on the transformed scale. Figure 4.10 shows the first two moments of the aggregated sample of flux contrast estimates. In the left panel, the binned estimates can indeed be seen to be approximately mean zero at all redshifts as desired. In the left panel, the mean standard error of the propagated estimates is approximately monotonically decreasing as a function of redshift due to the greater observational uncertainty at near-ultraviolet wavelengths.

Figure 4.11 illustrates the relative abundance of Lyα observations in our sample as a function of

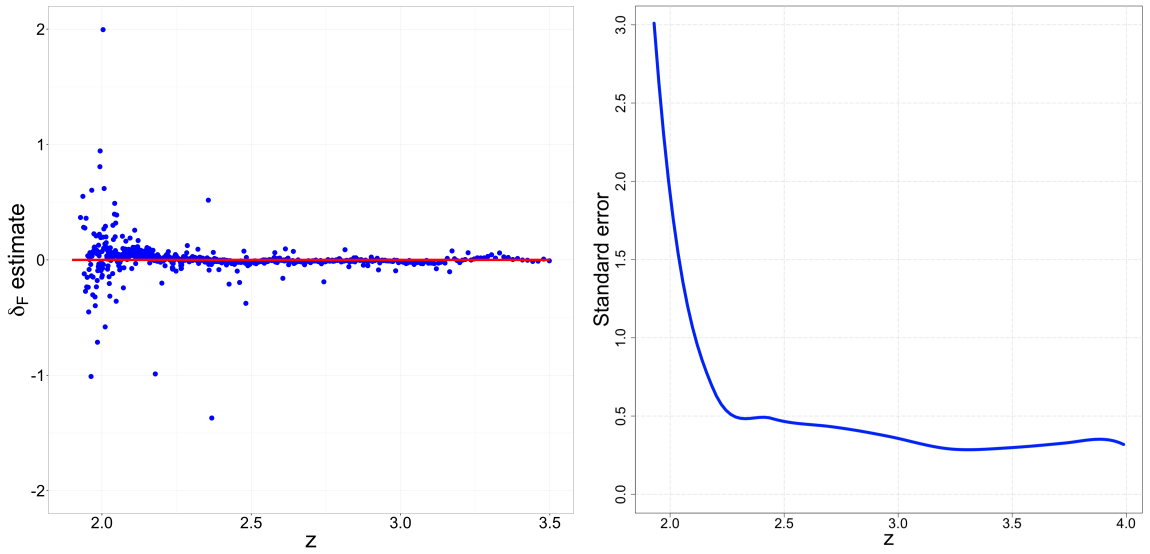


FIGURE 4.10: **Left:** Binned averages of the flux contrast estimates in our sample vs. redshift. Here we include the subtraction of the estimated scalar bias term. The flux contrast, by definition, is mean zero at all redshifts, which are estimates are in reasonable agreement with here. **Right:** Mean standard error of the flux contrast estimates in our sample vs. redshift. The uncertainty in the estimates increases at low redshifts where the Lyα wavelengths are still in the near-ultraviolet.

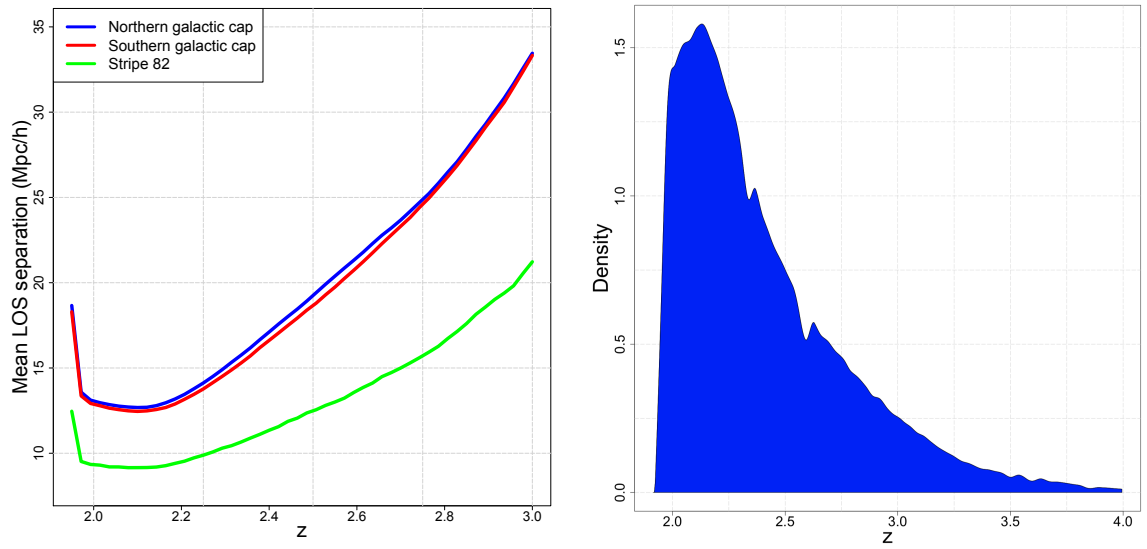


FIGURE 4.11: **Left:** The mean transverse sightline separation (in h^{-1} Mpc) of BOSS Lyα quasars as a function of redshift. This quantity is the primary constraint on the effective spatial resolution at which we are able to reconstruct the three-dimensional density field of the IGM across sightlines. The mean transverse sightline separation is effectively identical for the two contiguous regions that constitute the 10,400 deg^2 footprint. **Right:** Distribution of BOSS Lyα pixels as a function of redshift. In this chapter we limit our spatial reconstruction of the intergalactic medium to the redshift range $1.95 < z < 3$ due to the sparsity of observed quasars at higher redshifts.

redshift. In the left panel we show the mean transverse sightline separation[1] in comoving h^{-1} Mpc. This quantity serves as the primary constraint on the effective comoving spatial resolution on any reconstruction of the full three-dimensional absorption field because we cannot expect to recover structure on scales that are not suitably sampled in the transverse direction. The overall mean transverse sightline separation is approximately identical for the two contiguous regions in the BOSS footprint. However, Stripe 82 — a 220 deg^2 region in the Southern Galactic Cap along the celestial equator (declination $\delta \approx 0°$) — was repeatedly targeted to achieve higher source densities than the rest of the BOSS footprint. This 220 deg^2 region could therefore be mapped at higher spatial resolution than the rest of the sky coverage, but we do not pursue this here. The right panel of Figure 4.11 shows the redshift-wise distribution of the aggregated three-dimensional sample of flux contrast pixels. The sample becomes increasingly sparse at redshifts $z > 3$, so here we limit our focus to the redshift range $1.95 < z < 3$ when producing the three-dimensional reconstruction.

4.5 Methods

In this section we investigate two methods for reconstructing a three-dimensional map of the intergalactic medium from the aggregated set of BOSS Lyα absorptions detailed above. The first method is a multi-resolution Gaussian random field (GRF) model capable of scaling to relatively large ($n \approx 300,000$) spatial data sets [227]. The second is local polynomial regression [120], an approach that [25] recently proposed for Lyα forest tomography. In Section 4.6, we evaluate the performance of each method by attempting to recover the full three-dimensional absorption field of a large hydrodynamical cosmological simulation from a sparse set of quasars sightlines. The density of the sightlines is varied in order to mimic the heterogeneous density of the BOSS Lyα sightlines and we make a comparative assessment — both quantitative and qualitative — at each

[1] The perpendicular separation of the sightlines measured in terms of comoving arc length.

sightline density. We deem the performance of the multi-resolution GRF model to be superior, and we therefore proceed with this model for the reconstruction of the real IGM in Section 4.7.

Our three-dimensional models utilize Euclidean distances, which can be computed using the spectroscopic redshifts of the quasars and the assumed ΛCDM cosmological model. Specifically, for each Lyα flux contrast measurement observed at coordinates $(\alpha_i, \delta_i, z_i)$, we compute the comoving Cartesian coordinates $x_i = (x_i^1, x_i^2, x_i^3)$ according to the following equations:

$$x_i^1 = D_C \cdot \sin(90^\circ - \delta_i) \cdot \cos(\alpha_i) \tag{4.8}$$

$$x_i^2 = D_C \cdot \sin(90^\circ - \delta_i) \cdot \sin(\alpha_i) \tag{4.9}$$

$$x_i^3 = D_C \cdot \cos(90^\circ - \delta_i) \tag{4.10}$$

where the comoving distances are given by

$$D_C = D_H \int_0^{z_i} \frac{dz'}{\sqrt{\Omega_m(1+z')^3 + \Omega_k(1+z')^2 + \Omega_\Lambda}} \tag{4.11}$$

with $D_H = 3000\ h^{-1}$ Mpc and the assumed cosmological parameters $\Omega_m = 0.3$, $\Omega_k = 0$, $\Omega_\Lambda = 0.7$, and $h = H_0/(100\ \text{km s}^{-1}\ \text{Mpc}^{-1})$.

4.5.1 Linear smoothers

In this section we briefly discuss the broad *linear smoother* class of regression estimators. This serves as a fitting preface to our discussion of the LOESS estimator and the multi-resolution GRF model because, although they are motivated from seemingly disparate premises, they both belong to this common class. Recall the usual nonparametric regression setup where we observe

n pairs $(x_i, f(x_i))$, $i = 1, \ldots, n$, according to the data generating process

$$f(x_i) = f_0(x_i) + \epsilon_i, \qquad x_1, \ldots, x_n \in S \subset \mathbb{R}^d, \tag{4.12}$$

where $f(x_i)$ is a noisy measurement of a signal $f_0(x_i)$, $\mathbb{E}[\epsilon_i] = 0$, and $\mathrm{Var}(\epsilon_i) < \infty$. A statistical estimator $\widehat{f_0}$ of the signal f_0 is a linear smoother if, for any $x \in S$, we can write

$$\widehat{f_0}(x) = \sum_{i=1}^{n} \ell_i(x) f(x_i) \tag{4.13}$$

$$= \ell(x)^T f \tag{4.14}$$

for some weight vector $\ell(x)^T = (\ell_1(x), \ldots, \ell_n(x))$. That is, in words, linear smoothers are linear combinations of the observed data — not to be confused with *linear regression*, where we assume $\widehat{f_0}(x) = \beta_0 + x^T\beta$. Here, the weight vector may depend on the observed inputs $x_1, \ldots, x_n$, the location of the prediction x, and potentially *a priori* measurement variances $\sigma_i^2 = \mathrm{Var}(\epsilon_i)$, but not the observations $f(x_1), \ldots, f(x_n)$. Typically, it is desirable that the weight vector sums to one so constant signals are preserved. While nonparametric linear smoothers do not assume a rigid parametric form for the signal itself, each possesses a set of one or more hyperparameters — or *smoothing parameters* — that tune the flexibility of the estimated signal, e.g. kernel bandwidth, smoothing spline penalty coefficient, Gaussian process covariance parameters.

Kernel smoothing is directly motivated by the idea of taking locally-weighted averages of the observations. Given a smoothing kernel K satisfying

$$\int K(x)dt = 0, \quad \int xK(x)dt = 0, \quad 0 < \int x^2 K(x)dt < \infty, \tag{4.15}$$

the Nadaraya-Watson kernel regression estimator [135, 136] is defined via the weight vector

$$\ell^{\text{kernel}}(x) = \left(\frac{K\left(\frac{|x-x_1|}{h}\right) w_1}{\sum_{j=1}^{n} K\left(\frac{|x-x_j|}{h}\right) w_j}, \ldots, \frac{K\left(\frac{|x-x_n|}{h}\right) w_n}{\sum_{j=1}^{n} K\left(\frac{|x-x_j|}{h}\right) w_j} \right)^T \tag{4.16}$$

for some choice of kernel bandwidth $h > 0$. The bandwidth is the model hyperparameter that controls the smoothness of the fit and is typically chosen via an automatic, data-driven method such as K-fold cross validation [49] or minimization of Stein's unbiased risk estimate [228]. The kernel itself is traditionally fixed before fitting and not altered in model validation. Common choices for the kernel include the Epanechnikov, tricube, and Gaussian kernels. [72] showed that the Epanechnikov kernel is the optimal choice in a minimax sense (see Section 2.2.1.1).

The definition of kernel smoothers includes several familiar statistical estimators as special cases. For example, a sliding-mean regression is a kernel smoother with a rectangular (boxcar) kernel. Similarly, k-nearest neighbors is also a kernel smoother with a rectangular kernel, but with a design-adaptive bandwidth. Inverse distance weighting [IDW; 229] is a kernel smoother with the kernel being an inverse polynomial function of the Euclidean distance, modified to force the fit to interpolate through the observed data.

Kernel smoothing is unique among nonparametric linear smoothers in that it is explicitly motivated by the idea of taking locally-weighted averages of the data. The linearity of the other methods arises inadvertently in some sense. The weight vector of a linear smoother is therefore often referred to as the *effective kernel* (or *equivalent kernel*).

4.5.2 Local polynomial regression

Local polynomial regression (also known as locally weighted scatterplot smoothing, or LOESS; [120]) is the natural extension of kernel smoothing to higher-order local polynomials. For simplicity, let us first assume the dimension of the feature space is $d = 1$. Given a smoothing

kernel $K(\cdot)$ with bandwidth $h > 0$, the LOESS estimate at input x is obtained by minimizing

$$\sum_{i=1}^{n} \Big(f(x_i) - \phi_x(x_i; \beta_0, \ldots, \beta_p) \Big)^2 w_i \cdot K\left(\frac{|x_i - x|}{h} \right), \tag{4.17}$$

where $\phi_x(x_i; \beta_0, \ldots, \beta_p) = \beta_0 + \beta_1(x_i - x) + \cdots + \beta_d(x_i - x)^p/p!$ is a pth degree polynomial centered at x and $w_i = \widehat{\sigma}_i^{-2}$ (if such estimates are available). The LOESS estimate at input x is then given by $\widehat{f_0}(x) = \widehat{\beta}_0$. Setting $p = 0$ returns the kernel smoothing estimator.

More succinctly, the LOESS estimator is characterized by the effective kernel

$$\ell^{\text{LOESS}}(x) = e_1^T (\Phi_x^T W_x \Phi_x)^{-1} \Phi_x^T W_x \tag{4.18}$$

where $e_1^T = (1, 0, \ldots, 0)$,

$$W_x = \text{diag}\left(w_1 \cdot K\left(\frac{|x_1 - x|}{h} \right), \ \ldots, \ w_n \cdot K\left(\frac{|x_n - x|}{h} \right) \right), \tag{4.19}$$

and

$$\Phi_x = \begin{bmatrix} 1 & x_1 - x & \cdots & \frac{(x_1 - x)^p}{p!} \\ 1 & x_2 - x & \cdots & \frac{(x_2 - x)^p}{p!} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \cdots & \frac{(x_n - x)^p}{p!} \end{bmatrix}. \tag{4.20}$$

The order p of the local polynomial is typically fixed *a priori* and thus, like kernel smoothing, the kernel bandwidth is the lone hyperparameter that must be estimated for LOESS.

[230] and [121] showed that the local linear regression estimator ($p = 1$) eliminates the boundary bias of the kernel smoothing estimator. Taking $p > 1$ further reduces the bias of the estimator (e.g, near spikes and dips), but at the cost of increased variance throughout the input domain.

[230] and [121] showed the univariate LOESS estimator is minimax over L_2 Sobolev classes and [231] extended the results to multivariate feature spaces.

The multivariate LOESS estimator naturally arises by replacing the univariate ℓ_2 norm $|x - x_i|$ with the d-dimensional ℓ_2 norm $\|x - x_i\|_2$ and utilizing multivariate polynomials for the basis functions. One should typically scale the features before fitting a multivariate LOESS model but this is not necessary here since we are working with spatial covariates.

The computational complexity of an exact LOESS regression (evaluated at the observed inputs) is $\mathcal{O}(n^2)$ regardless of the compactness of the kernel because of the n distances that need to be computed for each local polynomial. However, the computational complexity can be improved to $\mathcal{O}(n \log n)$ by using k-d trees [232] or other space-partitioning data structures [e.g., 122] to quickly find an approximate set of nearest neighbors for each local polynomial fit. In this section we utilize the `locfit` R package [122].

4.5.3 Multi-resolution Gaussian random field regression

In this section we outline the multi-resolution Gaussian random field (GRF) model proposed by [227], which incorporates significant elements from fixed-rank kriging [233, 234] and stochastic partial differential equations literature [235–237].

4.5.3.1 Gaussian random field regression

Here we first briefly describe the generic setup of Gaussian random field regression. Recall the assumed observational data generating process

$$f(x) = f_0(x) + \epsilon(x) \tag{4.21}$$

In Gaussian random field regression we assume the underlying signal f_0 is itself stochastic — in particular, a realization of a Gaussian random field. Typically, the assumption is

$$f_0(x) \sim \mathcal{GP}(\mu(x), k(x, x')) \tag{4.22}$$

where $\mu(x)$ is a constant, linear, or very smooth mean function, $k(x, x')$ is the covariance function of the GRF, and ϵ is a white Gaussian noise process. The spatial covariance function is often assumed to be isotropic and defined by a small set of hyperparameters which can be estimated via maximum likelihood. A common choice for the covariance is the Gaussian (i.e. squared exponential) covariance

$$k(x, x') = \rho \cdot \exp\left(- \frac{\|x - x'\|_2^2}{\gamma} \right), \tag{4.23}$$

where γ is a range hyperparameter and ρ controls the marginal variance of the field. Given the observed data and estimated hyperparameter vector η, the point estimate of the GRF model is given by the mean of the posterior distribution

$$\mathbb{E}[f_0(x)|f, \eta] = k_x^T (K + W^{-1})^{-1} f, \tag{4.24}$$

where $k_x = \big(k(x, x_1), \ldots, k(x, x_n)\big)^T$, $K_{ij} = \mathrm{Cov}(f_0(x_i), f_0(x_j))$, W is diagonal with elements $W_{ii} = \sigma_i^{-2}$, and f is the vector of observations. Thus GRF regression is a linear smoother with effective kernel

$$\ell^{\mathrm{GRF}}(x) = k_x^T (K + W^{-1})^{-1} \tag{4.25}$$

In general, the computational complexity of computing the GRF point estimate is $\mathcal{O}(n^3)$.

4.5.3.2 Multi-resolution GRF model

Let $(x_1, \widehat{\delta}_F(x_1)), \ldots, (x_n, \widehat{\delta}_F(x_n))$ denote the flux contrast estimates defined in equation (4.6), with the Cartesian spatial coordinates given by equations (4.8)-(4.11). We assume the flux contrast estimates arise according to the data generating process

$$\widehat{\delta}_F(x_i) = g_0(x_i) + \epsilon_i \tag{4.26}$$

where g_0 is a realization of a smooth Gaussian random field (GRF) g and $\epsilon_1, \ldots, \epsilon_n$ are independent mean zero measurement errors. Since the flux contrast has mean zero we let g be a mean zero GRF. We expect the IGM to have multiple scales of structure so we let g be a sum of L independent GRFs $g_1, \ldots, g_L$ with marginal variances $\rho\alpha_1, \ldots, \rho\alpha_L$,

$$g(x) = \sum_{\ell=1}^{L} g_\ell(x), \tag{4.27}$$

where $\alpha_1, \ldots, \alpha_L > 0$ sum to one and $\rho > 0$. This representation allows g to adapt to the complex multi-scale dependence that we expect to see in the structure of the intergalactic medium. As in fixed-rank kriging, each individual GRF g_ℓ is then defined through a basis expansion

$$g_\ell(x) = \sum_{j=1}^{p_\ell} \phi_{\ell,j}(x)\beta_{\ell,j}, \tag{4.28}$$

where $\phi_{\ell,j}$, $j = 1, \ldots, p_\ell$, is a sequence of compactly supported radial basis functions organized on regular three-dimensional lattices of increasing resolution and β_ℓ is a vector of coefficients such that

$$\beta_\ell \sim MN(0, \rho\Sigma_\ell), \quad \ell = 1, \ldots, L. \tag{4.29}$$

Thus, the model for g is constructed as a sum of fixed basis functions coupled with stochastic coefficients. These two key elements of the model are further detailed in the sections below.

4.5.3.3 Radial basis functions

Each level of resolution in the multi-resolution GRF model is provided by a sequence of compactly supported radial basis functions (RBFs) organized on a regular lattice in three-dimensional Euclidean space. The node spacing of the lattices is given by

$$\delta_\ell = 2^{-(\ell-1)}\delta_1, \quad \ell = 1, \ldots, L, \tag{4.30}$$

where δ_1 and L are taken as hyperparameters of the model. At each level $\ell = 1, \ldots, L$, the RBFs are given by a real-valued radial function

$$\phi_{\ell,j}(x) = \phi\left(\frac{\|x - u_{\ell,j}\|_2}{\theta_\ell}\right), \quad j = 1, \ldots, p_\ell, \tag{4.31}$$

where $u_{\ell,1}, \ldots, u_{\ell,p_\ell} \in \mathbb{R}^3$ are the nodes of the lattice and θ_ℓ is a scaling parameter to adjust the amount of overlap in the RBFs at each level. In particular, we let ϕ be the three-dimensional Wendland polynomimal [238] given by

$$\phi(x) = \begin{cases} (1-x)^6(35x^2 + 18x + 3)/3 & 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.32}$$

and, following [227], we fix θ_ℓ to be 2.5 times the node spacing at level l so the RBFs have a sufficient amount of overlap to avoid artifacts in the covariance function from the organization of the RBFs on a lattice. Moreover, we add five extra layers of nodes beyond the edges of each

lattice to mitigate edge effects, which is particularly important for our distributed approximation algorithm discussed below.

4.5.3.4 Gaussian Markov random fields

To simplify notation we combine equations (4.27) and (4.28) so

$$g(x) = \sum_{j=1}^{p} \beta_j \phi_j(x), \tag{4.33}$$

where we have combined the multi-resolution coefficients into a single vector β and the multi-resolution bases into a single basis $\{\phi_j\}_{j=1}^{p}$. From equation (4.29) we have

$$\beta \sim MN(0, \rho\Sigma), \tag{4.34}$$

for some matrix Σ.

The foundation of the computational efficiency of this spatial model is the enforcement of sparsity in matrix computations in a way that does not sacrifice covariance models with many degrees of freedom and multi-scale correlations. In addition to the use of basis functions with compact support, this is accomplished by directly computing the precision matrix $\frac{1}{\rho}\Sigma^{-1}$ of the basis coefficients instead of the covariance matrix $\rho\Sigma$ and restricting Σ^{-1} to be sparse. This approach allows for the use of efficient sparse matrix algorithms and still permits Σ to be dense. Specifically, we first assume that coefficients between levels are independent, which gives the precision matrix

a convenient block diagonal structure

$$\frac{1}{\rho}\Sigma^{-1} = \frac{1}{\rho}\begin{bmatrix} \frac{1}{\alpha_1}\Sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\alpha_2}\Sigma_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{\alpha_L}\Sigma_L^{-1} \end{bmatrix}. \tag{4.35}$$

We then model the distribution of the basis coefficients at each level of resolution as a Gaussian Markov random field (GMRF) organized on the nodes of each lattice. The Markov property of each GMRF can be described by an undirected graph. In particular, given the set of lattice nodes V_ℓ at level ℓ we define an edge set E_ℓ such that

$$\left(\Sigma_\ell^{-1}\right)_{i,j} = 0 \quad \text{if} \quad \{i,j\} \notin E_\ell. \tag{4.36}$$

We assume the special case that each β_ℓ follows a first-order spatial autoregression with respect to the nodes of the lattice at level ℓ. Specifically, this means that for every $i = 1, \ldots, p_\ell$ the off-diagonal nonzero elements of the ith row of Σ_ℓ^{-1} correspond to the first and second order neighbors of node i. Given an autoregression matrix B_ℓ for level ℓ, we construct the distribution of β_ℓ according to $\beta_\ell = B_\ell^{-1}e$, where $e \sim N(0, \rho I)$. Following [237], we let

$$B_{i,j} = \begin{cases} 6 + \kappa^2 & i = j, \\ -1 & \{i,j\} \in E_\ell \text{ and } i \neq j, \\ 0 & \text{otherwise}, \end{cases} \tag{4.37}$$

where $\kappa \geq 0$. It follows that $\beta_\ell \sim N(0, \rho B^{-1} B^{-T})$. Moreover, the covariance matrix of the full coefficient basis is given by

$$\rho\Sigma = \rho \begin{bmatrix} \alpha_1 B_1^{-1} B_1^{-T} & 0 & \cdots & 0 \\ 0 & \alpha_2 B_2^{-1} B_2^{-T} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \alpha_L B_L^{-1} B_L^{-T} \end{bmatrix}. \tag{4.38}$$

4.5.3.5 Hyperparameters

Based on the introduced model setup, g is a mean zero Gaussian random field with covariance function

$$\mathrm{Cov}\big(g(x), g(x')\big) = \sum_{j=1}^{m} \sum_{k=1}^{m} \rho\Sigma_{j,k} \phi_j(x) \phi_k(x'). \tag{4.39}$$

We assume $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are independent with

$$\epsilon \sim MN(0, \sigma^2 W) \tag{4.40}$$

and $\mathrm{Cov}(\beta, \epsilon) = 0$. Here, W is diagonal with elements proportional to the estimated measurement variances of the flux contrast estimates and σ^2 is a hyperparameter of the measurement error distribution. Now letting Φ be the regression matrix with $\Phi_{i,j} = \phi_j(x_i)$ we can rewrite equation (4.26) in matrix vector notation

$$\widehat{\delta}_F = \Phi\beta + \epsilon, \tag{4.41}$$

where $\widehat{\delta}_F$ is the vector of flux contrast estimates lying along quasar sightlines. Thus, the assumed observational model is

$$\widehat{\delta}_F \sim MN(0, \rho\Pi), \tag{4.42}$$

where $\Pi = (\Phi\Sigma\Phi^T + \lambda W)$ and $\lambda = \frac{\sigma^2}{\rho}$. Here, λ is the "noise-to-signal ratio" and is used as a reparametrization of σ^2 to tune the smoothness of the maps. From equation (4.42) we have the likelihood

$$L(\rho, \Sigma, \lambda \mid \widehat{\delta}_F) = \frac{1}{(2\pi)^{n/2}|\rho\Pi|^{1/2}} e^{-\frac{1}{2}\widehat{\delta}_F^T(\rho\Pi)^{-1}\widehat{\delta}_F} \tag{4.43}$$

and log likelihood

$$\ell(\rho, \Sigma, \lambda \mid \widehat{\delta}_F) = -\frac{1}{2}\widehat{\delta}_F^T(\rho\Pi)^{-1}\widehat{\delta}_F - \frac{1}{2}\log|\rho\Pi| - \frac{n}{2}\log(2\pi). \tag{4.44}$$

In principle, we can use equation (4.44) to compute maximum likelihood estimates (MLEs) for the full set of hyperparameters: ρ, λ, and all hyperparameters that define Σ. Nevertheless, such an approach is not computationally feasible — that is, we did not have the computational resources at the time this was carried out. In particular, we can simplify computations by specifying the covariance Σ and focusing on MLEs for λ and ρ. The precision matrix Σ^{-1} is determined by the scale parameter κ and a smoothness parameter ν, which dictates the relative variances of the levels according to $\alpha_l \sim e^{l\cdot\nu}$ — thereby reducing the dimensionality of the hyperparameter space. The predicted map is relatively insensitive to the choice of κ and ν so we follow the practice of fixing these parameters and optimizing λ and ρ[2]. Details are provided below.

Lattice parameters

From the left panel of Figure 4.11 we see that the minimum mean sightline separation over all redshifts is approximately 12.5 h^{-1} Mpc for both the Northern and Southern Galactic Caps and occurs at $z \sim 2.1$. We therefore fix the finest level of basis functions to have a node spacing of 12.5 h^{-1} Mpc so the multi-resolution model has sufficient complexity to model structure at the scale of the mean sightline separation, but not so much that it will overfit the sightlines.

[2] Again, this was for lack of available computing power at the time of writing

Moreover we add three coarser levels with node spacings of 25, 50, and 100 h^{-1} Mpc to allow for larger scale correlations in the reconstructed absorption field.

Given the significantly higher density of sightlines in Stripe 82, the mapping of this region could benefit from a finer level of basis functions than that which we use here. Nevertheless, here we focus on optimizing the map with respect to the full 10,400 deg^2 footprint. We could potentially follow up with a separate higher resolution map of Stripe 82, with the hyperparameters optimized specifically to the 220 deg^2 Stripe 82 coverage.

κ and $\{\alpha_l\}_{l=1}^4$

The precision matrix Σ^{-1} is determined by the scale parameter κ and a smoothness parameter ν, which specifies the relative variances of the levels according to $\alpha_l \sim e^{l\cdot\nu}$. The predicted map is relatively insensitive to the choice of κ and ν so we follow the practice of fixing these parameters and optimizing λ and ρ.

λ and ρ

Let $\widehat{\Sigma}$ denote the covariance matrix determined by the hyperparameter choices outlined above. Substituting $\widehat{\Sigma}$ into (4.44), we first maximize (4.44) over ρ analytically, yielding

$$\widehat{\rho} = \widehat{\delta}_F^T \Pi^{-1} \widehat{\delta}_F / n. \tag{4.45}$$

This estimate is then substituted back into (4.44) to give a profile log likelihood,

$$\ell(\widehat{\rho}, \widehat{\Sigma}, \lambda \mid \widehat{\delta}_F) = -\frac{1}{2}\widehat{\delta}_F^T (\widehat{\rho}\Pi)^{-1} \widehat{\delta}_F - \frac{1}{2}\log|\widehat{\rho}\Pi| - \frac{n}{2}\log(2\pi), \tag{4.46}$$

which depends only on λ. The maximum likelihood estimate for λ is then computed via an iterative optimization, specifically a combination of golden section search and successive parabolic

interpolation [239].

Another significant computational cost arises in computing the determinant of Π when evaluating the likelihood (4.44). Sylvester's determinant theorem [240] provides the useful identity

$$|\Pi| = \frac{\lambda^{n-m}|\Gamma|}{|\Sigma^{-1}||W^{-1}|}. \tag{4.47}$$

Since Γ, Σ^{-1}, and W^{-1} are all sparse the determinants of each can then be computed efficiently from the product of the diagonal elements of the corresponding Cholesky decompositions.

4.5.3.6 Divide and conquer maximum likelihood estimation

The evaluation of equations (4.45) and (4.46) can be done much more efficiently by using the sparse matrix decompositions and matrix identities outlined above. Nevertheless, the matrices associated with a problem of this scale are too large to fit in memory on modern high-performance computing machines and we therefore must adapt our optimization of λ to a distributed ("divide and conquer") framework. Specifically, we partition the 10,400 deg^2 BOSS footprint into equal-area HEALPix pixels (shown in Figure 4.12), compute a "local-MLE" on each subset, and then produce a single global hyperparameter estimate for λ by studying the distribution of the local MLEs. The HEALPix pixelization of the BOSS footprint creates 768 subsets in total, each with an approximate sky area of 3.66 deg^2.

Here we describe the procedure for computing the global hyperparameter estimate $\widehat{\lambda}$. Let $S_1, \ldots, S_{768}$ be a partition of the targeted volume for three-dimensional reconstruction of the IGM induced by the HEALPix pixelization of the sky. Now define each local data set as

$$\mathcal{D}_j = \big\{(x_i, \widehat{\delta}_F(x_i)) : x_i \in S_j\big\}, \tag{4.48}$$

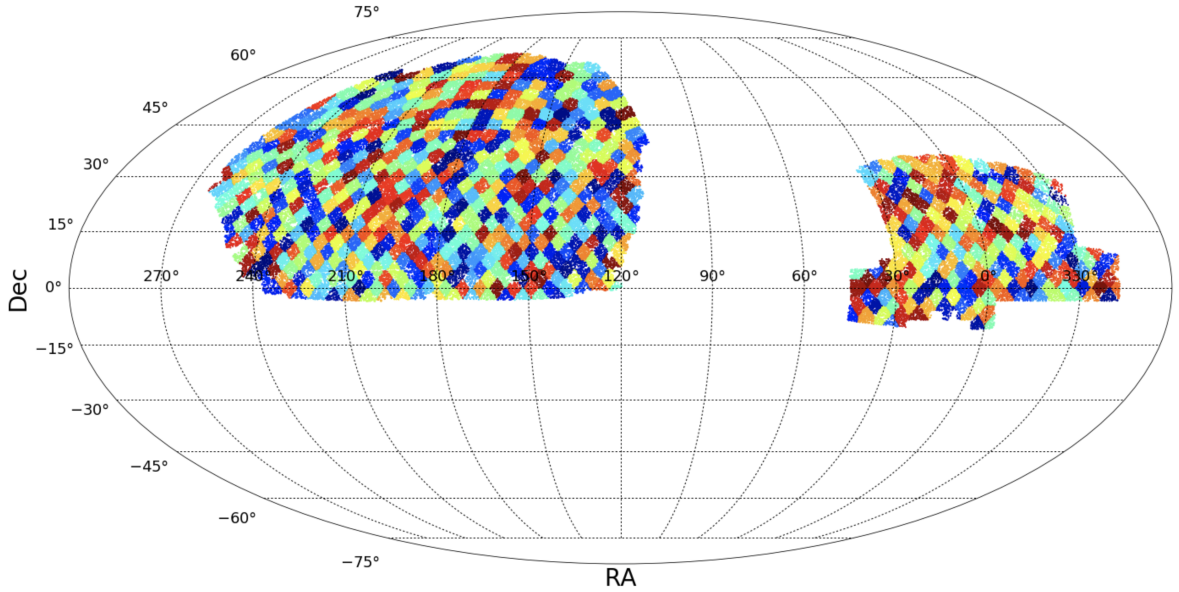


FIGURE 4.12: Partition of the BOSS footprint into 768 equal-area HEALPix subsets, which we utilize to produce an estimate for the model hyperparameter λ in a distributed fashion. Specifically, we compute a local maximum likelihood estimate of λ for each HEALPix subset and then produce a global estimate by taking the mode of a kernel density estimate fit to the local MLEs. Each HEALPix subset has area $\sim$3.66 deg^2.

where $n_j = |\mathcal{D}_j|$. Let $\widehat{\lambda}_j$ be the local-MLE obtained by optimizing the profile log likelihood (4.46) of the subsample $\mathcal{D}_j$. By definition, $\lambda > 0$ so we estimate the distribution of the local-MLEs with a truncated and scaled kernel density estimate

$$\widehat{p}(\lambda) = \begin{cases} \left(\int_0^\infty \widetilde{p}(\lambda) d\lambda \right)^{-1} \cdot \widetilde{p}(\lambda) & \lambda \geq 0 \\ 0 & \lambda < 0 \end{cases} \tag{4.49}$$

where

$$\widetilde{p}(\lambda) = \frac{1}{\sum_{i=1}^{768} n_i} \sum_{j=1}^{768} \frac{n_j}{h} K\left(\frac{\|\lambda - \widehat{\lambda}_j\|_2}{h} \right). \tag{4.50}$$

We select h by 10-fold cross validation and define the global parameter estimate to be the mode

$$\widehat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, \widehat{p}(\lambda). \tag{4.51}$$

Since the profile log likelihood of each subsample is a smooth function of λ, $\widehat{\lambda}$ will be nearly

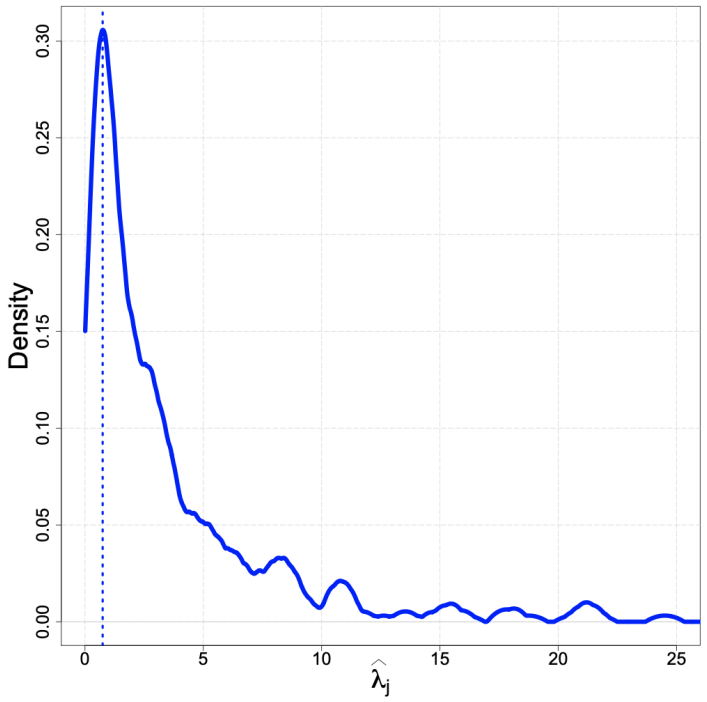


FIGURE 4.13: Kernel density estimate of the distributed sample of 768 local maximum likelihood estimates. We take the global parameter estimate to be the mode of the KDE $\widehat{\lambda} \approx 0.747$.

optimal for any subset S_j with MLE $\widehat{\lambda}_j$ lying in a neighborhood of $\widehat{\lambda}$. Figure 4.13 shows the kernel density estimate $\widehat{p}(\lambda)$ of the local-MLE distribution. The divide and conquer approach for computing $\widehat{\lambda}$ is summarized in the algorithm table below.

Algorithm 5 Divide and conquer estimation of λ

Require: $\mathcal{D}_1, \ldots, \mathcal{D}_{768}$, $\widehat{\boldsymbol{\Sigma}}$

1: **for all** j **do**

2: Compute $\widehat{\lambda}_j$ via optimization of the profile log likelihood for $\mathcal{D}_j$

3: **end for**

4: Let

$$\widehat{p}(\lambda) = \begin{cases} \left(\int_0^\infty \widetilde{p}(\lambda) d\lambda \right)^{-1} \cdot \widetilde{p}(\lambda) & \lambda \geq 0 \\ 0 & \lambda < 0 \end{cases} \tag{4.52}$$

where

$$\widetilde{p}(\lambda) = \frac{1}{\sum_{i=1}^{768} n_i} \sum_{j=1}^{300} \frac{n_j}{h} K\left(\frac{\|\lambda - \widehat{\lambda}_j\|_2}{h} \right) \tag{4.53}$$

and h is chosen by 10-fold cross validation

5: Define $\widehat{\lambda} = \text{argmax}_\lambda \, \widehat{p}(\lambda)$

4.5.3.7 Point estimate

For the point estimate of the three-dimensional absorption field, we consider the conditional distribution of the basis coefficients given the data and basis covariance. Fixing the covariance hyperparameters at their true values, it follows that

$$(\beta, \widehat{\delta}_F) \mid \{\lambda, \rho, \Sigma\} \sim MN\left(0, \rho \begin{bmatrix} \Sigma & \Sigma\Phi^T \\ \Phi\Sigma & \Pi \end{bmatrix}\right). \tag{4.54}$$

By basic normal theory, the conditional distribution of β given the flux contrast vector $\widehat{\delta}_F$ and all covariance parameters is

$$\beta \mid \{\widehat{\delta}_F, \lambda, \rho, \Sigma\} \sim MN(\beta_0, \rho\Sigma - \rho\Sigma\Phi^T\Pi^{-1}\Phi\Sigma). \tag{4.55}$$

The minimum squared-error basis vector is then given by the posterior mean

$$\beta_0 = \Sigma\Phi^T\Pi^{-1}\widehat{\delta}_F. \tag{4.56}$$

The true covariance parameters are not known in practice so we replace them with the maximum likelihood estimates and take the point estimate of the three-dimensional absorption field to be

$$\widehat{g}_0(x) = \Phi_x^T\widehat{\beta}, \tag{4.57}$$

where

$$\Phi_x = (\phi_1(x), \ldots, \phi_m(x)), \tag{4.58}$$

$$\widehat{\beta} = \widehat{\Sigma}\Phi^T\widehat{\Pi}^{-1}\widehat{\delta}_F, \tag{4.59}$$

$$\widehat{\Pi} = (\Phi\widehat{\Sigma}\Phi^T + \widehat{\lambda}W). \tag{4.60}$$

Therefore, the multi-resolution GRF model is a linear smoother with effective kernel

$$\ell^{GRF}(x) = \Phi_x^T\widehat{\Sigma}\Phi^T\widehat{\Pi}^{-1}. \tag{4.61}$$

The calculations in equations (4.45), (4.46), and (4.59) involve the inverse of Π, which is very large and dense. We can, however, avoid explicitly computing Π^{-1} by computing the vector $\Pi^{-1}\widehat{\delta}_F$ directly. Applying the Sherman-Morrison-Woodbury formula [241], we have

$$\Pi^{-1} = \left(\Phi\widehat{\Sigma}\Phi^T + \lambda W\right)^{-1} \tag{4.62}$$

$$= W^{-1} - (W^{-1}\Phi)\Gamma^{-1}(W^{-1}\Phi)^T \tag{4.63}$$

where

$$\Gamma = \Phi^T W^{-1}\Phi + \lambda\widehat{\Sigma}^{-1}. \tag{4.64}$$

By construction, Φ, W^{-1}, and $\widehat{\Sigma}^{-1}$ are all sparse so Γ is also sparse. Furthermore, Γ is positive definite since it is a sum of positive definite matrices. Therefore, we can utilize the sparse Cholesky decomposition of Γ to solve the linear system

$$\Gamma v = (W^{-1}\Phi)^T\widehat{\delta}_F \tag{4.65}$$

and it follows that

$$\Pi^{-1}y = W^{-1}\widehat{\delta}_F - W^{-1}\Phi v. \tag{4.66}$$

4.5.3.8 Distributed approximation of the point estimate

As with the evaluation of the likelihood, the size of the matrices associated with the calculation of the point estimate is too large to fit in memory and therefore must also be adapted to a distributed framework. We use a similar approach to the divide and conquer maximum likelihood estimation. Specifically, we partition the volume into contiguous three-dimensional subsets, compute a "sub-map" reconstruction on each, and combine them into a mosaic map. The only necessary provision is that sufficiently large margins are included when fitting each submap to avoid visible discontinuity artifacts at the seams of the partition. We partition the volume into $(400\ h^{-1}\ \text{Mpc})^3$ cubes and include all observations within $(200\sqrt{3}+125)\ h^{-1}$ Mpc of the center of each cube when producing each local map. Altogether, the mosaic map comprises 932 submaps in the Northern Galactic Cap and 422 submaps in the Southern Galactic Cap.

4.5.3.9 Simulation from the posterior

The multi-resolution GRF model has an analytical Gaussian posterior distribution, however it is much more computationally efficient to approximate the posterior standard errors via Monte Carlo simulation instead of the analytical calculation. In particular, holding the covariance hyperparameters fixed, we can generate an ensemble of three-dimensional models $g_1^*, \ldots, g_B^*$ using the Monte Carlo algorithm detailed in Algorithm 6. The estimated pointwise standard errors of the posterior distribution then follow as

$$\widehat{se}(\widehat{g}(x)) = \sqrt{\frac{1}{B}\sum_{j=1}^{B}\left(g_j^*(x) - \widehat{g}(x)\right)^2}, \tag{4.67}$$

from which we obtain the approximate pointwise $1 - \alpha$ credible interval

$$\widehat{g}(x) \pm z_{\alpha/2}\widehat{\mathrm{se}}(\widehat{g}(x)). \tag{4.68}$$

The algorithm for computing $\widehat{\mathrm{se}}(\widehat{g}(x))$ (shown below) involves a simple trick based on the linear statistics for the multivariate normal to generate draws from the conditional distribution.

Algorithm 6 Monte Carlo confidence intervals

Require: $\widehat{g}(x)$, Φ, W, $\widehat{\Sigma}^{-1}$, $\widehat{\rho}$, $\widehat{\sigma}^2$, $\widehat{\lambda}$

Compute the sparse Cholesky decomp. $\widehat{\rho\Sigma}^{-1} = B^T B$
for j in $1 : B$ **do**
  Solve $B\beta_j^* = e \quad$ where $\quad e \sim MN(0, I)$
  $y_j^* \leftarrow \Phi\beta_j^* + \epsilon^* \quad$ where $\quad \epsilon^* \sim MN(0, \widehat{\sigma}^2 W)$
  $\widehat{\beta}_j^* \leftarrow \widehat{\Sigma}\Phi^T\widehat{\Pi}^{-1}y_j^*$
  $g_j^*(x) \leftarrow \widehat{g}(x) + \Phi_x(\beta_j^* - \widehat{\beta}_j^*)$
end for
$\widehat{se}(\widehat{g}(x)) = \sqrt{\dfrac{1}{B}\displaystyle\sum_{j=1}^{B}\left(g_j^*(x) - \widehat{g}(x)\right)^2}$
$C(x) = \left(\widehat{g}(x) - z_{\alpha/2}\widehat{\mathrm{se}}(\widehat{g}(x)),\ \widehat{g}(x) + z_{\alpha/2}\widehat{se}(\widehat{g}(x))\right)$
return $C(x)$

4.6 Comparison of methods on a cosmological simulation

In this section we briefly compare the performance of the three-dimensional LOESS estimator and the multi-resolution GRF model on a hydrodynamical cosmological simulation of a Lyα absorption field so that we may choose one model to proceed with for analyzing the BOSS data. The simulation (shown in Figure 4.14) constitutes a cubic volume of $(400\ h^{-1}\ \mathrm{Mpc})^3$ and is output at redshift $z = 2$ with a voxel resolution of $11.7\ h^{-3}\ \mathrm{Mpc}^3$ / voxel (176^3 voxels in total). The cosmological parameters of the simulation were $h = 0.702$, $\Omega_m = 0.275$, $\Omega_\Lambda = 0.725$, $\Omega_b = 0.046$, spectral index $n_s = 0.968$, and amplitude of mass fluctuations, $\sigma_8 = 0.82$. See [25] for more details regarding the development of the simulation.

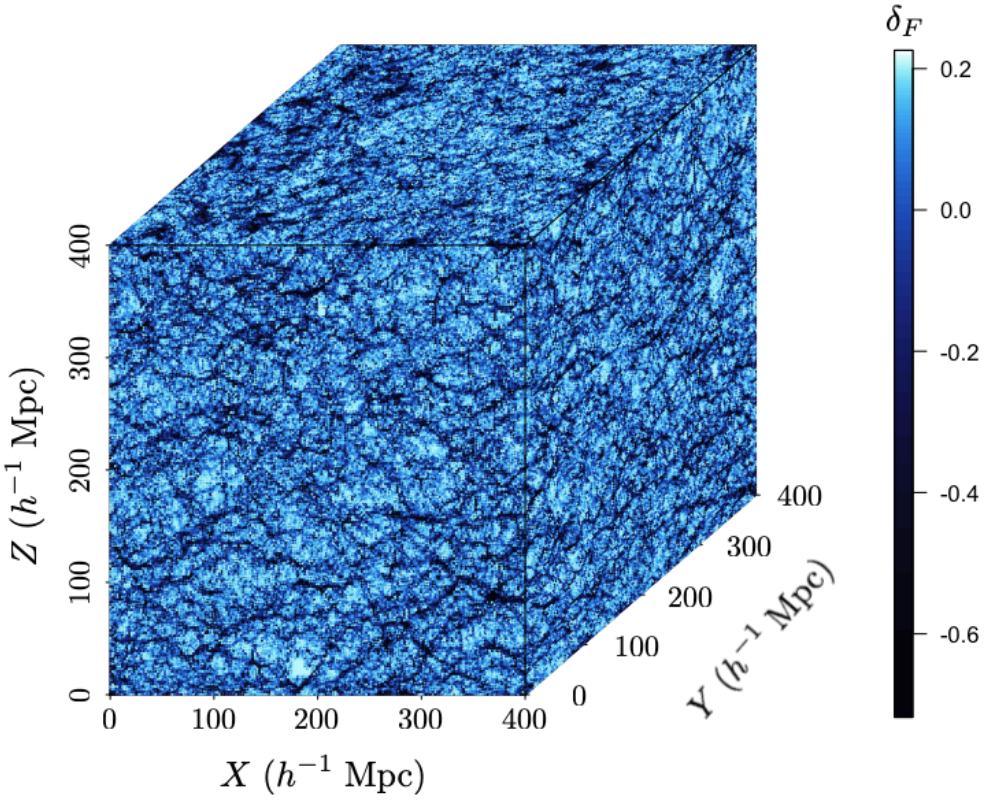


FIGURE 4.14: Cosmological hydrodynamical simulation of a (400 h^{-1} Mpc)^3 Lyα absorption field. We use this simulation to evaluate the performance of our statistical methods in this chapter. See [25] for details regarding the parameters of the simulation.

In order to mimic the heterogeneous Lyα quasar sightline density of the BOSS Lyα catalog, we attempt to reconstruct the simulation cube at three different sightline densities — sampling sets of 100, 300, and 500 sightlines uniformly over the X-Y plane of the cube. The coordinates of each sample of sightlines are shown in Figure 4.15.

Figure 4.16 shows the results of a 10-fold cross validation optimization of the kernel bandwidth for the LOESS estimator. The estimated risk curve appears troublingly flat, with optimal bandwidths (denoted in red) that severely oversmooth the absorption field. We provide a brief explanation for why this occurs below. In order to mitigate this issue we apply what we call a "reverse 1-SE rule" to choose the bandwidth, which selects the smallest bandwidth for which the estimated risk

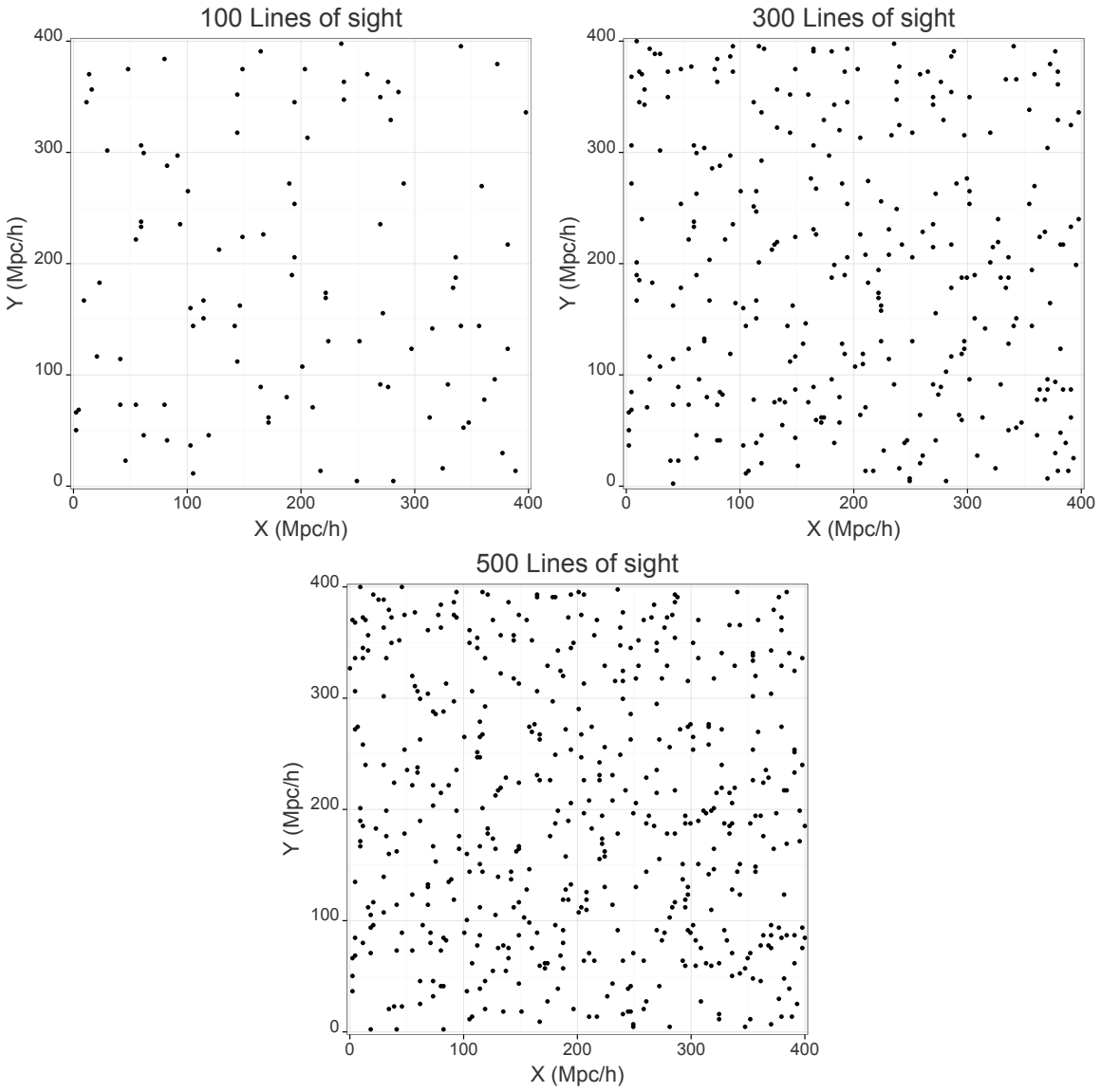


FIGURE 4.15: X-Y coordinates of the sample of one-dimensional sightlines used for each three-dimensional reconstruction.

is within one standard of the minimum risk. These reverse 1-SE bandwidths are shown in blue.

