Title
Mapping the Cosmic Evolution of Hydrogen: Analysis and Inference Techniques for Next-Generation 21 cm Cosmology

Author
Kern, Nicholas

Publication Date
2020

Peer reviewed|Thesis/dissertation

Mapping the Cosmic Evolution of Hydrogen: Analysis and Inference Techniques for Next-Generation 21 cm Cosmology

By

Nicholas Steven Kern

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Astrophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Aaron R. Parsons, Chair
Professor Martin White
Professor Uroš Seljak

Summer 2020

Mapping the Cosmic Evolution of Hydrogen: Analysis and Inference Techniques for Next-Generation 21 cm Cosmology

Abstract

Mapping the Cosmic Evolution of Hydrogen: Analysis and Inference Techniques for Next-Generation 21 cm Cosmology

by

Nicholas Steven Kern

Doctor of Philosophy in Astrophysics

University of California, Berkeley

Professor Aaron R. Parsons, Chair

The path towards detecting the cosmological 21 cm signal from the Epoch of Reionization and Cosmic Dawn has seen tremendous progress over the past decade: interferometric experiments have placed increasingly stringent upper-limits on the 21 cm power spectrum, and global signal experiments have made a tentative first-detection of the 21 cm monopole signal. As next-generation experiments are designed, built, and begin taking data, the road towards maximizing their scientific potential hinges crucially upon the development of robust data analysis techniques for isolating the 21 cm signal from foregrounds and low-level instrumental systematics. This thesis is a compilation of work that describes new and old techniques and applies them to the up-and-coming Hydrogen Epoch of Reionization Array (HERA), a second-generation low-frequency experiment that aims to make a detection of the 21 cm power spectrum from redshifts $6 < z < 20$. Such a feat would dramatically improve our understanding of the intergalactic medium at these redshifts, and would shed light on the formation and properties of the first stars, galaxies, and compact objects in the universe. Towards this goal, this work touches on some of the current challenges facing the analysis of 21 cm data, from instrument commissioning and calibration to astrophysical parameter inference. In particular, we develop a detailed understanding of low-level systematics in the HERA system, yielding improved sensitivity in the measured power spectrum by over two orders of magnitude. As HERA construction is finished and full-sensitivity observing commences, the algorithms and analysis frameworks discussed here are setting the stage for a future full-sensitivity HERA analysis and the next-generation of 21 cm cosmology.

to Elise

Contents

List of Figures

List of Tables

Acknowledgments

My days and evenings in the Astronomy Department at UC Berkeley have been some of the best of my life. It is a pleasure to acknowledge my family, friends, colleagues, and mentors who have helped me on my journey through graduate school, and ultimately made it possible for me to take on a PhD.

First, I want to thank my research advisor, Aaron Parsons, who taught me radio interferometry and whose mentorship guided me through my PhD. Working with experimental data can be as frustrating as it is gratifying, and Aaron's incredible intuition and ability to tackle a problem–any problem–has been an incalculable benefit. I hope to have gleaned some fraction of that ability during my graduate studies. Also, one of the things that I've enjoyed the most about Berkeley is the close-knit 21 cm group: thanks for fostering a friendly and collegial environment in the lab. I also need to thank my informal advisor, Adrian Liu, who helped me early in my career, is someone I could always count on to throw an idea around or walk through a problem, and has been an inspiration as a mentor.

I am also grateful to my fellow radio lab colleagues, scientists, postdocs and graduate students: Dave DeBoer, Josh Dillon (thanks for working closely with me and for being an astute code reviewer), Phil Bull, Ridhima Nunhokee, Jack Hickish, Aaron Ewall-Wice, Deepthi Gorthi (getting coffee or boba with you and chatting radio stuff was always the highlight of my afternoon), Zaki Ali (thanks for showing me the data analysis ropes), Carina Cheng (thanks for your friendship), Gerry Zhang, Kara Kundert, and Morgan Presley. Thank you for all of the pizza Fridays, beers after exams, and the moral support that is the highlight of working in the radio lab. I also need to thank members of the HERA Collaboration for always pushing me to do my best work, and for working tirelessly to make HERA work.

I also want like to thank my fellow 2015 graduate cohort colleagues for embarking on the PhD journey with me. When I think back on my graduate studies, I realize it is the friendships we've made by working and being together over the years that has made this so enjoyable, and is certainly what I will miss the most.

I also need to thank my parents, brother, and sister, who have always supported me in my academics and given me so much love throughout my life.

Finally, I want to thank and acknowledge my life partner, best friend, and wife, Elise: your love and companionship guides me through life and gives meaning to what I do.

Chapter 1

Introduction

1.1 The Cosmic Evolution of Hydrogen

After the Big Bang, the universe was filled with a hot, ionized plasma, with a baryonic content consisting mostly of free electrons and protons. As the universe expanded it cooled adiabatically until, after roughly 380,000 years, the plasma was cool enough for electrons to re-combine onto protons. With the diminishment of the free electron population, the thermal bath of photons–before confined to travel only short distances due to the high Thompson electron scattering cross section–were now allowed to free-stream. This event was known as *recombination* and is marked by the rapid neutralization of atomic hydrogen (HI) at a redshift $z \sim 1100$. The free-streaming photons from this epoch constitute the Cosmic Microwave Background (CMB), whose discovery has revolutionized modern cosmology (Penzias & Wilson 1965; Dicke et al. 1965). Detailed studies of the CMB photon temperature fluctuations have pulled back the curtain on the initial conditions of the universe–the early, pre-inflation quantum fluctuations that at later times become the seeds of cosmological large scale structure growth–and have become a pillar of the current cosmological ΛCDM standard model (Komatsu et al. 2011; Planck Collaboration et al. 2018). This model states that, to good approximation, our universe can be described by a handful of physical parameters, which tell us that the universe is nearly flat in curvature; has a present-day energy content that is dominated by a cosmological constant ("dark energy"), and has a matter content that is comprised predominantly by a non-baryonic "dark matter," with only 5% of the mass-energy budget being baryonic in nature.

Although hydrogen makes up only a small fraction of the total content of the universe, it is the most abundant baryonic element and is of critical importance to the formation of collapsed luminous structures, like stars and galaxies. Understanding the evolution of hydrogen is therefore crucial to our broader understanding of the formation process of galaxies, stars, planets, and, ultimately, intelligent life. In addition, as we will discuss further in this work, it can also be used a cosmological probe to trace the growth of large scale structure throughout cosmic history. In the current epoch 13.7 billion years after the Big Bang, hydrogen exists in an neutral atomic and molecular state within collapsed, self-shielded structures like galaxies, which are the sites of star formation. However, a large amount of hydrogen exists in the intervening space between galaxies,

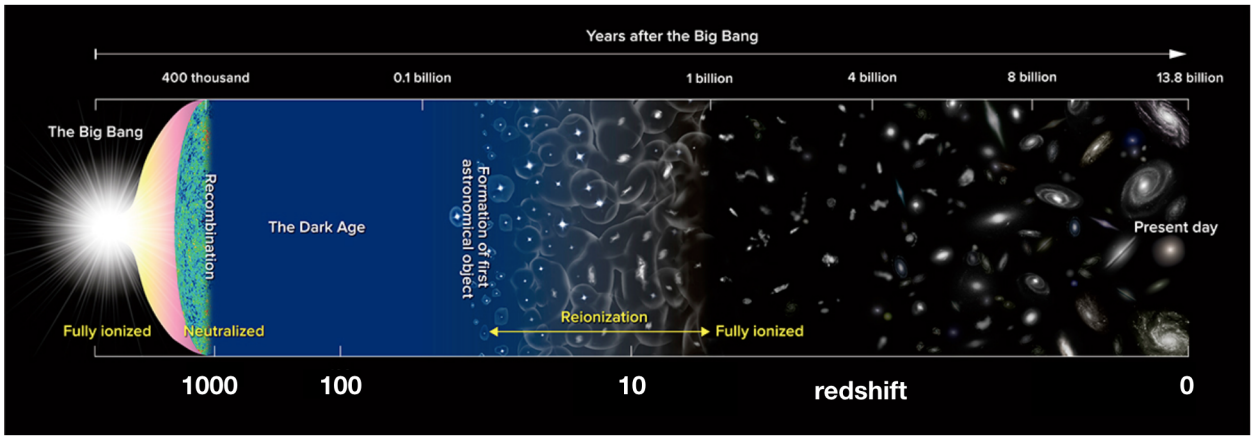

Figure 1.1: A cosmological timeline of the universe, highlighting the recombination epoch ($z \sim 1100$), the Epoch of Reionization ($z \sim 6$), and the growth of large scale structure to the present day. Reproduced from the National Astronomical Observatory of Japan (NAOJ).

known as the intergalactic medium (IGM), in a nearly completely ionized state (Pritchard & Loeb 2012). This marks a drastic change from the early recombination epoch, where the baryonic mass budget was filled predominately by neutral hydrogen. How did this happen? What series of events to the current era led to this drastic phase transition for the most important baryon? These are the broad scientific questions that underpin the thrust of this thesis.

Our current understanding of this processes is outlined by Figure 1.1, which shows the evolution of large scale structure from the Big Bang (left) up to the present day (right). As stated, we know that hydrogen atoms were largely neutralized at recombination ($z \sim 1100$), leading to the fraction of hydrogen atoms with bound electrons relative to the number of ionized hydrogen atoms (i.e. the neutral fraction $x_{\rm HI}$) being nearly one. Afterwards, diffusely separated baryonic matter began to coalesce under the influence of gravity. For some time this process occurred, known as the Dark Ages, without anything spectacular happening: the densest regions of the universe not being dense enough to actually form stars (Scott & Rees 1990). At some point, however, matter condensed enough to form the very first generations of stars, which injected high energy photons that both heated and ionized their surrounding environments, as well as enriching the media with heavy metals formed in the interiors of these first stars. This epoch, when the first stars formed in the densest regions of the universe and profoundly altered their ionization, temperature, and chemical state is known as *Cosmic Dawn* (CD). Eventually, the first generation of stellar populations and proto-galaxies formed, and continued to emit high energy photons, which heated and ionized sufficiently more of the neutral hydrogen in the IGM. At some point, this resulted in the near entire (re-)ionization of HI in the IGM, known as the *Epoch of Reionization* (EoR).

As shown in Figure 1.1, we think the EoR ended around 1 billion years after the Big Bang ($z \sim 6$). This has been inferred largely from observations of quasar spectra at high redshift, showing a dearth of Lyman-α transmission from quasars above $z \sim 6$ implying considerably larger neutral fractions in the intervening space (Gunn & Peterson 1965; Fan et al. 2006). However, we know

considerably less about when the EoR started, how long it took to reionize hydrogen in the IGM, and what luminous sources actually drove reionization. Furthermore, we know little about the properties of the first luminous objects at Cosmic Dawn, such as when they formed, what their radiative properties were, and the details of how they heated and enriched the IGM with metals. Although we have mapped the clustering of millions of galaxies in the local universe with optical spectroscopy (Blanton et al. 2017), and have studied the afterglow of the Big Bang in the universe's infancy through the CMB (Planck Collaboration et al. 2018), the Cosmic Dawn and the EoR remain a largely unexplored part of our cosmic history, due in part to the difficulty of observing it with conventional methods. However, an alternative technique promises to revolutionize our understanding of the EoR and Cosmic Dawn by systematically mapping out the structure of the IGM throughout these eras, known as 21 cm cosmology (Furlanetto et al. 2006; Morales & Wyithe 2010; Pritchard & Loeb 2012; Liu & Shaw 2019).

1.2 21 cm Cosmology: A 3D Probe of Large Scale Structure

The hyperfine splitting of the electron's ground state in a hydrogen atom emits or absorbs a photon with a wavelength of 21 cm (1.420 GHz), known as HI's 21 cm line. Although this is a forbidden transition, the sheer amount of neutral hydrogen in the universe makes this a readily observable signature, and has been used extensively to map the structure of the Milky Way and the distribution of nearby galaxies (Haynes et al. 2018). However, this signal can also be observed at cosmological distances, and between redshifts $300 > z > 6$, the 21 cm line traces the neutral hydrogen content in the IGM. As 21 cm photons traverse through the universe they experience cosmological redshift, meaning we observe them with our telescopes at a much lower frequency (or longer wavelength). While 21 cm radiation has a rest-frame wavelength $\lambda_{\rm rest} = 21$ cm, its observed wavelength is directly proportional to the redshift at which it was emitted,

$$\lambda_{\rm obs} = \frac{\lambda_{\rm rest}}{1+z}. \tag{1.1}$$

Thus, the 21 cm line allows us to trace the distribution of neutral hydrogen not only across the sky, but also *along the line-of-sight*, simply by scanning our telescopes up and down in frequency. This combination makes the 21 cm line a three-dimensional probe of the IGM, holding a tremendous information content that, if tapped, could transform our understanding of galaxy formation and cosmology.

What we observe with our telescopes is the 21 cm specific intensity I, which is equivalently expressed as a "brightness temperature," or the temperature of a blackbody emitting radiation of intensity $I(T)$. The observed brightness temperature of 21 cm radiation from a high-redshift HI cloud in the IGM relative to the background CMB photon field is proportional to the neutral fraction of the cloud, the density of the cloud, the excitation temperature of the HI hyperfine transition (also known as its spin temperature T_S), and cosmological parameters, written as

$$\delta T_b(z) \approx 9 x_{\rm HI}(1+\delta_b)(1+z)^{1/2}\left[1-\frac{T_\gamma(z)}{T_S}\right]\left[\frac{H(z)/(1+z)}{{\rm d}v_\parallel/{\rm d}r_\parallel}\right]\ {\rm mK}, \tag{1.2}$$

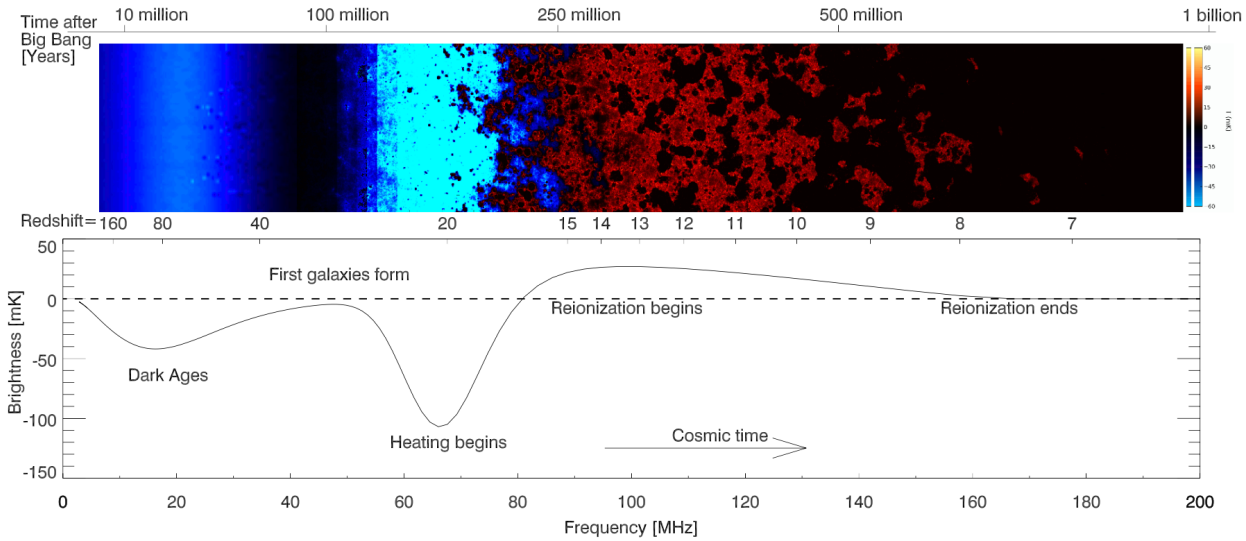

Figure 1.2: A prediction for the differential 21 cm brightness temperature during Cosmic Dawn and the Epoch of Reionization. The top panel shows a slice through a 3D simulation showing the spatial fluctuations of the signal, while the bottom panel shows the evolution of the monopole (global signal) brightness over time. The peaks and troughs of the global signal correspond to physically meaningful events in the process of galaxy formation, including when the first stars were formed, when they heated the large-scale IGM, and when they reionization neutral hydrogen in the IGM. Reproduced from Pritchard & Loeb (2012).

where δ_b is the density of the (baryons) gas, T_γ is the background CMB temperature, T_S is the HI spin temperature, $H(z)$ is the Hubble parameter, and $\mathrm{d}v_\parallel/\mathrm{d}r_\parallel$ is the gradient of the proper line-of-sight velocity of the HI cloud (Furlanetto et al. 2006). Importantly, if the neutral fraction is zero or if the spin temperature is in equilibrium with the photon temperature, $\delta T_b \rightarrow 0$. We can also see that if the spin temperature is greater than the CMB, we observe the HI cloud in emission, while if the spin temperature is less than the CMB, we observe the cloud in absorption.

Figure 1.2 shows a theoretical prediction for the δT_b signal during the Dark Ages, Cosmic Dawn, and reionization epochs. The top panels shows a slice through a 3D simulation revealing the spatial fluctuations of the signal, while the bottom panel shows the sky-averaged monopole signal (also known as the "global signal") as a function of redshift. We see that the evolution of the signal shows clear peaks and troughs that correspond to physically meaningful events in the formation of luminous structure. This is driven largely by the coupling of T_S to different physical temperatures (either the background CMB temperature or the kinetic temperature of the gas), which is dictated by the expansion of the universe and the injection of photons into the IGM by the first stars.

At ultra high redshifts ($z > 300$) T_S is collisionally coupled to the gas temperature, which itself is coupled to the photon temperature due to a small but effective residual free electron population after recombination. However, at $z < 300$ the gas de-couples from the photon temperature as the Thompson scattering mean-free path becomes large, and thus the gas cools adiabatically whereas T_γ cools only as $(1 + z)^{-1}$. This means that below $z \sim 300$ the 21 cm signal can be observed in

absorption, which coincides with the Dark Ages (Scott & Rees 1990). At some point, though, the universe expands enough to render collisional coupling ineffective, forcing T_S to re-couple with T_γ, and thus driving $\delta T_b \rightarrow 0$ around $z \sim 60$. This occurs until the formation of the first Population III stars, which inject a spectrum of photons into their immediate surroundings, and in particular dispere Lyman-α photons into the IGM. These photons trigger the Wouthuysen-Field effect (Field 1958), which actually re-couples T_S to the gas temperature via Lyman-α scattering. At this point the gas temperature is significantly cooler than the CMB, which drives the deep absorption trough seen at $z \sim 20$ in Figure 1.2.[1] This trend sharply reverses after the IGM gas is sufficiently heated from UV and X-ray photons sourced by the first stars and black holes, and is in fact enough to drive the signal into emission for the first time. IGM heating is expected to be inhomogenous, driving spatial fluctuations in the 21 cm signal. Hydrogen reionization begins when these sources begin to ionize an appreciable amount of HI in the IGM ($z \sim 15$), which is a highly inhomogenous process and drives more spatial fluctations. As the EoR ends and $x_{\rm HI} \rightarrow 0$, so too does the 21 cm signal. Due to its sensitive dependences on the temperature, density, and ionization state of the gas, 21 cm summary statistics like the power spectrum are unique probes into the astrophysics of Cosmic Dawn and the EoR (Pober et al. 2014; Ewall-Wice et al. 2016a; Greig & Mesinger 2017a), and may also be used in tandem with cosmological probes to break degeneracies and more tightly constrain ΛCDM parameters (Liu et al. 2016; Kern et al. 2017).

A considerable number of experiments over the past decade have been designed to measure the 21 cm signal from the EoR and Cosmic Dawn, and are largely distinct in their design depending on their targeting of either the 21 cm power spectrum or the global signal. Generally, interferometers with their higher angular resolution and smaller field of view are used to measure the power spectrum of spatial fluctuations, while single-antenna telescopes with wider field-of-views (FoV) are used to measure the monopole signal. Experiments targeting the power spectrum include the Giant Metrewave Radio Telescope (GMRT; Pen et al. 2009; Paciga et al. 2013), the Murchison Widefield Array (MWA; Bowman et al. 2013; Dillon et al. 2015; Ewall-Wice et al. 2016b; Beardsley et al. 2016; Barry et al. 2019b; Li et al. 2019; Trott et al. 2020), the Donald C. Backer Precision Array for Probing the Epoch of Reionization (PAPER; Parsons et al. 2010; Ali et al. 2015; Jacobs et al. 2015), the Low Frequency Array (LOFAR; Patil et al. 2017; Gehlot et al. 2018; Mertens et al. 2020), and the Long Wavelength Array (LWA; Eastwood et al. 2019). Together these experiments have set increasingly stringent upper limits on the EoR and Cosmic Dawn power spectrum, but none have yielded a convincing detection of the signal. None of the experiments are significantly noise-limited, and are hampered by instrumental and environment systematics that appear in the process of taking data over long seasons, calibrating the data, and mitigating foreground emission from the galaxy and extragalactic point sources.

Going forward, the next near-future experiment for the EoR and Cosmic Dawn power spectrum is the Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017). HERA is a next-generation, targeted 21 cm experiment that is built specifically for making a detection of the power

[1] Recently, a tentative detection of this monopole absorption trough was reported by the Experiment to Detect the Global EoR Signature (EDGES; Bowman et al. 2018), which was found to be significantly deeper than expected and, if not a consequence of an instrumental systematic, may need non-standard physics to explain (e.g. Barkana et al. 2018; Muñoz et al. 2018).

Figure 1.3: Left: A section of the HERA array under construction in South Africa in mid-2018. Right: An isolated HERA antenna in the field, showing the cross-dipole feed suspended above the dish.

spectrum at Cosmic Dawn and the EoR. It will be a closely-packed interferometer consisting of 350 dishes each 14-meters in diameter (Figure 1.3). While the experiment has the potential for a high-significance detection of the EoR power spectrum, instrumental systematics will still need to be modeled to very high precision in order to fulfill its science goals.

1.3 This Thesis

This thesis focuses on addressing a number of outstanding questions related to the analysis and interpretation of future 21 cm datasets, with a particular emphasis on applications to HERA.

Relating a 21 cm dataset to constraints on astrophysical and cosmological parameters is a non-trivial process for 21 cm cosmology at the EoR and Cosmic Dawn, as the signal is dependent on a host of complex astrophysical processes that are difficult to model from first principles. If we are to extract astrophysics and/or cosmology from our data, we need a method for accelerating our models. In this work, we demonstrate the application of machine learning techniques for hastening this process, and present a parameter forecast for HERA spanning across both astrophysical and cosmological parameters for the first time.

When working with real HERA data, instrumental systematics will need to be modeled to higher precision than ever before if it is to make a robust detection of the power spectrum. We present methods for modeling and mitigating an array of instrumental systematics, which, when demonstrated on real HERA data, show promising results. As 21 cm analysis pipelines become increasingly sophisticated, the need for independent validation testing of their ability to actually recover a mock EoR signal becomes increasingly important. In this work, we also show how such validation tests can instill confidence in the analysis, and help us better understand the statistical properties of the systematics themselves. We apply the techniques explored in this work spanning data reduction, calibration, systematic modeling, and power spectrum estimation to a deep integration from a subset of the HERA array and show that we recover noise-limited power spectra for $k \geq 0.25\ h\ \mathrm{Mpc}^{-1}$ with a peak-foreground-power dynamic range of a factor of 10^8. Due to the

fact that, for similar values of k, fiducial EoR signals are expected to lie around a foreground-power dynamic range of 10^10 for HERA (Thyagarajan et al. 2016), our analysis bodes well for HERA's ability to push down to fiducial EoR signal amplitudes when the full array is completed and usher-in the next-generation of 21 cm cosmology.

Chapter 2

Statistical Parameter Inference with Emulators

Here we describe what is normally the very last step in the process of a cosmological experiment: taking our reduced and compressed data and using it to infer the physical parameters of the universe. We discuss some of the challenges faced by 21 cm experiments in placing comprehensive and robust parameter constraints on the cosmology and astrophysics of the EoR. We then show that a surrogate model, or computer emulator (Habib et al. 2007), can help to overcome some of these challenges, and discuss some of their limitations. Drawing from Kern et al. (2017), we demonstrate the first application of these techniques to the problem of 21 cm EoR and Cosmic Dawn parameter inference. We show that, given a futuristic 21 cm dataset, we can constrain a wide range of physical parameters spanning UV-photon production at the EoR, X-ray photon production at Cosmic Dawn, and large-scale structure growth.

2.1 The Parameter Inference Problem

Given a theoretical model, parameter inference boils down to placing quantitative constraints on the model parameters given a series of observations and their associated uncertainties. To do this we seek to estimate the *posterior probability distribution*, $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$, of our model parameters ($\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_m\}$) given the data vector ($\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$), which Bayes' theorem describes as

$$\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\mathcal{P}(\boldsymbol{x})}, \tag{2.1}$$

where $\mathcal{L}$ is the *likelihood function* of our data, $\pi(\theta)$ is the *prior probability distribution* of our model parameters, and $\mathcal{P}(\boldsymbol{x})$ is the *probability distribution of our data*, also known as the Bayesian evidence (Sivia & Skilling 2006). The standard maximum likelihood estimate, $\hat{\boldsymbol{\theta}}_{\rm ML}$, is found by maximizing $\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta})$, whereas the maximum a posteroiri estimate, $\hat{\boldsymbol{\theta}}_{\rm MAP}$, is found by maximizing $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$. In this work, we will use $\hat{\boldsymbol{\theta}}$ to mean exclusively the maximum a posteriori (MAP) estimate. For the purposes of parameter estimation, we drop the Bayesian evidence term as it is simply an

overall normalization factor that does not change the values of $\hat{\boldsymbol{\theta}}$ that maximize $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x})$, and as such we can simply write Equation 2.1 as a proportionality. Assuming our data is Gaussian distributed, we can write its likelihood as

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi)^m \det|\boldsymbol{\Sigma}|}} \exp\left[-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))\right], \tag{2.2}$$

where $\boldsymbol{\Sigma}$ is our data covariance matrix of size $n \times n$, and $\boldsymbol{\mu}$ is our model's prediction given a set of parameter values $\boldsymbol{\theta}$.

To make our lives easier, we often seek to maximize the log probability distribution. Dropping constants, this leaves us with

$$\ln \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{x}) \propto -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) + \ln \pi(\boldsymbol{\theta}) \tag{2.3}$$

For simple models, like a linear model describing a polynomial fit, we can easily extremize the likelihood (yielding the classic linear least squares solution), and with well-behaved priors we can also extremize the posterior analytically. In most practical scenarios at the frontier of astrophysics, though, this is not the case: our model is either an analytic but complex, high dimensional function, or is non-analytic entirely, in which case we are forced to maximize it with a gradient descent algorithm. However, this is not the real challenge for parameter inference, as fast and efficient gradient descent computer algorithms have been around for many years. Instead the challenge is mapping the topology of the posterior distribution throughout its parameter space, and estimating the width of the distribution about its maximum. This involves stepping through a parameter space and evaluating the model many times in order to perform the characterization. This is the Markov chain Monte Carlo (MCMC) approach to sampling the posterior distribution and solving for $\hat{\boldsymbol{\theta}}$ and its credible intervals, or the narrowest parameter space bounds straddling $\hat{\boldsymbol{\theta}}$ and enclosing a fixed fraction of the total probability mass (Brooks et al. 2011). Evaluating the credible intervals, in effect our errorbars on $\hat{\boldsymbol{\theta}}$, involves integrating the posterior distribution across its parameter space, which can be a difficult task given a high dimensional and complex model, but is very straight-forward given a set of Markov chains.

While many different kinds of Markov chain schemes have been designed for MCMC and optimized for specific scenarios (multi-modal distributions, highly degenerate parameters, large parameter spaces, etc.), the fundamental requirement is that we are able to evaluate our physical model at many locations in the parameter space. Large dimensional parameter spaces make this requirement exponentially more strict. However, in many modern astrophysical contexts, our models are extremely sophisticated and computationally intensive to run, such as an N-body and/or hydrodynamic cosmological simulation. This places a fundamental limit on the total number of realizations we can make of the model given realistic computational resources. In this limit, direct MCMC is often not feasible, and to make parameter inference possible we need to further approximate the posterior distribution.

To give an example, let's assume we are working with a simulation that takes 48 core-hours to run and is embarrassingly parallel. The number of times we need to evaluate it depends on the convergence properties of our Markov chain, as well as the dimensionality of our parameter

space and the shape of the posterior itself. For example, the first few iterations of the sampler are strongly influenced by the sampler starting point, and not the posterior itself. Therefore, the sampler iterations made before reaching some sort of a stationary distribution (the burn-in phase) are typically discarded. For a handful of parameters and a fairly simple posterior distribution (i.e. one that is not highly degenerate), rule of thumb dictates on the order of $10^3 - 10^4$ MCMC iterations for sufficient convergence; however, convergence cannot be proven and a range of diagnostics can be employed to assess whether the chains are well-behaved. For our 48 core-hour astrophysical simulation, this implies a total walltime of 4.8×10^5 core-hours, or 20 days with a 1000-core machine. For many experimental groups, this may be an unfeasible amount of computational resources or wait-time for just a single MCMC run, which generally needs to be repeated multiple times to ensure consistency and for data quality management (e.g. for null tests). For context, even when factoring in Moore's law, direct MCMC of a state-of-the-art ~100-thousand core-hour hydrodynamical cosmological simulation with existing MCMC algorithms will not be possible in the foreseeable future.

One obvious solution to this problem is the use less sophisticated models that can be rapidly evaluated. However, there are many astrophysical problems where fast and simple models have only limited accuracy, and the push for sensitive datasets will necessitate increasingly accurate models. Another solution is to speed up the simulations via numerical approximation or computational optimization. For simulations that are borderline MCMC-able, this may prove fruitful and enable a more direct inference. However, in many cases, this still does not make the problem of parameter inference tractable. In the next section, we explore an alternative approach for approximating the posterior with surrogate modeling, also known as computer emulation. Other methods for accelerating the parameter inference problem can be found in approximate inference (e.g. Lintusaari et al. 2016) and posterior optimization (Blei et al. 2016; Seljak & Yu 2019).

2.2 Surrogate Modeling of Computer Simulations

A different approach for estimating the posterior distribution across parameters of a sophisticated model is to employ a technique known as computer emulation, or surrogate modeling. The premise behind an emulator is to construct a fast and flexible function that can learn the mapping between the simulation's input parameters and its output products as a function of the model parameters (Sacks et al. 1989; Kennedy & O'Hagan 2001; Gramacy & Lee 2009). We will define the vector containing the true simulation output as $\boldsymbol{\mu}(\boldsymbol{\theta})$ and its emulator prediction as $\boldsymbol{\mu}_{\rm E}(\boldsymbol{\theta})$. If we can constrain an accurate and fast function for $\boldsymbol{\mu}_{\rm E}$, then by substituting it into the likelihood we can accelerate MCMC convergence by many orders of magnitude. In recent years emulators have gained traction in a number of cosmological parameter inference problems, including the prediction of the non-linear matter overdensity power spectrum (Heitmann et al. 2006; Habib et al. 2007), the CMB angular power spectrum (Fendt & Wandelt 2007; Aslanyan et al. 2015), the weak lensing convergence field (Petri et al. 2015), and the gravitational waveforms of inspiralling compact objects (Field et al. 2014).

Building an emulator is effectively a three step process, involving i) generating a training set;

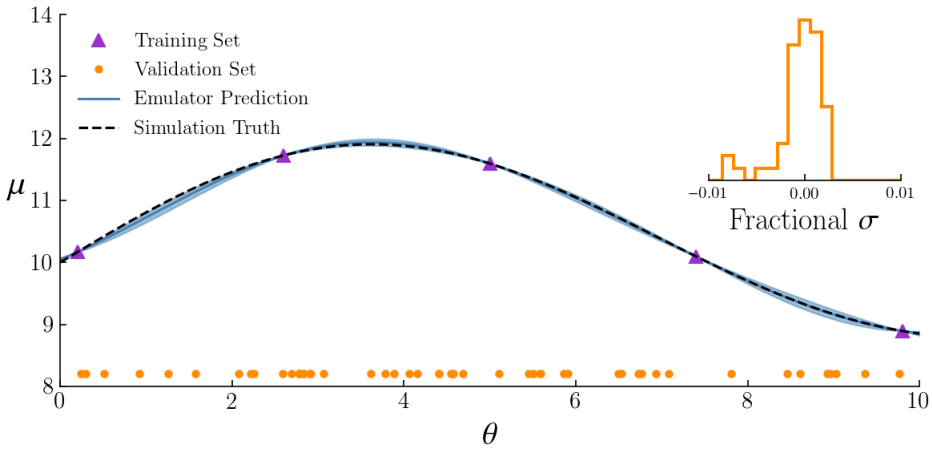

Figure 2.1: Mock emulator construction. We sample the training set (purple triangles) across a single parameter θ and evaluate the model $\mu = 10 + \sin(\theta/2) + \sin(\theta/3)$ at each point. Having trained the emulator, we can reconstruct any point in the parameter space (blue solid), which we compare against the true simulation response (black dashed). A separate cross validation set is drawn (orange) and evaluated with the emulator and the true simulation to assess the accuracy of the emulator, which is shown for this particular example to achieve sub 1% fractional uncertainty. Here we use a Gaussian process regression model, which also returns the expected emulator uncertainty as a function of θ (blue shaded).

ii) choosing a regression model for describing the training set; and iii) cross validating the trained emulator to verify its accuracy. To build an accurate surrogate model, we will need to sample the parameter space densely enough to capture the underlying fluctuations of the simulation output. If we sample too coarsely, we run the risk of missing important features of the simulation output with respect to changes in $\boldsymbol{\theta}$. In a sense, we need to sample at least as densely as the simulation's Nyquist limit across the parameters, or half of the length scale of the simulation's fastest spatial Fourier mode. Of course we do not know this length scale a priori, but we may have some physically-motivated guess, or we can make a fairly computationally cheap estimate by evaluating its first and second derivative at a fiducial point in parameter space. Either way, the training of an emulator is generally an iterative process: laying down a training set, training the emulator, cross validating it, and repeating until our emulator error has reached an acceptable threshold. Figure 2.1 shows an example of this on a mock dataset, demonstrating how an adequate fitting function can be an excellent approximation to a black-box simulation.

Assuming we have a flexible-enough regression model, the accuracy of an emulator is driven primarily by the sampling density of the training set. As such, the emulator error is effectively arbitrarily tunable, and can be decreased simply by generating more training set samples. At some point, however, the generation of a large training set may approach the number of likelihood evaluations that a direct MCMC would have made. Even in this limit, the emulator approach still has a few advantages. The first is that because we define the locations in parameter space to sample the training data up-front, the generation of our training set is an embarrassingly parallel

computation. Unlike MCMC, where the succeeding step depends on the step before it and thus must be run sequentially, we can evaluate the training set simultaneously for all samples. This means that the total time it takes to generate a training set of N samples is still generally less than the time it takes to run an MCMC chain with N iterations. The second is that MCMC chains often need to be re-run. Any time we change our analysis–either due to discovered errors, updates to our data vector or data covariance, or for data quality null tests–MCMC chains need to be re-run. With emulation, however, the heavy computation of generating a training set is needed only once, after which we can re-run our MCMC chains repeatedly with little computational cost.

Emulation techniques can allow for accurate parameter inference in situations where direct MCMC is wildly unfeasible: in many cases bringing factors of 10^3 or more in computational speed-up. However, emulators are not without their drawbacks, and deciding when to use emulator techniques depends on the specific problem at hand. Emulators, like most machine learning techniques, are always limited by the size and extent of their training data. That means when performing parameter inference with flat, uninformative priors, we generally need to still enact hard prior boundaries at the edges of the training set. In such a case, there is the concern that our training set does not actually span the true MAP estimate, which could lead us to a false or local maximum in the posterior distribution. Another difficulty emulators face is generating a sufficient training set in high dimensional spaces: because the hypervolume of a space scales exponentially with its dimensionality, it is extremely difficult to sufficiently sample a high dimensional parameter space blindly. In this case, MCMC algorithms like Hamiltonian Monte Carlo (Betancourt 2017) that are gradient-aware are better suited for high dimensional sampling. A variant of the emulator technique known as Learn-As-You-Go (Aslanyan et al. 2015) strikes this balance well, by starting with direct MCMC as a burn-in step but saves the likelihood calls and generates an emulator on the fly with the stored Markov chains. When the Markov chain sampling density is high enough, it substitutes the simulation with the dynamically compiled emulator, which speeds up the remainder of the MCMC convergence.

Yet another approach for handling high dimensional spaces is to cut out the corners of the hypervolume. For a high dimensional rectangular space, the majority of its hypervolume lies outside the boundaries of a circumscribed hypersphere. This can be done in practice by using a gradient descent approach to find the rough location of $\hat{\boldsymbol{\theta}}_{\rm MAP}$ and then using a Fisher matrix approach for deriving the approximate size and shape of the posterior about that point, which we can then use as a guess for where and how wide in the parameter space we should sample our emulator training set (Schneider et al. 2011).

2.2.1 Propagating Emulator Error into the Likelihood

Although emulator error is in practice nearly arbitrarily tunable, there will always be uncertainty on the emulator prediction, and we should propagate those errors into the likelihood. If the emulator error is on-par with the observational uncertainties of the data (or even a sizeable fraction of it, for that matter), taking the emulator error into account is crucial for ensuring that our inferences are not statistically biased.

The emulated model and the true model can be related via an error term,

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu}_{\mathrm{E}}(\boldsymbol{\theta}) - \boldsymbol{\delta}(\boldsymbol{\theta}). \tag{2.4}$$

Naturally we do not know $\boldsymbol{\delta}$ without evaluating the model at $\boldsymbol{\theta}$, but if we treat it as a Gaussian random variable and have an estimate of its covariance matrix $\boldsymbol{\Sigma}_{\mathrm{E}}$ such that $\boldsymbol{\delta}(\boldsymbol{\theta}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathrm{E}}(\boldsymbol{\theta}))$, we can propagate its effects into the likelihood by marginalizing over it. Note that we defined the emulator error covariance as a function of the model parameters, which can happen say if we have low training set sampling density in a particular region of parameter space, and can be easily computed with certain regression models, like Gaussian processes. We will drop this dependence here for brevity, but in principle it can be retained and the covariance recomputed for every draw of $\boldsymbol{\theta}$.

If we insert this into the log likelihood from Equation 2.3 we get

$$\ln \mathcal{L} \propto -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}}(\boldsymbol{\theta}) + \boldsymbol{\delta})^T \boldsymbol{\Sigma}_{\mathrm{S}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}}(\boldsymbol{\theta}) + \boldsymbol{\delta}), \tag{2.5}$$

where we define $\boldsymbol{\Sigma}_{\mathrm{S}}$ to be the covariance of the data vector, $\boldsymbol{x}$, of some experimental survey. Although we do not know the value of $\boldsymbol{\delta}$ precisely, we can account for its statistical influence by treating it as a parameter of the likelihood and marginalizing over it, yielding the marginal likelihood function. Taking $\boldsymbol{\delta}$ to be zero-mean, its prior distribution can be expressed as

$$\ln \pi \propto -\frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{\Sigma}_{\mathrm{E}}^{-1} \boldsymbol{\delta}. \tag{2.6}$$

Multiplying this into the likelihood and re-arranging we get

$$\begin{aligned} \ln \mathcal{L} + \ln \pi &\propto -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}}(\boldsymbol{\theta}) + \boldsymbol{\delta})^T \boldsymbol{\Sigma}_{\mathrm{S}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}}(\boldsymbol{\theta}) + \boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{\Sigma}_{\mathrm{E}}^{-1} \boldsymbol{\delta} \\ &= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}}(\boldsymbol{\theta}))^T \boldsymbol{\Sigma}_{\mathrm{S}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}}(\boldsymbol{\theta})) - \frac{1}{2}\boldsymbol{\delta}^T (\boldsymbol{\Sigma}_{\mathrm{S}}^{-1} + \boldsymbol{\Sigma}_{\mathrm{E}}^{-1})\boldsymbol{\delta} + (\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{E}})^T \boldsymbol{\Sigma}_{\mathrm{S}}^{-1} \boldsymbol{\delta}, \end{aligned} \tag{2.7}$$

where we factored-out terms proportional to $\boldsymbol{\delta}$ from the likelihood. Integrating over Equation 2.7 can be done analytically using Gaussian integrals, specifically

$$\int \exp\left[-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{b}\boldsymbol{x}\right] d^n \boldsymbol{x} = \sqrt{\frac{(2\pi)^n}{\det \boldsymbol{A}}} \exp\left[\frac{1}{2}\boldsymbol{b}^T \boldsymbol{A}^{-1} \boldsymbol{b}\right], \tag{2.8}$$

where $\boldsymbol{A}$ is an $n \times n$ real, symmetric matrix, and $\boldsymbol{b}$ and $\boldsymbol{x}$ are both vectors of length n. The log marginal likelihood can then expressed as

$$\ln \mathcal{L}_M \propto -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}})^T \left[\boldsymbol{\Sigma}_{\mathrm{S}}^{-1} - \boldsymbol{\Sigma}_{\mathrm{S}}^{-1}(\boldsymbol{\Sigma}_{\mathrm{S}}^{-1} + \boldsymbol{\Sigma}_{\mathrm{E}}^{-1})^{-1} \boldsymbol{\Sigma}_{\mathrm{S}}^{-1}\right] (\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}}), \tag{2.9}$$

which can be further simplified using the Woodbury matrix identity

$$(\boldsymbol{A} + \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \boldsymbol{A}^{-1} \tag{2.10}$$

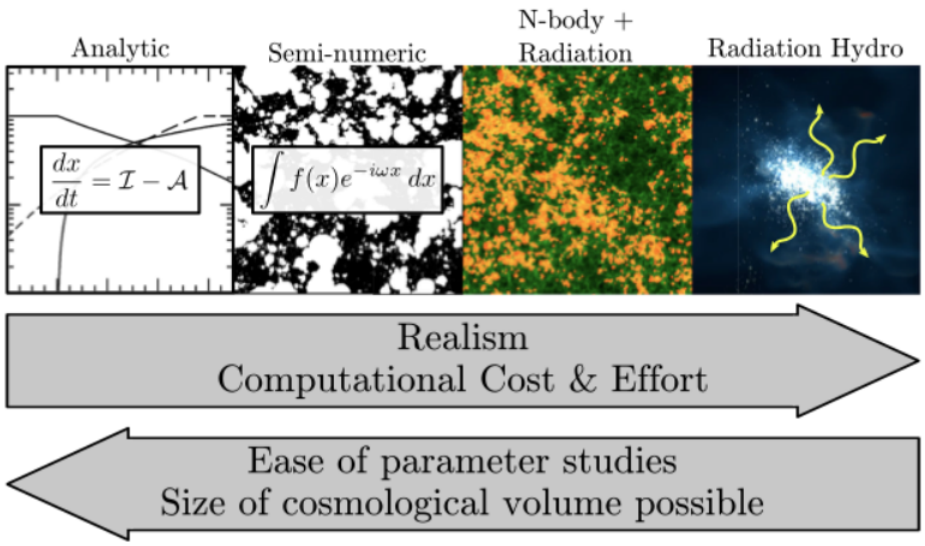

Figure 2.2: A range of techniques exist to model the heating and ionization of the IGM during the EoR, which strike different balances between realism and computational complexity. In this work, we focus on semi-numerical simulations, which are accurate enough to make reliable predictions of 21 cm summary statistics on large scales, and are considerably faster than full hydrodynamic simulations. Figure reproduced from Wise (2019).

to give

$$\ln \mathcal{L}_{\mathrm{M}} \propto -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}})^{T}(\boldsymbol{\Sigma}_{\mathrm{S}} + \boldsymbol{\Sigma}_{\mathrm{E}})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}}). \tag{2.11}$$

This result can also be reached by expressing $(\boldsymbol{x} - \boldsymbol{\mu}_{\mathrm{E}})$ as the sum of two normally distributed random variables $(\boldsymbol{x} - \boldsymbol{\mu})$ and $\boldsymbol{\delta}$. The convolution theorem then tells us that the variance of their sum is the sum of their variances, or $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathrm{S}} + \boldsymbol{\Sigma}_{\mathrm{E}}$. Note that by marginalizing the likelihood over the emulator error term we are effectively inflating the covariance of the likelihood, meaning that we expect the sensitivity of our final parameter constraints to be somewhat degraded. While this is not ideal, it is actually crucial for building confidence in our final parameter constraints. Addison et al. (2016), for example, demonstrated how not accounting for uncertainty on the emulated CMB angular power spectrum can lead to biased and over-constrained posterior distributions. If the emulator error truly is negligible compared to the survey error budget, then it should not noticeably inflate the final parameter constraints.

2.3 Choosing a Cosmic Dawn and EoR Simulation

The fundamental challenge in simulating the EoR is the dynamic range in size scales needed to properly simulate the 21 cm signal from first principles: large scale statistics are the observational

target, necessitating large volumes (Iliev et al. 2014). However, photon sources driving the heating and reionization of the IGM such as early stellar populations, quasars (QSO), and supernovae, are formed at the sub-galactic scale of parsecs to kilo-parsecs. An array of techniques have therefore been developed to study reionization at different levels, oulined in Figure 2.2 (reproduced from Wise (2019)). The most realistic are 3D coupled radiation + hydrodynamic simulations, which self-consistently track gravity, hydrodynamics and radiative transfer on cosmological scales, and are therefore extremely expensive ($\sim 10^6$ core-hours; Semelin et al. 2017). Although it should be noted that even these simulations cannot currently simultaneously resolve giant molecular clouds (the sites of star formation) and span cosmological scales ($L > 200$ Mpc), and therefore need to use some sub-grid approximations to model the production of stars, active galactic nuclei (AGN) and supernovae (Wise 2019). Other methods make the assumption that baryonic matter tracks the dark matter on large scales, and therefore post-process the output of an N-body gravitational simulation with a radiative transfer scheme, again using sub-grid models for photon production, but now also having to explicitly include a model for the galaxy-halo relationship ($\sim 10^5$ core-hours; Mellema et al. 2006). A major improvement in simulation runtime, however, is gained when making controlled numerical approximations that significantly decrease the computational cost of running the cosmological simulation. An example of this is the development of the excursion-set approach for modeling the hydrogen ionization field (Furlanetto et al. 2004), which has been optimized and improved upon to yield a plethora of "semi-numerical codes" that achieve 3D, large-scale EoR simulations at the $\sim 10^0$ core-hour runtime (Zahn et al. 2007; Mesinger & Furlanetto 2007; Santos et al. 2010; Fialkov et al. 2014). While being approximate, these simulations recover the EoR power spectrum at the $\sim 10\%$ level on large scales compared to numerical radiative transfer simulations (Zahn et al. 2011), but yield a factor of 10^5 in computational speed up. Lastly, pure analytic methods for modeling the ionization of hydrogen and 21 cm summary statistics also exist. These methods, while fast and flexible and therefore useful for parameter forecasting, are only realistic under certain conditions and are therefore highly model dependent (McQuinn et al. 2006; Mao et al. 2008).

Currently, the EoR and Cosmic Dawn astrophysical parameter space is only weakly constrained. Using the `21cmFAST` semi-numerical simulation (Mesinger et al. 2011), Greig & Mesinger (2017b) showed that combining observations of the CMB (Planck Collaboration et al. 2016) and the 'dark fraction' of the Lyman-α and Lyman-β forest (McGreer et al. 2015) lead to a rough constraint on the evolution of the hydrogen neutral fraction. This is accompanied by strong degeneracies between the semi-numerical parameters that govern the production of UV photons by EoR galaxies. In certain regimes, such semi-numerical simulations can be used for brute-force MCMC-driven parameter inference. In Greig & Mesinger (2017a), for example, they use `21cmFAST` and an MCMC sampler to forecast the improved constraints on EoR and Cosmic Dawn astrophysics that a fiducial 21 cm power spectrum detection would make. However, such computations spanning both the EoR and Cosmic Dawn epochs of the 21 cm signal, even for semi-numerical simulations, are generally still slow enough to make direct MCMC burdensome. With a semi-numerical approach to modeling EoR radiative transfer, `21cmFAST` can generate a large volume simulation of the 21 cm signal at a single redshift in the matter of minutes. However, when predicting the 21 cm signal across a range of redshifts, particularly at higher redshifts ($z \gtrsim 12$) when fluctuations in the spin temperature are

expected to be non-negligible, the runtime is considerably slower, on the order of hours to tens of hours depending on the simulation resolution, driven largely by the extra computation needed to account for IGM heating from X-ray sources. Nonetheless, Greig & Mesinger (2017a) showed that a brute-force MCMC-based parameter inference is on the cusp of feasibility: using a 200-core cluster over eight days to produce $\sim 8 \times 10^4$ posterior samples. However, this is still burdensome enough to merit acceleration via emulation techniques, especially if these MCMC runs need to be repeated. In this work, we will try to alleviate this by using `21cmFAST` to generate a large training set over a wide range of parameters that we can then use to train an emulator.