For the multi-resolution GRF model we show the results of the maximum likelihood optimization of the hyperparameter λ in Figure 4.17. Recall the definition $\lambda = \sigma^2/\rho$, so this is indeed a one-dimensional optimization, but is shown in the original two-parameter parametrization (where ρ has an analytical optimum given σ^2).

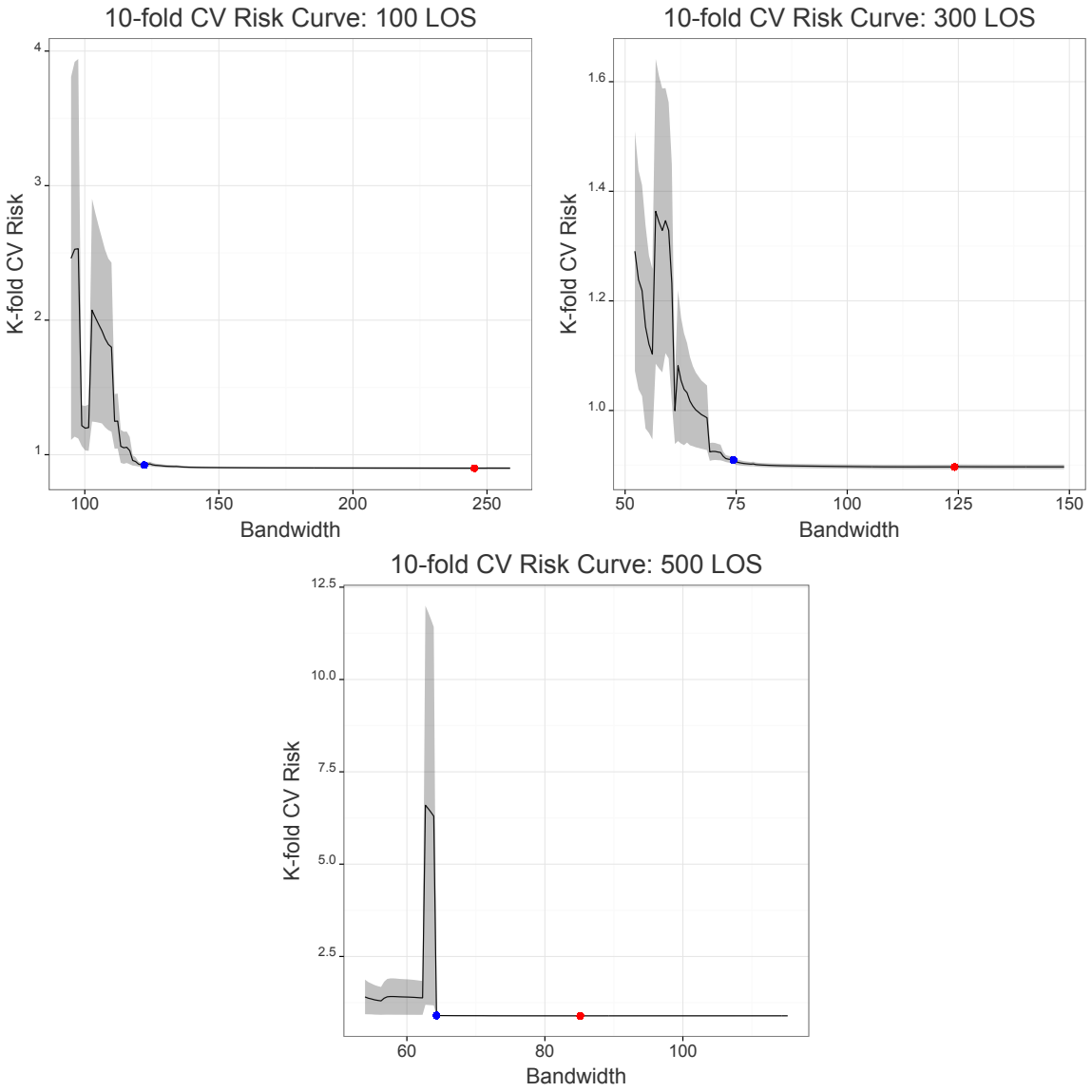


FIGURE 4.16: K-fold cross validation curve for 100, 500 LOS samples. The minimum CV risk bandwidth is designated in red while the reverse 1-SE bandwidth is shown in blue.

The point estimate maps for each three-dimensional model are displayed in Figure 4.18 and the two-point correlation function of each reconstruction is shown in Figure 4.19. Qualitatively speaking, the maximum-likelihood-optimized GRF model seems to produce a reasonable reconstruction of the large-scale structure of the simulation cube, while even the use of the reverse 1-SE bandwidth for the LOESS estimator only recovers structure on the $\sim$100 h^{-1} Mpc scale. Figure 4.20 shows the LOESS fit at a fixed Cartesian coordinate Z plus or minus one standard error. We omit the analogous plots for the GRF model here due to the computational expense of the Monte Carlo

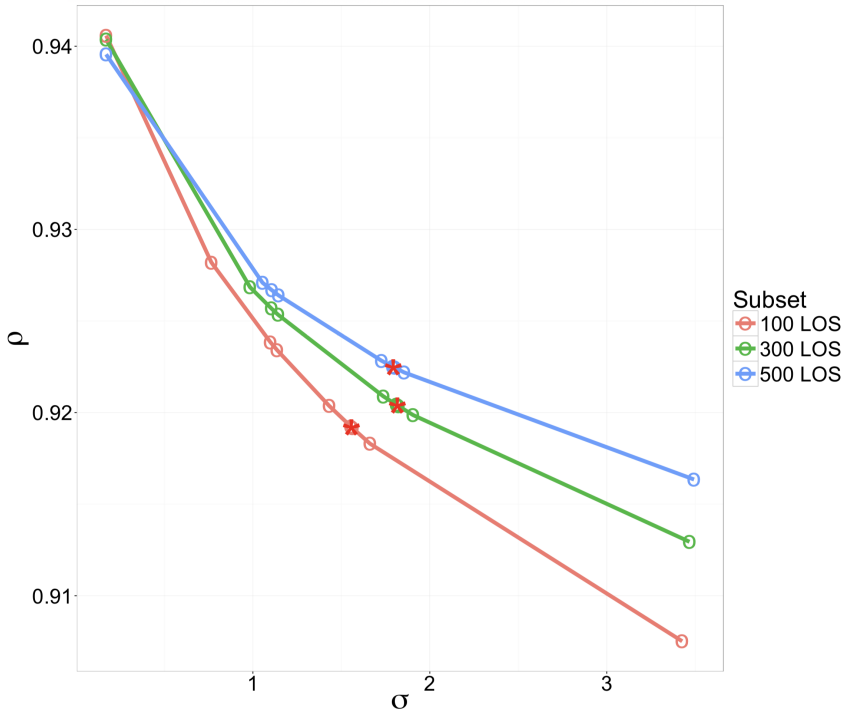


FIGURE 4.17: Maximum likelihood optimization over covariance parameters. The optimal pair for each sample of LOS is designated in red. *Note: What's actually going on here is a one-dimensional optimization of* $\lambda = \sigma^2/\rho$*, which suffices since the log-likelihood only depends on* σ^2 *and* ρ *through* λ*.*

standard errors.

The flatness of the LOESS cross validation curves arises due to the fact that most of the variation in the absorption field is below the scale on which we can reconstruct the field from the density of observed sightlines. An analogous issue arises when evaluating the performance of each absorption field reconstruction with the mean integrated squared error (MISE) with respect to the full resolution signal of the field (shown in Table 4.1). In terms of MISE, the models appear to perform equally poorly across all sightline samples because the small-scale variation of the field dominates the squared error. It is not clear how much the multi-resolution GRF model suffers from an analogous model validation issue as the LOESS estimator. That is, the assumption that the flux contrast observations are a realization of a multi-resolution GRF with structure on $\geq 12.5\ h^{-1}$ Mpc scales and an error distribution that is white Gaussian noise is very seriously violated because there is highly correlated structure on $< 12.5\ h^{-1}$ Mpc scales. The GRF itself

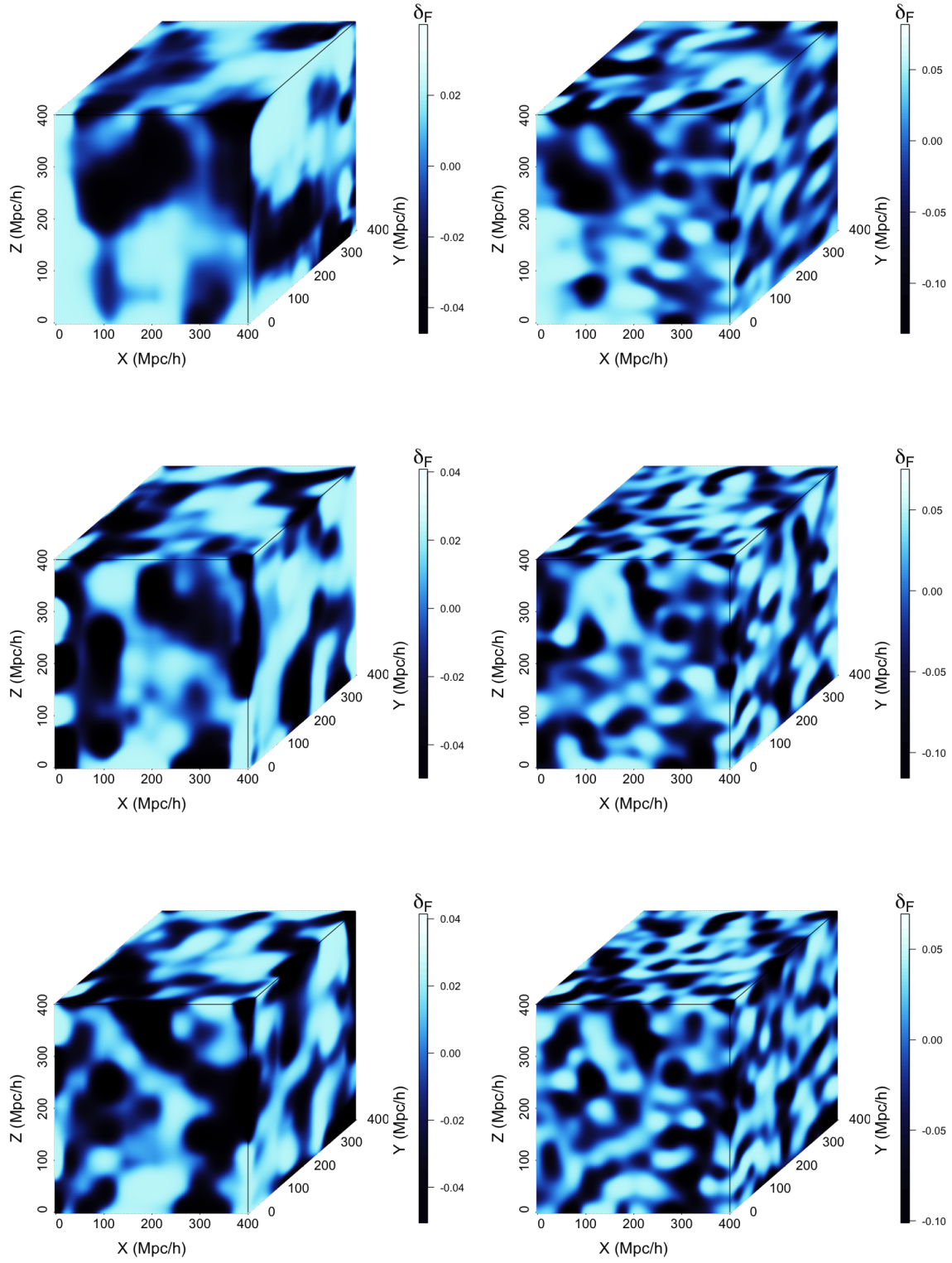


FIGURE 4.18: Predicted maps constructed by LOESS (left) and the GRF model (right) using the 100, 300, and 500 LOS samples, respectively.

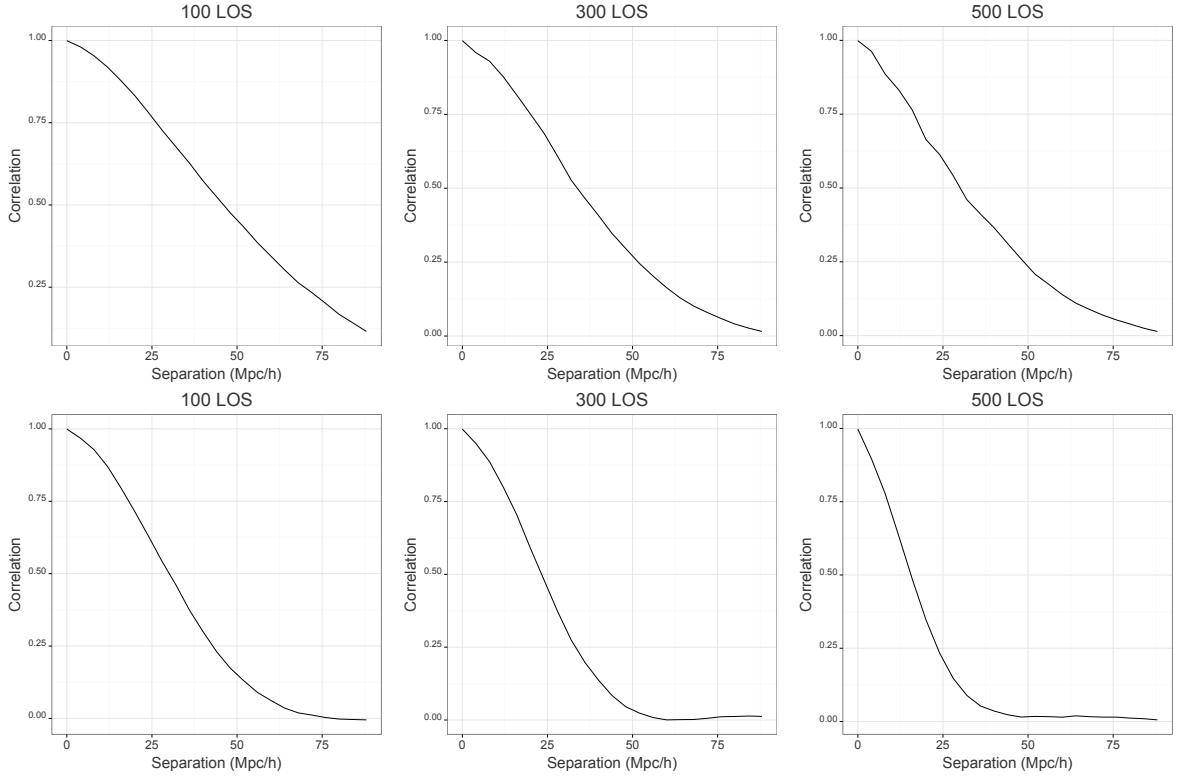


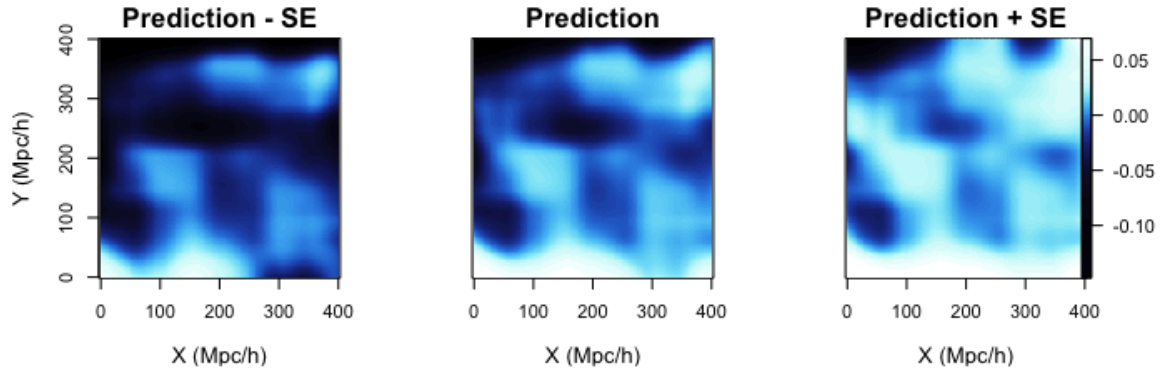
FIGURE 4.19: Autocorrelation functions of the LOESS and GRF density field reconstructions on 100, 300, and 500 LOS samples.

LOS sample	LOESS	GRF model
100	0.05489	0.05460
300	0.05470	0.05421
500	0.05454	0.05400

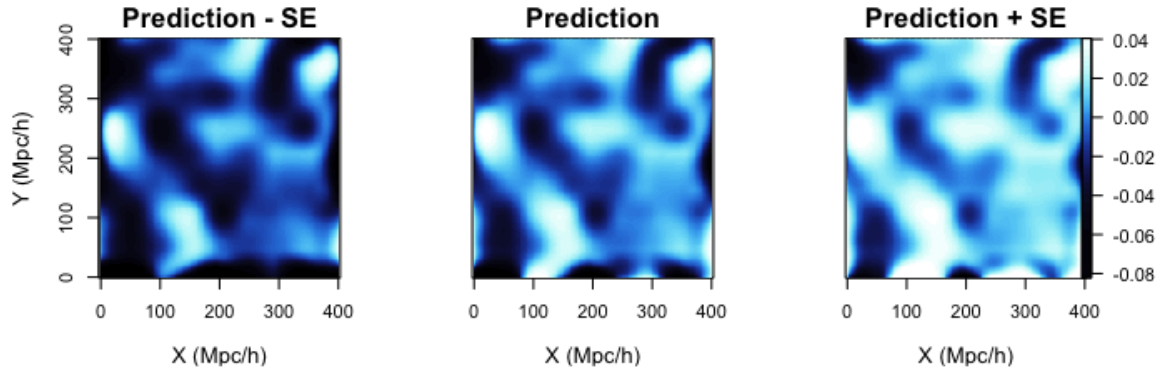
TABLE 4.1: Mean integrated squared errors of each three-dimensional model fit on each of the three line of sight samples. There is very little disparity between the performance of each method across all samples because the majority of the variation in the Lyα absorption field is below the scale of the mean transverse sightline separation, and therefore below the scale on which we can hope to reconstruct the signal. Squared-error with respect to the full resolution data is therefore not a suitable metric from which to gauge the performance of the reconstruction.

therefore suffers from severe model misspecification, which calls into question the maximum likelihood model validation procedure. At this point in time, these points served as valuable lessons learned — all of which we provide solutions for in our analysis in Chapter 5.

(a) 100 LOS



(b) 300 LOS



(c) 500 LOS

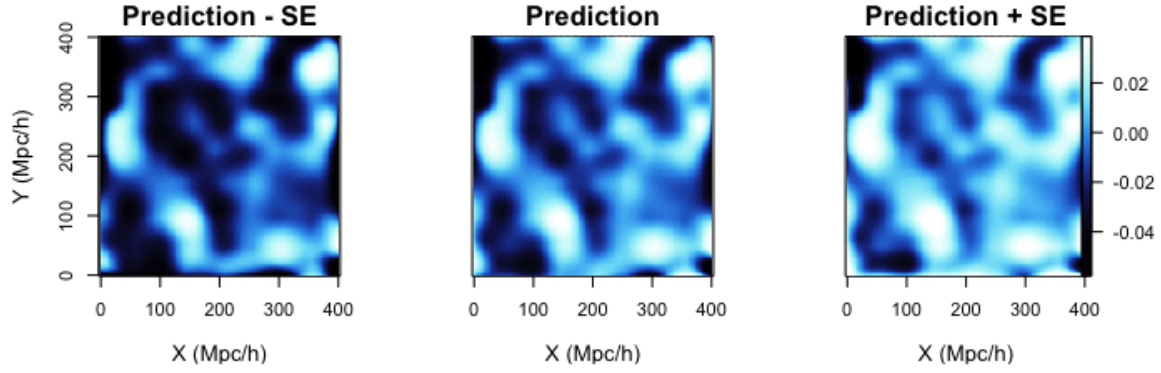


FIGURE 4.20: Slices of the LOESS-reconstructed simulation cube utilizing various LOS sample sizes (a) the 100 LOS sample, (b) the 300 LOS sample, and (c) the 500 LOS sample. The standard errors for the GRF model were not calculated here because of computational cost.

4.7 Application to BOSS Lyα quasar catalog

In this section we display the results of the multi-resolution GRF reconstruction of the 44.3 h^{-1} Gpc^3 volume sampled by the BOSS Lyα quasar sightlines over our selected redshift range.

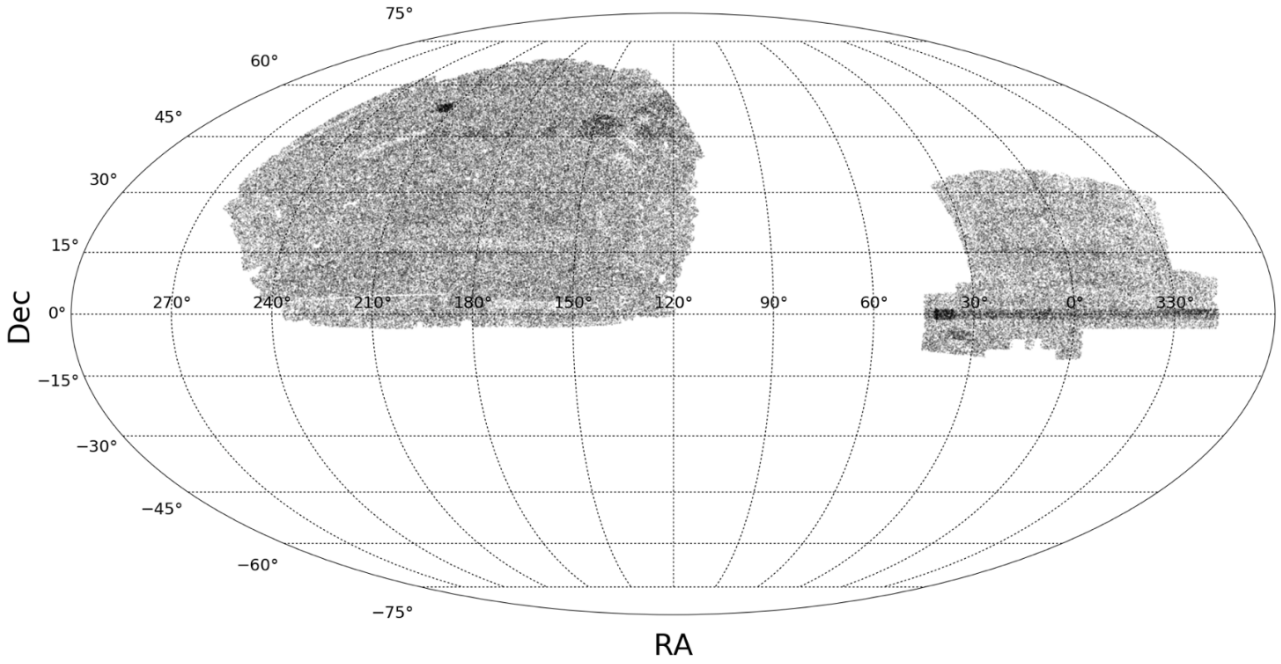


FIGURE 4.21: Footprint of the BOSS Lyα quasars used for the three-dimensional Lyα absorption field reconstruction in this chapter, shown in equatorial coordinates. The total footprint is 10,300 deg^2 ($\sim$25% sky coverage) and the total number quasar sightlines is 168, 953.

The map covers approximately 10,300 square degrees of the sky and spans the redshift range $1.95 < z < 3$. This map represents our first attempt at reconstructing the large-scale intergalactic medium over the unprecedented volumes made possible by the full BOSS Lyα quasar catalog. Our more recent work on this topic is detailed in Chapter 5. The map detailed in this chapter was never submitted for publication due to our dissatisfaction with flaws in the statistical modeling — both the three-dimensional modeling and the one-dimensional pipeline that propagates the observed spectra to the flux contrast scale. Going forward, we therefore proceeded with constructing our own one-dimensional modeling pipeline (Chapters 2 and 3), before ultimately returning to the problem of three-dimensional reconstruction (Chapter 5).

In total, the final sample we used to construct the map in this chapter consists of 168, 953 quasar sightlines. This number accounts for the removal of spectra with no Lyman-α data in the range $1.95 < z < 3$, broad absorption line quasars, and quasars outside of the $\sim$10,300 deg^2 densely sampled sky area. The footprint of the reduced sample is shown in Figure 5.2. Altogether,

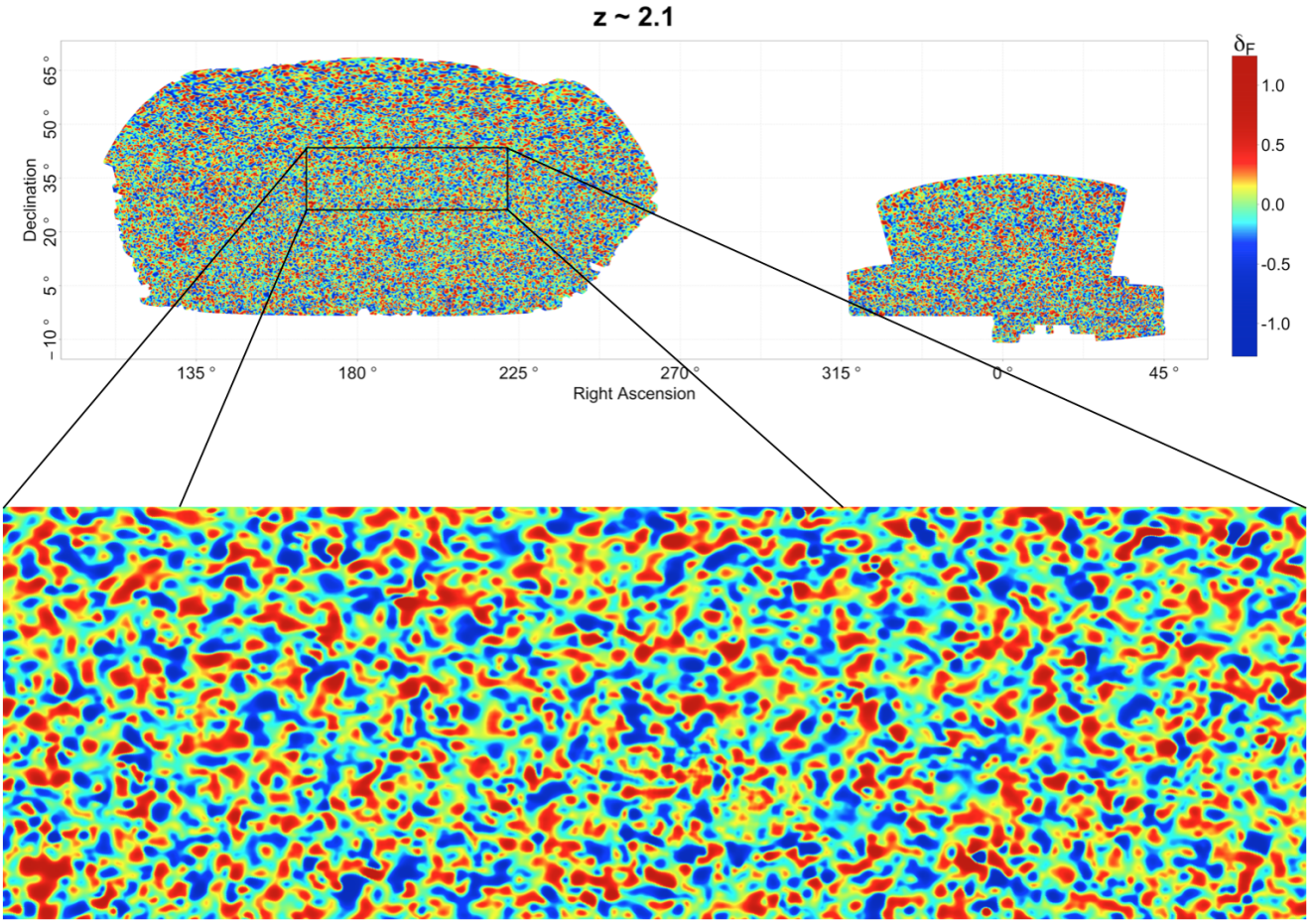


FIGURE 4.22: Two-dimensional sky map cross section of the reconstructed three-dimensional Lyα absorption field at redshift $z = 2.1$. The map has a sky coverage of approximately 10,300 deg^2 (∼25% sky coverage) and has a volume of 44.4 h^{-1} Gpc^3. This map represents our first attempt at reconstructing the large-scale intergalactic medium over the unprecedented volume made possible by the full BOSS Lyα quasar catalog. Our more recent work is detailed in Chapter 5.

this represents a ∼25% sky coverage. The sampling design of the data as a function of redshift is highly heterogeneous, as the collective data set is densest at $z \sim 2.1$ and sharply declines at both lower and higher redshifts. Additionally, although to a much lesser extent, there is sampling heterogeneity on the two-dimensional sky. In particular, the high density of spectra in Stripe 82 can be seen in the Southern Galactic Cap at declination $\delta \sim 0°$. The most significant heterogeneity is of course a byproduct of the sightline sampling, with observations along an individual sightline having an average separation of ∼600 h^{-1} kpc and the average separation between sightlines being anywhere between ~ 12.5 h^{-1} Mpc and ~ 32.5 h^{-1} Mpc, depending on the redshift. We display two-dimensional sky map cross sections of the reconstructed Lyα

absorption field at a variety of redshifts in Figures 4.22–4.25.

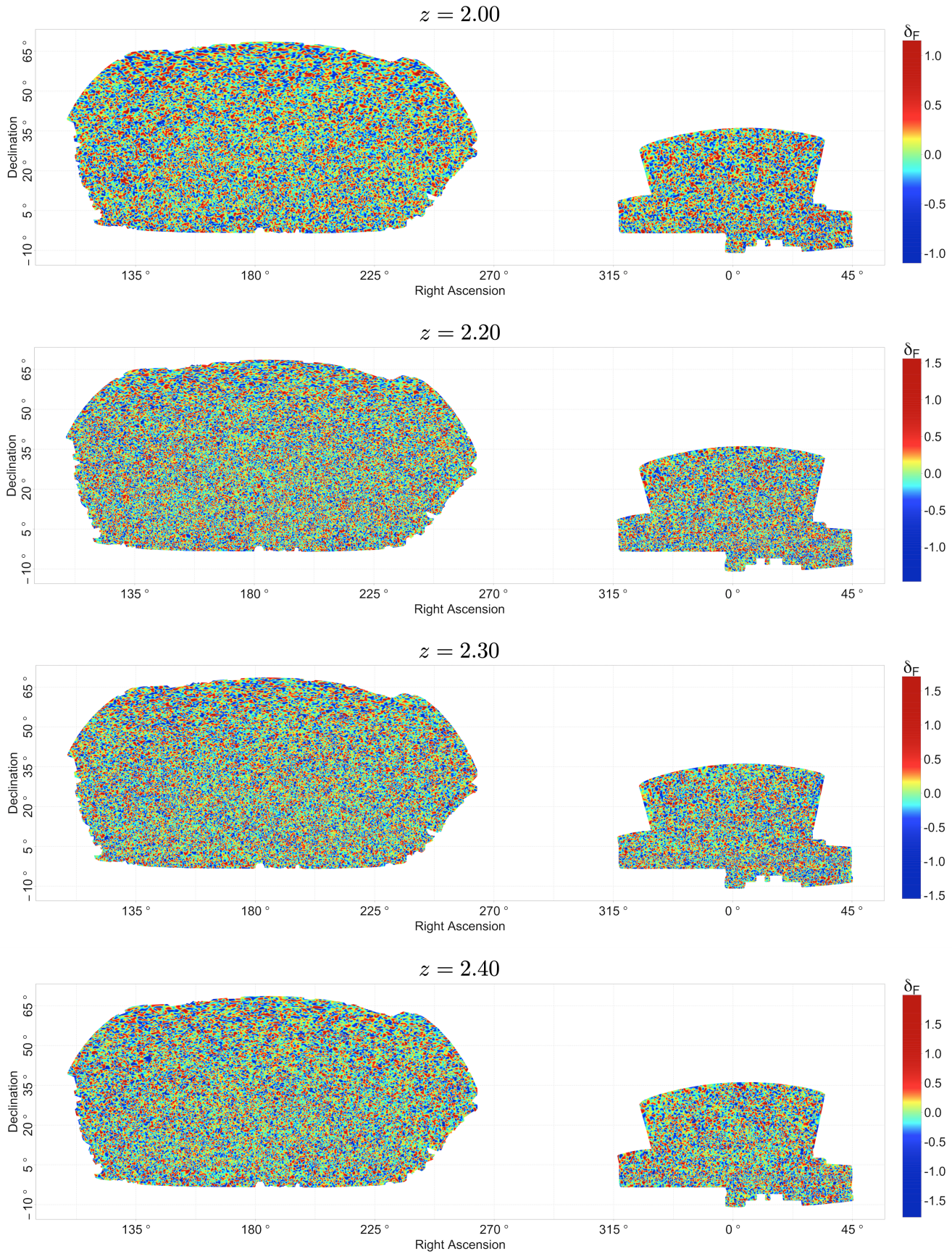


FIGURE 4.23: Two-dimensional sky map cross sections (continued).

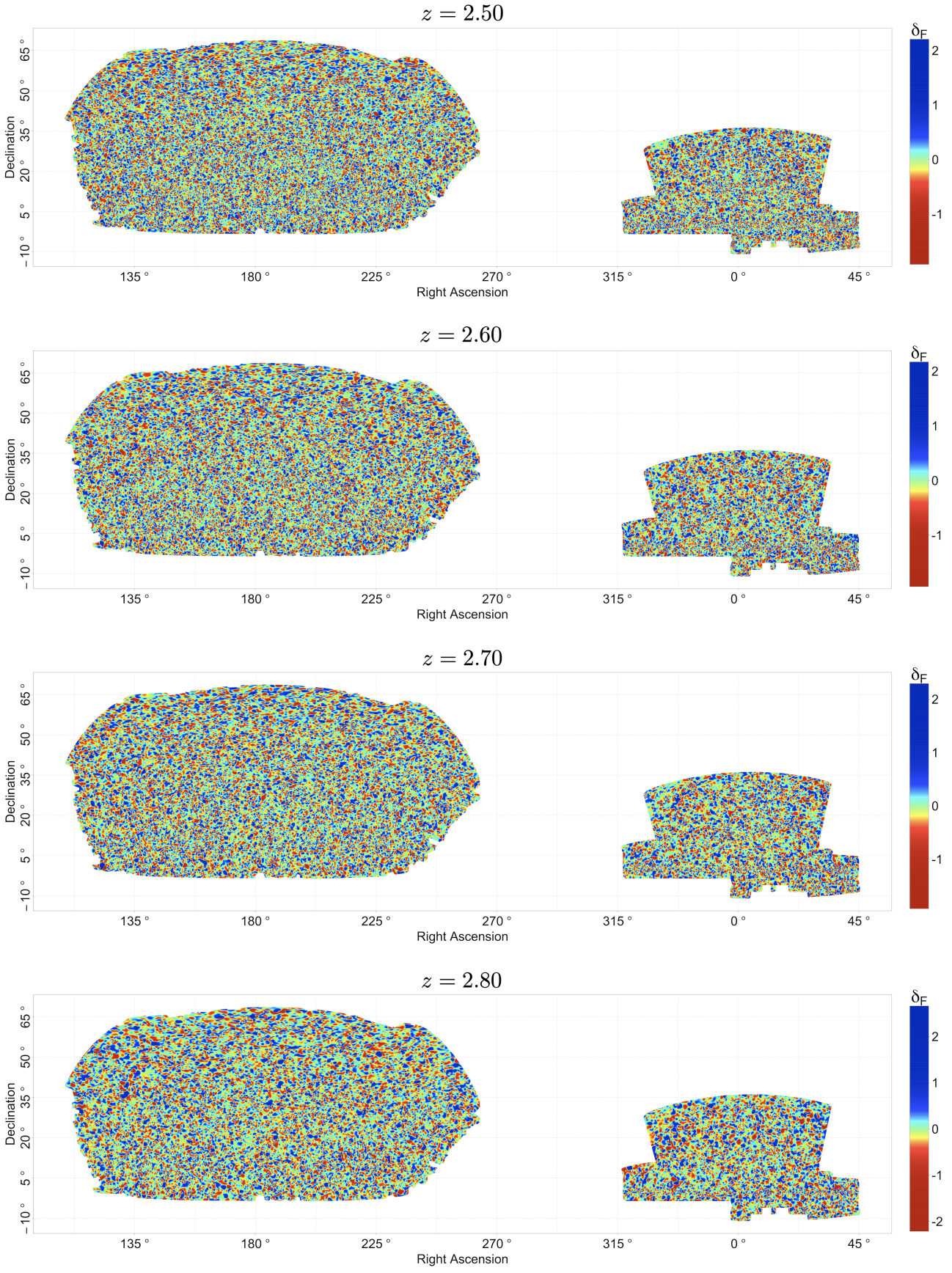


FIGURE 4.24: Two-dimensional sky map cross sections (continued).

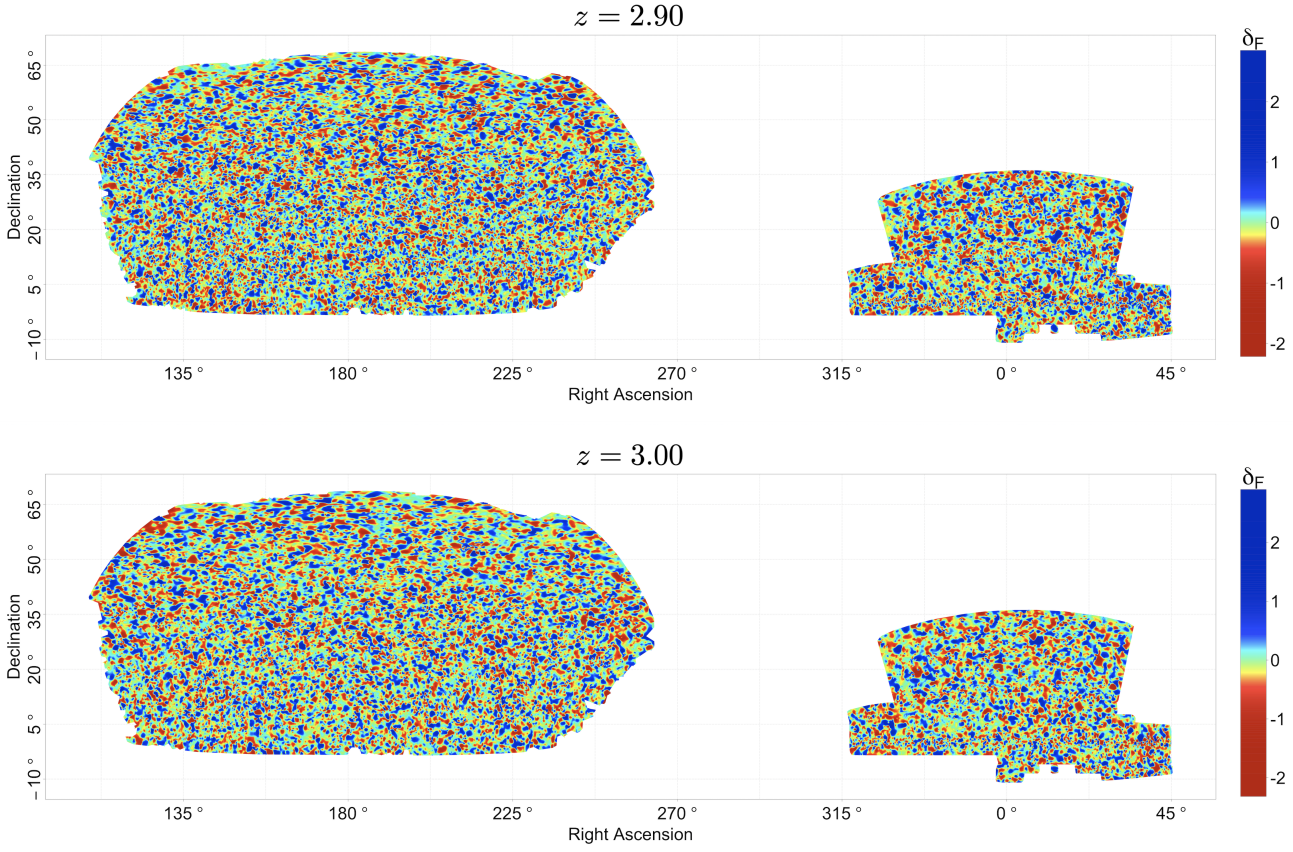


FIGURE 4.25: Two-dimensional sky map cross sections (continued).

Chapter 5

Three-dimensional cosmography of the high redshift Universe using intergalactic absorption: Mature Investigation

This chapter is based on our manuscript draft currently titled *Three-dimensional cosmography of the high redshift Universe using intergalactic absorption*, which was approved as a pre-submission inquiry at *Nature* in the spring of 2020 and will be submitted in full upon completion during the summer of 2020.

5.1 Introduction

The Lyman-α forest — a dense series of H I absorptions seen in the spectra of high redshift quasars — provides a unique cosmological probe of the large-scale structure of the redshift $z \gtrsim 2$ Universe [113, 163]. The density of observed quasars across the sky has recently risen to a level that, in principle, allows for a large-scale three-dimensional reconstruction of the full foreground matter density distribution transected by background quasar radiation [17–20]. However, until now, such a reconstruction has not been possible without the development of suitable statistical methods.

Using a sample of approximately 160,000 quasar sightlines collected by the SDSS-III Baryon Oscillation Spectroscopic Survey [2, 26, 27], here we present a 47 h^{-3} Gpc^3 Lyman-α absorption large-scale structure map, accompanied by rigorous error quantification and an extensive census of candidates for galaxy protoclusters and cosmic voids. The statistical reconstruction requires minimal assumptions on the underlying matter density field and is specifically optimized to recover three-dimensional structures lying between the one-dimensional sightlines backlit by quasars. The map covers approximately 25 percent of the sky and spans the redshift range $1.98 \leq z \leq 3.15$, allowing for studies of the cosmological matter distribution over unprecedented volumes.

As ultraviolet radiation from a quasar propagates through intergalactic space, photons at the Lyman-α (Lyα) wavelength of 1215.67 Å are scattered commensurate with the density of H I in the gaseous intergalactic medium (IGM). This realization was first reached with simple linear-theory models and later observed in numerical simulations [171, 242]. The gas causing forest absorption is pressure supported on small scales ($\sim$100 kpc), but on the $\gtrsim$10 Mpc scales studied in this work thermal pressure is negligible and the IGM traces the dark matter distribution with high accuracy [183, 184]. Winds from galaxies affect only a small fraction of the volume of intergalactic gas, and in simulations have been shown not to significantly disrupt the close overall relationship between forest absorption and matter density. By using the forest to map out structure in the Universe, we will therefore be tracing out the total distribution of gravitating matter.

The Lyα forest is now one of the standard cosmological probes used to measure the large-scale structure of the Universe and constrain cosmological parameters. Current and future sky surveys have significant elements devoted to the Lyα forest, and the number of quasars with suitably measured spectra has grown to over half a million [2, 189, 243]. The one-dimensional analysis of Lyα forest spectra, including the flux probablity distribution [13] and power spectrum [14–16],

provided some of the first cosmological results. As the number of surveyed sightlines has increased, the denser sampling has enabled three-dimensional clustering to be measured, first on $\lesssim 100\ h^{-1}$ Mpc scales [17], and then for measurements of baryon acoustic oscillations [BAOs; 18–20].

The fact that structure can be measured across sightlines means that interpolation techniques can be used to convert sets of one-dimensional sightlines into three-dimensional maps. The use of multiple Lyα forest sightlines to make three-dimensional maps of the IGM was pioneered by [21], in which the authors used N-body simulations of the IGM to demonstrate that the full three-dimensional Lyα absorption field can in principle be recovered given a dense set of quasar sightlines. In that case, extremely dense sampling was considered, and the motivation was to recover the predicted structure of the IGM over small volumes ($\sim$30,000 Mpc^3). [22] continued this theoretical work, using a Wiener interpolation applied to simulations to showed that filamentary structure can, in principle, be recovered given enough spectra. [244] explored the observational requirements in detail and [245] showed that galaxy protoclusters — large-scale overdensities that will evolve into galaxy clusters at lower redshifts [185] — can be recovered with high fidelity from such reconstructed maps. [25] demonstrated a different, nonparametric map-making technique — local polynomial smoothing — applying it to simulations and showing that it enables an investigation of the topology of the structure of the Universe from the Lyα forest at these previously unaccessible redshifts.

The first published observational three-dimensional map of the IGM from the Lyα forest was made by [23]. Those authors used star-forming galaxies instead of quasars as background spectra to achieve an extremely high (megaparsec scale) resolution, over a sky area of 70 square arcmin. In the present paper, we will focus on mapping a sky area approximately 1.5×10^5 larger, but with resolution of the order of $\sim$17.5 Mpc. The BOSS DR12 quasar Lyα dataset that we use was primarily designed for BAO measurement, and so covers appropriately large volumes. As no other current surveys of tracers with a higher space density than quasars exist that cover large

fractions of the sky at these ($z > 2$) redshifts, three-dimensional Lyα forest spectra offer a way to make the highest fidelity maps of these large comoving volumes. Such maps can enable us to explore the topology and clustering of the intergalactic medium, and by extension the dark matter structure which underlies it. We can also identify galaxy protoclusters and cosmic voids that have not been mapped before, and so for the first time carry out the type of science at higher redshifts that has been the domain of large galaxy surveys at redshifts $z < 1$.

5.2 Methods

5.2.1 First data reduction

First we report the BOSS quasar spectra that are entirely discarded from our sample and then we discuss the masking of individual pixels. We discard all spectra of quasars that are $\gtrsim$15 deg (150 h^{-1} Mpc at $z = 1.98$) outside the two contiguous regions of the sky targeted for mapping, where we retain the buffer to aid in reducing statistical boundary effects in the three-dimensional reconstruction near the edges of the targeted footprint. We remove all damped Lyα systems, Lyman-limit systems, and broad absorption line quasars. With respect to the SDSS spectral warning bitmask (`ZWARNING`[1]), we retain only the spectra with the bitmask integer equal to 0 or 16, indicating spectra with either no known issues or those only flagged as `MANY_OUTLIERS`, which is almost always indicative of a high signal-to-noise spectrum and no actual error. Finally, we retain only the spectra with `zerr` < 0.005, i.e. those for which the spectroscopic redshift of the quasar has been estimated to high precision by the BOSS spectroscopic pipeline [66].

Within each individual quasar spectrum, we consider a truncated restframe Lyα forest $\overline{\Lambda}_{\rm rest} = (1045\ \text{Å}, 1195\ \text{Å})$ in order to remove the Lyα and Lyβ emission peaks from the window, which allows for higher precision nonparametric estimation of the unabsorbed quasar continua [31].

[1] `https://www.sdss.org/dr12/algorithms/bitmasks/#ZWARNING`

Therefore, we mask all pixels with $\lambda_{\rm rest} \notin \overline{\Lambda}_{\rm rest}$. In regard to the SDSS pixel warning bitmask (`SPPIXMASK`[2]) we mask all pixels that have been flagged as potentially problematic in all exposures (`and_mask` $\neq 0$) or in any single exposure (`or_mask` $\neq 0$) or those declared altogether untrustworthy (`ivar` $= 0$). Finally, we mask pixels within 1.5×10^{-4} dex (in units of log base 10 Angstroms) of the strong sky lines documented in the data release 12 sky mask[3,4]. For the $\Delta \log_{10}(\lambda) = 10^{-4}$ dex pixel spacing of the SDSS and BOSS spectrographs [109], this corresponds to 1.5 spectral pixels masked on either side of each bright sky line. The reduced set of quasar Lyα forests are individually processed to the flux contrast scale with the procedure detailed below before undergoing an additional reduction that yields the sample ultimately used for three-dimensional modeling.

5.2.2 Transforming spectra to flux contrast

Three-dimensional reconstruction of Lyα forest absorption fields relies crucially on the ability to first accurately estimate the unabsorbed continuum of each coadded quasar Lyα forest and the broader procedure for transforming the raw flux observations to the Lyα flux contrast scale that traces H I density fluctuations. Furthermore, accurate statistical uncertainties at the stage of three-dimensional reconstruction necessarily must account for the added uncertainty introduced by the stochastic transformation to Lyα flux contrast. Our approach to processing the individual quasar spectra is described in detail in our previous work [29, 31] (Chapters 2 and 3), so we summarize it briefly here.

Given a BOSS spectrum of a quasar spectroscopically confirmed at redshift $z = z_0$, let $\Lambda(z_0) = (\lambda_{{\rm Ly}\beta},\ \lambda_{{\rm Ly}\alpha}) \cdot (1 + z_0)$ denote the redshifted Lyα forest interval that is truncated at the advent of the Lyβ forest. After correcting for extinction systematics [126, 127], the observational data

[2] `https://www.sdss.org/dr12/algorithms/bitmasks/#SPPIXMASK`
[3] `https://github.com/igmhub/picca/blob/master/etc/dr12-sky-mask.txt`
[4] `https://trac.sdss3.org/wiki/BOSS/LyaForestsurvey/SkyMask`

generating process (DGP) of the quasar flux in the Lyα forest can be assumed to follow the model

$$f(\lambda_i) = f_0(\lambda_i) + \epsilon_i, \qquad \lambda_i \in \Lambda(z_0), \tag{5.1}$$

$$= \overline{F}(\lambda_i) \cdot C(\lambda_i) \cdot \big(1 + \delta_F(\lambda_i)\big) + \epsilon_i, \tag{5.2}$$

where $f(\lambda_i)$ is the coadded flux at wavelength λ_i, $f_0(\cdot)$ is the flux signal, $\{\epsilon_i\}_{i=1}^n$ are uncorrelated Gaussian measurement errors attributable to photon noise, CCD readout noise, and sky-subtraction error (assumed $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma_i^2$), $\{\lambda_i\}_{i=1}^n$ form an equally spaced wavelength grid in log-space with $\Delta \log_{10}(\lambda_i) = 10^{-4}$ dex log-Angstroms, z_0 is the redshift of the quasar, $C(\cdot)$ is the flux of the unabsorbed quasar continuum, $F(\cdot) = f_0(\cdot)/C(\cdot)$ is the transmitted flux fraction, $\overline{F}(\lambda_i) = \mathbb{E}[F(\lambda_i)]$ is the mean Lyα transmitted flux fraction at redshift $z_i = \lambda_i/\lambda_{\mathrm{Ly}\alpha} - 1$, and $\delta_F(\lambda_i) = F(\lambda_i)/\overline{F}(\lambda_i) - 1$ is the transmitted flux contrast at redshift $z_i = \lambda_i/\lambda_{\mathrm{Ly}\alpha} - 1$. The flux contrast δ_F inversely traces H I density fluctuations in the intervening intergalactic medium and, by definition, is mean zero across the sky at each fixed redshift, with negative contrasts corresponding to H I densities above the cosmic mean and positive contrasts corresponding to H I densities below the cosmic mean.

We use trend filtering [1] to estimate the flux signal f_0 in each coadded quasar spectrum and local polynomial regression (LOESS) to estimate the smooth mean flux level $m(\lambda_i) = \overline{F}(\lambda_i) \cdot C(\lambda_i)$. We then define the Ly$\alpha$ flux contrast estimates along each quasar sightline as

$$\widehat{\delta}_F(z_i) = \frac{\widehat{f_0}(z_i)}{\widehat{m}(z_i)} - 1, \qquad z_i = \lambda_i/\lambda_{\mathrm{Ly}\alpha} - 1, \tag{5.3}$$

where $\widehat{f_0}$ is the trend filtering estimate of f_0 and $\widehat{m}$ is the LOESS estimate of m. We refer the reader to our previous work for details on the tuning of the model hyperparameters. An example of this transformation of the forest to the flux contrast scale is shown in Figure 5.1. Both

trend filtering and LOESS are nonparametric statistical methods that allow for highly flexible estimators of f_0 and m, respectively. In addition to the advantages discussed in previous work, defining our flux contrast estimates along the sightlines as in equation (5.3) before pooling the sightlines for a three-dimensional analysis yields two key benefits. First, denoising the observed spectra via trend filtering prior to normalizing with respect to the estimated mean flux level keeps the width of the $\widehat{\delta}_F$ sampling distribution significantly narrower than if we simply defined the flux contrast estimator as a normalization of the noisy observational spectrum. This specification is particularly important to make optimal use of the large number of faint $z \gtrsim 2.1$ quasar spectra collected by BOSS (i.e. when the mean flux level m is near zero), and in turn leads to a more precise three-dimensional absorption field reconstruction and a higher accompanying signal-to-noise ratio. Second, our use of the nonparametric LOESS estimator of the Lyα forest mean flux level m, instead of commonly used parametric estimators based on low-dimensional linear combinations of quasar spectral templates, significantly reduces the statistical bias of the intrinsic quasar continuum estimator. Statistical biases in the quasar continuum estimator are highly undesirable in this setting because they propagate as systematic biases into the data used for three-dimensional reconstruction. Therefore, by prioritizing low bias in the continuum estimator we can be more confident that significant deviations from the cosmic mean in the reconstructed three-dimensional absorption field — in particular, our candidates for galaxy protoclusters and voids — are real fluctuations in the density field, and not due to systematic biases originating from misspecification of the quasar continuum models. Furthermore, the nonparametric flexibility of the LOESS estimator also provides robustness to potential errors in the extinction corrections, which could otherwise introduce systematic biases into the reconstruction in the same way as biased continuum models.

The assumed Gaussian error distribution of the observational Lyα forest DGP, complete with the estimated measurement variances $\{\widehat{\sigma}_i^2\}_{i=1}^n$ provided by the BOSS spectroscopic pipeline [66],

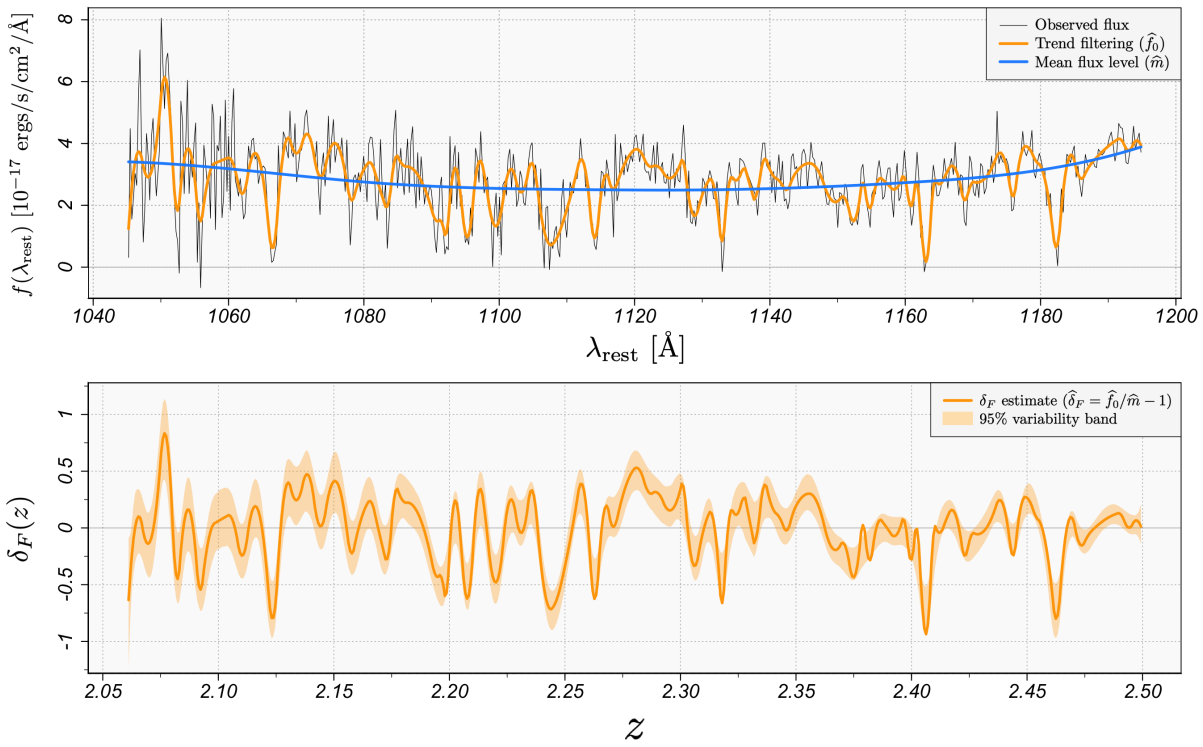


FIGURE 5.1: **Top**: Lyman-α forest of a quasar spectrum (in the restframe) from the twelfth data release of the Baryon Oscillation Spectroscopic Survey ([27]; Plate = 6487, MJD = 56362, Fiber = 647). The quasar is located at (RA, Dec, z) $\approx$ (196.680°, 31.078°, 2.560). A trend filtering estimate for the flux signal and a local linear regression estimate for the mean flux level are overlaid. **Bottom**: The estimated relative fluctuations in the Lyman-α transmitted flux fraction, due to the presence of absorbing neutral hydrogen in the intergalactic medium (shown in redshift space). A 95% variability band constructed from the parametric bootstrap procedure is superposed.

can then be sampled and propagated forward to determine the sampling distributions of the sightline flux contrast estimators in equation (5.3) and therefore obtain the uncertainty of the measurements on the transformed scale. We approximate the observational DGP of each coadded quasar spectrum with a parametric bootstrap [246] with samples drawn independently from the Gaussian distribution

$$f^*(\lambda_i) \sim N(\widehat{f}_0(\lambda_i), \widehat{\sigma}_i^2), \qquad i = 1, \ldots, n. \tag{5.4}$$

As discussed in the sections below, this parametric bootstrap approximation of each Lyα forest DGP is central to both our construction of an optimal pixel-weighting scheme for fitting the three-dimensional model and our further propagation of uncertainty through to the final stage

of three-dimensional reconstruction, thereby providing error estimates for the reconstructed absorption field that account for all statistical uncertainties in the full modeling pipeline.

Along one-dimensional quasar sightlines the resolution of the SDSS and BOSS twin spectrographs allows for studies of the intervening Lyα absorption field down to $\sim$1 h^{-1} Mpc scales — well below the scale on which three-dimensional reconstruction over large volumes is currently possible. It is therefore useful to formally define a decomposition of the underlying absorption field into two orthogonal signals

$$\delta_F(\alpha, \delta, z) = \delta_F^L(\alpha, \delta, z) + \delta_F^S(\alpha, \delta, z), \tag{5.5}$$

where δ_F^L is the large-scale structure of the Lyα absorption field δ_F — indexed here by angular equatorial coordinates (α, δ) and redshift z — and δ_F^S is the small-scale structure of δ_F. More precisely, let $r_\perp(z)$ be the comoving mean transverse sightline separation at redshift z, restricted to the set of unique background quasars in the sample and marginalized over the sky. We define $\delta_F^L(\alpha, \delta, z)$ to contain all frequencies of the Fourier series of δ_F (assumed to vary spatially) that are above the sky-averaged transverse scale $r_\perp(z)$ and we define $\delta_F^S(\alpha, \delta, z)$ to contain the frequencies that are below $r_\perp(z)$. We can approximately regard δ_F^L as the large-scale absorption field structure that is recoverable across sightlines given the density of background quasars across the sky. For this reason, we will denote our three-dimensional model $\widehat{\delta}_F^L$. For the sake of validating the three-dimensional model, which we discuss below, it is useful to first define a one-dimensional estimate of δ_F^L along each quasar sightline. Specifically, letting $\widehat{\delta}_F(z)$ be the continuous-time representation of the sightline Lyα flux contrast estimate defined in equation (5.3), we construct an estimate of δ_F^L along each sightline by computing a sliding-window wavelet transform of $\widehat{\delta}_F(z)$ and hard-thresholding the levels of the wavelet decomposition at each z that produce power at higher frequencies than $r_\perp(z)$ (in h^{-1} Mpc). We denote this one-dimensional estimate $\widetilde{\delta}_F^L$ in order to distinguish it from the three-dimensional reconstruction itself.

5.2.3 Second data reduction

Here we apply one further reduction to the processed set of spectra before proceeding with three-dimensional modeling. We discard all spectra for which at any point the estimated mean flux level $\widehat{m}$ becomes nonpositive, as well as any spectrum for which the flux signal estimate $\{\widehat{f_0}(z_i)\}_{i=1}^{n}$, reduced to a sliding-window resolution of $r_{\perp}(z_i)$, contains ten or more nonpositive individual pixels. We mask all pixels outside of the targeted redshift range $1.98 \leq z \leq 3.15$ plus a 150 h^{-1} Mpc buffer on each side. Finally, for any individual pixel with $\widehat{\mathrm{Var}}^{-1}(\widehat{\delta}_F) > 150$, as estimated via the parametric bootstrap procedure, we threshold the inverse variance to be exactly 150 to prevent a small fraction of the data set from having excessive leverage on the three-dimensional model.

Altogether, our sample used for three-dimensional reconstruction includes 159,581 spectra from 142,696 distinct quasars, totaling over 72 million processed Lyα flux contrast pixels and constituting densities of $\sim$15.4 sightlines/deg^2 and $\sim$13.8 quasars/deg^2 over the 10,332 deg^2 sky area targeted for mapping. The primary observational limitation on the effective spatial resolution of the three-dimensional map is the local transverse sightline separation of distinct background quasars. This quantity varies significantly as a function of redshift, both as quasars become increasingly faint on the high redshift end and as ultraviolet Lyα wavelengths become atmospherically opaque on the low redshift end. The sightline density of our sample peaks at $z \sim 2.16$, with a mean transverse separation of $r_{\perp}(2.16) = 10.7\ h^{-1}$ Mpc, compared to $r_{\perp}(1.98) = 12.2\ h^{-1}$ Mpc and $r_{\perp}(3.15) = 31.7\ h^{-1}$ Mpc at the extremes of our selected redshift range. At each fixed redshift, the sampling density of background quasars is approximately uniform across the sky with the primary exception being Stripe 82 — a 220 deg^2 region along the celestial equator (declination $\delta \approx 0^{\circ}$) in the Southern Galactic Cap that was repeatedly targeted to achieve higher source densities than the rest of the BOSS footprint. The mean comoving sightline density of Stripe

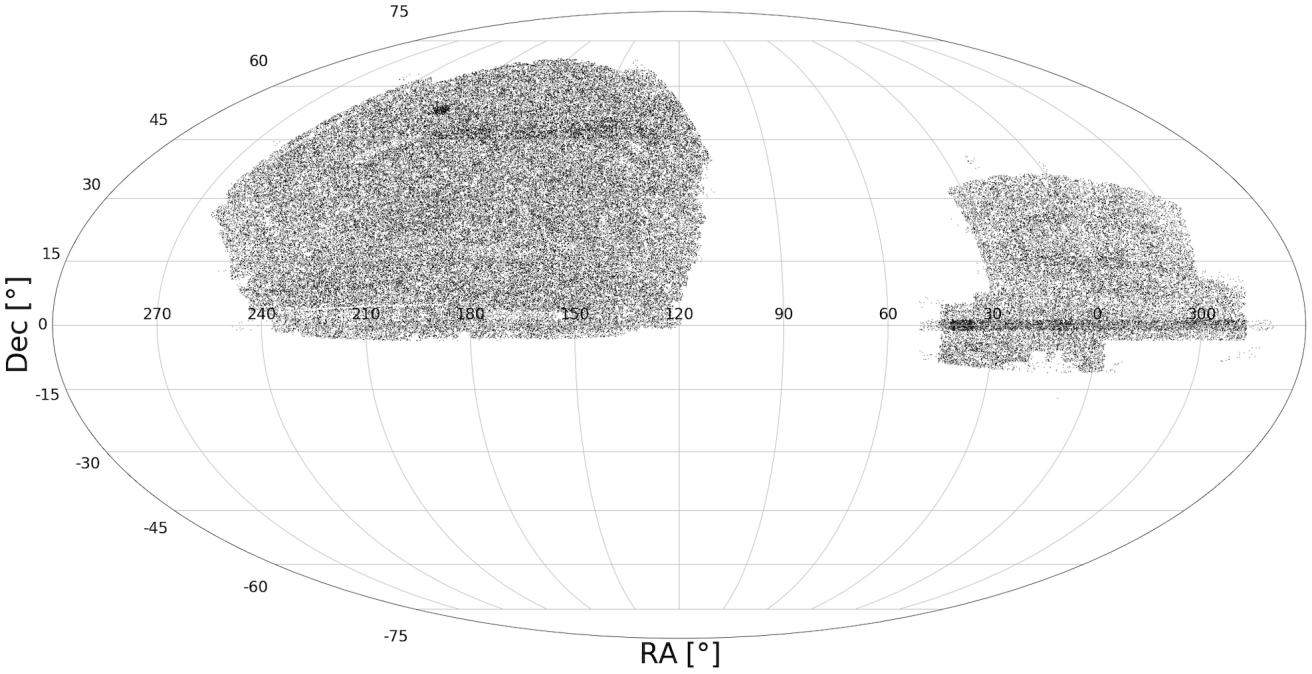


FIGURE 5.2: Footprint of the BOSS Lyα quasars used for the three-dimensional Lyα absorption field reconstruction in this chapter, shown in equatorial coordinates. Altogether, our sample used for three-dimensional reconstruction includes 159,581 spectra from 142,696 distinct quasars, constituting densities of $\sim$15.4 sightlines/deg^2 and $\sim$13.8 quasars/deg^2 over the 10,332 deg^2 sky area targeted for mapping.

82 quasars in our sample peaks at $r_\perp(2.24) = 7.2\ h^{-1}$ Mpc. The redshift distributions of the aggregated set of processed Lyα flux contrast pixels and the mean sightline separation of unique quasars in the final sample are shown in Figure 5.3. As detailed below, our three-dimensional absorption field reconstruction adapts to the local variations in the sightline density so the underlying absorption field is reconstructed at the highest spatial resolution allowed by the local density of background quasars.

5.2.4 Three-dimensional absorption field reconstruction

Given the collection of processed Lyα flux contrast measurements lying along sightlines, here we detail a novel and scalable nonparametric statistical technique for reconstructing the full three-dimensional Lyα forest absorption field — a problem often referred to as *Lyman-α forest tomography* or *intergalactic medium tomography*. We assume a spatially flat Λ-Cold Dark Matter

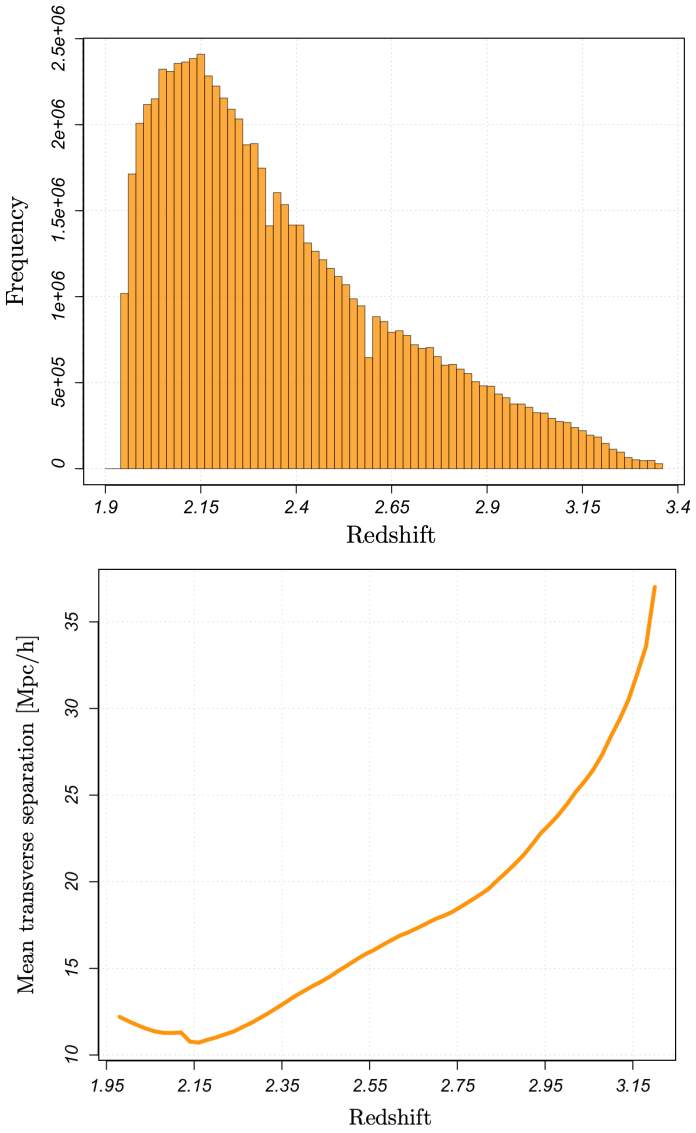


FIGURE 5.3: **Top:** Redshift distribution of the ∼72 million Lyα flux contrast pixels in our sample used for three-dimensional reconstruction of the 47 h^{-3} Gpc^3 absorption field over the redshift range $1.98 < z < 3.15$. **Bottom:** Comoving mean transverse sightline separation of the unique set of quasars in our sample used for three-dimensional reconstruction. This quantity is the primary constraint on the effective spatial resolution at which we are able to reconstruct the three-dimensional density field of the IGM across sightlines, and the resolution of the map therefore varies significantly across redshifts.

[170] (ΛCDM) cosmological model with parameters $\Omega_m = 0.315$, $\Omega_\Lambda = 0.685$, and $h = 0.674$. As we detail below, our three-dimensional model – a spatially-penalized kernel ridge regression (SKRR) on a multi-resolution set of basis functions – provides a number of statistical and computational advantages over the Wiener filtering approach that has been popularized for Lyα tomographic mapping on hydrodynamical simulations and small observational volumes [21–24]. The advantages of our proposed methodology fall into three main categories:

1. **Nonparametric/data-driven modeling.** We require no prior distributional assumptions on the underlying Lyα absorption field. Instead, we focus on directly learning the spatial dependence that is most predictive of the absorption structure lying *in between* the sightlines backlit by quasars. Furthermore, we make no simplifying assumptions on the sampling design of the observations (e.g. treating sightlines as parallel).

2. **Uncertainty quantification.** By further extending the parametric bootstrap procedure introduced in our previous work, we are able to provide rigorous estimates of the total statistical uncertainty in the reconstructed three-dimensional Lyα absorption field. In total, we produce 50 bootstrap realizations of the full 47 h^{-1} Gpc^3 map, with the variation in the bootstrap distribution accounting for the inherent observational uncertainty in the quasar spectra and all subsequent statistical uncertainties introduced by the pipeline of analyses used to reconstruct the full three-dimensional absorption field. This total statistical uncertainty — which we find to be in close agreement with a Gaussian — will allow for rigorous statistical error bounds on all future analyses of the map and in particular, here we utilize it to compile a census of high-significance candidates for high redshift galaxy protoclusters and cosmic voids.

3. **Computational efficiency.** We develop a parallelized approximation algorithm distributed over the spatial domain and enforce sparsity in the matrices central to computing

each local absorption field reconstruction in a way that allows us to carry out a spatial analysis of this magnitude, while also not compromising the complex spatial dependence needed to faithfully reconstruct the full absorption field down to the scale of the transverse quasar sightline separation. The computational efficiency and scalability provided by these two components in tandem is the central ingredient that allows us to: (i) reconstruct a three-dimensional Lyα forest absorption field over an unprecedented 47 h^{-1} Gpc^3 volume; (ii) optimize the accuracy of the reconstruction over a large number of hyperparameter combinations; and (iii) quantify the total statistical uncertainty of the reconstruction by producing 50 bootstrap reconstructions of the full absorption field.

The large-scale structure in Lyα forest absorption fields, which we have denoted δ_F^L, can be viewed as a spatial process with multiple scales of structure evolved from scale-invariant density fluctuations in the early Universe convolved with the multi-scale baryon acoustic oscillation (BAO) signature imprinted by pre-recombination acoustic waves [194]. It is therefore appropriate to regard reconstruction of three-dimensional Lyα absorption fields as a problem of multi-resolution analysis. Our three-dimensional SKRR model makes use of Euclidean distances; therefore, using the comoving radial distances computed under the assumed ΛCDM cosmological model, we convert the $n \approx 7.2 \times 10^7$ sightline flux contrast measurements defined in equation (5.3) to three-dimensional Cartesian coordinates through the usual spherical-to-Cartesian coordinate transformation. We consider a model space given by a multi-resolution basis expansion in three-dimensional Euclidean space

$$\delta_F^L(x) = \sum_{\ell,j} \beta_{\ell,j}\, \phi_{\ell,j}(x), \qquad x \in \mathbb{R}^3, \tag{5.6}$$

where $\{\phi_{\ell,j}\}_{j=1,\ldots,m_\ell}^{\ell=1,\ldots,d}$ is a multi-resolution set of compactly-supported radial basis functions (RBFs) organized on regular grids and $\beta_{\ell,j}$ are the basis coefficients that, given the choice of

basis, parametrize the three-dimensional model. We fix the finest level of RBFs to have a node separation just below the scale of the mean transverse quasar sightline separation, so the model possesses sufficient resolution to recover the large-scale absorption field δ_F^L across sightlines but will not overadapt to the small-scale structure δ_F^S that can be seen along each densely-sampled sightline. Each coarser level of the multi-resolution basis is then constructed by dyadically increasing the node separation of the previous level, and we let the total number of levels in the basis be a tunable hyperparameter of the model. At each level of resolution $\ell = 1, \ldots, d$, we construct the RBFs via a real-valued radial function

$$\phi_{\ell,j}(x) = \phi\left(\frac{\|x - u_{\ell,j}\|_2}{\theta_\ell}\right), \qquad j = 1, \ldots, m_\ell, \tag{5.7}$$

where $u_{\ell,1}, \ldots, u_{\ell,m_\ell} \in \mathbb{R}^3$ are the nodes of the RBFs organized on a regular three-dimensional grid, $\|\cdot\|_2$ is the Euclidean ℓ_2 norm, and θ_ℓ is a scaling parameter to adjust the amount of overlap in the RBFs at each level. For $\phi(\cdot)$ we adopt the compactly-supported Wendland polynomial [238]

$$\phi(y) = \begin{cases} (1-y)^6 \cdot (35y^2 + 18y + 3))/3 & 0 \leq y < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{5.8}$$

and, following literature on multi-resolution fixed rank kriging [227], we let the scale parameter θ_ℓ of each RBF be equal to 2.5 times the node spacing at its respective basis level, which ensures that the reconstruction resembles a continuous multi-resolution field with no visible artifacts from the finite basis. A one-dimensional example of the multi-resolution Wendland basis is shown in Figure 5.4.

Given the multi-resolution linear model introduced in equation (5.6), the problem of reconstructing the three-dimensional Lyα forest absorption field reduces to a problem of jointly optimizing the number of multi-resolution levels in the basis and the corresponding set of basis coefficients. Here,

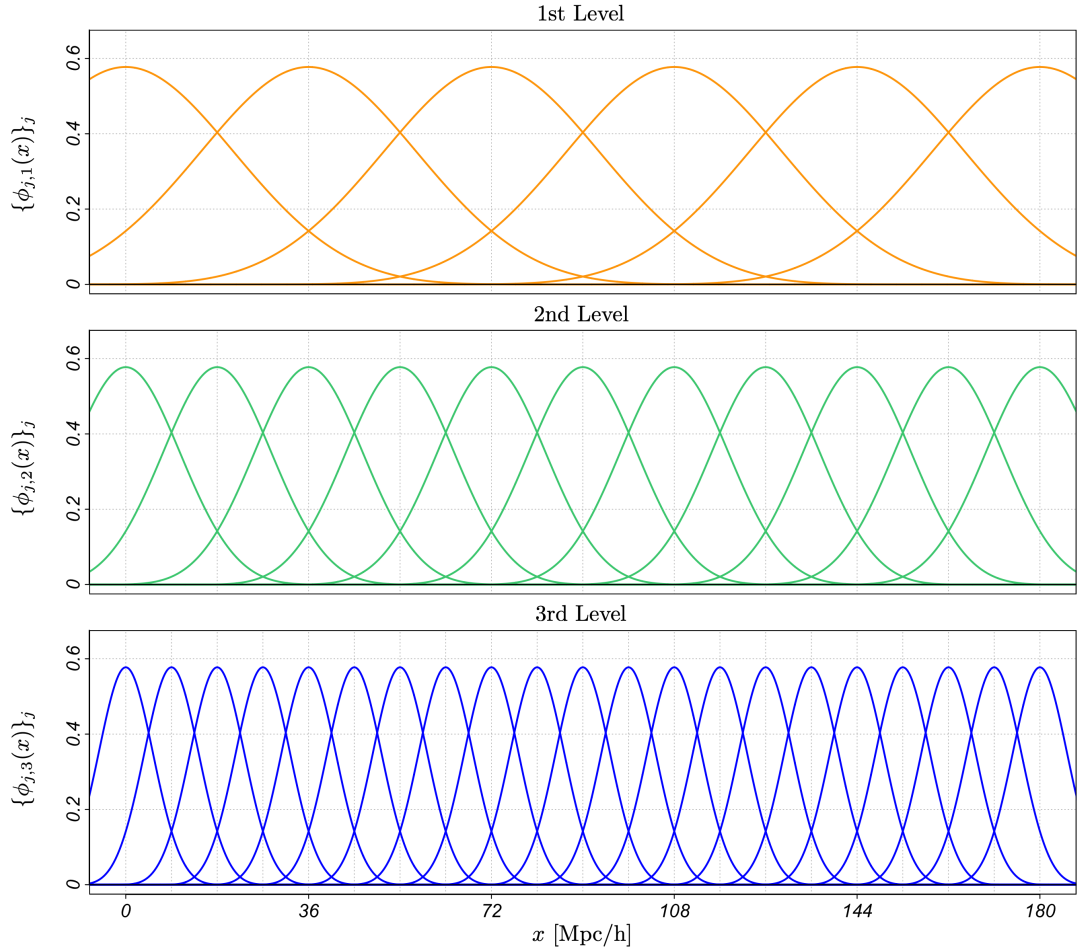


FIGURE 5.4: One-dimensional example of the multi-resolution Wendland basis used by the SKRR model. The three-level basis shown here corresponds to the optimized SKRR model on the redshift bin $z_1 = [1.98, 2.64)$.

we pose that for Lyα forest tomography, an optimal reconstruction should be taken to mean one that recovers the underlying absorption field structure as accurately as possible *in between* the one-dimensional sightlines probed by background quasars. We therefore approach the problem as one of optimizing the *out-of-sample* predictive accuracy of the model on held-out background quasars. In order to simplify notation, let $\{\phi_1, \ldots, \phi_m\}$ denote our multi-resolution basis with $m = \sum_{\ell=1}^{d} m_\ell$, and let $\beta \in \mathbb{R}^m$ be the corresponding vector of basis coefficients. Holding the multi-resolution basis fixed, the optimization that determines the basis coefficients of our SKRR

model can be concisely stated as a generalized ridge regression minimization problem

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^m}{\operatorname{argmin}} \, (\widehat{\delta}_F - \Phi\beta)^T W (\widehat{\delta}_F - \Phi\beta) + \|D\beta\|_2^2, \tag{5.9}$$

where $\widehat{\delta}_F \in \mathbb{R}^n$ is a vector of the aggregated flux contrast estimates defined in equation (5.3), $\Phi \in \mathbb{R}^{n \times m}$ with $\Phi_{ij} = \phi_j(x_i)$ is the regression matrix with respect to the basis, W is a diagonal pixel-weighting matrix with elements

$$W_{ii}^{-1} = \mathbb{E}[(\widehat{\delta}_F(x_i) - \delta_F^L(x_i))^2], \qquad i = 1, \ldots, n \tag{5.10}$$

estimated by parametric bootstrapping the quantity $(\widehat{\delta}_F - \widetilde{\delta}_F^L)^2$ along each sightline, and $D \in \mathbb{R}^{m \times m}$ is a structured penalty matrix that enforces spatial dependence among the basis coefficients at each level of resolution. The cost functional in equation (5.9) therefore states that the SKRR solution balances a tradeoff between minimizing the weighted sum of squared error with respect to the Lyα flux contrast estimates lying along observed sightlines and enforcing smoothness in the full three-dimensional reconstruction as prescribed by the generalized ridge penalty matrix D. We construct the penalty matrix to be block diagonal with blocks $D_\ell \in \mathbb{R}^{m_\ell \times m_\ell}$, $\ell = 1, \ldots, d$, where each block D_ℓ determines the spatial dependence of the basis coefficients at the corresponding level of resolution and the relative contribution of that level of resolution to the total model. Specifically, we define each block of the penalty matrix as

$$D_\ell = \sqrt{\gamma_\ell}(\Delta_\ell + \alpha_\ell I), \quad \gamma_\ell \geq 0, \; \alpha_\ell > 0, \tag{5.11}$$

where $\Delta_\ell \in \mathbb{R}^{m_\ell \times m_\ell}$ is a sparse three-dimensional fusion penalty that enforces spatial dependence among neighboring basis coefficients, I is the identity matrix that corresponds to the traditional pure-ridge shrinkage of the basis coefficients, α_ℓ is a range hyperparameter that determines the

range of the spatial dependence at level ℓ, and γ_ℓ is a penalty hyperparameter that determines the relative contribution of level ℓ to the total model. More explicitly, at each level ℓ, the fusion penalty is defined as

$$\Delta_{\ell_{ij}} = \begin{cases} 6 & i = j, \\ -1 & j \in \mathcal{N}_i, \\ 0 & \text{otherwise}, \end{cases} \tag{5.12}$$

where $\mathcal{N}_i$ is the set of indices of the basis nodes at level ℓ that are first-order neighbors of the node $u_{\ell,i}$. The decomposition of the spatial penalty follows as

$$\|D\beta\|_2^2 = \gamma_\ell \Big(\sum_{\ell=1}^{d} \beta_\ell^T \Delta_\ell^2 \beta_\ell + 2\alpha_\ell \beta_\ell^T \Delta_\ell \beta_\ell + \alpha_\ell^2 \beta_\ell^T \beta_\ell \Big), \tag{5.13}$$

where the first term consists of second-order differences between neighboring basis coefficients, the second term consists of first-order differences between neighboring basis coefficients, and the third term is the traditional pure-ridge sum-of-squared coefficients. We find the optimal spatial penalty to be consistent with a scale-invariant range hyperparameter $\alpha_\ell \equiv \alpha$, which in turn is helpful in reducing the dimensionality of the model hyperparameter space. Therefore, in total, the model possesses $d + 2$ free hyperparameters that allow the model to learn a highly complex spatial dependence. We discuss the validation of these hyperparameters in a dedicated section below.

We have thoroughly tested the optimality of the ℓ_2 (squared) norm in this setting as the penalization functional. The use of the ℓ_2 norm is favorable in this setting because the matter density distribution on the $\gtrsim 10\ h^{-1}$ Mpc scales studied in this work is spatially homogeneous (i.e. the smoothness is relatively uniform). As the density of background quasars continues to increase with future sky surveys and it becomes feasible to recover high-density filaments over large volumes in Lyα absorption fields, an ℓ_1 norm on the spatial penalty — corresponding to a

generalized lasso regression [28] — will provide an essential boost in spatial adaptivity. It will, however, require significantly greater computational expense.

The cost functional in equation (5.9) is strictly convex and differentiable everywhere so, for any fixed set of hyperparameters, the solution follows as

$$\widehat{\beta} = (\Phi^T W \Phi + D^T D)^{-1} \Phi^T W \widehat{\delta}_F \tag{5.14}$$

and is unique. By construction, $\Phi^T W \Phi$ and $D^T D$ are very sparse and symmetric positive definite so the solution can be computed very efficiently by utilizing a sparse Cholesky decomposition of $G = \Phi^T W \Phi + D^T D$ and specialized algorithms for the sparse matrix multiplications. Returning the basis coefficients to their original indexing, the SKRR three-dimensional reconstruction follows as

$$\widehat{\delta}_F^L(x) = \sum_{\ell,j} \widehat{\beta}_{\ell,j}\, \phi_{\ell,j}(x). \tag{5.15}$$

A two-dimensional example of the isotropic multi-resolution Mercer kernel of the SKRR estimator is shown in Figure 5.5 (with three levels of resolution).

An analysis of the effective smoothing kernel demonstrates that the SKRR model instrinsically adapts to both the local density of sightlines and the local signal-to-noise ratio of the pixels. It is useful, however, to provide additional design-adaptive flexibility to address the severe heterogeneity of background quasars as a function of redshift. Therefore, we partition the volume into three disjoint redshift bins $z_1 = [1.98, 2.64)$, $z_2 = [2.64, 2.96)$, $z_3 = [2.96, 3.15]$, and optimize the RBF spacing and model hyperparameters separately in each bin.

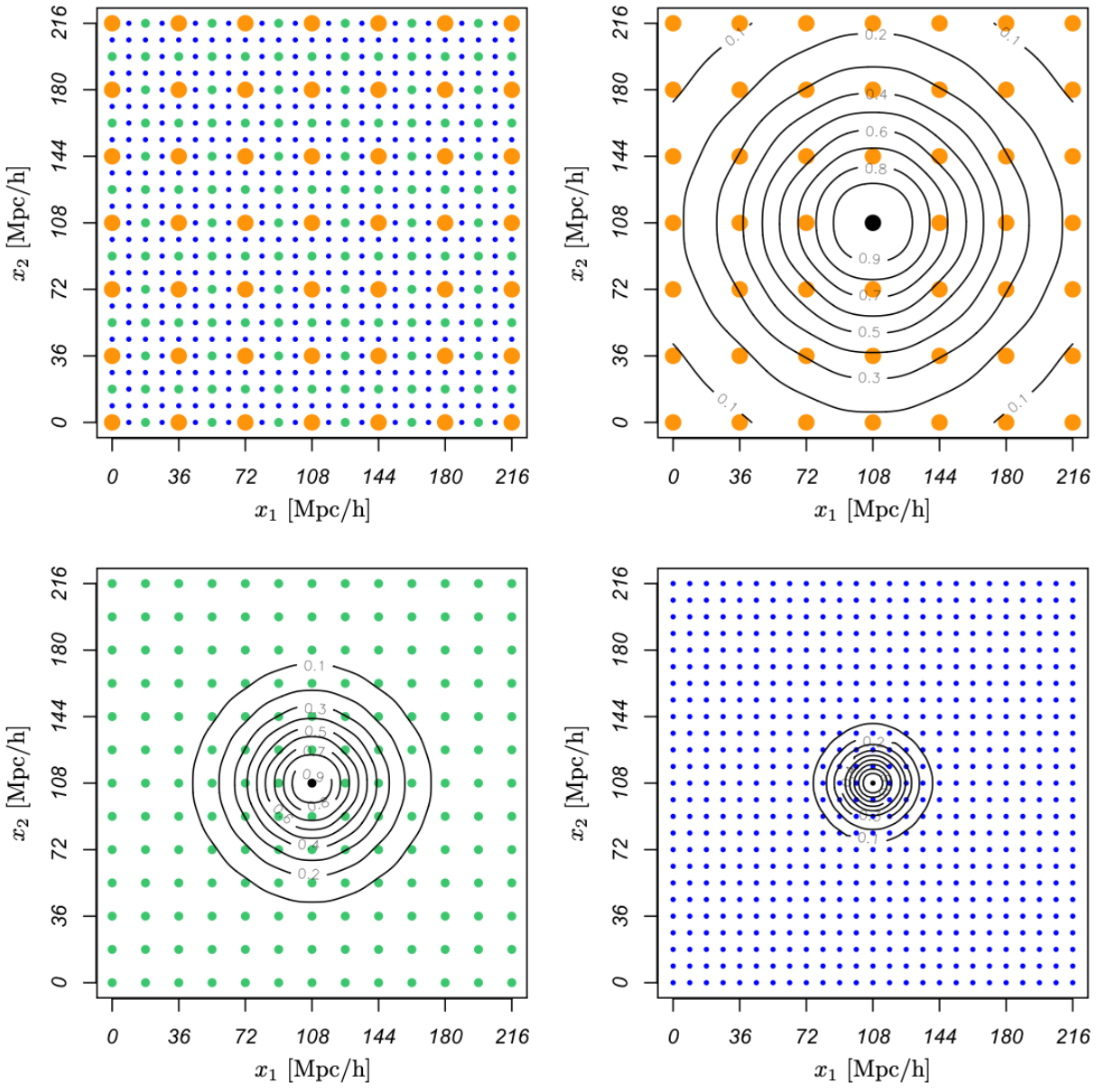


FIGURE 5.5: Two-dimensional example of the isotropic Mercer kernel induced by the SKRR model, with contours of the Mercer kernel shown for the central spatial location $(x_1, x_2) = (108, 108)$. The contours of each level of resolution in the Mercer kernel show the pairwise degree of similarity that is encouraged between each pair of points in the spatial reconstruction. A separate set of hyperparameters then controls the relative contribution of each level of spatial smoothness. The three-level Mercer kernel shown here corresponds to the optimized SKRR model on the redshift bin $z_1 = [1.98, 2.64)$.

5.2.5 Distributed computing

The calculation of the SKRR solution is dominated by the $\mathcal{O}(m^3)$ time complexity and $\mathcal{O}(m^2)$ memory complexity of the Cholesky decomposition of G. Our enforcement of extreme sparsity in the calculations allows problems on the order of $m \approx 3.0 \times 10^5$ to be solved efficiently on a single processor of a high-performance computing infrastructure. However, reconstruction of the full 47 h^{-1} Gpc^3 absorption field at the scale of the mean transverse sightline separation implies a total basis size of $4.0 \times 10^7 \lesssim m \lesssim 5.0 \times 10^7$. Therefore, in order to make the full reconstruction, hyperparameter validation, and uncertainty quantification computationally feasible, it is necessary to scale the SKRR estimator horizontally (i.e. over multiple machines) into a large-scale distributed system. Here, we develop a distributed approximation algorithm that allows us to adapt the SKRR model to the multi-machine setting, while maintaining close proximity to the computationally-infeasible global solution. This distributed algorithm operates locally by exploiting the trivial spatial dependence of the absorption field on $> 250\ h^{-1}$ Mpc scales. Let $\mathcal{F} \subset \mathbb{S}^2$ be the 10,332 deg^2 footprint targeted for mapping and let $C_j = \mathcal{F} \times z_j$, $j \in \{1, 2, 3\}$, be a partition of the 47 h^{-1} Gpc^3 volume into the three redshift bins over which the model hyperparameters are independently optimized. For each $j \in \{1, 2, 3\}$, let $B_{j,1}, \ldots, B_{j,r_j} \subset \mathbb{R}^3$ be a disjoint set of cubes in $\mathbb{R}^3$ such that

$$C_j \subset \bigcup_{k=1}^{r_j} B_{j,k}, \tag{5.16}$$

and for each cube $B_{j,k}$, let $\widetilde{B}_{j,k}$ be an augmented cube with a 250 h^{-1} Mpc buffer region added in each direction along the Cartesian coordinate axes. For each cube, we define a local absorption field reconstruction $\widehat{\delta}_F^{L,j,k}(x; \eta_j)$, $x \in \widetilde{B}_{j,k}$ by fitting the SKRR estimator with hyperparameter

vector $\eta_j = (d_j, \alpha_j, \gamma_{j,1}, \ldots, \gamma_{j,d})$ on the local basis

$$\mathcal{B}_{j,k} = \left\{ \phi_{\ell,j} : u_{\ell,j} \in \widetilde{B}_{j,k} \right\} \tag{5.17}$$

and the local set of observations

$$\mathcal{D}_{j,k} = \left\{ \left(x_i, \widehat{\delta}_F(x_i) \right) : x_i \in \widetilde{B}_{j,k} \right\}. \tag{5.18}$$

The distributed-SKRR three-dimensional reconstruction of the full absorption field then follows as the piecewise estimator

$$\widehat{\delta}_F^{L\|}(x) = \sum_{j=1}^{3} \sum_{k=1}^{r_j} \mathbb{1}(x \in A_{j,k}) \cdot \widehat{\delta}_F^{L,j,k}(x; \eta_j), \qquad x \in C, \tag{5.19}$$

where $A_{j,k} = B_{j,k} \cap C_j$ and $C = C_1 \cup C_2 \cup C_3$.

In total we utilize $1,664$ local models and construct the spatial partition so that each local model has a computationally-manageable $m \approx 3.0 \times 10^5$ basis functions. The use of buffer regions of 250 h^{-1} Mpc around each local model ensures that the distributed piecewise estimator closely approximates a continuous three-dimensional field with smooth low order derivatives. Although the use of buffer regions enables a close approximation of a continuous field, very small discontinuities can be visually resolved at some of the seams of the partition when examining the distributed reconstruction at the highest pixel resolution. Therefore, we post-process the final distributed-SKRR reconstruction with a Epanechnikov smoothing kernel with a bandwidth of $\zeta = 12$ h^{-1} Mpc, which we find to be sufficient to remove the small discontinuities while also not degrading the highest resolution structure recovered in the map.

The total computational expense of the analysis presented in this work was approximately 3.0×10^6 CPU hours – carried out a large array of machines, with each machine configured with

$\sim$200 GB of memory. The model validation and bootstrap uncertainty quantification of the three-dimensional absorption field reconstruction account for nearly all of the computational expense.

5.2.6 Model validation

In accordance with our objective to optimize the spatial dependence of the three-dimensional reconstruction to maximize predictive accuracy in between observed sightlines, we optimize all hyperparameters of the SKRR model by holding out 15% of the background quasars, training on the remaining 85%, and validating the accuracy of the three-dimensional reconstruction on the held-out set of one-dimensional absorption field skewers. Recall the one-dimensional estimate $\widetilde{\delta}_F^L$ of large-scale Lyα flux contrast, which we defined by degrading the high-resolution one-dimensional absorption field to match the scale of the transverse sightline separation. Let $\widetilde{\delta}_{F_i}^L$, $i = 1, \ldots, n'$ be the aggregated set of $\widetilde{\delta}_F^L$ pixels over all sightlines in the validation set and let $\widehat{\delta}_{F_i}^{L\|}$, $i = 1, \ldots, n'$ be the set of validation set predictions given by the distributed-SKRR model with hyperparameter vector $\eta = (d, \alpha, \gamma_1, \ldots, \gamma_d)$. We utilize the Kendall tau ranking distance to measure how far the SKRR model is from reproducing the discretized ordering of the flux contrast measurements corresponding to the unseen 15% of quasars in the validation set. In particular, letting $\rho(\cdot)$ be the sample rank operator, for any given hyperparameter vector η the Kendall tau ranking distance is given by

$$R(\eta) = |\mathcal{D}_1 \cup \mathcal{D}_2| \tag{5.20}$$

$$= |\mathcal{D}_1| + |\mathcal{D}_2| \tag{5.21}$$

where

$$\mathcal{D}_1 = \left\{(i,j)_{i<j} : \ \rho(\widehat{\delta}^{L\|}_{F_i}) < \rho(\widehat{\delta}^{L\|}_{F_j}) \ \cap \ \rho(\widetilde{\delta}^{L}_{F_i}) > \rho(\widetilde{\delta}^{L}_{F_j})\right\} \tag{5.22}$$

$$\mathcal{D}_2 = \left\{(i,j)_{i<j} : \ \rho(\widehat{\delta}^{L\|}_{F_i}) > \rho(\widehat{\delta}^{L\|}_{F_j}) \ \cap \ \rho(\widetilde{\delta}^{L}_{F_i}) < \rho(\widetilde{\delta}^{L}_{F_j})\right\}. \tag{5.23}$$

Optimizing the three-dimensional model with respect to the Kendall tau ranking distance therefore favors hyperparameter combinations that reproduce the ordering of the discretized large-scale absorption field in between the observed quasar sightlines — i.e. relative overdensities and underdensities in the large-scale absorption field are accurately predicted as such.

5.2.7 Uncertainty quantification

The total statistical uncertainty in the three-dimensional map can be quantified by studying the sampling distribution of the distributed-SKRR estimator, starting with the inherent observational uncertainty in the quasar spectra and propagating it through the subsequent stochastic transformation to Lyα flux contrast and finally to the stage of three-dimensional reconstruction. We utilize an extended version of the parametric bootstrap procedure described in our previous work [29, 31] (Chapter 3). Each of the 50 bootstrap reconstructions of the full 47 h^{-1} Gpc^3 Lyα absorption field is constructed as follows:

1. Construct a bootstrap sample of observational quasar spectra $f_1^*, \ldots, f_q^*$, $q = 159,581$ according to the parametric bootstrap Gaussian DGP in equation (5.4).

2. Define the Lyα flux contrast estimates along each sightline as in equation (5.3) using the trend filtering estimate of flux signal and the LOESS estimate of mean flux level fit to each spectrum in the bootstrap sample.

3. Pool the sightline flux contrast estimates and compute the bootstrap three-dimensional reconstruction $\widehat{\delta}_F^{L\|*}(x)$ via the distributed-SKRR model defined in equation (5.19).

The sampling distribution of the distributed-SKRR estimator does not directly inherit the Gaussianity of the observational error distributions of the quasar spectra because our intermediate transformation to the flux contrast scale is nonlinear. Nevertheless, a quantile-quantile analysis of the bootstrap sampling distribution reveals that the total statistical uncertainty in the three-dimensional reconstruction remains consistent with a Gaussian distribution to high significance. All publicly available data products of the three-dimensional absorption field reconstruction (Section 5.4) are accompanied by the pointwise standard error estimates computed through this bootstrap procedure.

5.3 Results

In this section we display various visualizations of our 47 h^{-1} Gpc^3 reconstruction of the IGM and document the number of candidates we have detected for galaxy protoclusters and cosmic voids. We define a candidate for a galaxy protocluster to be a statistically significant (at an n-σ level) contiguous overdensity in the reconstructed absorption field. Analogously, we define a candidate for a cosmic void to be a statistically significant contiguous underdensity in the reconstructed absorption field. Here, contiguity is defined through a three-dimensional friends-of-friends algorithm over the volume of the map on a $(1\ h^{-1}\ \text{Mpc})^3$/voxel Cartesian grid in which a voxel is compared with its 26 neighboring voxels that share a face, edge, or corner. Please refer to our upcoming paper for more details and a deeper discussion of the results.

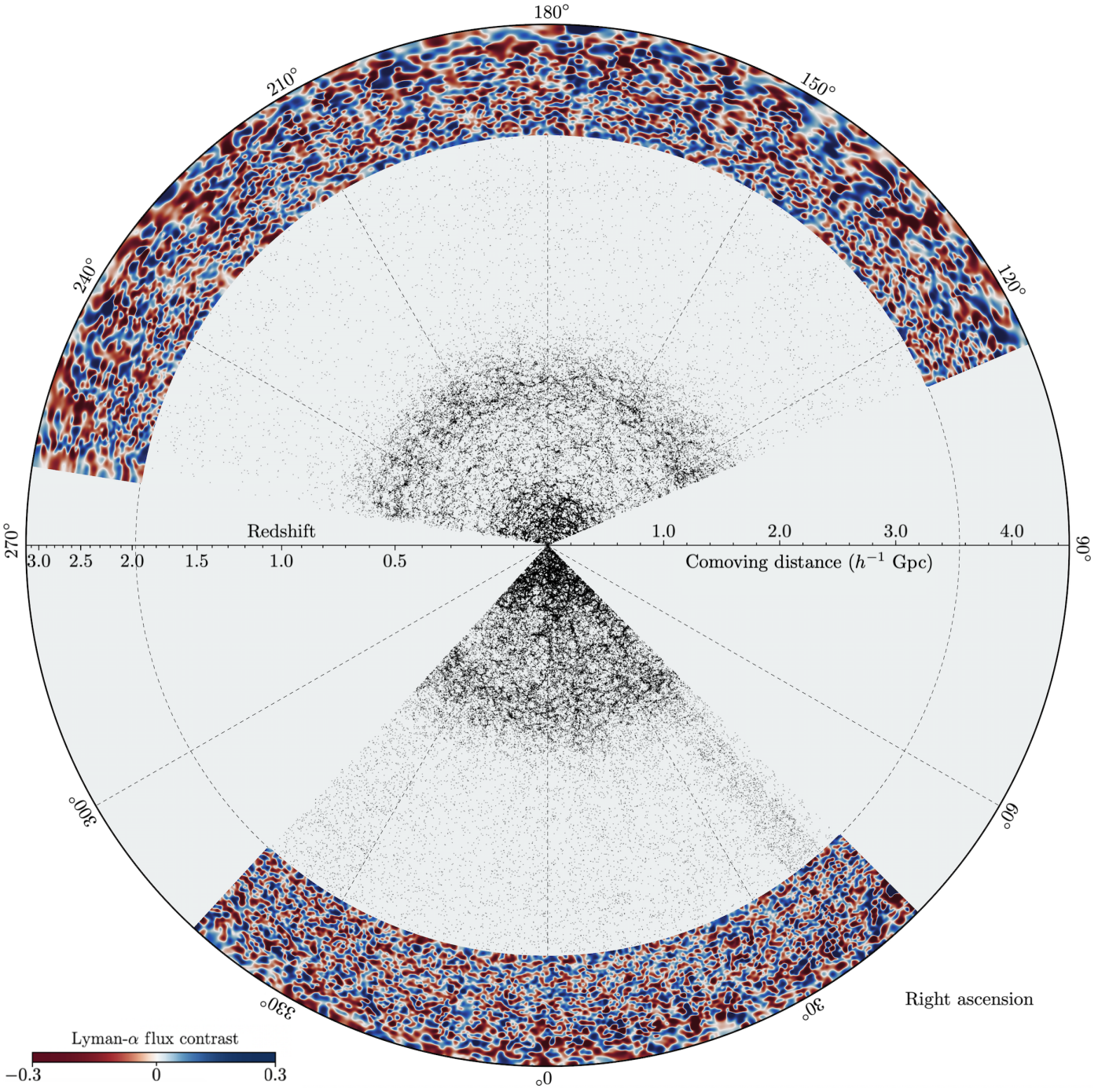


FIGURE 5.6: Earth-centric map of the observed large-scale matter distribution of the Universe out to 4481 h^{-1} comoving Mpc. The 47 h^{-3} Gpc^3 Lyα forest absorption field reconstructed in this work spans the redshift range $1.98 \leq z \leq 3.15$ (3560 h^{-1} Mpc $< d_{\parallel} <$ 4481 h^{-1} Mpc) and provides a continuous high redshift complement to the cosmological matter distribution traced by galaxies and non-Lyα quasars at low redshifts (black dots). Pictured here is a 113° field-of-view in the Northern Galactic Cap (top) at a celestial declination of $\delta = 40°$ and a 87.5° field-of-view in the Southern Galactic Cap (bottom) along the celestial equator ($\delta = 0°$). The effective spatial resolution of the reconstructed Lyα absorption field progressively degrades at the high redshift end due to sparser observations of background quasars, but remains well below the scale of baryon acoustic oscillations.