EoR Parameters

In `21cmFAST`, the production rate of UV photons is governed by the ionization efficiency of star-forming galaxies, ζ, which can be expressed as

$$\zeta = 30 \left(\frac{f_{\rm esc}}{0.15}\right) \left(\frac{f_\star}{0.1}\right) \left(\frac{N_\gamma}{4000}\right) \left(\frac{2}{1+n_{\rm rec}}\right), \tag{2.12}$$

where $f_{\rm esc}$ is the fraction of produced UV photons that escape the galaxy, $f_\star$ is the fraction of the gas collapsed in dark matter halos that make it into stars, N_γ is the number of ionizing photons produced per stellar baryon and $n_{\rm rec}$ is the average number of times a hydrogen atom in the IGM recombines with a free electron (Furlanetto et al. 2004). The splitting of ζ into these four constituent parameters is merely for clarity: the numerics of `21cmFAST` respond only to a change in ζ, regardless of what sub-parameter sourced that change. These sub-parameters are therefore completely degenerate with each other in the way they affect reionization in `21cmFAST`. Previous works have explored how to parameterize the mass and redshift evolution of ζ (Greig & Mesinger 2015; Sun & Furlanetto 2016). For the time being, we assume ζ to be constant for intuitive purposes. Some of the fiducial values for the terms in Equation 2.12 are physically motivated—$N_\gamma \sim 4000$ is the expectation from spectral models of Population II stars (Barkana & Loeb 2005), and both $f_\star$ and $f_{\rm esc}$ are thought to lie within a few factors of 0.1 (Kuhlen & Faucher-Giguère 2012; Robertson et al. 2015; Paardekooper et al. 2015; Xu et al. 2016; Sun & Furlanetto 2016)—however, these are not strongly constrained at high redshifts and are particularly unconstrained for low-mass halos.

Baryonic matter must cool in order for it to condense and allow for star formation. This can occur through radiative cooling from molecular hydrogen, although this is easily photodissociated by Lyman-Werner photons from stellar feedback (Haiman et al. 1997). Other cooling pathways exist, but in general, low mass mini-halos are thought to have poor star formation efficiencies due to stellar feedback (Haiman et al. 2000). We can parameterize the lower limit on halo mass for efficient star formation as a minimum halo virial temperature, $T_{\rm vir}^{\rm min}$ (K). Here we adopt a fiducial $T_{\rm vir}^{\rm min}$ of 5×10^4 K, above the atomic line cooling threshold of 10^4 K (Barkana & Loeb 2002).

As ionizing photons escape star forming galaxies and propagate through their local HII region, they are expected to encounter pockets of neutral hydrogen in highly shielded sub-structures (Lyman-limit systems). Without explicitly resolving these ionization sinks in the simulation, we can parameterize their effect on ionizing photons escaping a galaxy by setting an effective mean-free path through HII regions for UV photons, $R_{\rm mfp}$. In practice, this sets the maximum bubble size

around ionization sources. Consistent with previous forecasting work, we adopt a fiducial value of $R_{\rm mfp} = 15\,{\rm Mpc}$ (Greig & Mesinger 2017a).

X-ray Spectral Parameters

The sensitivity of the 21 cm power spectrum to cosmic X-rays during the IGM heating epoch may allow us to constrain the spectral properties of the X-ray generating sources. These are theorized to come predominately from either High Mass X-ray Binaries (HMXB) or a hot Interstellar Medium (ISM) component in galaxies heated by supernovae. In `21cmFAST`, the X-ray source emissivity is proportional to

$$\epsilon_X(\nu) \propto f_{\rm X} \left(\frac{\nu}{\nu_{\rm min}}\right)^{-\alpha_{\rm X}}, \tag{2.13}$$

where $f_{\rm X}$ is the X-ray efficiency parameter (an overall normalization), $\alpha_{\rm X}$ is the spectral slope parameter, and $\nu_{\rm min}$ is the obscuration frequency cutoff parameter, below which we take the X-ray emissivity to be zero due to ISM absorption. High-resolution hydrodynamic simulations of the X-ray opacity within the ISM have found that such a power-law model is a reasonable approximation of the emergent X-ray spectrum from the first galaxies (Das et al. 2017). Our fiducial choice of $f_{\rm X} = 1$ corresponds to an average of 0.1 X-ray photons produced per stellar baryon. HMXB spectra have typical spectral slopes $\alpha_{\rm X}$ of roughly 1, while a hot ISM component tends to have a spectral slope of roughly 3 (Mineo et al. 2012). Our fiducial choice of $\alpha_{\rm X} = 1.5$ straddles these expectations. The obscuration cutoff frequency, $\nu_{\rm min}$, parameterizes the X-ray optical depth of the ISM in the first galaxies and is dependent on their column densities and metallicities. We choose a fiducial value of $\nu_{\rm min} = 0.3\,{\rm keV}$, consistent with previous theoretical work (Pacucci et al. 2014; Ewall-Wice et al. 2016a). Because the model assumes the X-ray production comes from star forming halos, the EoR parameter $T_{\rm vir}^{\rm min}$ also affects the spatial distribution of X-ray sources, and is therefore also implicitly an X-ray heating parameter. For a more detailed description of the X-ray numerics in `21cmFAST`, see Mesinger et al. (2013).

Cosmological Parameters

A previous study utilizing a Fisher matrix approach found that even though cosmological parameters have strong constraints from other cosmological probes, such as the *Planck* satellite, their residual uncertainties introduce a non-negligible effect on the 21 cm power spectrum and thus degrade the constraints one can place on astrophysical parameters using 21 cm measurements (Liu et al. 2016). Stated another way, by excluding cosmological parameters from a joint fit, we would be falsely *overconstraining* the astrophysical parameters. Additionally, besides their degradation of astrophysical parameter constraints, we would also like to be able to constrain cosmology with the rich amount of information the 21 cm signal provides us. To do this, we pick $\{\sigma_8, H_0, \Omega_b h^2, \Omega_b h^2, n_s\}$ as our cosmological parameter set. This particular parameterization is selected to match the current `21cmFAST` cosmological inputs and is done merely for convenience. It may be worth investigating in future work if other ΛCDM parameterizations are more suitable

for 21 cm constraints. In terms of `21cmFAST`, all of the chosen parameters play a role in setting the initial conditions for the density field, and $\Omega_b h^2$, $\Omega_b h^2$ and H_0 are furthermore directly related to the definition of the 21 cm brightness temperature. Some of these cosmological parameters also play a role in transforming our observed coordinates on the sky into cosmological distance coordinates. While we do not include these effects into this study, a complete analysis would require such a consideration, which may be addressed in future work. Our fiducial values for the cosmological parameters are (σ_8, h, $\Omega_b h^2$, $\Omega_b h^2$, n_s) = (0.8159, 0.6774, 0.0223, 0.1188, 0.9667), which are consistent with recent *Planck* results (Planck Collaboration et al. 2016). Because σ_8 and H_0 are not directly constrained by *Planck* but are derived parameters in their ΛCDM parameterization, we use the `CAMB` code (Lewis et al. 2000) to map the parameter degeneracies of A_s (the normalization of the primordial perturbation power) and $\theta_{\rm MC}$ (the `CosmoMC` code approximation for the angular size of the sound horizon at recombination; Lewis & Bridle 2002) onto that of σ_8 and H_0 respectively.

2.4 Building a 21 cm Power Spectrum Emulator

2.4.1 Training Set Sampling

Deciding where in our model parameter space to build up our finite number of training samples is called training set "design." The goal in creating a particular training set design is to maximize the emulator's accuracy across the model parameter space, while minimizing the number of samples we need to generate. This is particularly crucial for computationally expensive simulations because the construction of the training set will be the most dominant source of overhead. Promising designs include variants of the Latin-Hypercube (LH) design, which seeks to produce uniform sampling densities when all points are marginalized onto any one dimension (McKay et al. 1979). Previous studies applying emulators to astrophysical contexts have shown LH designs to work particularly well for Gaussian-Process based emulators (Heitmann et al. 2009).

Of particular concern in training set design is the "curse of dimensionality", or the fact that a parameter space volume depends exponentially on its dimensionality. In other words, in order to sample a parameter space at constant density, the number of samples we need to generate depends exponentially on the dimensionality of the space. One way to partially mitigate this is to impose a prior on our parameter space, which allows us to ignore sampling in the corners of the hypervolume where the prior distribution has very small probability. In low dimensional spaces this form of cutting corners only marginally helps us; in two dimensions, for example, the area of a square is only $4/\pi$ greater than the area of its circumscribed circle. In ten dimensions, however, the volume of a hypercube is 400 times that of its circumscribed hypersphere. In eleven dimensions this increases to over a factor of 1000. This means that if we choose to restrict ourselves to a hypersphere instead of a hypercube in an eleven dimensional space, we have reduced the volume our training set needs to cover by over three orders of magnitude. Schneider et al. (2011) investigated the benefits of this technique, and used the Fisher Matrix formalism to inform the size of the hypersphere, which they call Latin-Hypercube Sampling Fisher Sphere (LHSFS). This technique works well in the limit that we already have relatively good prior distributions on our parameters. For parameters that are

weakly constrained, we may need to turn to other mechanisms for narrowing the parameter space before training set construction.

The parameter constraint forecast we present in section 2.5, for example, starts with a coarse rectangular LH design spanning a wide range in parameter values and thousands of training samples. We emulate at a highly approximate level and use the MCMC sampler to roughly locate the region of high probability in parameter space. We supplement this initial training set with more densely-packed, spherical training sets in order to further refine our estimate of the maximum a posteriori (MAP) point (section 2.5). The extent of the supplementary spherical training sets are informed from a Fisher matrix forecast, similar to (Schneider et al. 2011).

2.4.2 Dimensionality Reduction

After constructing a training set, our next task is to decide which simulation data products to emulate over the high dimensional parameter space. Let us collect the various outputs of our simulation into a column vector $\boldsymbol{d}$s, which in our case will be the 21 cm power spectrum Δ^2 over a range of k modes and a redshifts z. Because the power spectra are non-negative quantities, we will hereafter work with the log-transformed data, which makes it easier to regress over them. Suppose our training set consists of $m_{\rm tr}$ samples scattered across parameter space, each having its own data vector. Hereafter, we will index individual data vectors across the training samples $\{1, 2, \ldots, m_{\rm tr}\}$ with upper index j such that the data vector from the $j^{\rm th}$ training sample is identified as $\boldsymbol{d}^j$, located in parameter space at point $\boldsymbol{\theta}^j$. We will also index individual data elements across the data outputs $\{1, 2, \ldots, n\}$ with lower index i, such that the $i^{\rm th}$ data output is identified as d_i. The $i^{\rm th}$ data output from the $j^{\rm th}$ training sample is therefore uniquely identified as d_i^j.

Under the standard emulator algorithm, each data output, d_i, requires its own emulating function or surrogate model. If we are only interested in a handful of outputs, then constructing an emulating function for each data output (i.e., direct emulation) is typically not hard. However, we may wish to emulate upwards of hundreds or thousands of data outputs, say for example the 21 cm power spectrum at dozens of k modes over dozens of individual redshifts, in which case this process becomes sufficiently more complex. One way we can reduce this complexity is to compress our data. Instead of performing an element-by-element emulation of the data vectors, we may take advantage of the correlations between different components of the data vector. For example, with the smoothness of most power spectra, neighboring k and z bins will be highly correlated (example 21 cm power spectra are shown in Figure 2.5). There are thus fewer independent degrees of freedom than there are elements in a data vector, which is the basis behind dimensionality compression techniques like a truncated Principal Component Analysis (PCA).Transforming to the new basis and truncating low-eigenvalue components thus reduces the number of data points that must be emulated (Habib et al. 2007; Higdon et al. 2008).

To construct the principal components, we begin by taking the covariance of our training data, since it captures the typical ways in which the data vary over the parameter space. We also center the data (i.e., subtract the mean) and rescale the data (i.e., divide by a constant) such that the

covariance is given by

$$\boldsymbol{C} \equiv \left\langle \boldsymbol{R}^{-1} \left(\boldsymbol{d} - \overline{\boldsymbol{d}}\right) \left(\boldsymbol{d} - \overline{\boldsymbol{d}}\right)^T \boldsymbol{R}^{-1} \right\rangle \tag{2.14}$$

where $\overline{\boldsymbol{d}}$ is a vector containing the average of each data output across the training set, $\boldsymbol{R}$ is a diagonal $n \times n$ matrix containing our scaling constants, and the outer angle brackets $\langle \dots \rangle$ represent an average over all $m_{\rm tr}$ samples in the training set. The principal components are then found by performing an eigen decomposition of the covariance matrix, given as

$$\boldsymbol{C}\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Lambda}, \tag{2.15}$$

where $\boldsymbol{\Phi}$ is an $n \times n$ matrix with each column representing one of the n orthogonal eigenmodes (or principal components), and $\boldsymbol{\Lambda}$ is a diagonal matrix containing their corresponding eigenvalues. We can think of the eigenmode matrix Φ as a linear transformation from the basis of our centered and scaled data to a more optimal basis, given as

$$\boldsymbol{w}^j = \boldsymbol{\Phi}^T \left[\boldsymbol{R}^{-1}(\boldsymbol{d}^j - \overline{\boldsymbol{d}})\right], \tag{2.16}$$

where $\boldsymbol{w}$ is our data expressed in the new basis. This basis partitions our data into mutually exclusive, uncorrelated modes. Indeed, the covariance of our data in this basis is

$$\langle \boldsymbol{w}\boldsymbol{w}^T \rangle = \boldsymbol{\Lambda}, \tag{2.17}$$

i.e., our eigenvalue matrix from before, which is diagonal. We can rearrange Equation 2.16 into an expression for our original data vector, given as

$$\boldsymbol{d}^j = \overline{\boldsymbol{d}} + \boldsymbol{R}\boldsymbol{\Phi}\boldsymbol{w}^j, \tag{2.18}$$

where because $\boldsymbol{\Phi}$ is real and symmetric, its inverse is equal to its transpose. This gives us insight as to why the $\boldsymbol{w}$ vectors—the data expressed in the new basis—are called the eigenmode weights: to reconstruct our original data, we need to multiply our eigenmode matrix by an appropriate set of weights, $\boldsymbol{w}$, and then undo our initial scaling and centering. We note that our formulation of the eigenvectors through an eigen-decomposition of a covariance matrix is similar to the approach found in Habib et al. (2007); Higdon et al. (2008); Heitmann et al. (2009), who apply singular value decomposition (SVD) directly on the data matrix. In the case when our covariance matrix is centered and whitened (i.e., scaled by the standard deviation of the data), our two methods yield the same eigenvectors.

Although we have expressed our data in a new basis, we have not yet compressed the data because the length of $\boldsymbol{w}^j$, like $\boldsymbol{d}^j$, is n, meaning we are still using n numbers to describe our data. However, one benefit of working in our new basis is that we need not use all n eigenmodes to reconstruct our data vector. If we column-sort the n eigenmodes in $\boldsymbol{\Phi}$ by their eigenvalues, keep those with the top M eigenvalues and truncate the rest, we can approximately recover our original data vector as

$$\boldsymbol{d}^j \approx \overline{\boldsymbol{d}} + \boldsymbol{R}\boldsymbol{\Phi}\boldsymbol{w}^j, \tag{2.19}$$

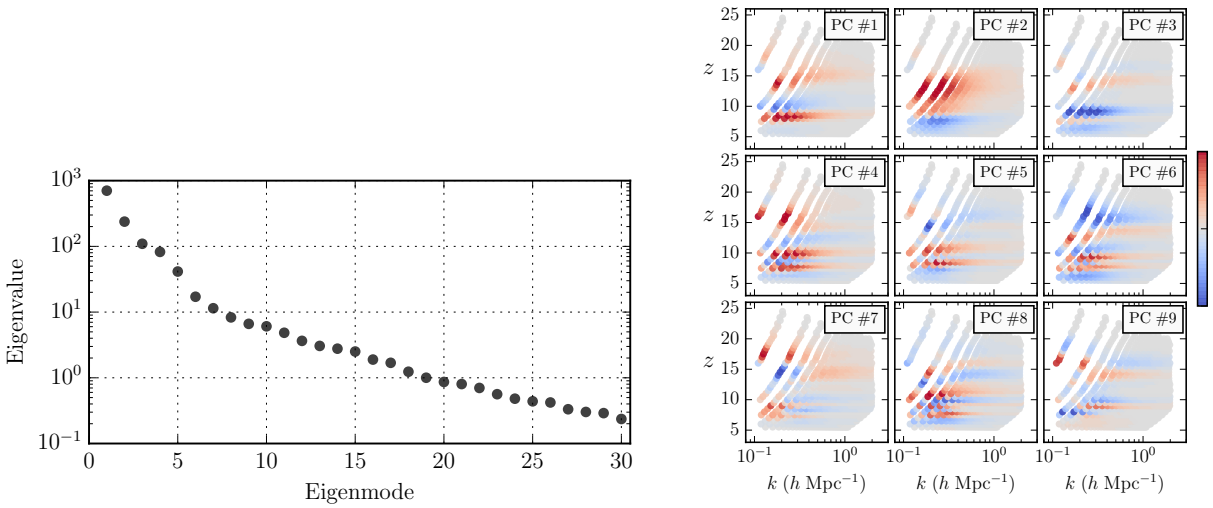

Figure 2.3: **Left**: The first thirty eigenvalues formed from training data of $\ln \Delta^2$. **Right**: The first nine principal components of the power spectrum data at each unique k-z combination. The color scale is artificially normalized to [-1, 1] for easier comparison.

where $\mathbf{\Phi}$ is now defined as the $n \times M$ truncated eigenmode matrix, and $\boldsymbol{w}^j$ is now defined as the length-M column vector where we have similarly sorted and then truncated the weights corresponding to the truncated eigenmodes. *Hereafter, we will use* $\mathbf{\Phi}$ *and* $\boldsymbol{w}$ *to exclusively mean the eigenmode matrix and weight vector respectively after truncation.* Because we are now expressing our data with M numbers where $M < n$, we have compressed our data by a factor of n/M. The precision of this approximation depends on the inherent complexity of the training set and the number of eigenmodes we choose to keep. For our use-case, we typically achieve percent-level precision with an order-of-magnitude of compression ($n/M \sim 10$).

In the case where our scaling matrix, $\boldsymbol{R}$, is the identity matrix, the formalism described above is the standard PCA or Karhunen-Loève transform (KLT). This means that PCA and KLT operate directly on the data covariance matrix formed from our unscaled data. However, not all of the k modes of our power spectrum data will be measured to the same fidelity by our experiment. For the k modes where our experiment will deliver higher precision measurements, our data compression technique should also yield higher precision data reconstructions. To do this, we can incorporate a non-identity scaling matrix, $\boldsymbol{R}$, which can take an arbitrary form such that we produce eigenmodes that are desirable for the given task at hand. A natural choice would be to use the noise (or, in general, the experimental errors) of our instrument. This has the effect of downweighting portions of the data where our measurements will have minimal influence due to larger experimental errors, and conversely upweights the parts of the data with the smallest experimental errors. In the context of our worked example, we also include a whitening term in our scaling matrix, σ_d, which is the standard deviation of the unscaled and centered data. After experimenting with various scaling

matrices, we find a scaling matrix of $R_{ij} = \delta_{ij}\sigma_d^i[\sigma_i/\exp(\overline{d}_i)]^{1/2}$ to work well, where δ_{ij} is the Kronecker delta, σ are the observational errors, and $\exp(\overline{d})$ is the average of the training set data, expressed in linear (not logarithmic) space.

An example set of principal components formed from training data discussed in the following sections is shown in Figure 2.3, where we display the first nine principal components (eigenmodes) of the logged, centered, and scaled Δ^2 training data. We discuss the simulation used to generate this training data in section 2.3. The amplitude of the PCs have been artificially normalized to one for easier comparison. We find in general that at a particular redshift, an individual PC tends to be smooth and positively correlated along k, and at a particular k shows negative and positive correlations across redshift. This is a reflection of the underlying smoothness of the power spectra across k, and the fact that physical processes such as reionization, X-ray heating and Lyman-α coupling tend to produce redshift-dependent peaks and troughs in the power spectrum. The reason why the PCs lose strength at high k is because our rescaling matrix $\boldsymbol{R}$ downweights our data covariance matrix at high k. As we will see in section 2.5, the bulk of a 21 cm experiment's sensitivity to the power spectrum is located at lower k.

2.4.3 Gaussian Process Regression

For the purposes of emulation, we are not interested in merely reconstructing our original training set data at their corresponding points in parameter space $\boldsymbol{\theta}^j$, but are also interested in constructing a prediction of the data vector, $\boldsymbol{d}^{\rm new}$, at any new position in our parameter space, $\boldsymbol{\theta}^{\rm new}$. We can construct a prediction of the data vector at this new point in parameter space by evaluating Equation 2.19 with $\boldsymbol{w}^{\rm new}$; however, we do not know this weight vector a priori. To estimate it at any point in parameter space, we require a predictive function for each element of $\boldsymbol{w}$ spanning the entire parameter space. In other words, we need to interpolate $\boldsymbol{w}$ over our parameter space. To do this, we adopt a Gaussian process (GP) model, which is a highly flexible and non-parametric regressor. See Rasmussen & Williams (2006) for a review on Gaussian Processes for regression.

A GP is fully specified by its mean function and covariance kernel. The GP mean function can be thought of as the global trend of the data, while its covariance kernel describes the correlated fluctuations about the mean trend. In practice, because we center our data about zero before constructing the principal components and their weights (Equation 2.14), we set our mean function to be identically zero. For the covariance kernel we employ a standard squared-exponential kernel, which is fully stationary, infinitely differentiable, produces smooth realizations of a correlated random Gaussian field, and is given in multivariate form as

$$k(\boldsymbol{\theta}, \boldsymbol{\theta}'|\boldsymbol{L}) = \sigma_A^2 \cdot \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}')^T \boldsymbol{L}^{-2} (\boldsymbol{\theta} - \boldsymbol{\theta}')\right], \tag{2.20}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ denote two position vectors in our parameter space, $\boldsymbol{L}$ is a diagonal matrix containing the characteristic scale length of correlations ℓ across each parameter, and σ_A is the characteristic amplitude of the covariance. $\boldsymbol{L}$ is a tunable hyperparameter of the kernel function that must be selected *a priori*. We discuss how we make these choices in section 2.4.3. We set $\sigma_A = 1$ and

therefore it is not a hyperparameter of our kernel. For this to be valid, we must scale the eigenmode weight training data to have variance of unity.

In our case, we have multiple GP regressors—one for each component of the eigenmode weight vector. Consider for example the weight for the first eigenmode. Suppose we group the training data for this weight into a vector $\boldsymbol{y}^{\rm tr}$, such that $y_j^{\rm tr} \equiv w_1^j/\lambda_1^{1/2}$, where λ_i is the variance of weight element $\boldsymbol{w}_i$ from Equation 2.17. Dividing by the standard deviation ensures that the variance of the weights are one, and therefore allows us to set $\sigma_{\boldsymbol{A}} = 1$. If we define an $m_{\rm tr} \times m_{\rm tr}$ matrix $\boldsymbol{K}_1^{\text{tr-tr}}$ such that $(K_1^{\text{tr-tr}})_{ij} \equiv k(\boldsymbol{\theta}_i^{\rm tr}, \boldsymbol{\theta}_j^{\rm tr}|\boldsymbol{L}_1)$, then the GP prediction for the weight at point $\boldsymbol{\theta}^{\rm new}$ is given by

$$w_1^{\rm new} = \lambda_1^{1/2} (\boldsymbol{k}_1^{\text{new-tr}})^T \left[\boldsymbol{K}_1^{\text{tr-tr}} + \sigma_n^2 \boldsymbol{I}\right]^{-1} \boldsymbol{y}^{\rm tr}, \tag{2.21}$$

where $\boldsymbol{k}_1^{\text{new-tr}}$ is a length-$m_{\rm tr}$ vector defined analogously to $\boldsymbol{K}_1^{\text{tr-tr}}$, i.e., $(k_1^{\text{new-tr}})_i \equiv k(\boldsymbol{\theta}^{\rm new}, \boldsymbol{\theta}_i^{\rm tr}|\boldsymbol{L}_1)$, $\boldsymbol{L}_1$ is the matrix containing the hyperparameters chosen a priori for the input training data and the subscript 1 specifies that the input training data are the weights of the first PC mode, w_1. The variance about this prediction is then given by[1]

$$\gamma_1^{\rm new} = 1 - (\boldsymbol{k}_1^{\text{new-tr}})^T \left(\boldsymbol{K}_1^{\text{tr-tr}} + \sigma_n^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_1^{\text{new-tr}}, \tag{2.22}$$

where $\boldsymbol{I}$ is the identity matrix, and σ_n^2 is the variance of random Gaussian noise possibly corrupting the training data from their underlying distribution and is a hyperparameter of the GP (Rasmussen & Williams 2006).

Evaluating Equation 2.21 for each PC weight yields a set of predicted weights that come together to form the vector $\boldsymbol{w}^{\rm new}$. This may then be inserted into Equation 2.19 to yield predictions for the quantities we desire. Similarly, evaluating Equation 2.22 for each PC weight and stacking them into a vector $\boldsymbol{\gamma}^{\rm new}$, we may propagate our GP's uncertainty on $\boldsymbol{w}^{\rm new}$ into an emulator covariance $\boldsymbol{\Sigma}_E$, which describes the uncertainty on the unlogged[2] emulator predictions $\exp(\boldsymbol{d}^{\rm new})$, and is given by

$$(\Sigma_E)_{ij} = \sum_k^M \exp(d_i^{\rm new}) \exp(d_j^{\rm new}) \Phi_{ik} \Phi_{jk} \gamma_k^{\rm new}, \tag{2.23}$$

where in deriving this expression we have assumed that the emulator errors are small. Importantly, note that because $\gamma_k^{\rm new}$ depends on $\boldsymbol{\theta}^{\rm new}$, the same is true for $\boldsymbol{\Sigma}_E$. This is to be expected. For instance, one would intuitively expect the emulator error to be larger towards the edge of our training region than at the center of it. In practice, it is helpful to complement estimates of emulator from Equation 2.23 with empirical estimates derived from cross validation.

So far we have been working towards constructing a set of GP models for each PC mode, each of which is a predictive function spanning the entire parameter space and uses all of the training data.

[1] In principle, one may perform a GP estimate over several points in parameter space at once. Equation 2.21 then predicts an entire vector of $w_1^{\rm new}$ values simultaneously, and Equation 2.22 generalizes to a full covariance matrix. Here we do not employ such a formalism since an MCMC chain explores parameter space one point at a time.

[2] Recall we defined the data vector to be the *logarithm* of the original quantities we wished to emulate.

A different regression strategy is called the Learn-As-You-Go method (Aslanyan et al. 2015). In this method, one takes a small subset of the training data immediately surrounding the point-of-interest, $\boldsymbol{\theta}^{\rm new}$, in order to construct localized predictive functions, which then get thrown away after the prediction is made. This is desirable when the training set becomes exceedingly large ($m_{\rm tr} \gtrsim 10^4$ samples), because the computational cost of GP regression naively scales as $m_{\rm tr}^3$.

Our emulator algorithm in `emupy` relies on code from the Gaussian process module in the publicly-available Python package Sci-Kit Learn,[3] which has an optimized implementation of Equation 2.21 and Equation 2.22 (Pedregosa et al. 2012). Another popular regression model for surrogate modeling are feed-forward artificial neural networks (ANN; e.g. Schmit & Pritchard 2018), which recent works have shown can out perform GP-based regression in the limit of a large training set (Jennings et al. 2019). Recent parameter inference applications to measurements of the 21 cm global signal (Bowman et al. 2018) and LOFAR upper limits on the power spectrum (Mertens et al. 2020) have used such neural-network based emulators (Cohen et al. 2019; Mondal et al. 2020).

GP Hyperparameter Solution

The problem we have yet to address is how to select the proper set of hyperparameters for our GP kernel function, in particular the characteristic scale length of correlations ℓ across each model parameter. We can do this through a model selection analysis, where we seek to find $\boldsymbol{L}$ such that the marginal likelihood of the training data given the model hyperparameters is maximized. From Rasmussen & Williams (2006), the GP log-marginal likelihood for a single PC mode is given (up to a constant) by

$$\ln \mathcal{L}_{\rm M}(\boldsymbol{y}^{\rm tr}|\boldsymbol{L}) \propto -\frac{1}{2}(\boldsymbol{y}^{\rm tr})^T(\boldsymbol{K}^{\rm tr\text{-}tr})^{-1}\boldsymbol{y}^{\rm tr} - \frac{1}{2}\det(\boldsymbol{K}^{\rm tr\text{-}tr}), \tag{2.24}$$

where $\boldsymbol{K}^{\rm tr\text{-}tr}$ has the same definition as in Equation 2.21, and thus carries with it a dependence on $\boldsymbol{\theta}^{\rm tr}$ and $\boldsymbol{L}$. In principle, one could also simultaneously vary the assumed noise variance (σ_n^2) of the target data as an additional hyperparameter and jointly fit for the combination of $[\boldsymbol{L}, \sigma_n^2]$. To find these optimal hyperparameters, we can use a gradient descent algorithm to explore the hyperparameter parameter space of $\boldsymbol{L}$ and σ_n^2 until we find a combination that maximizes $\ln \mathcal{L}_{\rm ML}$. When working with training data directly from simulations, we would expect σ_n^2 to be near-zero; we are not dealing with any observational or instrumental systematics that might introduce uncertainty into their underlying values. Depending on the simulation, there may be numerical noise or artifacts that introduce excess noise or outlier points into the target data, which may skew the resultant best-fit for $\boldsymbol{L}$ or break the hyperparameter regression entirely. This can be alleviated by keeping σ_n^2 as a free parameter and fitting for it and $\boldsymbol{L}$ jointly, or by setting σ_n^2 to a small but non-zero number.

In our eleven dimensional space, this calculation can become exceedingly slow when the number of samples in our training set exceeds ten thousand. In our initial broad parameter space exploration (section 2.5), for example, performing a hyperparameter gradient descent with all 15,000 samples is not attempted. To solve for the hyperparameters, we thus build a separate training set that slices the

[3] `http://scikit-learn.org/`

parameter space along each parameter and lays down samples along that parameter while holding all others constant. We then take this training set slice and train a 1D GP and fit for the optimal ℓ of that parameter by maximizing the marginal likelihood. We repeat this for each parameter and then form our $\boldsymbol{L}$ matrix by inserting our calculated ℓ along its diagonal. This is a method of constraining each dimension's ℓ individually, in contrast to the previous method of constraining ℓ across all dimensions jointly. While this is a more approximate method, it is computationally much faster.

In order to construct a fully hierarchical model, we should in principle not be selecting a single set of hyperparameters for our GP fit, but instead should be marginalizing over all allowed hyperparameter combinations informed by the marginal likelihood. That is to say, we should fold the uncertainty on the optimal choice of ℓ into our uncertainty on our predicted $\boldsymbol{w}^{\rm new}$ and thus our predicted $\boldsymbol{d}^{\rm new}$. In theory this would be ideal, but in practice, this quickly becomes computationally infeasible. This is because the time it takes to train a GP and make interpolation predictions naively scales as the number of training samples $m_{\rm tr}$ cubed. Optimized implementations, such as the one in Sci-Kit Learn, achieve better scaling for low $m_{\rm tr}$, but for large $m_{\rm tr}$ this efficiency quickly drops, to the point where having to marginalize over the hyperparameters to make a single prediction at a single point in parameter space can take upwards of minutes, if not hours, which begins to approach the run time of our original simulation. Furthermore, all of these concerns are exponentially exacerbated in high dimensional spaces. However, in the limit of a high training set sampling density this difference becomes small, which is to say that our marginal likelihood becomes narrow as a function of the hyperparameters. Lastly, and most importantly, we can always turn to diagnose the accuracy of our emulator (and calibrate out its failures) by cross validating it against a separate set of simulation evaluations. In doing so, we can ensure that the emulator is accurate within the space enclosed by our training set.

2.4.4 Emulator Cross Validation

Emulators are approximations to the data products of interest, and as such we need to be able to assess their performance if we are to trust the parameter constraints we produce with them. As discussed above, this can be accomplished empirically via cross validation (CV). In this paper, rather than computing extra simulation outputs to serve as CV samples, we elect to perform k-fold cross validation. In k-fold cross validation, we take a subset of our training samples and separate them out, train our emulator on the remaining samples, cross validate over the separated set, and then repeat this k times. This ensures we are not training on the cross validation samples and also means we do not have to use extra computational resources generating new samples.

We use two error metrics to quantify the emulator performance. The first metric is the absolute fractional error of the emulator prediction over the cross validation data, expressed as $\epsilon_{\rm abs} = ([\Delta^2]_E - [\Delta^2]_{\rm CV})/[\Delta^2]_{\rm CV}$. This gives us a sense of the absolute precision of our emulator. However, not all k modes contribute equally to our constraining power. Because our 21 cm experiment will measure some k modes to significantly higher signal-to-noise (S/N) than other k modes, we want to confirm that our emulator can at least emulate at a precision better than the S/N of our experiment, and is therefore not the dominant source of error at those k modes. The second metric we use is then the offset between the emulator and the CV, divided by the error bars of our 21 cm experiment,

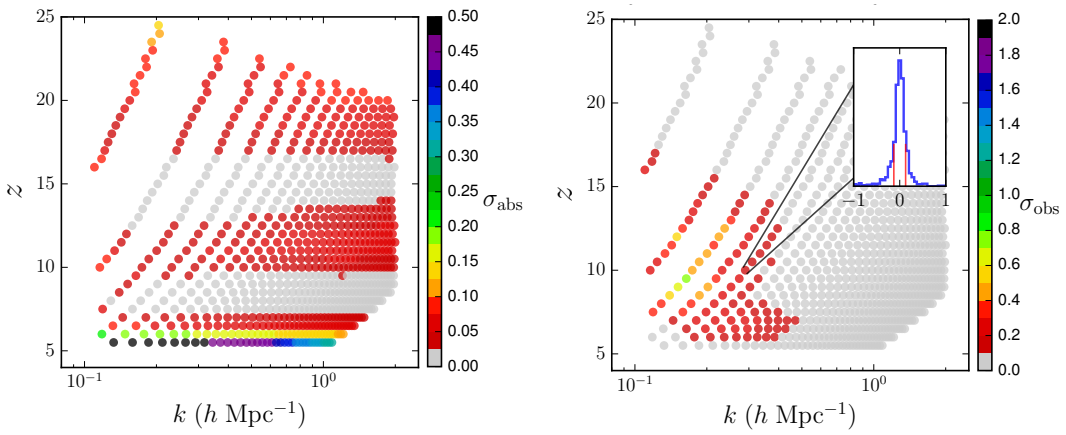

Figure 2.4: **Left**: Standard deviation of the absolute fractional emulator error ($\sigma_{\rm abs}$) with respect to the CV set. Grey color indicates an emulator precision of $\leq 2.5\%$. **Right**: Standard deviation of the offset between emulator prediction and CV data, divided by the experimental errors ($\sigma_{\rm obs}$). The grey color over the majority of the data signifies we can recover the data to $\leq 10\%$ relative to the experimental error bars. **Inset**: Error distribution $\epsilon_{\rm obs}$ for a data output, with its robust standard deviation marked as vertical bars.

given by $\epsilon_{\rm obs} \equiv ([\Delta^2]_E - [\Delta^2]_{\rm CV})/\sigma_S$, where σ_S are the 21 cm power spectrum error bars of our experiment. For this, we use the projected sensitivity error bars of the HERA experiment, discussed in section 2.5 and shown in Figure 2.5.

Cross validation leaves us with an error distribution of the CV samples at each unique k and z. Applying our error metrics, we are left with two sets of error distributions for the emulated power spectra, $\epsilon_{\rm abs}(k,z)$ and $\epsilon_{\rm obs}(k,z)$. To quantify their characteristic widths, we calculate their robust standard deviations, $\sigma_{\rm abs}$ and $\sigma_{\rm obs}$ respectively, using the bi-weight method of Beers et al. (1990). We show an example of these standard deviations in Figure 2.4 having trained our GP emulator on the training data of section 2.5. This training set contains a total of $\sim 5 \times 10^3$ samples. The error distributions are derived by taking half of the samples closest to the origin and performing 5-fold cross validation (i.e. training on 80% and cross validatin on 20%, repeated 5 times).The left panel shows the absolute error metric $\sigma_{\rm abs}$, and demonstrates our ability to emulate at $\leq 5\%$ precision for the majority of the power spectra, and $\leq 10\%$ for almost all of the power spectra. The right panel shows the observational error metric $\sigma_{\rm obs}$, and demonstrates that we can emulate to an average precision that is well below the observational error bars of a HERA-like experiment for virtually all k modes, keeping in mind that the highest S/N k modes for 21 cm experiments are at low-k and low-z for $z \gtrsim 6$. The inset shows the underlying error distribution and its robust standard deviation for one of the power spectra data output. Note that the the distribution of points on the k-z plane that are shown in Figure 2.4 are not determined by the emulator: one can easily emulate the power spectra at different values of k and z. Instead, these points were chosen to match the observational survey parameters and our choice of binning along our observational bandpass, which is discussed in more detail in section 2.5.

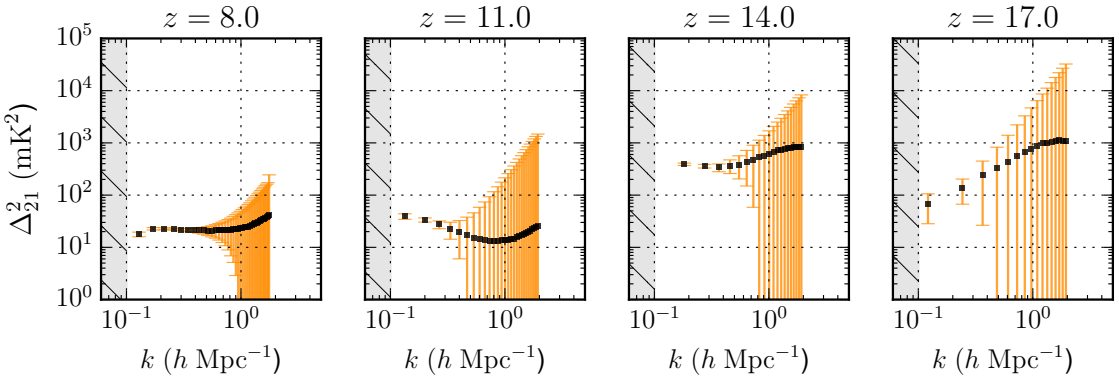

Figure 2.5: A mock observation of the 21 cm power spectrum created from an underlying "truth" realization of `21cmFAST` with error bars corresponding to the projected sensitivity of the HERA331 experiment after a single observing season. The grey-hatched region to the left denotes inaccessibility due to foreground dominated k modes. Although we display only four redshifts, the entire mock observation contains the 21 cm power spectrum from $5 < z < 25$ in steps of $\Delta z = 0.5$.

The observational error metric is of course dependent on the chosen 21 cm experiment and its power spectrum sensitivity. This particular emulator design, for example, may not be precise enough to emulate within the error bars of a futuristic experiment. If we need to boost our emulator's precision, we can do so to an almost arbitrary level by simply generating more training samples and packing the space more densely. The limiting factors of this is the need to generate an additional number of training samples which is unknown a priori, and the intrinsic $m^3_{\rm tr}$ scaling of the Gaussian Process regressor. However, with sufficient computational resources and novel emulation strategies like Learn-As-You-Go, increasing the emulator's precision to match an arbitrary experimental precision is in principle feasible.

2.5 Mock HERA Parameter Constraint Forecast

Mock HERA Observations

To create a mock 21 cm power spectrum observation, we run `21cmFAST` with "true" model parameters set to the fiducial values described in section 2.3. For this mock observation, the initial conditions of the density field are generated with a different random seed than what was used to construct the training set realizations.

Uncertainty in the 21 cm power spectrum at the EoR comes from three main sources, (i) thermal noise of the instrument, (ii) uncertainty in foreground subtraction, and (iii) sampling (or cosmic) variance of our survey. For portions of Fourier space that are clean of foregrounds, the variance of the power spectrum at an individual k mode from the two remaining sources of uncertainty can be

written as

$$\sigma^2(k) \approx \left[X^2 Y \frac{k^3}{2\pi^2} \frac{\Omega'}{2t} T_{\rm sys}^2 + \widehat{\Delta_{21}^2}(k) \right]^2 , \tag{2.25}$$

where the first term is the thermal noise ($k = |k|$), and the second term is the sampling variance uncertainty at each individual k mode (Pober et al. 2013b). In the first term, X^2Y are scalars converting angles on the sky and frequency spacings to transverse and longitudinal distances in h Mpc^{-1}, and Ω' is the ratio of the square of the solid angle of the primary beam divided by the solid angle of the square of the primary beam (Parsons et al. 2014). The total amount of integration time on a particular k mode is t, and $T_{\rm sys}$ is the system temperature taken to be the sum of a receiver temperature at 100 K and a sky temperature at $60\lambda^{2.55}$ K, where λ has units of meters (Parsons et al. 2014). To compute the variance on the 1D power spectrum, Var$[\Delta^2(k)]$, from the above variances on the 2D power spectrum, we bin into annuli of constant scalar k and add the variances reciprocally (Pober et al. 2013b).

We perform these calculations with the public Python package `21cmSense`,[4] which takes as input a specification of the interferometer design and survey parameters (see Parsons et al. 2012a; Pober et al. 2014). We assume a HERA-like instrument with a compact, hexagonal array of 331 dishes that each span 14-m in diameter (Dillon & Parsons 2016; DeBoer et al. 2017). We further assume the observations are conducted for 6 hours per day spanning a 180 day season for a total of 1080 observing hours. Within an instrumental bandpass spanning 50-250 MHz, we construct power spectra from $5 < z < 25$ in co-evolution redshift bins of $\Delta z = 0.5$. We also adopt the set of "moderate" foreground assumptions defined in `21cmSense`. This assumes that in a cylindrical Fourier space decomposed into wavenumbers perpendicular ($k_\perp$) and parallel ($k_\parallel$) to the observational line-of-sight, the foreground contaminants are limited to a characteristic "foreground wedge" at low $k_\parallel$ and high $k_\perp$(see e.g., Datta et al. 2010; Morales et al. 2012; Trott et al. 2012; Parsons et al. 2012b; Pober et al. 2013a; Liu et al. 2014a,b). One then pursues a strategy of foreground avoidance (rather than explicit subtraction) under the approximation that outside the foreground wedge there is a foreground-free "EoR window." To be conservative, we impose an additional buffer above the foreground wedge of $k_\parallel = 0.1\ h\,\mathrm{Mpc}^{-1}$ to control for foreground leakage due to inherent spectral structure of the foregrounds. The selection of this buffer is motivated by observations of Pober et al. (2013a), who made empirical measurements of the foreground wedge as seen by the PAPER experiment at redshift $z \sim 8.3$. For our sensitivity calculations, we impose a constant buffer at all redshifts, even though one would expect foreground wedge leakage to evolve with redshift just as the wedge itself evolves with redshift. Intuitively, we expect foreground leakage to have the same redshift dependence as the wedge: at higher redshifts foreground leakage reaches to higher $k_\parallel$ because the power spectrum window functions become more elongated along the $k_\parallel$ direction (see Liu et al. 2014a). This means that our assumed buffer of $k_\parallel = 0.1\ h$ Mpc^{-1} is over-conservative for $z < 8.3$ and under-conservative for $z > 8.3$.

We note that the sensitivity projections from `21cmSense` are assumed to be uncorrelated across k, meaning that our Σ_S is diagonal. While this is not strictly true for a real experiment it is often

[4] https://github.com/jpober/21cmSense

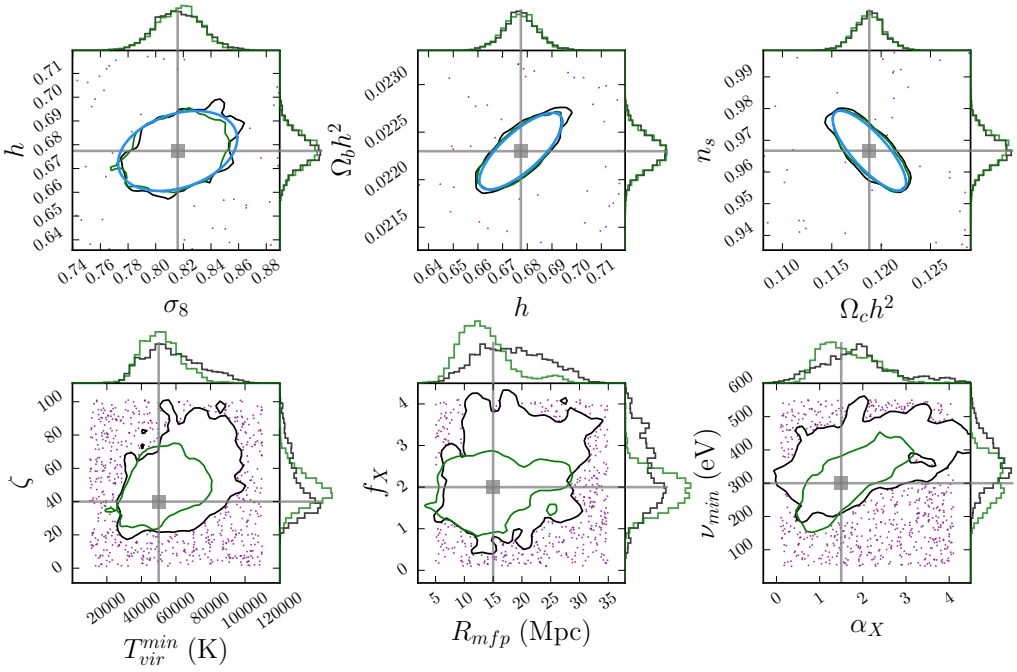

Figure 2.6: Posterior constraints for our initial parameter space exploration. The black contours represent 95% posterior credibility after emulating over our rectangular Latin Hypercube (LH) design training set (shown as purple points). The green contours represent 95% posterior credibility after emulating over the LH training set plus a second, spherical training set populated within the contours of the initial constraints. The blue ellipses over the cosmological parameters show the 95% probability contour of our *Planck* prior distribution. The grey square shows the true underlying parameters of the observation. The histograms adjacent to the contour plots show the marginalized posterior distribution across each model parameter.

an assumption made in parameter constraint forecast studies, with the reasoning that via careful binning in the $u - v$ plane informed by the extent of the telescope's primary beam response, one can minimize the correlation between different $u - v$ modes. For more detailed discussions of foreground avoidance and subtraction techniques for 21 cm interferometers, we defer the reader to Pober et al. (2014). Figure 2.5 shows the resulting sensitivity projection produced by applying the above calculations to our truth `21cmFAST` realization.