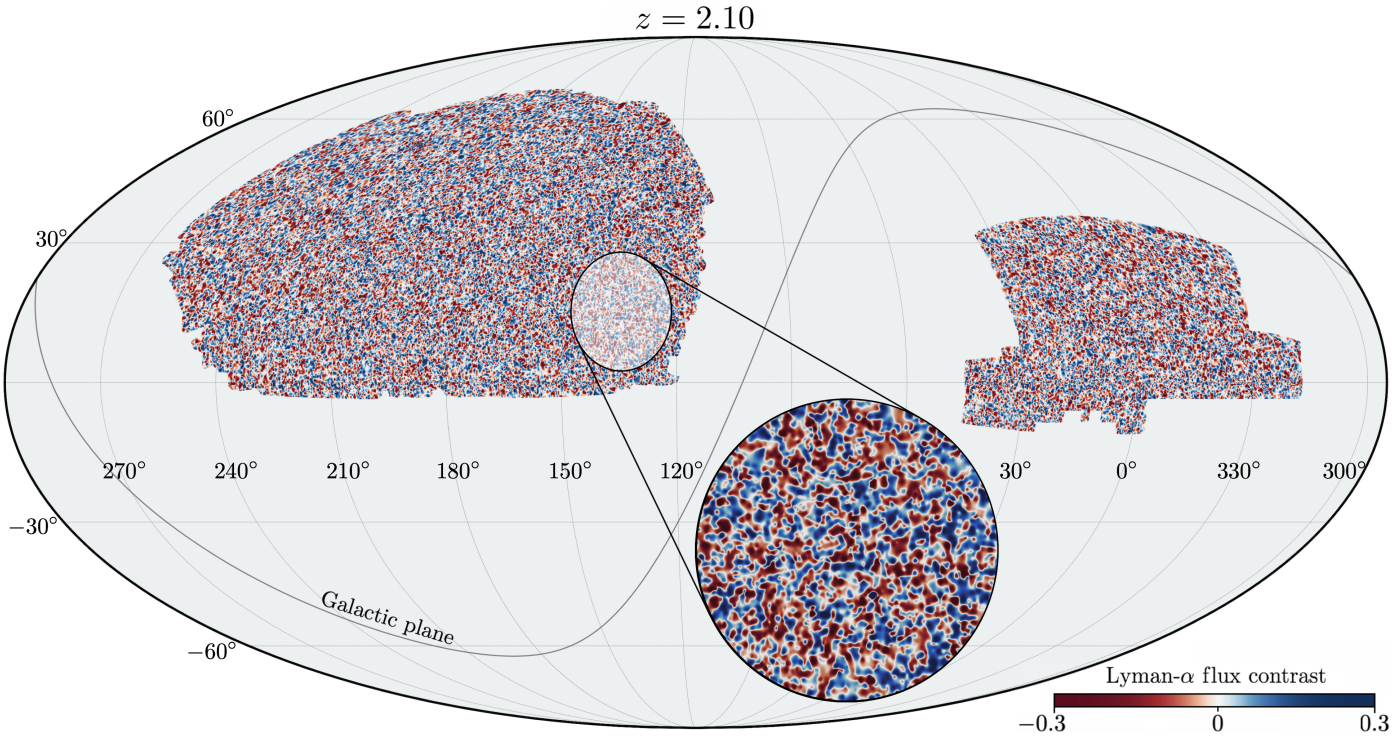


FIGURE 5.7: Cross-sectional sky map of the reconstructed three-dimensional Lyman-α absorption field. Shown here is the large-scale matter density distribution of the Universe at redshift $z = 2.10$ (3677 h^{-1} comoving Mpc from Earth), with equatorial coordinates in a Mollweide projection. At this epoch, the Universe was approximately 3.12 billion years old. The full 47 h^{-3} Gpc^3 comoving volume mapped in this work possesses a footprint of 10,332 deg^2 ($\sim$25% sky coverage) at all redshifts $1.98 \leq z \leq 3.15$, split between two contiguous regions of the sky — a 7,474 deg^2 region in the Northern Galactic Cap (left) and a 2,858 deg^2 region in the Southern Galactic Cap (right). Coherent structure is detected at multiple scales of resolution, with the smallest recoverable structure lower bounded by the sky-marginalized transverse sightline separation $r_\perp(z)$ of the set of background quasars used for the three-dimensional reconstruction (e.g. for $z = 2.10$, $r_\perp \approx 11.1$ h^{-1} Mpc). Two-dimensional sky maps and 1σ Gaussian uncertainties are made publicly available (see Section 5.4) for every 0.01 unit in the redshift interval $1.98 \leq z \leq 3.15$ in a HEALPix pixelization with $N_{\rm side} = 2048$ (1.7 arcmin pixel resolution).

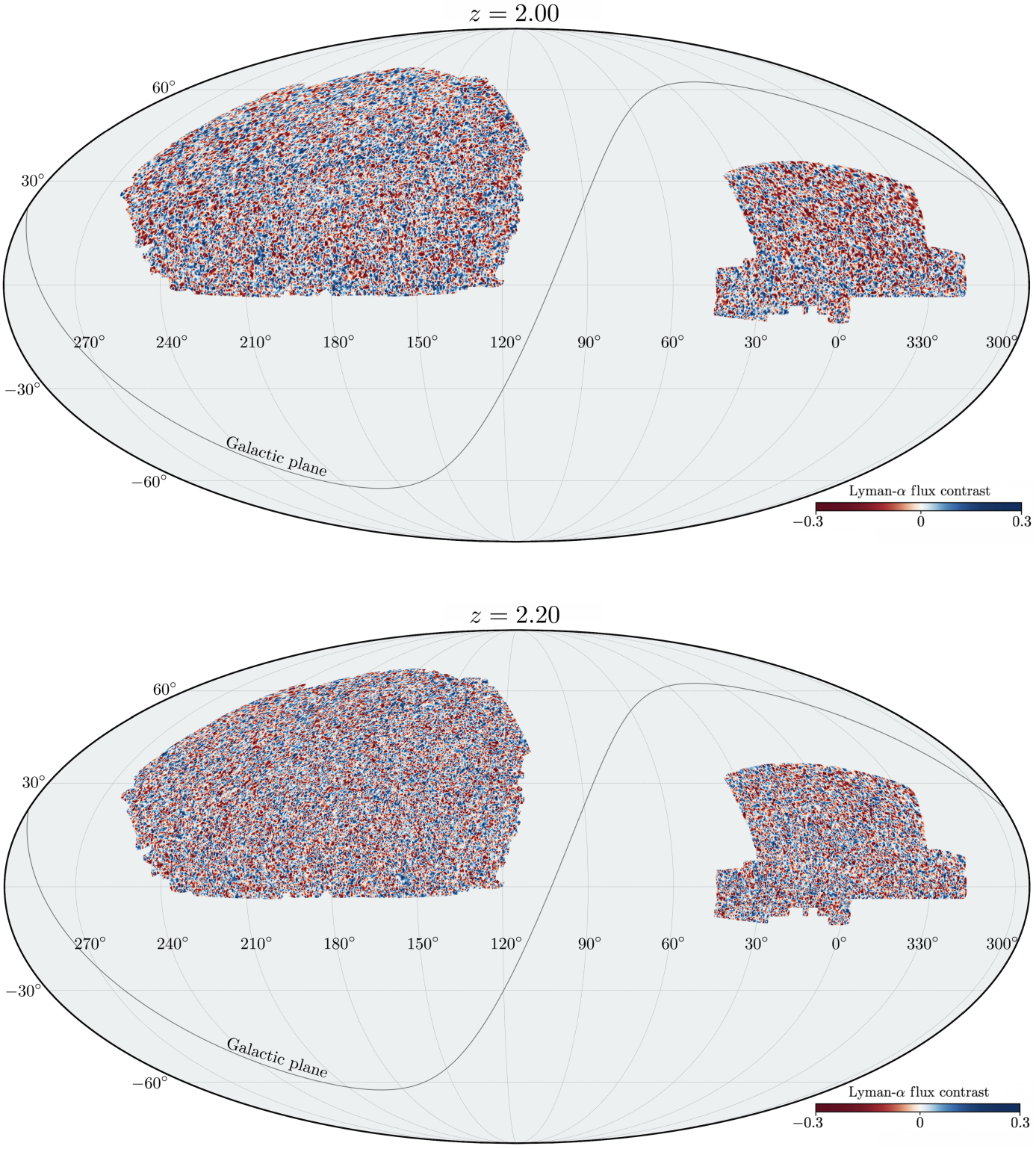


FIGURE 5.8: (Continued:) Cross-sectional sky maps of the reconstructed three-dimensional Lyman-α absorption field. **Top:** Redshift $z = 2.00$ (3580 h^{-1} comoving Mpc from Earth; Universe at age $t = 3.27$ billion years). **Bottom:** Redshift $z = 2.20$ (3769 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.98$ billion years).

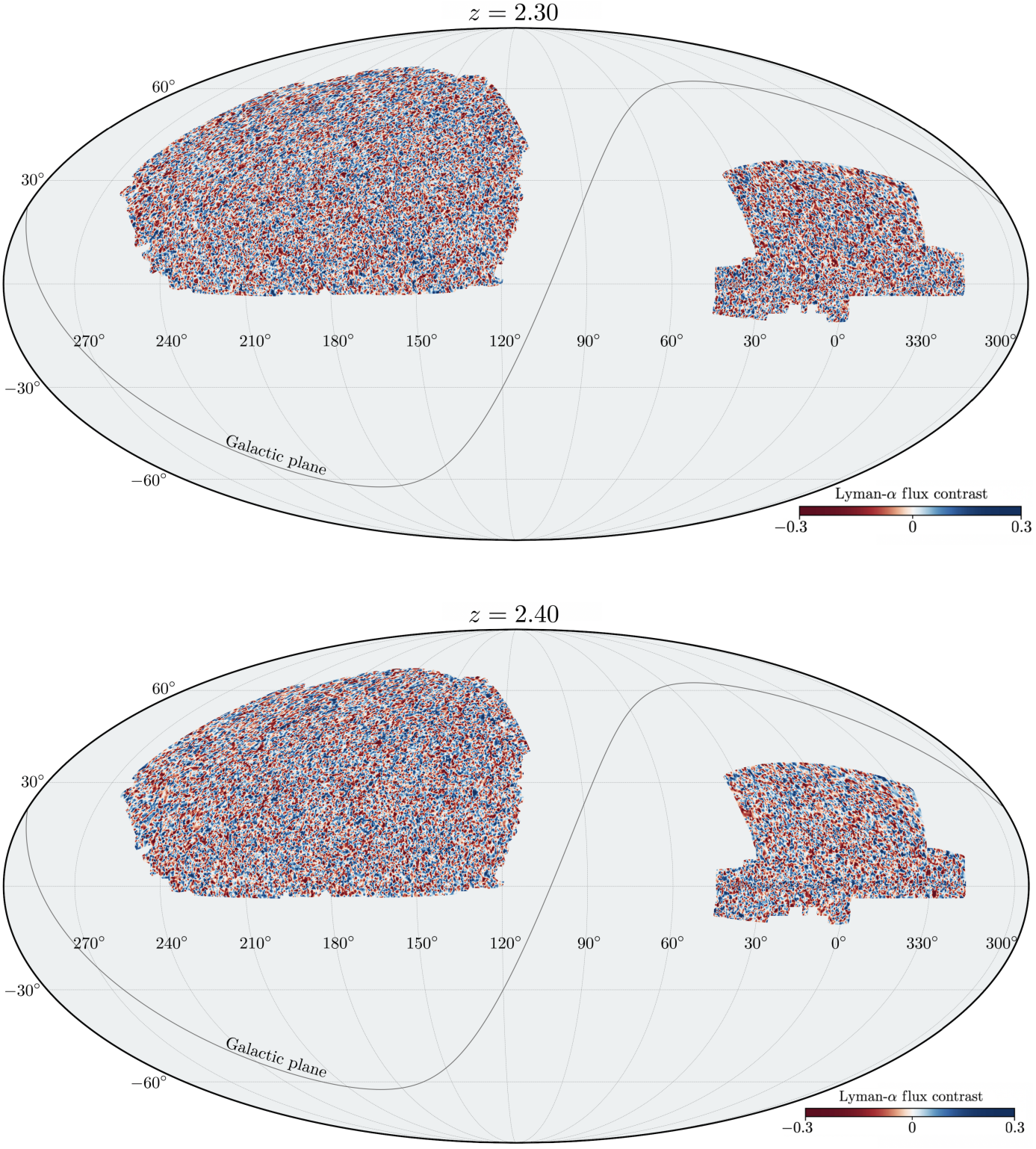


FIGURE 5.9: (Continued:) Cross-sectional sky maps of the reconstructed three-dimensional Lyman-α absorption field. **Top:** Redshift $z = 2.30$ (3858 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.85$ billion years). **Bottom:** Redshift $z = 2.40$ (3943 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.72$ billion years).

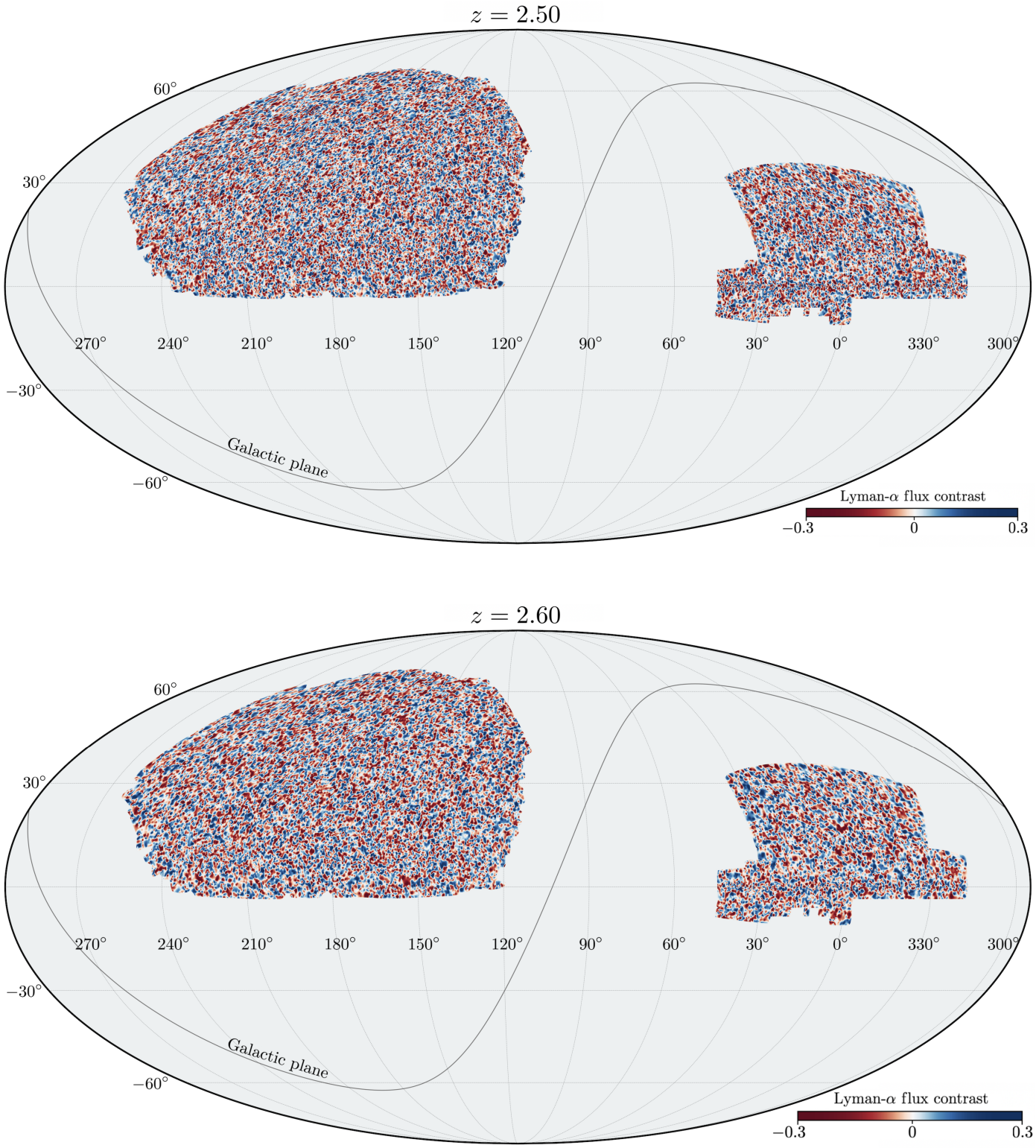


FIGURE 5.10: (Continued:) Cross-sectional sky maps of the reconstructed three-dimensional Lyman-α absorption field. **Top:** Redshift $z = 2.50$ (4024 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.61$ billion years). **Bottom:** Redshift $z = 2.60$ (4102 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.50$ billion years).

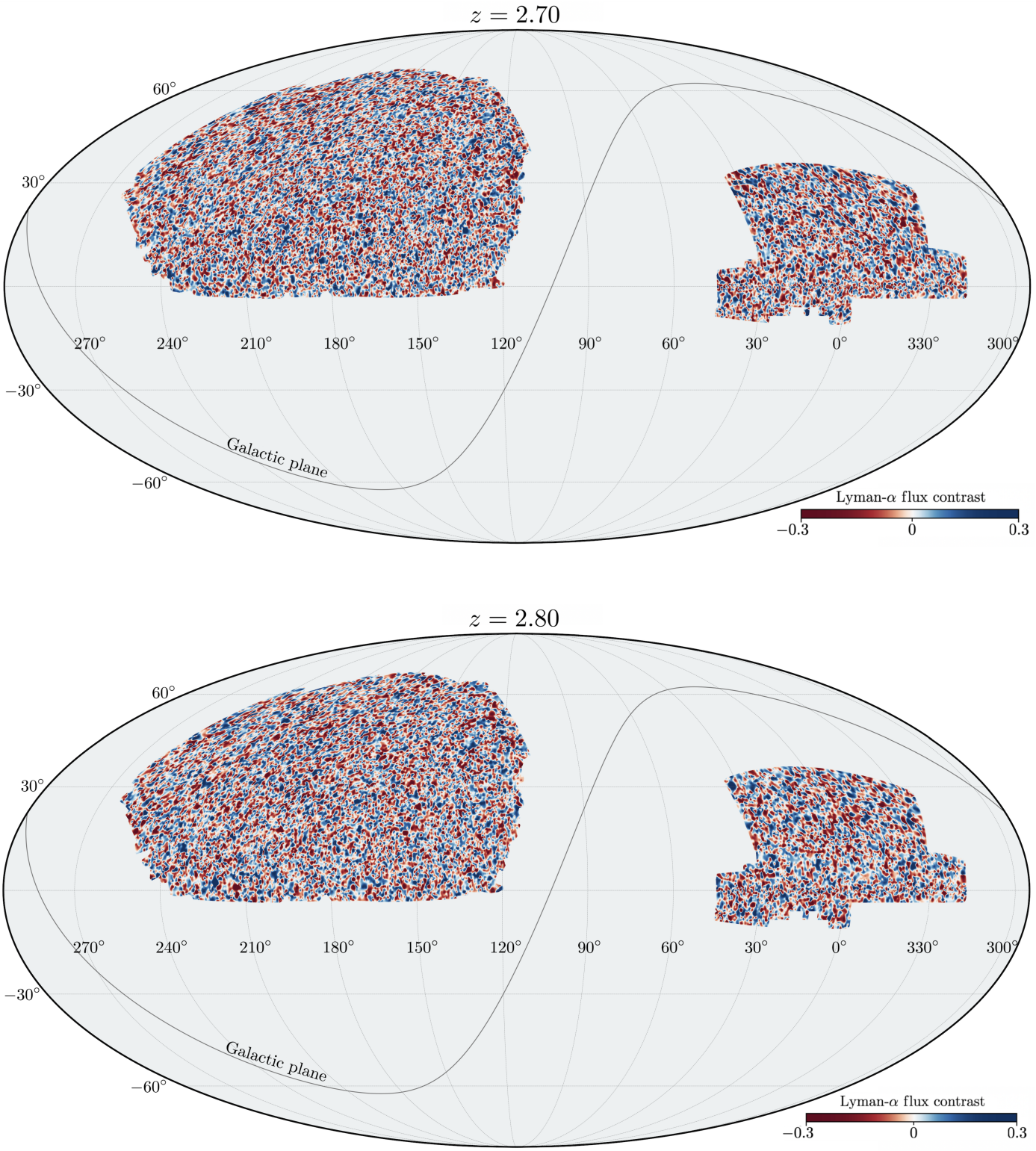


FIGURE 5.11: (Continued:) Cross-sectional sky maps of the reconstructed three-dimensional Lyman-α absorption field. **Top:** Redshift $z = 2.70$ (4177 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.40$ billion years). **Bottom:** Redshift $z = 2.80$ (4249 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.31$ billion years).

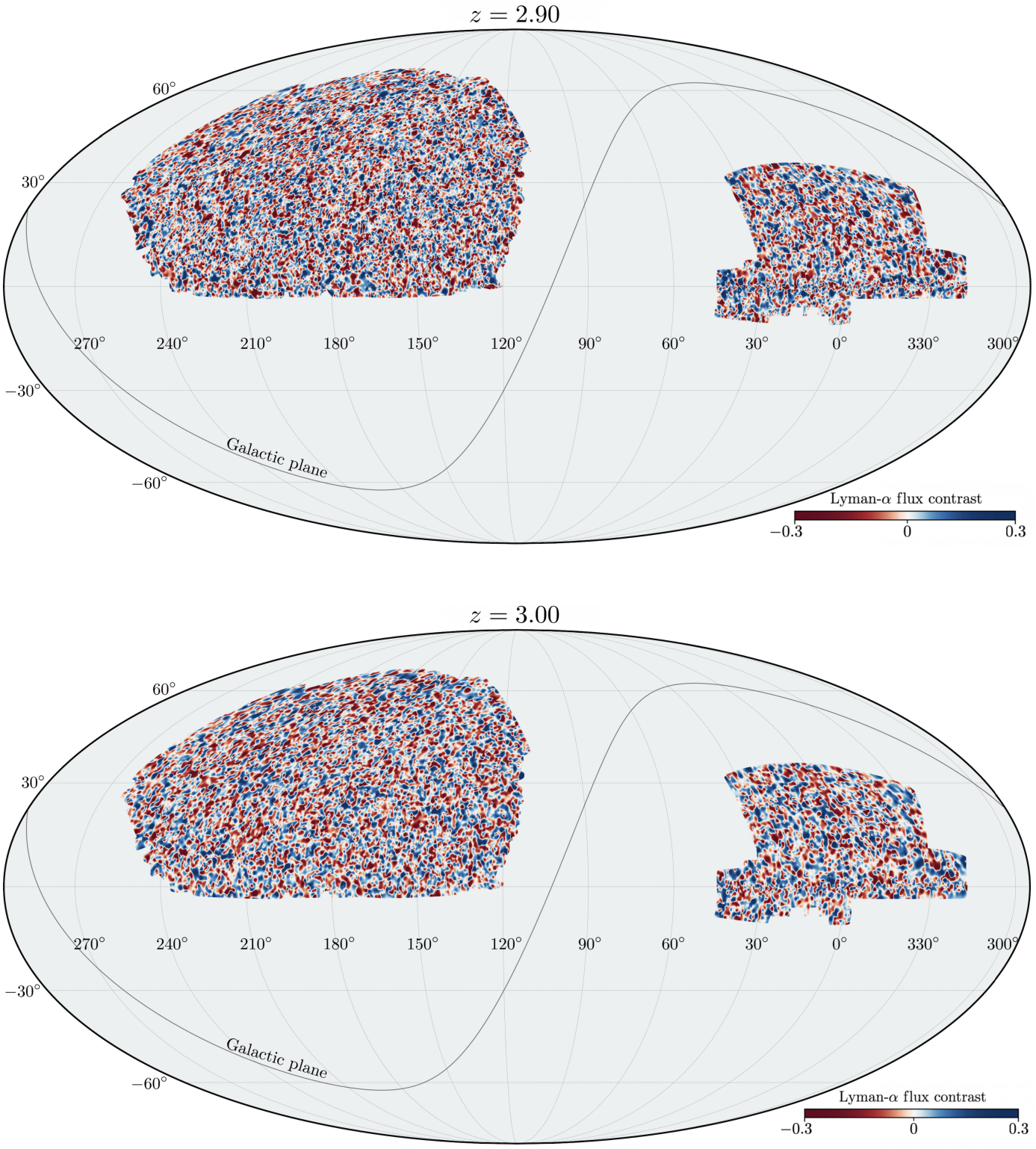


FIGURE 5.12: (Continued:) Cross-sectional sky maps of the reconstructed three-dimensional Lyman-α absorption field. **Top:** Redshift $z = 2.90$ (4318 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.22$ billion years). **Bottom:** Redshift $z = 3.00$ (4385 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.14$ billion years).

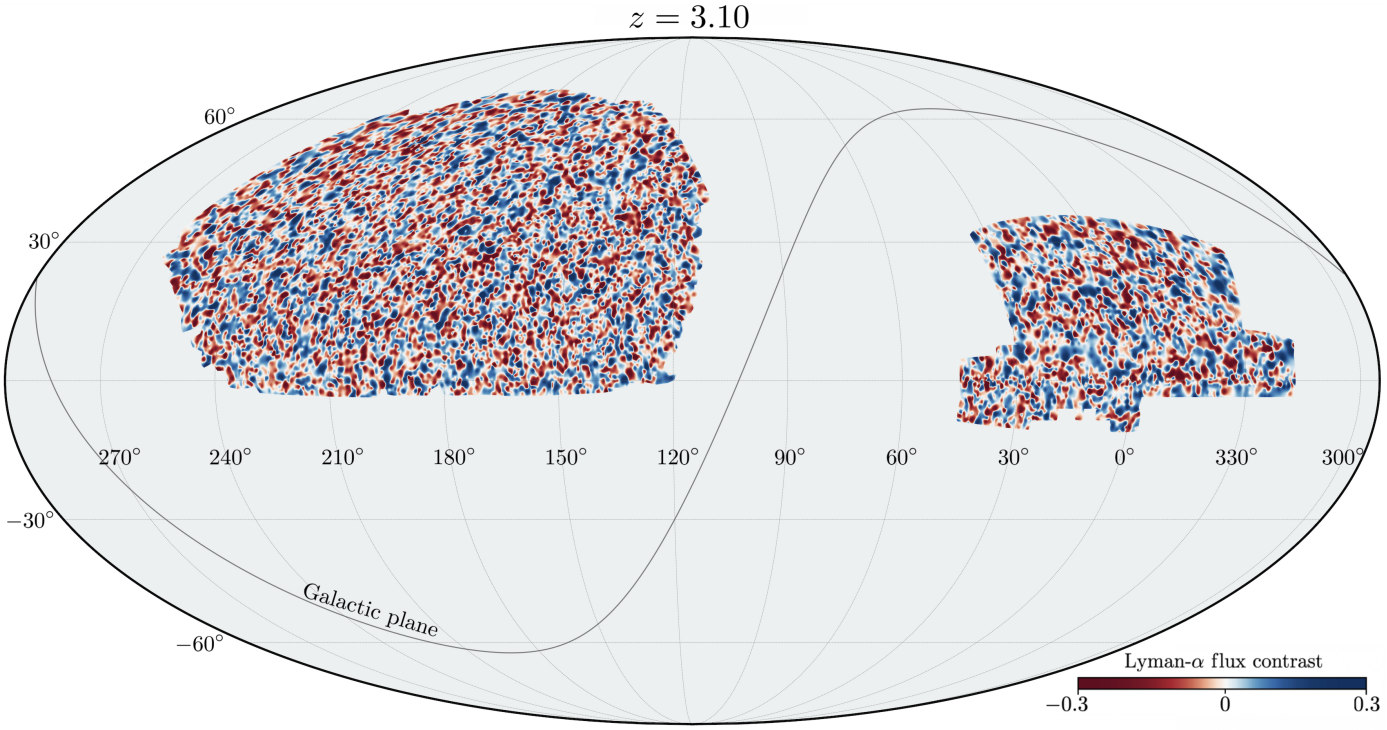


FIGURE 5.13: (Continued:) Cross-sectional sky maps of the reconstructed three-dimensional Lyman-α absorption field. Redshift $z = 3.10$ (4450 h^{-1} comoving Mpc from Earth; Universe at age $t = 2.06$ billion years).

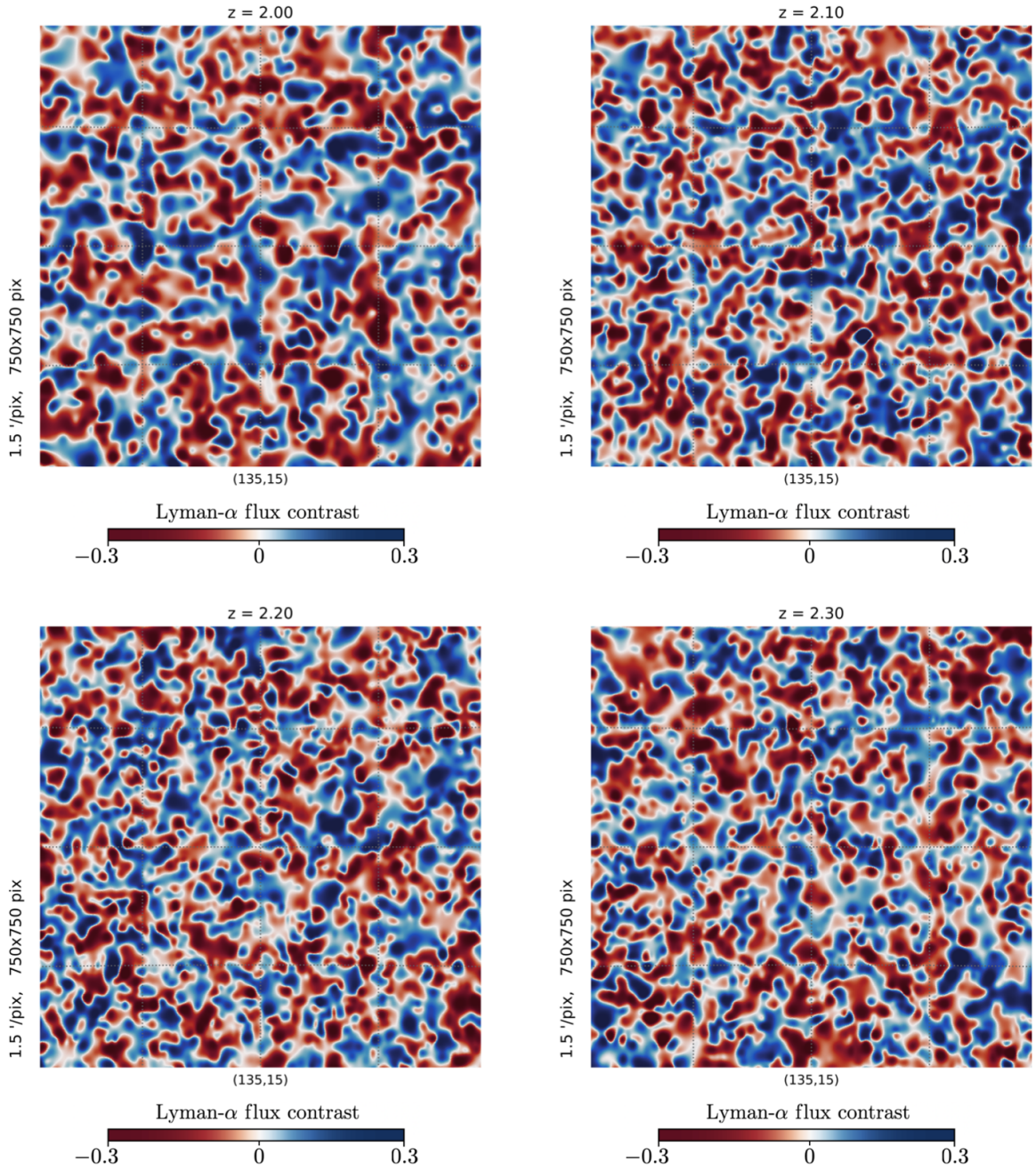


FIGURE 5.14: Gnomonic projections of a $(18.75 \text{ deg})^2$ slice of the reconstructed Lyα absorption field at various redshifts. The square slice is centered at equatorial coordinates $(\alpha, \delta) = (135°, 15°)$ (in the Northern Galactic Cap). Meridians and parallels are overlaid at 5° intervals.

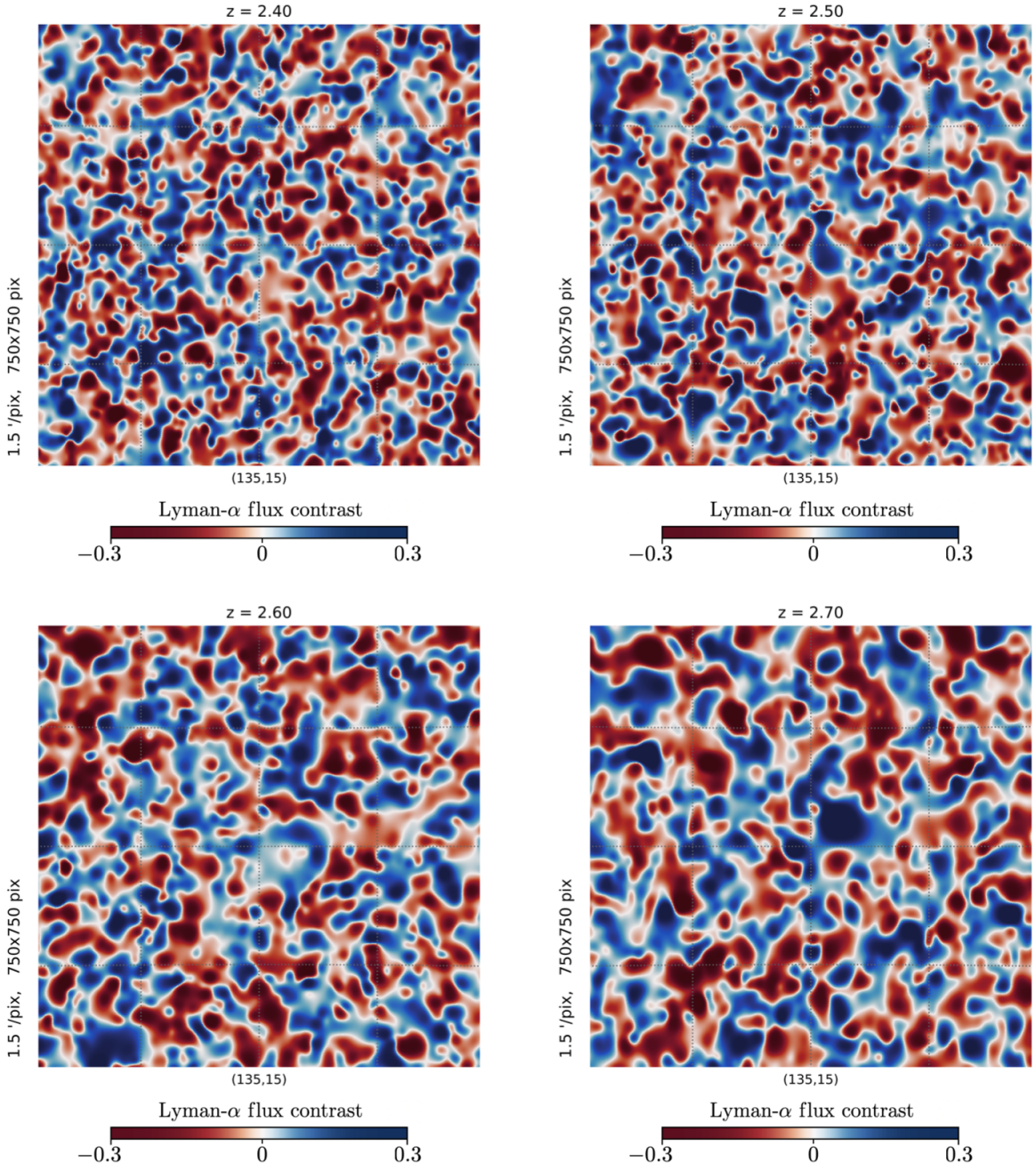


FIGURE 5.15: (Continued:) Gnomonic projections of a $(18.75\ \text{deg})^2$ slice of the reconstructed Lyα absorption field at various redshifts. The square slice is centered at equatorial coordinates $(\alpha, \delta) = (135°, 15°)$ (in the Northern Galactic Cap). Meridians and parallels are overlaid at 5° intervals.

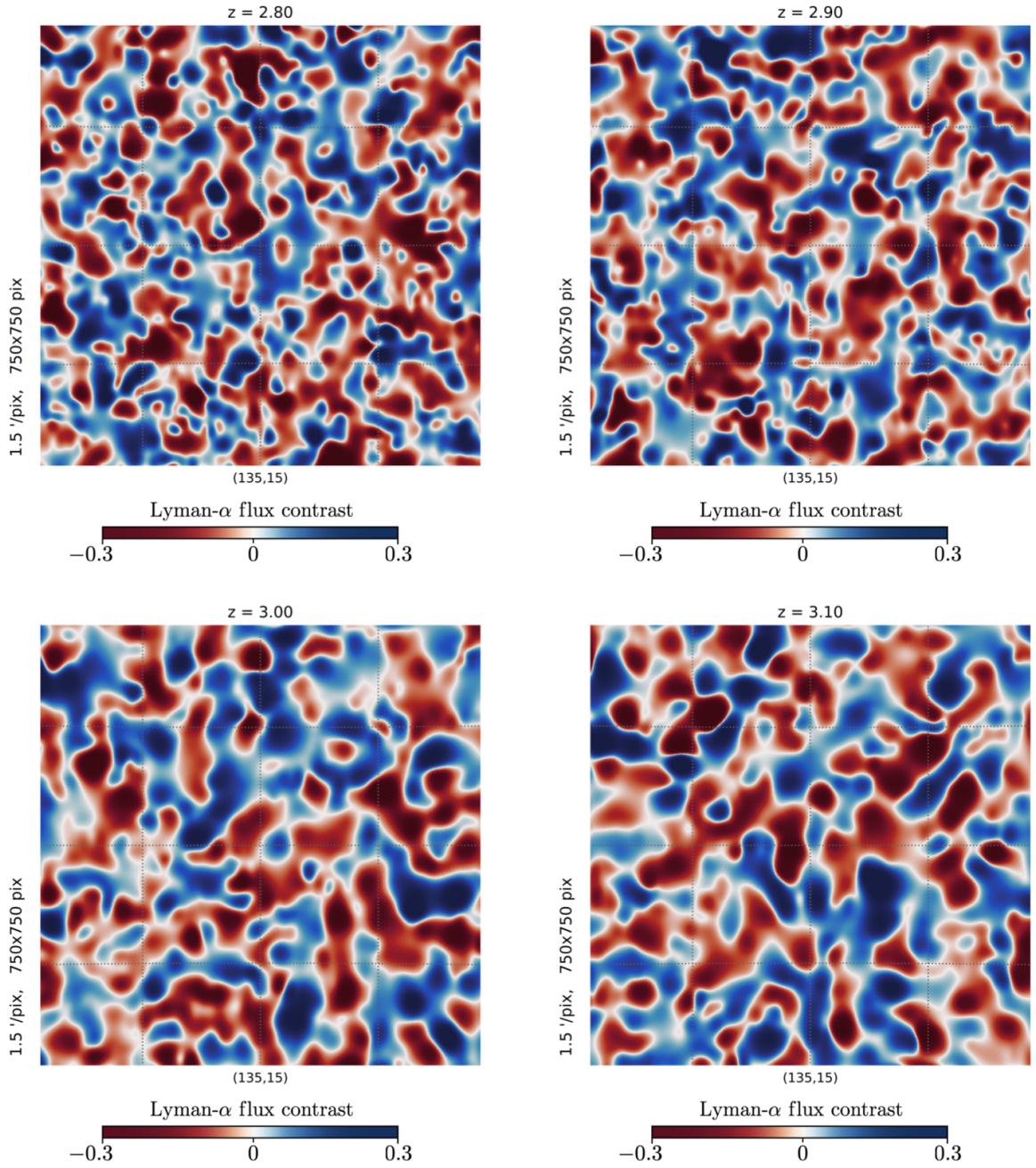


FIGURE 5.16: (Continued:) Gnomonic projections of a $(18.75\ \mathrm{deg})^2$ slice of the reconstructed Lyα absorption field at various redshifts. The square slice is centered at equatorial coordinates $(\alpha, \delta) = (135°, 15°)$ (in the Northern Galactic Cap). Meridians and parallels are overlaid at 5° intervals.

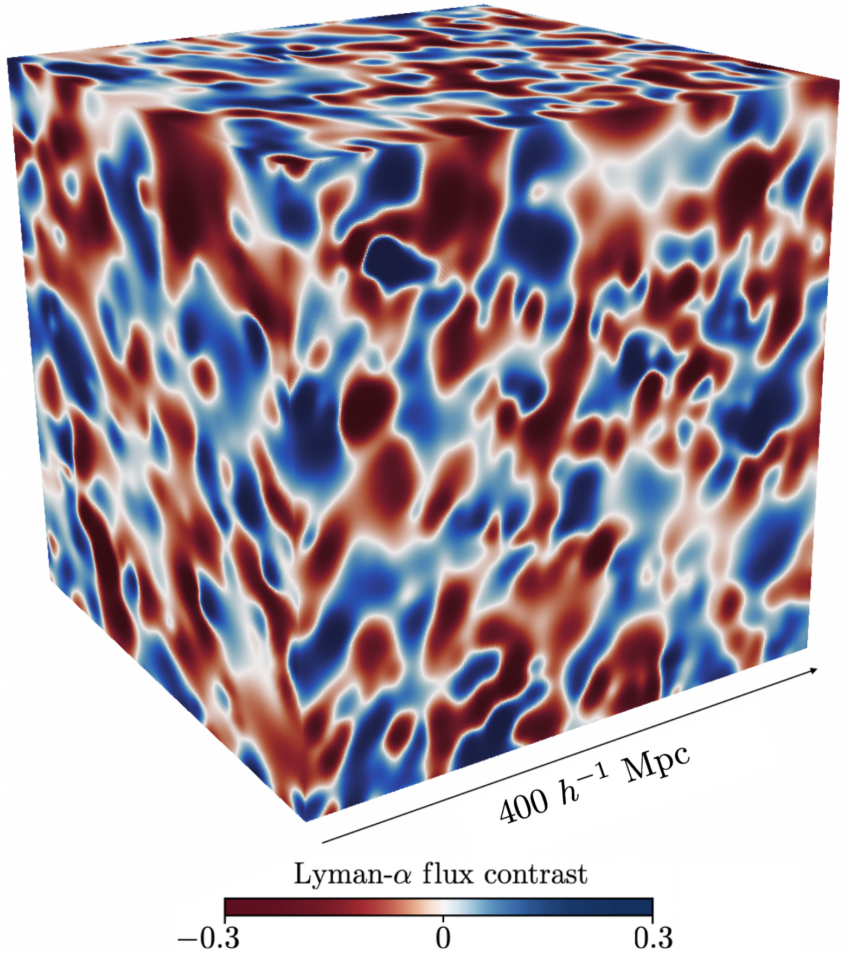


FIGURE 5.17: Three-dimensional reconstruction of the intergalactic medium matter density distribution, as traced by Lyα forest absorption in the spectra of background quasars. The (400 h^{-1} Mpc)^3 comoving volume pictured here, which is centered at redshift $z \sim 2.3$, constitutes $\sim$0.1% of the total cosmological volume mapped in this work. On the $\gtrsim 10$ h^{-1} Mpc scales mapped in this work, the dimensionless Lyα flux contrast (coloured by the sample quantiles) directly traces the distribution of gaseous H I in the intergalactic medium and, by extension, the total distribution of gravitating cosmological matter, with positive flux contrasts corresponding to underdensities and negative flux contrasts corresponding to overdensities.

	Significance level					
	3σ	4σ	5σ	6σ	7σ	8σ
Galaxy protoclusters	115,156	86,220	54,300	29,868	17,608	10,896
Cosmic voids	108,676	79,436	53,776	32,504	20,840	12,984

TABLE 5.1: Cosmic census of candidates for high redshift galaxy protoclusters and cosmic voids in the absorption field reconstruction, detected at increasing levels of statistical significance. We define galaxy protocluster and cosmic void candidates to be statistically significant contiguous overdensities and underdensities, respectively, as determined by a three-dimensional friends-of-friends algorithm over the volume of the map on a (1 h^{-1} Mpc)^3/voxel Cartesian grid. Shown here are estimates for the total numbers expected across the full 47 h^{-3} Gpc^3 volume, which we produced by taking the total number of candidates detected in the Southern Galactic Cap ($\sim$25% of the total volume) and multiplying by four. We are still waiting for the friends-of-friends clustering algorithm to complete in the Northern Galactic Cap.

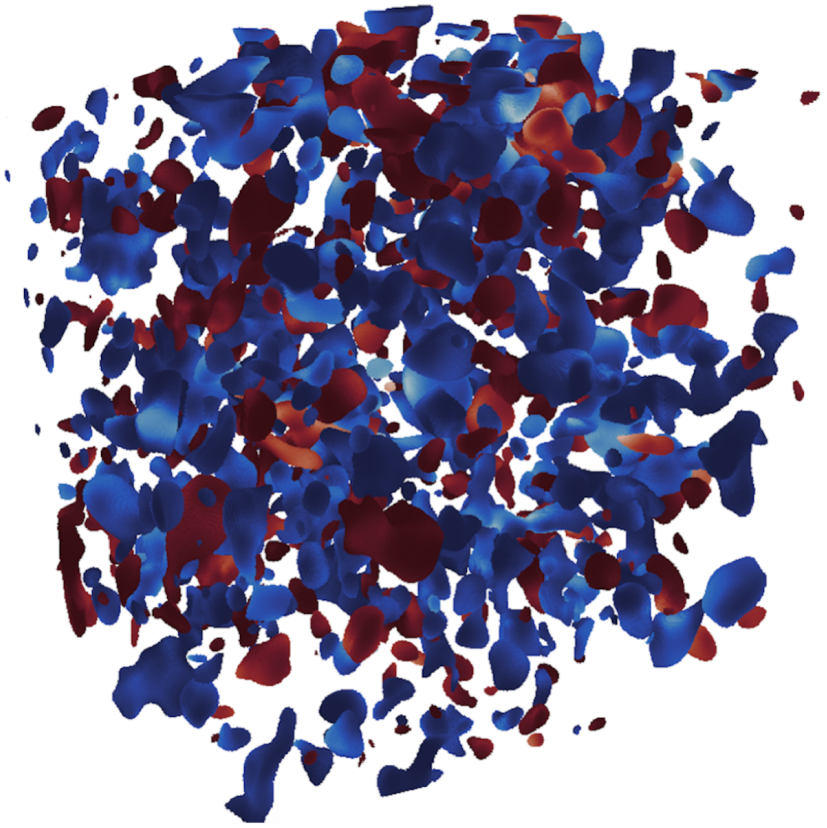


FIGURE 5.18: Statistically significant overdensities (red) and underdensities (blue) in the (400 h^{-1} Mpc)^3 cubic volume shown in Figure 5.17. Here, the overdensities and underdensities are significant at the 6σ level. We catalog each contiguous overdensity as a candidate for a galaxy protocluster and each contiguous underdensity as a candidate for a cosmic void.

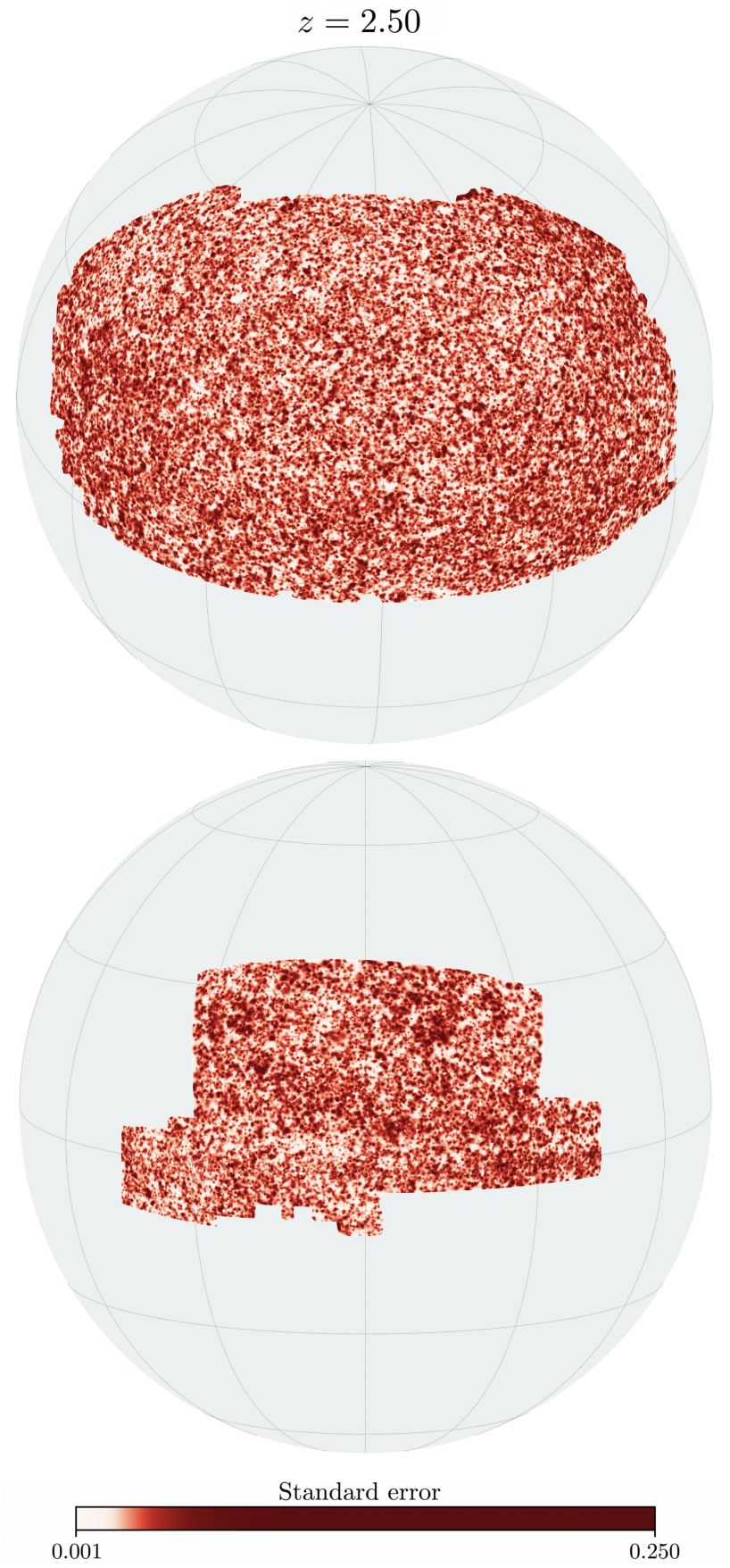


FIGURE 5.19: The estimated standard errors of the reconstructed Lyα absorption field at redshift $z = 2.50$ (in an orthographic projection), with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom. The pointwise standard errors are computed from a sample of 50 bootstrap reconstructions of the full absorption field. We find the bootstrap distribution to be consistent with a Gaussian to high significance.