Broad Parameter Space Search

To produce parameter constraints with an emulator, we must first construct a training set spanning the regions of parameter space we would like our MCMC sampler to explore. Due to the finite size of any training set, we need to set hard limits *a priori* on the breadth of the training set in parameter space. Our prior distribution on the model parameters is a straightforward way to make this choice. The astrophysical parameters of the EoR and EoH, however, are highly

unconstrained and in some cases span multiple orders of magnitude. In order to fully explore this vast parameter space with the emulator, we are left with a few options: (i) we could construct a sparse and wide training set, emulate at a highly approximate level, MCMC for the posterior distribution and then repopulate the parameter space with more training samples in the region of high probability and repeat, or (ii) use a gradient descent method to locate the general location of maximum probability. Both require direct evaluations of the simulation, but the former can be done in batch runs on a cluster and the latter is a sequential, iterative process (although it is typically not as computationally demanding as a full MCMC). For this work, we choose the former, and construct a wide rectangular training set from a Latin Hypercube design consisting of 15×10^3 points. For one parameter in particular, f_X, we do not cover the entire range of its currently allowed values. In order to exhaustively explore the prior range of f_X, one might consider performing an initial gradient descent technique to localize its value. Because gradient descent algorithms are common in the scientific literature, we do not perform this test and assume we have already localized the value of f_X to within the extent of our initial training set, or assume we are comfortable adopting a prior on f_X spanning the width of our initial training set.

We use this initial training set to solve for an estimate of the hyperparameters for our Gaussian Process kernel as detailed in section 2.4.3. With a training set consisting of over 10^4 points, we do not solve for a global predictive function of the eigenmode weights, but use a variant of the Learn-As-You-Go algorithm described in subsection 2.4.3 for emulation. We k-fold cross validate on the training set and find that we can emulate the power spectra to accuracies ranging in the 50%-100% level depending on the redshift and k mode. While this is by no means "high-precision" emulation (and will pale in comparison to the precision achieved in our final emulator runs for producing the ultimate parameter constraints), it is enough to refine our rough estimate of the location of the MAP point. We incorporate these projected emulator errors into our likelihood as described in subsection 2.2.1. We adopt flat priors over the astrophysical parameters and covarying priors on the cosmological parameters representative of the *Planck* base TT+TE+EE+low-ℓ constraint, whose covariance matrix can be found in the `CosmoMC` code (Lewis & Bridle 2002).

We show the results of our initial parameter space search in Figure 2.6, where the black contours represent the 95% credible region of our constraint and the histograms show the posterior distribution marginalized across all other model parameters. The purple points in Figure 2.6 show samples from the initial LH training set, demarcating its hard bounds. The blue contours on the cosmological parameters show the 95% credible region of the prior distribution, showing that at this level of emulator precision the posterior distribution across the cosmology is dominated by the strong *Planck* prior. Even while emulating to a highly approximate level, we find that we can recover a rough and unbiased localization of the underlying truth parameter values. After localization, we can choose to further refine the density of our training set to produce better estimates of the MAP point and ensure we are converging upon the underlying true parameters. To do this, we extend the training set with an extra 5,000 samples sampled from a spherical multivariate Gaussian located near the truth parameters with a size similar to the width of the posterior distribution. The 95% credible region of the parameter constraints produced using an updated training set of 20,000 samples is shown in Figure 2.6 as the green contours, which shows an improvement in the MAP localization.

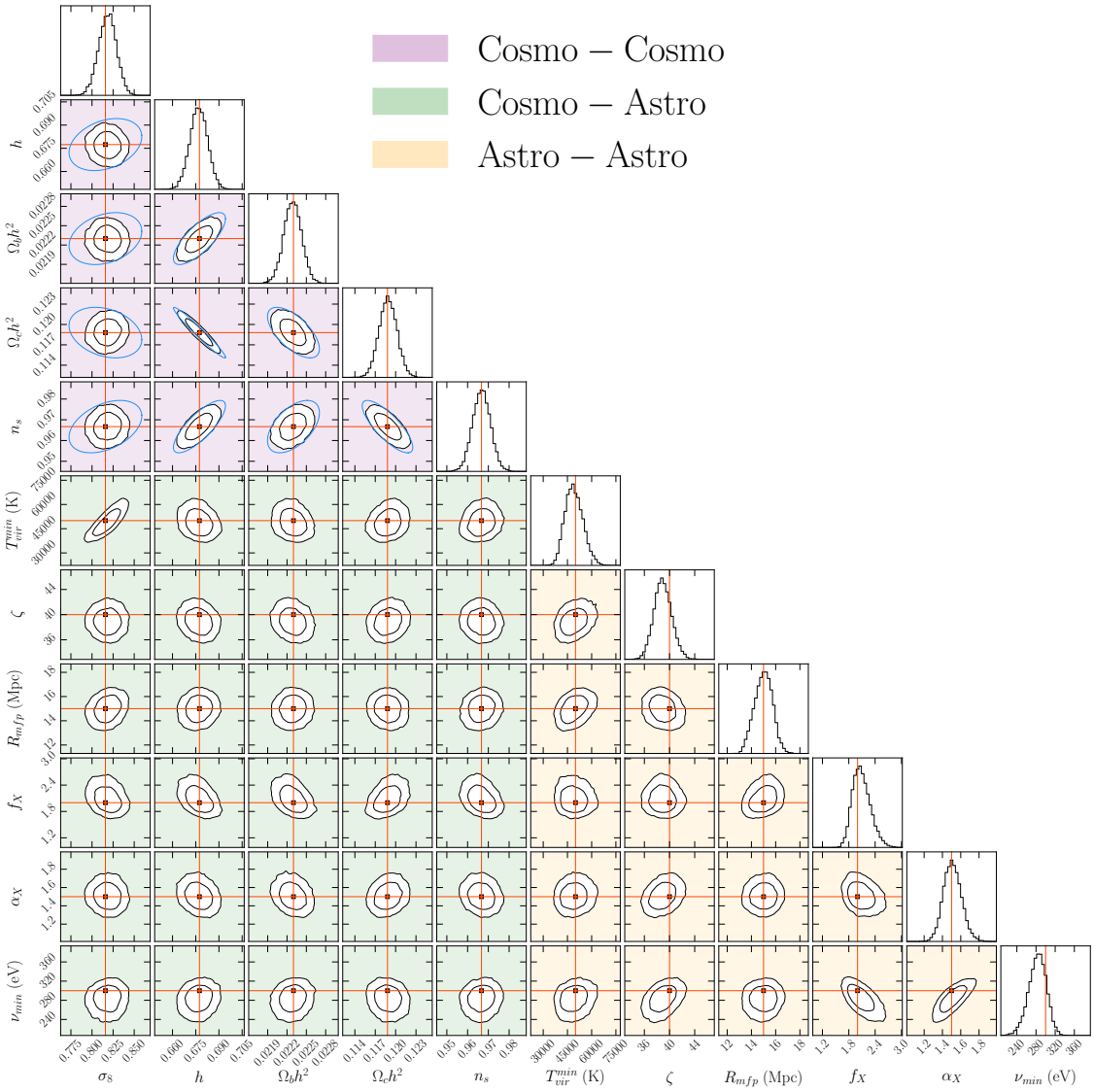

Figure 2.7: The joint posterior distribution of the eleven-parameter model, showing the 68% and 95% credible regions of the pairwise covariances (off-diagonal) and their marginalized distribution across each model parameter (diagonal). Purple-shaded boxes represent pairwise covariances between cosmological parameters; green-shaded boxes represent cosmological-astrophysical covariances, and yellow-shaded boxes represent astrophysical covariances. Blue contours on the cosmological covariances indicate the 95% credible region of the adopted prior distribution consistent with *Planck*. The underlying true parameters of the observation are marked as red squares with crosshairs.

Posterior Characterization

With a reasonable estimate of the MAP location in hand, we now construct a dense training set so that we may emulate to higher precision. To do this, we build another training set consisting of 5000 samples from a packed Gaussian distribution and 500 samples from a LHSFS design (see subsection 2.4.1) with a location near the truth parameters and size similar to the posterior distribution found in section 2.5. To assess the accuracy of the emulator trained over this training set, we 5-fold cross validate over a subset of the data in the core of the training set. The results can be found in Figure 2.4, which shows we can emulate the power spectra to an accuracy of $\sim 10\%$ for most of the data. More importantly, however, Figure 2.4 shows that the emulator error is always lower than the inherent observational survey error, and for the majority of the data is considerably lower. Nonetheless, we account for emulator errors by adding them in quadrature with the survey error bars as described in subsection 2.2.1. Our MCMC run setup involves 300 chains each run for ~5,000 steps, yielding over 10^6 posterior samples. On a MacPro Desktop computer, this entire calculation takes $\sim$ 12 hours and utilizes $\sim$ 10 GB of memory.

The final characterization of the posterior distribution is found in Figure 2.7, where we show its marginalized pairwise covariances between all eleven model parameters and its fully marginalized distributions along the diagonal. With the exception of σ_8, the cosmological constraints are mostly a reflection of the strong *Planck* prior distribution (shown as blue contours). Compared to previous EoR forecasts of Pober et al. (2014); Ewall-Wice et al. (2016a); Greig et al. (2016), the strength of the EoR parameter degeneracies are weakened due to the inclusion of cosmological physics that washes out part of the covariance structure. This importance is exemplified by the strong degeneracy between the amplitude of clustering, σ_8, and the minimum virial temperature, $T_{\rm vir}^{\rm min}$. At a particular redshift, an increase in σ_8 increases the number of collapsed dark matter halos. At the same time, an increase in $T_{\rm vir}^{\rm min}$ suppresses the number of collapsed halos that can form stars, meaning they balance each other out in terms of their effect on the number of star forming halos present at any particular redshift. Compared to the recent work of Greig & Mesinger (2017a), who performed a full MCMC over EoR and EoH parameters with `21cmFAST` assuming a HERA-331 experiment, our constraints are slightly stronger. This could be for a couple of reasons, including (i) the fact that they add an additional 20% modeling error onto their sampled power spectra and (ii) their choice of utilizing power spectra across 8 redshifts when fitting the mock observation, compared to our utilization of power spectra across 37 different redshifts when fitting to our mock observation.

The posterior distributions for each parameter marginalized across all others are shown in Figure 2.8, where they are compared against their input prior distributions. We see that the HERA-331 experiment, with a moderate foreground avoidance scheme, will nominally place strong constraints on the EoR and EoH parameters of `21cmFAST` with respect to our currently limited prior information. For the cosmological parameters, the HERA likelihood alone is considerably weaker than the *Planck* prior; however, we can see that a HERA likelihood combined with a *Planck* prior can help strengthen constraints on certain cosmological parameters. Because 21 cm experiments are particularly sensitive to the location of the redshift peaks of the 21 cm signal,[5] parameters

[5]Strong peaks and dips in the z evolution of Δ^2 mean that slight deviations along z produce large deviations in Δ^2.

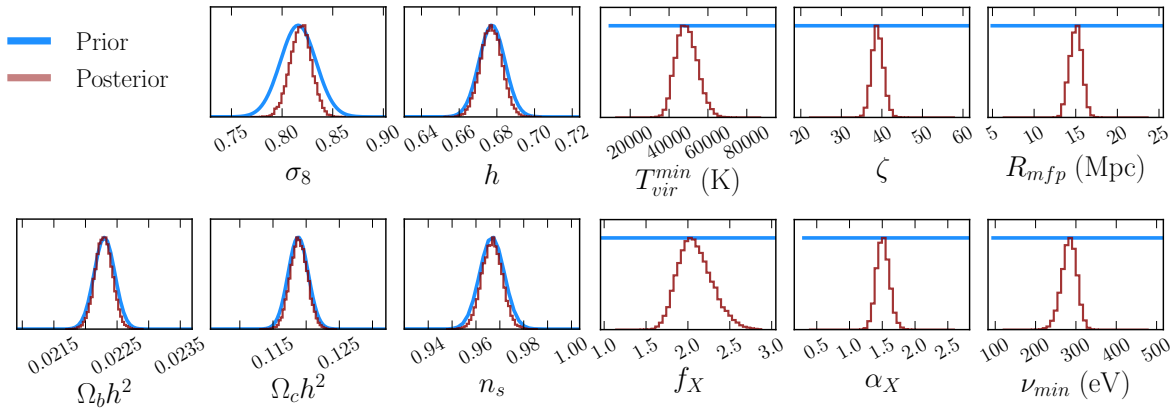

Figure 2.8: The posterior distribution of Figure 2.7 for each model parameter marginalized across all other parameters, compared against the adopted prior distributions. We adopt priors on the cosmological parameters consistent with *Planck* constraints, and adopt flat priors across the astrophysical parameters. We find that HERA will be able to produce $\sim$ 10% level constraints on the astrophysical parameters and will help strengthen constraints on σ_8.

like σ_8, which control the overall clustering and thus affect the timing of reionization, are more easily constrained. Going further, Liu et al. (2016) showed that one can produce improved CMB cosmological parameter constraints by using 21 cm data to constrain the prior range of τ, which is a CMB nuisance parameter that is strongly degenerate with σ_8 and thus degrades its constraining power. Our 21 cm power spectrum constraint on σ_8 shown above does not include this additional improvement one can achieve by jointly fitting 21 cm and CMB data, which is left for future work.

2.6 Validation Tests

Here we discuss performance tests that help to further validate the emulator performance. In particular, we address the issue of what happens when the underlying true parameters lie at the edges or outside of the hard bounds of our training set, and make a direct comparison of the constraints produced by our emulator and a traditional brute-force MCMC algorithm.

2.6.1 Comparison Against Direct MCMC

An important performance test of the emulator algorithm is to compare its parameter constraints against the constraints produced by brute-force, where we directly call the simulation in our MCMC sampler. Of course we cannot do this for the simulation we would like to use—hence the need for the emulator—but we can do this if we use a smaller and faster simulation. For this test, we still use `21cmFAST` but only generate the power spectra at $z = \{8, 9, 10\}$ and ignore the spin temperature contribution to Δ^2, which drastically speeds up the simulation. In addition, we use a

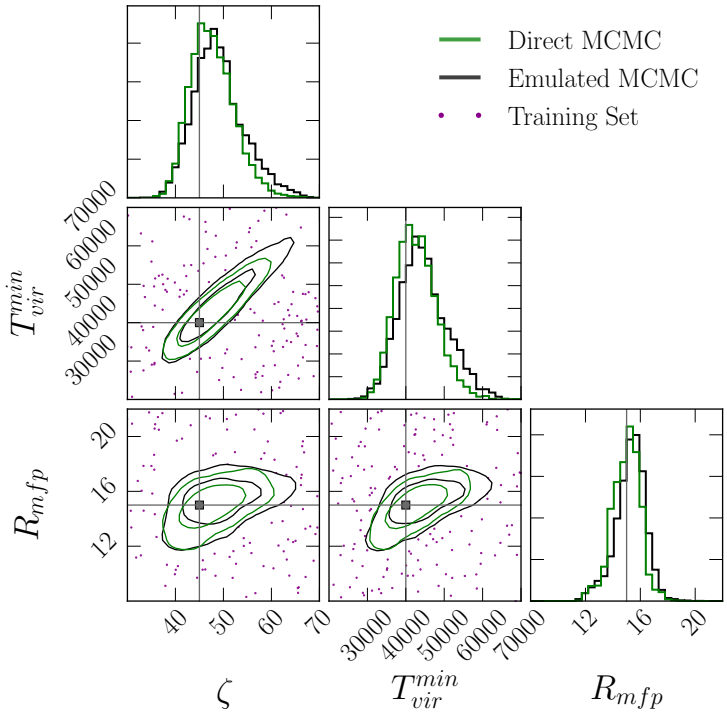

Figure 2.9: Emulator performance test comparing the constraints from the emulator (black) against brute-force constraints which directly evaluate the simulation (green). Both are able to produce unbiased constraints on the underlying "truth" parameters of the mock observation (square). The training set samples used to construct the emulator are shown in the background (purple points).

smaller simulation box-size and use a coarser resolution which yields additional speed-ups. We also restrict ourselves to varying only the three EoR astrophysics parameters described in section 2.3, meaning we achieve faster MCMC convergence. Using a coarser resolution and ignoring spin temperature fluctuations means the simulation is less accurate, but for the purposes of this test the simulation accuracy is irrelevant: we merely want to gauge if the emulator induces significant bias into constraints that we would otherwise produce by directly using the simulation.

Our mock observation is constructed using a realization of `21cmFAST` with fiducial EoR parameters $(\zeta, T_{\rm vir}^{\rm min}, R_{\rm mfp}) = (45, 4\times10^3\ {\rm K}, 15\ {\rm Mpc})$, and with the same fiducial cosmological parameters of section 2.5. We place error bars over the fiducial realization using the same prescription of that described in section 2.5, corresponding to the nominal sensitivity projections for the HERA331 experiment under "moderate" foreground avoidance. Similar to section 2.5, we fit the power spectra across $0.1 \leq k \leq 2\ h\ {\rm Mpc}^{-1}$ in our MCMC likelihood function calls.

The result of the test is shown in Figure 2.9, where we plot the emulator and brute-force posterior constraints, as well as the training set samples used to construct the emulator. We find that the emulator constraints are in excellent agreement with the constraints achieved by brute-force. In the case where the emulator constraints slightly deviate from the brute-force constraints (in this

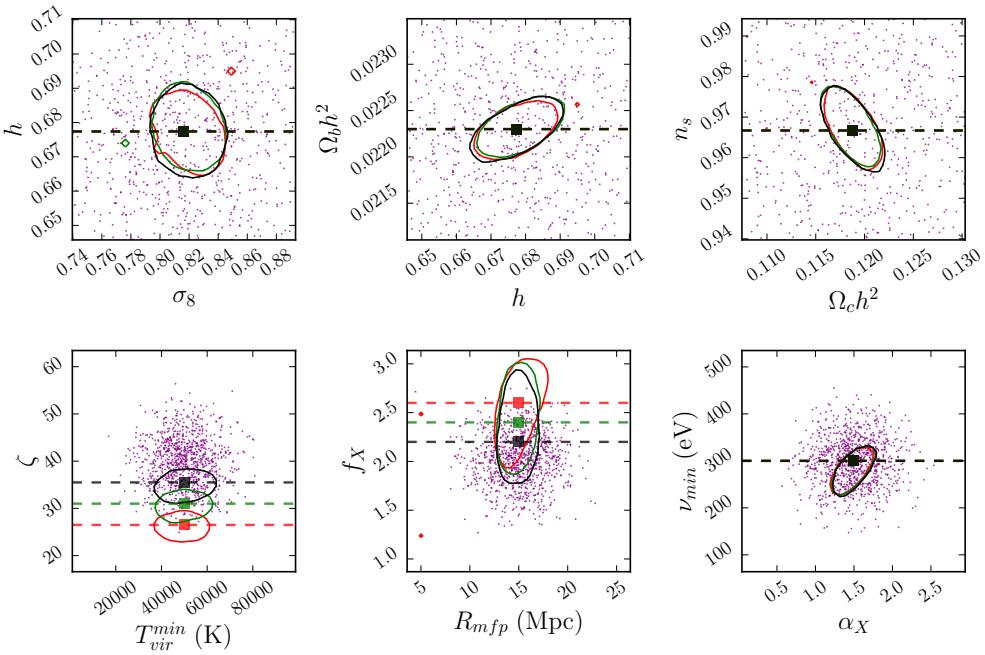

Figure 2.10: 95% credible regions of the posterior distribution while moving the true parameters of the mock observation away from the center of the training set, demonstrating the ability of the emulator to recover unbiased MAP constraints even when the training set does not directly overlap with the underlying "truth" parameters.

case high ζ and high $T_{\rm vir}^{\rm min}$), the emulator deviations are conservative relative to the brute-force contours. In other words, the emulator constraints are always equal to or broader than the brute-force constraints, and do not falsely over-constrain the parameter space or induce systematic bias into the recovered MAP.

2.6.2 Training Set Miscentering

The ability of the emulator to produce reliable parameter constraints hinges principally on the assumption that the true parameter values lie within the bounds of the training set. If this is not the case, the emulator cannot make accurate predictions of the simulation behavior and is making a best guess based on extrapolation. In the case that emulator errors are not accounted for, this can lead to artificial truncation of the posterior distribution and create a false, over-constraining of the parameter space. This was observed to be problematic for a small number of figures in the 2015 *Planck* papers. Though the underlying cosmological constraints were unaffected, some illustrative plots employed an emulator-based method that seemed to be in tension with a more accurate direct MCMC method because the underlying parameters lay outside of the emulator's training set (Addison et al. 2016). It is therefore crucial to be able to assess if our training set encompasses the underlying truth parameters or if the training set has been miscentered. If the

emulator can alert us when this is the case, we can repopulate a new training set in a different location and have greater confidence that the emulator is not falsely constraining the parameter space due to the finite width of the training set.

Given our method in section 2.5 for localizing the parameter space via a sequence of training sets that iteratively converge upon the general location of the underlying true parameters, it is natural to ask, what if we made our final, compact training set a little too compact and missed the underlying MAP? How can we assess if this is the case, and if so, where do we populate the new training set? The most straightforward answer is to look at the posterior constraints compared to the width of the training set: if the posterior constraints run-up against the edge of the training set significantly, this may be an indication that we need to move the training set in that direction.

We perform such a test using our final compact training set and shift the position of mock observation's underlying truth parameters to the edges of the training set for parameters ζ and $f_{\rm X}$: two particularly unconstrained parameters. Figure 2.10 shows the result, demonstrating the emulator's ability to shift the posterior contours when it senses that the MAP lies at the edge of the training set. In this case, we would know to generate more training samples near the region of high probability and retrain our emulator.

Chapter 3

HERA Instrument Calibration

In this section we discuss techniques and challenges related to the calibration of 21 cm data, drawing primarily from Kern et al. (2020a). First, we present a sky-based calibration pipeline for HERA and demonstrate it on Phase I data. We show where systematic uncertainties arise in the process and present a method for mitigating them. Next, we use this pipeline to show how one can fill-in the missing components of the gain solutions leftover after just redundant calibration, one of the factors motivating HERA's redundant design. Demonstrating this on Phase I data, we show the impact these techniques and considerations have on measurements of the 21 cm power spectrum under a foreground avoidance strategy.

3.1 Introduction

Over the past decade, first-generation 21 cm EoR experiments like the Donald C. Backer Precision Array for Probing the Epoch of Reionization (PAPER; Parsons et al. 2014; Jacobs et al. 2015; Cheng et al. 2018; Kolopanis et al. 2019), the Murchison Widefield Array (MWA; Tingay et al. 2013; Dillon et al. 2014; Ewall-Wice et al. 2016b; Beardsley et al. 2016; Barry et al. 2019b; Li et al. 2019), the Low Frequency Array (LOFAR; van Haarlem et al. 2013; Patil et al. 2017; Gehlot et al. 2018), the Giant Metre Wave Radio Telescope (GMRT; Paciga et al. 2013), and the Long Wavelength Array (LWA; Eastwood et al. 2019) have set increasing stringent limits on the Cosmic Dawn 21 cm power spectrum. Meanwhile, global signal experiments have placed constraints on the 21 cm monopole (Bernardi et al. 2016; Singh et al. 2017), with a reported first detection of the signal at Cosmic Dawn from the Experiment to Detect the Global EoR Signature (EDGES; Bowman et al. 2018). 21 cm experiments face the challenge of separating-out the weak cosmological signal from galactic and extra-galactic foreground emission that is many orders of magnitude brighter. However, the 21 cm signal is expected to be highly spectrally variant due to inhomogeneities in the density, ionization state and temperature of the IGM along the line-of-sight, while non-thermal foreground emission is expected to be spectrally smooth. This provides a means for separating foreground emission from the desired cosmological signal. However, even small instrumental effects can distort these foregrounds and contaminate the region in Fourier space

occupied nominally only by the EoR signal and thermal noise, known as the EoR window (Morales et al. 2012). High dynamic range instrumental gain calibration is therefore critical to 21 cm science.

Per-antenna gain calibration is the task of solving for a single complex number per antenna and feed polarization (as a function of both time and frequency) that best satisfies the antenna-based calibration equation for a visibility V_{ij} defined between antenna i and antenna j,

$$V_{ij}^{\rm measured}(\nu, t) = V_{ij}^{\rm true}(\nu, t) g_i(\nu, t) g_j^*(\nu, t), \tag{3.1}$$

where $V_{ij}^{\rm measured}$ is the raw data, $V_{ij}^{\rm true}$ is the true visibility that would be measured by an uncorrupted instrument, and g_i and g_j are the instrumental gains for antenna i and j, respectively (Hamaker et al. 1996). Recent work has shown how incomplete models in sky-based calibration (Barry et al. 2016; Ewall-Wice et al. 2017; Byrne et al. 2019) and non-redundancies in redundant calibration (Joseph et al. 2018; Orosz et al. 2019) can lead to gain calibration errors that contaminate the EoR window. Foreground and instrument simulations for HERA indicate that the fiducial EoR signal at $k \sim 0.2\ h\,{\rm Mpc}^{-1}$ is expected to be roughly 10^5 times weaker than the peak foreground amplitude at $k \sim 0\ h\,{\rm Mpc}^{-1}$ in the visibility (Thyagarajan et al. 2016). Because gain calibration is multiplicative in frequency it can equivalently be thought of as a convolution in delay space, the Fourier dual of frequency. This means that each antenna's gain kernel, or the gain's footprint in delay space, must be nominally suppressed by at least five orders of magnitude at delay scales of $\tau \gtrsim 400$ ns (400 ns equals $k_\parallel = 0.2\ h\,{\rm Mpc}^{-1}$ at $z = 10$ or $\nu \sim 130$ MHz for the 21 cm line). In this case we have chosen to represent the gains as *direction-independent*, which is the component of gains we are concerned with in this work, although much work has been devoted to *direction-dependent* gain calibration (e.g. Bhatnagar et al. 2008; Intema 2014).

Redundant calibration has been hailed as a powerful alternative calibration strategy for 21 cm experiments that sidesteps some of the requirements of sky-based calibration (Liu et al. 2010; Zheng et al. 2014). However, redundant calibration still needs a sky model to pin down certain degenerate parameters it cannot solve for (Dillon et al. 2018; Li et al. 2018; Joseph et al. 2018; Byrne et al. 2019). In this work, we explore hybrid redundant-absolute calibration strategies using the `hera_cal` package.[1] Applying them to HERA Phase I, we show that redundant calibration seems to mitigate some errors associated with sky-based calibration, however, it also has its own set of uncertainties due to inherent non-redundancies that need to be mitigated. For low delay modes in the gains, we find that redundant and sky calibration yield very similar results.

In this work, we use the term *absolute calibration* to refer to the components of the full antenna-based gains that are constant across the array (note these are still frequency dependent). One example of this is the average antenna gain amplitude, which sets the overall flux scale of the data. Indeed, these are exactly the terms that are degenerate in redundant calibration. In sky-based calibration these terms are automatically solved for, whereas in redundant calibration these terms are degenerate and unconstrained.

The data used in this section were taken with the HERA Phase I instrument (DeBoer et al. 2017) in a 56-element configuration on Dec. 10, 2017 (Julian Date 2458098). HERA is located in the Karoo Desert, South Africa, at the South African Karoo Radio Astronomy Reserve. Data were

[1] `https://github.com/HERA-Team/hera_cal`

taken in drift-scan mode for roughly 12 hours per night starting at 5pm South African Standard Time, of which roughly 9 hours are deemed good quality data when the Sun is below the horizon.

The Phase I instrument repurposed many of the older PAPER experiment components, including its signal chains, correlator, feeds and front-end modules (FEM), and attached them to newly designed HERA antennas. The HERA antenna is a 14-meter dish with an optimized version of the dual linear polarization PAPER feed and FEM hoisted 4.9 meters to its focal height (Figure 1.3). The optimized feed and dish were designed to minimize reflections within the antenna, and thus limit excess chromaticity induced by the signal chain (Neben et al. 2016; Thyagarajan et al. 2016; Ewall-Wice et al. 2016; Patra et al. 2018). From the FEM, which houses an initial stage of amplification, the analog chain consists of a 150-meter coaxial cable connected to a node unit in the field where the signals are fed through a post amplification stage (PAM) and a filtering stage. From there, the signals travel through another 20-meter coaxial cable to a container where they are digitized, Fourier transformed and then cross-multiplied with all other antenna and linear polarization streams. Additional observational parameters are detailed in Table 3.1.

Not all of the PAPER signal chains could be salvaged for the HERA Phase I instrument. As a temporary stopgap, additional FEMs, cables and PAMs were manufactured for Phase I data collection. We refer to the new set of signal chains as "Type 1" and the old set of signal chains as "Type 2," which are colored blue and red in Figure 3.1, respectively. The transmission properties of the signal chains are studied in more detail in Kern et al. (2020b). For more details on the HERA Phase I signal chain and electronics, we refer the reader to Parsons et al. (2010); DeBoer et al. (2017); Fagnoni et al. (2019).

Before calibration, the data are pre-processed with part of the HERA analysis pipeline. Specifically, faulty antennas are identified and flagged at a quality metrics stage and radio frequency interference (RFI) is excised from the data using median filtering and a watershed algorithm (Kerrigan et al. 2019). The data are written to disk in the Miriad file format post-correlation, which are then converted to `UVFITS` using the `pyuvdata` software (Hazelton et al. 2017) and imported to CASA Measurement Sets via CASA's `importuvfits` task.

3.2 Sky-Based Calibration

Standard sky-based calibration is typically done by choosing a bright, well-characterized point source for the model visibilities. This is made difficult for HERA because it is a drift-scan array, meaning it cannot be pointed to an arbitrary location on the sky. Furthermore, the larger collecting area provided by a dish, as opposed to a lone dipole, means HERA's primary beam response is more compact on the sky compared to other experiments like PAPER or the MWA: at 150 MHz, HERA's primary beam FWHM is roughly 10°, compared to roughly 45° for the PAPER experiment. This means that the number of bright, well-characterized radio sources that transit our field of view is low. In fact, not a single point source within 5° of HERA's declination exceeds 20 Jy in flux density in the cold part of the radio sky (far from the galactic plane). Implementing self-calibration to high dynamic range is also difficult for HERA given its highly redundant sampling of the *uv* plane, making HERA's narrow-band grating lobes very severe. This is compounded by the poor

Table 3.1: HERA Observation Parameters

Parameter	Value
Array Configuration	Phase I
Number of Antennas	56
Array Coordinates	-30.7° S, 21.4°E
Observing Mode	drift-scan
Correlator Integration	10.7 seconds
Frequency Range	100 - 200 MHz
Channel Width	97.65 kHz
Dish Diameter	14 meter
Feed Type	dual polarization X & Y dipoles
Visibility Polarizations	XX, XY, YX, YY
Shortest / Longest Baseline	14.6 / 139.3 meters

Table 3.2: Observational parameters of HERA Phase I data.

angular resolution of the Phase I instrument, making it quickly confusion noise limited. Redundant calibration somewhat skirts the problem of an inadequate sky model, and indeed exploiting the power of redundant calibration was a motivating factor behind HERA's redundant design (Dillon & Parsons 2016). However, redundant calibration operates only within a specific subspace of the full antenna-based calibration equations, meaning a model of the sky is still fundamentally needed to fill in the few remaining degenerate modes (Liu et al. 2010; Zheng et al. 2014; Dillon et al. 2018; Li et al. 2018; Byrne et al. 2019; Dillon et al. 2020).

For power spectrum estimators that do not attempt to subtract the dominant foreground emission in the data (at the expense of losing low k modes), the stringent requirement of high dynamic range source modeling is relaxed because we are not interested in recovering modes inherently occupied by foreground emission. Hybrid techniques also exist, which try to reap the benefits of both foreground removal and avoidance (Kerrigan et al. 2019). For foreground avoidance estimators, a path towards achieving deep, noise-limited power spectrum integrations at intermediate spatial modes of $k \gtrsim 0.2\ h\ \mathrm{Mpc}^{-1}$ with a calibration derived from the sky may be possible even with the challenges faced by the HERA Phase I instrument. In this section we describe a sky-based calibration strategy for HERA using custom pipelines for calibration and imaging [2] built around the Common Astronomy Software Applications (CASA; McMullin et al. 2007) package. We start by discussing the construction of our flux density model, and then describe our calibration methodology and its validation via imaging and source extraction.

[2] https://github.com/HERA-Team/casa_imaging

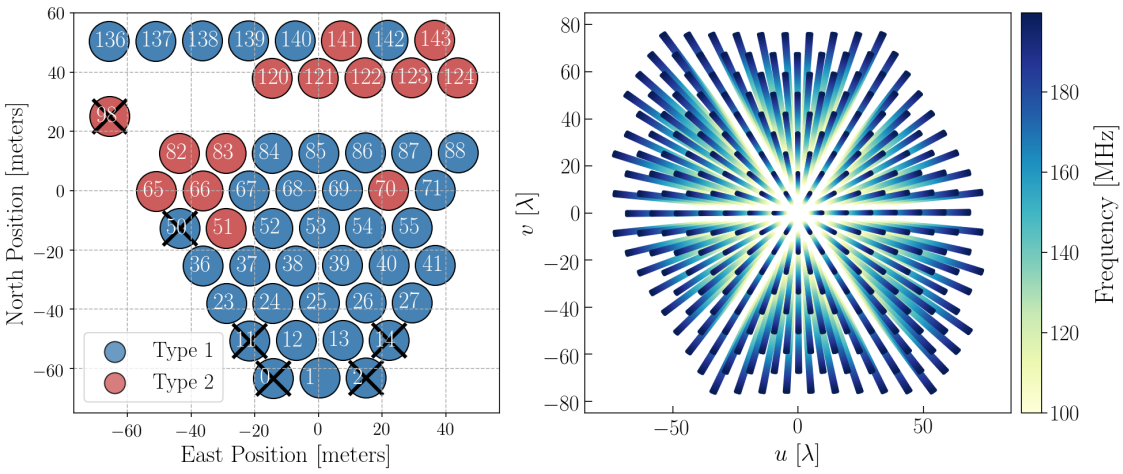

Figure 3.1: **Left**: The HERA Phase I array layout with 56 connected antennas and 50 operational antennas. Antennas determined to be problematic are marked with crosses. **Right**: The corresponding *uv* sampling of the array over a 10-minute time window and a frequency range of 100 – 200 MHz, highlighting HERA's highly redundant *uv* sampling. The color gradient represents independent *uv* samples throughout the total bandwidth.

3.2.1 Building a Sky Model

Our ideal model for sky-based calibration would involve a single, bright point source located at the pointing center of the field-of-view (FoV). Because HERA is a drift-scan array, this means our ideal calibrator would be located at $\delta \sim -30.7°$ and would transit zenith at some point in the night. Ideally this calibrator would be so bright that other off-axis point sources or diffuse emission would contribute a vastly subdominant component of the measured visibilities. Unfortunately this is not the case for HERA, so we are forced to make compromises. Figure 3.2 is a map of radio foregrounds at 150 MHz from the Global Sky Model (de Oliveira-Costa et al. 2008) and shows the HERA stripe (white-dashed), which denotes the track of the FWHM of HERA's primary beam (10° at 150 MHz). We see that the HERA stripe covers a fairly small part of the sky, demonstrating how limited we are in the amount of sky available for identifying bright calibrators.

To select the best calibration field given our limitations, we can identify some key criteria that a good field should satisfy. The first criterion is that the field should have most of its radio emission contained in the main-lobe of the primary beam. Off-axis sources located in the far side-lobes of the primary beam are troublesome because primary beam side-lobes are hard to model accurately. One workaround is to peel these sources from the visibilities before calibration (e.g. Hurley-Walker et al. 2017; Eastwood et al. 2019) but that requires one to image them at a fine frequency resolution to capture primary beam chromaticity and also with high dynamic range, which as stated is challenging for HERA Phase I. Additionally, we want our direction-independent calibration to be representative of the instrument response *at zenith*, because that is where most of the measured EoR signal comes from. Said another way, we do not want our direction-independent

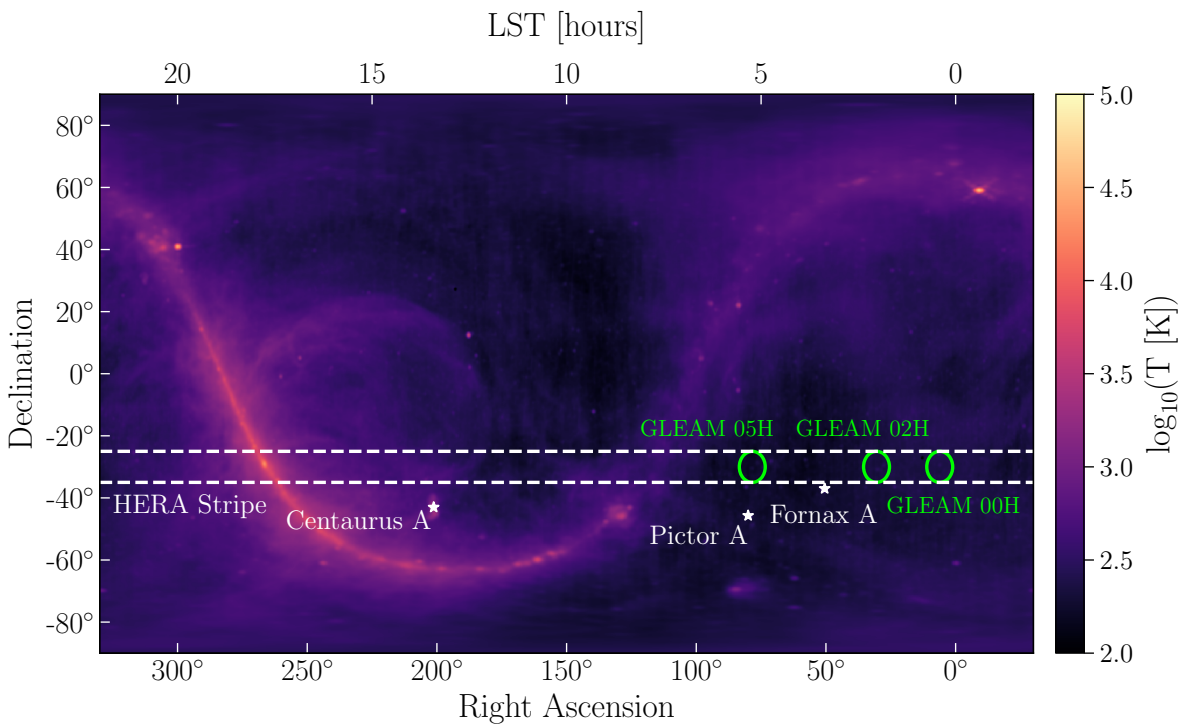

Figure 3.2: The radio sky at 150 MHz from the GSM (de Oliveira-Costa et al. 2008), showing the bright galactic and extra-galactic foregrounds that stand in the way of cosmological 21 cm experiments. The HERA stripe is shown in dashed lines centered at HERA's declination of -30.7° with a width of 10°, which is the FWHM of the primary beam at 150 MHz. The three fields identified as ideal calibration fields are shown in green circles, and some bright extended sources in the vicinity are marked as stars.

calibration to soak up structure from direction-dependent effects introduced by off-axis sources. One example of this is diffuse emission coming from the plane of the galaxy, which extends across the entire FoV when it transits.

The second criterion for a good calibration field is that it should have sources that are well characterized at the observing frequencies. Furthermore, it should have a relatively bright source very close to the FoV pointing center so that we can confirm via imaging that our calibration at zenith yields a good match to the input model. Such a source can also be useful for empirically characterizing the primary beam response with drift-scan source tracks (Pober et al. 2012; Eastwood et al. 2018; Nunhokee et al. 2020).

Recently, the MWA constructed the GLEAM point source catalogue (Hurley-Walker et al. 2017) from a deep, low-frequency survey spanning the Southern Hemisphere, overlapping with the HERA stripe. We searched the GLEAM catalogue for all point sources within 2.5° degrees of $\delta = -30.7°$ with a flux density above 15 Jy at 150 MHz, located in the cold part of the radio sky (LST < 6 hours). We find three such sources in the GLEAM catalog, J0024-2929 at 0 hours LST, J0200-3053 at 2 hours LST and J0455-3006 at 5 hours LST. Their positions, flux densities, and spectral indices

Table 3.3: HERA Calibrator Candidates from GLEAM

Name	RA (J2000)	Dec (J2000)	S_{peak}	S_{int}	α
J0024-2929	6.126	-29.48	16.45	16.10	-0.867
J0200-3053	30.05	-30.89	19.50	17.95	-0.863
J0455-3006	73.81	-30.11	16.34	17.11	-0.781

Note. — All GLEAM (Hurley-Walker et al. 2017) sources above 15 Jy, with LST < 6 hours and $-33.2 < \delta < -28.2$. Equatorial coordinates are in degrees, flux densities are in Jy at 151 MHz and α is the spectral index anchored at 151 MHz.

are reported in Table 3.3. Jacobs et al. (2016) performed a similar exercise with the TGSS ADR catalog (Intema et al. 2017). They also find J0200-3053 as a possible calibrator, but do not identify the other two sources we quote from the GLEAM catalog. For the shared source, the quoted values between the GLEAM and TGSS ADR catalogs agree to within 15%, which is roughly in-line with the overall accuracy of the survey flux scales. The green circles in Figure 3.2 are centered on each of these three calibration fields, and have diameter equal to the 10° FWHM of the HERA primary beam at 150 MHz. Stars mark the location of the nearby bright, extended sources like Pictor A and Fornax A.

Even though a ~20 Jy primary calibrator source exists at the pointing center of each field, they themselves make up only a fraction of the total flux density measured by the instrument at those LSTs. For short baselines the dominant sky component is diffuse galactic emission, while longer baselines are dominated by point sources spread across the FoV. Although models of the diffuse galactic emission exist (de Oliveira-Costa et al. 2008; Zheng et al. 2017) they are only accurate at the ~15% and furthermore extend across the entire FoV, filling the hard-to-model sidelobes. At the moment, we only use point sources in our flux density model and cut short baselines (< 40 meters) that have significant amounts of diffuse foreground emission. Our starting model for each field is made up of all GLEAM point sources down to 0.1 Jy in flux density extending 20° in radius from the pointing center, which typically results in ~10,000 sources in the flux density model. We take the GLEAM-reported flux density of each point source at 151 MHz and their spectral index and insert them into a CASA component list. All sources are assumed to be unpolarized and their fluxes are inserted purely as Stokes I. For GLEAM sources without a spectral index, we take the reported flux density of the source at 122, 130, 143, 151, 158, 166, and 174 MHz and fit our own spectral index. After constructing a component list with all of the relevant GLEAM sources we make a 1024-channel spectral cube image of the component list with the CASA `Image.modify` task, matching the channelization of HERA data, and export it to `FITS` format. The image has a pixel resolution of 300 arcseconds, which is 6 times smaller than the synthesized beam FWHM of ~0.5 degrees.

Note that the GLEAM catalogue does not include bright, extended sources like Fornax A and Pictor A. As shown in Figure 3.2, the calibration fields are chosen such that these sources are

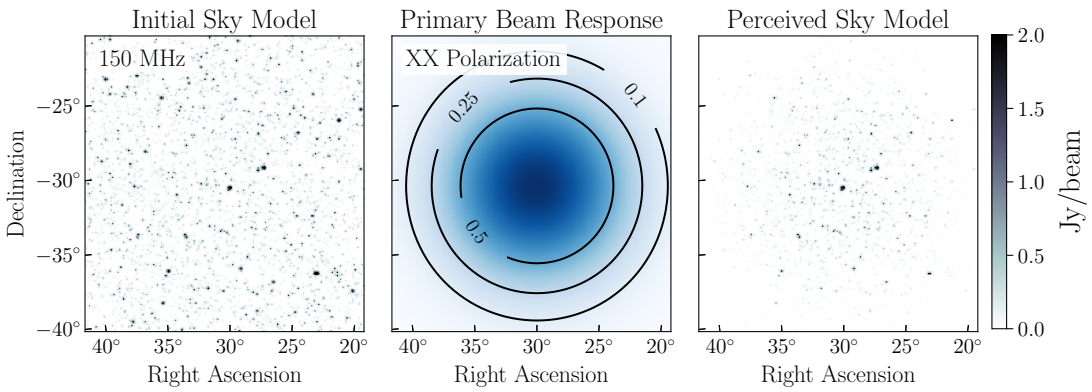

Figure 3.3: Construction of the GLEAM-02H field sky model for calibration at 150 MHz. Each frequency channel in the model is constructed independently in the same manner. **Left**: All GLEAM point sources in Stokes I polarization above 0.1 Jy within 20° of the pointing center. In this figure, the point sources have been convolved with a narrow 2D Gaussian merely for visual clarity. **Center**: The peak-normalized primary beam response for the XX instrumental linear polarization at 150 MHz (Fagnoni et al. 2019). **Right**: The Stokes I model multiplied by the XX primary beam response yields a perceived flux density model that is then converted into visibilities for calibration.

heavily attenuated by the primary beam, but even still these sources can be seen at the level of a few Jansky for the 02-hour and 05-hour fields, for example. Fornax A and Pictor A can be included in the component list model for the GLEAM-02H and GLEAM-05H fields, respectively, by adopting point source models with spectral indices informed by recent low-frequency studies (Jacobs et al. 2013; McKinley et al. 2015). Although these sources have a non-zero angular extent to them, for HERA Phase I angular resolutions a point source model is a fair approximation.

Next we incorporate the effects of the direction and frequency-dependent antenna primary beam response to create a perceived flux density model. We use an electromagnetic simulation of the HERA primary beam from Fagnoni et al. (2019), which includes effects from the dish and feed. That work also explored the effects of mutual coupling on the primary beam response given an element embedded in the array, finding second-order effects on the beam response near the horizon at the level of 10^{-2} in power. Empirical studies by Kern et al. (2020b) find similar levels of mutual coupling in the data, and present post-calibration methods for mitigating their effects. In this work we only use the Fagnoni et al. (2019) beam model of the antenna and feed, and defer using the embedded element pattern in calibration for future work. Each linear dipole in the feed, X and Y, is assigned its own beam model, where one is simply a 90° rotation of the other. The beams are peak-normalized at boresight independently at each frequency, and we then multiply the beam response at each pixel on the sky separately for the X and Y dipoles. This results in two spectral cubes, one for both the XX and YY instrumental visibility polarizations, which constitutes our perceived model. In this work we do not construct models for the cross-polarized XY and YX visibilities as we will not perform polarization calibration, although this can be done with polarized beam models (Martinot et al. 2018). Figure 3.3 demonstrates this for the GLEAM 02H field in XX

instrumental polarization, showing the initial sources (left), the XX primary beam response (or the squared X-dipole response) at 150 MHz (center), and the product of the two (right). Lastly, the model cubes are transformed from the image to the *uv* domain via CASA's `ft` task and are inserted into the model column of the Measurement Sets for calibration.

3.2.2 Calibration

Next we will describe our approach for deriving complex, direction-independent antenna gains with CASA. For simplicity, we will focus our discussion specifically to the GLEAM-02H field, but calibration on any other field would follow the same procedures outlined below. As noted, the data are first processed for faulty antennas and RFI flagging by the HERA analysis pipeline. We then take five minutes of drift-scan data centered at the transit of the primary calibrator, apply a fringe-stop phasing to the transit LST and then time-average the data. Averaging five minutes of data allows us to increase the signal-to-noise ratio (SNR) of the derived gains and is still a fairly short time interval compared to the FWHM primary beam crossing time (at 150 MHz) of ~46 minutes, at which point sky source decorrelation will begin to be a problem. Due to the inherent stability of the drift-scan observing mode, we do not expect the gains to vary substantially over such short time scales (although see subsection 3.3.2 for higher-order effects).

Before proceeding with calibration, we enact a minimum baseline cut such that all baselines shorter than 40 meters ($\sim 20\lambda$) are excluded, leaving 65% of the visibilities for calibration. HERA's shortest baselines are most sensitive to the diffuse galactic emission that is not included in our point source model. After experimenting with various baseline cuts we find a 40-meter cut to be a good compromise between keeping as much data as possible for maximal gain SNR and eliminating diffuse foreground flux in the data that is not included in our model.

Our process for deriving antenna gains uses a series of standard routines in CASA. Before each calibration step, we apply all previous calibration steps to the data on-the-fly. The final calibration is then simply the product of all steps in our calibration chain. We start by performing delay calibration using the `gaincal` task, which is done to calibrate out the cable delay of each antenna. Next we perform mean-phase and mean-amplitude calibration (which consist of two numbers for each antenna-polarization across the entire bandwidth) also using the `gaincal` task. This removes any residual phase offset after delay calibration and sets the overall flux scale of the data. Up to now all calibration steps are smooth across frequency and therefore do not contain significant spectral structure. Finally, we derive complex antenna bandpasses using the `bandpass` task, which solves each frequency channel independently from all others. This last step has the possibility of introducing an arbitrary amount of spectral structure into the gain solutions and therefore deserves closer attention, which we revisit in section 3.3.

In this work we do not make any attempt to correct for effects due to the ionosphere. This is less of a concern given the higher frequency range of 100 – 200 MHz, in addition to the fact that the array has limited angular resolution and that observations are only taken at night when the sun is below the horizon leading to calmer ionospheric conditions. We also do not attempt to calibrate the relative phase between dipole polarizations in this work, which is difficult due to the dearth of bright polarized sky sources (Moore et al. 2017; Lenc et al. 2017), although this can still

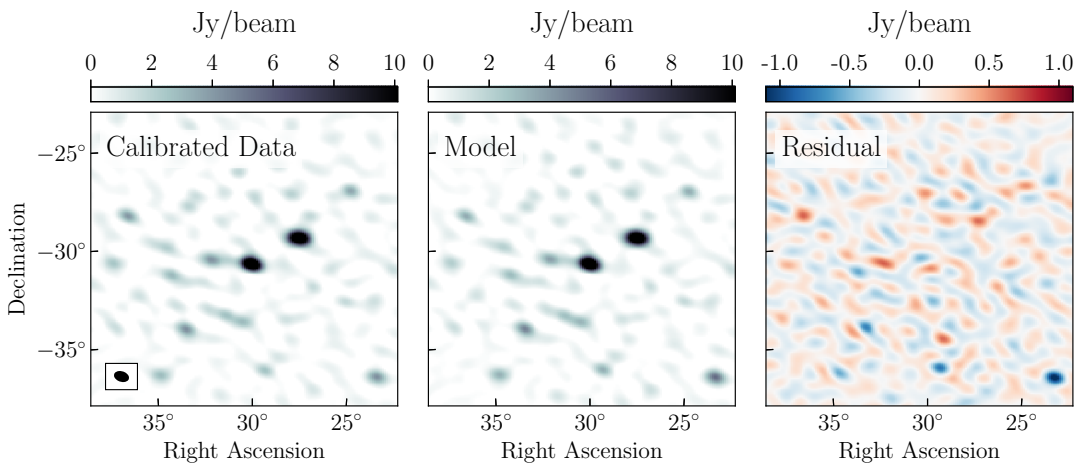

Figure 3.4: Multi-frequency synthesis image of the GLEAM-02H field in XX polarization spanning 120 — 180 MHz of the calibrated visibilities (left), model visibilities (center) and the residual visibilities (right). Each image is CLEANed with the same parameters down to 0.5 Jy, with the restoring beam shown in the lower left. The model and calibrated data show good agreement in the main lobe of the primary beam. At larger zenith angles the residual image shows evidence for mis-calibration, likely due to primary beam errors.

be partially constrained if we assert that the Stokes V visibilities be consistent with thermal noise (Kohn et al. 2016). This is also less of a concern because in this work we are mostly interested in the parallel-hand (i.e. XX and YY) dipole and Stokes I data products, which are not as sensitive to this term as the Stokes U & V data products. While previous work has shown that ionospheric leakage of point source foregrounds can in principle be significant (Nunhokee et al. 2017), ionospheric-induced leakage terms have also been shown to average down night-to-night (Martinot et al. 2018). As we will show in subsection 3.2.3, the amount of intrinsic polarization leakage observed in the data is quite small, even without performing any kind of polarization calibration. Future HERA observations that i) extend below 100 MHz or ii) are interested in polarized data products will need to revisit these topics. For an investigation into direction-dependent effects and polarization leakage from the HERA-19 commissioning array see Kohn et al. (2019).

3.2.3 Imaging

To test the fidelity of the calibration, we make multi-frequency synthesis (MFS) images of the calibrated data, the calibration model and their residual visibility as a visual assessment of their agreement. The MFS images use five minutes of data and a 60 MHz bandwidth spanning 120 – 180 MHz. All images are made from only the baselines involved in the calibration ($|b| > 40$ m), employ robust weighting with `robust = -1` and adopt the Hogbom deconvolution algorithm (Högbom 1974) using the `tclean` task. All images are CLEANed independently down to a threshold of 0.5 Jy in the polarization they are imaged in. CLEAN masks are used around the brightest sources initially and then the CLEAN mask is opened up to the entire field. We produce images in instrumental XX

and YY polarization and also pseudo-Stokes I, Q, U & V polarization.

The HERA array is not perfectly co-planar, which will introduce artifacts into wide-field images made with CASA. This can be mitigated with W-projection (Cornwell et al. 2008), however, given the field of view and modest angular resolution of the Phase I array, we do not expect non-co-planar effects to generate an appreciable amount of error. Therefore we do not perform W-projection in the process of imaging, which also reduces its overall computational cost.

Figure 3.4 shows the GLEAM-02H field in XX polarization and images of its calibrated data (left), model (center) and their residual visibility (right). The size of the synthesized beam is shown in the lower left. We see good agreement between the data and model down to a few percent. The residual image appears noise-like in the main lobe, but further away from the pointing center we can begin to correlate point sources in the data with point sources in the residual. This is a result of an improper perceived flux density model (either with the inherent source fluxes or, more likely, the adopted primary beam response). This will introduce spectrally-dependent errors into the gain solutions at some level (Barry et al. 2016; Ewall-Wice et al. 2017) which we explore in the following section. This can be partially mitigated by self-calibration or redundant calibration, although redundant calibration still suffers from this effect to some degree (Byrne et al. 2019).

We also make images of the pseudo-Stokes visibilities as a diagnostic. The pseudo-Stokes visibilities (Hamaker et al. 1996) are a linear sum of the linear polarization visibilities, defined as

$$\begin{pmatrix} V_I \\ V_Q \\ V_U \\ V_V \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & -i & i & 0 \end{pmatrix}\begin{pmatrix} V_{\mathrm{XX}} \\ V_{\mathrm{XY}} \\ V_{\mathrm{YX}} \\ V_{\mathrm{YY}} \end{pmatrix}. \tag{3.2}$$

Note these are not true Stokes parameters, which are only properly defined in the image plane, but can be thought of as approximations to the true Stokes visibility one would form by Fourier transforming the true Stokes parameter from the image plane to the uv plane. In the limit that the instrumental (direction-dependent) Mueller matrix is the identity matrix, then the pseudo-Stokes visibility defined in Equation 3.2 is identical to the true Stokes visibility. In practice, we do not expect this to be the case except for possibly near the pointing center in the image where, after having performed direction-independent calibration, one might expect direction-dependent terms to be minimal.

We do not expect appreciable levels of polarized sources in the GLEAM-02H field. For a recent study by the MWA see Lenc et al. (2017). Given an ideal telescope with no instrumental leakage, we would therefore expect the pseudo Q, U and V visibilities to look noise-like. However, we know that the primary beam response at a given point on the sky for the X and Y dipoles are not the same at low zenith angles, which will by itself cause polarization leakage of observed off-axis sources into Stokes Q (Moore et al. 2017). Furthermore, we have not attempted to calibrate feed D-terms (Hamaker et al. 1996) or the unconstrained relative X-Y phase parameter leftover after Stokes I calibration (Sault et al. 1996). We also know from previous studies that mutual coupling exists at a non-negligible level (Fagnoni et al. 2019; Kern et al. 2020b), which is in principle a direction-dependent term in the Mueller matrix. This means that we wholly expect that images formed from pseudo-Stokes visibilities will i) not necessarily be representative of the true

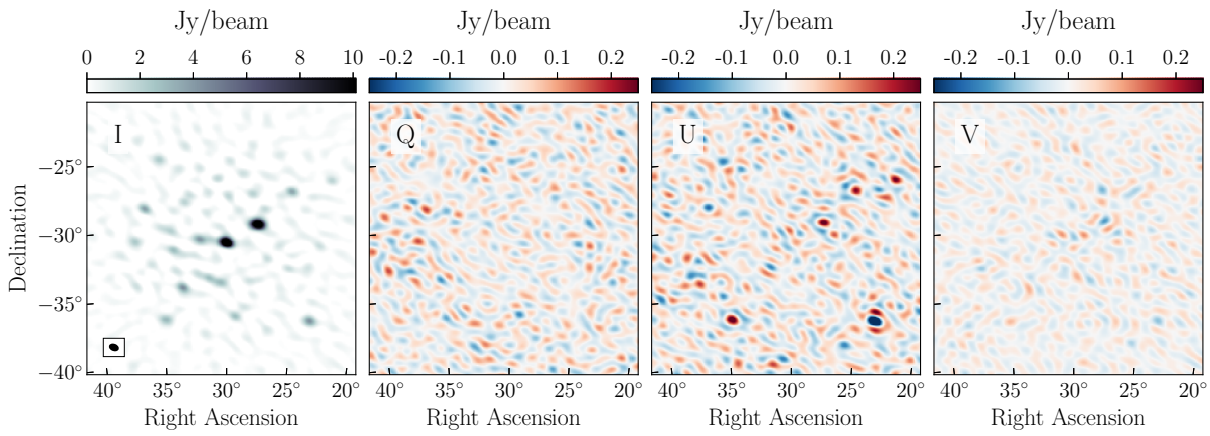

Figure 3.5: Multi-frequency synthesis images (120 – 180 MHz) of the GLEAM-02H field in all pseudo-Stokes I (far-left), Q (center-left), U (center-right) and V (far-right) polarizations. Each image is CLEANed with the same parameters down to 1 Jy, with the CLEAN beam shown in the lower left. Even with no polarization calibration, the observed leakage from I → Q, U & V is a few percent.

Stokes parameters in the image plane, except for maybe near the pointing center and ii) that we should observe non-negligible amounts of polarization leakage from Stokes I → Stokes Q, U & V. To properly make true Stokes parameters one would image each of the linear dipole visibilities and perform direction-dependent corrections in the image plane before adding them in a similar manner as Equation 3.2. At the moment we defer this to future works that more carefully consider polarization calibration.

Figure 3.5 shows MFS images of the GLEAM-02H field in pseudo-Stokes I, Q, U & V (left to right). The first thing to note is that the observed leakage of Stokes I to Q, U and V is on the order of a few percent, which is quite low given we did not apply a polarization or direction-dependent calibration. Looking at the pseudo-Stokes Q image we can see the effects of primary beam asymmetry between X and Y dipoles: without a primary beam correction (which is not applied here), the asymmetry will cause leakage of I → Q (Moore et al. 2017), which is exacerbated the more discrepant the primary beam responses are at a given point on the sky. Although nearly azimuthally symmetric, the X-dipole beam is elongated along the North-South direction while the Y-dipole is elongated along the East-West direction (Fagnoni et al. 2019; Martinot et al. 2018). This means we might expect the relative amplitude of the X and Y beams to attain a better match in the corner of our images, and would therefore expect to see more I → Q leakage in a quadrupolar pattern on the sky. Indeed, this is observed in the pseudo-Stokes Q image to some degree (Figure 3.5).

The pseudo-Stokes U and V images also exhibit interesting behaviors, in particular the sources in the pseudo-Stokes U image that are clearly correlated with true Stokes I sources, as well as the rumble in the pseudo-Stokes V image that seems to be concentrated near the main lobe. This could be due to polarization leakage stemming from the uncalibrated X-Y phase term, however, further work is needed to identify its exact cause.

Having shown that our calibration does a fairly good job bringing our data in-line with our

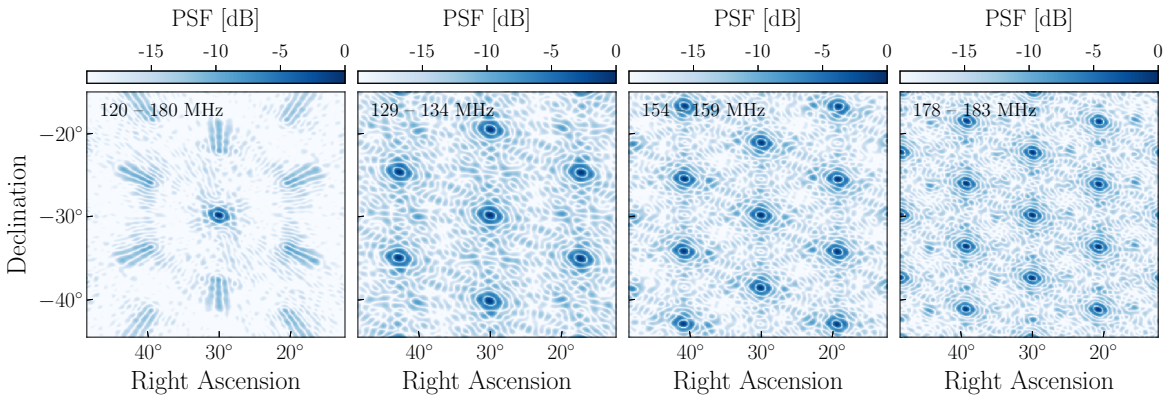

Figure 3.6: The HERA Phase I point spread function (without primary beam correction) from a 5-minute observation across a wide band (left), and a narrow band located in a low-band spectral window (center-left), mid-band spectral window (center-right) and high-band spectral window (right). The grating lobes of the narrow band spectral windows appear in hexagonal patterns reflecting the (un-)sampled *uv* spacings on the array, and reach upwards of 50% of the peak PSF response at image-center.

model (Figure 3.4) and that, even without polarization calibration, polarization leakage is observed at a few percent (Figure 3.5), we should also show that our derived bandpass is an accurate solution as a function of frequency. To do this we can make a spectral cube of our calibrated data and compare to the original catalogue used for calibration. However, making a spectral cube with fine frequency resolution means that the point-spread function (PSF) sidelobes and grating lobes become increasingly a problem due to the sparse sampling of the *uv* plane. Figure 3.6 shows the HERA Phase I PSF across a wide 60 MHz band (left) and three narrower 5 MHz bands (center and right). For wide-band imaging the PSF grating lobes are smeared out due to the large bandwidth. However, for narrow-band imaging the grating lobes rise to above 50% the peak PSF response at image center; for narrower spectral windows this is only exacerbated. Such strong grating lobes make performing deconvolution to high dynamic range difficult, especially in a confusion-limited regime.

We can partially work around this by applying CLEAN masks around bright sources and then CLEANing down iteratively while opening up the mask to dimmer and dimmer sources. Indeed, this is what we do to make a coarse-channel spectral cube, which consists of MFS images with 5 MHz in bandwidth using iterative CLEAN runs targeting successively dimmer sources. However, in the case of single-channel imaging even this does not work: the grating lobes are just too severe to deconvolve them from the image without misplacing source flux in un-modeled sidelobes. Figure 3.7 shows the result of a coarse, 5 MHz wide spectral cube CLEAN for a spectral window centered at 155 MHz (greyscale, left). We also show all GLEAM sources in the original model with fluxes above 0.5 Jy in purple, which demonstrates the high degree of confusion given our modest angular resolution: each "source" in our images are generally two or more GLEAM sources blended together. We therefore cannot easily relate the source flux in our images to one or even multiple sources in the GLEAM catalogue, as each GLEAM source will have a different contribution to a

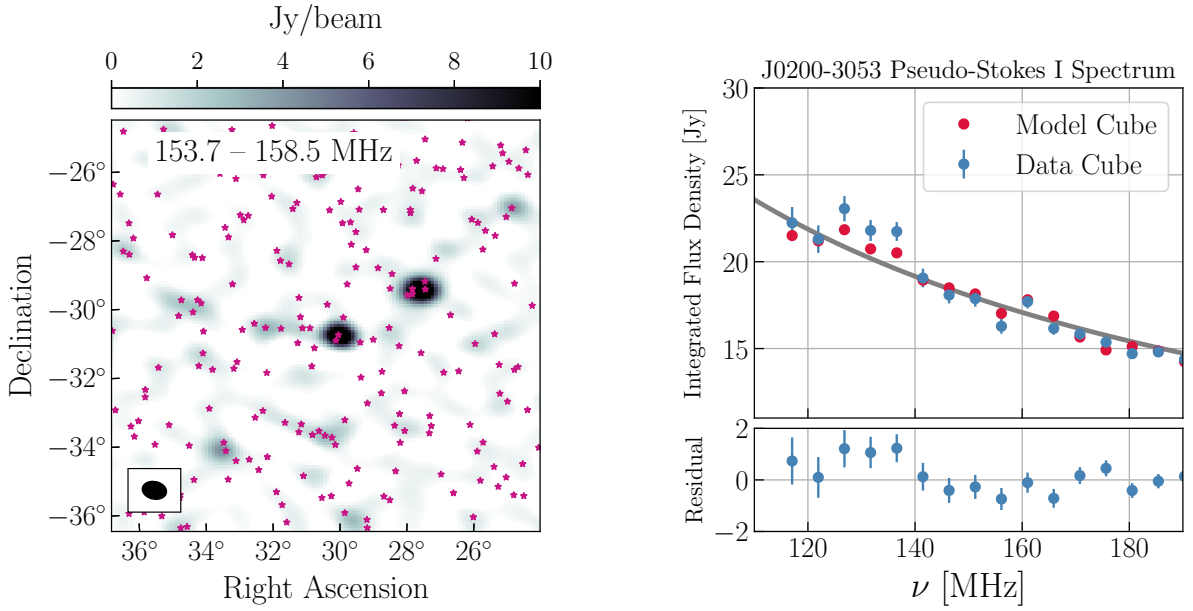

Figure 3.7: The extracted spectrum of the primary calibrator GLEAM J0200-3053 from the GLEAM-02H field. **Left**: CLEANed MFS image of the data (colorscale) across a narrow band (153.7–158.5 MHz). The purple markers show each GLEAM point source above 0.5 Jy used in the initial model, demonstrating the degree of source confusion given the Phase I angular resolution. **Right**: Extracted spectrum of J0200-3053 (center of left image) across each channel in the data spectral cube (blue) and model spectral cube (red). The data and model are in good agreement with each other, and are well-fit by the original input GLEAM J0200-3053 power law model (grey). Large-scale frequency deviations from the power-law fit are partially reflected in both the data and model, suggesting that they are not due to mis-calibration but due to imperfect PSF sidelobe removal in the process of imaging. The data cube – model cube difference shows residual structure at the ~ 5% level.

HERA source given its distance from it and the HERA PSF. If our goal is to compare extracted fluxes between the data and a point-source model we should take the PSF out of the equation. The deconvolution on the data attempts to do this at some level, but is limited fundamentally in precision by the width of the synthesized beam. Another way is to simply add the PSF into the model by imaging the model visibilities and then CLEANing and running a source extraction in the same way as is done for the data. This means that the inherent shortcomings of the deconvolution and the limitations of the PSF (both things not really relevant for validating gain calibration done in the *uv* domain rather than the image domain) are kept constant between data and model, so we can make a better comparison between the two.

Source extraction is done on a source-by-source basis with custom software. First we select the coordinates of a desired source in the data, then the extraction process makes a postage-stamp cutout in the shape of the synthesized beam with twice its FWHM around the desired source and fits a 2D Gaussian of variable major axis length, eccentricity, amplitude and position angle using

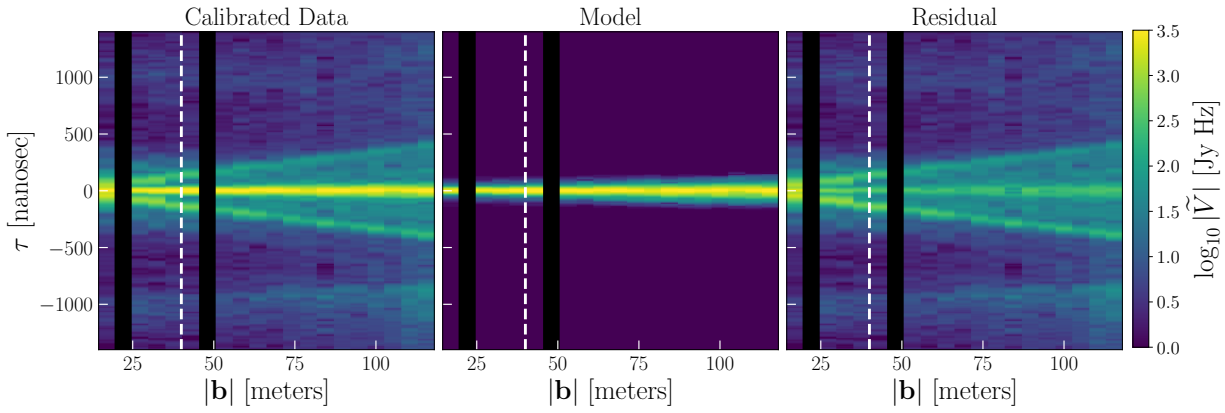

Figure 3.8: Redundantly-averaged pseudo-Stokes I visibilities in delay space transformed across a 120 – 180 MHz spectral window, ordered according to baseline length. We show the calibrated data (left), the point source calibration model (center), and their residual (right). Short baselines to the left of the white dashed line are not used in calibration. Black regions represent a lack of data at those baseline lengths. The data clearly show a pitchfork-like foreground wedge predicted by Thyagarajan et al. (2016). Note that the edges of the pitchfork are not reflected in the calibration model, which will generate calibration errors. The residual power of the main foreground lobe in the wedge is suppressed by about a factor of 10 compared to the data, but is still seen above the noise floor of the data. Additional power at large delays ($\tau \sim 1000$ ns) are the same systematics seen in Kern et al. (2020b).

the `astropy.modeling` module. It then records the integrated flux of the fit in Jy and computes the fit error by taking the RMS of the image in an annulus outside the cutout and dividing by the square-root of the synthesized beam area (Condon 1997).

This is done for the GLEAM-02H field primary calibrator J0200-3053 for each 5 MHz-wide channel in the coarse spectral cube of the data and model, shown in Figure 3.7. The data (blue) and model (red) are in good agreement with each other across the entirety of the band, and are in relatively good agreement with the primary calibrator's original power law model from the GLEAM catalogue (grey, Table 3.3). Both the data and model exhibit sinusoidal frequency fluctuations about the power law model; however, because this structure is represented in model spectra we can conclude that some of these fluctuations are due to imperfect PSF sidelobe removal in the CLEAN process, rather than calibration errors. If we take the difference between the extracted data and model fluxes then we see residual deviations at $\sim 5\%$ the source's intrinsic flux. However, these deviations look similar in form to the first-order sinusoidal variations about the smooth power law, possibly suggesting that some of these features in the residual are too due to imperfect PSF sidelobe removal. One possibility is that the CLEAN deconvolution achieved better sidelobe removal on the model cube compared to the data cube, which would generate the kind of observed sinusoidal variations in the data-to-model residual. This would not be entirely surprising given the extra terms in the data that are not in the model, including diffuse foregrounds, which would make deconvolution more difficult. The second-order fluctuations in the data-to-model residual (channel-to-channel) hovers at roughly 1% of the intrinsic source flux. Overall, these lines of

evidence suggest that the quality of the spectral calibration across the band is on the order of a few percent.

However, the leading uncertainty in our absolute calibration is the determination of the overall flux scale. By adopting the GLEAM point source catalogue as our model, we have set the flux scale of our calibration to GLEAM, which themselves tie their flux scale to the VLA Low-Frequency Sky Survey redux (VLSSr; Lane et al. 2014), the NRAO VLA Sky Survey (NVSS; Condon et al. 1998) and the Molongo Reference Catalogue (MRC; Large et al. 1981). When comparing their measured source fluxes to sources from these catalogues, their flux scaling appears to be unbiased with an uncertainty of $\sim 10\%$. One concern about our usage of a single GLEAM field to set the flux scale is the fact that GLEAM's J0200-3053 source may be an outlier in that distribution, implying that our flux scale could be significantly biased. This concern is tempered by the residual image of Figure 3.4, which shows that not only J0200-3053, but all sources in the main lobe of the beam have an unbiased residual, meaning that our final flux scale agrees with the GLEAM flux scale for all sources in the main-lobe of our primary beam.

To better understand the match between the data and the flux density model, we take the full gain solutions from the GLEAM-02H field and use them to calibrate all baselines in the data. We then form pseudo-Stokes I visibilities and coherently average all baselines within a redundant group (i.e. with the same baseline length and orientation). Then we take the Fourier transform of the visibilities across a wide bandwidth spanning 120 – 180 MHz, having first applied a Blackman-Harris windowing function (Blackman & Tukey 1958) to limit spectral leakage in the discrete Fourier transform (DFT). Before we do this, however, we must first account for the frequency channels that have been flagged due to RFI. These will create strong sidelobes in the Fourier transform if not accounted for. To overcome this, we employ a 1-dimensional deconvolution algorithm that deconvolves the sidelobes due to the RFI, which is conceptually identical to the CLEAN algorithm employed by radio interferometric imaging to interpolate over missing uv samples (Högbom 1974), and can be found in the `hera_cal` package. In our case, we build a CLEAN model in delay space out to $\tau = 2000$ ns, and interpolate over the flagged channels with the CLEAN model before taking the final DFT to the delay domain, which we do for the data and model visibilities in an identical fashion. We then coherently time average the 5 minutes of data around the GLEAM-02H calibration field, take the absolute value of the averaged visibilities and average all baselines of the same length, regardless of orientation. This is the same procedure one would take to form 2D cylindrically averaged power spectra, but in this case we are working with just the visibilities in the Fourier domain.

Figure 3.8 shows this for the calibrated data (left), model (center) and shows their residual (right). From it, we can clearly see the pitchfork-like foreground wedge with a main component centered at $\tau = 0$ ns and then branches at positive and negative delay, tracking the foreground horizon limit as a function of baseline length. Recall that baselines shorter than 40-meters in length (white dashed) are not used in calibration. The pitchfork branches are caused by the foreshortening of a baseline's separation vector at the horizon, thus increasing its sensitivity to diffuse emission (Thyagarajan et al. 2015). The point source model, lacking diffuse foregrounds, clearly does not have a strong pitchfork feature. This discrepancy will create gain errors in the calibration solutions at the delay scale of the pitchfork, which for baselines above 40-meters begins at around 150 ns

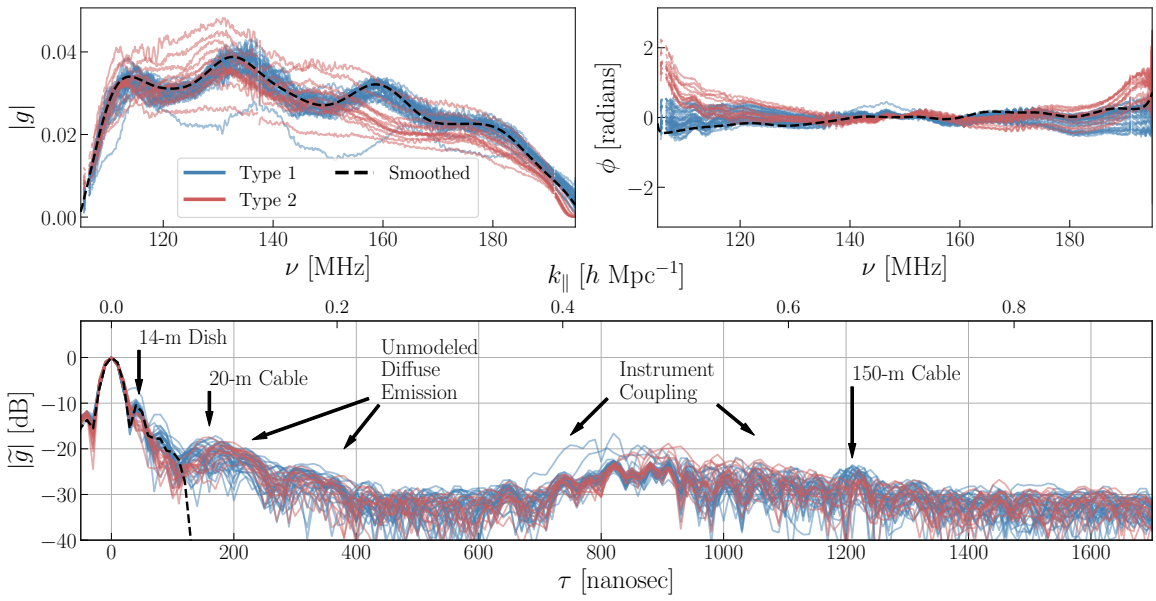

Figure 3.9: Antenna gains derived from the GLEAM-02H field. Type 1 & 2 signal chains are plotted in blue and red, respectively. The phase of the gains (top-right) are plotted after taking out the cable delay from each antenna for visual clarity. The peak-normalized delay response of the gains show structure at delays representative of elements in the signal chain (bottom), and also show contamination by terms that are not antenna based, like un-modeled diffuse emission and instrumental coupling systematics (Kern et al. 2020b). We also show one of the Type 1 gains smoothed at a 100 nanosecond scale for reference (dashed-black).

and extends beyond that for longer baselines. This is explored in the following section. Lastly, the data and model are somewhat well matched at $\tau \sim 0$ ns, with the residual power being suppressed by a factor of 10 compared to the data but still above the noise floor of the data outside the wedge. This residual power can come from un-modeled diffuse flux in the main lobe of the primary beam, but is also likely to be from calibration errors due to mis-modeled point sources. An increase in observed power in the data at large delays $|\tau| > 800$ ns is a cross coupling systematic, and is not foreground signal (Kern et al. 2020b).

3.3 Gain Stability

In this section we characterize the spectral and temporal properties of the derived complex antenna gains and discuss their impact on downstream analyses. Gain calibration is a multiplicative term in the frequency and time domain meaning it can equivalently be thought of as convolution in the Fourier domains of delay and fringe-rate, the Fourier duals of frequency and time respectively, by a "gain kernel," or the Fourier transform of the gain response. Solving for and applying antenna-based gains can therefore be thought of as trying to deconvolve the inherent gain kernel imparted by the instrument. For 21 cm experiments aiming to uncover a signal buried under noise and

systematics, the principal concern when applying gain solutions to the data is understanding how this gain kernel may or may not be smearing foreground signal to spectral modes that are otherwise foreground-free: any kind of deviation in the derived gain solution from the true underlying gain will cause such smearing, at some level.

3.3.1 Spectral Response

Works investigating sky-based calibration in the limit of an incomplete sky model showed it results in gains with erroneous spectral structure that can fundamentally limit 21 cm studies (Barry et al. 2016; Ewall-Wice et al. 2017; Byrne et al. 2019). Similar effects have been shown to exist for redundant calibration, where inherent non-redundancies of the array create a similar type of spectrally-dependent gain error (Orosz et al. 2019). What has yet to be studied in detail is how other kinds of instrumental systematics, such as mutual coupling or crosstalk, get picked up in the process of gain calibration and what their effect is in shaping the inherent and estimated gain kernel. For systematics like crosstalk and mutual coupling which are highly baseline-dependent, one would naively expect that the antenna-based gains would not significantly pick up on these terms due to their decoherence when averaged across different baselines; however, it would not be surprising to see them at some level reflected in the gain solutions, even if they are averaged down to some degree. Furthermore, Figure 3.8 shows us that there is a non-negligible data-to-model discrepancy caused by un-modeled diffuse emission even for baselines above our 40-meter cut, which will also create gain errors.

To summarize Kern et al. (2020b), the HERA Phase I system shows evidence for cross-coupling systematics at large delays $|\tau| > 800$ ns, and also shows evidence for diffuse flux and / or mutual coupling at smaller delays corresponding to a baseline's geometric horizon (for $|\mathbf{b}| = 45$ m, this is ~ 150 ns). In Figure 3.9 we show the frequency and delay response of the CASA-derived, sky-based gains from section 3.2. We plot the gain amplitude (upper-left), gain phase after removing the cable delay for each antenna (upper-right) and the Fourier transform of the gains in delay space (bottom) normalized to their peak power at $\tau = 0$ ns. We categorize the gains into Type 1 (blue) and Type 2 (red) signal chains (Figure 3.1), which shows a clear bi-modality in the spectral structure of the gains between these groups. This bi-modality is also seen in the reflection properties of the signal chains and is discussed in more detail in Kern et al. (2020b). Arrows mark the expected regions in delay space where certain electromagnetic elements in the signal chain can create systematics, such as reflections within the 14-meter dish and reflections in the 20-meter and 150-meter coaxial cables. The gain kernels of each antenna (Figure 3.9-bottom) also clearly shows that instrumental cross coupling systematics at $|\tau| > 800$ ns are being picked up by the gain solutions. It also shows un-modeled diffuse emission at lower delays $|\tau| > 200$ ns, which may also have contributions from mutual coupling systematics that appear at similar delays. Because instrumental cross coupling and diffuse emission are baseline-based and not antenna-based, they cannot be calibrated out of the data with antenna-based, direction-independent gains, and must be removed on a per-baseline basis. This means that the presence of these structures in the gains will only spread the systematics around: in the worst case spreading them to baselines that may have been systematic free to begin with.

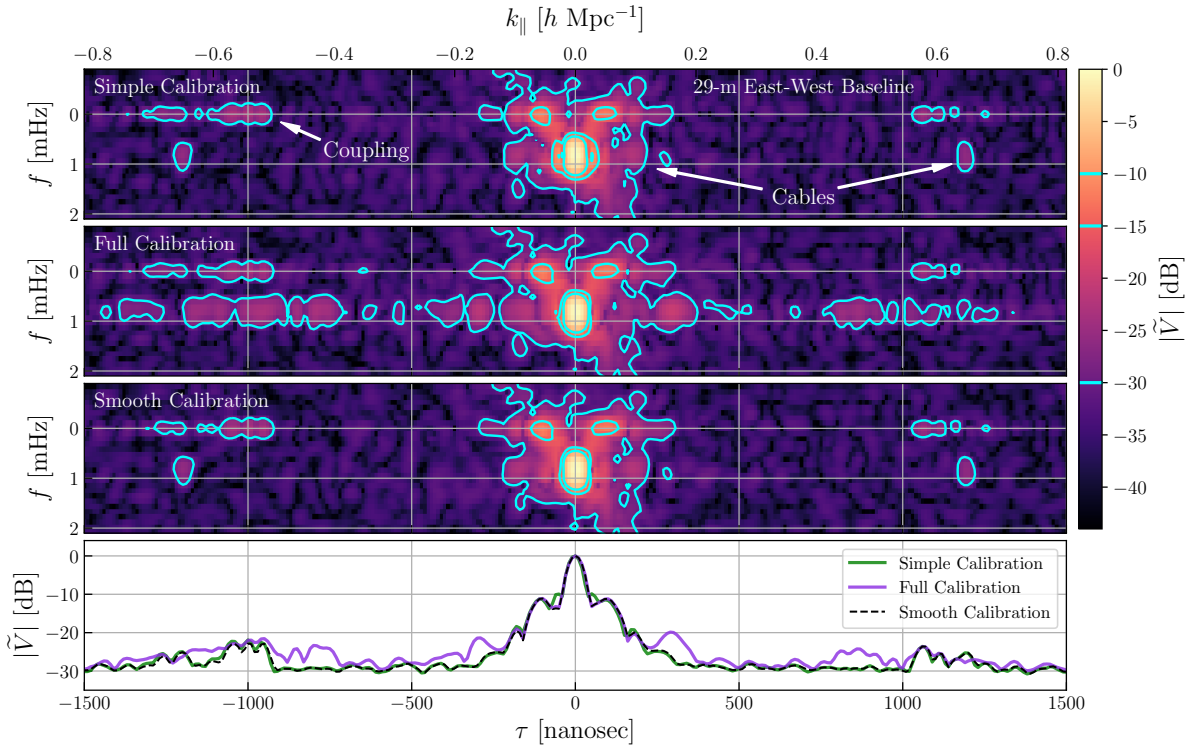

Figure 3.10: Sky-based gains applied to a single 29-meter East-West visibility over 8-hours of LST and transformed to delay and fringe-rate space (see text for details). The data are peak-normalized, and contours show -30, -15, and -10 dB levels. The time-averaged delay responses are shown in the bottom panel. Sources of cable reflection and instrument coupling are marked. The full gain applied to the data leads to significant contamination of coupling systematics due to the full gain kernel smearing the foreground horizontally in fringe-rate and delay space. Smoothing the calibration allows us to calibrate out the features at low delays we know to be calibrate-able (e.g. dish reflections) and toss out features in the gain kernel above 100 nanoseconds.

Figure 3.10 shows the result of applying sky-based gains to the visibility data and transforming to the Fourier domains of delay and fringe-rate space. We apply the gains to 8-hours of drift-scan data from a single 29-meter East-West visibility, and we do this having filtered the gains in three different ways: 1) the first method (simple calibration) takes only the band-averaged amplitude and cable delay component of the gain 2) the second method (full calibration) just takes the full gain solution as-is, and 3) the third method (smooth calibration) smooths the gains across frequency out to a 100 ns scale, which is also plotted in Figure 3.9 (black-dashed). The bottom panel shows the time-averaged delay response of the panels shown above. In the simple calibrated data, the foregrounds are contained to low delays and appear predominately at positive fringe-rates, which we expect because the sky rotates in a single coherent direction in the main lobe of primary beam (Parsons et al. 2016). Foregrounds can also occupy near-zero and negative fringe-rate modes, which correspond to structures on the sky near the South celestial pole and near the horizon, but

are attenuated by the primary beam response. If the data were nominal then the rest of the Fourier space would be dominated by thermal noise; however, this is not what we observe. We also see cable reflection signatures, which should appear as reflected copies of the foregrounds at the same fringe-rates but at positive and negative delays (marked). And we see strong cross coupling features at large positive and negative delays occupying near-zero fringe-rate modes (marked).

When we go to apply the full calibration we find a large amount of excess structure at intermediate and large delays occupying positive fringe-rates, which is not surprising given the gain kernels shown in Figure 3.9. We see that other baselines that happened to have the systematics at intermediate delays have contaminated this baseline at the same delays. What is more, these systematics are now occupying the same positive fringe-rate modes as the sky,[3] and therefore cannot be easily removed with standard cross coupling removal techniques (Kern et al. 2020a). There are some benefits of the full calibration, though. One is that it can calibrate out signal chain reflections because those factor as antenna based terms. This can be seen in the data as the suppression of the cable reflections at large positive and negative delays, and also in the suppression of the dish reflection at $\tau = \pm 50$ ns (which is most apparent as the tightening up of the contours in the brightest spots of the foregrounds, or the drop in the shoulder power in the time-averaged spectra). While cable reflections at high delays can be calibrated out with sky-agnostic modeling (Ewall-Wice et al. 2016b; Kern et al. 2019), calibrating out reflections at low delays that bleed into the main foreground lobe is harder, and thus better suited to correction via standard gain calibration.

The ideal compromise, then, is to smooth our gains to keep the gain kernel at low delays and suppress its power at delays that we no longer trust its response. For the calibration at hand, this seems to be at roughly 100 nanoseconds, which enables the calibration to pickup on the dish reflection at 50 ns but suppresses the spurious terms in the gains at 150 ns and beyond. Given our 100 MHz bandwidth with 1024 channelization, a maximum delay range of 100 ns leads to about 15 free delay modes in the smoothed gains, which can be thought of as a smoothed gain with 15 spectral degrees of freedom per antenna and dipole polarization. Applying this gain to the data (last two panels in Figure 3.10), we see that we recover the best of both scenarios: the dish reflection is suppressed as desired and we also do not spread more instrument coupling at intermediate and high delays over what is already present in the data. To perform the smoothing we use the same delay-domain deconvolution technique described before as a low-pass Fourier filter, which is useful given that the gains are also flagged at certain frequency channels due to RFI. Although this calibration is performed for a single time, one can also take time and frequency dependent calibration solutions and smooth across both the temporal and spectral axes with this technique.

We can also show the effects of the smooth and full calibration on the full dataset. We do this by applying the calibration to the data and transforming them to the delay-domain in a similar manner as was done for Figure 3.8. In this case, Figure 3.11 plots this for the smooth calibrated data (left), the full calibrated data (center) and their fractional residual (right). Note that the calibrated data are plotted on the same colorscale as Figure 3.8, but are plotted with a smaller delay range to highlight

[3] This happens because the gains are a multiplicative term, meaning that although the systematics originally occupied $f \sim 0$ Hz, the contaminated gains spread them to $f > 0$ Hz modes.

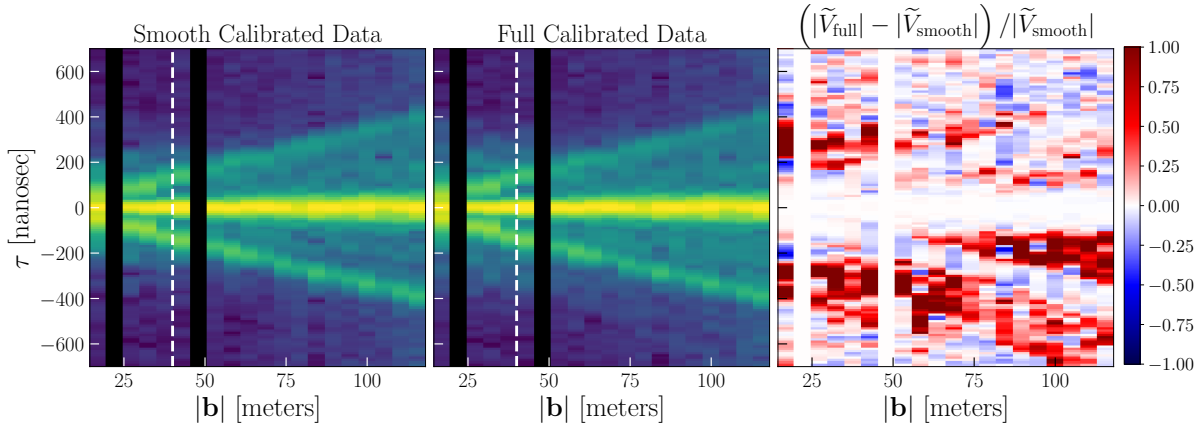

Figure 3.11: Redundantly-averaged pseudo-Stokes I visibilities in delay space transformed across a 120 – 180 MHz spectral window, ordered according to baseline length. We show the smooth calibrated data (left), full calibrated data (center) and their fractional residual. The calibrated data (left and center) are plotted on the same colorscale as Figure 3.8 over a smaller delay range to highlight the features within the foreground wedge. Within the smoothing scale of 100 ns, the fractional residual shows the two are in good agreement as expected. Outside the smoothing scale, however, the residual shows significant excess structure (red) in the full calibrated data not seen in the smooth calibrated data, which suggests that the structures are not real and are errors in the gain solution.

features within the foreground wedge. We see that the two calibrations achieve a good match at low delays, as expected, but for delays beyond the smoothing scale we find that the full calibrated data has significant excess structure (red) compared to the smooth calibrated data. This is indicative of the full calibration *introducing* spectral features into the data, rather than calibrating them out, which is highly suggestive of gain errors on these scales and further motivates the $\tau \sim 100$ ns smoothing scale of the gains derived in section 3.2.

Philosophically, this kind of approach to gain calibration, in other words keeping only degrees of freedom like low delay modes that we trust and filtering out the rest, is conservative from the perspective of not introducing structure into the data that was not already there. The cost of this approach is that we are not calibrating out gain structure at these delays inherently introduced by the instrument, if it exists in the first place. At the moment, however, we do not really have much of a choice: providing a constrained calibration with a few degrees of freedom is the best we can currently do, and until we have evidence that structure in the gain kernel at higher delays is real gain structure, we should not attempt to calibrate it out. This approach makes interpreting a fiducial detection in the power spectrum at similar, intermediate delays somewhat convoluted, and a suite of null tests and jacknives will be necessary to try to tease out whether said detection is residual calibration structure or real sky structure.

The obvious question moving forward for HERA then is, do we believe there to be true gain structure at low and intermediate delays that we need to calibrate out? The answer to this depends on the required dynamic range. For low delays we generally need 10^5 in dynamic range performance

of the gain kernel due to the foreground-to-EoR amplitude ratio: for larger delays this requirement becomes more stringent as the EoR signal is expected to weaken. Therefore, do we think HERA has true gain structure at some level above -50 dB at τ ~200 ns? Based on simulations (Fagnoni et al. 2019) and a rough extrapolation of Figure 3.9 the answer is probably yes, and therefore, we need a way to remove the cross coupling systematics from the data *before* performing antenna gain calibration. Cross coupling systematic removal is done by applying a high-pass filter in fringe-rate space (Kern et al. 2019; Kolopanis et al. 2019). This removes cross coupling, which occupies low fringe rates, but it also removes a component of the foregrounds as well, which we need for calibration. Doing this only on the data and not on the model would create a discrepancy in the data that would act as its own form of systematic. Fringe rate filtering therefore needs to be done on the model and data before calibration in order to probe the true instrument gain kernel to higher and higher delays. Achieving high fringe-rate resolution for a high-pass filter means simulating a large LST coverage with a wide-field flux density model. Unfortunately, the CASA-based calibration methodology presented in this work does not easily lend itself to this as it only reliably simulates short time intervals near the calibration field. This kind of analysis is best done using a numerical visibility simulator with wide-field diffuse and point-source maps, which we defer to future work.

Other smoothing algorithms have been investigated in the literature, which has been motivated due to a recent understanding of how incomplete sky models cause gain errors in sky-based calibration (Barry et al. 2016; Byrne et al. 2019) and non-redundancies cause gain errors in redundant calibration (Ewall-Wice et al. 2017; Orosz et al. 2019). The MWA, for example, uses low-order polynomials to smooth their sky-based gain solutions to limit gain error spectral structure in 21 cm power spectral analyses (Beardsley et al. 2016; Barry et al. 2019a). The reason we opt for direct Fourier filtering of the gains in this work is because a truncated polynomial basis is not able to encapsulate arbitrary gain fluctuations on large scales; in other words they do not form a complete basis in the Fourier domain for low-delay modes. This is fine for mitigating small-scale structure but means one runs the risk of not calibrating out large-scale modes that can cause biases in narrow band power spectrum analyses, although in simulated MWA analyses there is no evidence for such biases (Barry et al. 2016).

3.3.2 Temporal Response

In this section we use the data to assess the temporal stability of the instrumental gain. HERA observations are taken in drift-scan mode, meaning the array does not change or move over the course of observations. This lends itself to a fairly stable instrument as a function of time, and we therefore do not expect large deviations in the gains over short time intervals. However, effects such as ambient temperature drift and the cooling cycle within signal chain nodes are known to cause slight drifts in the calibration over the course of a night (e.g. Jacobs et al. 2013). In this section we investigate the data to quantify the amplitude of these gain drift terms and confirm they can be calibrated out if necessary. Note we do not actually apply time-dependent gains to the data in the remainder of this work: we merely present ways in which these terms can be calibrated-out for deep integrations if necessary.