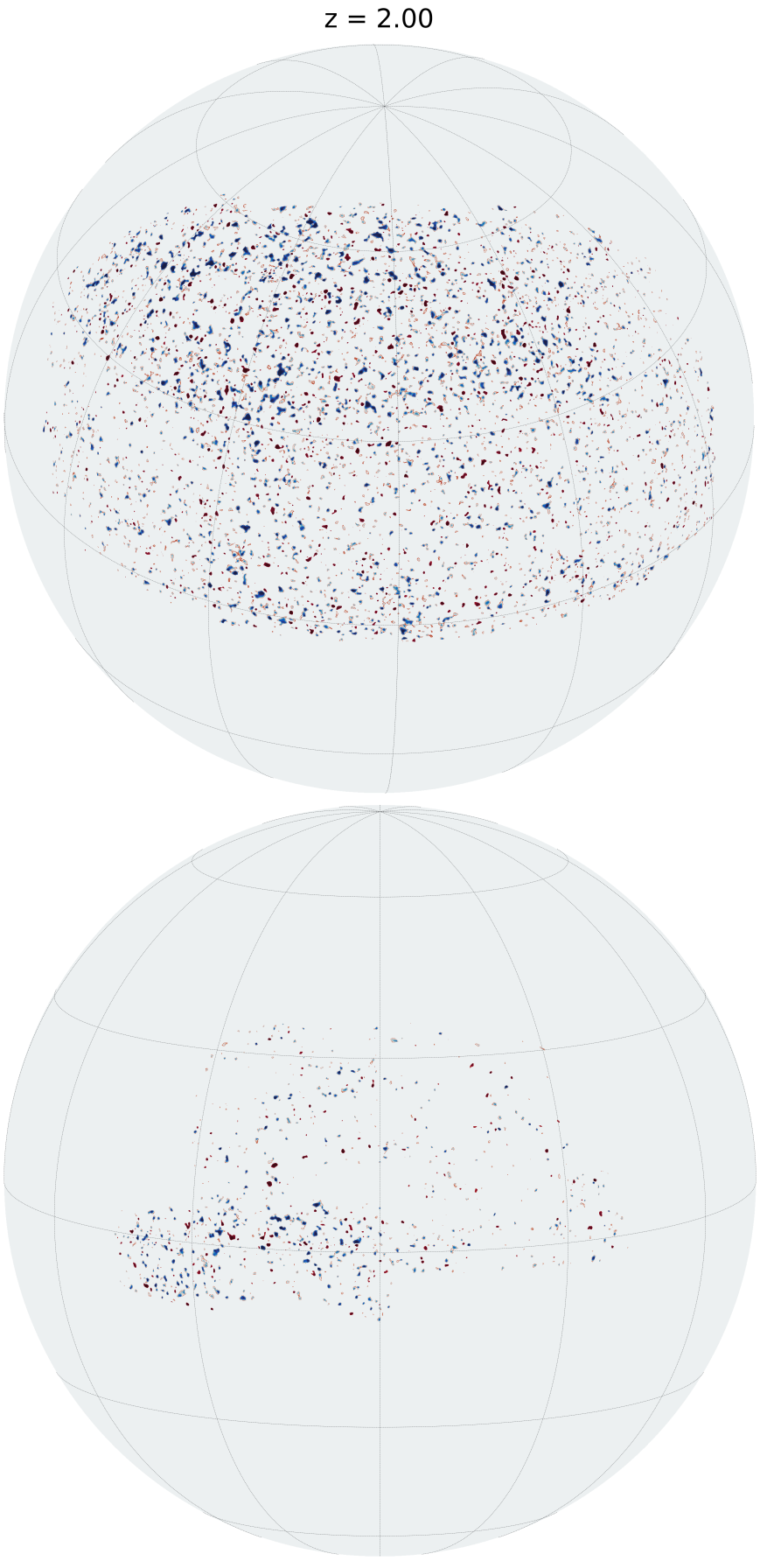


FIGURE 5.20: Orthographic projection of the redshift $z = 2.00$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

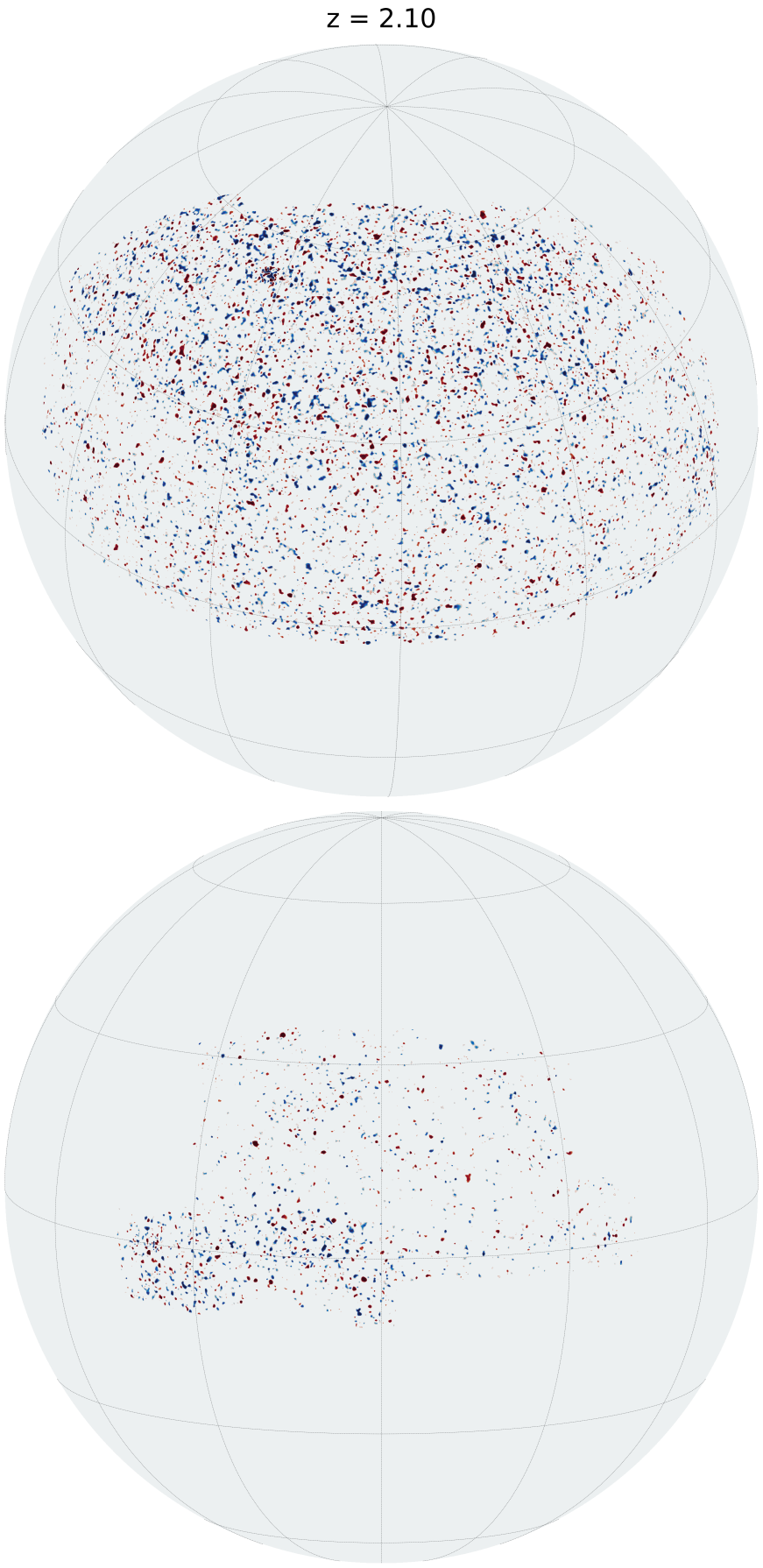


FIGURE 5.21: (Continued:) Orthographic projection of the redshift $z = 2.10$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

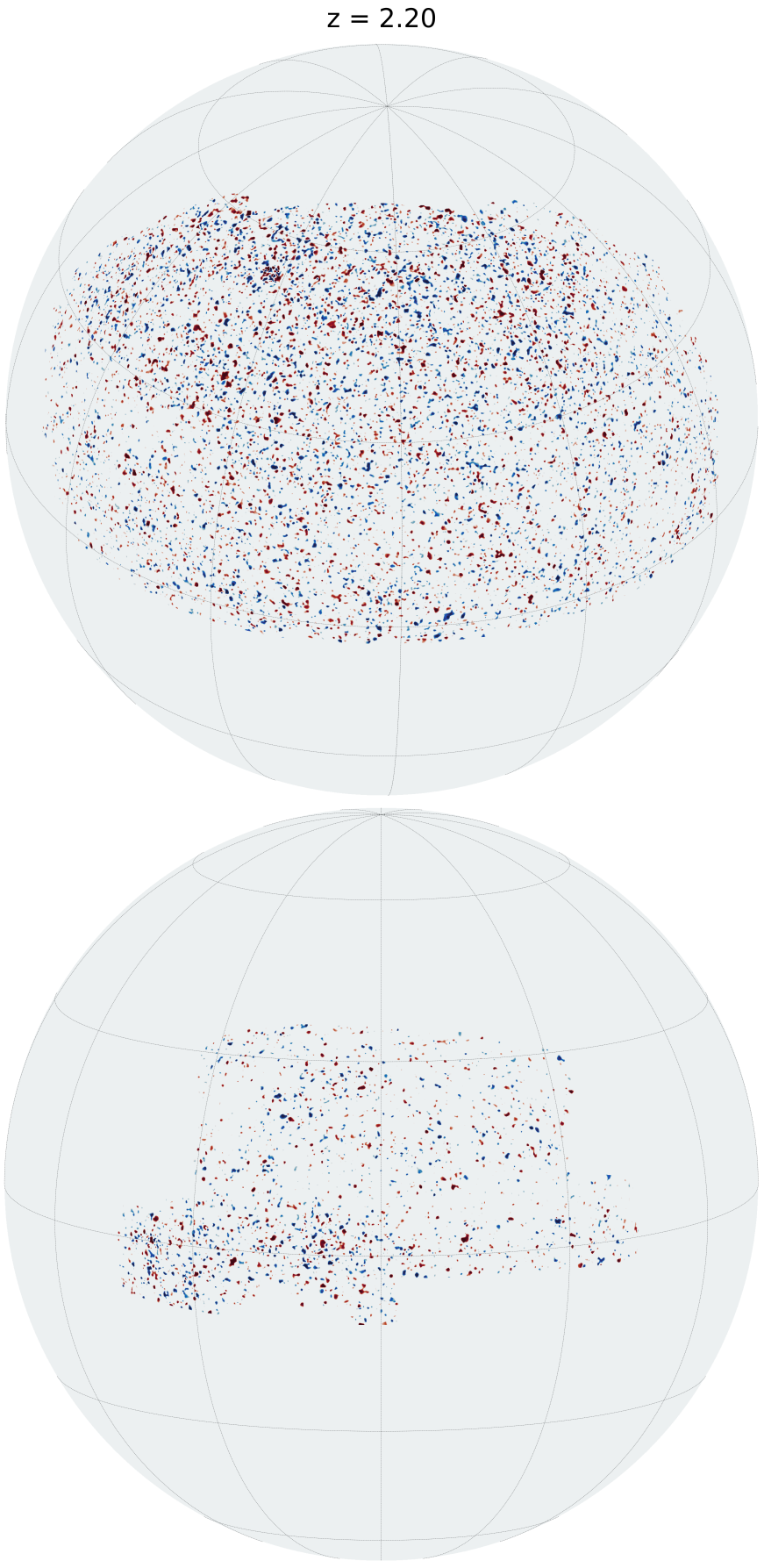


FIGURE 5.22: (Continued:) Orthographic projection of the redshift $z = 2.20$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at the 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

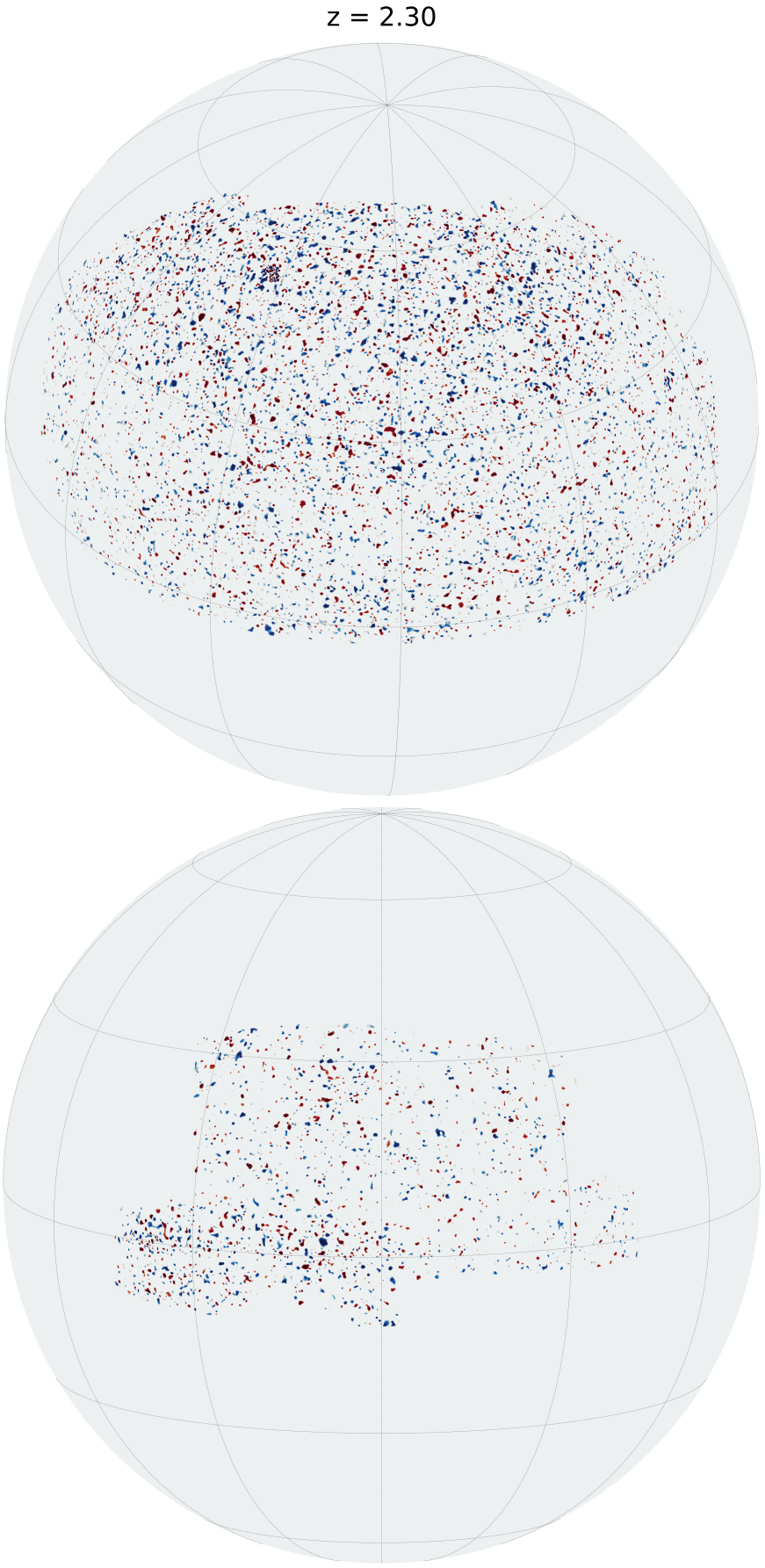


FIGURE 5.23: (Continued:) Orthographic projection of the redshift $z = 2.30$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

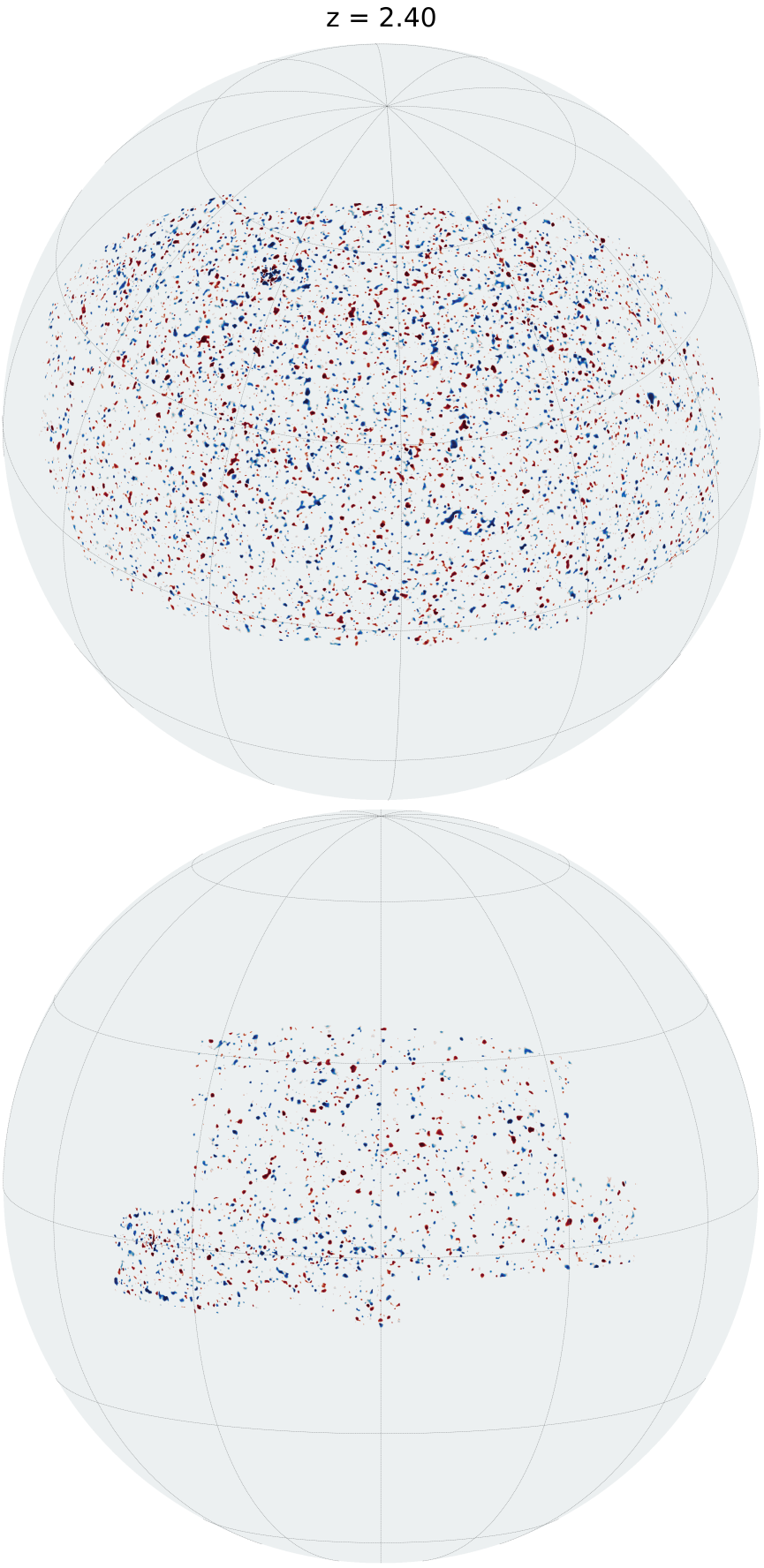


FIGURE 5.24: (Continued:) Orthographic projection of the redshift $z = 2.40$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

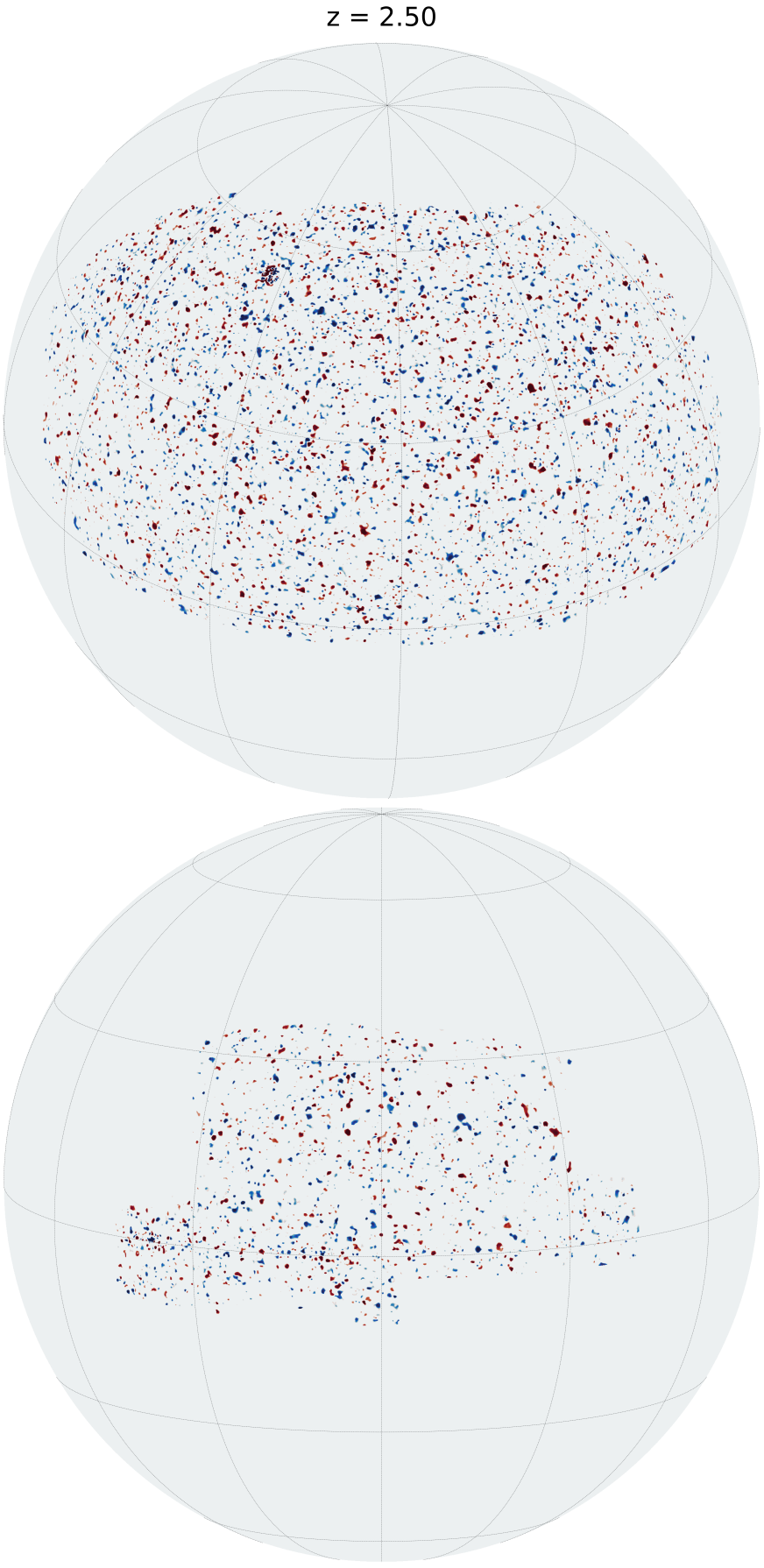


FIGURE 5.25: (Continued:) Orthographic projection of the redshift $z = 2.50$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

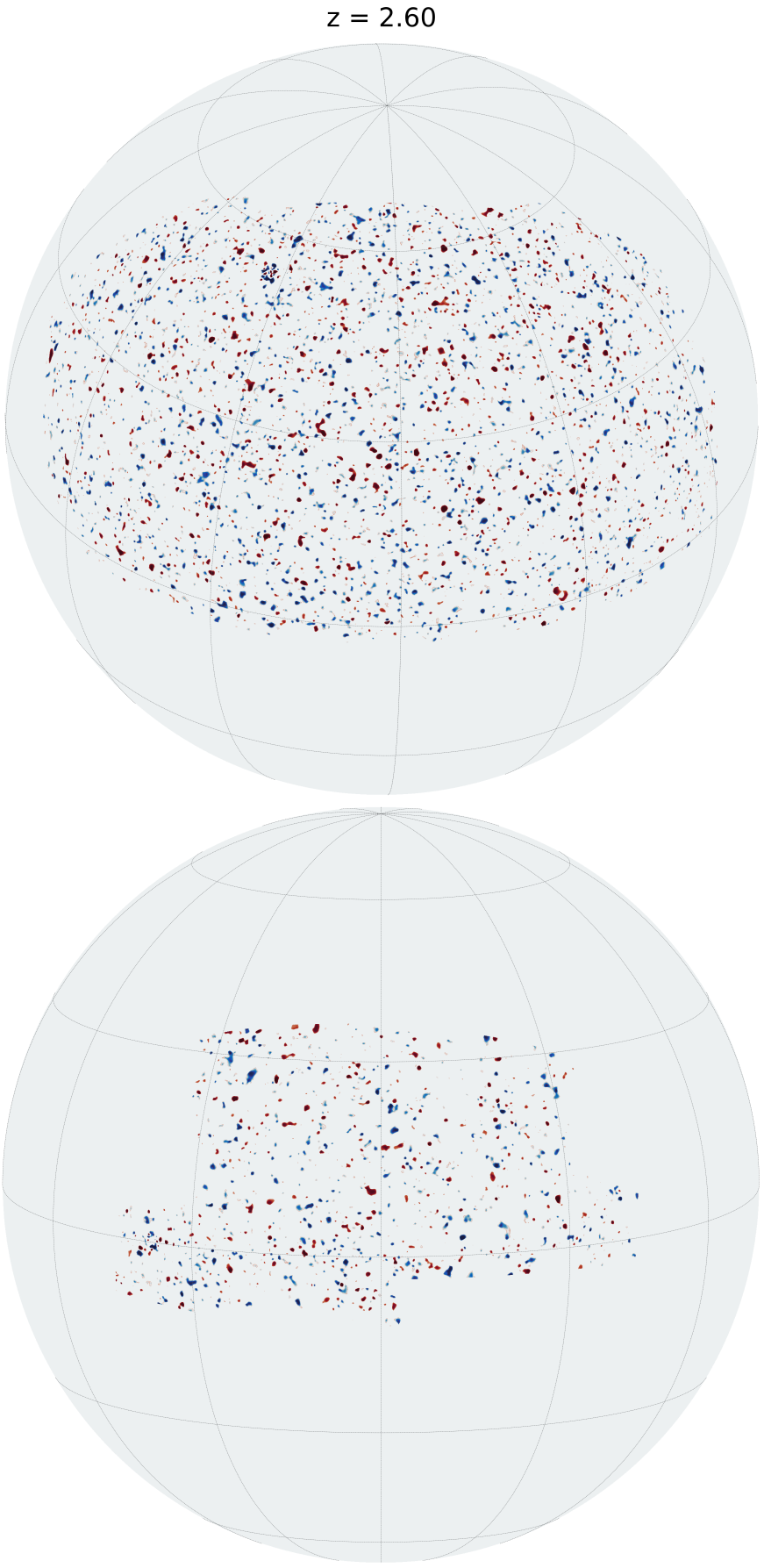


FIGURE 5.26: (Continued:) Orthographic projection of the redshift $z = 2.60$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

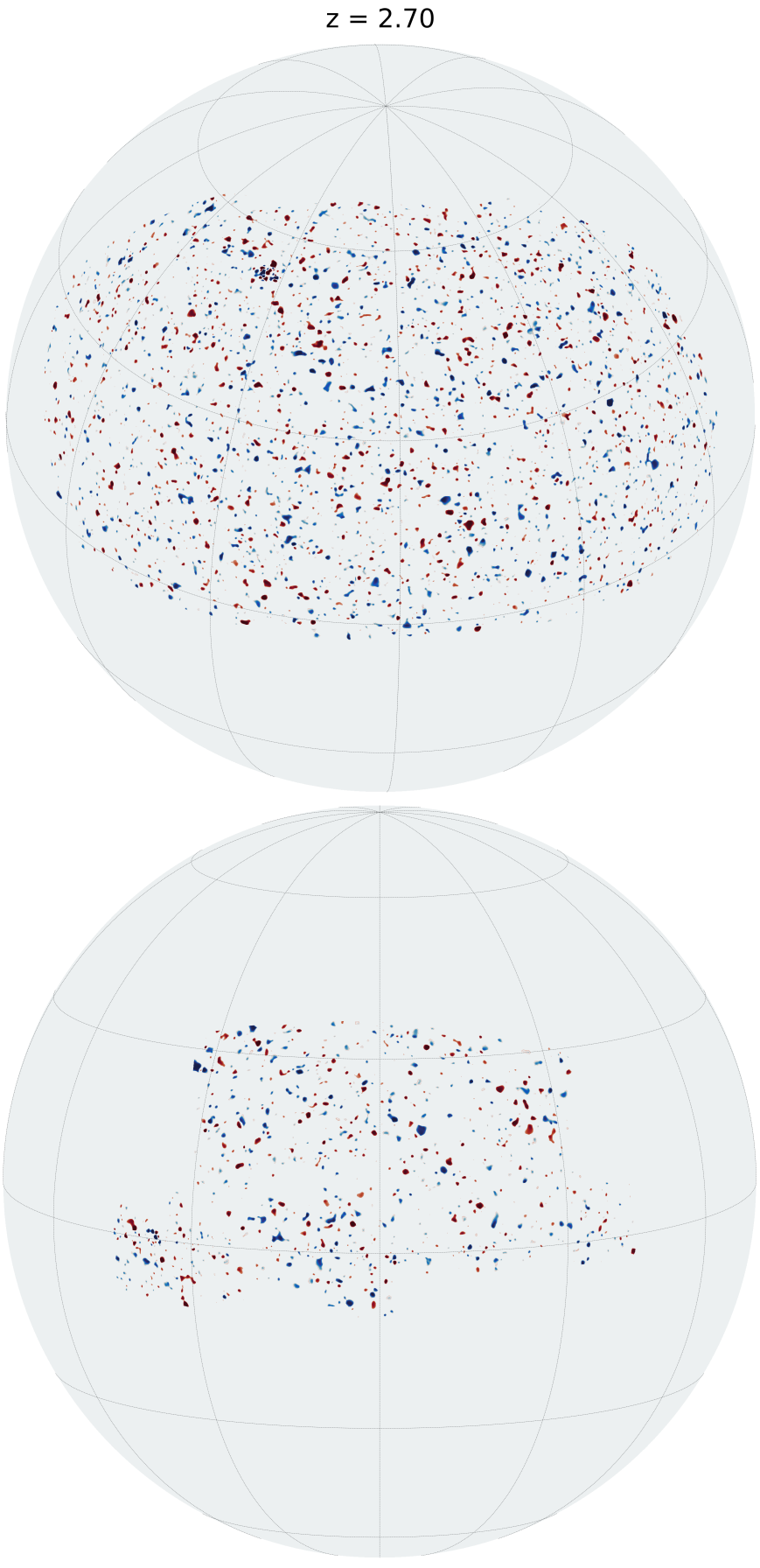


FIGURE 5.27: (Continued:) Orthographic projection of the redshift $z = 2.70$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

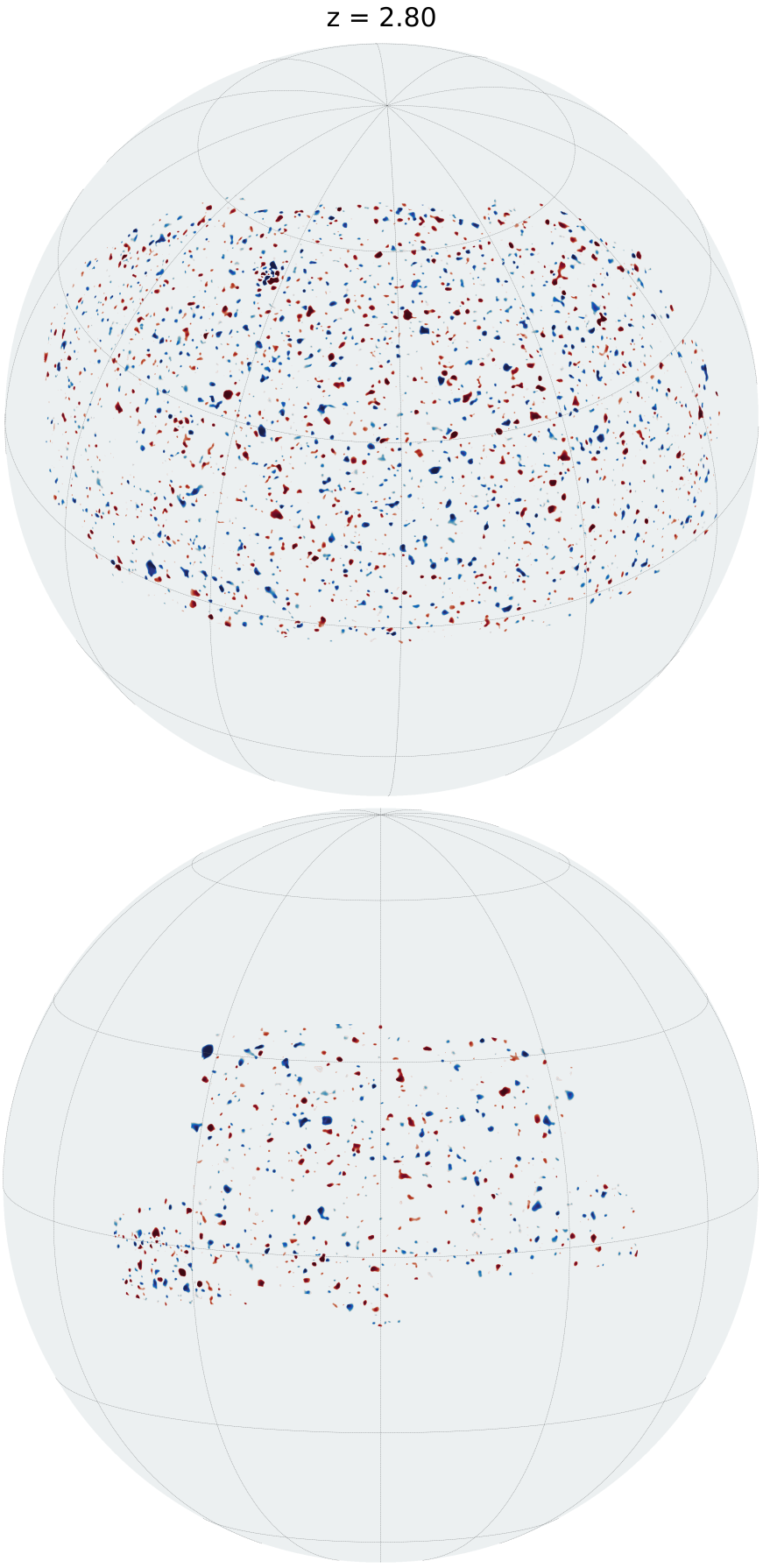


FIGURE 5.28: (Continued:) Orthographic projection of the redshift $z = 2.80$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

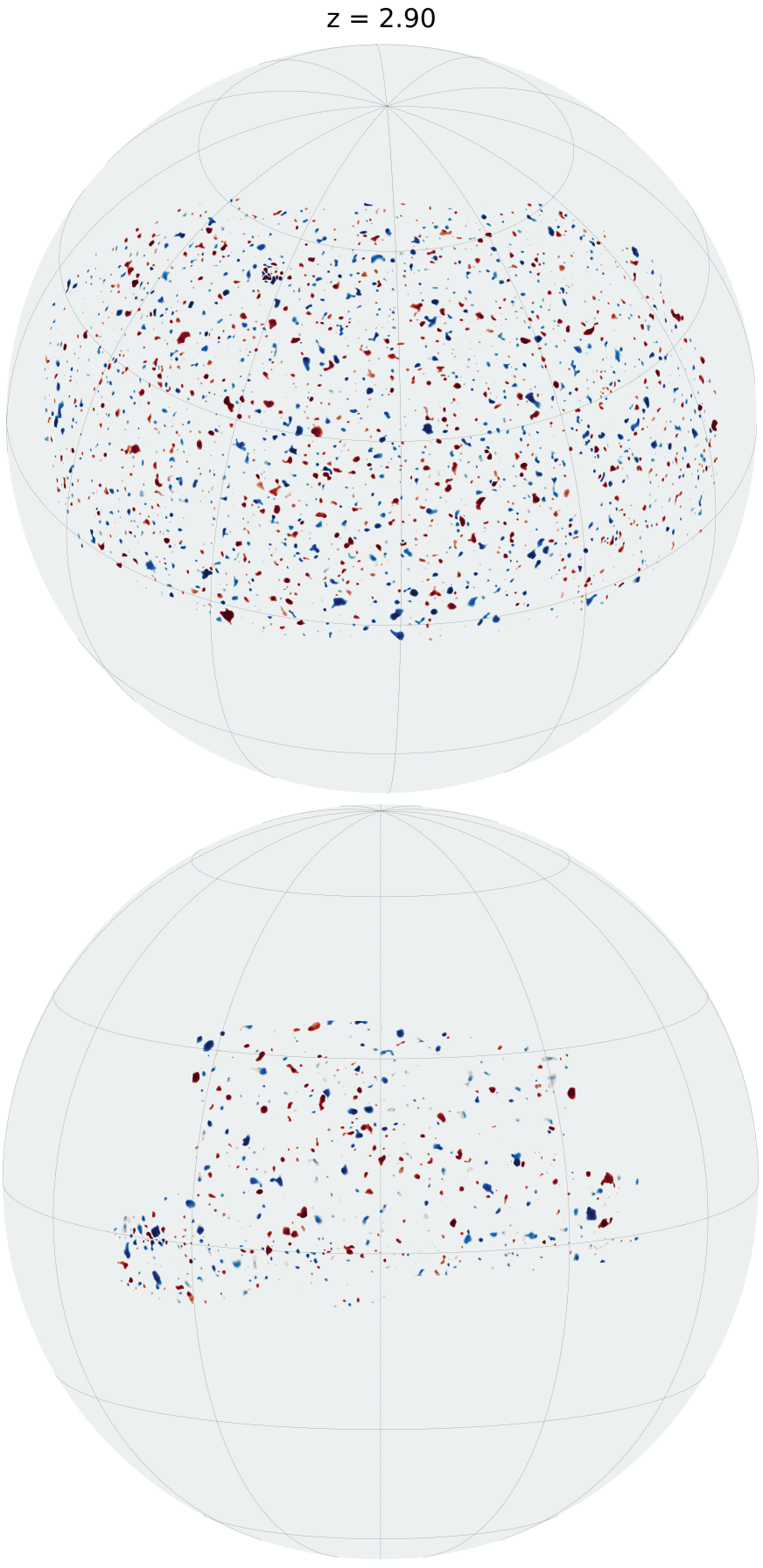


FIGURE 5.29: (Continued:) Orthographic projection of the redshift $z = 2.90$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

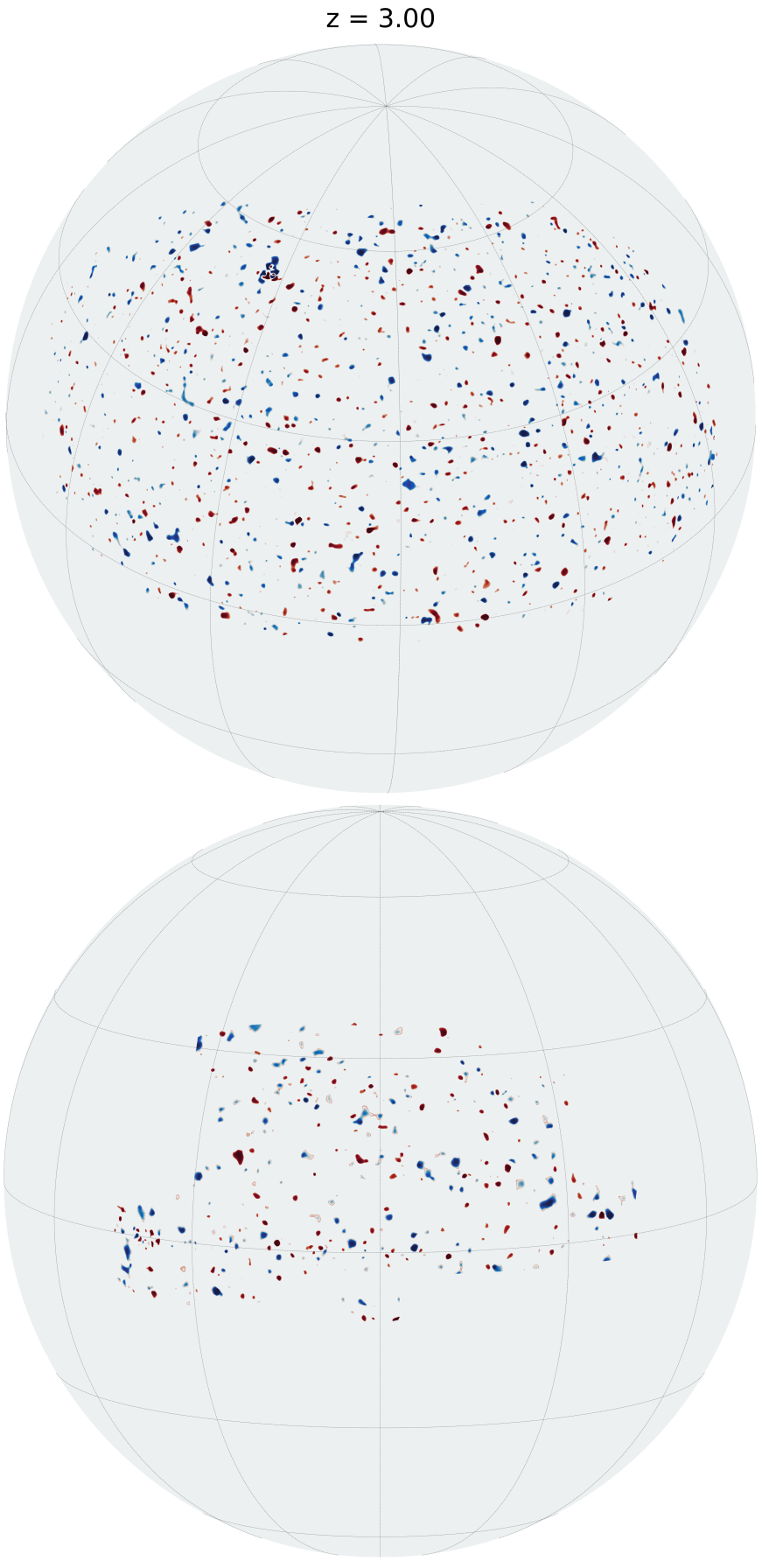


FIGURE 5.30: (Continued:) Orthographic projection of the redshift $z = 3.00$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

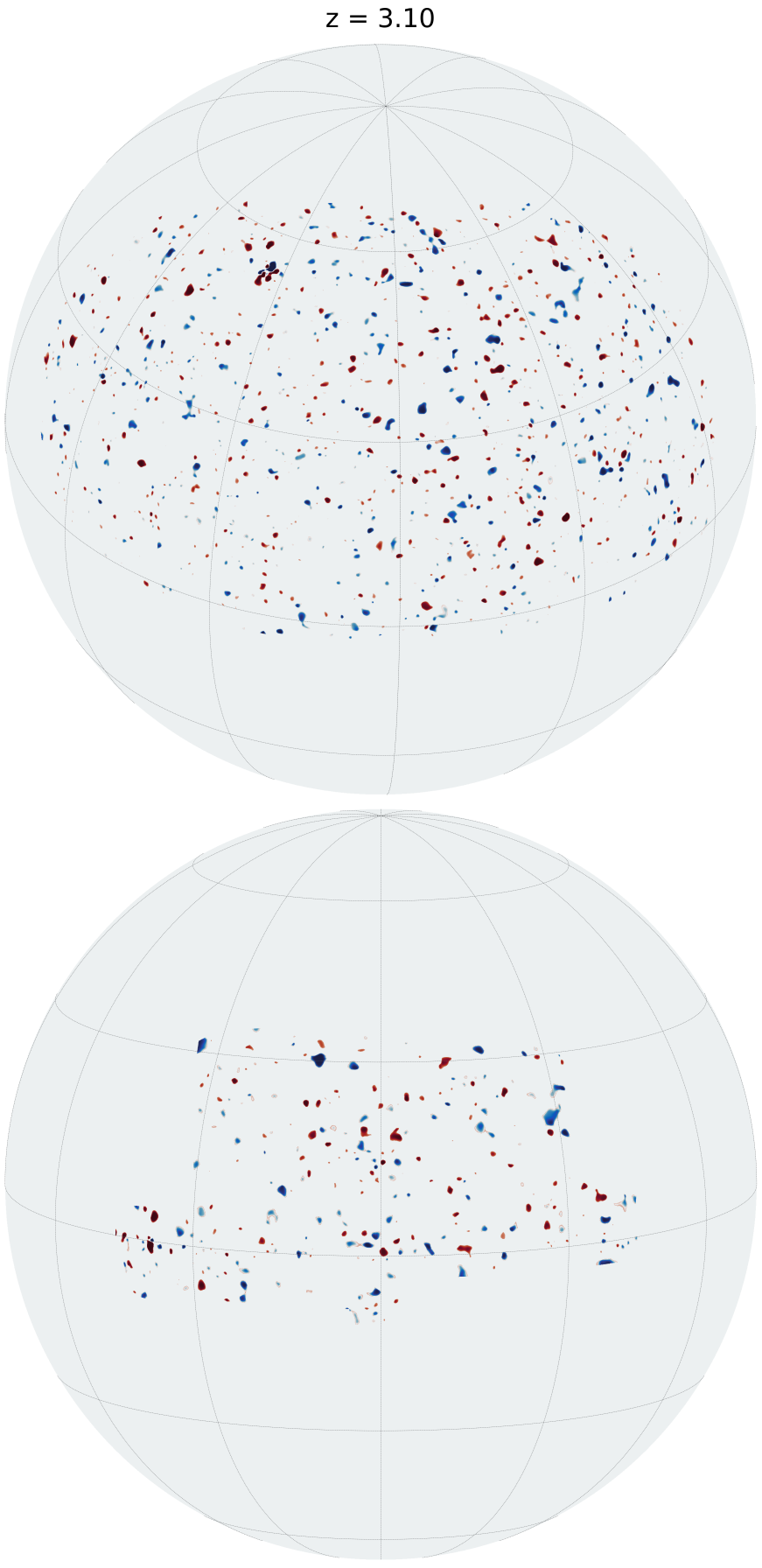


FIGURE 5.31: (Continued:) Orthographic projection of the redshift $z = 3.10$ candidates for galaxy protoclusters (red) and cosmic voids (blue) detected at a 4σ significance level, with the Northern Galactic Cap sky coverage shown on top and Southern Galactic Cap on bottom.

5.4 Data availability

All products of this work, including the Lyα absorption map, the associated standard error measurements, and the catalog of candidates for galaxy protoclusters and cosmic voids will be made publicly available at the webpage `http://stat.cmu.edu/Lyman-alpha-cosmos-map`. Two-dimensional HEALPix sky maps are available for download in FITS format at a 1.8 arcmin resolution ($N_{\rm side} = 2048$) and the full 47 h^{-3} Gpc^3 absorption field can be queried through a SQL database on a regular Cartesian grid up to a (1 h^{-1} Mpc)/voxel resolution.

Acknowledgements

We gratefully thank the Yale Center for Research Computing (YCRC) for accommodating our taxing use of their high-performance computing infrastructure, without which this work would not have been possible. This work was partially funded by the National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF), under NASA ATP grant NNX17AK56G, NASA ATP grant 80NSSC18K1015, and NSF grant AST1615940. This dissertation contains data collected by the Sloan Digital Sky Survey III (SDSS-III). Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is `http://www.sdss3.org/`. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University. This dissertation has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. This dissertation includes data collected by the *Kepler* mission. Funding for the *Kepler* mission is provided by the NASA Science Mission Directorate. This research contains data that was accessed from the Open Supernova Catalog, a centralized, open repository for supernova metadata, light curves, and

spectra. The Open Supernova Catalog web site is `https://sne.space`. The original sources of the supernova light curve studied in this paper are credited in the text.

Bibliography

[1] Ryan J. Tibshirani. Adaptive Piecewise Polynomial Estimation via Trend Filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

[2] Kyle S. Dawson, David J. Schlegel, Christopher P. Ahn, Scott F. Anderson, Éric Aubourg, Stephen Bailey, Robert H. Barkhouser, Julian E. Bautista, Alessand ra Beifiori, Andreas A. Berlind, Vaishali Bhardwaj, Dmitry Bizyaev, Cullen H. Blake, Michael R. Blanton, Michael Blomqvist, Adam S. Bolton, Arnaud Borde, Jo Bovy, W. N. Brandt, Howard Brewington, Jon Brinkmann, Peter J. Brown, Joel R. Brownstein, Kevin Bundy, N. G. Busca, William Carithers, Aurelio R. Carnero, Michael A. Carr, Yanmei Chen, Johan Comparat, Natalia Connolly, Frances Cope, Rupert A. C. Croft, Antonio J. Cuesta, Luiz N. da Costa, James R. A. Davenport, Timothée Delubac, Roland de Putter, Saurav Dhital, Anne Ealet, Garrett L. Ebelke, Daniel J. Eisenstein, S. Escoffier, Xiaohui Fan, N. Filiz Ak, Hayley Finley, Andreu Font-Ribera, R. Génova-Santos, James E. Gunn, Hong Guo, Daryl Haggard, Patrick B. Hall, Jean-Christophe Hamilton, Ben Harris, David W. Harris, Shirley Ho, David W. Hogg, Diana Holder, Klaus Honscheid, Joe Huehnerhoff, Beatrice Jordan, Wendell P. Jordan, Guinevere Kauffmann, Eyal A. Kazin, David Kirkby, Mark A. Klaene, Jean-Paul Kneib, Jean-Marc Le Goff, Khee-Gan Lee, Daniel C. Long, Craig P. Loomis, Britt Lundgren, Robert H. Lupton, Marcio A. G. Maia, Martin Makler, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Marc Manera, Claudia Maraston, Daniel Margala, Karen L. Masters, Cameron K. McBride, Patrick McDonald, Ian D. McGreer, Richard G. McMahon, Olga Mena, Jordi Miralda-Escudé, Antonio D. Montero-Dorta, Francesco Montesano, Demitri Muna, Adam D. Myers, Tracy Naugle, Robert C. Nichol, Pasquier Noterdaeme, Sebastián E. Nuza, Matthew D. Olmstead, Audrey Oravetz, Daniel J. Oravetz, Russell Owen, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, John K. Parejko, Isabelle Pâris, Will J. Percival, Ismael Pérez-Fournon, Ignasi Pérez-Ràfols, Patrick Petitjean, Robert Pfaffenberger, Janine Pforr, Matthew M. Pieri, Francisco Prada, Adrian M. Price-Whelan, M. Jordan Raddick, Rafael Rebolo, James Rich, Gordon T.