All signal chains in the HERA array are brought via coaxial cable to an RFI-shielded and air-

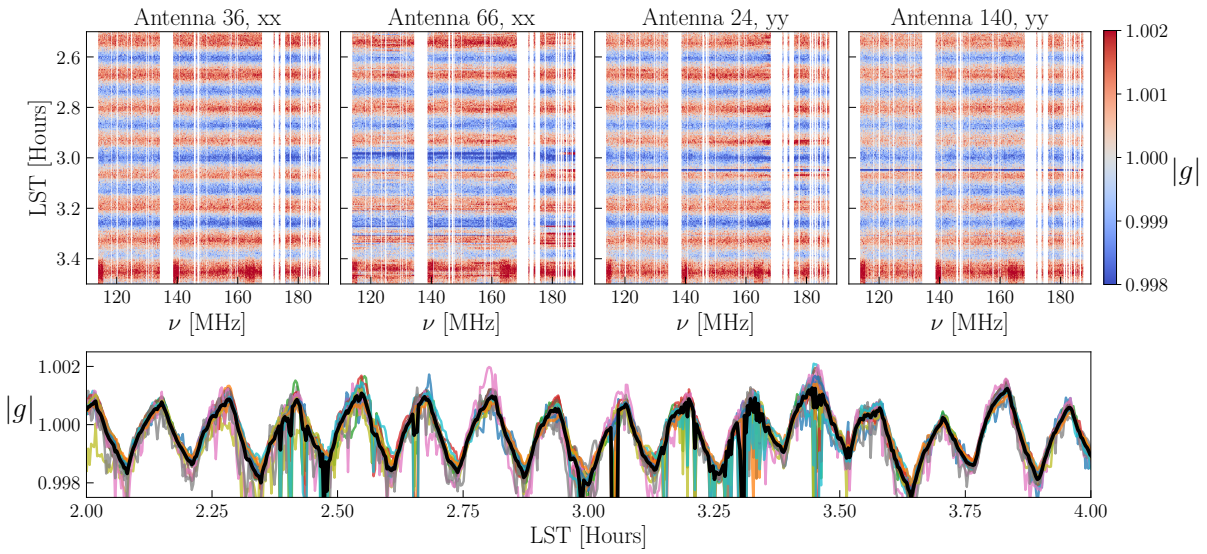

Figure 3.12: Temperature oscillations in the instrumental gain due to an air-conditioning cycle in the field container housing the ADC are a 0.1% effect. Top panels show the square-root of the ratio of the raw auto-correlations to the time-smoothed auto-correlations for a few antennas and both XX and YY polarization. The oscillation looks to be of roughly the same amplitude across different antennas, polarizations and frequencies. The bottom panel shows the frequency-averaged oscillation for a handful of antennas (colored lines) and their average (black). This shows a saw-tooth time profile that also matches temperature data collected in the container.

conditioned container in the field where the data are converted from analog to digital signals and are then correlated. Due to the air-conditioning cycle within this container, which cycles at roughly a 6-minute period, we expect the overall amplitude of the gains to drift at the same timescales. We can estimate the amplitude of this drift using a smoothed version of the auto-correlations, which is the approach adopted by the LWA (Eastwood et al. 2019) who face the same issue. Assuming that the only temporal structure in the auto-correlations occurs intrinsically at the time-scale of the beam crossing time (∼40 minutes), we can probe time structure from the gains by taking a time-smoothed version of the auto-correlation and dividing it by the un-smoothed auto-correlation. We smooth a handful of auto-correlation visibilities on a 20-minute timescale, divide their un-smoothed visibility counterparts by them and take their square-root, which leaves us with a set of ratio waterfalls as a function of frequency and time for each antenna-polarization. We show some of these in Figure 3.12, which plots the square-root ratio for each time and frequency bin for four antennas (top row) and also their frequency-average as a function of LST (bottom panel). We see that the gain fluctuations induced by the air-conditioning cycle in the container has a coherent phase and amplitude across all antennas and polarizations, and is also fairly constant across frequency. The frequency-average of each antenna and their respective average is shown to reflect a sawtooth profile as a function of time, whose profile inversely matches temperature data collected within the container. Figure 3.12 shows us that the 6-minute gain oscillations are a very small effect at

the 0.1% level, and can be decently well-calibrated by a single number as a function of time for all antennas, polarizations and frequency channels in the array. The HERA Phase II configuration will have a forced air cycling system that will better control fast temperature variations in container units.

A steady decrease in ambient temperature after sunset can cause slow evolution in the performance of the exposed part of the signal chains, in particular the low-noise amplifier in the FEM, which is attached to the feed. This kind of gain drift is expected to be slow but could add up over the course of an entire night of observing, especially if we choose to calibrate the data once at either the beginning or end of the night. To test this, we calibrate a single night of data at three different fields (Figure 3.2) at different times of a single night, and compare the average gain amplitude derived from each field. Figure 3.13 shows this drift having normalized the gains to the 2-hour field, demonstrating a slow drift that over the course of ~5 hours leads to about a 10% drift in the gain amplitude. Also plotted is the ambient temperature measured by a nearby weather station, which shows an expected inverse correlation with the antenna gain. Similarly, the band-averaged gain phase drift (after taking out the cable delay) is kept to within 0.2 radians over the same time interval, but unlike the average amplitude the phase drift does not appear monotonic in time.

Using the temperature data we can derive an ambient temperature coefficient for the change in the average gain as a function of temperature difference (Jacobs et al. 2013). We can represent a relationship for the difference in ambient temperature relative to the ratio of the derived gain response of the analog system as

$$10 \cdot \log_{10} \left| \frac{g_{\rm new}}{g_{\rm norm}} \right| \ [\rm dB] = C \cdot (T_{\rm new} - T_{\rm norm}) \ [\rm K], \tag{3.3}$$

where $T_{\rm norm}$ and $g_{\rm norm}$ are the ambient temperature and average gain amplitude at the time of gain calibration (i.e. our normalization time), $T_{\rm new}$ and $g_{\rm new}$ are the temperature and gain at any new time in LST, and C is the temperature coefficient in units dB $\rm K^{-1}$. In Figure 3.13, for example, we have chosen the normalization to be at 2 hours LST. Using the three data points from Figure 3.13, we derive a temperature coefficient of -0.031 dB $\rm K^{-1}$ for the gains. With a similar approach, Pober et al. (2012) also derive a gain temperature coefficient of -0.03 dB $\rm K^{-1}$ for the PAPER system, which used similar front-end hardware as HERA Phase I. Jacobs et al. (2013) also used data from two different seasons to derive an auto-correlation temperature coefficient for the PAPER system of -0.06 dB $\rm K^{-1}$, which, when divided by a factor of two in order to map it to a gain temperature coefficient, is also in agreement with these results.

3.4 Combining Redundant Calibration

A key part of HERA's design is to exploit its inherent redundant sampling of the *uv* plane for precision redundant calibration (Dillon & Parsons 2016). Redundant calibration asserts that all visibilities of the same baseline length and orientation (uniquely defining a "baseline type") measure the same visibility, which with enough redundant baselines allows for an overconstrained system of equation while keeping the true visibility a free parameter (Wieringa 1992; Liu et al. 2010). This

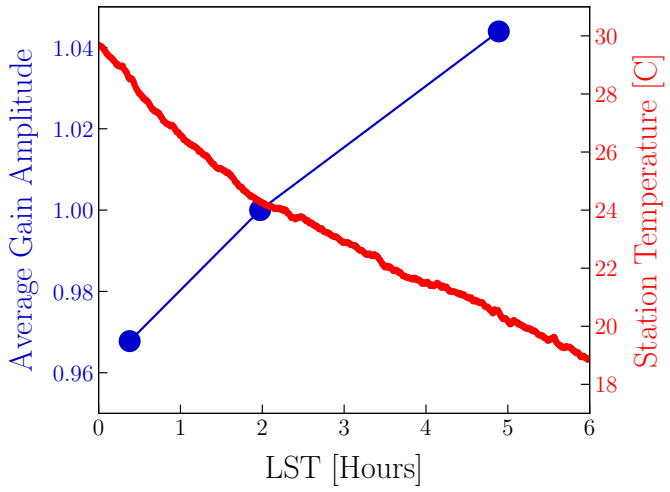

Figure 3.13: The average gain amplitude drift (blue) throughout the 2458098 observing night, derived from three independent calibration fields and normalized to the field at 2 hours. We also overplot the ambient temperature measured by a nearby weather station (red), showing an expected inverse correlation with the gain drift. Using Equation 3.3 these data yield a gain temperature coefficient of -0.031 dB K^{-1}.

means that redundant calibration does not need an estimate of the true model visibilities, and thus temporarily skirts some of the issues with incomplete or inaccurate sky models. In practice this is never exactly true, and slight antenna position and primary beam uncertainties therefore generate gain errors in redundant calibration (Ewall-Wice et al. 2017; Orosz et al. 2019). Nonetheless, we would like to explore options for combining redundant and absolute calibration to exploit their complementary advantages, either as an alternative or hybrid calibration pipeline.

For a baseline between antennas i and j and another between antennas j and k, both belonging to the same baseline type of ij (for example antenna pairs 23 & 24 and 24 & 25 from Figure 3.1), the redundant calibration equations are

$$\begin{aligned} V_{ij}^{\text{data}} &= g_i V_{ij}^{\text{model}} g_j^* + n_{ij} \\ V_{jk}^{\text{data}} &= g_j V_{ij}^{\text{model}} g_k^* + n_{jk} \\ &\vdots \end{aligned} \tag{3.4}$$

Note that the model visibility for V_{jk} is now V_{ij}^{model}. In this case, we are left with four free parameters, g_i, g_j, g_k, and V_{ij}^{model}, which we can solve for by minimizing their chi-square,

$$\chi^2 = \sum_{i,j} \frac{|V_{ij}^{\text{data}} - g_i V_{ij}^{\text{model}} g_j^*|}{\sigma_{ij}^2}, \tag{3.5}$$

where σ_{ij}^2 is the noise variance on baseline ij and the sum is over all antenna pairs in the array. Although a two-baseline array like the one in Equation 3.4 is not redundantly calibrate-able, we can

see that increasing the number of redundant baselines will turn this into an overconstrained system of equations (Liu et al. 2010). However, redundant calibration is not the final answer for antenna-based calibration, as there exist fundamental degeneracies that redundant calibration simply cannot constrain. One of these degeneracies is the average gain and model visibility amplitude. Looking at Equation 3.5, we can see that if we multiply all antenna gains by some fraction A, and then divide all model visibilities by A^2 we leave the final χ^2 unchanged. Recall we are free to do this because, unlike in sky-based calibration, the model visibility is a free parameter. Thus it can perfectly counteract such deviations in the gains and implies that the full system of equations is insensitive to their average amplitude. In addition to the average gain amplitude, the other major degeneracy associated with redundant calibration is known as the "tip-tilt" phase gradient across the East-West and North-South coordinates of the array (Zheng et al. 2014; Dillon et al. 2018). If each antenna is assigned a vector $\mathbf{r}_i$ originating from the center of the array to its topocentric coordinates of East & North, we can insert a "tip-tilt" phase gradient into the gains as

$$\begin{aligned} g_i &\rightarrow g_i \exp(i\mathbf{\Phi}\mathbf{r}_i) \\ \text{where } \mathbf{\Phi} &= (\Phi_E,\ \Phi_N)\,. \end{aligned} \tag{3.6}$$

The coefficient $\mathbf{\Phi}$ is therefore a phase gradient coefficient with units of radians per meter, with separate coefficients for the East and North directions. Such a perturbation to the gains is a degeneracy in redundant calibration because we can exactly cancel this out by applying the opposite term to the model visibilities. For example, we can express the second term in the chi-square metric of Equation 3.5 as

$$\begin{aligned} &g_i V_{ij}^{\text{model}} g_j^* \rightarrow \\ &\quad g_i \exp(i\mathbf{\Phi}\mathbf{r}_i) V_{ij}^{\text{model}} \exp(-i\mathbf{\Phi}\mathbf{r}_{ij}) g_j^* \exp(-i\mathbf{\Phi}\mathbf{r}_j) \\ &\quad = g_i V_{ij}^{\text{model}} g_j^* \exp(i\mathbf{\Phi}\mathbf{r}_{ij}) \exp(-i\mathbf{\Phi}\mathbf{r}_{ij}) \\ &\quad = g_i V_{ij}^{\text{model}} g_j^*, \end{aligned} \tag{3.7}$$

where we have made use of the fact that $\mathbf{r}_i - \mathbf{r}_j = \mathbf{r}_{ij}$, and can see that after substitutions the term is unchanged. This amounts to a total of three parameters, the average gain amplitude and the east and north phase gradient, that need to be solved for after redundant calibration and require a sky model to pin down (per frequency, time and polarization). Thus, the issues of inaccurate sky models are somewhat mitigated but not totally circumvented by redundant calibration (Byrne et al. 2019). From a sky-based calibration perspective, these degeneracies can roughly be thought of as the overall flux scale of the array and its pointing center on the sky.

There are multiple ways to fill-in the missing degenerate parameters of redundant calibration. One approach is to take the redundant calibration solutions and project only its degenerate components onto the degenerate modes in the sky-based calibration solutions (Li et al. 2018). One can also take model visibilities and setup a new calibration equation that solves explicitly for the degenerate parameters (partial absolute calibration). Finally, one can take the sky-based calibrations as a starting point by applying them to the data and then run redundant calibration. The latter two of these, along with standard sky calibration, are shown in Figure 3.14, outlining the

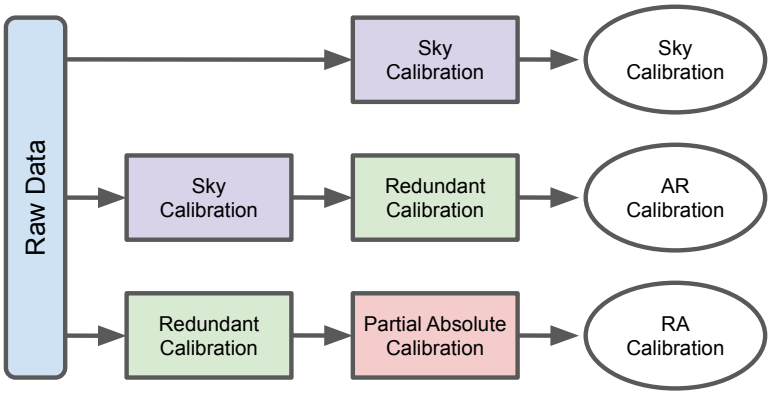

Figure 3.14: Schematic showing the order of operations for three related calibration strategies, similar to that of Li et al. (2018). For AR and RA calibration, the gains from the first step are applied to the data before proceeding to the second step. In addition, the gains derived by redundant calibration have their degenerate modes projected out before proceeding.

order of operations of the three proposed calibration schemes: sky calibration, sky + redundant (AR) calibration and redundant + partial absolute (RA) calibration. Both AR and RA calibration schemes are built into the `redcal` module of the `hera_cal` software package (which we use here), including setting up and solving a system of equations that specifically picks out the degenerate parameters of redundant calibration given a set of sky model visibilities, which we discuss in more detail in section 3.6. For the RA approach discussed here, the model visibilities used for extracting the degenerate modes are simply the raw data calibrated with the sky-based gains.

Because redundant calibration cannot constrain the degenerate modes inherent to its system of equations, the output gains will generally have some random combination of degenerate vectors, which will be influenced by the convergence of the calibration solver and its starting point from the raw data. To fix this, we can *project out* these degeneracies by fixing them to some a priori chosen position, which will then get filled in by absolute calibration (Dillon et al. 2018; Li et al. 2018). The simplest thing is to re-scale the gains such that the average amplitude is 1.0 and the phase gradient is 0.0, which is done to just the redundant calibration portion of the gains in both RA and AR calibration.

We saw in Figure 3.9 the presence of antenna-based structures that we expect to appear in the gains, like the dish reflection and the 20-meter and 150-meter cable reflections, but we also saw significant contamination by instrumental coupling across a wide range of delays. To understand the kinds of structures picked up by redundant calibration we can inspect the gains in a similar manner. Figure 3.15 shows the distribution of the gains at each step in the AR calibration scheme in amplitude (left) and in phase (right) having removed the cable delay for each antenna. It also shows the gains Fourier transformed across frequency in delay space (right) and peak-normalized. The top panels of Figure 3.15 show just sky calibration (the same as Figure 3.9). The middle panel shows just the redundant calibration component of the gain, where in deriving them we first apply the sky calibration gains to the data, and the bottom panels shows the final product of the

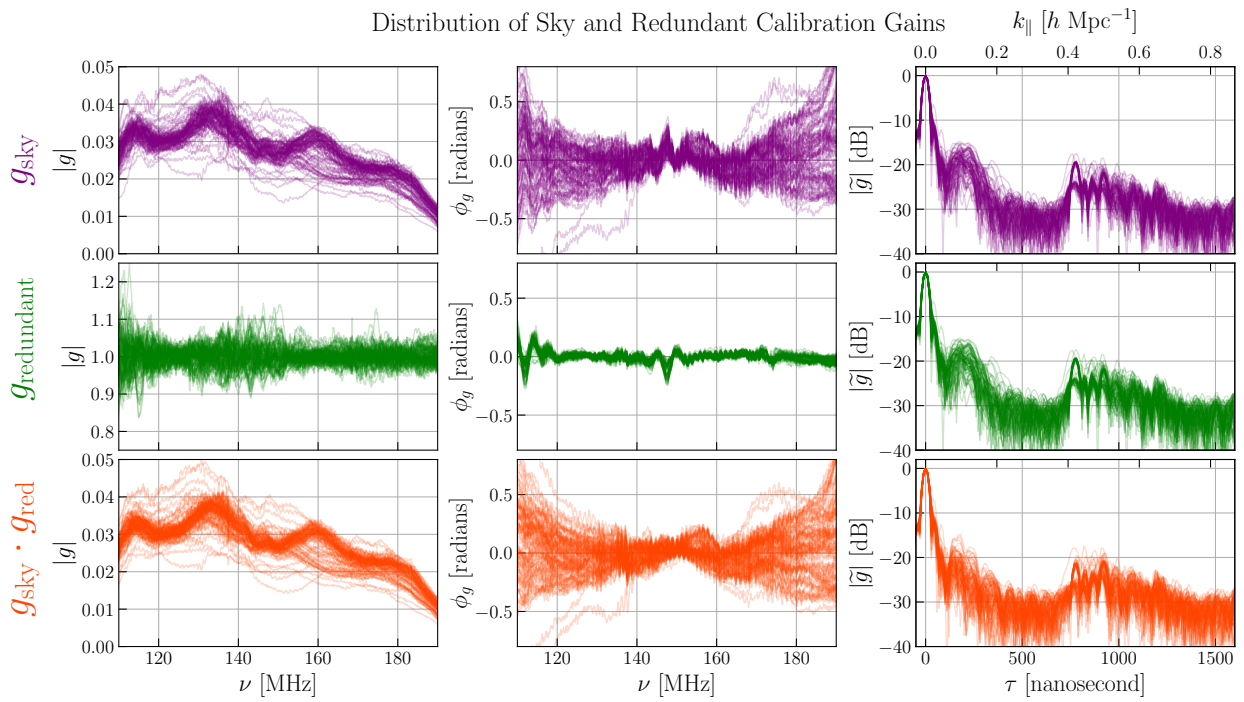

Figure 3.15: The distribution of gain solutions from AR calibration. The top panels show just the sky calibration (similar to Figure 3.9) in amplitude, phase (after removing their cable delay) and in delay space. Middle panels show just the redundant calibration portion of the gains in AR calibration. Bottom panels show the product of the two steps, which forms the full AR calibration gain solutions. Note the notch in the phase plots that is canceled out by redundant calibration, which leads to some suppression of the 200 nanosecond feature.

two gains. Note that the redundant calibration gains derived here use the same baselines as the sky calibration of $|\mathbf{b}| > 40$ meters. For the redundant calibration gains, we can see that its average amplitude is one as expected, and has similar kinds of spectral structure as the sky calibration gains. Looking at their product, or the AR calibration gains, we can see some of the benefits of redundant calibration. Compared to the sky calibration delay response (purple), the AR delay response (orange) has a slightly suppressed bump at ~200 ns, which can also be seen as the negation of the coherent ripple in the center phase plots. We observe this ripple in the sky-based gain phases (top-center), which seems to be corrected-for by redundant calibration (middle-center) such that their product (bottom-center) demonstrates less of a ripple. One possible explanation is that this ripple is caused by an imperfect sky model that creates spectral errors in the sky-based gain that is then corrected by redundant calibration. However, we still see significant power at $\tau \gtrsim 200$ ns, which could originate from non-redundancies between nominally redundant baselines specifically at the horizon, where diffuse emission generates the pitchfork effect in the data but also where the per-antenna primary beams are likely the least redundant with each other. Similar to how unmodeled diffuse emission created gain errors in sky calibration, these kinds of non-redundancies will create

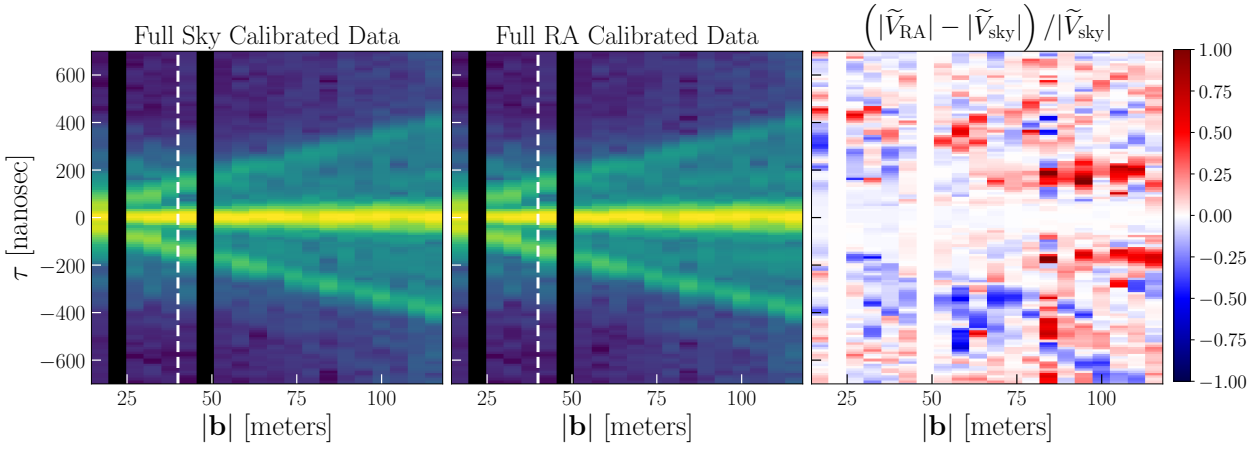

Figure 3.16: Redundantly-averaged pseudo-Stokes I visibilities in delay space transformed across a 120 – 180 MHz spectral window and ordered according to baseline length, having applied the full sky calibration gains (left) and the full RA calibration gains (center). These are plotted on the same colorscale as Figure 3.8. Taking their fractional difference (right) shows that the RA calibration introduces less structure into the data at $|\tau| \sim 200$ ns for shorter baselines (blue regions), although it also seems to introduce new structure at slightly smaller delays for long baselines (red regions).

errors in redundant calibration and will appear at similar delays (Orosz et al. 2019). Previous work showed that non-redundancy seems to be worse for short baselines (Carilli et al. 2018), but quantifying this in more detail is still in progress (Dillon et al. 2020). The AR calibration gains also show significant power at $\tau \gtrsim 800$ ns, which shows that redundant calibration is not immune to picking up cross coupling instrumental systematics. The RA gain solutions show nearly the same structure as solutions derived from AR calibration down to below 1% in fractional difference, so we do not plot them here for brevity.

To further show the effects of redundant calibration on the data, we take the full gains (this time from the RA calibration scheme) and apply them to the data. We then redundantly average and Fourier transform them similar to Figure 3.8. Figure 3.16 shows this process having applied the full sky calibration and full RA calibration, and also shows the fractional difference between the two (right). Areas where the RA calibration is introducing new structures show up as red, and areas where sky calibration is introducing structure where RA is not show up at blue. We see that for shorter baselines and delays near $|\tau| \sim 200$ ns, that RA calibration is inserting less structure into the data compared to sky calibration, which agrees with our observation earlier that spurious gain structure at those delay scales seem suppressed. However, we see that at slightly smaller delays and larger baseline lengths, RA calibration is also inserting additional power compared to sky calibration, which is likely a result of its own gain errors. Additionally, at small delays ($|\tau| < 100$ ns) the two are in good agreement with each other.

The take-away from this section is: 1) all three calibration schemes yield gains that are similar at low delays; 2) hybrid redundant calibration seems to correct for some of the errors in the sky-based calibration but still introduces its own set of errors; 3) both sky and redundant calibration suffer

from gain errors that are induced by baseline-dependent instrumental systematics. Moving forward, future analyses will benefit from attempting to model diffuse emission and removing instrumental cross coupling systematics before calibration in order to calibrate intermediate delay scales and exploit the full power of a combined redundant and absolute calibration approach.

Li et al. (2018) performed a similar comparison with the MWA, using the Fast Holographic Deconvolution package (FHD; Sullivan et al. 2012) for sky-based calibration and the `omnical` package (Zheng et al. 2014) for redundant calibration. Similar to this work, they find marginal improvements with a combined sky + redundant calibration approach.

3.5 Power Spectrum Performance

We use the visibility-based, delay spectrum estimator of the 21 cm power spectrum to further assess the quality of the calibration and the overall stability of the array. The delay transform is simply the Fourier transform of the visibilities across frequency into the delay domain

$$\widetilde{V}(\mathbf{u}, \tau) = \int d\nu \; e^{2\pi i \nu \tau} V(\mathbf{u}, \nu), \tag{3.8}$$

where $\mathbf{u} = \mathbf{b}/\lambda$ is the uv vector of the baseline and λ is the observing wavelength (Parsons et al. 2012a; Liu et al. 2014a; Parsons et al. 2014). The Fourier dual of frequency, τ, is not a direct mapping of the line-of-sight spatial wavevector $k_\parallel$ but under certain assumptions it is a fairly good approximation. This is known as the "delay approximation" and has been shown to be fairly accurate for short baselines (Parsons et al. 2012a). The delay spectrum estimate of the 21 cm power spectrum is the delay transformed-visibilities squared, multiplied by the appropriate scaling factors,

$$\widehat{P}_{21}(k_\perp, k_\parallel) \approx |\widetilde{V}(\mathbf{u}, \tau)|^2 \frac{X^2 Y}{\Omega_{pp} B_p} \left(\frac{c^2}{2 k_B \bar{\nu}^2} \right)^2, \tag{3.9}$$

where X and Y convert angles on the sky and delay modes to cosmological length scales, Ω_{pp} is the sky-integral of the squared primary beam, $\bar{\nu}$ is the average frequency in the delay transform window and B_p is the delay transform bandwidth, as defined in Appendix B of Parsons et al. (2014). The relationships between the Fourier domains inherent to the telescope, $\mathbf{u}$ and τ, and the cosmological Fourier domains are

$$\begin{aligned} k_\parallel &= \frac{2\pi}{X} \tau \\ k_\perp &= \frac{2\pi}{Y} \frac{b}{\lambda}, \end{aligned} \tag{3.10}$$

where $X = c(1+z)^2 \nu_{21}^{-1} H(z)^{-1}$, $Y = D(z)$, $\nu_{21} = 1.420$ GHz, $H(z)$ is the Hubble parameter, $D(z)$ is the transverse comoving distance, b is the baseline length and λ is the observing wavelength (Parsons et al. 2012b; Liu et al. 2014a). For this analysis, we adopt a ΛCDM cosmology with parameters derived from the *Planck* 2015 analysis (Planck Collaboration et al. 2016), namely $\Omega_\Lambda = 0.6844$, $\Omega_b = 0.04911$, $\Omega_c = 0.26442$ and $H_0 = 67.27$ km/s/Mpc.

Due to the chromaticity of an interferometer, foreground emission that is inherently spectrally smooth (such as galactic synchrotron) will have increased spectral structure in the measured visibilities. The delay at which the instrument imparts this spectral structure is dependent on the geometric delay of the source signal between the two antennas that make up a baseline, given as

$$\tau = \frac{|\mathbf{b}| \sin(\theta)}{c}, \tag{3.11}$$

where θ is the zenith angle of the incident foreground emission and $\mathbf{b}$ is the baseline separation vector. We can see that spectrally-smooth foregrounds incident from zenith will appear at lower delays and therefore have less induced chromaticity, while foregrounds incident from large zenith angles will have more induced chromaticity. The maximum delay a smooth spectrum foreground can appear at is called the horizon limit, in which case $\tau_{\text{horizon}} = \tau(\theta = 90°)$. If we could perfectly image the interferometric data we could also reconstruct the smooth spectrum foregrounds. However, this is in practice never the case, as effects like missing uv samples and imaging via gridded Fourier transforms create low-level chromatic sidelobes that corrupt the images with spectrally-dependent residual foregrounds. Visibility-based power spectrum estimators that do not even attempt to image the data are stuck with the most severe amounts of instrument-induced chromaticity, generally out to the baseline horizon delay. The horizon limit is a function of baseline length (Equation 3.11), and as such it forms a wedge-like shape in the data's Fourier domain and has come to be known as the foreground wedge (Datta et al. 2010; Morales et al. 2012; Parsons et al. 2012a; Thyagarajan et al. 2013; Liu et al. 2014a; Morales et al. 2019). Because HERA has a fairly compact primary beam we expect most foreground power to lie within $\tau \leq |\mathbf{b}| \sin(\theta = 5°)/c$; however, the vast amounts of diffuse emission near the horizon means that we still expect to see some amount of foreground power out to the horizon limit, even though it is significantly attenuated by the primary beam (Thyagarajan et al. 2016).

The issue of whether foregrounds actually appear tightly confined within the foreground wedge is an open question: 21 cm foreground studies seem to indicate that *supra-horizon* foreground power tends to extend only slightly beyond the horizon (Pober et al. 2013a; Bernardi et al. 2013; Gehlot et al. 2018; Lanman et al. 2020), but whether this is truly the case down to EoR sensitivities is not known. There are a number of effects that can contribute to measured supra-horizon foreground emission, including intrinsic foreground spectral structure, unflagged RFI, primary beam chromaticity, and also gain calibration errors. As discussed in section 3.3, the intrinsic gain kernel of the instrument may have a non-negligible extent to large delay modes, which if left uncalibrated will push foreground power out to higher delays. Similarly, gain errors will introduce structure at these scales and have the same effect. Smoothing the gains eliminates the latter concern but still leaves the possibility of the former effect. To asses the degree of foreground containment we can form wide-band, visibility-based power spectra as a diagnostic.

This is complemented by an understanding of how thermal noise appears in the power spectra. Given our knowledge of the noise properties of our antennas, we can compute a theoretical estimate of the noise power spectrum, P_{N}, which is equivalent to the root-mean square (RMS) of the power spectrum if the only component in the data were noise. This is one measure of the uncertainty on the power spectra due to noise, but also represents the theoretical amplitude of the power spectra

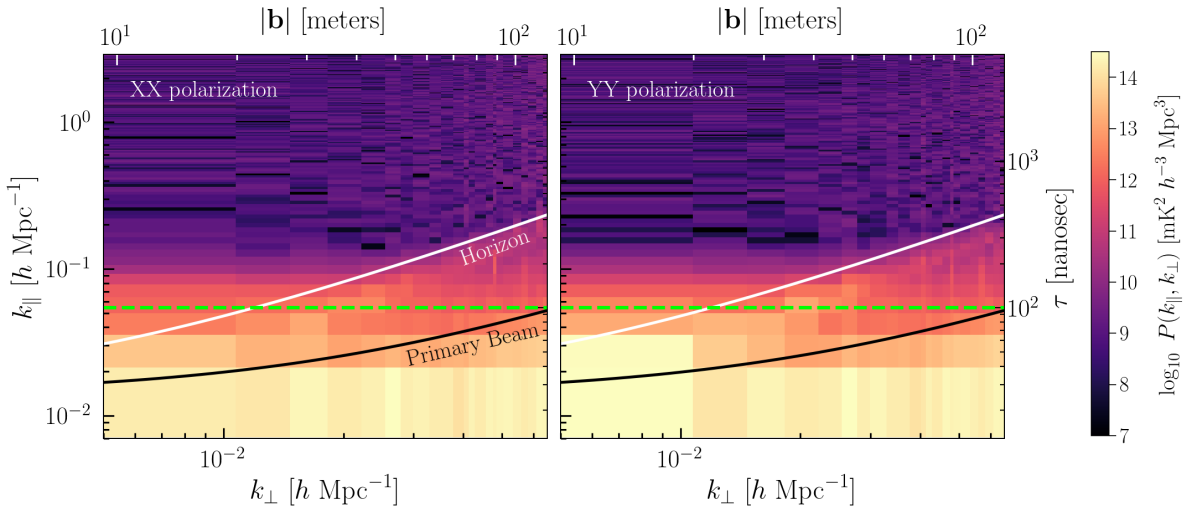

Figure 3.17: Wide-band, two-dimensional power spectra of each linear dipole polarization XX (left) and YY (right) having applied the smooth sky-based calibration, and after systematic removal and an incoherent average (i.e. after squaring the visibilities) from 0 to 2 hours LST. Power spectra are formed between 139 – 178 MHz having applied a Blackman window to limit spectral leakage in the discrete Fourier transform. The black line marks the FWHM of the primary beam (±5° from zenith) and the white line marks the baseline horizon. Both lines have an additive buffer of $k_\parallel = 0.014$ h Mpc^{-1} to account for the width of the Blackman kernel in Fourier space. The dashed green line marks the maximum delay scale of the smoothed gain solutions. Most of the foreground power is confined within the horizon limit of the array, however there is evidence for some supra-horizon leakage at short baselines.

in the limit that they are noise dominated (as opposed to signal or systematic dominated). This is given in Cheng et al. (2018) as

$$P_{\rm N} = \frac{X^2 Y \Omega_{\rm eff} T_{\rm sys}^2}{t_{\rm int} N_{\rm coherent} \sqrt{2 N_{\rm incoherent}}}, \tag{3.12}$$

where the X and Y scalars are the same as before, $T_{\rm sys}$ is the system temperature in milli-Kelvin, $t_{\rm int}$ is the correlator integration time in seconds, $N_{\rm coherent}$ is the number of sample averages done at the visibility level (i.e. before visibility squaring), and $N_{\rm incoherent}$ is the number of sample averages done at the power spectrum level (i.e. after visibility squaring). $\Omega_{\rm eff}$ is the effective beam area given by $\Omega_{\rm eff} = \Omega_p^2/\Omega_{pp}$, where Ω_p is the integral of the beam across the sky in steradians, and Ω_{pp} is the integral of the squared-beam across the sky in steradians (Pober et al. 2013a; Parsons et al. 2014). Using similar data products, Kern et al. (2020b) showed that the HERA Phase I system achieves an antenna-averaged $T_{\rm sys} \sim 250$ K at 160 MHz, which we adopt in this work.

The raw data are flagged for radio frequency interference (RFI) and are thus nulled at the flagged channels. This leads to a highly discontinuous windowing function that when taking a Fourier transform will spread foreground power and contaminate the EoR window. To prevent this, we employ the same 1D delay domain deconvolution as the gain smoothing filter (section 3.3) on

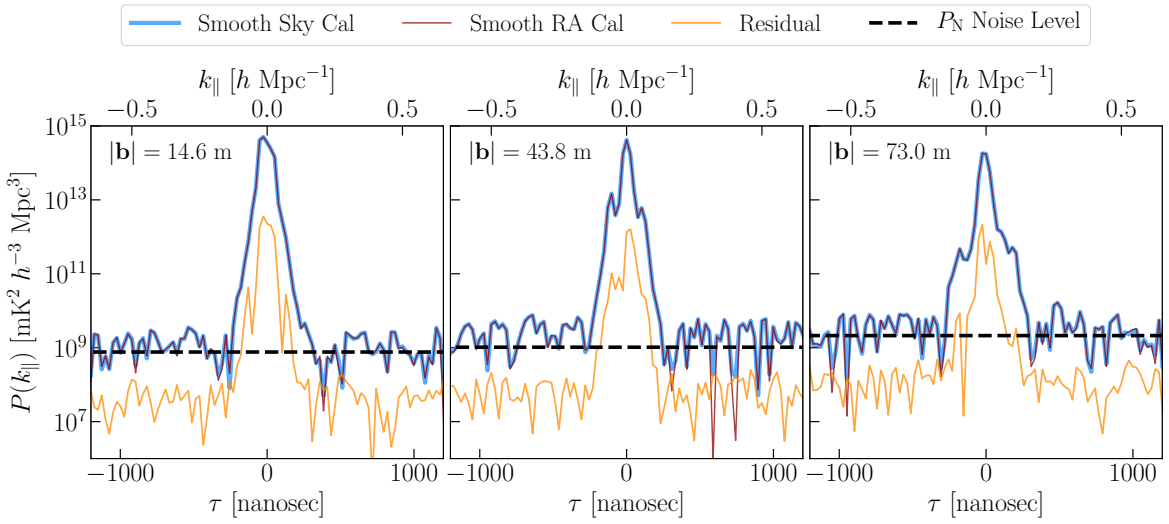

Figure 3.18: Delay spectra of three redundantly averaged East-West baseline types for the instrumental YY polarization, showing the data calibrated with the smooth sky calibration (blue), the smooth RA calibration (red) and their residual (red), along with the thermal noise floor (dashed-black) assuming a T_{sys} = 250 K. The two calibration yield nearly the same averaged power spectra across all delays, which show consistency with the theoretical noise floor outside $k_\parallel \gtrsim 0.2\ h$ Mpc^{-1}.

each visibility, filling in model CLEAN components out to 2000 nanoseconds. HERA Phase I data are contaminated by cable reflection and cross coupling instrumental systematics (Kern et al. 2020b). Because we are concerned with foreground leakage due to calibration in this work, we remove these systematics before forming power spectra for visual clarity. Specifically, we apply a time-domain filter to suppress cross coupling in all visibilities with a projected East-West length greater than 14 meters (throwing out all other visibilities). This time-domain filter is performed in the delay domain, and isolates a rectangle spanning $|\tau| > 0.8\tau_{\text{horizon}}$ and fringe-rates given by the 99% EoR power bounds in Kern et al. (2019) for each baseline independently. We also calibrate out a single cable reflection term for each of the 20-meter and 150-meter cables in the analog system per dipole polarization also using the methods in Kern et al. (2019).

We form power spectra across a spectral window from 139 – 178 MHz and apply a Blackman window prior to taking the Fourier transform to limit spectral leakage in the discrete Fourier transform. Baselines are only cross-multiplied with themselves, and are not cross-multiplied with other baselines in a redundant group. Normally this would produce a noise-bias in the power spectra, so instead we cross-multiply baselines with themselves at adjacent time integrations, having first rephased them to the same pointing center (Pober et al. 2013a). We do this for all baselines for each time integration pair in the range of 0 – 2 hours LST. After squaring the visibilities, we incoherently average the power spectra across LST and then average all power spectra of the same baseline length (regardless of orientation), which is equivalent to cylindrically gridding $\mathbf{k}$ space into $k_\perp$ and $k_\parallel$ annuli.

Figure 3.17 shows the 2D power spectra in instrumental XX and YY visibility polarizations

with the smooth sky calibration gains applied to the data (smoothed out to $\tau = 100$ ns). We also show the primary beam FWHM limit (black) and the full horizon limit (white) in both instrumental XX and YY visibility polarization. Both lines have an additive buffer of $k_{\parallel} = 0.014\ h\ \mathrm{Mpc}^{-1}$ to account for the width of the Blackman kernel in Fourier space. The dashed green line shows the maximum delay scale of the applied gains after smoothing. We find that most of the foreground power is contained within the horizon limit, with some amounts of supra-horizon leakage for short baselines. The strong pitchfork feature of the foreground emission tracing the horizon line is not as prominent in this plot as it was in Figure 3.8, which is due to the fact that it was partially removed in the cross coupling filter applied to the data. Kern et al. (2020b) showed that the edges of the pitchfork are slowly time variable and thus can be separated from the cosmological 21 cm signal and filtered out with a high-pass time filter. Figure 3.10 also demonstrates this, showing that the two lobes at $\pm\tau = 100$ ns are also centered at $f = 0$ mHz, meaning they primarily contain slowly time-variable terms. This means that the time filter designed to eliminate cross coupling also helps to reduce some of the strongest foreground emission straddling the boundary of the foreground wedge and EoR window.

To first order, Figure 3.17 tells us that our single field, smoothed sky-based calibration with restricted degrees of freedom has done a fairly good job calibrating the data and has largely kept foreground power contained within the foreground wedge. For short baselines, however, we can begin to see evidence for some amount of supra-horizon emission that could be due to uncalibrated gain terms or imperfectly removed cross coupling, the latter of which is harder to remove for shorter baselines. This supra-horizon emission is located beyond the smoothing scale of the gains and appears in amplitude slightly larger than predictions of the high-order dish reflections (Patra et al. 2018), but is contained within $k_{\parallel} \lesssim 0.2\ h\ \mathrm{Mpc}^{-1}$ down to nearly $\sim 10^6$ in dynamic range against the foreground peak. Note that, possibly coincidentally, this supra-horizon excess seems more prevalent for baselines shorter than our initial baseline cut of 40 meters. Deeper integrations will help to discriminate whether the observed supra-horizon emission extends further out in $k_{\parallel}$ space at lower noise levels.

Figure 3.18 shows the same power spectra but focuses on three unique baseline types: purely East-West baselines of 14.6 m, 43.8 m, and 73 m in length. In addition to showing power spectra of the data with the smooth sky calibration (blue), we also show the smooth RA calibration (also smoothed out to $\tau = 100$ ns) and the residual between the two. This demonstrates that the calibration strategies, post-smoothing, have nearly the same impact on the averaged power spectrum. We also show the theoretical noise-floor of the data (dashed-black), which more clearly demonstrates the agreement of the data with the noise floor outside $k_{\parallel} \gtrsim 0.2\ h\ \mathrm{Mpc}^{-1}$. Note that the noise floor for longer baseline types is higher because there are fewer physical baselines, meaning less averaging is done in the (coherent) redundant average.

3.6 Partial Absolute Calibration

Partial absolute calibration is the process of taking a set of sky model visibilities and setting up a system of equations that solves for just the degenerate components of redundant calibration. The

number of degenerate modes in redundant calibration depends on the kind of redundant calibration being employed (Dillon et al. 2018). Here we discuss the degeneracies associated with the "2-pol" scheme, which calibrates the X and Y dipoles separately and ignores cross-feed polarization terms. As shown in section 3.4, there are three main degeneracies in redundant calibration *for each dipole polarization*: the average gain amplitude (or the absolute flux scale of the instrument) and a "tip-tilt" phase gradient as a function of distance from the center of the array for both the East and North spatial axes (or the overall pointing center of the instrument). Each of these parameters has an arbitrary frequency dependence, meaning that various kinds of spectral structure can occupy these degenerate modes. We can express these parameters in the ith antenna gain of the X dipole as

$$g_{i,X}(\nu) = \exp(\eta_{\mathrm{abs},X}(\nu) + i2\pi\nu(T_{E,X}r_{i,E} + T_{N,X}r_{i,N}) + i(\Phi_{E,X}(\nu)r_{i,E} + \Phi_{N,X}(\nu)r_{i,N})), \tag{3.13}$$

where $r_{i,E}$ is the East distance of antenna i from the center of the array in meters and we have explicitly included the frequency dependence of the gain and its parameters. Note that we have redefined the phase component into the sum of two terms, a spatial delay gradient $\mathbf{T}_X = (T_{E,X}, T_{N,X})$, and a spatial phase gradient $\mathbf{\Phi}_X = (\Phi_{E,X}, \Phi_{N,X})$. Note that the delay gradient parameter has units of nanoseconds / meter and is itself frequency-independent, but has the effect of creating a phase slope in the gain across frequency. The delay gradient manifests as a delay plane across the array that sets the phase center. It forms a subspace of the original phase gradient space so we simply pull it out and redefine the phase gradient term Φ to be a deviation about the delay plane. This is important because when we go to solve the calibration equation we want the phase measurements to be near zero or at least considerably less than 2π to mitigate phase wrapping (Liu et al. 2010). Phase wrapping creates local minima that confuse the calibration phase solver, which can be alleviated through pre-conditioning of the system of equations by first solving for and eliminating the delay gradient term.

Using a logarithm to linearize the calibration equation, the average antenna amplitude for the X dipole is found by solving the following system of equations

$$\mathbf{y} = \begin{pmatrix} \ln\left|\frac{V^{\mathrm{data}}_{ij,XX}}{V^{\mathrm{model}}_{ij,XX}}\right| \\ \ln\left|\frac{V^{\mathrm{data}}_{jk,XX}}{V^{\mathrm{model}}_{jk,XX}}\right| \\ \vdots \end{pmatrix} = \mathbf{A}\hat{\mathbf{x}} = \begin{pmatrix} 2 \\ 2 \\ \vdots \end{pmatrix} \begin{pmatrix} \hat{\eta}_{\mathrm{abs},X} \end{pmatrix}, \tag{3.14}$$

where we have specified now that the visibilities are from the XX instrumental polarization. We use the linear and non-linear equation solving package `linsolve` to solve these systems of equations.

The delay gradient parameter can be isolated by taking the phase of the data-model ratio:

$$\mathrm{angle}\left(\frac{V^{\mathrm{data}}_{ij,XX}}{V^{\mathrm{model}}_{ij,XX}}\right)(\nu) = 2\pi\nu\mathbf{T}_X\mathbf{r}_{ij}, \tag{3.15}$$

where the angle$(\times)$ operator is $\tan^{-1}(\mathrm{Im}(\times)/\mathrm{Re}(\times))$. However, we can see that the delay gradient parameter is not inherently a function of frequency, so instead of solving this equation at each

frequency we should re-cast it in a form that is frequency independent and then solve that equation. This can be expressed as

$$\text{delay}\left(\frac{V_{ij,XX}^{\text{data}}}{V_{ij,XX}^{\text{model}}}\right) = \mathbf{T}_X \mathbf{r}_{ij}, \tag{3.16}$$

where the delay($\times$) operator takes the Fourier transform of its argument and identifies the delay of its peak in amplitude via quadratic interpolation of the three strongest Fourier modes. The system of equations for the delay gradient of the X dipole is then

$$\mathbf{y} = \begin{pmatrix} \text{delay}\left(\frac{V_{ij,XX}^{\text{data}}}{V_{ij,XX}^{\text{model}}}\right) \\ \text{delay}\left(\frac{V_{jk,XX}^{\text{data}}}{V_{jk,XX}^{\text{model}}}\right) \\ \vdots \end{pmatrix} = \mathbf{A}\hat{\mathbf{x}} = \begin{pmatrix} r_{ij,E} & r_{ij,N} \\ r_{jk,E} & r_{jk,N} \\ \vdots & \end{pmatrix} \begin{pmatrix} \hat{T}_{E,X} \\ \hat{T}_{N,X} \end{pmatrix}. \tag{3.17}$$

The estimated delay gradient gain is then expressed as $\hat{g}_{i,X}(\nu) = \exp(i2\pi\nu\hat{\mathbf{T}}_X\mathbf{r}_i)$.

After dividing the data visibilities by the estimated delay gradient gains, we can solve for the leftover phase gradient parameter for the X dipole with the system of equations

$$\mathbf{y} = \begin{pmatrix} \text{angle}\left(\frac{\widetilde{V}_{ij,XX}^{\text{data}}}{V_{ij,XX}^{\text{model}}}\right) \\ \text{angle}\left(\frac{\widetilde{V}_{jk,XX}^{\text{data}}}{V_{jk,XX}^{\text{model}}}\right) \\ \vdots \end{pmatrix} = \mathbf{A}\hat{\mathbf{x}} = \begin{pmatrix} r_{ij,E} & r_{ij,N} \\ r_{jk,E} & r_{jk,N} \\ \vdots & \end{pmatrix} \begin{pmatrix} \hat{\Phi}_{E,X} \\ \hat{\Phi}_{N,X} \end{pmatrix}, \tag{3.18}$$

where $\widetilde{V}^{\text{data}}$ denotes the fact that we have first divided the data visibilities with the delay gradient gain (or equivalently multiplied the model visibilities by the delay gradient gain). The average amplitude parameter, being orthogonal to the phase parameters, does not necessarily need to precede these steps. The full partial absolute calibration gain is then simply Equation 3.13 filled with our estimates of the degenerate parameters.

One interesting feature about delay and phase gradient calibration is that they do not require a reference antenna. Because phase is a periodic coordinate system, sky-based phase calibration requires that we select a reference antenna whose phase is identically zero, which is a way to constrain the overall phase parameter which does not have a physical meaning. If the array coordinates are defined in East-North-Up coordinates, then for delay and phase gradient calibration we can change the overall phase of the gain solutions by moving the estimated delay plane or phase plane up or down along the Up axis (i.e. the z-axis if the array is defined in X-Y-Z coordinates). Moving the delay and phase planes up or down does not change the relative delay and phase between antennas, which is really what we care about. We can pin this free parameter by selecting the Up-axis intercept of the plane, which is equivalent to setting the origin of the $\mathbf{r}_i$ vector coordinates that, up to now, we have defined as the center of the array. We can see now that setting it at the center

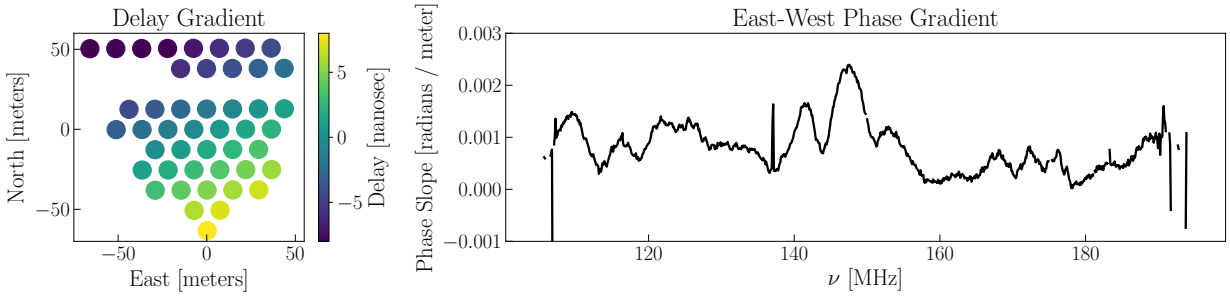

Figure 3.19: The delay gradient (left) and East-West phase gradient (right) derived by partial absolute calibration for the X dipole polarization using the GLEAM-02H field flux density model. We observe a significant amount of spectral structure in the phase gradient parameter, meaning it cannot be overlooked in partial absolute calibration.

of the array is not strictly required, but it does make computations easier if done so. Therefore, the act of setting the origin of the antenna position coordinate system plays the same role in delay and phase gradient calibration as the reference antenna does in standard sky-based phase calibration.

In Figure 3.19 we show the derived delay gradient (left) and phase gradient (right) parameters across the array for the X dipole polarization. The phase gradient terms shows significant amounts of spectral structure, highlighting its ability to pickup on non-trivial spectral terms in the data. For a large, 350+ element array these steps may take too long to calibrate the data in real time using all $\sim N^2$ baselines. However, for partial absolute calibration we may get away with only using some of the baselines instead of all of them. The degenerate parameters of redundant calibration outlined above are not actually specific to any individual antenna in the array: they are only properties of the array itself. We could, for example, use only a fraction of the longest baselines in the array for partial absolute calibration, which gives us a lever-arm advantage for estimating the delay and phase gradient terms. Concern about this approach are 1) increased noise in the gains due to less data points in our $\mathbf{y}$ vector and 2) if the baselines selected are drawn from a unique population of antennas relative to all other antennas in the array, in which case the average amplitude and phase gradients estimated with the longest baselines (which will preferably come from antennas near the edge of the array) will be mis-estimates for the other antennas not represented in the system of equations. One could devise strategies for mitigating these kinds of concerns by, say, ensuring that while only a fraction of the baselines are used in calibration, every antenna is at least somewhat represented in the system of equations.

Chapter 4

Instrumental Coupling Systematics: Modeling and Mitigation for HERA

In this section we tackle the problem of residual systematics leftover after our initial first round of calibration discussed in chapter 3. Specifically, we investigate a class of systematics that we refer to as "instrumental coupling" systematics, which form the majority of the observed systematics in HERA Phase I data. In particular, we break this down into two kinds of systematics: reflections within the signal chain of an antenna (i.e. antenna auto-coupling) and coupling *between* signal chains of different antennas (i.e. antenna cross-coupling). In contrast to the last section, we start by providing a purely theoretical description of these systematics–both analytic and numerical–and describe their phenomenology in the interferometric visibilities of drift-scan experiments. We use this information to construct data models for modeling the subtracting these terms in the data, and present a series of injection and recovery trials on purely simulated data. Having described the systematics and validated our mitigation techniques on simulations, we then demonstrate an application of these methods on HERA Phase I data. In doing so, we provide a thorough detailing of the observed systematics in the HERA Phase I system, and attempt to discern their physical origin within the HERA system. This section draws primarily from a two-part series of papers: Kern et al. (2019) detailing the theory of the systematics and the techniques used for their mitigation, and Kern et al. (2020b) detailing their application to HERA data.

4.1 Mathematical Overview

In this section, we describe how signal chain reflections and antenna cross couplings appear in interferometric data products. To begin, we start with a the two-element interferometer (Hamaker et al. 1996; Smirnov 2011), consisting of two antennas, 1 and 2, whose feeds measure an incident electric field and convert it into a voltage. In Figure 4.1, we show a schematic of the HERA analogue system and mark possible sources of internal instrument coupling. These signals travel from the feeds through each antenna's signal chain to the correlator, and along the way are amplified, digitized, channelized, and Fourier transformed into the frequency domain. The correlator then

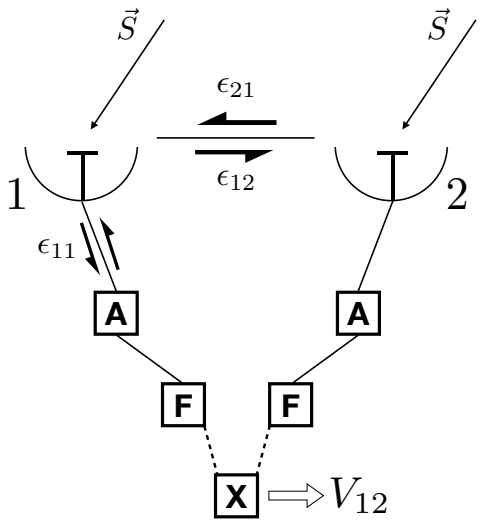

Figure 4.1: A schematic of two HERA signal chains, 1 & 2, with possible sources of systematics demarcated. Sky signal ($\vec{S}$) enters each antenna's feed, is converted into a voltage and travels down their signal chains where it is first processed at a node housing an amplifier (**A**). It is then directed to an engine that digitizes and Fourier transforms the signal (**F**) and sent to the correlator (**X**), which produces the visibility V_{12}. A possible cable reflection in antenna 1's signal chain is marked as ϵ_{11}, traversing up and down the cable connecting the feed to the node, and possible cross-coupilng is marked as ϵ_{12}, where radiation is reflected off of antenna 2 and into antenna 1, or vice versa. Dashed lines indicate a signal pathway after digitization, where internal instrument coupling is no longer a major concern.

cross multiplies voltage spectra to form the fundamental interferometric data product: the cross-correlation visibility, V_{12}, between antenna 1 and 2, written as

$$V_{12}(\nu, t) = v_1(\nu, t)v_2^*(\nu, t). \tag{4.1}$$

The correlator can also produce the auto-correlation visibility by correlating an antenna voltage with itself (e.g. V_{11}). Here we have chosen to define the visibility as the product of two antenna voltage spectra, rather than the correlation of voltage time streams: although the two are equivalent given the convolution theorem, the former will prove to be an easier basis when working with reflections. In addition, we have been explicit about the frequency and time dependence of each antenna's voltage spectrum v and, by extension, the complex visibility V, although we drop these throughout the text for brevity. We have also dropped the time averaging done by any real correlator, which is done for brevity and does not alter our results in this section. While we could have cast the visibility equation (Equation 4.1) in matrix form (e.g. Hamaker et al. 1996; Smirnov 2011), we find it easier to understand the impact that the specific sytematics we discuss in this paper have on the resultant data products using a simpler, algebraic form for the visibility equation.

The correlator outputs time-ordered visibilities as a function of local sidereal time (LST; denoted as t) and frequency (denoted as ν). When Fourier transforming the data across the frequency axis,

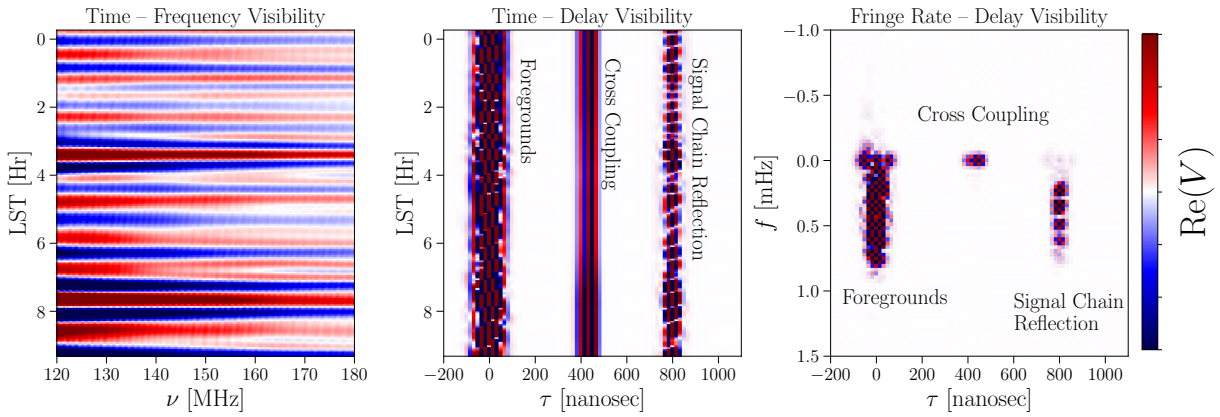

Figure 4.2: The real component of a simulated cross-correlation visibility with foregrounds, a signal chain reflection inserted at $\tau = 800$ ns and a cross coupling term inserted at $\tau = 400$ ns, plotted in dimensionless units for visual clarity. **Left:** Visibility in time and frequency space. **Center:** Visibility in time and delay space. **Right:** Visibility in fringe-rate and delay space. Different components of the visibility–in particular systematics–are usually better separated in delay and fringe-rate space than in the original time and frequency space.

we put the data into a temporal domain. To separate this from the original time domain, we refer to the Fourier dual of frequency as the *delay domain* (denoted as τ). Similarly, the Fourier transform of our data across time puts the data into a spectral domain, which we refer to as the *fringe-rate domain* (denoted as f), using similar notation as Parsons et al. (2016). In the absence of explicit markers, we will use V to mean the visibility in time and frequency space, and use $\widetilde{V}$ to mean the visibility in one or both of the Fourier domains: which should be clear based on context, otherwise we will use explicit notation.

Different components of the visibility are generally more localized in Fourier space. Foreground signal, for example, is intrinsically spectrally smooth and will therefore occupy low delay modes, whereas a fiducial EoR model, being non-spectrally smooth, occupies low and high delay modes. Figure 4.2 shows a simulated foreground + systematic visibility in real and Fourier space, demonstrating how systematics are usually better separated in Fourier spaces (center & right panel). We present the figure here to guide the reader's intuition about the phenomenology of the systematics in real and Fourier space while we discuss their mathematical form below. Note that the parameters of the systematics as simulated in Figure 4.2, for example the delays they show up at, have been chosen merely for visual clarity, and are not necessarily the systematic parameters seen in actual data. We describe the simulations used throughout this work in section 4.2.

4.1.1 Describing Signal Chain Reflections

A reflection in the signal chain of an antenna inserts a copy of the original signal with a time lag. An example of this is a reflection at the end of an analogue cable, where the signal travels back up and cable, reflects again at the start of the cable and travels back down and is transmitted through

the system along with the original signal. The time lag, or delay, the reflected signal has acquired is two times the cable length divided by the speed of light in the cable. The reflected signal also acquires an amplitude suppression meaning it is generally only a fraction of the input signal, but even a small fraction of the foreground signal in the data can dwarf the expected EoR signal and therefore needs to be accounted for. If v_1 is antenna 1's voltage spectrum without a signal chain reflection, then the presence of a reflection can be encapsulated as

$$v_1'(\nu, t) = v_1(\nu, t) + \epsilon_{11}(\nu) v_1(\nu, t) \tag{4.2}$$

where v_1' is the voltage spectrum of antenna 1 with the reflection component, and ϵ is a coupling coefficient describing the reflection in antenna 1's signal chain (denoted as 11 because it is coupling the signal with itself). The coupling coefficient can be broken into three constituent parameters as

$$\epsilon_{11}(\nu) = A_{11} e^{2\pi i \tau_{11} \nu + i\phi_{11}}, \tag{4.3}$$

where A is the amplitude, τ is the delay offset (the total time it takes to be reflected) and ϕ is the phase offset the reflected signal may have acquired relative to the original signal. In Equation 4.2 we have assumed time and frequency stability of the reflection parameters, although in practice these parameters will have some variation with time and frequency.

If we insert the corrupted voltage spectra into the visibility equation (Equation 4.1), we get

$$V_{12}' = v_1 v_2^* + \epsilon_{11} v_1 v_2^* + v_1 \epsilon_{22}^* v_2^* + \epsilon_{11} v_1 \epsilon_{22}^* v_2^*. \tag{4.4}$$

We can see that in addition to the original cross-correlation term ($v_1 v_2^*$) we now also have copies of it at positive and negative delay offsets that are suppressed in amplitude by a factor of A_{11} and A_{22}, respectively. The time-behavior of a reflection mimics that of the original data, in that it shows the same temporal oscillation (i.e. fringing) as the foregrounds, and thus also appears at the same fringe-rate modes as the foregrounds (e.g. right panel of Figure 4.2). The conjugation of ϵ_{22} means that the reflected signal from antenna 2 appears at negative delays in V_{12}, while the reflected signal from antenna 1 appears at positive delays.

The resultant auto-correlation visibility can also be computed, and is given by

$$V_{11}' = v_1 v_1^* + \epsilon_{11} v_1 v_1^* + v_1 \epsilon_{11}^* v_1^* + |\epsilon_{11}|^2 v_1 v_1^*, \tag{4.5}$$

where we see that the first order reflections show up at $\pm\tau_{11}$, while the second-order reflection appears at $\tau = 0$ ns due to the conjugation of the coupling coefficient with itself. If we assume that the first order reflections are at sufficiently high delay and neglect the second-order term, we can approximate the visibility at $\tau = 0$ ns as $\widetilde{V}_{11}'(\tau = 0) \approx (v_1 v_1^*)(\tau = 0)$. Similarly, near the reflection delay of τ_{11} the auto-correlation visibility simplifies to

$$\widetilde{V}_{11}'(\tau = \tau_{11}) \approx \epsilon_{11} v_1 v_1^*. \tag{4.6}$$

This means that one can estimate the reflection coefficient amplitude in delay space as

$$A_{11} = \frac{\left|\widetilde{V}_{11}'(\tau = \pm\tau_{11})\right|}{\left|\widetilde{V}_{11}'(\tau = 0)\right|}, \tag{4.7}$$

which will be useful when modeling reflection systematics.

If one can estimate their parameters from the data, reflections can be removed via standard (direction-independent) antenna based calibration. In this paradigm, the raw voltage spectrum of antenna 1 corrupted by the instrument is related to its true value as

$$v_1^{\text{raw}} = v_1 g_1, \tag{4.8}$$

which when inserted into the visibility equation yields the standard antenna based calibration equation,

$$V_{12}^{\text{raw}} = V_{12} g_1 g_2^* = \langle v_1 v_2^* \rangle g_1 g_2^*. \tag{4.9}$$

The g term is called the antenna gain, and accounts for amplitude and phase errors introduced by the various stages of the signal chain from the feed all the way to the correlator. Note that this form of the calibration equation does not account for polarization leakage induced by cross-feed coupling, which is generally a higher order correction (Hamaker et al. 1996; Sault et al. 1996). By re-arranging Equation 4.2 as

$$v_1' = v_1(1 + \epsilon_{11}) = v_1 g_1, \tag{4.10}$$

we can see that signal chain reflections can be completely encompassed in this gain term, and hence corrected for by dividing the corrupted data by a gain constructed from an estimate of the reflection coefficient.

4.1.2 Describing Antenna Cross Coupling

We now turn our attention to another systematic we refer to as antenna cross coupling, which acts to couple one antenna's voltage stream with another antenna's voltage stream before reaching the correlation stage. Note that our model for cross coupling is different than "capacitive crosstalk" created by the electric field of two parallel signal chains interacting with each other within cabling, receivers, and analogue-to-digital conversion (ADC) units, which is a common systematic for radio interferometers (Parsons & Backer 2009; Zheng et al. 2014; Ali et al. 2015; Patil et al. 2017; Cheng et al. 2018). There are well established hardware solutions for suppressing crosstalk, such as phase switching (Chaudhari et al. 2017). However, if residual crosstalk remains, or if phase switching is not implemented in the system, we need to model and remove it for robust EoR measurements.

Our cross cross systematic model simply states that, before correlation, one antenna's voltage is added to another antenna's voltage with a coupling coefficient that determines the amplitude and relative delay with which the voltage is added. For the purposes of our description, we additionally assume that this coupling coefficient can be decomposed into the same three parameters as before, technically making it a form of reflection systematic.[1] While this model may indeed be capable of describing certain some forms of capacitive crosstalk, we do not expect all forms of capacitive crosstalk to necessarily fall within the bounds of these assumptions.

[1] While we adopt this assumption in this section to make the algebra simpler, the algorithm we present in section 4.3 is more general and does not rely on this assumption.

To write down how this affects the interferometric visibility, we can start by writing the corrupted antenna voltages as

$$\begin{aligned} v_1' &= v_1 + \epsilon_{21} v_2 \\ v_2' &= v_2 + \epsilon_{12} v_1, \end{aligned} \tag{4.11}$$

where ϵ_{21} describes the voltage coupling of antenna 2 into antenna 1 and vice versa for ϵ_{12}. Substituting these equations into Equation 4.1, we get

$$V_{12}' = v_1 v_2^* + v_1 \epsilon_{12}^* v_1^* + \epsilon_{21} v_2 v_2^* + \epsilon_{21} v_2 \epsilon_{12}^* v_1^*. \tag{4.12}$$

We can see that the cross-correlation visibility now contains the auto-correlation visibility terms $v_1 v_1^*$ and $v_2 v_2^*$ at the first-order level, which are purely real quantities and thus have identically zero phase. In the complex plane, the cross-correlation term $v_1 v_2^*$ winds around the origin as a function of time because its phase varies temporally. Because the auto-correlation has no phase, the act of the cross coupling terms is to introduce an additive bias to the data with an arbitrary phase set by the coupling coefficient itself. Assuming the coupling coefficient is slowly variable (if not completely stable), the first order systematic terms in Equation 4.12 only change in amplitude over time set by the natural variation in the amplitude of the auto-correlation, (e.g. $v_1 v_1^*$). This variation is generally fairly slow on timescales of a beam crossing, which for HERA is roughly 1 hour. This leads us to two critical insights about the behavior of the cross coupling terms: 1) their time variability is slow, thus occupying low-fringe rate modes (e.g. see right of Figure 4.2) and 2) they have a time-stable phase determined solely by the phase of the coupling coefficient.

In a more generalized case, the cross coupling between antenna 1 and 2 may have an angular dependence on the sky. Take for example the case of mutual coupling (or feed-to-feed reflections), where part of the radiation incident on antenna 1's feed is reflected and received by antenna 2's feed. This behavior will be highly angular dependent due to the non-trivial electromagnetic properties of the feed itself. Nonetheless, we can reason that the systematic phenomenology will be similar to as before. We can think of this angular dependence as a windowing function on the primary beam of the underlying auto-correlation, meaning that the first-order terms in Equation 4.12 will be proportional to only a fraction of $v_1 v_1^*$, such that they have a smaller amplitude. It may also mean that these terms will have a slightly faster time dependence in the data, as the "effective beam" created by the angular windowing function is smaller on the sky than the total primary beam, and thus leads to a faster "effective beam crossing time."

We can also compute the effects of cross coupling on the measured auto-correlation visibility, V_{11}, which yields

$$V_{11}' = v_1 v_1^* + v_1 \epsilon_{21}^* v_2^* + \epsilon_{21} v_2 v_1^* + |\epsilon_{21}|^2 v_2 v_2^*. \tag{4.13}$$

In this case, we find that the cross-correlation is inserted into the measured auto-correlation at the first order level with a delay offset of τ_{21}. These terms are likely many order of magnitudes below the peak auto-correlation visibility amplitude, given that the cross-correlation visibilities are generally a few orders of magnitude below the auto-correlation inherently, which is further compounded by the amplitude suppression from ϵ_{21}.

We can see simply from Equation 4.12 that the corruption of V'_{12} by cross coupling *cannot* be factorized into antenna based gains, based simply on the presence of the ϵ_{12}-like terms, which are baseline-dependent. Removal of cross coupling terms in the data must therefore be done on a per-baseline basis by constructing a model of the systematic in each visibility and then subtracting it.

4.1.3 Summary

To summarize, reflections along a single antenna's signal chain produces a duplicate of the signal with suppressed amplitude and some delay offset. This is true for both the cross and auto-correlation visibility products. Example mechanisms include cable reflections and dish-to-feed reflections within the confines of a single antenna. Reflections in the cross-correlation visibility have the same time structure as the un-reflected visibility, meaning reflected foreground signal occupies the same fringe-rate modes as un-reflected foreground signal, but is shifted to high delays (e.g. Figure 4.2). Reflections can be removed from the raw data by creating a model of the reflections and incorporating them into the per-antenna calibration gains.

Another systematic we describe is created by antenna-to-antenna cross coupling, which mixes the voltage signals between the antennas. This has the effect of introducing a copy of the auto-correlation visibility into the measured cross-correlation visibility at positive and negative delay offsets, and similarly introduces copies of the cross-correlation visibility into the measured auto-correlation visibility. In the measured cross-correlation visibility, the first-order coupling terms are slowly time variable and occupy low fringe-rate modes centered at $f = 0$ Hz. Cross coupling terms cannot be removed via antenna based calibration, and must be modeled and subtracted at the per-baseline level.

4.2 Simulated Visibilities with `healvis` and `hera_sim`

We use the numerical visibility simulation package `healvis`[2] to compute mock observations of foreground and EoR sky models. `healvis` numerically integrates the measurement equation (Equation 4.1) by representing the sky and direction-dependent antenna primary beam response as HEALpix maps (Górski et al. 2005), and summing their product with the baseline fringe pattern to compute the visibility. The mechanics of `healvis` as a simulator is described in more detail in Lanman & Pober (2019). Our simulations use HEALpix sky maps with NSIDE = 128 and use a frequency and angular-dependent electromagnetic simulation of the HERA dish and feed primary beam response (Fagnoni et al. 2019). The adopted beam model does not include mutual coupling effects. The beam has been smoothed across frequency to limit excess spectral structure above 250 ns, mimicking an idealized HERA beam response. We simulate visibilities for HERA baselines of various orientations and separations ranging from 0 meters in separation (i.e. the auto-correlation) out to 60 meters in separation. The sky resolution provided by a HEALpix NSIDE = 128 map is

[2] https://github.com/RadioAstronomySoftwareGroup/healvis

roughly ten times smaller than the fringe wavelength of the longest baseline in consideration at the highest simulated frequency of $\nu = 180$ MHz.

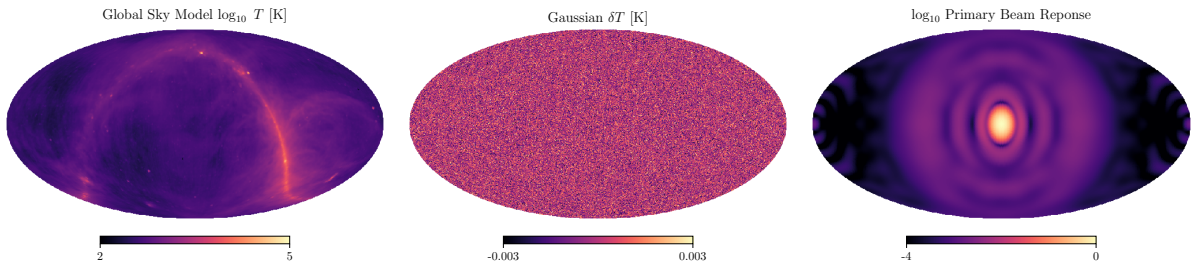

Figure 4.3: HEALpix sky maps at $\nu = 120$ MHz used for simulating diffuse foregrounds (left) and an uncorrelated EoR field (center). The antenna primary beam response (right) is taken from an electromagnetic simulation of the HERA dish and feed (Fagnoni et al. 2019).

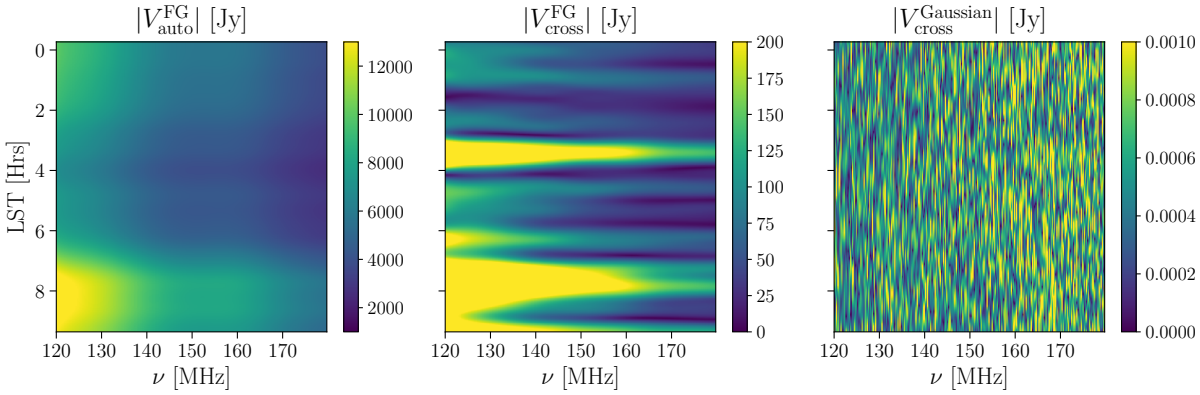

Figure 4.4: Mock visibilities computed with `healvis` and `hera_sim` showing (left) an auto-correlation visibility of diffuse foregrounds, (center) a cross-correlation visibility of diffuse foregrounds, and (right) a cross-correlation visibility of an EoR model.

Foreground visibilities are simulated with HEALpix maps of the diffuse, low-frequency radio sky from the `PyGSM` package[3], which is a repackaging of the original 2008 Global Sky Model (GSM; de Oliveira-Costa et al. 2008). We simulate a bandwidth spanning 120 – 180 MHz with 256 channels and an LST range of roughly 0 - 8 hours with a time cadence of 30 seconds (about 1000 time bins), which corresponds to the transit of the cold part of the radio sky. Figure 4.3 shows the diffuse radio sky from the GSM, a mock EoR realization, and the adopted antenna primary beam response at $\nu = 120$ MHz. The EoR model is constructed as an uncorrelated δT field with a variance of 25 mK^2, consistent with fiducial expectations for the signal at EoR redshifts (Mesinger et al. 2011). Its Fourier correlations are modeled as a flat spectrum in $P(k)$: while fiducial EoR

[3] https://github.com/telegraphic/PyGSM

models tend to favor EoR as a roughly flat spectrum field in $\frac{k^3}{2\pi^2}P(k)$, for the purpose of using these simulations as mock visibilities for validation and testing we require only a plausible EoR model. The pixel size of an NSIDE 128 map corresponds to a transverse comoving length scale of $\sim$ 70 cMpc at $z = 8$, which is larger than the size scales where EoR is correlated during the beginning and middle of reionization. The frequency axis is simulated with a channel resolution of 234 kHz, which at redshift $z = 8$ corresponds to a comoving length scale of $\sim$ 4 cMpc, which is roughly the scale at which our uncorrelated Gaussian field approximation begins to break down. However, we also believe this to have a negligible impact on our performance and signal loss tests: the most sensitive part of our analysis is the computation of the EoR PSD functions (Figure 4.8) which relies on the time correlations of the EoR model, not its frequency correlations. Figure 4.4 shows simulated `healvis` visibilities of the diffuse foreground (left) and EoR model (center) described above.

We also use the visibility simulation toolbox `hera_sim`[4] to model signal chain reflection and cross coupling systematics. `hera_sim` is a general purpose toolbox for creating mock observations with realistic instrumental and environmental effects, like thermal noise, reflections and RFI. For inserting systematics into the data, `hera_sim` uses the equations outlined in section 4.1, in particular Equation 4.5 and Equation 4.12. For the signal loss trials described in section 4.4, we simulate 100 independent EoR visibilities coming from EoR sky maps generated with unique random seeds, and add the GSM foreground visibility to each one. We then add in the relevant systematic to be tested, and use this set of EoR + foreground + systematic visibilities to perform the ensemble average needed for quantifying signal loss.

4.2.1 EoR Power Spectral Density Functions for HERA

We use `healvis`-simulated HERA observations to calculate the expected power spectral density (PSD) functions of an EoR-like signal for HERA visibilities. At any given time, a point source locked to the celestial sphere generates a complex sinusoid in the inteferometric visibility with a time period determined by its position on the sky and the fringe profile of the baseline at hand. Over a short time interval, its Fourier transform across time is nearly a delta function at a fringe-rate set by the inverse of its time period. Thus, over short time intervals, points on the sky map to specific fringe-rates in the visibility (Parsons et al. 2016), which is also related to the m-mode analysis (Shaw et al. 2014). The analytically-derived "optimal fringe-rate filter" in Parsons et al. (2016) is a filter that minimizes the noise-component of the visibility-based power spectrum errorbars, and is related to the PSD of EoR-like signals in the visibility. However, the analytic derivation hinges crucially on the assumption that the beam crossing time is much longer than the fringe-crossing time for a point source on the sky, which was a fairly valid assumption for a wide-field experiment like PAPER. This is not the case for HERA, which has both shorter baselines with wider fringes and also a more compact primary beam. Therefore, calculating the power spectral density of an EoR-like sky signal is more easily done numerically. We do this by generating over 100 independent EoR sky models with the same variance but different initial seeds. We perform HERA visibility simulations of each

[4] https://github.com/HERA-Team/hera_sim

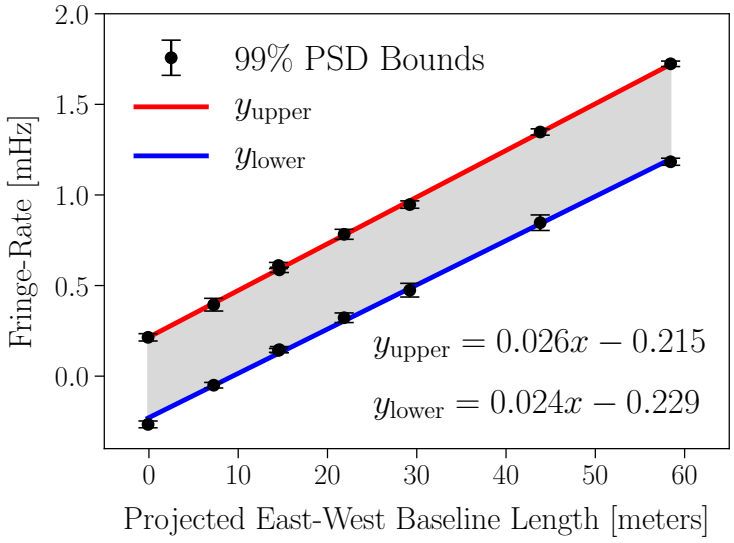

Figure 4.5: Power spectral density (PSD) bounds of EoR sky models for HERA baselines in fringe-rate space at $\nu = 120$ MHz. The PSD curves are shown in Figure 4.8, and the bounds quoted correspond to 99% of the total power. Best-fit lines are shown for extrapolation to other HERA baselines.

sky using `healvis`, Fourier transform them from LST into the fringe-rate domain and then take their absolute value and average across each realization. The square of these profiles is shown in Figure 4.8, which represents a numerically-derived PSD of a theoretical EoR-like signal in HERA visibilities. These profiles allow us to tailor visibility Fourier filters to do things like minimize EoR signal loss in systematic subtraction, or to minimize the thermal noise component in the power spectrum errorbars. We tabulate the 99% power bounds of these curves in Table 4.1 for the few baselines presented in this study. We also plot these bounds and provide a fitting formula as a function of projected East-West baseline length in Figure 4.5, such that one can extrapolate these data points to other baselines in the HERA array. Errorbars on the 99% bounds in Figure 4.5 are calculated via bootstrap resampling over the independent realizations (Efron & Tibshirani 1994).

4.3 Systematic Modeling

Next we discuss our approach for modeling reflection and cross coupling systematics in the data. We use sky and instrument simulations to generate mock visibilities of diffuse foregrounds and systematics, which we use to test our algorithms and provide benchmarks on their performance. More details on the construction of the simulated data products used in this work can be found in section 4.2. The fiducial parameters of our systematic simulations are chosen to roughly reflect the behavior of systematics seen in HERA Phase I data (Kern et al. 2020b). Systematics in the HERA Phase I system can be found at variable amplitudes and delays in the data depending on the baseline or antenna at hand, but are generally seen at an amplitude of $\sim 3 \times 10^{-3}$ times the peak

foreground amplitude at $\tau = 0$ ns and at delays spanning 200 – 1500 ns. For these simulations we do not include instrumental thermal noise so that we can test the underlying performance of our algorithms to high dynamic range.

4.3.1 Modeling Signal Chain Reflections

Modeling signal chain reflections can in theory be done simultaneously with standard gain calibration because reflections factor as an antenna-based effect (see section 4.1). However, there are reasons why we might be wary of using standard calibration techniques for deriving reflection parameters. Principally, standard bandpass calibration typically operates on the $\sim N_{\rm ant}^2$ number of cross-correlation visibilities, and generally allows each frequency channel's gain to be solved independently from other channels. Frequency dependent calibration errors will therefore set a fundamental floor to the precision with which reflections can be calibrated via standard antenna-based bandpass techniques (Barry et al. 2016; Ewall-Wice et al. 2017; Orosz et al. 2019). Furthermore, the dynamic range of signal to noise is considerably higher in the auto-correlation than in the cross-correlation visibility, as it is a measurement of the total power received by an antenna: in many cases a signal chain reflection cannot even be seen above the noise floor in a cross-correlation visibility but is highly apparent in the auto-correlation visibility. This latter point is important, because it implies that reflection parameters estimated from the auto-correlation visibilities will have a signal-to-noise ratio (SNR) that is drastically higher than the SNR of the cross-correlation visibilities, meaning that, when calibrating out systematics in the cross-correlation visibilities, the SNR of the derived reflection parameters will never be a limiting factor. Of course other real-world factors can limit the precision of the derived reflection parameters such as non-trivial frequency evolution, which is discussed in more detail in the context of HERA in Kern et al. (2020b).

Reflection parameters must be estimated to high precision in order to get even a modest suppression of their systematic power in the visibilities. For example, Ewall-Wice et al. (2016b) employed a reflection fitting algorithm on MWA auto-correlation visibilities by fitting sinusoids in the frequency domain and was able to suppress reflection systematics by a couple orders of magnitude in the power spectrum, although their end result band powers were still systematic limited at some k modes. Similarly, Beardsley et al. (2016) explore reflection calibration on MWA data as an extension to their restricted polynomial gain calibration scheme and also achieve a couple of orders of magnitude of suppression in the 2D power spectrum, although their more deeply integrated power spectra show its re-emergence.

The algorithm we present here also operates on the auto-correlation visibility, but we choose to model the reflection in the delay domain rather than the frequency domain. Recall from section 4.1 that in the auto-correlation visibility, V_{11}, a signal chain reflection appears as a shifted copy of the original visibility at symmetric positive and negative delay offsets. The Fourier transformed auto-correlation visibility, $|\widetilde{V}_{11}|$, is intrinsically quite peaky in delay space, meaning that a reflection is essentially a narrow spike appearing at its corresponding reflection delay. The reflection amplitude is then estimated via Equation 4.7, and the reflection phase is estimated by transforming back to frequency space and aligning reflection templates while varying their phase until a squared error metric is minimized. This yields an initial estimate of the reflection parameters, but it needs to be

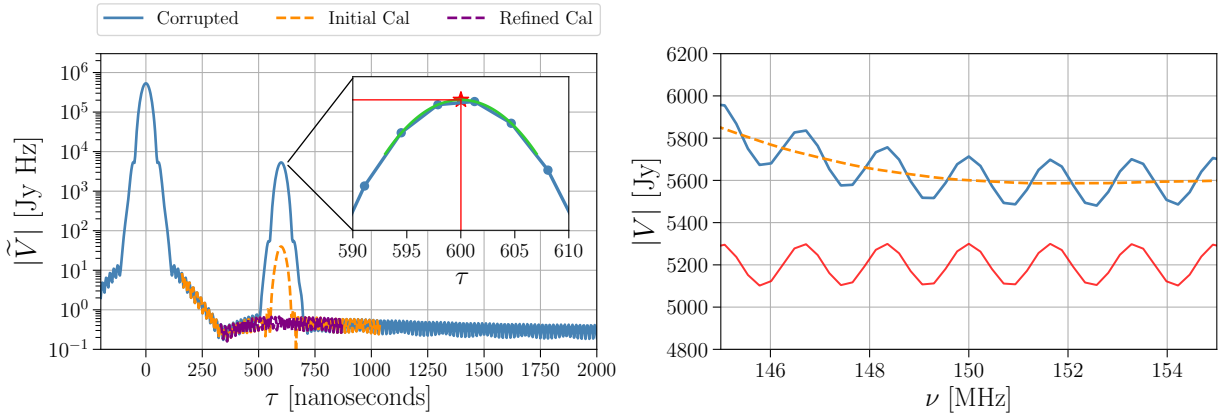

Figure 4.6: Reflection modeling and removal on a simulated auto-correlation visibility. **Left**: Foreground-only auto-correlation in delay space with a simulated cable reflection at 600 ns (blue). Dashed orange shows the visibility after initial reflection calibration, demonstrating roughly two orders of magnitude of suppression. The inset highlights the reflection bump, showing a spectral fit via quadratic interpolation (green curve) to achieve more precise estimates of the reflection delay and amplitude (red star). The calibration is then refined using an iterative technique until the reflection bump is minimized (purple). **Right**: Simulated visibility with reflection in frequency space (blue), a scaled version of the fitted reflection coefficient (red) highlighting its phase coherence with the reflection ripple in the data, and the visibility after initial calibration (dashed orange).

refined in order to suppress the systematic to high dynamic range. In order to refine our parameter estimates, we setup a non-linear optimization system that perturbs the initial guesses, applies the calibration to the data in frequency space, transforms to Fourier space, estimates the amplitude of the residual reflection bump and repeats until it is minimized or a stopping threshold is reached. Solving this with an iterative minimization technique allows us to estimate the reflection parameters with sufficient accuracy to suppress the reflection in our systematic simulations by eight orders of magnitude in the power spectrum.

To summarize, our approach for estimating the reflection parameters from the data takes the following steps:

1. Zero-pad the auto-correlation in frequency space and apply a windowing function before Fourier transforming to delay space to minimize sidelobe power
2. Fit for the peak of the reflection bump in $|\widetilde{V}|$ via quadratic interpolation of its nearest neighbors
3. The estimated reflection delay, τ, is equal to τ_{peak}
4. The estimated reflection amplitude, A, is the ratio $|\widetilde{V}(\tau = \tau_{\text{peak}})|/|\widetilde{V}(\tau = 0)|$ (Equation 4.7)
5. Set all modes of $\widetilde{V}$ to zero except the modes nearest τ_{peak} and Fourier transform back to frequency space to get V_{filt}

6. The estimated reflection phase, ϕ, is found by minimizing $|V_{\rm filt} - Ae^{2\pi i\nu\tau + i\phi}|^2$ while varying ϕ from 0 - 2π.

7. Setup a non-linear optimization that perturbs the initial reflection parameter estimates, applies the calibration and transforms to Fourier space until the residual near the original reflection bump is minimized or a stopping threshold is reached.

We demonstrate this algorithm on a simulated HERA auto-correlation visibility corrupted by a cable reflection with a (frequency independent) amplitude of 10^{-2} at a delay of 600 nanoseconds (Figure 4.6). The natural delay resolution of HERA data is 10 nanoseconds, which is much too coarse to achieve precision estimates of the reflection delay. Zero-padding the data by a factor of three gets us to a delay resolution of 3 ns, but this is still not precise enough for accurate reflection delay estimates. By employing quadratic interpolation on the spectral peak, we can recover the input cable delay to roughly ±0.1 ns (left of Figure 4.6). In this idealized, noise-free simulation, the initial reflection calibration estimates the reflection delay to within ±0.1 ns of its true value, and its phase to within ±0.01 radians, yet we only see systematic power suppression of two orders of magnitude in the visibilities. This is representative of the precision needed to achieve even modest systematic suppression. With the refined reflection calibration, however, we find we can achieve reflection systematic suppression of up to four orders of magnitude in the visibility, and recover the reflection parameters to within 1 part in 10^6 of their true value.

Fiducial EoR levels are expected to be roughly 10^{-5} times the peak cross-correlation foreground power in the visibility at $k_\parallel \sim 0.1\ h\ \mathrm{Mpc}^{-1}$, and is generally thought to be even weaker at higher $k_\parallel$ (Mesinger et al. 2011). If reflection systematics have inherent amplitudes of around 10^{-3}, then a few orders of magnitude of further suppression will push them below expected EoR levels at low k. In practice, non-ideal effects like frequency evolution of the reflection parameters will limit the precision of reflection calibration, which has been observed in real instruments (Ewall-Wice et al. 2016b). Indeed, inclusion of such effects in our systematic simulation will likely degrade our algorithm performance. However, if frequency evolution is a limiting factor for reflection calibration, one simple strategy is to split the full bandwidth into multiple sub-bands and perform reflection calibration independently on each of them, with the caveat that non-negligible frequency evolution within each sub-band may need to be mitigated in other ways. One can also do this more self-consistently by estimating the reflection parameters and their frequency dependence across the full band, however, we defer this to future work.

4.3.2 Modeling Cross Coupling

Antenna cross coupling systematics are a baseline-dependent effect and as such must be modeled and subtracted for each cross-correlation visibility independently. In other works, cross coupling has been modeled as a phase-stable term in the data that can be removed by applying a finite impulse response (FIR) high-pass filter to the data (Parsons et al. 2010; Ali et al. 2015; Kolopanis et al. 2019). The algorithm described in this work is conceptually quite similar in that we take advantage of cross coupling's slow time variability to model it, but is different in its methodology. Note that the method presented here does not assume that the instrumental bandpass has been calibrated out:

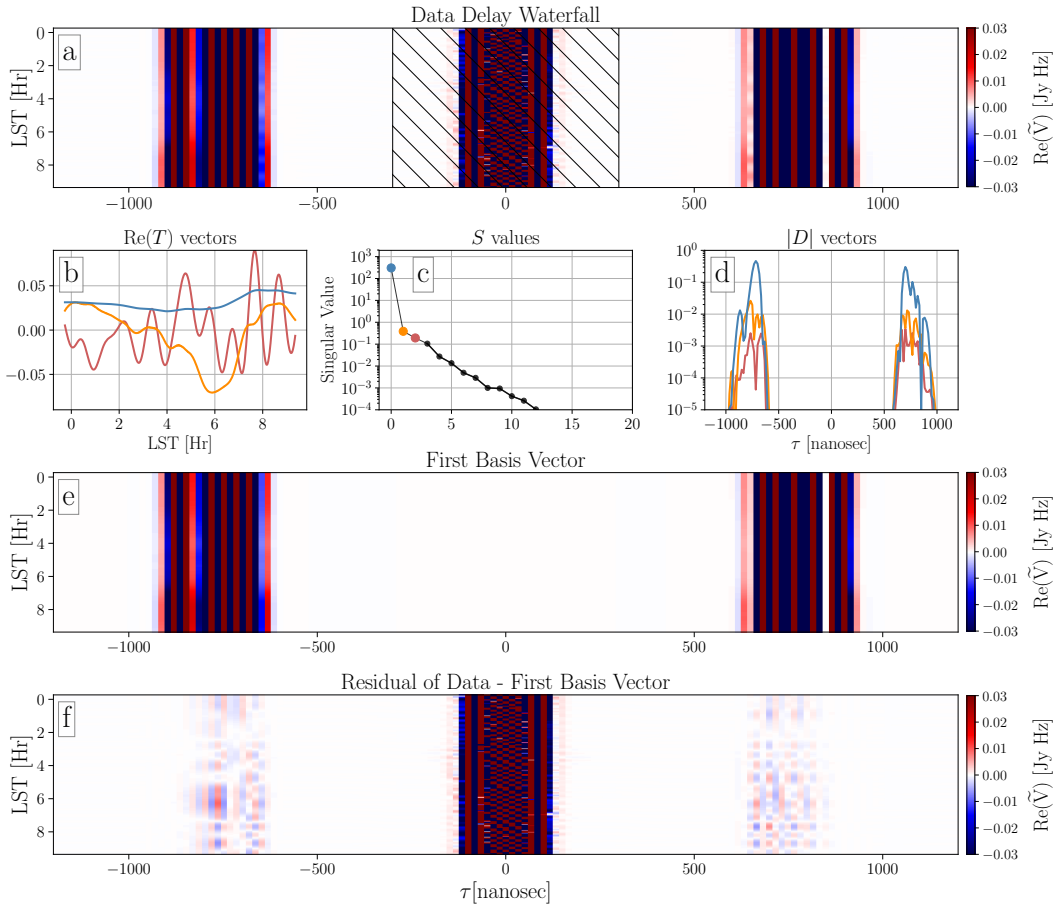

Figure 4.7: Semi-empirical modeling and removal of cross coupling systematics from a simulated cross-correlation visibility. **(a)** A simulated visibility with foregrounds (center) and cross coupling systematics (left & right). The hatched region of $|\tau| < 300$ ns is assigned zero weight before taking the SVD. **(b, c, d)** The resulting **T** modes, singular values and **D** modes after factorization via SVD. The **D** modes are artificially offset for visual clarity. **(e)** The outer product of the first **T** and **D** mode multiplied with its singular value yields the first basis vector having the shape of the original data matrix. **(f)** The difference of the systematic model and the original data shows decent subtraction of the systematic, but isn't enough to completely remove it from the data.

these two steps are in principle interchangeable. In practice, it helps if at least the antenna cable delays are calibrated out such that the main foreground lobe shows up at the expected delays of $\tau \approx 0$ ns, but again this is not strictly necessary.

Our semi-empirical approach starts with the cross-correlation visibility Fourier transformed across frequency, $\widetilde{V}_{12}$, such that it is in the time and delay domain. If we think of the visibility as a 2D rectangular matrix, we can use Singular Value Decomposition (SVD) to decompose the matrix

as

$$\widetilde{\mathbf{V}} = \mathbf{TSD}^{\dagger}, \tag{4.14}$$

where **T** is a unitary matrix containing basis vectors (also referred to as eigenmodes) across time, **D** is a unitary matrix containing basis vectors across delay, and **S** is a diagonal matrix containing the weight (or their singular values) of each mode in the data. There may be components of our data matrix, $\widetilde{V}$, that are inherently low-rank, like a slowly time-variable systematic for example. Thermal noise in the visibility, on the other hand, occupies the full-rank of the data matrix. SVD can help us model and pull out the low-rank components of the matrix, thus providing an approach for systematic removal on a per-baseline basis.

Our SVD-based systematic removal algorithm operating on an individual visibility takes the following steps:

1. Fourier transform the visibility waterfall to delay space.
2. Apply a rectangular band-stop window across delay to down-weight foregrounds at low delays.
3. Decompose the visibility via SVD.
4. Choose the first N modes to describe the systematic and truncate the rest.
5. Take the outer product of the remaining **T** and **D** modes to form N data-shaped templates.
6. Multiply each template with their corresponding singular value in **S** and sum them to generate the full time and delay-dependent systematic model.
7. Fourier transform the systematic model from delay space back to frequency space and subtract it from the data.

A demonstration of this process on simulated visibilities is shown in Figure 4.7. In this example, the simulation contains foregrounds and cross coupling systematics, but is free of both thermal noise and EoR components. We start with a simulated visibility spanning roughly 9 hours in LST with 1000 time bins and spanning a bandwidth from 120 – 180 MHz with 256 frequency bins. By Fourier transforming it to the delay domain (a), we see that foregrounds are confined to low delays while cross coupling systematics span a wide range of delays at positive and negative delay offsets. Applying a band-stop windowing function across delay before taking the SVD (hatched region) assigns zero weight to delay modes dominated by foreground signal at $|\tau| < 300$ ns. The result of the SVD shows significant isolation of the information content of the systematic in the visibility into the first eigenmode (c). The first time eigenmode (b) indeed shows it to be slowly time variable, as we would expect for a cross coupling systematic (see section 4.1). The outer product of the first time and delay eigenmodes multiplied with their singular value yields a systematic template with the shape of our original data matrix (d). Taking this single template as our systematic model (equivalent to setting $N = 1$) and subtracting it from the data yields the systematic-subtracted data (f).