Richards, Constance M. Rockosi, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, J. A. Rubiño-Martin, Lado Samushia, Ariel G. Sánchez, Conor Sayres, Sarah J. Schmidt, Donald P. Schneider, C. G. Scóccola, Hee-Jong Seo, Alaina Shelden, Erin Sheldon, Yue Shen, Yiping Shu, Anže Slosar, Stephen A. Smee, Stephanie A. Snedden, Fritz Stauffer, Oliver Steele, Michael A. Strauss, Alina Streblyanska, Nao Suzuki, Molly E. C. Swanson, Tomer Tal, Masayuki Tanaka, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Christy A. Tremonti, M. Vargas Magaña, Licia Verde, Matteo Viel, David A. Wake, Mike Watson, Benjamin A. Weaver, David H. Weinberg, Benjamin J. Weiner, Andrew A. West, Martin White, W. M. Wood-Vasey, Christophe Yeche, Idit Zehavi, Gong-Bo Zhao, and Zheng Zheng. The Baryon Oscillation Spectroscopic Survey of SDSS-III. *The Astronomical Journal*, 145(1):10, 2013. doi: 10.1088/0004-6256/145/1/10.

[3] Joshua N. Winn. Transits and Occultations. *https://arxiv.org/abs/1001.2010*, 2010. URL `https://arxiv.org/abs/1001.2010`.

[4] Auvergne, M., Bodin, P., Boisnard, L., Buey, J.-T., Chaintreuil, S., Epstein, G., Jouret, M., Lam-Trong, T., Levacher, P., Magnan, A., Perez, R., Plasson, P., Plesseria, J., Peter, G., Steller, M., Tiphène, D., Baglin, A., Agogué, P., Appourchaux, T., Barbet, D., Beaufort, T., Bellenger, R., Berlin, R., Bernardi, P., Blouin, D., Boumier, P., Bonneau, F., Briet, R., Butler, B., Cautain, R., Chiavassa, F., Costes, V., Cuvilho, J., Cunha-Parro, V., De Oliveira Fialho, F., Decaudin, M., Defise, J.-M., Djalal, S., Docclo, A., Drummond, R., Dupuis, O., Exil, G., Fauré, C., Gaboriaud, A., Gamet, P., Gavalda, P., Grolleau, E., Gueguen, L., Guivarc'h, V., Guterman, P., Hasiba, J., Huntzinger, G., Hustaix, H., Imbert, C., Jeanville, G., Johlander, B., Jorda, L., Journoud, P., Karioty, F., Kerjean, L., Lafond, L., Lapeyrere, V., Landiech, P., Larqué, T., Laudet, P., Le Merrer, J., Leporati, L., Leruyet, B., Levieuge, B., Llebaria, A., Martin, L., Mazy, E., Mesnager, J.-M., Michel, J.-P., Moalic, J.-P., Monjoin, W., Naudet, D., Neukirchner, S., Nguyen-Kim, K., Ollivier, M., Orcesi, J.-L., Ottacher, H., Oulali, A., Parisot, J., Perruchot, S., Piacentino, A., Pinheiro da Silva, L., Platzer, J., Pontet, B., Pradines, A., Quentin, C., Rohbeck, U., Rolland, G., Rollenhagen, F., Romagnan, R., Russ, N., Samadi, R., Schmidt, R., Schwartz, N., Sebbag, I., Smit, H., Sunter, W., Tello, M., Toulouse, P., Ulmer, B., Vandermarcq, O., Vergnault, E., Wallner, R., Waultier, G., and Zanatta, P. The CoRoT satellite in flight: description and performance. *Astronomy & Astrophysics*, 506(1):411–424, 2009. doi: 10.1051/0004-6361/200810860. URL `https://doi.org/10.1051/0004-6361/200810860`.

[5] William J. Borucki, David Koch, Gibor Basri, Natalie Batalha, Timothy Brown, Douglas Caldwell, John Caldwell, Jørgen Christensen-Dalsgaard, William D. Cochran, Edna DeVore, Edward W. Dunham, Andrea K. Dupree, Thomas N. Gautier, John C. Geary, Ronald Gilliland, Alan Gould, Steve B. Howell, Jon M. Jenkins, Yoji Kondo, David W. Latham, Geoffrey W. Marcy, Søren Meibom, Hans Kjeldsen, Jack J. Lissauer, David G. Monet, David Morrison, Dimitar Sasselov, Jill Tarter, Alan Boss, Don Brownlee, Toby Owen, Derek Buzasi, David Charbonneau, Laurance Doyle, Jonathan Fortney, Eric B. Ford, Matthew J. Holman, Sara Seager, Jason H. Steffen, William F. Welsh, Jason Rowe, Howard Anderson, Lars Buchhave, David Ciardi, Lucianne Walkowicz, William Sherry, Elliott Horch, Howard Isaacson, Mark E. Everett, Debra Fischer, Guillermo Torres, John Asher Johnson, Michael Endl, Phillip MacQueen, Stephen T. Bryson, Jessie Dotson, Michael Haas, Jeffrey Kolodziejczak, Jeffrey Van Cleve, Hema Chandrasekaran, Joseph D. Twicken, Elisa V. Quintana, Bruce D. Clarke, Christopher Allen, Jie Li, Haley Wu, Peter Tenenbaum, Ekaterina Verner, Frederick Bruhweiler, Jason Barnes, and Andrej Prsa. Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327(5968):977–980, 2010.

[6] Steve B. Howell, Charlie Sobeck, Michael Haas, Martin Still, Thomas Barclay, Fergal Mullally, John Troeltzsch, Suzanne Aigrain, Stephen T. Bryson, Doug Caldwell, William J. Chaplin, William D. Cochran, Daniel Huber, Geoffrey W. Marcy, Andrea Miglio, Joan R. Najita, Marcie Smith, J. D. Twicken, and Jonathan J. Fortney. The K2 Mission: Characterization and Early Results. *Publications of the Astronomical Society of the Pacific*, 126 (938):398–408, April 2014.

[7] George R. Ricker, Joshua N. Winn, Roland Vanderspek, David W. Latham, Gáspár. Á. Bakos, Jacob L. Bean, Zachory K. Berta-Thompson, Timothy M. Brown, Lars Buchhave, Nathaniel R. Butler, R. Paul Butler, William J. Chaplin, David Charbonneau, Jørgen Christensen-Dalsgaard, Mark Clampin, Drake Deming, John Doty, Nathan De Lee, Courtney Dressing, E. W. Dunham, Michael Endl, Francois Fressin, Jian Ge, Thomas Henning, Matthew J. Holman, Andrew W. Howard, Shigeru Ida, Jon Jenkins, Garrett Jernigan, John A. Johnson, Lisa Kaltenegger, Nobuyuki Kawai, Hans Kjeldsen, Gregory Laughlin, Alan M. Levine, Douglas Lin, Jack J. Lissauer, Phillip MacQueen, Geoffrey Marcy, P. R. McCullough, Timothy D. Morton, Norio Narita, Martin Paegert, Enric Palle, Francesco Pepe, Joshua Pepper, Andreas Quirrenbach, S. A. Rinehart, Dimitar Sasselov, Bun'ei Sato, Sara Seager, Alessandro Sozzetti, Keivan G. Stassun, Peter Sullivan, Andrew Szentgyorgyi, Guillermo Torres, Stephane Udry, and Joel Villasenor. Transiting Exoplanet Survey Satellite

(TESS). In *Proceedings of SPIE*, volume 9143 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 914320, August 2014. doi: 10.1117/12.2063489.

[8] A. Prša, E. F. Guinan, E. J. Devinney, M. DeGeorge, D. H. Bradstreet, J. M. Giammarco, C. R. Alcock, and S. G. Engle. Artificial Intelligence Approach to the Determination of Physical Properties of Eclipsing Binaries. I. The EBAI Project. *The Astrophysical Journal*, 687(1):542–565, November 2008.

[9] Andrej Prša, Natalie Batalha, Robert W. Slawson, Laurance R. Doyle, William F. Welsh, Jerome A. Orosz, Sara Seager, Michael Rucker, Kimberly Mjaseth, Scott G. Engle, Kyle Conroy, Jon Jenkins, Douglas Caldwell, David Koch, and William Borucki. Kepler Eclipsing Binary Stars. I. Catalog and Principal Characterization of 1879 Eclipsing Binaries in the First Data Release. *The Astronomical Journal*, 141(3):83, February 2011.

[10] S. E. Woosley, D. Kasen, S. Blinnikov, and E. Sorokina. Type Ia Supernova Light Curves. *The Astrophysical Journal*, 662(1):487–503, June 2007.

[11] Jonathan Tennyson. *Astronomical Spectroscopy: an Introduction to the Atomic and Molecular Physics of Astronomical Spectroscopy*. World Scientific, 3rd edition, 2019.

[12] Anthony Aguirre, Lars Hernquist, Joop Schaye, Neal Katz, David H. Weinberg, and Jeffrey Gardner. Metal Enrichment of the Intergalactic Medium in Cosmological Simulations. *The Astrophysical Journal*, 561(2):521–549, November 2001. ISSN 1538-4357. doi: 10.1086/323370. URL `http://dx.doi.org/10.1086/323370`.

[13] Khee-Gan Lee, Joseph F. Hennawi, David N. Spergel, David H. Weinberg, David W. Hogg, Matteo Viel, James S. Bolton, Stephen Bailey, Matthew M. Pieri, William Carithers, David J. Schlegel, Britt Lundgren, Nathalie Palanque-Delabrouille, Nao Suzuki, Donald P. Schneider, and Christophe Yèche. IGM Constraints from the SDSS-III/BOSS DR9 Lyα Forest Transmission Probability Distribution Function. *The Astrophysical Journal*, 799(2):196, 2015. doi: 10.1088/0004-637X/799/2/196.

[14] Patrick McDonald, Uroš Seljak, Renyue Cen, David Shih, David H. Weinberg, Scott Burles, Donald P. Schneider, David J. Schlegel, Neta A. Bahcall, John W. Briggs, J. Brinkmann, Masataka Fukugita, Željko Ivezić, Stephen Kent, and Daniel E. Vanden Berk. The Linear Theory Power Spectrum from the Lyα Forest in the Sloan Digital Sky Survey. *The Astrophysical Journal*, 635(2):761–783, December 2005.

[15] Nathalie Palanque-Delabrouille, Christophe Yèche, Arnaud Borde, Jean-Marc Le Goff, Graziano Rossi, Matteo Viel, Éric Aubourg, Stephen Bailey, Julian Bautista, Michael Blomqvist, Adam Bolton, James S. Bolton, Nicolás G. Busca, Bill Carithers, Rupert A. C. Croft, Kyle S. Dawson, Timothée Delubac, Andreu Font-Ribera, Shirley Ho, David Kirkby, Khee-Gan Lee, Daniel Margala, Jordi Miralda-Escudé, Demitri Muna, Adam D. Myers, Pasquier Noterdaeme, Isabelle Pâris, Patrick Petitjean, Matthew M. Pieri, James Rich, Emmanuel Rollinde, Nicholas P. Ross, David J. Schlegel, Donald P. Schneider, Anže Slosar, and David H. Weinberg. The one-dimensional Lyα forest power spectrum from BOSS. *Astronomy & Astrophysics*, 559:A85, 2013. doi: 10.1051/0004-6361/201322130.

[16] Andreu Arinyo i Prats, Jordi Miralda-Escudé, Matteo Viel, and Renyue Cen. The non-linear power spectrum of the Lyman alpha forest. *Journal of Cosmology and Astroparticle Physics*, 2015(12):017–017, December 2015. doi: 10.1088/1475-7516/2015/12/017. URL `https://doi.org/10.1088%2F1475-7516%2F2015%2F12%2F017`.

[17] Anže Slosar, Andreu Font-Ribera, Matthew M Pieri, James Rich, Jean-Marc Le Goff, Éric Aubourg, Jon Brinkmann, Nicolas Busca, Bill Carithers, Romain Charlassier, Marina Cortês, Rupert Croft, Kyle S Dawson, Daniel Eisenstein, Jean-Christophe Hamilton, Shirley Ho, Khee-Gan Lee, Robert Lupton, Patrick McDonald, Bumbarija Medolin, Demitri Muna, Jordi Miralda-Escudé, Adam D Myers, Robert C Nichol, Nathalie Palanque-Delabrouille, Isabelle Pâris, Patrick Petitjean, Yodovina Piškur, Emmanuel Rollinde, Nicholas P Ross, David J Schlegel, Donald P Schneider, Erin Sheldon, Benjamin A Weaver, David H Weinberg, Christophe Yeche, and Donald G York. The Lyman-α forest in three dimensions: measurements of large scale flux correlations from BOSS 1st-year data. *Journal of Cosmology and Astroparticle Physics*, 2011(09):001, September 2011.

[18] N. G. Busca et al. Baryon Acoustic Oscillations in the Ly-α forest of BOSS quasars. *Astron. Astrophys.*, 552:A96, 2013. doi: 10.1051/0004-6361/201220724. URL `https://doi.org/10.1051/0004-6361/201220724`.

[19] A. Slosar et al. Measurement of baryon acoustic oscillations in the Lyman-α forest fluctuations in BOSS data release 9. *Journal of Cosmology and Astroparticle Physics*, 2013 (04):026–026, April 2013. doi: 10.1088/1475-7516/2013/04/026. URL `https://iopscience.iop.org/article/10.1088/1475-7516/2013/04/026`.

[20] Bautista, Julian E., Busca, Nicolás G., Guy, Julien, Rich, James, Blomqvist, Michael, du Mas des Bourboux, Hélion, Pieri, Matthew M., Font-Ribera, Andreu, Bailey, Stephen,

Delubac, Timothée, Kirkby, David, Le Goff, Jean-Marc, Margala, Daniel, Slosar, Anze, Vazquez, Jose Alberto, Brownstein, Joel R., Dawson, Kyle S., Eisenstein, Daniel J., Miralda-Escudé, Jordi, Noterdaeme, Pasquier, Palanque-Delabrouille, Nathalie, Pâris, Isabelle, Petitjean, Patrick, Ross, Nicholas P., Schneider, Donald P., Weinberg, David H., and Yèche, Christophe. Measurement of baryon acoustic oscillation correlations at $z = 2.3$ with SDSS DR12 Lyα-Forests. *Astronomy & Astrophysics*, 603:A12, 2017. doi: 10.1051/0004-6361/201730533. URL `https://doi.org/10.1051/0004-6361/201730533`.

[21] C. Pichon, J. L. Verily, E. Rolland, S. Columbi, and P. Petitjean. Inversion of the Lyman α forest: three-dimensional investigation of the intergalactic medium. *Monthly Notices of the Royal Astronomical Society*, 326:597–620, 2001.

[22] S. Caucci, S. Colombi, C. Pichon, E. Rollinde, P. Petitjean, and T. Sousbie. Recovering the topology of the intergalactic medium at $z \sim 2$. *Monthly Notices of the Royal Astronomical Society*, 386:211–229, 2008.

[23] Khee-Gan Lee, Joseph F. Hennawi, Casey Stark, J. Xavier Prochaska, Martin White, David J. Schlegel, Anna-Christina Eilers, Andreu Arinyo i Prats, Nao Suzuki, Rupert A. C. Croft, Karina I. Caputi, Paolo Cassata, Olivier Ilbert, Bianca Garilli, Anton M. Koekemoer, Vincent Le Brun, Olivier Le Fèvre, Dario Maccagni, Peter Nugent, Yoshiaki Taniguchi, Lidia A. M. Tasca, Laurence Tresse, Gianni Zamorani, and Elena Zucca. Lyman-alpha Forest Tomography from Background Galaxies: The First Megaparsec-Resolution Large-Scale Structure Map at $z > 2$. *The Astrophysical Journal Letters*, 795(1):12, October 2014.

[24] Khee-Gan Lee, Alex Krolewski, Martin White, David Schlegel, Peter E. Nugent, Joseph F. Hennawi, Thomas Müller, Richard Pan, J. Xavier Prochaska, Andreu Font-Ribera, Nao Suzuki, Karl Glazebrook, Glenn G. Kacprzak, Jeyhan S. Kartaltepe, Anton M. Koekemoer, Olivier Le Fèvre, Brian C. Lemaux, Christian Maier, Themiya Nanayakkara, R. Michael Rich, D. B. Sanders, Mara Salvato, Lidia Tasca, and Kim-Vy H. Tran. First Data Release of the COSMOS Lyα Mapping and Tomography Observations: 3D Lyα Forest Tomography at $2.05 < z < 2.55$. *The Astrophysical Journal Supplement Series*, 237(2):31, August 2018. doi: 10.3847/1538-4365/aace58.

[25] Jessi Cisewski, Rupert A. C. Croft, Peter E. Freeman, Christopher R. Genovese, Nishikanta Khandai, Melih Ozbek, and Larry Wasserman. Non-parametric 3D map of the intergalactic

medium using the Lyman-alpha forest. *Monthly Notices of the Royal Astronomical Society*, 440(3):2599–2609, April 2014.

[26] Daniel J. Eisenstein, David H. Weinberg, Eric Agol, Hiroaki Aihara, Carlos Allende Prieto, Scott F. Anderson, James A. Arns, Éric Aubourg, Stephen Bailey, Eduardo Balbinot, Robert Barkhouser, Timothy C. Beers, Andreas A. Berlind, Steven J. Bickerton, Dmitry Bizyaev, Michael R. Blanton, John J. Bochanski, Adam S. Bolton, Casey T. Bosman, Jo Bovy, W. N. Brandt, Ben Breslauer, Howard J. Brewington, J. Brinkmann, Peter J. Brown, Joel R. Brownstein, Dan Burger, Nicolas G. Busca, Heather Campbell, Phillip A. Cargile, William C. Carithers, Joleen K. Carlberg, Michael A. Carr, Liang Chang, Yanmei Chen, Cristina Chiappini, Johan Comparat, Natalia Connolly, Marina Cortes, Rupert A. C. Croft, Katia Cunha, Luiz N. da Costa, James R. A. Davenport, Kyle Dawson, Nathan De Lee, Gustavo F. Porto de Mello, Fernando de Simoni, Janice Dean, Saurav Dhital, Anne Ealet, Garrett L. Ebelke, Edward M. Edmondson, Jacob M. Eiting, Stephanie Escoffier, Massimiliano Esposito, Michael L. Evans, Xiaohui Fan, Bruno Femenía Castellá, Leticia Dutra Ferreira, Greg Fitzgerald, Scott W. Fleming, Andreu Font-Ribera, Eric B. Ford, Peter M. Frinchaboy, Ana Elia García Pérez, B. Scott Gaudi, Jian Ge, Luan Ghezzi, Bruce A. Gillespie, G. Gilmore, Léo Girardi, J. Richard Gott, Andrew Gould, Eva K. Grebel, James E. Gunn, Jean-Christophe Hamilton, Paul Harding, David W. Harris, Suzanne L. Hawley, Frederick R. Hearty, Joseph F. Hennawi, Jonay I. González Hernández, Shirley Ho, David W. Hogg, Jon A. Holtzman, Klaus Honscheid, Naohisa Inada, Inese I. Ivans, Linhua Jiang, Peng Jiang, Jennifer A. Johnson, Cathy Jordan, Wendell P. Jordan, Guinevere Kauffmann, Eyal Kazin, David Kirkby, Mark A. Klaene, G. R. Knapp, Jean-Paul Kneib, C. S. Kochanek, Lars Koesterke, Juna A. Kollmeier, Richard G. Kron, Hubert Lampeitl, Dustin Lang, James E. Lawler, Jean-Marc Le Goff, Brian L. Lee, Young Sun Lee, Jarron M. Leisenring, Yen-Ting Lin, Jian Liu, Daniel C. Long, Craig P. Loomis, Sara Lucatello, Britt Lundgren, Robert H. Lupton, Bo Ma, Zhibo Ma, Nicholas MacDonald, Claude Mack, Suvrath Mahadevan, Marcio A. G. Maia, Steven R. Majewski, Martin Makler, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Claudia Maraston, Daniel Margala, Paul Maseman, Karen L. Masters, Cameron K. McBride, Patrick McDonald, Ian D. McGreer, Richard G. McMahon, Olga Mena Requejo, Brice Ménard, Jordi Miralda-Escudé, Heather L. Morrison, Fergal Mullally, Demitri Muna, Hitoshi Murayama, Adam D. Myers, Tracy NAugustle, Angelo Fausti Neto, Duy Cuong Nguyen, Robert C. Nichol, David L. Nidever, Robert W. O'Connell, Ricardo L. C. Ogando, Matthew D. Olmstead,

Daniel J. Oravetz, Nikhil Padmanabhan, Martin Paegert, Nathalie Palanque-Delabrouille, Kaike Pan, Parul Pandey, John K. Parejko, Isabelle Pâris, Paulo Pellegrini, Joshua Pepper, Will J. Percival, Patrick Petitjean, Robert Pfaffenberger, Janine Pforr, Stefanie Phleps, Christophe Pichon, Matthew M. Pieri, Francisco Prada, Adrian M. Price-Whelan, M. Jordan Raddick, Beatriz H. F. Ramos, I. Neill Reid, Celine Reyle, James Rich, Gordon T. Richards, George H. Rieke, Marcia J. Rieke, Hans-Walter Rix, Annie C. Robin, Helio J. Rocha-Pinto, Constance M. Rockosi, Natalie A. Roe, Emmanuel Rollinde, Ashley J. Ross, Nicholas P. Ross, Bruno Rossetto, Ariel G. Sánchez, Basilio Santiago, Conor Sayres, Ricardo Schiavon, David J. Schlegel, Katharine J. Schlesinger, Sarah J. Schmidt, Donald P. Schneider, Kris Sellgren, Alaina Shelden, Erin Sheldon, Matthew Shetrone, Yiping Shu, John D. Silverman, Jennifer Simmerer, Audrey E. Simmons, Thirupathi Sivarani, M. F. Skrutskie, Anže Slosar, Stephen Smee, Verne V. Smith, Stephanie A. Snedden, Keivan G. Stassun, Oliver Steele, Matthias Steinmetz, Mark H. Stockett, Todd Stollberg, Michael A. Strauss, Alexander S. Szalay, Masayuki Tanaka, Aniruddha R. Thakar, Daniel Thomas, Jeremy L. Tinker, Benjamin M. Tofflemire, Rita Tojeiro, Christy A. Tremonti, Mariana Vargas Magaña, Licia Verde, Nicole P. Vogt, David A. Wake, Xiaoke Wan, Ji Wang, Benjamin A. Weaver, Martin White, Simon D. M. White, John C. Wilson, John P. Wisniewski, W. Michael Wood-Vasey, Brian Yanny, Naoki Yasuda, Christophe Yèche, Donald G. York, Erick Young, Gail Zasowski, Idit Zehavi, and Bo Zhao. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. *The Astronomical Journal*, 142(3):72, August 2011.

[27] Shadab Alam, Franco D. Albareti, Carlos Allende Prieto, F. Anders, Scott F. Anderson, Timothy Anderton, Brett H. Andrews, Eric Armengaud, Éric Aubourg, Stephen Bailey, Sarbani Basu, Julian E. Bautista, Rachael L. Beaton, Timothy C. Beers, Chad F. Bender, Andreas A. Berlind, Florian Beutler, Vaishali Bhardwaj, Jonathan C. Bird, Dmitry Bizyaev, Cullen H. Blake, Michael R. Blanton, Michael Blomqvist, John J. Bochanski, Adam S. Bolton, Jo Bovy, A. Shelden Bradley, W. N. Brandt, D. E. Brauer, J. Brinkmann, Peter J. Brown, Joel R. Brownstein, Angela Burden, Etienne Burtin, Nicolás G. Busca, Zheng Cai, Diego Capozzi, Aurelio Carnero Rosell, Michael A. Carr, Ricardo Carrera, K. C. Chambers, William James Chaplin, Yen-Chi Chen, Cristina Chiappini, S. Drew Chojnowski, Chia-Hsun Chuang, Nicolas Clerc, Johan Comparat, Kevin Covey, Rupert A. C. Croft, Antonio J. Cuesta, Katia Cunha, Luiz N. da Costa, Nicola Da Rio, James R. A. Davenport, Kyle S. Dawson, Nathan De Lee, Timothée Delubac, Rohit Deshpande, Saurav Dhital, Letícia

Dutra-Ferreira, Tom Dwelly, Anne Ealet, Garrett L. Ebelke, Edward M. Edmondson, Daniel J. Eisenstein, Tristan Ellsworth, Yvonne Elsworth, Courtney R. Epstein, Michael Eracleous, Stephanie Escoffier, Massimiliano Esposito, Michael L. Evans, Xiaohui Fan, Emma Fernández-Alvar, Diane Feuillet, Nurten Filiz Ak, Hayley Finley, Alexis Finoguenov, Kevin Flaherty, Scott W. Fleming, Andreu Font-Ribera, Jonathan Foster, Peter M. Frinchaboy, J. G. Galbraith-Frew, Rafael A. García, D. A. García-Hernández, Ana E. García Pérez, Patrick Gaulme, Jian Ge, R. Génova-Santos, A. Georgakakis, Luan Ghezzi, Bruce A. Gillespie, Léo Girardi, Daniel Goddard, Satya Gontcho A Gontcho, Jonay I. González Hernández, Eva K. Grebel, Paul J. Green, Jan Niklas Grieb, Nolan Grieves, James E. Gunn, Hong Guo, Paul Harding, Sten Hasselquist, Suzanne L. Hawley, Michael Hayden, Fred R. Hearty, Saskia Hekker, Shirley Ho, David W. Hogg, Kelly Holley-Bockelmann, Jon A. Holtzman, Klaus Honscheid, Daniel Huber, Joseph Huehnerhoff, Inese I. Ivans, Linhua Jiang, Jennifer A. Johnson, Karen Kinemuchi, David Kirkby, Francisco Kitaura, Mark A. Klaene, Gillian R. Knapp, Jean-Paul Kneib, Xavier P. Koenig, Charles R. Lam, Ting-Wen Lan, Dustin Lang, Pierre Laurent, Jean-Marc Le Goff, Alexie Leauthaud, Khee-Gan Lee, Young Sun Lee, Timothy C. Licquia, Jian Liu, Daniel C. Long, Martín López-Corredoira, Diego Lorenzo-Oliveira, Sara Lucatello, Britt Lundgren, Robert H. Lupton, Claude E. Mack III, Suvrath Mahadevan, Marcio A. G. Maia, Steven R. Majewski, Elena Malanushenko, Viktor Malanushenko, A. Manchado, Marc Manera, Qingqing Mao, Claudia Maraston, Robert C. Marchwinski, Daniel Margala, Sarah L. Martell, Marie Martig, Karen L. Masters, Savita Mathur, Cameron K. McBride, Peregrine M. McGehee, Ian D. McGreer, Richard G. McMahon, Brice Ménard, Marie-Luise Menzel, Andrea Merloni, Szabolcs Mészáros, Adam A. Miller, Jordi Miralda-Escudé, Hironao Miyatake, Antonio D. Montero-Dorta, Surhud More, Eric Morganson, Xan Morice-Atkinson, Heather L. Morrison, Benoit Mosser, Demitri Muna, Adam D. Myers, Kirpal Nandra, Jeffrey A. Newman, Mark Neyrinck, Duy Cuong Nguyen, Robert C. Nichol, David L. Nidever, Pasquier Noterdaeme, Sebastián E. Nuza, Julia E. O'Connell, Robert W. O'Connell, Ross O'Connell, Ricardo L. C. Ogando, Matthew D. Olmstead, Audrey E. Oravetz, Daniel J. Oravetz, Keisuke Osumi, Russell Owen, Deborah L. Padgett, Nikhil Padmanabhan, Martin Paegert, Nathalie Palanque-Delabrouille, Kaike Pan, John K. Parejko, Isabelle Pâris, Changbom Park, Petchara Pattarakijwanich, M. Pellejero-Ibanez, Joshua Pepper, Will J. Percival, Ismael Pérez-Fournon, Ignasi Perez-Rafols, Patrick Petitjean, Matthew M. Pieri, Marc H. Pinsonneault, Gustavo F. Porto de Mello, Francisco Prada, Abhishek Prakash, Adrian M. Price-Whelan, Pavlos Protopapas, M. Jordan Raddick, Mubdi Rahman, Beth A. Reid, James Rich, Hans-Walter Rix, Annie C.

Robin, Constance M. Rockosi, Thaíse S. Rodrigues, Sergio Rodríguez-Torres, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, John J. Ruan, J. A. Rubiño-Martín, Eli S. Rykoff, Salvador Salazar-Albornoz, Mara Salvato, Lado Samushia, Ariel G. Sánchez, Basílio Santiago, Conor Sayres, Ricardo P. Schiavon, David J. Schlegel, Sarah J. Schmidt, Donald P. Schneider, Mathias Schultheis, Axel D. Schwope, C. G. Scóccola, Caroline Scott, Kris Sellgren, Hee-Jong Seo, Aldo Serenelli, Neville Shane, Yue Shen, Matthew Shetrone, Yiping Shu, V. Silva Aguirre, Thirupathi Sivarani, M. F. Skrutskie, Anže Slosar, Verne V. Smith, Flávia Sobreira, Diogo Souto, Keivan G. Stassun, Matthias Steinmetz, Dennis Stello, Michael A. Strauss, Alina Streblyanska, Nao Suzuki, Molly E. C. Swanson, Jonathan C. Tan, Jamie Tayar, Ryan C. Terrien, Aniruddha R. Thakar, Daniel Thomas, Neil Thomas, Benjamin A. Thompson, Jeremy L. Tinker, Rita Tojeiro, Nicholas W. Troup, Mariana Vargas-Magaña, Jose A. Vazquez, Licia Verde, Matteo Viel, Nicole P. Vogt, David A. Wake, Ji Wang, Benjamin A. Weaver, David H. Weinberg, Benjamin J. Weiner, Martin White, John C. Wilson, John P. Wisniewski, W. M. Wood-Vasey, Christophe Yeche, Donald G. York, Nadia L. Zakamska, O. Zamora, Gail Zasowski, Idit Zehavi, Gong-Bo Zhao, Zheng Zheng, Xu Zhou, Zhimin Zhou, Hu Zou, and Guangtun Zhu. The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. *The Astrophysical Journal Supplement Series*, 219(1):12, July 2015.

[28] Ryan J. Tibshirani and Jonathan Taylor. The Solution Path of the Generalized Lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

[29] Collin A. Politsch, Jessi Cisewski-Kehe, Rupert A. C. Croft, and Larry Wasserman. Trend Filtering - I: A Modern Statistical Tool for Time-Domain Astronomy and Astronomical Spectroscopy. *Monthly Notices of the Royal Astronomical Society*, 492(3):4005–4018, March 2020.

[30] Aaditya Ramdas and Ryan J. Tibshirani. Fast and Flexible ADMM Algorithms for Trend Filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858, 2016.

[31] Collin A. Politsch, Jessi Cisewski-Kehe, Rupert A. C. Croft, and Larry Wasserman. Trend Filtering - II: Denoising Astronomical Signals with Varying Degrees of Smoothness. *Monthly Notices of the Royal Astronomical Society*, 492(3):4019–4032, March 2020.

[32] G. Dimitriadis, M. Sullivan, W. Kerzendorf, A. J. Ruiter, I. R. Seitenzahl, S. Taubenberger, G. B. Doran, A. Gal-Yam, R. R. Laher, K. Maguire, P. Nugent, E. O. Ofek, and J. Surace.

The late-time light curve of the Type Ia supernova SN 2011fe. *Monthly Notices of the Royal Astronomical Society*, 468(4):3798–3812, July 2017.

[33] Alexey Tolstov, Ken'ichi Nomoto, Elena Sorokina, Sergei Blinnikov, Nozomu Tominaga, and Yoshiaki Taniguchi. Light-curve Modeling of Fast-evolving Supernova KSN 2015K: Explosion in Circumstellar Matter of a Super-AGB Progenitor. *The Astrophysical Journal*, 881(1):35, August 2019.

[34] Patrick B. Hall, Scott F. Anderson, Michael Strauss, Donald York, Gordon T. Richards, Xiaohui Fan, G R. Knapp, Donald P. Schneider, Daniel E. Vanden Berk, Thomas Geballe, A Bauer, Robert H. Becker, Marc Davis, Hans-Walter Rix, R C. Nichol, N Bahcall, Jon Brinkmann, Robert Brunner, A J. Connolly, and Wei Zheng. Unusual Broad Absorption Line Quasars from the Sloan Digital Sky Survey. *The Astrophysical Journal Supplement Series*, 141, March 2002.

[35] Rupert A. C. Croft, David H. Weinberg, Mike Bolte, Scott Burles, Lars Hernquist, Neal Katz, David Kirkman, and David Tytler. Toward a Precise Measurement of Matter Clustering: Lyα Forest Data at Redshifts 2-4. *The Astrophysical Journal*, 581(1):20–52, December 2002.

[36] Jason L. Maron and Gregory G. Howes. Gradient Particle Magnetohydrodynamics: A Lagrangian Particle Code for Astrophysical Magnetohydrodynamics. *The Astrophysical Journal*, 595(1):564–572, September 2003.

[37] S. E. Persson, B. F. Madore, W. Krzemiński, W. L. Freedman, M. Roth, and D. C. Murphy. New Cepheid Period-Luminosity Relations for the Large Magellanic Cloud: 92 Near-Infrared Light Curves. *The Astronomical Journal*, 128:2239–2264, November 2004.

[38] Hiranya V. Peiris and Licia Verde. The Shape of the Primordial Power Spectrum: A Last Stand Before Planck. *Physical Review D*, 81:021302, January 2010.

[39] Carlos Contreras, Mario Hamuy, M. M. Phillips, Gastón Folatelli, Nicholas B. Suntzeff, S. E. Persson, Maximilian Stritzinger, Luis Boldt, Sergio González, Wojtek Krzeminski, Nidia Morrell, Miguel Roth, Francisco Salgado, María José Maureira, Christopher R. Burns, W. L. Freedman, Barry F. Madore, David Murphy, Pamela Wyatt, Weidong Li, and Alexei V. Filippenko. The Carnegie Supernova Project: First Photometry Data Release of Low-Redshift Type Ia Supernovae. *The Astronomical Journal*, 139(2):519–539, January 2010.

[40] S. Dhawan, B. Leibundgut, J. Spyromilio, and K. Maguire. Near-infrared light curves of Type Ia supernovae: studying properties of the second maximum. *Monthly Notices of the Royal Astronomical Society*, 448(2):1345–1359, April 2015.

[41] Neale P. Gibson, Suzanne Aigrain, Scott Roberts, Thomas M. Evans, Michael Osborne, F. Pont University of Oxford, and University of Exeter. A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy. *Monthly Notices of the Royal Astronomical Society*, 419(3):2683–2694, January 2012.

[42] S. Aigrain, H. Parviainen, and B. J. S. Pope. K2SC: Flexible systematics correction and detrending of K2 light curves using Gaussian Process regression. *Monthly Notices of the Royal Astronomical Society*, 459(3):2408–2419, April 2016.

[43] Adrià Gómez-Valent and Luca Amendola. H_0 from cosmic chronometers and Type Ia supernovae, with Gaussian Processes and the novel Weighted Polynomial Regression method. *Journal of Computational and Astroparticle Physics*, 2018(04):51, April 2018.

[44] M. Fligge and S. K. Solanki. Noise Reduction in Astronomical Spectra: A New Wavelet-Based Method. *Astronomy & Astrophysics Supplement Series*, 124:579–587, September 1997.

[45] Tom Theuns and Saleem Zaroubi. A wavelet analysis of the spectra of quasi-stellar objects. *Monthly Notices of the Royal Astronomical Society*, 317(4):989–995, October 2000.

[46] Vahid Z. Golkhou and Nathaniel R. Butler. Uncovering the Intrinsic Variability of Gamma-Ray Bursts. *The Astrophysical Journal*, 787(1):90, May 2014.

[47] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics, 2002.

[48] Larry Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics, 2006.

[49] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

[50] Arkadi Nemirovskii, Boris Polyak, and Alexandre Tsybakov. Rate of convergence of non-paramtric estimates of maximum-likelihood type. *Problems of Information Transmission*, 21, January 1985.

[51] Arkadi Nemirovskii. Nonparametric estimation of smooth regression function. *Izv. Akad. Nauk. SSSR Tekhn. Kibernet. (in Russian)*, 3:50–60, 1985.

[52] David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(8):879–921, 1998.

[53] Hans-Georg Muller and Ulrich Stadtmuller. Estimation of Heteroscedasticity in Regression Analysis. *The Annals of Statistics*, 15(2):610–625, June 1987.

[54] Jianqing Fan and Irene Gijbels. Variable Bandwidth and Local Linear Regression Smoothers. *The Annals of Statistics*, 20:2008–2036, 1992.

[55] Jianqing Fan and Irene Gijbels. Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(2):371–394, 1995.

[56] O. V. Lepski, E. Mammen, and V. G. Spokoiny. Optimal Spatial Adaptation to Inhomogeneous Smoothness: An Approach Based on Kernel Estimates with Variable Bandwidth Selectors. *The Annals of Statistics*, 25(3):929–947, June 1997.

[57] Irene Gijbels and Enno Mammen. Local Adaptivity of Kernel Estimates with Plug?in Local Bandwidth Selectors. *Scandinavian Journal of Statistics*, 25(3):503–520, 1998.

[58] Carl De Boor. Good Approximation by Splines with Variable Knots. II. In *Conference on the Numerical Solution of Differential Equations*, pages 12–20. Springer, 1974.

[59] David L. B. Jupp. Approximation to Data by Splines with Free Knots. *SIAM Journal on Numerical Analysis*, 15(2):328–343, 1978.

[60] Ilaria Dimatteo, Christopher R. Genovese, and Robert E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, December 2001.

[61] Alexandra M. Schmidt and Anthony O'Hagan. Bayesian Inference for Non-Stationary Spatial Covariance Structure via Spatial Deformations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(3):743–758, 2003.

[62] Christopher J. Paciorek and Mark J. Schervish. Nonstationary Covariance Functions for Gaussian Process Regression. In *Advances in neural information processing systems*, pages 273–280, 2004.

[63] Christopher J. Paciorek and Mark J. Schervish. Spatial Modelling Using a New Class of Nonstationary Covariance Functions. *Environmetrics*, 17(5):483–506, 2006.

[64] David L. Donoho and Iain M. Johnstone. Minimax Risk over l_p-Balls for l_q-error. *Probability Theory and Related Fields*, 99:277–303, June 1994.

[65] D. A. Howell, M. Sullivan, K. Perrett, T. J. Bronder, I. M. Hook, P. Astier, E. Aubourg, D. Balam, S. Basa, R. G. Carlberg, S. Fabbro, D. Fouchez, J. Guy, H. Lafoux, J. D. Neill, R. Pain, N. Palanque-Delabrouille, C. J. Pritchet, N. Regnault, J. Rich, R. Taillet, R. Knop, R. G. McMahon, S. Perlmutter, and N. A. Walton. Gemini Spectroscopy of Supernovae from the Supernova Legacy Survey: Improving High-Redshift Supernova Selection and Classification. *The Astrophysical Journal*, 634(2):1190–1201, December 2005.

[66] Adam S. Bolton, David J. Schlegel, Éric Aubourg, Stephen Bailey, Vaishali Bhardwaj, Joel R. Brownstein, Scott Burles, Yan-Mei Chen, Kyle Dawson, Daniel J. Eisenstein, James E. Gunn, G. R. Knapp, Craig P. Loomis, Robert H. Lupton, Claudia Maraston, Demitri Muna, Adam D. Myers, Matthew D. Olmstead, Nikhil Padmanabhan, Isabelle Pâris, Will J. Percival, Patrick Petitjean, Constance M. Rockosi, Nicholas P. Ross, Donald P. Schneider, Yiping Shu, Michael A. Strauss, Daniel Thomas, Christy A. Tremonti, David A. Wake, Benjamin A. Weaver, and W. Michael Wood-Vasey. Spectral Classification and Redshift Measurement for the SDSS-III Baryon Oscillation Spectroscopic Survey. *The Astronomical Journal*, 144(5):144, October 2012.

[67] A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[68] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[69] I. A. Ibragimov and R. Z. Hasminiskii. Asymptotic efficiency bounds for non-parametric estimation of a regression function in l_p. *Zapiski Nauchnykh Seminarov LOMI (in Russian)*, 97:88–101, 1980.

[70] Charles J. Stone. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040–1053, December 1982.

[71] Jianqing Fan. Local Linear Regression Smoothers and Their Minimax Efficiencies. *The Annals of Statistics*, 21(1):196–216, March 1993.

[72] Jianqing Fan, Theo Gasser, Irene Gijbels, Michael Brockmann, and Joachim Engel. Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency. *Annals of the Institute of Statistical Mathematics*, 49:79–99, 1997.

[73] Michael Nussbaum. Spline Smoothing in Regression Models and Asymptotic Efficiency in L_2. *The Annals of Statistics*, 13(3):984–997, 1985.

[74] Ryan J. Tibshirani. Degrees of freedom and model search. *Statistica Sinica*, pages 1265–1296, 2015.

[75] Gabriele Steidl, Stephan Didas, and Julia Neumann. Splines in Higher Order TV Regularization. *International Journal of Computer Vision*, 70(3):241–255, 2006.

[76] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 Trend Filtering. *SIAM Review*, 51(2):339–360, 2009.

[77] Yu-Xiang Wang, James Sharpnack, Alexander J. Smola, and Ryan J. Tibshirani. Trend Filtering on Graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016.

[78] Géza Kovács, Gáspár Bakos, and Robert W. Noyes. A Trend Filtering Algorithm for wide-field variability surveys. *Monthly Notices of the Royal Astronomical Society*, 356(2): 557–567, January 2005.

[79] Carl De Boor. A Practical Guide to Splines. Applied Mathematical Sciences, 1978.

[80] Grace Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, 1990.

[81] Yu-Xiang Wang, Alex Smola, and Ryan Tibshirani. The Falling Factorial Basis and Its Statistical Applications. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 730–738, Bejing, China, June 2014.

[82] Leonid I. Rudin, Stanely Osher, and Emad Faterni. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.

[83] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67 (1):91–108, 2005.

[84] Stephanie R. Land and Jerome H. Friedman. Variable Fusion: A New Method of Adaptive Signal Regression. Technical report, Department of Statistics, Stanford University, 1996.

[85] Julian E. Bautista, Stephen Bailey, Andreu Font-Ribera, Matthew M. Pieri, Nicolas G. Busca, Jordi Miralda-Escudé, Nathalie Palanque-Delabrouille, James Rich, Kyle Dawson, Yu Feng, and et al. Mock Quasar-Lyman-α Forest Data-sets for the SDSS-III Baryon Oscillation Spectroscopic Survey. *Journal of Computational and Astroparticle Physics*, 1505 (5):60, 2015.

[86] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[87] Taylor B. Arnold and Ryan J. Tibshirani. Efficient Implementations of the Generalized Lasso Dual Path Algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016.

[88] Taylor B. Arnold and Ryan J. Tibshirani. Path algorithm for generalized lasso problems, July 2018. URL `https://github.com/glmgen`.

[89] Simon Kornblith. Lasso/Elastic Net linear and generalized linear models, 2014. URL `https://github.com/JuliaStats/Lasso.jl`.

[90] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. `l1_tf`: Software for l1 Trend Filtering, May 2008. URL `http://stanford.edu/~boyd/l1_tf/`.

[91] Steven Diamond and Stephen Boyd. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[92] Charles M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151, November 1981.

[93] Bradley Efron. How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association*, 81(394):461–470, June 1986.

[94] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2003.

[95] Ryan J. Tibshirani and Jonathan Taylor. Degrees of Freedom in Lasso Problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

[96] Karl Pearson. On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[97] William G. Cochran. The χ^2 Test of Goodness of Fit. *The Annals of Mathematical Statistics*, 23(3):315–345, September 1952.

[98] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1):1–26, January 1979.

[99] Hans R. Kunsch. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, 17(3):1217–1241, September 1989.

[100] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

[101] C. F. J. Wu. Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics*, 14:1261–1295, 1986.

[102] R. Y. Liu. Bootstrap Procedures under some Non-I.I.D. Models. *The Annals of Statistics*, 16:1696–1708, 1988.

[103] Enno Mammen. Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics*, 21(1):255–285, March 1993.

[104] Mário A. T. Figueiredo. Adaptive Sparseness for Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–1159, 2003.

[105] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[106] James R. Faulkner and Vladimir N. Minin. Locally Adaptive Smoothing with Markov Random Fields and Shrinkage Priors. *Bayesian Analysis*, 13(1):225–252, March 2018.

[107] Nicolai Meinshausen. Relaxed Lasso. *Computational Statistics & Data Analysis*, 52(1): 374–393, 2007.

[108] François Fressin, Guillermo Torres, Jean-Michel Désert, David Charbonneau, Natalie M. Batalha, Jonathan J. Fortney, Jason F. Rowe, Christopher Allen, William J. Borucki, Timothy M. Brown, Stephen T. Bryson, David R. Ciardi, William D. Cochran, Drake

Deming, Edward W. Dunham, Daniel C. Fabrycky, Thomas N. Gautier III, Ronald L. Gilliland, Christopher E. Henze, Matthew J. Holman, Steve B. Howell, Jon M. Jenkins, Karen Kinemuchi, Heather Knutson, David G. Koch, David W. Latham, Jack J. Lissauer, Geoffrey W. Marcy, Darin Ragozzine, Dimitar D. Sasselov, Martin Still, Peter Tenenbaum, and Kamal Uddin. Kepler-10 c: a 2.2 Earth Radius Transiting Planet in a Multiple System. *The Astrophysical Journal Supplement Series*, 197(1):5, October 2011.

[109] Donald G. York, J. Adelman, Jr. John E. Anderson, Scott F. Anderson, James Annis, Neta A. Bahcall, J. A. Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, William N. Boroski, Steve Bracker, Charlie Briegel, John W. Briggs, J. Brinkmann, Robert Brunner, Scott Burles, Larry Carey, Michael A. Carr, Francisco J. Castander, Bing Chen, Patrick L. Colestock, A. J. Connolly, J. H. Crocker, István Csabai, Paul C. Czarapata, John Eric Davis, Mamoru Doi, Tom Dombeck, Daniel Eisenstein, Nancy Ellman, Brian R. Elms, Michael L. Evans, Xiaohui Fan, Glenn R. Federwitz, Larry Fiscelli, Scott Friedman, Joshua A. Frieman, Masataka Fukugita, Bruce Gillespie, James E. Gunn, Vijay K. Gurbani, Ernst de Haas, Merle Haldeman, Frederick H. Harris, J. Hayes, Timothy M. Heckman, G. S. Hennessy, Robert B. Hindsley, Scott Holm, Donald J. Holmgren, Chi hao Huang, Charles Hull, Don Husby, Shin-Ichi Ichikawa, Takashi Ichikawa, Željko Ivezić, Stephen Kent, Rita S. J. Kim, E. Kinney, Mark Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, John Korienek, Richard G. Kron, Peter Z. Kunszt, D. Q. Lamb, B. Lee, R. French Leger, Siriluk Limmongkol, Carl Lindenmeyer, Daniel C. Long, Craig Loomis, Jon Loveday, Rich Lucinio, Robert H. Lupton, Bryan MacKinnon, Edward J. Mannery, P. M. Mantsch, Bruce Margon, Peregrine McGehee, Timothy A. McKay, Avery Meiksin, Aronne Merelli, David G. Monet, Jeffrey A. Munn, Vijay K. Narayanan, Thomas Nash, Eric Neilsen, Rich Neswold, Heidi Jo Newberg, R. C. Nichol, Tom Nicinski, Mario Nonino, Norio Okada, Sadanori Okamura, Jeremiah P. Ostriker, Russell Owen, A. George Pauls, John Peoples, R. L. Peterson, Donald Petravick, Jeffrey R. Pier, Adrian Pope, Ruth Pordes, Angela Prosapio, Ron Rechenmacher, Thomas R. Quinn, Gordon T. Richards, Michael W. Richmond, Claudio H. Rivetta, Constance M. Rockosi, Kurt Ruthmansdorfer, Dale Sandford, David J. Schlegel, Donald P. Schneider, Maki Sekiguchi, Gary Sergey, Kazuhiro Shimasaku, Walter A. Siegmund, Stephen Smee, J. Allyn Smith, S. Snedden, R. Stone, Chris Stoughton, Michael A. Strauss, Christopher Stubbs, Mark SubbaRao, Alexander S. Szalay, Istvan Szapudi, Gyula P. Szokoly, Anirudda R. Thakar, Christy Tremonti, Douglas L. Tucker, Alan Uomoto, Dan Vanden Berk, Michael S. Vogeley, Patrick Waddell, Shui Wang, Masaru

Watanabe, David H. Weinberg, Brian Yanny, and Naoki Yasuda. The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3):1579–1587, September 2000.

[110] James Guillochon, Jerod Parrent, Luke Zoltan Kelley, and Raffaella Margutti. An open catalog for supernova data. *The Astrophysical Journal*, 835(1):64, January 2017.

[111] D. E. Osterbrock and G. J. Ferland. *Astrophysics of Gaseous Nebulae and Active Galactic Nuclei*. 2006.

[112] P. Marziani, J. W. Sulentic, D. Dultzin-Hacyan, M. Calvani, and M. Moles. Comparative Analysis of the High- and Low-Ionization Lines in the Broad-Line Region of Active Galactic Nuclei. *The Astrophysical Journal Supplement Series*, 104:37, May 1996.

[113] M. Rauch. The Lyman Alpha Forest in the Spectra of Quasistellar Objects. *Annual Review of Astronomy and Astrophysics*, 36(1):267–316, 1998. doi: 10.1146/annurev.astro.36.1.267. URL `https://doi.org/10.1146/annurev.astro.36.1.267`.

[114] Nickolay Y. Gnedin and Lam Hui. Probing the Universe with the Lyα forest – I. Hydrodynamics of the low-density intergalactic medium. *Monthly Notices of the Royal Astronomical Society*, 296(1):44–55, May 1998.

[115] Molly S. Peeples, David H. Weinberg, Romeel Davé, Mark A. Fardal, and Neal Katz. Pressure Support vs. Thermal Broadening in the Lyman-alpha Forest I: Effects of the Equation of State on Longitudinal Structure. *Monthly Notices of the Royal Astronomical Society*, 404(3):1281–1294, May 2010.

[116] Uffe Hellsten, Lars Hernquist, Neal Katz, and David H. Weinberg. The Observability of Metal Lines Associated with the Lyα Forest. *The Astrophysical Journal*, 499(1):172–180, May 1998.

[117] Matthew M. Pieri, Stephan Frank, David H. Weinberg, Smita Mathur, and Donald G. York. The Composite Spectrum of Strong Lyα Forest Absorbers. *The Astrophysical Journal Letters*, 724(1):L69–L73, November 2010.

[118] Nathalie Palanque-Delabrouille, Christophe Yèche, Julien Baur, Christophe Magneville, Graziano Rossi, Julien Lesgourgues, Arnaud Borde, Etienne Burtin, Jean-Marc LeGoff, James Rich, Matteo Viel, and David Weinberg. Neutrino masses and cosmology with

Lyman-alpha forest power spectrum. *Journal of Computational and Astroparticle Physics*, 2015(11):011, November 2015.

[119] Rupert A. C. Croft, David H. Weinberg, Neal Katz, and Lars Hernquist. Intergalactic Helium Absorption in Cold Dark Matter Models. *The Astrophysical Journal*, 488(2): 532–549, October 1997.

[120] W. S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

[121] Jianqing Fan and Irene Gijbels. *Local Polynomial Modeling and Its Applications*. Chapman and Hall, 1996.

[122] C. Loader. *Local Regression and Likelihood*. Springer-Verlag, 1999.

[123] Hiroaki Aihara, Carlos Allende Prieto, Deokkeun An, Scott F. Anderson, Éric Aubourg, Eduardo Balbinot, Timothy C. Beers, Andreas A. Berlind, Steven J. Bickerton, Dmitry Bizyaev, Michael R. Blanton, John J. Bochanski, Adam S. Bolton, Jo Bovy, W. N. Brandt, J. Brinkmann, Peter J. Brown, Joel R. Brownstein, Nicolas G. Busca, Heather Campbell, Michael A. Carr, Yanmei Chen, Cristina Chiappini, Johan Comparat, Natalia Connolly, Marina Cortes, Rupert A. C. Croft, Antonio J. Cuesta, Luiz N. da Costa, James R. A. Davenport, Kyle Dawson, Saurav Dhital, Anne Ealet, Garrett L. Ebelke, Edward M. Edmondson, Daniel J. Eisenstein, Stephanie Escoffier, Massimiliano Esposito, Michael L. Evans, Xiaohui Fan, Bruno Femenía Castellá, Andreu Font-Ribera, Peter M. Frinchaboy, Jian Ge, Bruce A. Gillespie, G. Gilmore, Jonay I. González Hernández, J. Richard Gott, Andrew Gould, Eva K. Grebel, James E. Gunn, Jean-Christophe Hamilton, Paul Harding, David W. Harris, Suzanne L. Hawley, Frederick R. Hearty, Shirley Ho, David W. Hogg, Jon A. Holtzman, Klaus Honscheid, Naohisa Inada, Inese I. Ivans, Linhua Jiang, Jennifer A. Johnson, Cathy Jordan, Wendell P. Jordan, Eyal A. Kazin, David Kirkby, Mark A. Klaene, G. R. Knapp, Jean-Paul Kneib, C. S. Kochanek, Lars Koesterke, Juna A. Kollmeier, Richard G. Kron, Hubert Lampeitl, Dustin Lang, Jean-Marc Le Goff, Young Sun Lee, Yen-Ting Lin, Daniel C. Long, Craig P. Loomis, Sara Lucatello, Britt Lundgren, Robert H. Lupton, Zhibo Ma, Nicholas MacDonald, Suvrath Mahadevan, Marcio A. G. Maia, Martin Makler, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Claudia Maraston, Daniel Margala, Karen L. Masters, Cameron K. McBride, Peregrine M. McGehee, Ian D. McGreer, Brice Ménard, Jordi Miralda-Escudé, Heather L. Morrison, F. Mullally, Demitri Muna, Jeffrey A. Munn, Hitoshi Murayama, Adam D. Myers, Tracy NAugustle,

Angelo Fausti Neto, Duy Cuong Nguyen, Robert C. Nichol, Robert W. O'Connell, Ricardo L. C. Ogando, Matthew D. Olmstead, Daniel J. Oravetz, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, Parul Pandey, Isabelle Pâris, Will J. Percival, Patrick Petitjean, Robert Pfaffenberger, Janine Pforr, Stefanie Phleps, Christophe Pichon, Matthew M. Pieri, Francisco Prada, Adrian M. Price-Whelan, M. Jordan Raddick, Beatriz H. F. Ramos, Céline Reylé, James Rich, Gordon T. Richards, Hans-Walter Rix, Annie C. Robin, Helio J. Rocha-Pinto, Constance M. Rockosi, Natalie A. Roe, Emmanuel Rollinde, Ashley J. Ross, Nicholas P. Ross, Bruno M. Rossetto, Ariel G. Sánchez, Conor Sayres, David J. Schlegel, Katharine J. Schlesinger, Sarah J. Schmidt, Donald P. Schneider, Erin Sheldon, Yiping Shu, Jennifer Simmerer, Audrey E. Simmons, Thirupathi Sivarani, Stephanie A. Snedden, Jennifer S. Sobeck, Matthias Steinmetz, Michael A. Strauss, Alexander S. Szalay, Masayuki Tanaka, Aniruddha R. Thakar, Daniel Thomas, Jeremy L. Tinker, Benjamin M. Tofflemire, Rita Tojeiro, Christy A. Tremonti, Jan Vandenberg, M. Vargas Magaña, Licia Verde, Nicole P. Vogt, David A. Wake, Ji Wang, Benjamin A. Weaver, David H. Weinberg, Martin White, Simon D. M. White, Brian Yanny, Naoki Yasuda, Christophe Yeche, and Idit Zehavi. The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III. *The Astrophysical Journal Supplement Series*, 193(2):29, March 2011.

[124] T.-S. Kim, M. Viel, M. G. Haehnelt, R. F. Carswell, and S. Cristiani. The power spectrum of the flux distribution in the Lyman-alpha forest of a Large sample of UVES QSO Absorption Spectra (LUQAS). *Monthly Notices of the Royal Astronomical Society*, 347(2):355–366, January 2004.

[125] Nathalie Palanque-Delabrouille, Yèche, Christophe, Borde, Arnaud, Le Goff, Jean-Marc, Rossi, Graziano, Viel, Matteo, Aubourg, Éric, Bailey, Stephen, Bautista, Julian, Blomqvist, Michael, Bolton, Adam, Bolton, James S., Busca, Nicolás G., Carithers, Bill, Croft, Rupert A. C., Dawson, Kyle S., Delubac, Timothée, Font-Ribera, Andreu, Ho, Shirley, Kirkby, David, Lee, Khee-Gan, Margala, Daniel, Miralda-Escudé, Jordi, Muna, Demitri, Myers, Adam D., Noterdaeme, Pasquier, Pâris, Isabelle, Petitjean, Patrick, Pieri, Matthew M., Rich, James, Rollinde, Emmanuel, Ross, Nicholas P., Schlegel, David J., Schneider, Donald P., Slosar, Anze, and Weinberg, David H. The one-dimensional Ly-alpha forest power spectrum from BOSS. *Astronomy & Astrophysics*, 559:A85, 2013.

[126] Jason A. Cardelli, Geoffrey C. Clayton, and John S. Mathis. The Relationship between Infrared, Optical, and Ultraviolet Extinction. *The Astrophysical Journal*, 345:245–256, October 1989.

[127] E.F. Schlafly and D.P. Finkbeiner. Measuring reddening with Sloan Digital Sky Survey stellar spectra and recalibrating SFD. *The Astrophysical Journal*, 737(2):103, 2011.

[128] Rupert A. C. Croft, David H. Weinberg, Max Pettini, Lars Hernquist, and Neal Katz. The Power Spectrum of Mass Fluctuations Measured from the Lyα Forest at Redshift z = 2.5. *The Astrophysical Journal*, 520(1):1–23, July 1999. doi: 10.1086/307438.

[129] Lam Hui, Scott Burles, Uroš Seljak, Robert E. Rutledge, Eugene Magnier, and David Tytler. On Estimating the QSO Transmission Power Spectrum. *The Astrophysical Journal*, 552(1):15–35, May 2001. doi: 10.1086/320436.

[130] F. Natali, E. Giallongo, S. Cristiani, and F. La Franca. The Optical-Ultraviolet Continuum of a Sample of QSOs. *The Astronomical Journal*, 115(2):397–404, February 1998. ISSN 0004-6256. doi: 10.1086/300211. URL `http://dx.doi.org/10.1086/300211`.

[131] Daniel E. Vanden Berk, Gordon T. Richards, Amanda Bauer, Michael A. Strauss, Donald P. Schneider, Timothy M. Heckman, Donald G. York, Patrick B. Hall, Xiaohui Fan, G. R. Knapp, Scott F. Anderson, James Annis, Neta A. Bahcall, Mariangela Bernardi, John W. Briggs, J. Brinkmann, Robert Brunner, Scott Burles, Larry Carey, Francisco J. Castander, A. J. Connolly, J. H. Crocker, István Csabai, Mamoru Doi, Douglas Finkbeiner, Scott Friedman, Joshua A. Frieman, Masataka Fukugita, James E. Gunn, G. S. Hennessy, Željko Ivezić, Stephen Kent, Peter Z. Kunszt, D. Q. Lamb, R. French Leger, Daniel C. Long, Jon Loveday, Robert H. Lupton, Avery Meiksin, Aronne Merelli, Jeffrey A. Munn, Heidi Jo Newberg, Matt Newcomb, R. C. Nichol, Russell Owen, Jeffrey R. Pier, Adrian Pope, Constance M. Rockosi, David J. Schlegel, Walter A. Siegmund, Stephen Smee, Yehuda Snir, Chris Stoughton, Christopher Stubbs, Mark SubbaRao, Alexander S. Szalay, Gyula P. Szokoly, Christy Tremonti, Alan Uomoto, Patrick Waddell, Brian Yanny, and Wei Zheng. Composite Quasar Spectra from the Sloan Digital Sky Survey. *The Astronomical Journal*, 122(2):549–564, August 2001. doi: 10.1086/321167.

[132] Shane W. Davis, Jong-Hak Woo, and Omer M. Blaes. The uv continuum of quasars: Models and sdss spectral slopes. *The Astrophysical Journal*, 668(2):682–698, October 2007. ISSN 1538-4357. doi: 10.1086/521393. URL `http://dx.doi.org/10.1086/521393`.

[133] Isabelle Pâris, P Petitjean, Emmanuel Rollinde, Eric Aubourg, N.G. Busca, Romain Charlassier, T Delubac, Hamilton Jean-Christophe, J M Le Goff, N Palanque-Delabrouille, S Peirani, Ch Pichon, Jonathan Rich, Mariana Magana, and Ch Yèche. A Principal

Component Analysis of quasar UV spectra at $z \sim 3$. *Astronomy & Astrophysics*, 530, April 2011.

[134] Khee-Gan Lee, Nao Suzuki, and David N. Spergel. Mean-flux-regulated Principal Component Analysis Continuum Fitting of Slian Digital Sky Survey Lyα Forest Spectra. *The Astronomical Journal*, 143(2):51, January 2012. ISSN 1538-3881. doi: 10.1088/0004-6256/143/2/51. URL `http://dx.doi.org/10.1088/0004-6256/143/2/51`.

[135] E. A. Nadaraya. On estimating regression. *Theory of Probability & its Applications*, 9: 141–142, 1964.

[136] Geoffrey S. Watson. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26:359–372, 1964.

[137] V.A. Epanechnikov. Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability & its Applications*, 14(1):153–158, 1967.

[138] Mariangela Bernardi, Ravi K. Sheth, Mark SubbaRao, Gordon T. Richards, Scott Burles, Andrew J. Connolly, Joshua Frieman, Robert Nichol, Joop Schaye, Donald P. Schneider, Daniel E. Vanden Berk, Donald G. York, J. Brinkmann, and Don Q. Lamb. A Feature at $z \sim 3.2$ in the Evolution of the Lyα Forest Optical Depth. *The Astronomical Journal*, 125 (1):32–52, January 2003.

[139] Claude-André Faucher-Giguère, Adam Lidz, Lars Hernquist, and Matias Zaldarriaga. Evolution of the Intergalactic Opacity: Implications for the Ionizing Background, Cosmic Star Formation, and Quasar Activity. *The Astrophysical Journal*, 688(1):85–107, November 2008.

[140] Aldo Dall'Aglio, Lutz Wisotzki, and Gabor Worseck. The UV background photoionization rate at $2.3 < z < 4.6$ as measured from the Sloan Digital Sky Survey, 2009.

[141] George D. Becker, Paul C. Hewett, Gabor Worseck, and J. Xavier Prochaska. A Refined Measurement of the Mean Transmitted Flux in the Ly-alpha Forest over $2 < z < 5$ Using Composite Quasar Spectra. *Monthly Notices of the Royal Astronomical Society*, 430: 2067–2081, 2013.

[142] Natalie M. Batalha, William J. Borucki, Stephen T. Bryson, Lars A. Buchhave, Douglas A. Caldwell, Jørgen Christensen-Dalsgaard, David Ciardi, Edward W. Dunham, Francois Fressin, Thomas N. Gautier, Ronald L. Gilliland, Michael R. Haas, Steve B. Howell, Jon M.

Jenkins, Hans Kjeldsen, David G. Koch, David W. Latham, Jack J. Lissauer, Geoffrey W. Marcy, Jason F. Rowe, Dimitar D. Sasselov, Sara Seager, Jason H. Steffen, Guillermo Torres, Gibor S. Basri, Timothy M. Brown, David Charbonneau, Jessie Christiansen, Bruce Clarke, William D. Cochran, Andrea Dupree, Daniel C. Fabrycky, Debra Fischer, Eric B. Ford, Jonathan Fortney, Forrest R. Girouard, Matthew J. Holman, John Johnson, Howard Isaacson, Todd C. Klaus, Pavel Machalek, Althea V. Moorehead, Robert C. Morehead, Darin Ragozzine, Peter Tenenbaum, Joseph Twicken, Samuel Quinn, Jeffrey VanCleve, Lucianne M. Walkowicz, William F. Welsh, Edna Devore, and Alan Gould. Kepler's First Rocky Planet: Kepler-10b. *The Astrophysical Journal*, 729(1):27, February 2011.

[143] Jon M. Jenkins, Douglas A. Caldwell, Hema Chandrasekaran, Joseph D. Twicken, Stephen T. Bryson, Elisa V. Quintana, Bruce D. Clarke, Jie Li, Christopher Allen, Peter Tenenbaum, Hayley Wu, Todd C. Klaus, Christopher K. Middour, Miles T. Cote, Sean McCauliff, Forrest R. Girouard, Jay P. Gunter, Bill Wohler, Jeneen Sommers, Jennifer R. Hall, AKM K. Uddin, Michael S. Wu, Paresh A. Bhavsar, Jeffrey Van Cleve, David L. Pletcher, Jessie A. Dotson, Michael R. Haas, Ronald L. Gilliland, David G. Koch, and William J. Borucki. Overview of the Kepler Science Processing Pipeline. *The Astrophysical Journal Letters*, 713(2):87–91, March 2010.

[144] R. L. Akeson, X. Chen, D. Ciardi, M. Crane, J. Good, M. Harbut, E. Jackson, S. R. Kane, A. C. Laity, S. Leifer, and et al. The NASA Exoplanet Archive: Data and Tools for Exoplanet Research. *Publications of the Astronomical Society of the Pacific*, 125(930): 989–999, August 2013.

[145] Kaisey Mandel and Eric Agol. Analytic Light Curves for Planetary Transit Searches. *The Astrophysical Journal Letters*, 580(2):171–175, December 2002.

[146] Robert W. Slawson, Andrej Prša, William F. Welsh, Jerome A. Orosz, Michael Rucker, Natalie Batalha, Laurance R. Doyle, Scott G. Engle, Kyle Conroy, Jared Coughlin, and et al. Kepler Eclipsing Binary Stars. II. 2165 Eclipsing Binaries in the Second Data Release. *The Astronomical Journal*, 142(5):160, October 2011.

[147] Gal Matijevič, Andrej Prša, Jerome A. Orosz, William F. Welsh, Steven Bloemen, and Thomas Barclay. Keplereclipsing binary stars. iii. classification of kepler eclipsing binary light curves with locally linear embedding. *The Astronomical Journal*, 143(5):123, April 2012.

[148] T. W.-S. Holoien, K. Z. Stanek, J. S. Brown, C. S. Kochanek, D. Godoy-Rivera, U. Basu, B. J. Shappee, J. L. Prieto, D. Bersier, S. Dong, P. Chen, and J. Brimacombe. ASASSN-16fp: Discovery of A Probable Supernova in UGC 11868. *The Astronomer's Telegram*, 9086, May 2016.

[149] P. J. Brown, A. A. Breeveld, S. Holland, P. Kuin, and T. Pritchard. SOUSA: the Swift Optical/Ultraviolet Supernova Archive. *Astrophysics and Space Science*, 354:89–96, November 2014.

[150] S. J. Prentice, C. Ashall, P. A. Mazzali, J.-J. Zhang, P. A. James, X.-F. Wang, J. Vinkó, S. Percival, L. Short, A. Piascik, F. Huang, J. Mo, L.-M. Rui, J.-G. Wang, D.-F. Xiang, Y.-X. Xin, W.-M. Yi, X.-G. Yu, Q. Zhai, T.-M. Zhang, G. Hosseinzadeh, D. A. Howell, C. McCully, S. Valenti, B. Cseh, O. Hanyecz, L. Kriskovics, A. Pál, K. Sárneczky, Á. Sódor, R. Szakáts, P. Székely, E. Varga-Verebélyi, K. Vida, M. Bradac, D. E. Reichart, D. Sand, and L. Tartaglia. SN 2016coi/ASASSN-16fp: An example of residual helium in a type Ic supernova? *Monthly Notices of the Royal Astronomical Society*, 478:4162–4192, August 2018.

[151] M. M. Phillips. The Absolute Magnitudes of Type IA Supernovae. *The Astrophysical Journal Letters*, 413:L105, August 1993.

[152] Christopher R. Burns, Maximilian Stritzinger, M. M. Phillips, ShiAnne Kattner, S. E. Persson, Barry F. Madore, Wendy L. Freedman, Luis Boldt, Abdo Campillay, Carlos Contreras, Gaston Folatelli, Sergio Gonzalez, Wojtek Krzeminski, Nidia Morrell, Francisco Salgado, and Nicholas B. Suntzeff. The Carnegie Supernova Project: Light Curve Fitting with SNooPy. *The Astronomical Journal*, 141(1):19, December 2010.

[153] C. R. Burns, M. Stritzinger, M. M. Phillips, E. Y. Hsiao, C. Contreras, S. E. Persson, G. Folatelli, L. Boldt, A. Campillay, S. Castellón, W. L. Freedman, B. F. Madore, N. Morrell, F. Salgado, and N. B. Suntzeff. The Carnegie Supernova Project: Intrinsic Colors of Type Ia Supernovae. *The Astrophysical Journal*, 789:32, July 2014.

[154] Vasily Belokurov, N. Wyn Evans, and Yann Le Du. Light-curve classification in massive variability surveys – II. Transients towards the Large Magellanic Cloud. *Monthly Notices of the Royal Astronomical Society*, 352(1):233–242, 2004.

[155] B. J. Shappee, J. L. Prieto, D. Grupe, C. S. Kochanek, K. Z. Stanek, G. De Rosa, S. Mathur, Y. Zu, B. M. Peterson, R. W. Pogge, S. Komossa, M. Im, J. Jencson, T.W-S. Holoien,

U. Basu, J. F. Beacom, D. M. Szczygieł, J. Brimacombe, S. Adams, A. Campillay, C. Choi, C. Contreras, M. Dietrich, M. Dubberley, M. Elphick, S. Foale, M. Giustini, C. Gonzalez, E. Hawkins, D. A. Howell, E. Y. Hsiao, M. Koss, K. M. Leighly, N. Morrell, D. Mudd, D. Mullins, J. M. Nugent, J. Parrent, M. M. Phillips, G. Pojmanski, W. Rosing, R. Ross, D. Sand, D. M. Terndrup, S. Valenti, Z. Walker, and Y. Yoon. The Man Behind the Curtain: X-rays Drive the UV through NIR Variability in the 2013 AGN Outburst in NGC 2617. *The Astrophysical Journal*, 788(1):48, May 2014.

[156] N. Elias-Rosa, S. Mattila, P. Lundqvist, M. Stritzinger, H. Kuncarayakti, J. Harmanen, A. Pastorello, S. Benetti, E. Cappellaro, N. Blagorodnova, S. Davis, S. Dong, M. Fraser, C. Gall, D. Harrison, S. Hodgkin, E. Y. Hsiao, P. Jonker, T. Kangas, E. Kankare, Z. Kostrzewa-Rutkowska, M. Nielsen, P. Ochner, J. L. Prieto, T. Reynolds, C. Romero-Canizales, F. Taddia, L. Tartaglia, G. Terreran, L. Tomasella, and L. Wyrzykowski. Spectroscopic classification of ASASSN-16fp with the Nordic Optical Telescope. *The Astronomer's Telegram*, 9090:1, May 2016.

[157] Brajesh Kumar, A. Singh, S. Srivastav, D. K. Sahu, and G. C. Anupama. ASASSN-16fp (SN 2016coi): a transitional supernova between Type Ic and broad-lined Ic. *Monthly Notices of the Royal Astronomical Society*, 473(3):3776–3788, January 2018.

[158] G. Terreran, R. Margutti, D. Bersier, J. Brimacombe, D. Caprioli, P. Challis, R. Chornock, D. L. Coppejans, Subo Dong, C. Guidorzi, K. Hurley, R. Kirshner, G. Migliori, D. Milisavljevic, D. M. Palmer, J. L. Prieto, L. Tomasella, P. Marchant, A. Pastorello, B. J. Shappee, K. Z. Stanek, M. D. Stritzinger, S. Benetti, Ping Chen, L. DeMarchi, N. Elias-Rosa, C. Gall, J. Harmanen, and S. Mattila. SN 2016coi (ASASSN-16fp): An Energetic H-stripped Core-collapse Supernova from a Massive Stellar Progenitor with Large Mass Loss. *The Astrophysical Journal*, 883(2):147, October 2019. doi: 10.3847/1538-4357/ab3e37.

[159] Collin A. Politsch and Rupert A.C. Croft. Mapping the Large-Scale Universe through Intergalactic Silhouettes. *CHANCE*, 32(3):14–19, 2019. doi: 10.1080/09332480.2019.1662696. URL `https://doi.org/10.1080/09332480.2019.1662696`.

[160] Matthew McQuinn. The Evolution of the Intergalactic Medium. *Annual Review of Astronomy and Astrophysics*, 54(1):313–362, September 2016. ISSN 1545-4282. doi: 10.1146/annurev-astro-082214-122355. URL `http://dx.doi.org/10.1146/annurev-astro-082214-122355`.

[161] Gregory A. Shields. A Brief History of Active Galactic Nuclei. *Publications of the Astronomical Society of the Pacific*, 111(760):661–678, June 1999. ISSN 1538-3873. doi: 10.1086/316378. URL `http://dx.doi.org/10.1086/316378`.

[162] James E. Gunn and Bruce A. Peterson. On the Density of Neutral Hydrogen in Intergalactic Space. *The Astrophysical Journal*, 142:1633–1636, November 1965. doi: 10.1086/148444.

[163] R. Lynds. The Absorption-Line Spectrum of 4c 05.34. *The Astrophysical Journal Letters*, 164:L73, March 1971. doi: 10.1086/180695.

[164] David H. Weinberg. The Lyman-α Forest as a Cosmological Tool. *AIP Conference Proceedings*, 2003. ISSN 0094-243X. doi: 10.1063/1.1581786. URL `http://dx.doi.org/10.1063/1.1581786`.

[165] S. S. Vogt, S. L. Allen, B. C. Bigelow, L. Bresee, B. Brown, T. Cantrall, A. Conrad, M. Couture, C. Delaney, H. W. Epps, D. Hilyard, D. F. Hilyard, E. Horn, N. Jern, D. Kanto, M. J. Keane, R. I. Kibrick, J. W. Lewis, J. Osborne, G. H. Pardeilhan, T. Pfister, T. Ricketts, L. B. Robinson, R. J. Stover, D. Tucker, J. Ward, and M. Z. Wei. *HIRES: the high-resolution echelle spectrometer on the Keck 10-m Telescope*, volume 2198 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 362. 1994. doi: 10.1117/12.176725.

[166] Steven S. Vogt. *An Overview of Science Results from HIRES: the First 6 Years*, volume 270 of *Astronomical Society of the Pacific Conference Series*, page 5. 2002.

[167] Nao Suzuki, David Tytler, David Kirkman, John M. O'Meara, and Dan Lubin. Relative Flux Calibration of Keck HIRES Echelle Spectra. *Publications of the Astronomical Society of the Pacific*, 115(811):1050–1067, September 2003. doi: 10.1086/376849.

[168] Scott Burles and David Tytler. The Neutral Hydrogen Column Density Towards Q1937-1009 From the Unabsorbed Intrinsic Continuum in the Lyman-alpha Forest. *The Astronomical Journal*, 114:1330, October 1997. doi: 10.1086/118566.

[169] Mattia Ciollaro, Jessi Cisewski, Peter Freeman, Christopher Genovese, Jing Lei, Ross O'Connell, and Larry Wasserman. Functional Regression for Quasar Spectra, 2014.

[170] Planck Collaboration, Y. Akrami, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R.

Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, J. Carron, B. Casaponsa, A. Challinor, H. C. Chiang, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J. M. Delouis, F. X. Desert, E. Di Valentino, C. Dickinson, J. M. Diego, S. Donzelli, O. Dore, M. Douspis, A. Ducout, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enblin, H. K. Eriksen, E. Falgarone, Y. Fantaye, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Genova-Santos, M. Gerbino, T. Ghosh, J. Gonzalez-Nuevo, K. M. Gorski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, G. Helou, D. Herranz, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihanen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J. M. Lamarre, M. Langer, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, J. P. Leahy, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. Lopez-Caniego, P. M. Lubin, Y. Z. Ma, J. F. Macias-Perez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, E. Martinez-Gonzalez, S. Matarrese, N. Mauri, J. D. McEwen, P. D. Meerburg, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschenes, D. Molinari, A. Moneti, L. Montier, G. Morgante, A. Moss, S. Mottet, M. Munchmeyer, P. Natoli, H. U. Norgaard-Nielsen, C. A. Oxborrow, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, T. J. Pearson, M. Peel, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J. L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubino-Martin, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, M. Shiraishi, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A. S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Terenzi, L. Toffolatti, M. Tomasi, T. Trombetti, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca. Planck 2018 results. I. Overview and the cosmological legacy of Planck, 2018.

[171] R. Cen, J. Miralda-Escudé, J. P. Ostriker, and M. Rauch. Gravitational Collapse of Small-Scale Structure as the Origin of the Lyman-Alpha Forest. *The Astrophysical Journal Letters*, 437:L9, December 1994.

[172] Renyue Cen and Jeremiah P. Ostriker. Where Are the Baryons? *The Astrophysical Journal*, 514(1):1–6, March 1999. doi: 10.1086/306949.

[173] Renyue Cen and Jeremiah P. Ostriker. Where Are the Baryons? II. Feedback Effects. *The Astrophysical Journal*, 650(2):560–572, October 2006. doi: 10.1086/506505.

[174] Renyue Cen and Taotao Fang. Where Are the Baryons? III. Nonequilibrium Effects and Observables. *The Astrophysical Journal*, 650(2):573–591, October 2006. doi: 10.1086/506506.

[175] A. A. Penzias and R. W. Wilson. A Measurement of Excess Antenna Temperature at 4080 Mc/s. *The Astrophysical Journal*, 142:419–421, July 1965. doi: 10.1086/148307.

[176] R. H. Dicke, P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson. Cosmic Black-Body Radiation. *The Astrophysical Journal*, 142:414–419, July 1965. doi: 10.1086/148306.

[177] C. L. Bennett, A. J. Banday, K. M. Gorski, G. Hinshaw, P. Jackson, P. Keegstra, A. Kogut, G. F. Smoot, D. T. Wilkinson, and E. L. Wright. Four-Year COBE DMR Cosmic Microwave Background Observations: Maps and Basic Results. *The Astrophysical Journal Letters*, 464:L1, June 1996. doi: 10.1086/310075.

[178] C. L. Bennett, D. Larson, J. L. Weiland, N. Jarosik, G. Hinshaw, N. Odegard, K. M. Smith, R. S. Hill, B. Gold, M. Halpern, E. Komatsu, M. R. Nolta, L. Page, D. N. Spergel, E. Wollack, J. Dunkley, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, and E. L. Wright. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results. *The Astrophysical Journal Supplement Series*, 208(2):20, October 2013. doi: 10.1088/0067-0049/208/2/20.

[179] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. I. R. Alves, C. Armitage-Caplan, M. Arnaud, M. Ashdown, F. Atrio-Barandela, J. Aumont, H. Aussel, C. Baccigalupi, A. J. Banday, R. B. Barreiro, R. Barrena, M. Bartelmann, J. G. Bartlett, N. Bartolo, S. Basak, E. Battaner, R. Battye, K. Benabed, A. Benoît, A. Benoit-Lévy, J. P. Bernard, M. Bersanelli, B. Bertincourt, M. Bethermin, P. Bielewicz, I. Bikmaev, A. Blanchard, J. Bobin, J. J. Bock, H. Böhringer, A. Bonaldi, L. Bonavera, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, H. Bourdin, J. W. Bowyer, M. Bridges, M. L. Brown, M. Bucher, R. Burenin, C. Burigana, R. C. Butler, E. Calabrese, B. Cappellini, J. F. Cardoso, R. Carr, P. Carvalho, M. Casale, G. Castex, A. Catalano, A. Challinor, A. Chamballu, R. R. Chary, X. Chen, H. C. Chiang, L. Y. Chiang, G. Chon, P. R. Christensen, E. Churazov, S. Church, M. Clemens, D. L. Clements, S. Colombi, L. P. L. Colombo, C. Combet, B. Comis, F. Couchot, A. Coulais, B. P. Crill, M. Cruz, A. Curto, F. Cuttaia, A. Da

Silva, H. Dahle, L. Danese, R. D. Davies, R. J. Davis, P. de Bernardis, A. de Rosa, G. de Zotti, T. Déchelette, J. Delabrouille, J. M. Delouis, J. Démoclès, F. X. Désert, J. Dick, C. Dickinson, J. M. Diego, K. Dolag, H. Dole, S. Donzelli, O. Doré, M. Douspis, A. Ducout, J. Dunkley, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, O. Fabre, E. Falgarone, M. C. Falvella, Y. Fantaye, J. Fergusson, C. Filliard, F. Finelli, I. Flores-Cacho, S. Foley, O. Forni, P. Fosalba, M. Frailis, A. A. Fraisse, E. Franceschi, M. Freschi, S. Fromenteau, M. Frommert, T. C. Gaier, S. Galeotta, J. Gallegos, S. Galli, B. Gandolfo, K. Ganga, C. Gauthier, R. T. Génova-Santos, T. Ghosh, M. Giard, G. Giardino, M. Gilfanov, D. Girard, Y. Giraud-Héraud, E. Gjerløw, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gregorio, A. Gruppuso, J. E. Gudmundsson, J. Haissinski, J. Hamann, F. K. Hansen, M. Hansen, D. Hanson, D. L. Harrison, A. Heavens, G. Helou, A. Hempel, S. Henrot-Versillé, C. Hernández-Monteagudo, D. Herranz, S. R. Hildebrand t, E. Hivon, S. Ho, M. Hobson, W. A. Holmes, A. Hornstrup, Z. Hou, W. Hovest, G. Huey, K. M. Huffenberger, G. Hurier, S. Ilić, A. H. Jaffe, T. R. Jaffe, J. Jasche, J. Jewell, W. C. Jones, M. Juvela, P. Kalberla, P. Kangaslahti, E. Keihänen, J. Kerp, R. Keskitalo, I. Khamitov, K. Kiiveri, J. Kim, T. S. Kisner, R. Kneissl, J. Knoche, L. Knox, M. Kunz, H. Kurki-Suonio, F. Lacasa, G. Lagache, A. Lähteenmäki, J. M. Lamarre, M. Langer, A. Lasenby, M. Lattanzi, R. J. Laureijs, A. Lavabre, C. R. Lawrence, M. Le Jeune, S. Leach, J. P. Leahy, R. Leonardi, J. León-Tavares, C. Leroy, J. Lesgourgues, A. Lewis, C. Li, A. Liddle, M. Liguori, P. B. Lilje, M. Linden-Vørnle, V. Lindholm, M. López-Caniego, S. Lowe, P. M. Lubin, J. F. Macías-Pérez, C. J. MacTavish, B. Maffei, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, D. Marinucci, M. Maris, F. Marleau, D. J. Marshall, P. G. Martin, E. Martínez-González, S. Masi, M. Massardi, S. Matarrese, T. Matsumura, F. Matthai, L. Maurin, P. Mazzotta, A. McDonald, J. D. McEwen, P. McGehee, S. Mei, P. R. Meinhold, A. Melchiorri, J. B. Melin, L. Mendes, E. Menegoni, A. Mennella, M. Migliaccio, K. Mikkelsen, M. Millea, R. Miniscalco, S. Mitra, M. A. Miville-Deschênes, D. Molinari, A. Moneti, L. Montier, G. Morgante, N. Morisset, D. Mortlock, A. Moss, D. Munshi, J. A. Murphy, P. Naselsky, F. Nati, P. Natoli, M. Negrello, N. P. H. Nesvadba, C. B. Netterfield, H. U. Nørgaard-Nielsen, C. North, F. Noviello, D. Novikov, I. Novikov, I. J. O'Dwyer, F. Orieux, S. Osborne, C. O'Sullivan, C. A. Oxborrow, F. Paci, L. Pagano, F. Pajot, R. Paladini, S. Pand olfi, D. Paoletti, B. Partridge, F. Pasian, G. Patanchon, P. Paykari, D. Pearson, T. J. Pearson, M. Peel, H. V. Peiris, O. Perdereau, L. Perotto, F. Perrotta, V. Pettorino, F. Piacentini, M. Piat, E. Pierpaoli, D. Pietrobon, S. Plaszczynski, P. Platania, D. Pogosyan, E. Pointecouteau, G. Polenta, N. Ponthieu, L. Popa, T. Poutanen,

G. W. Pratt, G. Prézeau, S. Prunet, J. L. Puget, A. R. Pullen, J. P. Rachen, B. Racine, A. Rahlin, C. Räth, W. T. Reach, R. Rebolo, M. Reinecke, M. Remazeilles, C. Renault, A. Renzi, A. Riazuelo, S. Ricciardi, T. Riller, C. Ringeval, I. Ristorcelli, G. Robbers, G. Rocha, M. Roman, C. Rosset, M. Rossetti, G. Roudier, M. Rowan-Robinson, J. A. Rubiño-Martín, B. Ruiz-Granados, B. Rusholme, E. Salerno, M. Sandri, L. Sanselme, D. Santos, M. Savelainen, G. Savini, B. M. Schaefer, F. Schiavon, D. Scott, M. D. Seiffert, P. Serra, E. P. S. Shellard, K. Smith, G. F. Smoot, T. Souradeep, L. D. Spencer, J. L. Starck, V. Stolyarov, R. Stompor, R. Sudiwala, R. Sunyaev, F. Sureau, P. Sutter, D. Sutton, A. S. Suur-Uski, J. F. Sygnet, J. A. Tauber, D. Tavagnacco, D. Taylor, L. Terenzi, D. Texier, L. Toffolatti, M. Tomasi, J. P. Torre, M. Tristram, M. Tucci, J. Tuovinen, M. Türler, M. Tuttlebee, G. Umana, L. Valenziano, J. Valiviita, B. Van Tent, J. Varis, L. Vibert, M. Viel, P. Vielva, F. Villa, N. Vittorio, L. A. Wade, B. D. Wandelt, C. Watson, R. Watson, I. K. Wehus, N. Welikala, J. Weller, M. White, S. D. M. White, A. Wilkinson, B. Winkel, J. Q. Xia, D. Yvon, A. Zacchei, J. P. Zibin, and A. Zonca. Planck 2013 results. I. Overview of products and scientific results. *Astronomy & Astrophysics*, 571:A1, November 2014. doi: 10.1051/0004-6361/201321529.

[180] Planck Collaboration, Y. Akrami, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, K. Benabed, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. R. Bond, J. Borrill, F. R. Bouchet, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, B. Casaponsa, A. Challinor, H. C. Chiang, L. P. L. Colombo, C. Combet, B. P. Crill, F. Cuttaia, P. de Bernardis, A. de Rosa, G. de Zotti, J. Delabrouille, J. M. Delouis, E. Di Valentino, J. M. Diego, O. Dore, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enblin, H. K. Eriksen, Y. Fantaye, J. Fergusson, R. Fernandez-Cobos, F. Finelli, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, K. Ganga, R. T. Genova-Santos, M. Gerbino, J. Gonzalez-Nuevo, K. M. Gorski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, G. Jung, E. Keihanen, R. Keskitalo, K. Kiiveri, J. Kim, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, J. M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, V. Lindholm, M. Lopez-Caniego, Y. Z. Ma, J. F. Macias-Perez, G. Maggio, D. Maino, N. Mandolesi, A. Marcos-Caballero, M. Maris, P. G. Martin, E. Martinez-Gonzalez, S. Matarrese, N. Mauri, J. D. McEwen, P. D. Meerburg, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. A. Miville-Deschenes,

D. Molinari, A. Moneti, L. Montier, G. Morgante, A. Moss, M. Munchmeyer, P. Natoli, F. Oppizzi, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, F. Perrotta, V. Pettorino, F. Piacentini, G. Polenta, J. L. Puget, J. P. Rachen, B. Racine, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, J. A. Rubino-Martin, B. Ruiz-Granados, L. Salvati, M. Savelainen, D. Scott, E. P. S. Shellard, M. Shiraishi, C. Sirignano, G. Sirri, K. Smith, L. D. Spencer, L. Stanco, R. Sunyaev, A. S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, J. Valiviita, B. Van Tent, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, A. Zacchei, and A. Zonca. Planck 2018 results. IX. Constraints on primordial non-Gaussianity, 2019.

[181] Albert Einstein. The Foundation of the General Theory of Relativity. *Annalen Phys.*, 49(7):769–822, 1916. doi: 10.1002/andp.200590044,10.1002/andp.19163540702. [Annalen Phys.14,517(2005); Annalen Phys.354,no.7,769(1916)].

[182] Scott Dodelson and Fabian Schmidt. *Modern cosmology*. Academic press, 2020.

[183] R. A. C. Croft, D. H. Weinberg, N. Katz, and Lars H. Recovery of the Power Spectrum of Mass Fluctuations from Observations of the Lyα Forest. *The Astrophysical Journal*, 495 (1):44–62, March 1998. doi: 10.1086/305289.

[184] S. Peirani, D. H. Weinberg, S. Colombi, J. Blaizot, Y. Dubois, and C. Pichon. LyMAS: Predicting Large-scale Lyα Forest Statistics from the Dark Matter Density Field. *The Astrophysical Journal*, 784(1):11, Feb. 2014.

[185] R.A. Overzier. The realm of the galaxy protoclusters. *The Astronomy and Astrophysics Review*, 24(1):14, 2016. doi: 10.1007/s00159-016-0100-3.

[186] R. A. C. Croft, A. J. Banday, and L. Hernquist. Lyman-alpha forest-CMB cross-correlation and the search for the ionized baryons at high redshift. *Monthly Notices of the Royal Astronomical Society*, 369(3):1090?1102, July 2006. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2006.10292.x. URL `http://dx.doi.org/10.1111/j.1365-2966.2006.10292.x`.

[187] Michael R. Blanton, Matthew A. Bershady, Bela Abolfathi, Franco D. Albareti, Carlos Allende Prieto, Andres Almeida, Javier Alonso-García, Friedrich Anders, Scott F. Anderson, Brett Andrews, Erik Aquino-Ortíz, Alfonso Aragón-Salamanca, Maria Argudo-Fernández, Eric Armengaud, Eric Aubourg, Vladimir Avila-Reese, Carles Badenes, Stephen Bailey,

Kathleen A. Barger, Jorge Barrera-Ballesteros, Curtis Bartosz, Dominic Bates, Falk Baumgarten, Julian Bautista, Rachael Beaton, Timothy C. Beers, Francesco Belfiore, Chad F. Bender, Andreas A. Berlind, Mariangela Bernardi, Florian Beutler, Jonathan C. Bird, Dmitry Bizyaev, Guillermo A. Blanc, Michael Blomqvist, Adam S. Bolton, Médéric Boquien, Jura Borissova, Remco van den Bosch, Jo Bovy, William N. Brandt, Jonathan Brinkmann, Joel R. Brownstein, Kevin Bundy, Adam J. Burgasser, Etienne Burtin, Nicolás G. Busca, Michele Cappellari, Maria Leticia Delgado Carigi, Joleen K. Carlberg, Aurelio Carnero Rosell, Ricardo Carrera, Nancy J. Chanover, Brian Cherinka, Edmond Cheung, Yilen Gómez Maqueo Chew, Cristina Chiappini, Peter Doohyun Choi, Drew Chojnowski, Chia-Hsun Chuang, Haeun Chung, Rafael Fernando Cirolini, Nicolas Clerc, Roger E. Cohen, Johan Comparat, Luiz da Costa, Marie-Claude Cousinou, Kevin Covey, Jeffrey D. Crane, Rupert A. C. Croft, Irene Cruz-Gonzalez, Daniel Garrido Cuadra, Katia Cunha, Guillermo J. Damke, Jeremy Darling, Roger Davies, Kyle Dawson, Axel de la Macorra, Flavia Dell'Agli, Nathan De Lee, Timothée Delubac, Francesco Di Mille, Aleks Diamond-Stanic, Mariana Cano-Díaz, John Donor, Juan José Downes, Niv Drory, Hélion du Mas des Bourboux, Christopher J. Duckworth, Tom Dwelly, Jamie Dyer, Garrett Ebelke, Arthur D. Eigenbrot, Daniel J. Eisenstein, Eric Emsellem, Mike Eracleous, Stephanie Escoffier, Michael L. Evans, Xiaohui Fan, Emma Fernández-Alvar, J. G. Fernandez-Trincado, Diane K. Feuillet, Alexis Finoguenov, Scott W. Fleming, Andreu Font-Ribera, Alexander Fredrickson, Gordon Freischlad, Peter M. Frinchaboy, Carla E. Fuentes, Lluís Galbany, R. Garcia-Dias, D. A. García-Hernández, Patrick Gaulme, Doug Geisler, Joseph D. Gelfand, Héctor Gil-Marín, Bruce A. Gillespie, Daniel Goddard, Violeta Gonzalez-Perez, Kathleen Grabowski, Paul J. Green, Catherine J. Grier, James E. Gunn, Hong Guo, Julien Guy, Alex Hagen, ChangHoon Hahn, Matthew Hall, Paul Harding, Sten Hasselquist, Suzanne L. Hawley, Fred Hearty, Jonay I. Gonzalez Hernández, Shirley Ho, David W. Hogg, Kelly Holley-Bockelmann, Jon A. Holtzman, Parker H. Holzer, Joseph Huehnerhoff, Timothy A. Hutchinson, Ho Seong Hwang, Héctor J. Ibarra-Medel, Gabriele da Silva Ilha, Inese I. Ivans, KeShawn Ivory, Kelly Jackson, Trey W. Jensen, Jennifer A. Johnson, Amy Jones, Henrik Jönsson, Eric Jullo, Vikrant Kamble, Karen Kinemuchi, David Kirkby, Francisco-Shu Kitaura, Mark Klaene, Gillian R. Knapp, Jean-Paul Kneib, Juna A. Kollmeier, Ivan Lacerna, Richard R. Lane, Dustin Lang, David R. Law, Daniel Lazarz, Youngbae Lee, Jean-Marc Le Goff, Fu-Heng Liang, Cheng Li, Hongyu Li, Jianhui Lian, Marcos Lima, Lihwai Lin, Yen-Ting Lin, Sara Bertran de Lis, Chao Liu, Miguel Angel C. de Icaza Lizaola, Dan Long, Sara Lucatello, Britt Lundgren, Nicholas K. MacDonald, Alice Deconto Machado, Chelsea L. MacLeod,

Suvrath Mahadevan, Marcio Antonio Geimba Maia, Roberto Maiolino, Steven R. Majewski, Elena Malanushenko, Viktor Malanushenko, Arturo Manchado, Shude Mao, Claudia Maraston, Rui Marques-Chaves, Thomas Masseron, Karen L. Masters, Cameron K. McBride, Richard M. McDermid, Brianne McGrath, Ian D. McGreer, Nicolás Medina Peña, Matthew Melendez, Andrea Merloni, Michael R. Merrifield, Szabolcs Meszaros, Andres Meza, Ivan Minchev, Dante Minniti, Takamitsu Miyaji, Surhud More, John Mulchaey, Francisco Müller-Sánchez, Demitri Muna, Ricardo R. Munoz, Adam D. Myers, Preethi Nair, Kirpal Nandra, Janaina Correa do Nascimento, Alenka Negrete, Melissa Ness, Jeffrey A. Newman, Robert C. Nichol, David L. Nidever, Christian Nitschelm, Pierros Ntelis, Julia E. O'Connell, Ryan J. Oelkers, Audrey Oravetz, Daniel Oravetz, Zach Pace, Nelson Padilla, Nathalie Palanque-Delabrouille, Pedro Alonso Palicio, Kaike Pan, John K. Parejko, Taniya Parikh, Isabelle Pâris, Changbom Park, Alim Y. Patten, Sebastien Peirani, Marcos Pellejero-Ibanez, Samantha Penny, Will J. Percival, Ismael Perez-Fournon, Patrick Petitjean, Matthew M. Pieri, Marc Pinsonneault, Alice Pisani, Radosław Poleski, Francisco Prada, Abhishek Prakash, Anna Bárbara de Andrade Queiroz, M. Jordan Raddick, Anand Raichoor, Sand ro Barboza Rembold, Hannah Richstein, Rogemar A. Riffel, Rogério Riffel, Hans-Walter Rix, Annie C. Robin, Constance M. Rockosi, Sergio Rodríguez-Torres, A. Roman-Lopes, Carlos Román-Zúñiga, Margarita Rosado, Ashley J. Ross, Graziano Rossi, John Ruan, Rossana Ruggeri, Eli S. Rykoff, Salvador Salazar-Albornoz, Mara Salvato, Ariel G. Sánchez, D. S. Aguado, José R. Sánchez-Gallego, Felipe A. Santana, Basílio Xavier Santiago, Conor Sayres, Ricardo P. Schiavon, Jaderson da Silva Schimoia, Edward F. Schlafly, David J. Schlegel, Donald P. Schneider, Mathias Schultheis, William J. Schuster, Axel Schwope, Hee-Jong Seo, Zhengyi Shao, Shiyin Shen, Matthew Shetrone, Michael Shull, Joshua D. Simon, Danielle Skinner, M. F. Skrutskie, Anže Slosar, Verne V. Smith, Jennifer S. Sobeck, Flavia Sobreira, Garrett Somers, Diogo Souto, David V. Stark, Keivan Stassun, Fritz Stauffer, Matthias Steinmetz, Thaisa Storchi-Bergmann, Alina Streblyanska, Guy S. Stringfellow, Genaro Suárez, Jing Sun, Nao Suzuki, Laszlo Szigeti, Manuchehr Taghizadeh-Popp, Baitian Tang, Charling Tao, Jamie Tayar, Mita Tembe, Johanna Teske, Aniruddha R. Thakar, Daniel Thomas, Benjamin A. Thompson, Jeremy L. Tinker, Patricia Tissera, Rita Tojeiro, Hector Hernandez Toledo, Sylvain de la Torre, Christy Tremonti, Nicholas W. Troup, Octavio Valenzuela, Inma Martinez Valpuesta, Jaime Vargas-González, Mariana Vargas-Magaña, Jose Alberto Vazquez, Sandro Villanova, M. Vivek, Nicole Vogt, David Wake, Rene Walterbos, Yuting Wang, Benjamin Alan Weaver, Anne-Marie Weijmans, David H. Weinberg, Kyle B. Westfall, David G. Whelan, Vivienne Wild, John Wilson, W. M. Wood-Vasey,

Dominika Wylezalek, Ting Xiao, Renbin Yan, Meng Yang, Jason E. Ybarra, Christophe Yèche, Nadia Zakamska, Olga Zamora, Pauline Zarrouk, Gail Zasowski, Kai Zhang, Gong-Bo Zhao, Zheng Zheng, Zheng Zheng, Xu Zhou, Zhi-Min Zhou, Guangtun B. Zhu, Manuela Zoccali, and Hu Zou. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. *The Astronomical Journal*, 154(1):28, July 2017. doi: 10.3847/1538-3881/aa7567.

[188] Bela Abolfathi, D. S. Aguado, Gabriela Aguilar, Carlos Allende Prieto, Andres Almeida, Tonima Tasnim Ananna, Friedrich Anders, Scott F. Anderson, Brett H. Andrews, Borja Anguiano, Alfonso Aragón-Salamanca, Maria Argudo-Fernández, Eric Armengaud, Metin Ata, Eric Aubourg, Vladimir Avila-Reese, Carles Badenes, Stephen Bailey, Christophe Balland, Kathleen A. Barger, Jorge Barrera-Ballesteros, Curtis Bartosz, Fabienne Bastien, Dominic Bates, Falk Baumgarten, Julian Bautista, Rachael Beaton, Timothy C. Beers, Francesco Belfiore, Chad F. Bender, Mariangela Bernardi, Matthew A. Bershady, Florian Beutler, Jonathan C. Bird, Dmitry Bizyaev, Guillermo A. Blanc, Michael R. Blanton, Michael Blomqvist, Adam S. Bolton, Médéric Boquien, Jura Borissova, Jo Bovy, Christian Andres Bradna Diaz, William Nielsen Brandt, Jonathan Brinkmann, Joel R. Brownstein, Kevin Bundy, Adam J. Burgasser, Etienne Burtin, Nicolás G. Busca, Caleb I. Cañas, Mariana Cano-Díaz, Michele Cappellari, Ricardo Carrera, Andrew R. Casey, Bernardo Cervantes Sodi, Yanping Chen, Brian Cherinka, Cristina Chiappini, Peter Doohyun Choi, Drew Chojnowski, Chia-Hsun Chuang, Haeun Chung, Nicolas Clerc, Roger E. Cohen, Julia M. Comerford, Johan Comparat, Janaina Correa do Nascimento, Luiz da Costa, Marie-Claude Cousinou, Kevin Covey, Jeffrey D. Crane, Irene Cruz-Gonzalez, Katia Cunha, Gabriele da Silva Ilha, Guillermo J. Damke, Jeremy Darling, Jr. Davidson, James W., Kyle Dawson, Miguel Angel C. de Icaza Lizaola, Axel de la Macorra, Sylvain de la Torre, Nathan De Lee, Victoria de Sainte Agathe, Alice Deconto Machado, Flavia Dell'Agli, Timothée Delubac, Aleksandar M. Diamond-Stanic, John Donor, Juan José Downes, Niv Drory, Hélion du Mas des Bourboux, Christopher J. Duckworth, Tom Dwelly, Jamie Dyer, Garrett Ebelke, Arthur Davis Eigenbrot, Daniel J. Eisenstein, Yvonne P. Elsworth, Eric Emsellem, Michael Eracleous, Ghazaleh Erfanianfar, Stephanie Escoffier, Xiaohui Fan, Emma Fernández Alvar, J. G. Fernandez-Trincado, Rafael Fernand o Cirolini, Diane Feuillet, Alexis Finoguenov, Scott W. Fleming, Andreu Font-Ribera, Gordon Freischlad, Peter Frinchaboy, Hai Fu, Yilen Gómez Maqueo Chew, Lluís Galbany, Ana E. García Pérez, R. Garcia-Dias, D. A. García-Hernández, Luis Alberto Garma Oehmichen, Patrick Gaulme, Joseph Gelfand ,

Héctor Gil-Marín, Bruce A. Gillespie, Daniel Goddard, Jonay I. González Hernández, Violeta Gonzalez-Perez, Kathleen Grabowski, Paul J. Green, Catherine J. Grier, Alain Gueguen, Hong Guo, Julien Guy, Alex Hagen, Patrick Hall, Paul Harding, Sten Hasselquist, Suzanne Hawley, Christian R. Hayes, Fred Hearty, Saskia Hekker, Jesus Hernand ez, Hector Hernandez Toledo, David W. Hogg, Kelly Holley-Bockelmann, Jon A. Holtzman, Jiamin Hou, Bau-Ching Hsieh, Jason A. S. Hunt, Timothy A. Hutchinson, Ho Seong Hwang, Camilo Eduardo Jimenez Angel, Jennifer A. Johnson, Amy Jones, Henrik Jönsson, Eric Jullo, Fahim Sakil Khan, Karen Kinemuchi, David Kirkby, IV Kirkpatrick, Charles C., Francisco-Shu Kitaura, Gillian R. Knapp, Jean-Paul Kneib, Juna A. Kollmeier, Ivan Lacerna, Richard R. Lane, Dustin Lang, David R. Law, Jean-Marc Le Goff, Young-Bae Lee, Hongyu Li, Cheng Li, Jianhui Lian, Yu Liang, Marcos Lima, Lihwai Lin, Dan Long, Sara Lucatello, Britt Lundgren, J. Ted Mackereth, Chelsea L. MacLeod, Suvrath Mahadevan, Marcio Antonio Geimba Maia, Steven Majewski, Arturo Manchado, Claudia Maraston, Vivek Mariappan, Rui Marques-Chaves, Thomas Masseron, Karen L. Masters, Richard M. McDermid, Ian D. McGreer, Matthew Melendez, Sofia Meneses-Goytia, Andrea Merloni, Michael R. Merrifield, Szabolcs Meszaros, Andres Meza, Ivan Minchev, Dante Minniti, Eva-Maria Mueller, Francisco Muller-Sanchez, Demitri Muna, Ricardo R. Muñoz, Adam D. Myers, Preethi Nair, Kirpal Nand ra, Melissa Ness, Jeffrey A. Newman, Robert C. Nichol, David L. Nidever, Christian Nitschelm, Pasquier Noterdaeme, Julia O'Connell, Ryan James Oelkers, Audrey Oravetz, Daniel Oravetz, Erik Aquino Ortíz, Yeisson Osorio, Zach Pace, Nelson Padilla, Nathalie Palanque-Delabrouille, Pedro Alonso Palicio, Hsi-An Pan, Kaike Pan, Taniya Parikh, Isabelle Pâris, Changbom Park, Sebastien Peirani, Marcos Pellejero-Ibanez, Samantha Penny, Will J. Percival, Ismael Perez-Fournon, Patrick Petitjean, Matthew M. Pieri, Marc Pinsonneault, Alice Pisani, Francisco Prada, Abhishek Prakash, Anna Bárbara de Andrade Queiroz, M. Jordan Raddick, Anand Raichoor, Sandro Barboza Rembold, Hannah Richstein, Rogemar A. Riffel, Rogério Riffel, Hans-Walter Rix, Annie C. Robin, Sergio Rodríguez Torres, Carlos Román-Zúñiga, Ashley J. Ross, Graziano Rossi, John Ruan, Rossana Ruggeri, Jose Ruiz, Mara Salvato, Ariel G. Sánchez, Sebastián F. Sánchez, Jorge Sanchez Almeida, José R. Sánchez-Gallego, Felipe Antonio Santana Rojas, Basílio Xavier Santiago, Ricardo P. Schiavon, Jaderson S. Schimoia, Edward Schlafly, David Schlegel, Donald P. Schneider, William J. Schuster, Axel Schwope, Hee-Jong Seo, Aldo Serenelli, Shiyin Shen, Yue Shen, Matthew Shetrone, Michael Shull, Víctor Silva Aguirre, Joshua D. Simon, Mike Skrutskie, Anže Slosar, Rebecca Smethurst, Verne Smith, Jennifer Sobeck, Garrett Somers, Barbara J. Souter, Diogo Souto, Ashley Spindler, David V.