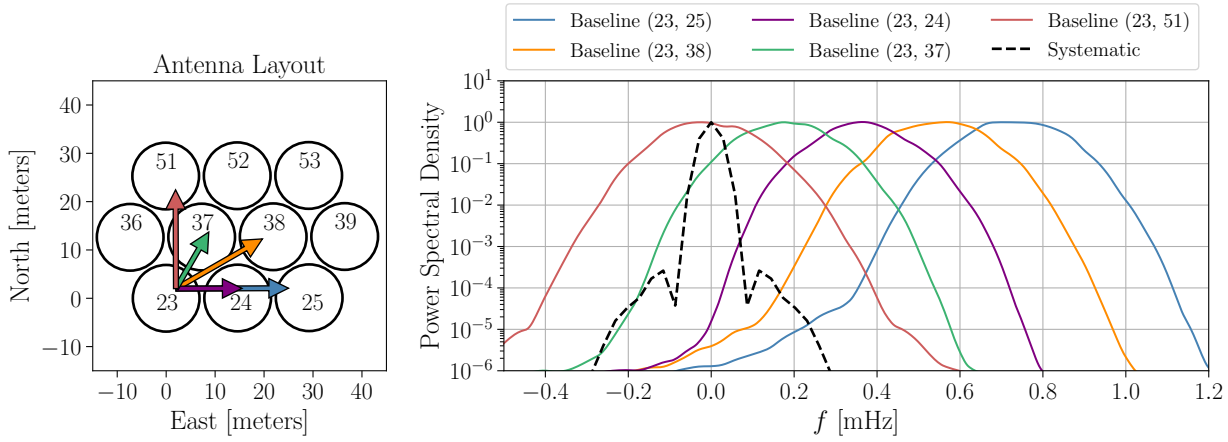

Figure 4.8: Peak-normalized PSDs in fringe-rate space of an uncorrelated Gaussian EoR sky model for HERA baselines at 120 MHz. **Left:** A subset of the HERA array showing its shortest baselines arranged in a hexagonal pattern. The arrows denote four unique baseline orientations. **Right:** Peak-normalized PSDs describing how EoR sky signal populates the inteferometric visibility in the fringe-rate domain, as well as a cross-coupling systematic (black dashed).

We see in (d) that a cross coupling model with $N = 1$ provides good subtraction of the systematic, but is not enough to completely remove it: based on Figure 4.7 (c) we can see that it provides effectively three orders of magnitude of suppression, which, depending on the inherent amplitude of the systematic, may or may not be enough to push it below fiducial EoR levels. We can remove more and more of the systematic by increasing the number of SVD modes we incorporate into the systematic model. However, as is the case with empirically-based models, this has the side-effect of possibly introducing structure from other components of the visibility that we do not want in our systematic model, such as the EoR itself. If EoR signal was somehow soaked up by our systematic model, then by subtracting the model from the data we are inducing EoR signal loss, which is highly undesirable.

[h!]

To limit EoR signal loss in the process of systematic subtraction, we can filter the systematic model to reject Fourier modes that we know hold EoR power. In general, if a signal occupies the visibility in the fringe-rate domain with variance given by $\sigma(f)^2$ and we enact a filter on it by multiplying by a weighting function $w(f)$, then the total power of the signal before filtering, P_{before}, can be related to the total power after filtering, P_{after}, as

$$\frac{P_{\text{before}}}{P_{\text{after}}} = \frac{\int df\ \sigma(f)^2}{\int df\ \sigma(f)^2 w(f)^2}. \tag{4.15}$$

Therefore, if we know statistically how the EoR will populate the visibilities in the fringe-rate domain–in a sense deriving their power spectral density functions (PSDs)–we can construct a Fourier filter that is tailored to reject Fourier modes in the systematic model that we know hold EoR power.

Table 4.1: EoR Visibility Power Bounds

Baseline Length	Baseline Angle	99% Power Bounds
29.2 meters	0°	$0.46 < f < 0.95$ mHz
25.3 meters	30°	$0.31 < f < 0.77$ mHz
14.6 meters	0°	$0.14 < f < 0.58$ mHz
14.6 meters	60°	$-0.05 < f < 0.40$ mHz
25.3 meters	90°	$-0.27 < f < 0.21$ mHz

Note. — Power bounds are defined at $\nu = 120$ MHz. Baseline angle is defined in East North Up (ENU) coordinates as ϕ° North of East (e.g. left of Figure 4.8).

This is closely related to the optimal fringe-rate formalism outlined in Parsons et al. (2016). In Figure 4.8 we show peak-normalized power spectral densities (PSD) of an EoR sky model at 120 MHz for various baselines in the array, which describe the relative amount of signal power occupied by different fringe rates. We derive these via ensemble simulations of the same EoR sky while varying the initial seed (section 4.2). The PSD of a simulated HERA cross coupling systematic is also plotted (black dashed). While centered at a fringe-rate of 0 mHz, the systematic has a tail that extends out to negative and positive fringe-rates, the latter being where most of the power from the EoR also lies. Conversely, while most of the EoR power lies at positive fringe-rates for many baselines, some of its power also extends to zero and negative fringe-rates. Baselines that are longer along the East-West direction have more EoR power pushed to higher fringe-rates and thus are more naturally separated from cross coupling systematics, while baselines that are purely North-South in orientation see an EoR signal that is centered at $f \sim 0$ mHz, and almost completely overlaps the cross coupling systematic.

Each curve in Figure 4.8 integrated out to 99% of their total area yields the domain in fringe-rate space where 99% of the EoR power is contained in the visibility. We tabulate these bounds in Table 4.1 for a few HERA baselines at $\nu = 120$ MHz. If we low-pass filter any of the visibilities in Figure 4.8 in fringe-rate space by applying a symmetric top-hat filter with a maximum extent $f_{\rm max}$ given by these lower bounds, then Equation 4.15 tells us we will retain 99% of the EoR power in the data after filtering, which for our purposes is an effectively lossless operation given other more dominant sources of error. This result means that our ability to safely remove cross coupling systematics is baseline-dependent: for baselines with large East-West lengths (e.g. blue), we can filter out the vast majority of the systematic without attenuating the EoR. For shorter baselines (e.g. green), we may only be able to remove part of the systematic, and for baselines oriented along the North-South direction (e.g. red), we may not be able to remove any cross coupling systematics, if they exist.

Armed with the ability to filter an arbitrary signal without attenuating its EoR component, we can return to the problem of choosing the appropriate number of eigenmodes to use in describing

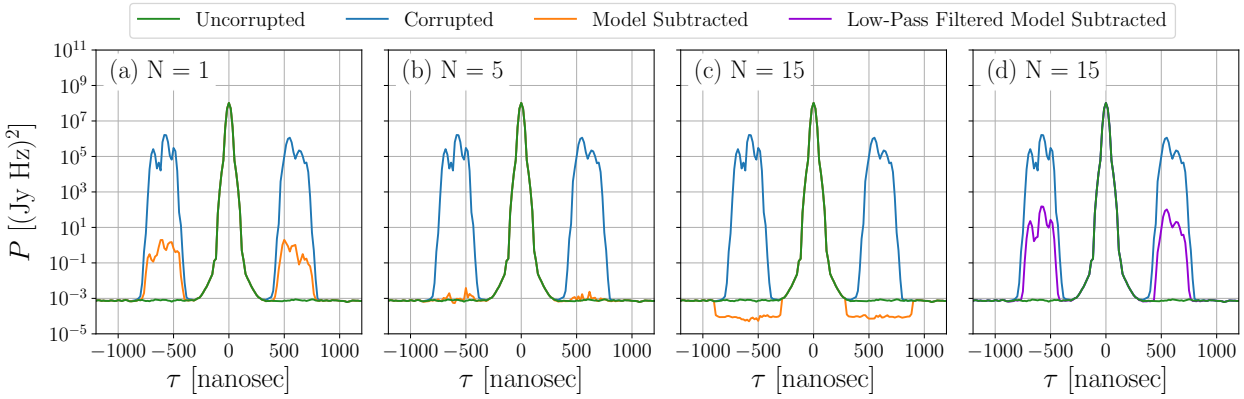

Figure 4.9: Cross coupling systematic removal on simulated EoR + foreground visibilities for a 15-meter East-West baseline with various choices of N. We show the visibility amplitude averaged over LST for the uncorrupted data (green), the data corrupted by a cross-coupling systematic (blue) and the systematic-model subtracted data (orange) for $N = 1, 5,$ & 15 (a, b & c respectively). In the last panel, we show the result of low-pass filtering the SVD **T** modes before forming the full systematic model and subtracting it from the data. For the baseline at hand, we do this with a fringe-rate cutoff of $f_{\rm max} = 0.14$ mHz (Table 4.1). This shows that by low-pass filtering the systematic model, we can constrain it such that it removes the systematic as much as possible while not attenuating the EoR, and is therefore optimal even if it leaves some of the systematic in the data.

a cross coupling systematic in the data. We noted in Figure 4.7 that by increasing the number of SVD eigenmodes used to describe our systematic model, we might be able to remove more of the systematic from the data. Figure 4.9 proves this on a simulated visibility, now simulated with an EoR and foreground component (green) corrupted by a cross coupling systematic at high delays (blue). We also show SVD-based systematic removal with increasingly more eigenmodes used to describe the systematic (orange). We can see that going from $N = 1$ (a) to $N = 5$ (b) enables us to subtract more of the systematic from the data. However, at some point we expect low-level eigenmodes to be influenced by other components of the data, such as EoR, foregrounds or noise, which raises the possibility of removing those components along with the systematic. This is shown in (c), where with $N = 15$ we have over subtracted the systematic and caused signal loss of EoR power at high delays. One might conclude from this that $N = 5$ is the "sweet spot" choice for the number of eigenmodes to use, but this choice is conditional on the relative amplitude between EoR and systematic: without knowing the amplitude of EoR in the data a priori, we have no way of knowing the appropriate number of eigenmodes to use that would enable us to subtract the systematic without attenuating EoR. This makes the algorithm as originally described in effect unusable, because it is an operation that is dangerously lossy to EoR signal.

The solution to this problem is to apply a low-pass time filter to the systematic model that is tailored to reject fringe-rate modes occupied by the EoR. Specifically, we can apply a filter to the systematic model **T** matrix that only keeps structure below some pre-defined maximum fringe-rate, $f_{\rm max}$, such that in the process of subtracting it from the data, all fringe-rates $|f| > |f_{\rm max}$ are left

unaffected. For example, if we could tolerate a maximum of 1% attenuation of EoR power in the process of systematic removal, then $f_{\rm max}$ would be set at the lower bounds tabulated in Table 4.1. The result of applying such a filter to the SVD eigenmodes is demonstrated in Figure 4.9 (d), which shows the systematic-subtracted data with $N = 15$ having first applied a low-pass filter to $\mathbf{T}$. Although a significant amount of systematic remains, we can now be confident that we have not attenuated the EoR signal in the data, even while using a large number of eigenmodes to describe the systematic.

In this section we have argued and shown via sky and instrument simulations that we can construct a cross coupling model that removes the vast majority of the systematic while remaining lossless to the EoR signal (for certain baseline orientations). In real data, however, the fidelity of this model will be fundamentally limited by the thermal noise floor of the observation, as is the case for any signal term modeled on a per-baseline basis. If the cross coupling systematic is truly baseline-dependent and is uncorrelated between baselines, then the residual systematic term will integrate down like thermal noise when we combine visibilities and we would not expect it to re-appear in the integrated power spectra. In HERA, for example, there is evidence that the observed cross coupling systematics are at least partially uncorrelated between baselines (Kern et al. 2020a).

4.4 Signal Loss

We have thus far presented an overview of instrumental systematics that can hinder if not prohibit the detection of the EoR for current and future 21 cm intensity mapping surveys, and have outlined algorithms for modeling and removing them from the data. However, any experiment that wishes to use a systematic removal technique on the data must show that the subtraction did not attenuate the desired signal in the data. In other words, one must quantify and account for possible sources of signal loss in a data reduction pipeline. In this work, we use signal loss to refer specifically to the inadvertent subtraction of sky signal (EoR or foreground) from the visibilities.

Quantifying signal loss can be done in a variety of ways depending on the nature of the algorithm one wants to test (Cheng et al. 2018; Mouri Sardarabadi & Koopmans 2019). In general, however, we can quantify the amount of signal loss induced by an algorithm by generating two identical mock datasets, introducing a systematic to one of them, attempting to remove it, and then comparing the end-result power between the two datasets. For this analysis, we generate mock observations using the same simulations used in section 4.3, but now in addition to diffuse foregrounds and systematics we introduce an EoR component. The EoR sky model is an uncorrelated Gaussian random field across both the spatial and frequency axes with a variance of 25 mK^2 (see section 4.2 for details).

We simulate the EoR and foreground visibilities separately, and then assign their sum as V_1. Next we create and add in a systematic visibility and assign their sum as V_2. Lastly, we create a visibility model of the systematic using our algorithms presented above, remove it from the data

and assign the residual as V_3, which can be summarized as follows:

$$\begin{aligned} V_1 &= V_{\rm eor} + V_{\rm fg} \\ V_2 &= V_1 + V_{\rm sys} \\ V_3 &= F(V_2) = V_2 - V_{\rm mod} \end{aligned} \tag{4.16}$$

where F is a systematic removal algorithm whose signal loss properties we would like to quantify, and whose effect is to subtract a model of the systematic, $V_{\rm mod}$, from the corrupted visibility. Note we do not include a thermal noise term, which is done so that we can probe the signal loss properties of the algorithms down to the extremely weak levels of a fiducial EoR signal.

Each visibility has an associated total power, which is a real-valued quantity and can be calculated as the square of the Fourier transformed visibility. In the ideal EoR + foreground case, this is

$$\begin{aligned} P_1 &= \widetilde{V}_1 \widetilde{V}_1^* \\ &= P_{\rm eor} + P_{\rm fg} + 2{\rm Re}(P_{\rm eor,fg}) \end{aligned} \tag{4.17}$$

where $\widetilde{V}$ signifies the visibility Fourier transformed from frequency to delay space, and $P_{\rm eor,fg}$ represents the cross-power between $\widetilde{V}_{\rm eor}$ and $\widetilde{V}_{\rm fg}$. Similarly, we can write the power of the systematic-subtracted visibility as,

$$\begin{aligned} P_3 &= \widetilde{V}_3 \widetilde{V}_3^* \\ &= P_{\rm eor} + P_{\rm fg} + 2{\rm Re}(P_{\rm eor,fg}) \\ &\quad + P_{\rm sys} + P_{\rm mod} - 2{\rm Re}(P_{\rm sys,mod}) \\ &\quad + 2{\rm Re}(P_{\rm eor,sys}) - 2{\rm Re}(P_{\rm eor,mod}) \\ &\quad + 2{\rm Re}(P_{\rm fg,sys}) - 2{\rm Re}(P_{\rm fg,mod}). \end{aligned} \tag{4.18}$$

In the case where we have *perfectly subtracted* the systematic from the visibility (i.e. $V_{\rm mod} = V_{\rm sys}$), we see that the systematic power terms cancel with the model power terms such that, not surprisingly, we get that $P_3 = P_1$. In the case where we have *imperfectly subtracted* the systematic–either by incorrect estimation of its phase and/or amplitude–the cross terms no longer cancel. What this means for the total power of the resultant visibility, P_3, depends on how well-matched the model visibility is to the systematic, in addition to the relative inherent amplitude of the sky signal versus the systematic. For example, in the case of an imperfect systematic model, then we can see that this will always be true: $P_{\rm sys} + P_{\rm mod} > 2{\rm Re}(P_{\rm sys,mod})$, meaning their difference results in excess power. Whether or not the EoR and foreground cross term residuals in Equation 4.18 result in overall positive or negative power depends on how well matched the systematic model is to either EoR or foregrounds.

Given this, we can construct a simple metric,

$$R_3(\tau) = \frac{\langle P_3(\tau) \rangle}{\langle P_1(\tau) \rangle} \tag{4.19}$$

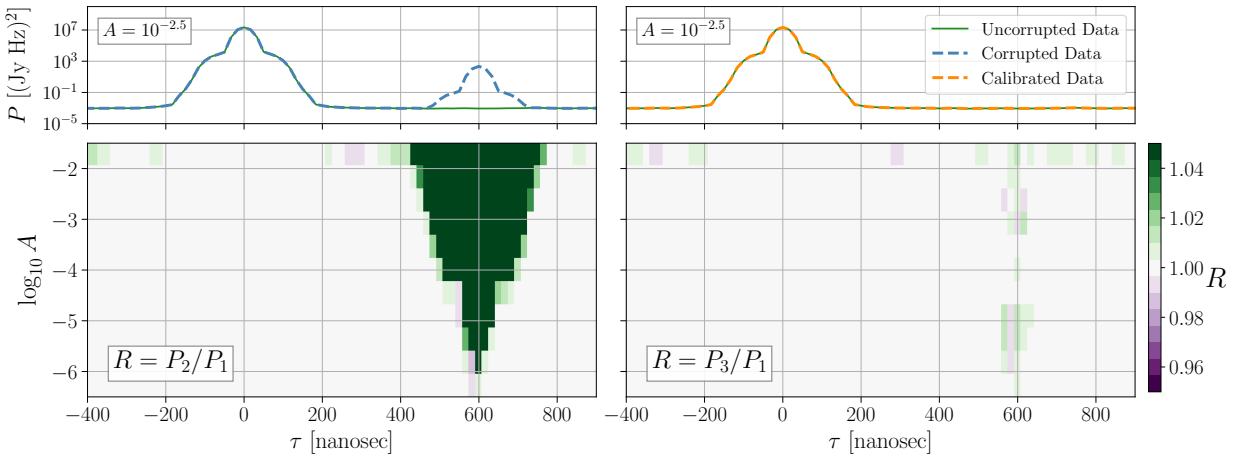

Figure 4.10: Signal loss trials of reflection calibration with noise-free, foreground + EoR simulated visibilities. **Top Row:** Power spectra of the corrupted visibility V_2 (blue-dashed), the uncorrupted visibility V_1 (green-solid) and the calibrated visibility V_3 (orange-dashed) from a signal loss trial with a reflection amplitude of $A = 10^{-2.5}$. **Bottom Row:** Heatmaps of the signal loss R metric computed for the corrupted data (left) and the calibrated data (right) as a function of reflection amplitude (y-axis) and delay (x-axis). The residual fluctuation about $R = 1$ in the right panel is encompassed within the $1/\sqrt{N}$ sample variance of our finite ensemble average.

to determine whether or not a step in our data analysis induces signal loss. Here the $\langle\rangle$ denotes an ensemble average over many realizations of the visibilities with the same kind of sky signal and systematic. Taking the ensemble average before taking the ratio is done to ensure the power spectra are properly normalized. Signal loss occurs anytime $R_3(\tau) < 1$, meaning that our model-subtracted visibility, V_3, has less power in it than our uncorrupted visibility, V_1. Specifically, *EoR signal loss* occurs anytime $R_3(\tau) < 1$ at delays we know to be dominated by EoR over foregrounds, such as all delays significantly outside the geometric delay of the baseline. In the case of $R_3(\tau) > 1$, the resultant visibility is systematic limited but, importantly, is not under reporting the power in the data relative to the pure sky signal visibility. We can also form the metric $R_2(\tau)$ using P_2 instead of P_3, which informs us of the relative amplitude of the raw systematic (without any removal) compared to the underlying sky signals.

Whether or not an algorithm is lossy in practice can depend on the relative amplitude between the signal and systematic present in the data. As such, we need to compute the R metric while varying the relative amplitude between the EoR and systematics. Cheng et al. (2018), for example, do this by repeatedly injecting mock EoR signals into their analysis pipeline with increasing amplitude. In this study we take the opposite approach. We adopt a fixed EoR amplitude consistent with rough theoretical expectations and insert systematics at amplitudes below, equal to and above the adopted EoR amplitude and compute R. This approach is more consistent with what we expect to find in the real data: at certain times, frequencies, or baselines, we may find systematics to be heavily dominant, while at other times, frequencies or baselines, there may be no systematics at all.

4.4.1 Signal Loss in Reflection Calibration

To test signal loss in the context of reflection calibration, we precompute the visibilities for a single diffuse foreground model and 100 independent EoR models, with each simulation spanning 8 hours of LST and a thousand individual time integrations. Because HERA has a beam crossing time of about 1 hour, this yields an effective number of independent foreground + EoR simulations of $\sim$ 800. The adopted EoR model is an uncorrelated Gaussian field across angular position and frequency with a variance of 25 mK^2 (section 4.2).

A single signal loss trial takes the following steps. First we choose a random EoR model from our library of pre-computed visibilities and add it to our foreground visibility (V_1). We then make a copy of it and insert a reflection with a delay of 600 nanoseconds, a random phase, and a single amplitude across frequency using Equation 4.4 (V_2). We then model the reflection in the simulated auto-correlation knowing only the approximate delay range at which it appears, and then apply the derived gain solution to the cross-correlation visibilities (V_3). Next we compute power spectra of each data product (P_1, P_2, & P_3). We then repeat this on the order of 100 times, each with a different random EoR model, and then take their average to approximate the ensemble average in Equation 4.19. We then form the R_2 and R_3 metrics as a function of time and delay–i.e. $R_3(t, \tau)$–and average over time to collapse them onto a single axis across delay. This entire procedure produces one signal loss trial, which is defined uniquely by the amplitude, A, of the reflection inserted into the visibility.

Figure 4.10 shows multiple trials for different reflection amplitudes in the range of 10^{-6} to 10^{-1}. The top row shows power spectra of each of the three visibility products as a function of delay for one trial when $A = 10^{-2.5}$. Recall that the reflection amplitude is defined with respect to the visibility, meaning that the observed reflection amplitude in the power spectrum is A^2. The bottom row shows a heatmap of the signal loss R metric as a function of delay (x-axis) and each trial's reflection amplitude (y-axis). The left panels shows R_2 and the right panel shows R_3, highlighting the amount of delay space that is brought down to $R \sim 1$ after reflection calibration, with negligible amounts of signal loss (purple shaded regions). Furthermore, the weak levels of fluctuating residual systematic and signal loss observed at the $\sim 2\%$ level are within the $1/\sqrt{N}$ sample variance of our finite ensemble average. For context, HERA cable reflection amplitudes are seen at around 10^{-3} (Kern et al. 2020a).

Multi-Reflection Regime

Above we probed for signal loss when performing reflection calibration on a single reflection that was isolated in delay space. Next, we relax this assumption and test how the the algorithm performance and signal loss properties change when we add in more reflections, which is relevant for any instrument with multiple cables, or with cables that have sub-reflections along the length of the cable (Ewall-Wice et al. 2016b; Kern et al. 2020b). We choose to model the relative amplitude of these reflections as an inverse power law as a function of delay with a nominal reflection amplitude of $A = 3 \times 10^3$. Our algorithm models and calibrates out each reflection one-at-a-time, starting with the reflection with the largest amplitude. We do not feed the algorithm the position of each reflection (it searches for it automatically within a specified range of delays), but we do assume we

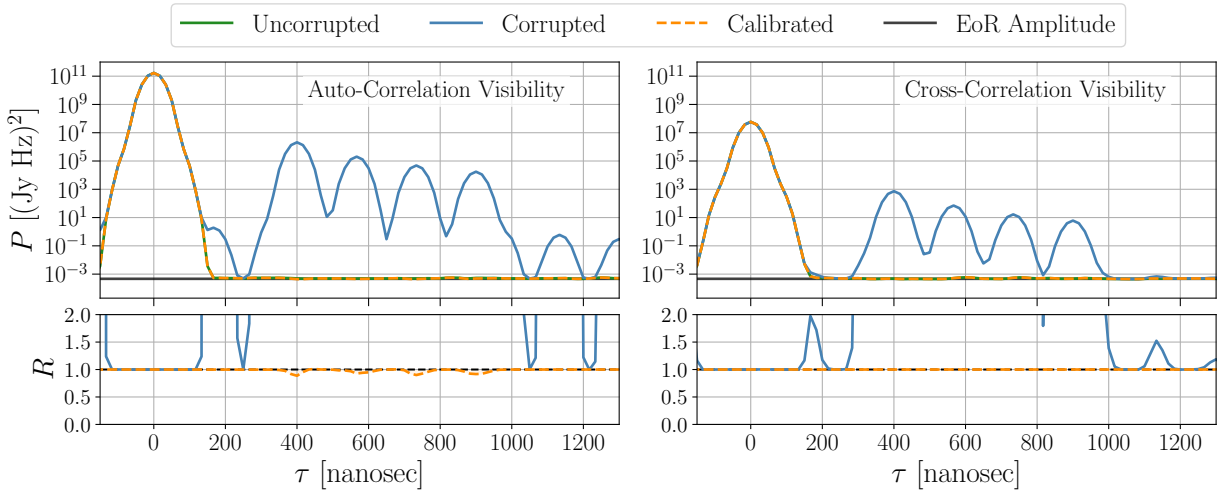

Figure 4.11: Reflection calibration run on an auto-correlation (left) in a low-confusion, multi-reflection regime, where we then apply the resultant gains to a cross-correlation visibility with a 29-meter baseline length (right) and repeat a few dozen times. In the top panels, the (ensemble average) power spectrum of the uncorrupted data (P_1; green), corrupted data (P_2; blue) and calibrated data (P_3; orange-dashed) are plotted along with a line denoting the underlying EoR amplitude in the data (grey). Signal loss metrics R_2 (blue) and R_3 (orange) are shown in the bottom panels. We see that while reflection calibration can lead to a slight amount of signal loss at the reflection delays of the auto-correlation visibility (bottom-left, dashed), signal loss is not observed to an appreciable degree in the cross-correlation visibility (bottom-right, dashed).

know the number of reflection inherent in the data, which controls how many times we iterate the algorithm.

Our first test shown in Figure 4.11 involves only five reflections inserted across a relatively wide region in delay, such that they can be considered non-overlapping (or un-confused). In the top panels, we show power spectra of the uncorrupted data (P_3; green), the corrupted data (P_2; blue) and the reflection calibrated data (P_3; dashed-orange), along with a line marking the EoR amplitude in the data (grey). We show this for the auto-correlation visibility (left) and a 29-meter cross-correlation visibility (right). In the bottom panels we show the signal loss metrics R_2 (blue) and R_3 (dashed-orange). Recall that reflection calibration builds up a set of gains strictly from the auto-correlation, and then applies those gains to the cross-correlations. We find that our algorithm performs exceptionally well in this regime: removing reflections down to the inherent sidelobe floor of the auto-correlation, which is more than enough to bring the systematics to the EoR level in the cross-correlation visibility. We find that while reflection calibration can lead to slight signal loss in the auto-correlation visibility, we find no appreciable levels of signal loss in the cross-correlation.

Our next test shown in Figure 4.12 increases the number of reflection modes inserted into the same region, such that they become almost entirely overlapping. In this case we can see that our algorithm fails to perfectly calibrate out the reflections in both the auto-correlation (left) and cross-correlation (right) due to the partial confusion. Nonetheless, we still find that while imperfect

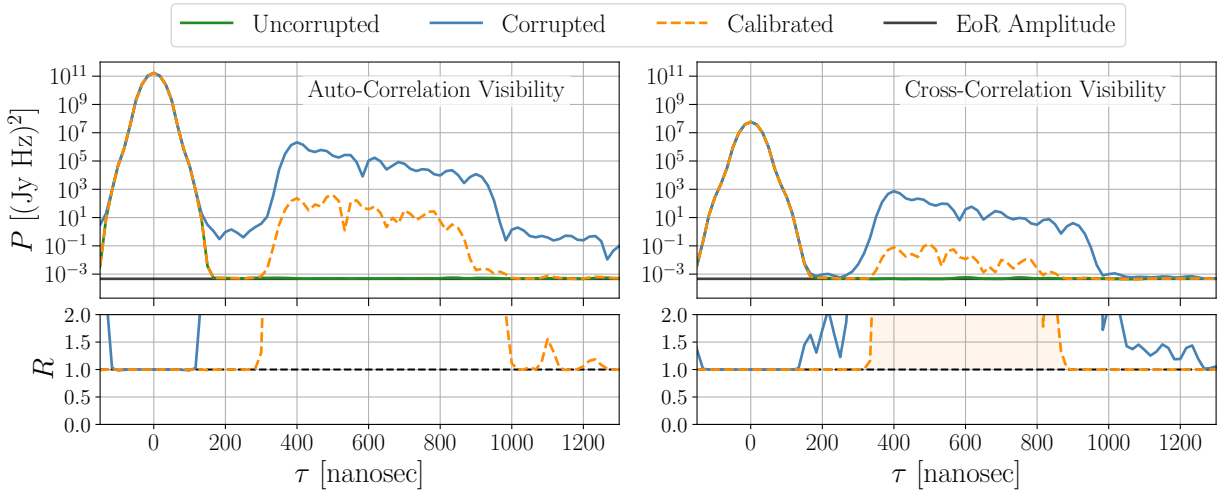

Figure 4.12: Same figure as Figure 4.11 but now in a higher-confusion, multi-reflection regime. Importantly, even when reflection calibration encounters confusion in its peak finding algorithm and fails to perfectly model the reflection, it still does not induce appreciable signal loss in the cross-correlation visibility (bottom-right, orange).

reflection calibration can lead to slight signal loss in the auto-correlation, the cross-correlation is resistant to signal loss. Reflection calibration's resistance to signal loss in the cross-correlations is perhaps not surprising, given that reflection calibration operates in an antenna-based space while EoR and other sky signals live in a baseline-based space. Furthermore, our algorithm solely uses the auto-correlations to estimate the reflection parameters.

In total, our findings suggest that 1) our reflection calibration algorithm performs moderately well even in the many-reflection regime, and that 2) even if a broad delay region is contaminated by reflections, we can still in principle use this region for EoR measurements (or upper limits in the case of imperfect systematic removal) after reflection calibration because it does not suffer appreciable levels of signal loss.

4.4.2 Signal Loss in Cross Coupling Subtraction

In this section, we quantify signal loss for cross coupling subtraction in a similar manner. Based on our conclusions from subsection 4.3.2, we expect our cross coupling subtraction to be effectively lossless to EoR by construction, but testing this against ensemble signal loss trials is a good double-check and validation of our arguments. The rough functional form of the simulated cross coupling inserted into the visibilities is informed by cross coupling systematics observed in the HERA Phase I system. The simulations used are fundamentally the same as those in subsection 4.4.1, except simulated with cross coupling systematics rather than cable reflections. To simulate cross coupling in a cross-correlation visibility, we use Equation 4.12 to insert ~ 25 modes spanning $700 < |\tau| < 900$ nanoseconds with a decaying power law as a function of $|\tau|$

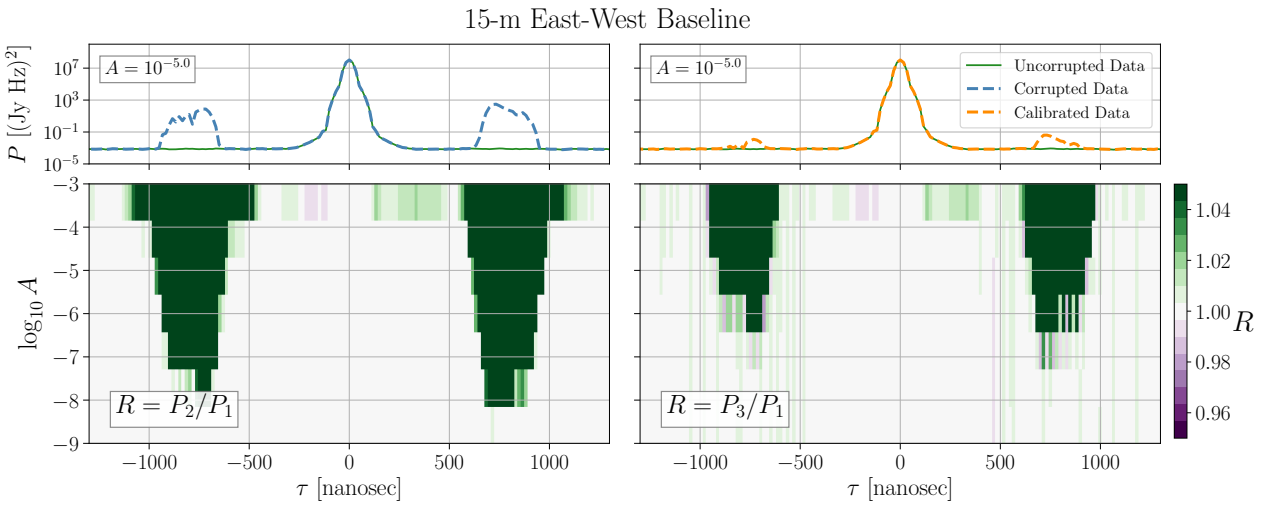

Figure 4.13: Signal loss trials of cross coupling removal with noise-free, foreground + EoR simulated visibilities for a 15 meter East-West HERA baseline. The systematic model is formed using 20 SVD modes and applies a low-pass time filter with an $f_{\max} = 0.14$ mHz. **Top Row:** Power spectra of the corrupted visibility P_2 (blue-dashed), the uncorrupted visibility P_1 (green-solid) and the systematic model-subtracted visibility P_3 (orange–dashed) from a signal loss trial with a coupling amplitude $A = 10^{-4}$. **Bottom Row:** Signal loss R metric computed for the corrupted visibility (left) and the model-subtracted visibility (right) as a function of coupling amplitude (y-axis) and delay (x-axis), with the model-subtracted visibility (right). No appreciable amounts of signal loss is observed, and the model-subtracted data show roughly four orders of systematic suppression in the power spectrum.

for their relative amplitudes, normalized such that the maximum amplitude relative to the peak *auto-correlation* foreground power equals a predefined amplitude, A. In the process of systematic removal, we model the systematic with $N = 20$ SVD eigenmodes and apply a low-pass fringe-rate filter on the SVD **T** modes using a Gaussian process smoothing with a maximum fringe-rate given by the lower bound in Table 4.1. We then Fourier transform the data from frequency to delay space using a 7-term Blackman-Harris window, and average across ensemble trials. After forming the R_2 and R_3 metrics as a function of time and delay, we truncate 5% of the time bins on either edge of the time axis before taking their time average to limit the influence of boundary effects in the smoothing process.

The result is shown in Figure 4.13 and Figure 4.14, which shows signal loss trials run on a 15-meter East-West baseline and a 29-meter East-West baseline, respectively. As expected, we see better systematic suppression for the longer baseline, where the EoR signal is inherently more isolated from the systematic in fringe-rate space. We also see that in the case of strong coupling amplitudes we cannot completely suppress the systematic down to EoR levels; however, we can nonetheless suppress the systematic by roughly four orders of magnitude in power for the 15-meter baseline and eight orders of magnitude in power for the 29-meter baseline. Importantly, we show that the algorithm presented does not significantly attenuate EoR in the data, with residual

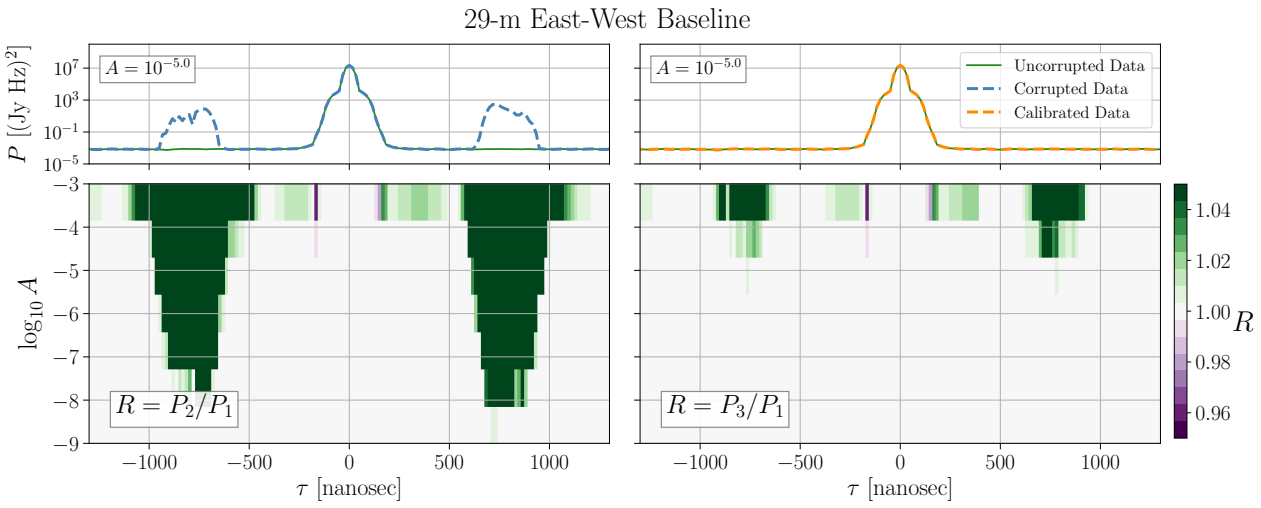

Figure 4.14: The same signal loss trials for removal of a cross-coupling systematic as described in Figure 4.13, but for a 29-meter East-West baseline using a low-pass time filter with $f_{\text{max}} = 0.46$ mHz. In this case we get upwards of six orders of magnitude in systematic suppression in the power spectrum.

fluctuations about $R = 1$ at the 1% level in power. While this result is not surprising given the fact that in section 4.3 we constructed our systematic model to explicitly be lossless to EoR, it is still a useful cross check on our algorithm and its implementation on the data.

4.5 Observed HERA Systematics

Next we turn to diagnosing the observed systematics in HERA Phase I data. The observations presented in this work come from a single night spanning 8 hours of local sidereal time (LST) on Julian Date 2458101. The array layout and correlator settings are the same as those described in section 3.1. At the time, the signal chains of the array front-end were split into two categories: Type 1 that used newly manufactured FEMs, PAMs, and coaxial cables specifically for HERA Phase I, and Type 2 that re-purposed the PAPER FEMs, PAMs and coaxial cables. In this analysis, we only use North-South ("YY") linear dipole polarization data, although all four auto and cross-feed polarization data products are recorded by the correlator.

The data have been pre-processed with part of the HERA reduction and calibration pipeline. Specifically, the data are first flagged for radio frequency interference (RFI) using a median filter and watershed algorithm operating on the cross correlation visibilities (Kerrigan et al. 2019). In this work, we also enact two additional steps for RFI flagging. The first takes stacked auto-correlation visibilities and differences them across time and frequency, normalizes them by their median absolute deviation and flags the residual at the 4 sigma level. Our second step runs a delay-based, iterative deconvolution on a subset of the auto-correlation visibilities, which attempts to deconvolve the discontinuous windowing function created by flagged data. This is similar in concept to the image-based CLEAN deconvolution (Högbom 1974), except applied to the frequency and delay

domains rather than the uv and lm domains, and with the missing data coming from RFI rather than incomplete uv sampling. We then normalize the filtered residual in frequency space by its median absolute deviation, and again enact RFI cuts at the 4 sigma level. Flags from each of the three independent steps are combined with a logical OR and then broadcasted across time and/or frequency if a 15% flagged threshold is met for any individual time bin or frequency channel. In total roughly 30% of the data volume is flagged, although this likely contains a decent amount of over-flagging.

Next we calibrate the data using a highly simplified antenna-based calibration. The full HERA calibration pipeline computes complex antenna gains for each time integration over the entire night from a combination of redundant calibration (Dillon 2017) and a constrained absolute calibration with the resultant gains smoothed across time and frequency (section 3.3). In this work, we take the gains derived from these steps and 1) average them across the entire night into a single spectrum, 2) average their amplitude across frequency to a single number, and 3) fit for a phase-slope across frequency (i.e. a single antenna delay). We are left with a single amplitude and delay for each antenna, which we apply to all times of the night. This has the effect of properly setting the flux scale of the data and also calibrates out the antenna cable delay, but ensures the gain itself we apply to the data has little to no spectral structure.

Because of our highly simplified calibration, the instrumental bandpass is not corrected for and still exists in the data. Calibration, being multiplicative in frequency space, can be thought of as a convolution in delay space. The true response of the visibilities in delay space is therefore initially convolved by the bandpass kernel upon measurement by the telescope. Assuming the bandpass is composed primarily of large-scale modes, its impact will be a slight smoothing-out of the true sky delay response and features created by systematics. Bandpass calibration performed beforehand may therefore sharpen systematics in delay space and actually make it easier to model and remove them.

4.5.1 Signal Chain Reflections

In this section we inspect the data for evidence of signal chain reflections. To do this, we take the auto-correlation visibility from each antenna and look for peaks in delay space. The calibrated data are filled with flags due to RFI and are thus nulled to zero at the flagged channels. This is not ideal for inspecting the data in delay space, as the Fourier transform of such a discontinuous windowing function creates strong sidelobes. To mitigate this we employ the same delay-based, iterative deconvolution algorithm from before to subtract these sidelobes, effectively interpolating across the nulled gaps in the data due to RFI (Parsons & Backer 2009). We allow the deconvolution to place model components out to delays of $|\tau| < 1600$ ns, and iterate until the process reaches 5× the noise floor of the data. We then make a copy of the data, and with the first copy we average the absolute value of the deconvolved visibilities in delay space across a few hours of LST. With the second copy we average the full complex-valued, deconvolved visibilities across the same time range, which will have a lower noise floor due to the complex average. Finally, we apply a Blackman window to the CLEANed visibilities before taking its Fourier transform.

Figure 4.15 shows these data products for the Type 1 (left) and Type 2 (right) signal chain, with

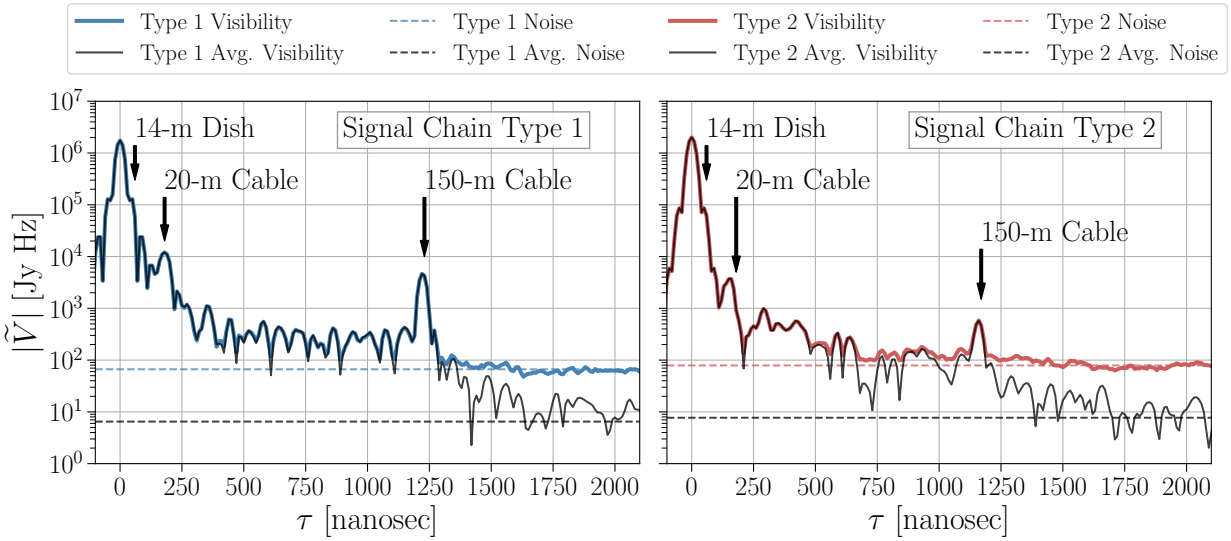

Figure 4.15: Auto-correlation visibilities for signal chain Type 1 (left) and Type 2 (right) with absolute time averaging (blue & red) and with complex time averaging (black), and their associated noise floors (dashed). Antennas 84 and 121 were used for the two auto-correlation visibilities. Delays for relevant length scales in the analogue system are marked with arrows. Resonances in the dish and reflections in the cables tend to be worse for signal chain Type 1. Additionally, we see evidence for a systematic tail in both signal chain types spanning a wide range of delays that does not integrate down like noise.

the absolute time-averaged data shown in blue or red, and the complex time-averaged data shown in solid black. Additionally, the thermal noise floors of each data product is plotted as dashed lines, which is estimated from the data via adjacent time and frequency differencing, and then divided by $1/\sqrt{N_{\rm avg}}$ where $N_{\rm avg}$ is the number of complex averages performed on the data. We find that Type 2 signal chains achieve a better overall impedance match with the analogue system, leading to slightly less structure in the auto-correlations across a wide range of delays. Nonetheless, we do see evidence for reflections from both the 20-meter and 150-meter cables, with reflection amplitudes in the range of roughly 3×10^{-3} and $1{\times}10^{-3}$, respectively. Of major concern is the tail of the auto-correlation response, which starts at low delays and slopes down to the noise floor out to the 150-meter cable delay. This tail is over an order of magnitude larger than that predicted by simulations of the HERA dish and feed (Ewall-Wice et al. 2016).

In this case the noise floor has been integrated down (solid black), we see that delays outside the 150-meter cable delay seem to effectively integrate down with the noise, while delays inside the 150-m cable-delay do not. This means that the features at low and intermediate delays are coherent on long timescales of at least a few hours. The abrupt change at $\sim$ 1250 nanoseconds is also possibly suggestive that tailed response might in part be originating within the 150-m cable. A possible mechanism for this could be sub-reflections within the cable due to intrinsic cable imperfections or environmental wear and damage along the cable. Another explanation is the effect of mutual coupling between neighboring antennas, which we explore in more detail in cross-correlation

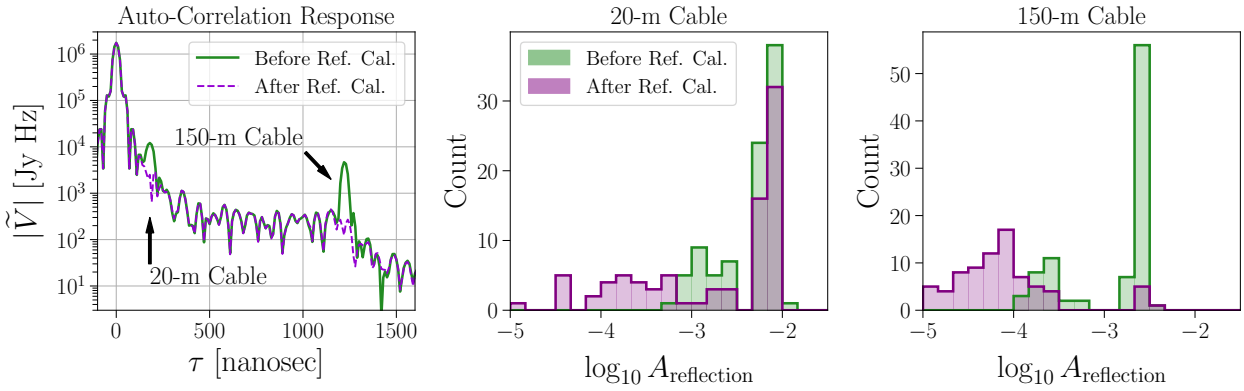

Figure 4.16: Reflection calibration performed over the full band (120 – 180 MHz) and applied to the auto-correlation visibilities. **Left**: The auto-correlation response before calibration (green) and after calibration (purple) demonstrates suppression of reflection systematics by roughly an order of magnitude in the visibility. **Center**: Histogram of derived 20-m reflection amplitudes before and after calibration. In the majority of cases we only see suppression by a factor of a few. **Right**: Histogram of derived 150-m reflection amplitudes before and after calibration. In the majority of cases we see suppression by at least an order of magnitude. Less suppression for the 20-m cable is likely attributable to more significant frequency evolution in the reflection parameters.