Stark, Keivan Stassun, Matthias Steinmetz, Dennis Stello, Thaisa Storchi-Bergmann, Alina Streblyanska, Guy S. Stringfellow, Genaro Suárez, Jing Sun, Laszlo Szigeti, Manuchehr Taghizadeh-Popp, Michael S. Talbot, Baitian Tang, Charling Tao, Jamie Tayar, Mita Tembe, Johanna Teske, Aniruddha R. Thakar, Daniel Thomas, Patricia Tissera, Rita Tojeiro, Christy Tremonti, Nicholas W. Troup, Meg Urry, O. Valenzuela, Remco van den Bosch, Jaime Vargas-González, Mariana Vargas-Magaña, Jose Alberto Vazquez, Sandro Villanova, Nicole Vogt, David Wake, Yuting Wang, Benjamin Alan Weaver, Anne-Marie Weijmans, David H. Weinberg, Kyle B. Westfall, David G. Whelan, Eric Wilcots, Vivienne Wild, Rob A. Williams, John Wilson, W. M. Wood-Vasey, Dominika Wylezalek, Ting Xiao, Renbin Yan, Meng Yang, Jason E. Ybarra, Christophe Yèche, Nadia Zakamska, Olga Zamora, Pauline Zarrouk, Gail Zasowski, Kai Zhang, Cheng Zhao, Gong-Bo Zhao, Zheng Zheng, Zheng Zheng, Zhi-Min Zhou, Guangtun Zhu, Joel C. Zinn, and Hu Zou. The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment. *The Astrophysical Journal Supplement Series*, 235(2):42, April 2018. doi: 10.3847/1538-4365/aa9e8a.

[189] DESI Collaboration, Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E. Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, Christophe Balland, Otger Ballester, Charles Baltay, Lucas Beaufore, Chris Bebek, Timothy C. Beers, Eric F. Bell, Jose Luis Bernal, Robert Besuner, Florian Beutler, Chris Blake, Hannes Bleuler, Michael Blomqvist, Robert Blum, Adam S. Bolton, Cesar Briceno, David Brooks, Joel R. Brownstein, Elizabeth Buckley-Geer, Angela Burden, Etienne Burtin, Nicolas G. Busca, Robert N. Cahn, Yan-Chuan Cai, Laia Cardiel-Sas, Raymond G. Carlberg, Pierre-Henri Carton, Ricard Casas, Francisco J. Castander, Jorge L. Cervantes-Cota, Todd M. Claybaugh, Madeline Close, Carl T. Coker, Shaun Cole, Johan Comparat, Andrew P. Cooper, M. C. Cousinou, Martin Crocce, Jean-Gabriel Cuby, Daniel P. Cunningham, Tamara M. Davis, Kyle S. Dawson, Axel de la Macorra, Juan De Vicente, Timothee Delubac, Mark Derwent, Arjun Dey, Govinda Dhungana, Zhejie Ding, Peter Doel, Yutong T. Duan, Anne Ealet, Jerry Edelstein, Sarah Eftekharzadeh, Daniel J. Eisenstein, Ann Elliott, Stephanie Escoffier, Matthew Evatt, Parker Fagrelius, Xiaohui Fan, Kevin Fanning, Arya Farahi, Jay Farihi, Ginevra Favole, Yu Feng, Enrique Fernandez, Joseph R. Findlay, Douglas P. Finkbeiner, Michael J. Fitzpatrick, Brenna Flaugher, Samuel Flender, Andreu Font-Ribera, Jaime E. Forero-Romero, Pablo Fosalba, Carlos S. Frenk, Michele Fumagalli, Boris T. Gaensicke,

Giuseppe Gallo, Juan Garcia-Bellido, Enrique Gaztanaga, Nicola Pietro Gentile Fusillo, Terry Gerard, Irena Gershkovich, Tommaso Giannantonio, Denis Gillet, Guillermo Gonzalez de Rivera, Violeta Gonzalez-Perez, Shelby Gott, Or Graur, Gaston Gutierrez, Julien Guy, Salman Habib, Henry Heetderks, Ian Heetderks, Katrin Heitmann, Wojciech A. Hellwing, David A. Herrera, Shirley Ho, Stephen Holland, Klaus Honscheid, Eric Huff, Timothy A. Hutchinson, Dragan Huterer, Ho Seong Hwang, Joseph Maria Illa Laguna, Yuzo Ishikawa, Dianna Jacobs, Niall Jeffrey, Patrick Jelinsky, Elise Jennings, Linhua Jiang, Jorge Jimenez, Jennifer Johnson, Richard Joyce, Eric Jullo, Stephanie Juneau, Sami Kama, Armin Karcher, Sonia Karkar, Robert Kehoe, Noble Kennamer, Stephen Kent, Martin Kilbinger, Alex G. Kim, David Kirkby, Theodore Kisner, Ellie Kitanidis, Jean-Paul Kneib, Sergey Koposov, Eve Kovacs, Kazuya Koyama, Anthony Kremin, Richard Kron, Luzius Kronig, Andrea Kueter-Young, Cedric G. Lacey, Robin Lafever, Ofer Lahav, Andrew Lambert, Michael Lampton, Martin Landriau, Dustin Lang, Tod R. Lauer, Jean-Marc Le Goff, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Su-Jeong Lee, Daniela Leitner, Michael Lesser, Michael E. Levi, Benjamin L'Huillier, Baojiu Li, Ming Liang, Huan Lin, Eric Linder, Sarah R. Loebman, Zarija Luki?, Jun Ma, Niall MacCrann, Christophe Magneville, Laleh Makarem, Marc Manera, Christopher J. Manser, Robert Marshall, Paul Martini, Richard Massey, Thomas Matheson, Jeremy McCauley, Patrick McDonald, Ian D. McGreer, Aaron Meisner, Nigel Metcalfe, Timothy N. Miller, Ramon Miquel, John Moustakas, Adam Myers, Milind Naik, Jeffrey A. Newman, Robert C. Nichol, Andrina Nicola, Luiz Nicolati da Costa, Jundan Nie, Gustavo Niz, Peder Norberg, Brian Nord, Dara Norman, Peter Nugent, Thomas O'Brien, Minji Oh, Knut A. G. Olsen, Cristobal Padilla, Hamsa Padmanabhan, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Antonella Palmese, Daniel Pappalardo, Isabelle Paris, Changbom Park, Anna Patej, John A. Peacock, Hiranya V. Peiris, Xiyan Peng, Will J. Percival, Sandrine Perruchot, Matthew M. Pieri, Richard Pogge, Jennifer E. Pollack, Claire Poppett, Francisco Prada, Abhishek Prakash, Ronald G. Probst, David Rabinowitz, Anand Raichoor, Chang Hee Ree, Alexandre Refregier, Xavier Regal, Beth Reid, Kevin Reil, Mehdi Rezaie, Constance M. Rockosi, Natalie Roe, Samuel Ronayette, Aaron Roodman, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, Eduardo Rozo, Vanina Ruhlmann-Kleider, Eli S. Rykoff, Cristiano Sabiu, Lado Samushia, Eusebio Sanchez, Javier Sanchez, David J. Schlegel, Michael Schneider, Michael Schubnell, Aurelia Secroun, Uros Seljak, Hee-Jong Seo, Santiago Serrano, Arman Shafieloo, Huanyuan Shan, Ray Sharples, Michael J. Sholl, William V. Shourt, Joseph H. Silber, David R. Silva, Martin M. Sirk, Anze Slosar, Alex Smith, George F. Smoot, Debopam Som, Yong-Seon Song, David Sprayberry,

Ryan Staten, Andy Stefanik, Gregory Tarle, Suk Sien Tie, Jeremy L. Tinker, Rita Tojeiro, Francisco Valdes, Octavio Valenzuela, Monica Valluri, Mariana Vargas-Magana, Licia Verde, Alistair R. Walker, Jiali Wang, Yuting Wang, Benjamin A. Weaver, Curtis Weaverdyck, Risa H. Wechsler, David H. Weinberg, Martin White, Qian Yang, Christophe Yeche, Tianmeng Zhang, Gong-Bo Zhao, Yi Zheng, Xu Zhou, Zhimin Zhou, Yaling Zhu, Hu Zou, and Ying Zu. The DESI Experiment Part I: Science,Targeting, and Survey Design, 2016.

[190] P. E. Dewdney, P. J. Hall, R. T. Schilizzi, and T. J. L. W. Lazio. The Square Kilometre Array. *Proceedings of the IEEE*, 97(8):1482–1496, 2009.

[191] Bolun Wang, Xiurui Zhu, Chunmei Gao, Yingguo Bai, J. W. Dong, and Li Jing Wang. Square kilometre array telescope — precision reference frequency synchronisation via 1f-2f dissemination. *Scientific Reports*, 5, 2015.

[192] Miguel F. Morales and J. Stuart B. Wyithe. Reionization and Cosmology with 21-cm Fluctuations. *Annual Review of Astronomy and Astrophysics*, 48(1):127–171, 2010. doi: 10.1146/annurev-astro-081309-130936. URL `https://doi.org/10.1146/annurev-astro-081309-130936`.

[193] M.G. Santos. Cosmology with the 21cm signal. In *51st Rencontres de Moriond on Cosmology*, pages 307–314. ARISF, 2016.

[194] P. J. E. Peebles and J. T. Yu. Primeval Adiabatic Perturbation in an Expanding Universe. *The Astrophysical Journal*, 162:815, December 1970. doi: 10.1086/150713.

[195] Yun Wang. Dark energy constraints from baryon acoustic oscillations. *The Astrophysical Journal*, 647(1):1–7, August 2006. doi: 10.1086/505384. URL `https://doi.org/10.1086%2F505384`.

[196] Eyal A. Kazin, Michael R. Blanton, Román Scoccimarro, Cameron K. McBride, and Andreas A. Berlind. Regarding the Line-of-sight Baryonic Acoustic Feature in the Sloan Digital Sky Survey and Baryon Oscillation Spectroscopic Survey Luminous Red Galaxy Samples. *The Astrophysical Journal*, 719(2):1032–1044, 2010. doi: 10.1088/0004-637X/719/2/1032.

[197] Martin White, M. Blanton, A. Bolton, D. Schlegel, J. Tinker, A. Berlind, L. da Costa, E. Kazin, Y. T. Lin, M. Maia, C. K. McBride, N. Padmanabhan, J. Parejko, W. Percival, F. Prada, B. Ramos, E. Sheldon, F. de Simoni, R. Skibba, D. Thomas, D. Wake, I. Zehavi,

Z. Zheng, R. Nichol, Donald P. Schneider, Michael A. Strauss, B. A. Weaver, and David H. Weinberg. The Clustering of Massive Galaxies at $z \sim 0.5$ from the First Semester of BOSS Data. *The Astrophysical Journal*, 728(2):126, 2011. doi: 10.1088/0004-637X/728/2/126.

[198] Ashley J. Ross, Will J. Percival, Ariel G. Sánchez, Lado Samushia, Shirley Ho, Eyal Kazin, Marc Manera, Beth Reid, Martin White, Rita Tojeiro, Cameron K. McBride, Xiaoying Xu, David A. Wake, Michael A. Strauss, Francesco Montesano, Molly E. C. Swanson, Stephen Bailey, Adam S. Bolton, Antonio Montero Dorta, Daniel J. Eisenstein, Hong Guo, Jean-Christophe Hamilton, Robert C. Nichol, Nikhil Padmanabhan, Francisco Prada, David J. Schlegel, Mariana Vargas Magaña, Idit Zehavi, Michael Blanton, Dmitry Bizyaev, Howard Brewington, Antonio J. Cuesta, Elena Malanushenko, Viktor Malanushenko, Daniel Oravetz, John Parejko, Kaike Pan, Donald P. Schneider, Alaina Shelden, Audrey Simmons, Stephanie Snedden, and Gong-bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: analysis of potential systematics. *Monthly Notices of the Royal Astronomical Society*, 424(1):564–590, 2012. doi: 10.1111/j.1365-2966.2012.21235.x.

[199] Rita Tojeiro, Will J. Percival, Jon Brinkmann, Joel R. Brownstein, Daniel J. Eisenstein, Marc Manera, Claudia Maraston, Cameron K. McBride, Demitri Muna, Beth Reid, Ashley J. Ross, Nicholas P. Ross, Lado Samushia, Nikhil Padmanabhan, Donald P. Schneider, Ramin Skibba, Ariel G. Sánchez, Molly E. C. Swanson, Daniel Thomas, Jeremy L. Tinker, Licia Verde, David A. Wake, Benjamin A. Weaver, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measuring structure growth using passive galaxies. *Monthly Notices of the Royal Astronomical Society*, 424(3):2339–2344, 2012. doi: 10.1111/j.1365-2966.2012.21404.x.

[200] Ariel G. Sánchez, C. G. Scóccola, A. J. Ross, W. Percival, M. Manera, F. Montesano, X. Mazzalay, A. J. Cuesta, D. J. Eisenstein, E. Kazin, C. K. McBride, K. Mehta, A. D. Montero-Dorta, N. Padmanabhan, F. Prada, J. A. Rubiño-Martín, R. Tojeiro, X. Xu, M. Vargas Magaña, E. Aubourg, N. A. Bahcall, S. Bailey, D. Bizyaev, A. S. Bolton, H. Brewington, J. Brinkmann, J. R. Brownstein, J. Richard Gott, J. C. Hamilton, S. Ho, K. Honscheid, A. Labatie, E. Malanushenko, V. Malanushenko, C. Maraston, D. Muna, R. C. Nichol, D. Oravetz, K. Pan, N. P. Ross, N. A. Roe, B. A. Reid, D. J. Schlegel, A. Shelden, D. P. Schneider, A. Simmons, R. Skibba, S. Snedden, D. Thomas, J. Tinker, D. A. Wake, B. A. Weaver, David H. Weinberg, Martin White, I. Zehavi, and G. Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological

implications of the large-scale two-point correlation function. *Monthly Notices of the Royal Astronomical Society*, 425(1):415–437, 2012. doi: 10.1111/j.1365-2966.2012.21502.x.

[201] Beth A. Reid, Lado Samushia, Martin White, Will J. Percival, Marc Manera, Nikhil Padmanabhan, Ashley J. Ross, Ariel G. Sánchez, Stephen Bailey, Dmitry Bizyaev, Adam S. Bolton, Howard Brewington, J. Brinkmann, Joel R. Brownstein, Antonio J. Cuesta, Daniel J. Eisenstein, James E. Gunn, Klaus Honscheid, Elena Malanushenko, Viktor Malanushenko, Claudia Maraston, Cameron K. McBride, Demitri Muna, Robert C. Nichol, Daniel Oravetz, Kaike Pan, Roland de Putter, N. A. Roe, Nicholas P. Ross, David J. Schlegel, Donald P. Schneider, Hee-Jong Seo, Alaina Shelden, Erin S. Sheldon, Audrey Simmons, Ramin A. Skibba, Stephanie Snedden, Molly E. C. Swanson, Daniel Thomas, Jeremy Tinker, Rita Tojeiro, Licia Verde, David A. Wake, Benjamin A. Weaver, David H. Weinberg, Idit Zehavi, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measurements of the growth of structure and expansion rate at $z = 0.57$ from anisotropic clustering. *Monthly Notices of the Royal Astronomical Society*, 426(4):2719–2737, 2012. doi: 10.1111/j.1365-2966.2012.21779.x.

[202] Shirley Ho, Antonio Cuesta, Hee-Jong Seo, Roland de Putter, Ashley J. Ross, Martin White, Nikhil Padmanabhan, Shun Saito, David J. Schlegel, Eddie Schlafly, Uros Seljak, Carlos Hernández-Monteagudo, Ariel G. Sánchez, Will J. Percival, Michael Blanton, Ramin Skibba, Don Schneider, Beth Reid, Olga Mena, Matteo Viel, Daniel J. Eisenstein, Francisco Prada, Benjamin A. Weaver, Neta Bahcall, Dimitry Bizyaev, Howard Brewinton, Jon Brinkman, Luiz Nicolaci da Costa, John R. Gott, Elena Malanushenko, Viktor Malanushenko, Bob Nichol, Daniel Oravetz, Kaike Pan, Nathalie Palanque-Delabrouille, Nicholas P. Ross, Audrey Simmons, Fernando de Simoni, Stephanie Snedden, and Christophe Yeche. Clustering of Sloan Digital Sky Survey III Photometric Luminous Galaxies: The Measurement, Systematics, and Cosmological Implications. *The Astrophysical Journal*, 761(1):14, 2012. doi: 10.1088/0004-637X/761/1/14.

[203] Lauren Anderson, Eric Aubourg, Stephen Bailey, Dmitry Bizyaev, Michael Blanton, Adam S. Bolton, J. Brinkmann, Joel R. Brownstein, Angela Burden, Antonio J. Cuesta, Luiz A. N. da Costa, Kyle S. Dawson, Roland de Putter, Daniel J. Eisenstein, James E. Gunn, Hong Guo, Jean-Christophe Hamilton, Paul Harding, Shirley Ho, Klaus Honscheid, Eyal Kazin, David Kirkby, Jean-Paul Kneib, Antoine Labatie, Craig Loomis, Robert H. Lupton, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Marc Manera, Claudia

Maraston, Cameron K. McBride, Kushal T. Mehta, Olga Mena, Francesco Montesano, Demetri Muna, Robert C. Nichol, Sebastián E. Nuza, Matthew D. Olmstead, Daniel Oravetz, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, John Parejko, Isabelle Pâris, Will J. Percival, Patrick Petitjean, Francisco Prada, Beth Reid, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Lado Samushia, Ariel G. Sánchez, David J. Schlegel, Donald P. Schneider, Claudia G. Scóccola, Hee-Jong Seo, Erin S. Sheldon, Audrey Simmons, Ramin A. Skibba, Michael A. Strauss, Molly E. C. Swanson, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas Magaña, Licia Verde, Christian Wagner, David A. Wake, Benjamin A. Weaver, David H. Weinberg, Martin White, Xiaoying Xu, Christophe Yèche, Idit Zehavi, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 427(4): 3435–3467, 2012. doi: 10.1111/j.1365-2966.2012.22066.x.

[204] Ashley J. Ross, Will J. Percival, Aurelio Carnero, Gong-bo Zhao, Marc Manera, Alvise Raccanelli, Eric Aubourg, Dmitry Bizyaev, Howard Brewington, J. Brinkmann, Joel R. Brownstein, Antonio J. Cuesta, Luiz A. N. da Costa, Daniel J. Eisenstein, Garrett Ebelke, Hong Guo, Jean-Christophe Hamilton, Mariana Vargas Magaña, Elena Malanushenko, Viktor Malanushenko, Claudia Maraston, Francesco Montesano, Robert C. Nichol, Daniel Oravetz, Kaike Pan, Francisco Prada, Ariel G. Sánchez, Lado Samushia, David J. Schlegel, Donald P. Schneider, Hee-Jong Seo, Alaina Sheldon, Audrey Simmons, Stephanie Snedden, Molly E. C. Swanson, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, and Idit Zehavi. The clustering of galaxies in the SDSS-III DR9 Baryon Oscillation Spectroscopic Survey: constraints on primordial non-Gaussianity. *Monthly Notices of the Royal Astronomical Society*, 428(2):1116–1127, 2013. doi: 10.1093/mnras/sts094.

[205] John K. Parejko, Tomomi Sunayama, Nikhil Padmanabhan, David A. Wake, Andreas A. Berlind, Dmitry Bizyaev, Michael Blanton, Adam S. Bolton, Frank van den Bosch, Jon Brinkmann, Joel R. Brownstein, Luiz Alberto Nicolaci da Costa, Daniel J. Eisenstein, Hong Guo, Eyal Kazin, Marcio Maia, Elena Malanushenko, Claudia Maraston, Cameron K. McBride, Robert C. Nichol, Daniel J. Oravetz, Kaike Pan, Will J. Percival, Francisco Prada, Ashley J. Ross, Nicholas P. Ross, David J. Schlegel, Don Schneider, Audrey E. Simmons, Ramin Skibba, Jeremy Tinker, Rita Tojeiro, Benjamin A. Weaver, Andrew Wetzel, Martin White, David H. Weinberg, Daniel Thomas, Idit Zehavi, and Zheng Zheng. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: the

low-redshift sample. *Monthly Notices of the Royal Astronomical Society*, 429(1):98–112, 2013. doi: 10.1093/mnras/sts314.

[206] Lado Samushia, Beth A. Reid, Martin White, Will J. Percival, Antonio J. Cuesta, Lucas Lombriser, Marc Manera, Robert C. Nichol, Donald P. Schneider, Dmitry Bizyaev, Howard Brewington, Elena Malanushenko, Viktor Malanushenko, Daniel Oravetz, Kaike Pan, Audrey Simmons, Alaina Shelden, Stephanie Snedden, Jeremy L. Tinker, Benjamin A. Weaver, Donald G. York, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III DR9 Baryon Oscillation Spectroscopic Survey: testing deviations from Λ and general relativity using anisotropic clustering of galaxies. *Monthly Notices of the Royal Astronomical Society*, 429(2):1514–1528, 2013. doi: 10.1093/mnras/sts443.

[207] Sebastián E. Nuza, Ariel G. Sánchez, Francisco Prada, Anatoly Klypin, David J. Schlegel, Stefan Gottlöber, Antonio D. Montero-Dorta, Marc Manera, Cameron K. McBride, Ashley J. Ross, Raul Angulo, Michael Blanton, Adam Bolton, Ginevra Favole, Lado Samushia, Francesco Montesano, Will J. Percival, Nikhil Padmanabhan, Matthias Steinmetz, Jeremy Tinker, Ramin Skibba, Donald P. Schneider, Hong Guo, Idit Zehavi, Zheng Zheng, Dmitry Bizyaev, Olena Malanushenko, Viktor Malanushenko, Audrey E. Oravetz, Daniel J. Oravetz, and Alaina C. Shelden. The clustering of galaxies at z $\approx$ 0.5 in the SDSS-III Data Release 9 BOSS-CMASS sample: a test for the ΛCDM cosmology. *Monthly Notices of the Royal Astronomical Society*, 432(1):743–760, 2013. doi: 10.1093/mnras/stt513.

[208] Ariel G. Sánchez, Eyal A. Kazin, Florian Beutler, Chia-Hsun Chuang, Antonio J. Cuesta, Daniel J. Eisenstein, Marc Manera, Francesco Montesano, Robert C. Nichol, Nikhil Padmanabhan, Will Percival, Francisco Prada, Ashley J. Ross, David J. Schlegel, Jeremy Tinker, Rita Tojeiro, David H. Weinberg, Xiaoying Xu, J. Brinkmann, Joel R. Brownstein, Donald P. Schneider, and Daniel Thomas. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological constraints from the full shape of the clustering wedges. *Monthly Notices of the Royal Astronomical Society*, 433(2):1202–1222, 2013. doi: 10.1093/mnras/stt799.

[209] Chia-Hsun Chuang, Francisco Prada, Antonio J. Cuesta, Daniel J. Eisenstein, Eyal Kazin, Nikhil Padmanabhan, Ariel G. Sánchez, Xiaoying Xu, Florian Beutler, Marc Manera, David J. Schlegel, Donald P. Schneider, David H. Weinberg, Jon Brinkmann, Joel R. Brownstein, and Daniel Thomas. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: single-probe measurements and the strong power of

$f(z)\sigma_8(z)$ on constraining dark energy. *Monthly Notices of the Royal Astronomical Society*, 433(4):3559–3571, 2013. doi: 10.1093/mnras/stt988.

[210] Claudia G. Scóccola, Ariel G. Sánchez, J. A. Rubiño-Martín, R. Génova-Santos, R. Rebolo, A. J. Ross, W. J. Percival, M. Manera, D. Bizyaev, J. R. Brownstein, G. Ebelke, E. Malanushenko, V. Malanushenko, D. Oravetz, K. Pan, D. P. Schneider, and A. Simmons. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: constraints on the time variation of fundamental constants from the large-scale two-point correlation function. *Monthly Notices of the Royal Astronomical Society*, 434(2):1792–1807, 2013. doi: 10.1093/mnras/stt1143.

[211] Eyal A. Kazin, Ariel G. Sánchez, Antonio J. Cuesta, Florian Beutler, Chia-Hsun Chuang, Daniel J. Eisenstein, Marc Manera, Nikhil Padmanabhan, Will J. Percival, Francisco Prada, Ashley J. Ross, Hee-Jong Seo, Jeremy Tinker, Rita Tojeiro, Xiaoying Xu, J. Brinkmann, Brownstein Joel, Robert C. Nichol, David J. Schlegel, Donald P. Schneider, and Daniel Thomas. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measuring $H(z)$ and $D_A(z)$ at $z = 0.57$ with clustering wedges. *Monthly Notices of the Royal Astronomical Society*, 435(1):64–86, 2013. doi: 10.1093/mnras/stt1261.

[212] Lauren Anderson, Eric Aubourg, Stephen Bailey, Florian Beutler, Adam S. Bolton, J. Brinkmann, Joel R. Brownstein, Chia-Hsun Chuang, Antonio J. Cuesta, Kyle S. Dawson, Daniel J. Eisenstein, Shirley Ho, Klaus Honscheid, Eyal A. Kazin, David Kirkby, Marc Manera, Cameron K. McBride, O. Mena, Robert C. Nichol, Matthew D. Olmstead, Nikhil Padmanabhan, N. Palanque-Delabrouille, Will J. Percival, Francisco Prada, Ashley J. Ross, Nicholas P. Ross, Ariel G. Sánchez, Lado Samushia, David J. Schlegel, Donald P. Schneider, Hee-Jong Seo, Michael A. Strauss, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Licia Verde, David Wake, David H. Weinberg, Xiaoying Xu, and Christophe Yeche. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measuring D_A and H at $z = 0.57$ from the baryon acoustic peak in the Data Release 9 spectroscopic Galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 439(1):83–101, 2014. doi: 10.1093/mnras/stt2206.

[213] Lauren Anderson, Éric Aubourg, Stephen Bailey, Florian Beutler, Vaishali Bhardwaj, Michael Blanton, Adam S. Bolton, J. Brinkmann, Joel R. Brownstein, Angela Burden, Chia-Hsun Chuang, Antonio J. Cuesta, Kyle S. Dawson, Daniel J. Eisenstein, Stephanie Escoffier, James E. Gunn, Hong Guo, Shirley Ho, Klaus Honscheid, Cullan Howlett, David

Kirkby, Robert H. Lupton, Marc Manera, Claudia Maraston, Cameron K. McBride, Olga Mena, Francesco Montesano, Robert C. Nichol, Sebastián E. Nuza, Matthew D. Olmstead, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, John Parejko, Will J. Percival, Patrick Petitjean, Francisco Prada, Adrian M. Price-Whelan, Beth Reid, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Cristiano G. Sabiu, Shun Saito, Lado Samushia, Ariel G. Sánchez, David J. Schlegel, Donald P. Schneider, Claudia G. Scoccola, Hee-Jong Seo, Ramin A. Skibba, Michael A. Strauss, Molly E. C. Swanson, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas Magaña, Licia Verde, David A. Wake, Benjamin A. Weaver, David H. Weinberg, Martin White, Xiaoying Xu, Christophe Yèche, Idit Zehavi, and Gong-Bo Zhao. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Releases 10 and 11 Galaxy samples. *Monthly Notices of the Royal Astronomical Society*, 441(1):24–62, 2014. doi: 10.1093/mnras/stu523.

[214] Héctor Gil-Marín, Jorge Noreña, Licia Verde, Will J. Percival, Christian Wagner, Marc Manera, and Donald P. Schneider. The power spectrum and bispectrum of SDSS DR11 BOSS galaxies - I. Bias and gravity. *Monthly Notices of the Royal Astronomical Society*, 451(1):539–580, 2015. doi: 10.1093/mnras/stv961.

[215] Héctor Gil-Marín, Licia Verde, Jorge Noreña, Antonio J. Cuesta, Lado Samushia, Will J. Percival, Christian Wagner, Marc Manera, and Donald P. Schneider. The power spectrum and bispectrum of SDSS DR11 BOSS galaxies - II. Cosmological interpretation. *Monthly Notices of the Royal Astronomical Society*, 452(2):1914–1921, 2015. doi: 10.1093/mnras/stv1359.

[216] Sukhdeep Singh, Rachel Mandelbaum, and Surhud More. Intrinsic alignments of SDSS-III BOSS LOWZ sample galaxies. *Monthly Notices of the Royal Astronomical Society*, 450(2):2195–2216, 2015. doi: 10.1093/mnras/stv778.

[217] Hironao Miyatake, Surhud More, Rachel Mandelbaum, Masahiro Takada, David N. Spergel, Jean-Paul Kneib, Donald P. Schneider, J. Brinkmann, and Joel R. Brownstein. The Weak Lensing Signal and the Clustering of BOSS Galaxies. I. Measurements. *The Astrophysical Journal*, 806(1):1, 2015. doi: 10.1088/0004-637X/806/1/1.

[218] Surhud More, Hironao Miyatake, Rachel Mandelbaum, Masahiro Takada, David N. Spergel, Joel R. Brownstein, and Donald P. Schneider. The Weak Lensing Signal and the Clustering of BOSS Galaxies. II. Astrophysical and Cosmological Constraints. *The Astrophysical Journal*, 806(1):2, 2015. doi: 10.1088/0004-637X/806/1/2.

[219] Martin White, Adam D. Myers, Nicholas P. Ross, David J. Schlegel, Joseph F. Hennawi, Yue Shen, Ian McGreer, Michael A. Strauss, Adam S. Bolton, Jo Bovy, X. Fan, Jordi Miralda-Escude, N. Palanque-Delabrouille, I. Paris, P. Petitjean, D. P. Schneider, M. Viel, David H. Weinberg, Ch. Yeche, I. Zehavi, K. Pan, S. Snedden, D. Bizyaev, H. Brewington, J. Brinkmann, V. Malanushenko, E. Malanushenko, D. Oravetz, A. Simmons, A. Sheldon, and Benjamin A. Weaver. The clustering of intermediate-redshift quasars as measured by the Baryon Oscillation Spectroscopic Survey. *Monthly Notices of the Royal Astronomical Society*, 424(2):933–950, 2012. doi: 10.1111/j.1365-2966.2012.21251.x.

[220] Shirley Ho, Nishant Agarwal, Adam D. Myers, Richard Lyons, Ashley Disbrow, Hee-Jong Seo, Ashley Ross, Christopher Hirata, Nikhil Padmanabhan, Ross O'Connell, Eric Huff, David Schlegel, Anže Slosar, David Weinberg, Michael Strauss, Nicholas P. Ross, Donald P. Schneider, Neta Bahcall, J. Brinkmann, Nathalie Palanque-Delabrouille, and Christophe Yèche. Sloan Digital Sky Survey III photometric quasar clustering: probing the initial conditions of the Universe. *Journal of Cosmology and Astroparticle Physics*, 2015(5):040, May 2015. doi: 10.1088/1475-7516/2015/05/040.

[221] David Kirkby, Daniel Margala, Anže Slosar, Stephen Bailey, Nicolás G. Busca, Timothée Delubac, James Rich, Julian E. Bautista, Michael Blomqvist, Joel R. Brownstein, Bill Carithers, Rupert A. C. Croft, Kyle S. Dawson, Andreu Font-Ribera, Jordi Miralda-Escudé, Adam D. Myers, Robert C. Nichol, Nathalie Palanque-Delabrouille, Isabelle Pâris, Patrick Petitjean, Graziano Rossi, David J. Schlegel, Donald P. Schneider, Matteo Viel, David H. Weinberg, and Christophe Yèche. Fitting methods for baryon acoustic oscillations in the Lyman-α forest fluctuations in BOSS data release 9. *Journal of Cosmology and Astroparticle Physics*, 2013(3):024, March 2013. doi: 10.1088/1475-7516/2013/03/024.

[222] Timothée Delubac, Julian E. Bautista, Nicolás G. Busca, James Rich, David Kirkby, Stephen Bailey, Andreu Font-Ribera, Anže Slosar, Khee-Gan Lee, Matthew M. Pieri, Jean-Christophe Hamilton, Éric Aubourg, Michael Blomqvist, Jo Bovy, Jon Brinkmann, William Carithers, Kyle S. Dawson, Daniel J. Eisenstein, Satya Gontcho A. Gontcho, Jean-Paul Kneib, Jean-Marc Le Goff, Daniel Margala, Jordi Miralda-Escudé, Adam D. Myers, Robert C. Nichol, Pasquier Noterdaeme, Ross O'Connell, Matthew D. Olmstead, Nathalie Palanque-Delabrouille, Isabelle Pâris, Patrick Petitjean, Nicholas P. Ross, Graziano Rossi, David J. Schlegel, Donald P. Schneider, David H. Weinberg, Christophe Yèche, and Donald G. York. Baryon acoustic oscillations in the Lyα forest of BOSS DR11 quasars. *Astronomy & Astrophysics*, 574:A59, 2015. doi: 10.1051/0004-6361/201423969.

[223] Andreu Font-Ribera, Eduard Arnau, Jordi Miralda-Escudé, Emmanuel Rollinde, J. Brinkmann, Joel R. Brownstein, Khee-Gan Lee, Adam D. Myers, Nathalie Palanque-Delabrouille, Isabelle Pâris, Patrick Petitjean, James Rich, Nicholas P. Ross, Donald P. Schneider, and Martin White. The large-scale quasar-Lyman α forest cross-correlation from BOSS. *Journal of Cosmology and Astroparticle Physics*, 2013(5):018, May 2013. doi: 10.1088/1475-7516/2013/05/018.

[224] Vid Iršič, Anže Slosar, Stephen Bailey, Daniel J. Eisenstein, Andreu Font-Ribera, Jean-Marc Le Goff, Britt Lundgren, Patrick McDonald, Ross O'Connell, Nathalie Palanque-Delabrouille, Patrick Petitjean, Jim Rich, Graziano Rossi, Donald P. Schneider, Erin S. Sheldon, and Christophe Yèche.

[225] Yue Shen, Cameron K. McBride, Martin White, Zheng Zheng, Adam D. Myers, Hong Guo, Jessica A. Kirkpatrick, Nikhil Padmanabhan, John K. Parejko, Nicholas P. Ross, David J. Schlegel, Donald P. Schneider, Alina Streblyanska, Molly E. C. Swanson, Idit Zehavi, Kaike Pan, Dmitry Bizyaev, Howard Brewington, Garrett Ebelke, Viktor Malanushenko, Elena Malanushenko, Daniel Oravetz, Audrey Simmons, and Stephanie Snedden. Cross-correlation of SDSS DR7 Quasars and DR10 BOSS Galaxies: The Weak Luminosity Dependence of Quasar Clustering at $z \sim 0.5$. *The Astrophysical Journal*, 778(2):98, 2013. doi: 10.1088/0004-637X/778/2/98.

[226] Andreu Font-Ribera, David Kirkby, Nicolas Busca, Jordi Miralda-Escudé, Nicholas P. Ross, Anže Slosar, James Rich, Éric Aubourg, Stephen Bailey, Vaishali Bhardwaj, Julian Bautista, Florian Beutler, Dmitry Bizyaev, Michael Blomqvist, Howard Brewington, Jon Brinkmann, Joel R. Brownstein, Bill Carithers, Kyle S. Dawson, Timothée Delubac, Garrett Ebelke, Daniel J. Eisenstein, Jian Ge, Karen Kinemuchi, Khee-Gan Lee, Viktor Malanushenko, Elena Malanushenko, Moses Marchante, Daniel Margala, Demitri Muna, Adam D. Myers, Pasquier Noterdaeme, Daniel Oravetz, Nathalie Palanque-Delabrouille, Isabelle Pâris, Patrick Petitjean, Matthew M. Pieri, Graziano Rossi, Donald P. Schneider, Audrey Simmons, Matteo Viel, Christophe Yeche, and Donald G. York. Quasar-Lyman α forest cross-correlation from BOSS DR11: Baryon Acoustic Oscillations. *Journal of Cosmology and Astroparticle Physics*, 2014(5):027, May 2014. doi: 10.1088/1475-7516/2014/05/027.

[227] Douglas Nychka, Soutir Bandyopadhyay, Dorit Hammerling, Finn Lindgren, and Stephan Sain. A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, June 2015.

[228] Charles Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151, 1981.

[229] Donald Shepard. A Two-dimensional Interpolation Function for Irregularly-spaced Data. In *Proceedings of the 1968 23rd ACM National Conference*, ACM '68, pages 517–524, New York, NY, USA, 1968. ACM.

[230] Jianqing Fan. Design-adaptive Nonparametric Regression. *Journal of the American Statistical Association*, 87:998–1004, 1992.

[231] D. Ruppert and M. P. Wand. Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*, 22:1346–1370, 1994.

[232] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Trans. Math. Softw.*, 3(3):209–226, September 1977.

[233] Matthias Katzfuss and Noel Cressie. Spatio-temporal smoothing and em estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, 32(4):430–446, 2011. doi: 10.1111/j.1467-9892.2011.00732.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.2011.00732.x`.

[234] Noel Cressie and Gardar Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, 2008. doi: 10.1111/j.1467-9868.2007.00633.x. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00633.x`.

[235] Finn Lindgren and Havard Rue. *Explicit construction of GMRF approximations to generalised Matern fields on irregular grids*, volume 12 of *Preprints in Mathematical Sciences*. 2007.

[236] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.

[237] Finn Lindgren, Håvard Rue, Johan Lindstrom, John T. Kent, Peter J. Diggle, J. B. Illian, D. P. Simpson, Tilmann Gneiting, Michael Scheuerer, R. Furrer, E. Furrer, D. Nychka, Paul Fearnhead, Peter Challenor, Yiannis Andrianakis, Gemma Stephenson, Jesper Moller, Xiangping Hu, Daniel Simpson, David Bolin, Patrick E. Brown, Michela Cameletti, Sara Martino, Daniel Cooley, Jennifer A. Hoeting, Rosa M. Crujeiras, Andres Prieto, Marco

A. R. Ferreira, Geir-Arne Fuglstad, Andrew Gelman, Peter Guttorp, Barnali Das, Ben Haaland, David J. Nott, John Haslett, Chaitanya Joshi, Vincent Garreta, Michael Hohle, L. Ippoliti, R. J. Martin, R. J. Bhansali, Venkata K. Jandhyala, Stergios B. Fotopoulos, Mikyoung Jun, Havard Wahl Kongsgard, Giovanna Jona Lasinio, Alessio Pollice, Chihoon Lee, Wayne T. Lee, Cari G. Kaufman, Bo Li, Marc G. Genton, Georg Lindgren, K. V. Mardia, Jorge Mateu, Debashis Mondal, Werner G. Muller, Helmut Waldl, Alessandro Ottavi, Omiros Papaspiliopoulos, Emilio Porcu, Marc Saez, Alexandra M. Schmidt, Alfred Stein, Paul Switzer, Kamil Turkman, Christopher K. Wikle, and Mevin B. Hooten. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach [with Discussion]. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73(4):423–498, 2011. ISSN 13697412, 14679868. URL `http://www.jstor.org/stable/41262260`.

[238] Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4:389–396, 1995.

[239] R. Brent. *Algorithms for Minimization without Derivatives*. Englewood Cliffs N.J.: Prentice-Hall, 1973.

[240] J. Sylvester. On the Relation between the Minor Determinants of Linearly Equivalent Quadratic Functions. *Philosophical Magazine*, 1:295–305, 1851.

[241] H. V. Henderson and S. R. Searle. On Deriving the Inverse of a Sum of Matrices. *SIAM Review*, 23:53–60, 1981.

[242] Hongguang Bi. Lyman-Alpha Absorption Spectrum of the Primordial Intergalactic Medium. *The Astrophysical Journal*, 405:479, March 1993. doi: 10.1086/172380.

[243] Romina Ahumada, Carlos Allende Prieto, Andres Almeida, Friedrich Anders, Scott F. Anderson, Brett H. Andrews, Borja Anguiano, Riccardo Arcodia, Eric Armengaud, Marie Aubert, Santiago Avila, Vladimir Avila-Reese, Carles Badenes, Christophe Balland, Kat Barger, Jorge K. Barrera-Ballesteros, Sarbani Basu, Julian Bautista, Rachael L. Beaton, Timothy C. Beers, B. Izamar T. Benavides, Chad F. Bender, Mariangela Bernardi, Matthew Bershady, Florian Beutler, Christian Moni Bidin, Jonathan Bird, Dmitry Bizyaev, Guillermo A. Blanc, Michael R. Blanton, Mederic Boquien, Jura Borissova, Jo Bovy, W. N. Brandt, Jonathan Brinkmann, Joel R. Brownstein, Kevin Bundy, Martin Bureau, Adam Burgasser, Etienne

Burtin, Mariana Cano-Diaz, Raffaella Capasso, Michele Cappellari, Ricardo Carrera, Solene Chabanier, William Chaplin, Michael Chapman, Brian Cherinka, Cristina Chiappini, Peter Doohyun Choi, S. Drew Chojnowski, Haeun Chung, Nicolas Clerc, Damien Coffey, Julia M. Comerford, Johan Comparat, Luiz da Costa, Marie-Claude Cousinou, Kevin Covey, Jeffrey D. Crane, Katia Cunha, Gabriele da Silva Ilha, Yu Sophia Dai, Sanna B. Damsted, Jeremy Darling, James W. Davidson Jr., Roger Davies, Kyle Dawson, Nikhil De, Axel de la Macorra, Nathan De Lee, Anna Barbara de Andrade Queiroz, Alice Deconto Machado, Sylvain de la Torre, Flavia Dell'Agli, Helion du Mas des Bourboux, Aleksandar M. Diamond-Stanic, Sean Dillon, John Donor, Niv Drory, Chris Duckworth, Tom Dwelly, Garrett Ebelke, Sarah Eftekharzadeh, Arthur Davis Eigenbrot, Yvonne P. Elsworth, Mike Eracleous, Ghazaleh Erfanianfar, Stephanie Escoffier, Xiaohui Fan, Emily Farr, Jose G. Fernandez-Trincado, Diane Feuillet, Alexis Finoguenov, Patricia Fofie, Amelia Fraser-McKelvie, Peter M. Frinchaboy, Sebastien Fromenteau, Hai Fu, Lluis Galbany, Rafael A. Garcia, D. A. Garcia-Hernandez, Luis Alberto Garma Oehmichen, Junqiang Ge, Marcio Antonio Geimba Maia, Doug Geisler, Joseph Gelfand, Julian Goddy, Jean-Marc Le Goff, Violeta Gonzalez-Perez, Kathleen Grabowski, Paul Green, Catherine J. Grier, Hong Guo, Julien Guy, Paul Harding, Sten Hasselquist, Adam James Hawken, Christian R. Hayes, Fred Hearty, S. Hekker, David W. Hogg, Jon Holtzman, Danny Horta, Jiamin Hou, Bau-Ching Hsieh, Daniel Huber, Jason A. S. Hunt, J. Ider Chitham, Julie Imig, Mariana Jaber, Camilo Eduardo Jimenez Angel, Jennifer A. Johnson, Amy M. Jones, Henrik Jonsson, Eric Jullo, Yerim Kim, Karen Kinemuchi, Charles C. Kirkpatrick IV, George W. Kite, Mark Klaene, Jean-Paul Kneib, Juna A. Kollmeier, Hui Kong, Marina Kounkel, Dhanesh Krishnarao, Ivan Lacerna, Ting-Wen Lan, Richard R. Lane, David R. Law, Henry W. Leung, Hannah Lewis, Cheng Li, Jianhui Lian, Lihwai Lin, Dan Long, Penelope Longa-Pena, Britt Lundgren, Brad W. Lyke, J. Ted Mackereth, Chelsea L. MacLeod, Steven R. Majewski, Arturo Manchado, Claudia Maraston, Paul Martini, Thomas Masseron, Karen L. Masters, Savita Mathur, Richard M. McDermid, Andrea Merloni, Michael Merrifield, Szabolcs Meszaros, Andrea Miglio, Dante Minniti, Rebecca Minsley, Takamitsu Miyaji, Faizan Gohar Mohammad, Benoit Mosser, Eva-Maria Mueller, Demitri Muna, Andrea Munoz-Gutierrez, Adam D. Myers, Seshadri Nadathur, Preethi Nair, Kirpal Nandra, Janaina Correa do Nascimento, Rebecca Jean Nevin, Jeffrey A. Newman, David L. Nidever, Christian Nitschelm, Pasquier Noterdaeme, Julia E. O'Connell, Matthew D Olmstead, Daniel Oravetz, Audrey Oravetz, Yeisson Osorio, Zachary J. Pace, Nelson Padilla, Nathalie Palanque-Delabrouille, Pedro A. Palicio, Hsi-An Pan, Kaike Pan, James Parker, Romain Paviot, Sebastien Peirani, Karla Pena Ramrez,

Samantha Penny, Will J. Percival, Ismael Perez-Fournon, Ignasi Perez-Rafols, Patrick Petitjean, Matthew M. Pieri, Marc Pinsonneault, Vijith Jacob Poovelil, Joshua Tyler Povick, Abhishek Prakash, Adrian M. Price-Whelan, M. Jordan Raddick, Anand Raichoor, Amy Ray, Sandro Barboza Rembold, Mehdi Rezaie, Rogemar A. Riffel, Rogerio Riffel, Hans-Walter Rix, Annie C. Robin, A. Roman-Lopes, Carlos Roman-Zuniga, Benjamin Rose, Ashley J. Ross, Graziano Rossi, Kate Rowlands, Kate H. R. Rubin, Mara Salvato, Ariel G. Sanchez, Laura Sanchez-Menguiano, Jose R. Sanchez-Gallego, Conor Sayres, Adam Schaefer, Ricardo P. Schiavon, Jaderson S. Schimoia, Edward Schlafly, David Schlegel, Donald P. Schneider, Mathias Schultheis, Axel Schwope, Hee-Jong Seo, Aldo Serenelli, Arman Shafieloo, Shoaib Jamal Shamsi, Zhengyi Shao, Shiyin Shen, Matthew Shetrone, Raphael Shirley, Victor Silva Aguirre, Joshua D. Simon, M. F. Skrutskie, Anze Slosar, Rebecca Smethurst, Jennifer Sobeck, Bernardo Cervantes Sodi, Diogo Souto, David V. Stark, Keivan G. Stassun, Matthias Steinmetz, Dennis Stello, Julianna Stermer, Thaisa Storchi-Bergmann, Alina Streblyanska, Guy S. Stringfellow, Amelia Stutz, Genaro Suarez, Jing Sun, Manuchehr Taghizadeh-Popp, Michael S. Talbot, Jamie Tayar, Aniruddha R. Thakar, Riley Theriault, Daniel Thomas, Zak C. Thomas, Jeremy Tinker, Rita Tojeiro, Hector Hernandez Toledo, Christy A. Tremonti, Nicholas W. Troup, Sarah Tuttle, Eduardo Unda-Sanzana, Marica Valentini, Jaime Vargas-Gonzalez, Mariana Vargas-Magana, Jose Antonio Vazquez-Mata, M. Vivek, David Wake, Yuting Wang, Benjamin Alan Weaver, Anne-Marie Weijmans, Vivienne Wild, John C. Wilson, Robert F. Wilson, Nathan Wolthuis, W. M. Wood-Vasey, Renbin Yan, Meng Yang, Christophe Yeche, Olga Zamora, Pauline Zarrouk, Gail Zasowski, Kai Zhang, Cheng Zhao, Gongbo Zhao, Zheng Zheng, Zheng Zheng, Guangtun Zhu, and Hu Zou. The Sixteenth Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra, 2019.

[244] Khee-Gan Lee, Joseph F. Hennawi, Martin White, Rupert A. C. Croft, and Melih Ozbek. Observational Requirements for Lyα Forest Tomographic Mapping of Large-scale Structure at $z \sim 2$. *The Astrophysical Journal*, 788(1):49, may 2014. doi: 10.1088/0004-637x/788/1/49. URL `https://doi.org/10.1088%2F0004-637x%2F788%2F1%2F49`.

[245] Casey W. Stark, Martin White, Khee-Gan Lee, and Joseph F. Hennawi. Protocluster discovery in tomographic Lyα forest flux maps. *Monthly Notices of the Royal Astronomical Society*, 453(1):311–327, 08 2015. ISSN 0035-8711. doi: 10.1093/mnras/stv1620. URL `https://doi.org/10.1093/mnras/stv1620`.

[246] B. Efron and R. Tibshirani. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1(1):54–75, February 1986. doi: 10.1214/ss/1177013815. URL `https://doi.org/10.1214/ss/1177013815`.