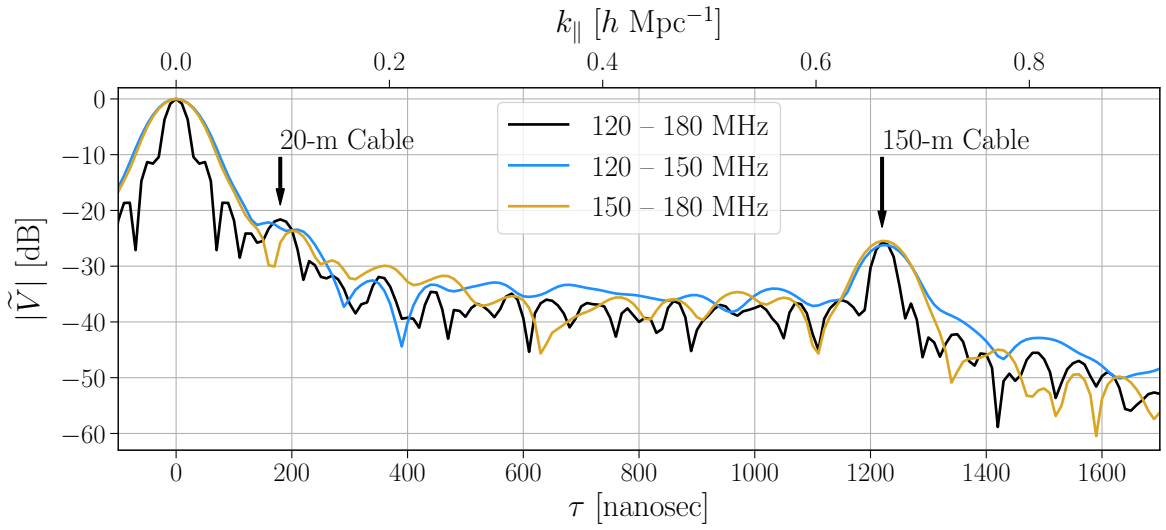

Figure 4.17: Auto-correlation visibility after complex time-averaging, transformed over the full band (120–180 MHz; black), just the low side of the band (120–150 MHz; blue) and just the high side of the band (150–180 MHz; gold) for a Type 1 signal chain. The 150-m cable reflection parameters are fairly consistent between both sides of the band, while the 20-m cable reflection shows significantly more frequency evolution. The smaller peaks along the systematic tail also shows significant frequency evolution.

visibilities in the following section. It is not easy to distinguish between these two effects in the auto-correlation visibilities alone. Direct electromagnetic simulations of mutual coupling in the HERA system provide mixed evidence: predicting it to appear at a similar amplitude and slope in the auto-correlations, but also predicting it to truncate at lower delays of ~600 ns (Fagnoni et al. 2019).

The fact that the auto-correlations show a systematic tail that, for $\tau > 300$ ns or $k_{\parallel} > 0.2$ h Mpc^{-1}, shows only three to four orders of magnitude of dynamic range is concerning, given that fiducial EoR amplitudes are generally assumed to lie at or below five orders of magnitude in dynamic range in the visibility for similar k (Thyagarajan et al. 2016). Furthermore, the observed systematic tail extends over a wide range of delays that covers essentially all of the $k_{\parallel}$ modes of interest ($0.2 < k < 0.6$ h Mpc^{-1}). These systematics need to be well-understood and mitigated if the data are to be used for stringent EoR limits.

Next we attempt to model some of these features and calibrate them out. One needs to proceed carefully when doing this because calibrating out structure that is inherent to the true data will actually *create* systematics. To be conservative, we only target the two features that we know to correlate with the expected delays of the 20-m and 150-m coaxial cables at ~ 200 and ~ 1250 ns. We use the method described in subsection 4.3.1 to derive reflection parameters across the full bandwidth excluding the band edges (120 – 180 MHz) and then apply them to the data in frequency space. Figure 4.16 shows the result, demonstrating the delay response of an auto-correlation before (green) and after (purple) reflection calibration, and also showing the derived reflection amplitudes of the 20-m and 150-m cable reflections before and after calibration. We find that in general we can suppress the 150-m cable reflection by a couple orders of magnitude (in the visibility), whereas for the 20-m cable reflection we get on average only a factor of a few suppression.

Often a limiting factor in reflection modeling is frequency evolution of the reflection parameters (Ewall-Wice et al. 2016b). In Figure 4.17 we plot the auto-correlation response having taken the Fourier transform of the data over a low-band (120–150 MHz; blue) and a high-band (150–180 MHz; gold), plotted in decibels relative to their peak value. Between the split sub-bands, we observe only slight evolution for the 150-m cable bump (labeled) but see more evolution in the 20-m cable bump (labeled). Furthermore, compared to the 150-m bump feature, the low-level fluctuations from 200 – 1200 ns are also substantially different between the low and high band spectral windows. Recall that the noise floor of the data in this case is around -50 dB, meaning the observed fluctuations at -35 dB are over an order of magnitude above the noise floor. This is likely at least part of the reason why we achieve less suppression for the 20-m cable reflection, and suggests that to mitigate reflections to higher dynamic range we will need to perform reflection calibration at the sub-band level. Because we find the suppression achieved by modeling these reflection across the full band is sufficient for this analysis (section 4.6), we defer sub-band reflection modeling to future studies.

4.5.2 Antenna Cross Coupling

Next we turn our attention to HERA's cross-correlation visibilities in order to probe for antenna cross coupling systematics. Specifically, we look at the North-South instrumental polarization (also

denoted as 'YY') for baselines (11, 12), (11, 13) & (11, 14), which are three East-West baselines with lengths of 15, 29 and 44 meters, respectively. These baselines display some of the strongest cross coupling systematics seen in the data, but are otherwise fairly nominal baselines.

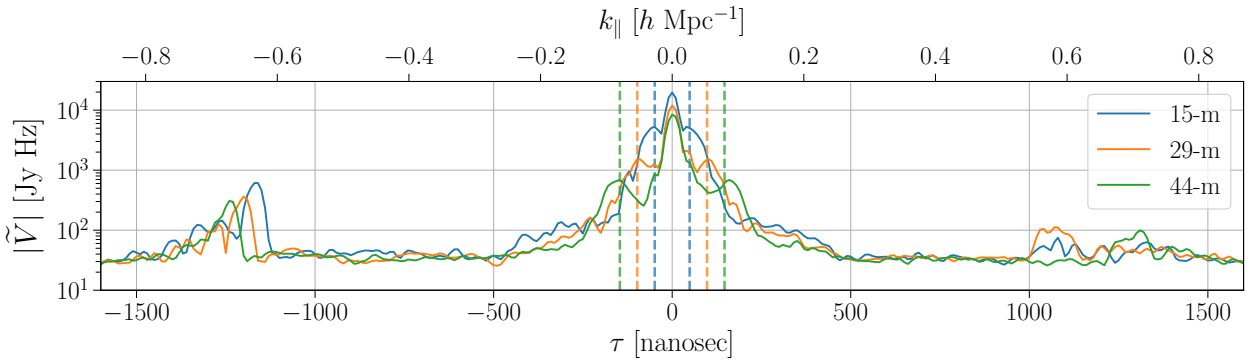

Figure 4.18: HERA cross correlation visibilities averaged in amplitude across LST for three East-West baselines of increasing length: 15 meters, 29 meters and 44 meters (blue, orange and green, respectively). The dashed vertical lines represent the geometric delay of the horizon for each baseline, within which foreground emission is nominally bounded. We see spikes in amplitude at the geometric horizon ("low-delay spikes") and also at higher delays of $|\tau| > 700$ ns ("high-delay spikes"). The low-delay spikes are thought to be either a pitchfork-effect as predicted by Thyagarajan et al. (2015) or antenna cross coupling. Evidence suggests the high-delay features to be some kind of cross coupling systematic.

Next we window the visibilities from 120 – 180 MHz with a Blackman-Harris function (Blackman & Tukey 1958) to limit spectral leakage, and then Fourier transform the visibilities to delay space. At the moment we are only interested in diagnosing systematics, so we do not square the Fourier amplitudes as we would in forming power spectra, meaning the visibilities are in units of Jansky Hz. Figure 4.18 shows the result for the 15-meter baseline (blue), 29-meter baseline (orange) and 44-meter baseline (green). Also plotted as dashed vertical lines are the geometric horizons for each baseline. The nearly-symmetric peaks at each baseline's geometric horizon could be due to the "pitchfork" effect predicted to exist for wide-field radio interferometers (Thyagarajan et al. 2015). The pitchfork effect is not a systematic in the context of this work: it is a natural phenomenon from diffuse foregrounds, and is explained as the boosting of measured diffuse sky power near the horizon, where sky signal shows up in the visibilities with delays of the baseline's geometric horizon. While HERA has a more compact primary beam compared to other low-frequency 21 cm experiments (e.g. MWA, PAPER), the pitchfork effect was nonetheless predicted to exist from simulations of the HERA dish and feed (Thyagarajan et al. 2016). However, these features could also be due to sky emission reflecting off the feed of one antenna and entering the feed of a neighboring antenna (i.e. feed-to-feed reflections or mutual coupling), which is a form of antenna cross-coupling that we would also expect to appear at the delay of each baseline's geometric horizon. While both are expected to produce power at a baseline's geometric horizon, both are also expected to be slowly time-variable, meaning they will occupy similar modes in the delay &

fringe-rate Fourier domains.[5]

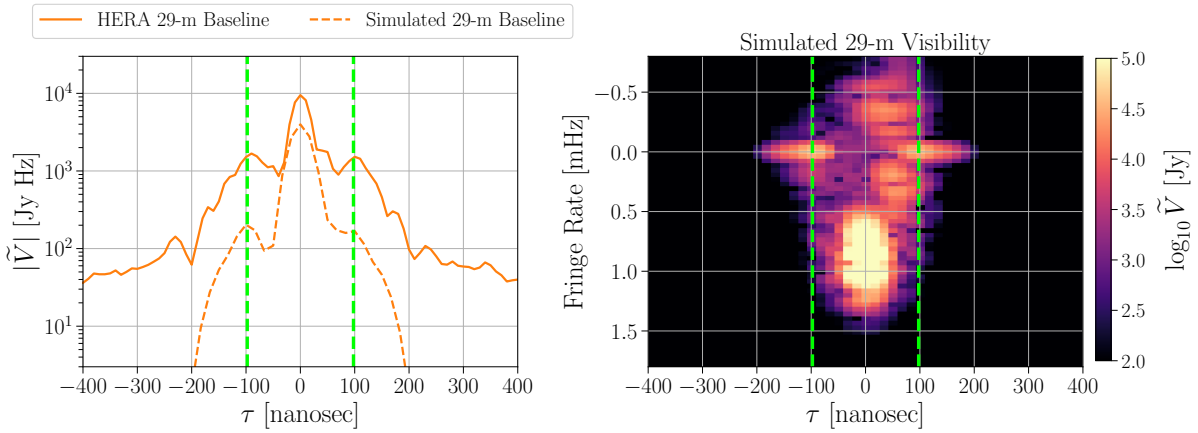

Figure 4.19: Comparison of HERA data with a simulated foreground visibility using the diffuse GSM sky for a 29-meter East-West baseline. **Left:** Averaged HERA cross correlation visibility amplitude in delay space (solid) with an equivalent data product from a simulated foreground visibility with matching LST range (dashed). The geometric baseline horizon is shown at $\sim$ 100 ns (dashed green). While we see some evidence for a slight pitchfork-like structure in the simulated visibility, it is significantly weaker than the power bumps at equivalent delays in the real data. **Right:** The simulated visibility transformed to fringe-rate and delay space, with the geometric baseline horizon over-plotted (dashed green). We can more clearly see the existence of the pitchfork effect in this plot, which is centered at $f = 0$ mHz, extends out to the geometric horizon and falls off after.

In Figure 4.19 we compare the data against a simulated diffuse foreground visibility from Kern et al. (2019), which uses the Global Sky Model (de Oliveira-Costa et al. 2008) as the foreground model and a simulated direction-dependent primary beam response for HERA (Fagnoni et al. 2019). While we do see evidence for a slight pitchfork effect in the simulated data at the geometric horizon delay, its amplitude is considerably weaker than what is observed in the data. There is also some total power missing from the $\tau = 0$ mode, which is likely due to our exclusion of point sources in the simulation. The simulated pitchfork can be seen more clearly when transforming the simulated visibility into fringe-rate and delay space (right of Figure 4.19), where indeed we see the pitchfork occupying $f \sim 0$ mHz modes as expected. This comparison needs further study to better understand the nature of excess power at the horizon delay: the pitchfork feature is highly dependent on the adopted primary beam response at the horizon, which is typically the least accurate aspect of the simulated primary beam response and is also hard to characterize in the field (Lanman et al. 2020). Future work using a combination of perturbed primary beam simulations and empirical beam constraints will help us better disentangle these effects in HERA data.

[5]Cross coupling produces slowly time-variable signals in the visibility because it inserts a copy of the auto-correlation, which is slowly time variable). The pitchfork mechanism is a mimicking of the auto-correlation at declinations near the horizon, thus we expect it to have a slow time variability like the auto-correlation.

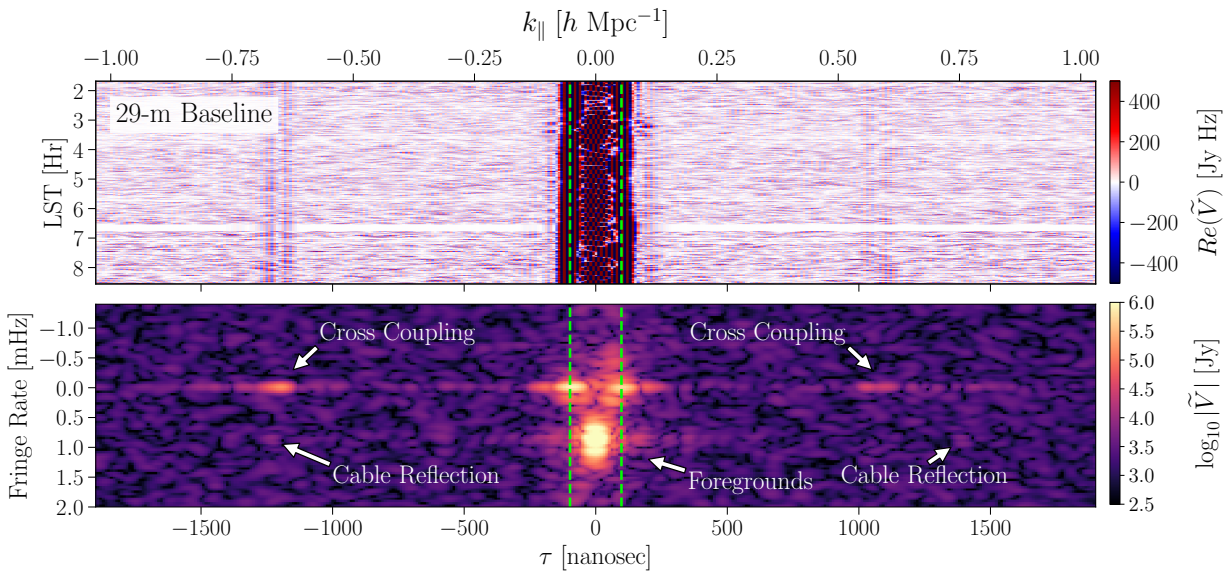

Figure 4.20: A HERA cross correlation visibility showing foregrounds, cable reflections and cross coupling systematics. **Top:** Real component of the visibility in time and delay space, showing foreground power falling within the geometric horizon (green dashed). Notice that power well within the horizon fringes quickly as a function of time, while power near the geometric horizon shows much slower time variability and has spillover to outside the baseline's horizon. **Bottom:** Visibility amplitude in fringe-rate and delay space. Here, we can see the slowly time variable systematics confined to $f \sim 0$ mHz fringe-rate modes, while foreground power is boosted to positive fringe rates. In addition, although not visible in the top plot, we can see the cable reflection just barely visible from the background noise, which appears at positive fringe-rates because it is merely a copy of the intrinsic foreground signal.

We also see evidence in Figure 4.18 for non-negligible amounts of spillover of foreground emission (or supra-horizon emission) beyond the baseline's geometric horizon, which has also been observed by other 21 cm experiments (e.g. Pober et al. 2013b; Beardsley et al. 2016). Supra-horizon emission can come naturally from intrinsic spectral structure of the foregrounds. It can also be created by chromaticity of the instrumental gain that pushes out structure inherently contained within the geometric horizon, or from low-level artifacts in the data which have a similar effect (Offringa et al. 2019). As noted above, the antenna-based gains we apply to the data are simplified to a single flux scaling and a single delay, meaning part of this supra-horizon emission may be due to uncalibrated instrumental gain terms, which we do not explore in this work. For a foreground-avoidance approach to estimating the 21 cm power spectrum, the presence of supra-horizon emission is highly concerning because it limits our ability to measure the low k modes that in theory probe the EoR at the highest signal-to-noise ratio. The upside is if supra-horizon emission is slowly time-variable (as are both the pitchfork effect and antenna cross coupling systematics), then regardless of its origin we can mitigate it by filtering it off in Fourier space. Indeed, this is exactly the principle that cross-coupling subtraction algorithms are founded upon.

Another striking feature in Figure 4.18 is the large amount of excess power above the noise floor at high delay ($|\tau| > 700$ ns). These features, which we refer to as the "high-delay" spikes, exhibit some very peculiar behavior. First, these features seem to be highly baseline-dependent: the three baselines shown in this section are all tied to antenna 11, yet their structures do not seem to be significantly correlated between the baselines. Second, their profile as a function of delay does not show isolated, individual peaks as one might expect from one or a few feed-to-feed reflections, but rather shows a wide range of delays corrupted by excess power. Third, while the structures show up roughly near the delays where we would expect reflections from the 150-m cable to appear, they also show up at delays significantly smaller, enough to necessitate a considerably shorter cable length than 150 meters, which is unlikely. The high-delay spikes exhibit show slow time-variability with their power centered at $f = 0$ mHz, as we would expect from a cross coupling systematic. Figure 4.20 shows the cross-correlation visibility from the 29-meter baseline in time & delay space (top) as well as in fringe-rate & delay space (bottom), where recall the latter is merely the Fourier transform of the former across time. We can clearly see that the high-delay structures are slowly variable, both by their slow movement as a function of time in the top plot, but also by the fact that their power is centered at $f = 0$ mHz in the bottom plot. This is in contrast to the foreground power centered at $\tau = 0$ ns, which oscillates rapidly as a function of time and is therefore boosted to positive fringe-rates, with the exception of the power at the baseline's geometric horizon (dashed green), which, like the systematics at high delay, exhibits slow time variability centered at $f = 0$ mHz.

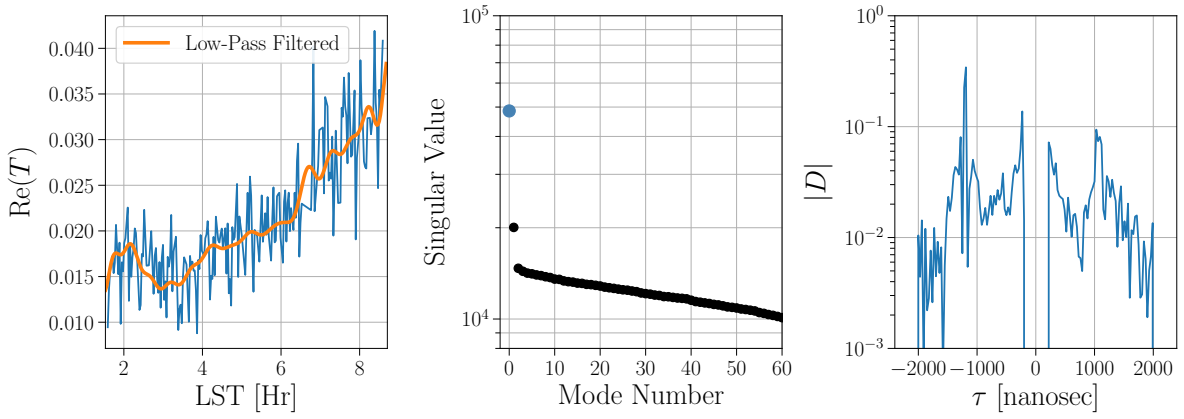

Figure 4.21: Singular value decomposition of the 29-m East-West baseline visibility from Figure 4.20. **Left:** The first **T** eigenvector across time showing its raw form (blue) and its low-pass filtered form (orange), having filtering out modes with $f > 0.46$ mHz with a Gaussian Process model (Kern et al. 2019). **Center:** The first sixty singular values, showing that most of the variance in the systematic-prone regions can be described with a handful of modes before a noise plateau is reached. **Right:** The first **D** eigenvector across delay, showing it picking up on the slowly variable structure at large delays ($|\tau| \sim 1200$ ns) and also some structure near the baseline horizon ($|\tau| \sim 200$ ns).

What we cannot see by looking at the visibility in time & delay but can barely begin to discern

when we transform to the fringe-rate domain are the cable reflections at $|\tau| \sim 1300$ ns. As we saw in Figure 4.15, the measured reflection amplitudes are roughly 3×10^{-3} times the peak power in the visibility. Because the high-delay spikes at $f = 0$ mHz also show up at similar delays and are stronger in amplitude, we cannot see the cable reflections in Figure 4.18 or in the top panel of Figure 4.20 buried under the other systematics. Reflections have the same time-structure as the unreflected signal, so by transforming to fringe-rate space we can isolate them from the slowly time variable systematics, and indeed we can just barely seem them above the noise floor of the cross correlation visibilities at roughly 3×10^{-3} times the main foreground power as expected. Figure 4.20 also shows evidence for the supra-horizon emission having two distinct components: one that is has fast time variability like foregrounds from the main-lobe of the primary beam, and another that is slowly fringing like a cross coupling systematic or a pitchfork effect, and both extend considerably beyond the baseline's geometric horizon.

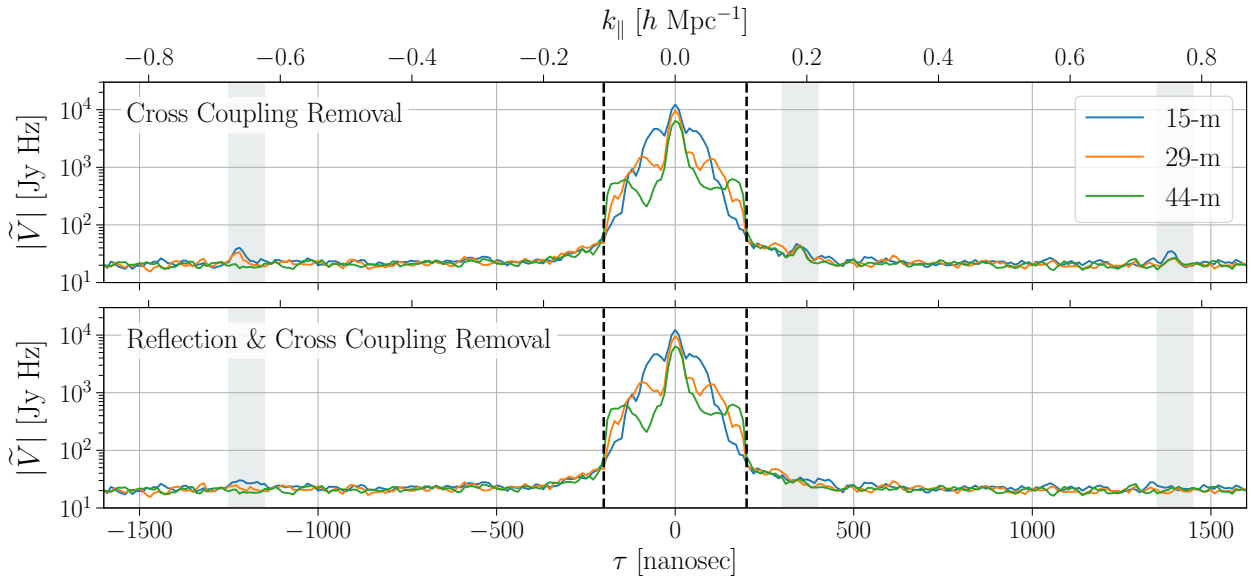

Figure 4.22: HERA cross correlation visibilities from Figure 4.18 after cross coupling subtraction but before reflection calibration (solid) and after both cross coupling subtraction and reflection calibration (dashed). The black-dashed line represents the lower delay boundary of the cross coupling model. Grey shaded regions indicate expected delays for reflection systematics having inspected the auto-correlations for peaks. Joint systematic suppression yields cross correlations visibly free of systematics at the level of the per-baseline noise floor.

Currently, there is not a single physical model for the origin of the high-delay spikes that can explain all of its behavior observed in the data. In section 4.7, we explore some simple physical models for the systematic and show that we can tentatively rule them out; however, further work is needed to more fully understand their origin. Nonetheless, their temporal behavior is suggestive of some kind of antenna cross coupling that occurs at some point along the signal chain. At present, what we can say with certainty is that their time-dependence is highly inconsistent with an EoR

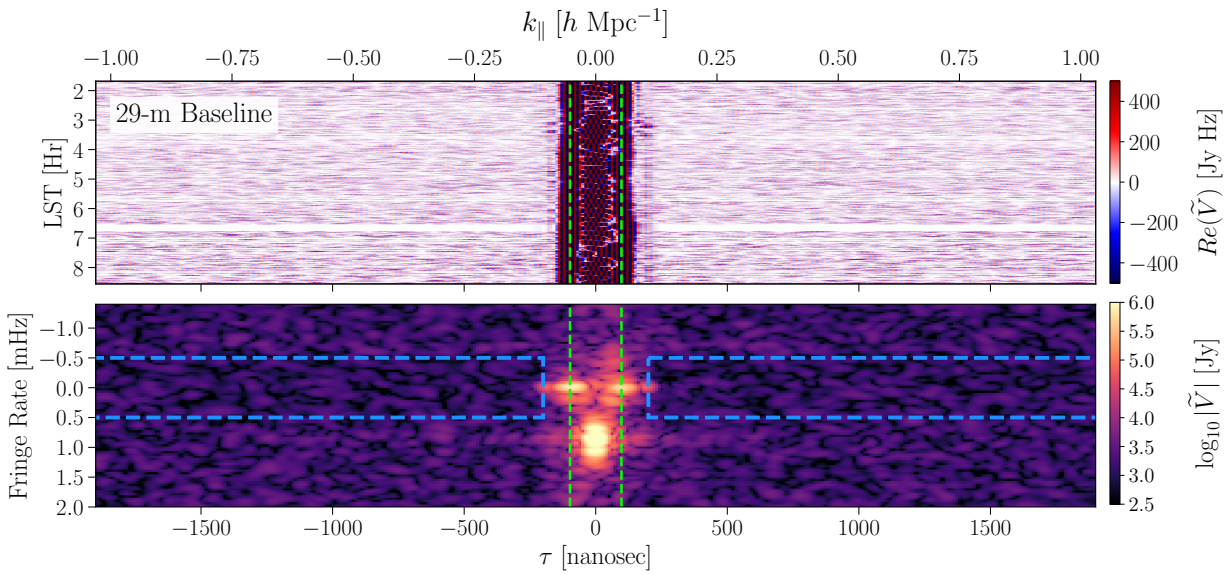

Figure 4.23: Same 29-m visibility in fringe-rate and delay space as shown in Figure 4.20 but now with reflection and cross coupling systematics removed. The blue-dashed region shows where the cross coupling algorithm modeled and removed systematics, and the green-dashed line marks the baseline's geometric horizon.

signal, and as such we can suppress it by filtering the data in fringe-rate space before forming power spectra.

With that in mind, Figure 4.21 shows the result of running an SVD-based cross coupling model (Kern et al. 2019) on the 29-meter baseline data, which decomposes the matrix shown in the top panel of Figure 4.20 into orthogonal time eigenmodes (**T**), orthogonal delay eigenmodes (**D**) and their singular values (**S**). Before taking the SVD we apply a bandstop window on the data matrix that assigns zero weight to all delay modes outside of the range $200 < |\tau| < 2000$ ns, which was chosen to encompass most of the observed cross-coupling systematics and to reject the foregrounds at very low delays. The left panel plots the first **T** eigenmode across time, showing the raw eigenmode (blue) and the eigenmode after low-pass filtering it out to $f_{\rm max} = 0.46$ mHz (orange). We use the Gaussian Process-based filter explored in Kern et al. (2019) to low-pass filter these time-modes. The center panel shows the first 60 singular values, giving us a sense for how much the information content is isolated into the first few eigenmodes. We find that most of the structure can be described with only a handful of modes before reaching a plateau. In forming the systematic model we keep the top 30 modes out of ~ 1000 and truncate the rest. We choose 30 based on inspection of Figure 4.21, which gets the strongest first few modes but also tries to get some of the modes after the plateau as they may be picking up on the systematic at a low level. Lastly, the right panel shows the first **D** eigenmode across delay, showing it picking up the high-delay cross coupling systematic and some of the supra-horizon emission at low delay. In addition to picking up on the systematic, the SVD will pick up on the noise of the data as well. However, because we keep only a small

fraction of the eigenmodes and additionally smooth them across time, we do not suspect that we are subtracting a significant component of the noise in the process of systematic removal. For the 29-meter baseline, the number of Fourier modes kept in the low-pass smoothing filter is ∼1/80 the total number of Fourier modes in the data.

We repeat this for the other baselines at hand, low-pass filtering the $\mathbf{T}$ basis vectors from the 15-meter and 44-meter baselines with $f_{\rm max} = 0.14$ and 0.83 mHz respectively, using a Gaussian-process-based smoothing for the low-pass filter. Figure 4.22 shows the baselines in Figure 4.18 after cross-coupling subtraction, with the vertical dashed line showing the minimum delay of the cross coupling model at $\tau = 200$ ns. The top panel shows only cross coupling subtraction, where we see significant suppression of the high-delay spikes and the outer edge of the low-delay spikes. As expected, after subtracting the strong cross coupling terms at high-delay we are left with the appearance of localized bumps that mark the cable reflections (marked in grey bands), which recall were not subtracted out with the cross coupling because they occupy fringe-rate modes that were filtered out of the systematic model in the process of smoothing. The bottom panel shows the data after applying reflection calibration from subsection 4.5.1 and cross coupling subtraction, showing that the data is now consistent with a scale-independent thermal noise floor for all delays outside $|\tau| > 500$ ns.

There is, however, still a slight slope in the data at intermediate delays of $200 < |\tau| < 500$ ns, which is part of the supra-horizon emission we observed earlier. To ensure that this tail is not coming from the cross-coupling component that we attempted to filter out, we can plot the systematic-subtracted data in fringe-rate & delay space, which is shown in Figure 4.23 with the blue-dashed region showing the region of Fourier space where cross coupling subtraction was performed. Figure 4.23 confirms that the excess signal between $200 < |\tau| < 500$ ns observed in Figure 4.22 does not come from modes that should have been subtracted in the process of cross coupling removal, and originates from the second supra-horizon component at higher fringe-rates. As discussed above, this supra-horizon emission can come from uncalibrated bandpass terms or from low-level artifacts in the data, which push foregrounds out in delay that were intrinsically contained within the geometric horizon. These effects can be somewhat mitigated with better bandpass calibration and data flagging, but are still active areas of research in the literature. Additionally, the slight overlap of low fringe-rate power inside the dashed region at $|\tau| = 200$ ns is produced by a windowing function applied to the data before taking its Fourier transform.

Signal loss is a principal concern when applying any baseline-dependent operation to the data, as we have done with cross coupling subtraction. In Kern et al. (2019) we vet our cross coupling modeling algorithms for EoR signal loss against numerical visibility simulations of the HERA Phase I system. We show that by low-pass filtering the systematic model along time (Figure 4.21), we can harden our systematic model against EoR signal loss to an almost arbitrary level. In our case, we chose the fringe-rate bounds above by adopting a signal loss tolerance of 1% in EoR power, which is below the expected measurement error of the full HERA array. We refer the reader to our analysis and discussion in that paper for more details on signal loss quantification in the context of cross coupling removal.

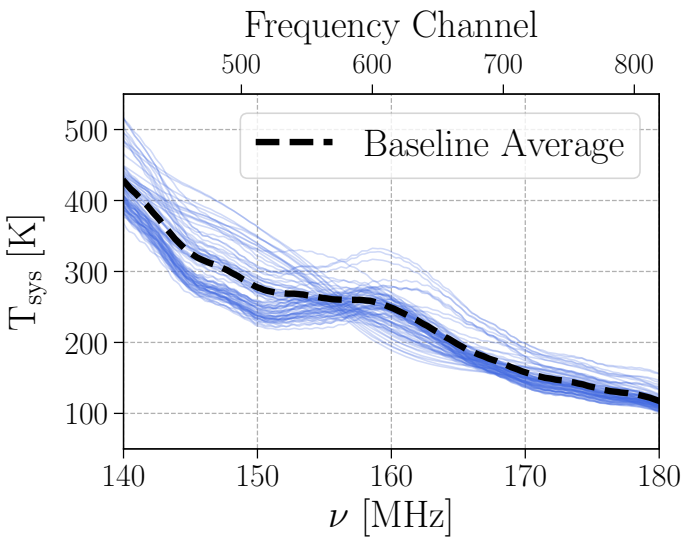

Figure 4.24: System temperature curves for all baselines used in the power spectral analysis (colored points), and their average (black dashed). Delay spectra presented in this section are formed between channels 450 and 650 (144 – 163 MHz) with an effective system temperature of $\sim$ 270 K.

4.6 Power Spectrum Estimation

Now that we've demonstrated that we can suppress reflection and cross coupling systematics for a few baselines down to their individual noise floors, we would like to prove that we can similarly do this for baselines across the entire array, and confirm that these systematics are a non-limiting factor in the power spectrum even after redundant baseline averaging. We will focus on the same three baseline orientations (14-m, 29-m and 44-m East West baselines), but now look at all baselines within the array that fall within each baseline group. To estimate the three-dimensional 21 cm power spectrum, $P_{21}(\mathbf{k})$, we use the delay spectrum estimator (Parsons et al. 2012a; Liu et al. 2014a; Parsons et al. 2014). This is the same estimator described in section 3.5.

Cross multiplying a visibility with itself to form a delay spectrum will result in an overall bias in power due to the noise present in the data. To avoid this, we take visibility spectra adjacent to each other in LST separated by 10.7 seconds and apply a phasing term to align their phase centers before cross multiplication (Pober et al. 2013a). This means the two visibilities to leading order measure the same cosmological mode on the sky but have uncorrelated noise realizations, such that they do not produce a noise bias upon cross correlation.

Thermal noise in interferometric visibilities is mean-zero, Gaussian distributed, and is statistically uncorrelated on all time and frequency scales; however, it generally is non-stationary, and will have an amplitude dependence as a function of LST and frequency. A signal chain's *system temperature* is proportional to the total amount of noise power received by the analogue system, and is the sum of the sky noise and receiver noise,

$$T_{\rm sys}(\nu, t) = T_{\rm sky}(\nu, t) + T_{\rm rcvr}(\nu, t) \text{ [K]}. \tag{4.20}$$

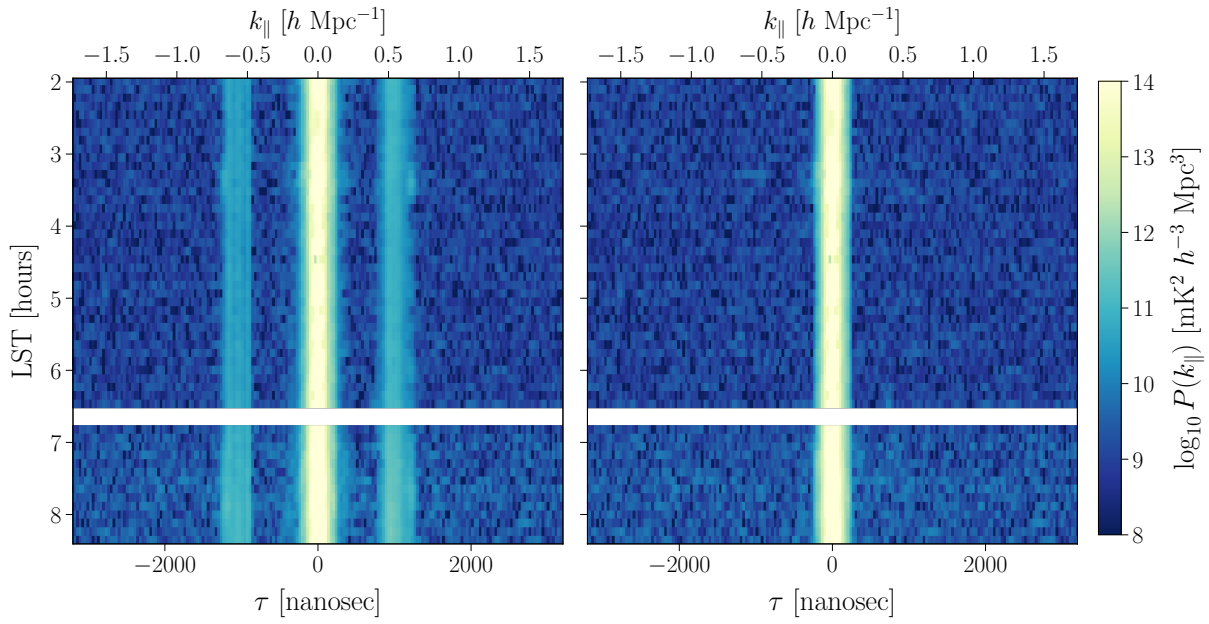

Figure 4.25: An averaged power spectrum waterfall of the East-West 15-m group showing the absolute value of the real component of the power spectra, having first incoherently averaged 35 separate baseline-pairs in the group. We plot the data with systematics in (left) and with systematics removed (right).

In practice, antenna signal chains will have variable system temperatures due to different angular primary beam responses and different receiver properties. A visibility-based system temperature can therefore be estimated, which is the system temperature as measured by a particular baseline. This can be estimated by taking differences of adjacent pixels in time and frequency and relating its RMS to a system temperature via the radiometer equation,

$$\sigma^{ij}_{\text{rms}} = \frac{2k_b\nu^2}{c^2\Omega_p}\frac{T^{ij}_{\text{sys}}}{\sqrt{\Delta\nu\Delta t}}, \tag{4.21}$$

where σ^{ij}_{rms} is the RMS of the visibility between antennas i and j in Jansky, k_b is the Boltzmann constant, ν is the average observing frequency, Ω_p is the angular integral of the peak-normalized primary beam response in steradians, $\Delta\nu$ is the correlator channel width in Hz and Δt is the correlator integration time in seconds (Thompson et al. 2017). Another estimate of the noise comes directly from the auto-correlation visibility, which itself is a measurement of the total power received by a particular antenna. For a cross-correlation visibility between antenna i and j, we can estimate the baseline's system temperature as

$$\sqrt{V_{ii}V_{jj}} = \frac{2k_b\nu^2}{c^2\Omega_p}T^{ij}_{\text{sys}}, \tag{4.22}$$

where V_{ii} is the auto-correlation visibility of antenna i. While both methods are comparable, we

defer to using the auto-correlations, which in practice generally lead to more stable and cleaner noise models.

Figure 4.24 shows system temperature estimates for each baseline participating in the analysis (blue) and the averaged system temperature, which is each baseline's system temperature averaged in quadrature. Again, because we have not corrected for the bandpass structure of the gains, the large-scale fluctuations in Figure 4.24 are not unexpected, and would be smoothed-out after solving for and applying the appropriate instrumental gains. The presence of such structure in the noise curves does not change the fundamental results of this section. Power spectra presented in this section are formed between channels 450 – 650 (144 – 163 MHz) with an effective system temperature of ~ 270 K.

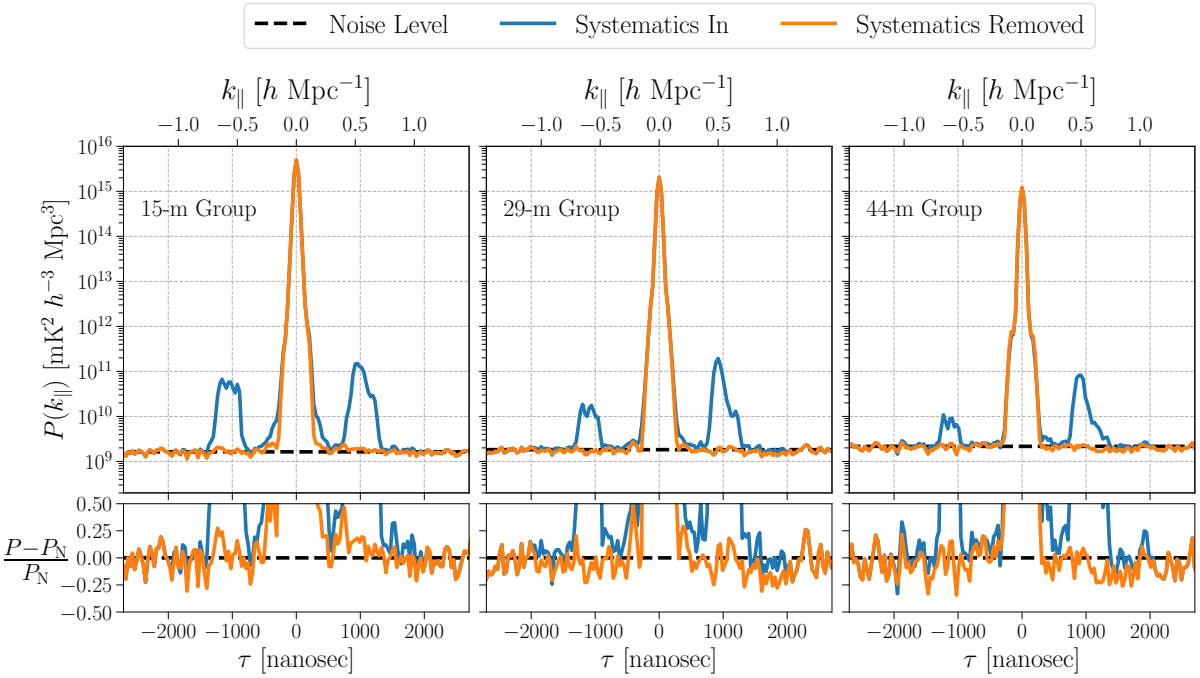

Figure 4.26: Delay spectra for three unique baseline lengths oriented along the East-West axis without systematic removal (blue) and with systematic removal (orange). The power spectra are formed directly from the visibilities for each baseline in the array, are incoherently averaged within each redundant group, and then their absolute value is averaged across the remaining bins in LST. We see suppression of high delay systematics down to the integrated noise floor, and get some suppression of supra-horizon power at low delay.

With an understanding of the noise properties of our data, we can compute a theoretical estimate of the noise power spectrum, P_N, which is equivalent to the root-mean square (RMS) of the power spectrum if the only component in the data were noise. This is one way to measure the uncertainty on the estimated power spectra, but also represents the theoretical amplitude of the power spectra in the limit that they are noise dominated (as opposed to signal or systematic dominated). This is

given in Cheng et al. (2018) as

$$P_{\rm N} = \frac{X^2 Y \Omega_{\rm eff} T_{\rm sys}^2}{t_{\rm int} N_{\rm coherent} \sqrt{2 N_{\rm incoherent}}}, \tag{4.23}$$

where the X and Y scalars are the same as before, $T_{\rm sys}$ is the system temperature in milli-Kelvin, $t_{\rm int}$ is the correlator integration time in seconds, $N_{\rm coherent}$ is the number of sample averages done at the visibility level (i.e. before visibility squaring), and $N_{\rm incoherent}$ is the number of sample averages done at the power spectrum level (i.e. after visibility squaring). $\Omega_{\rm eff}$ is the effective beam area given by $\Omega_{\rm eff} = \Omega_p^2/\Omega_{pp}$, where Ω_p is the integral of the beam across the sky in steradians, and Ω_{pp} is the integral of the squared-beam across the sky in steradians (Pober et al. 2013a; Parsons et al. 2014). We calculate P_N for each redundant group using the baseline-averaged system temperature.

The data are natively sampled at a 10.7 second cadence. Before forming power spectra, we coherently average each visibility across LST for 3.6 minutes (20 samples), applying a fringe-stop in each averaging window to limit sky signal attenuation. We select a wide spectral window between channels 400 to 700 (139 – 168 MHz), and apply a Blackman-Harris windowing function before transforming to Fourier space. Because the cosmological signal undergoes non-negligible evolution within such a bandwidth we would not normally use such a wide bandwidth for setting upper limits, however, we do this to achieve better resolution in delay space for diagnostic purposes. We then cross-multiply the visibilities and apply the necessary normalization factors. For simplicity, we only form power spectra by cross multiplying baselines with themselves (at adjacent times), and do not cross correlate different baselines within redundant groups. Then we average the power spectra within each redundant group (i.e. an incoherent average). For the 15-m, 29-m and 44-m groups this involves averaging 35, 28, and 20 independent baselines, respectively.

What we are left with is a single complex-valued power spectrum waterfall for each redundant group as a function of LST and delay, consisting of 60 leftover time bins and 200 delay bins. In Figure 4.25 we show this for the 15-m group with and without systematic removal (right & left). In our final step, we take the real component of each power spectrum waterfall and average its absolute value over the remaining time bins with uniform weighting for each bin. This is done to make a higher signal-to-noise measurement of the noise floor at the level of the power spectrum waterfall: we could have gained more sensitivity by not taking the absolute value before averaging, but our point here is to make a visually clearer comparison with the known noise level rather than gain increased sensitivity. Figure 4.26 shows the power spectra of the data without systematic removal (blue), with systematic removal (orange) and also shows the theoretical noise level given our visibility noise estimates and taking into account the various forms of averaging before and after squaring the visibilities (black dashed). In this case, the systematic removal includes both cross coupling subtraction and reflection calibration. We find that we can suppress the observed systematics by roughly two orders of magnitude in power, enabling us to achieve six orders of magnitude in dynamic range with respect to the peak foreground power for $|k_\parallel| > 0.2\ h\ {\rm Mpc}^{-1}$.

The power spectra in Figure 4.26 show general agreement with our prediction of the thermal noise floor for delays considerably outside of the foreground wedge. The bottom panels show the fractional offset of the data with respect to the analytic noise, showing that $P_{\rm N}$ is broadly consistent with the data at the $\sim$ 15% level. While the 29-meter and 44-meter group seem to exhibit slight

systematic offsets between in the observed and predicted P_{N}, this offset is contained within the majority of the random fluctuations seen in the noise floor: without noise propagation trials this offset is hard to quantify rigorously, which we defer to future work. Although the geometric horizon for these short baselines is on the order of 50 – 150 ns, the Blackman-Harris windowing function pushes this out by about +100 ns, such that their effective horizon is on the order of 150 – 250 ns. However, we can still see some amount of positive power near the transition region, particularly for the 15-meter group. This could be due to uncalibrated bandpass terms in the data, low-level artifacts in the data missed by RFI flagging, or residual reflection and cross coupling systematics. More complete gain calibration and deeper integrations will allow us to investigate this at higher SNR levels.

Gosh et al. (in prep.) also propose methods for subtracting systematics observed in the HERA Phase I instrument using a Gaussian Process based model. With their model, they find good subtraction of the systematic down to similar dynamic ranges (10^6 in power), at the cost of possible signal loss at the $\sim 10\%$ level. Systematics of a similar nature were also observed in the HERA-19 Commissioning array (Kohn et al. 2019). However, a direct comparison with this work is difficult because the array was re-configured en route to the Phase I configuration.

As a final note, we would like to clarify how we came to the noise level plotted in Figure 4.26. Noise in the interferometric visibility is a complex Gaussian random variable, meaning that when we form power spectra by squaring the visibilities we are left with a noise component that is drawn from a complex normal-product distribution. A real-valued normal-product distribution can be shown to be described by a modified Bessel function of the second kind of order 0 (Wells et al. 1962; Cui et al. 2016). A complex-valued normal-product random variable is simply the sum of two real-valued normal-product random variables, which means its probability density function (PDF) is a convolution of the Bessel function with itself and turns out to be a double-sided exponential distribution. Therefore, after squaring the visibilities, noise in the power spectrum is drawn exponentially.

However, most power spectrum pipelines will average the data after squaring the visibilities (i.e. incoherent averaging), which will re-Gaussianize the data due to the Central Limit Theorem. Indeed, to create Figure 4.25 we perform a few dozen incoherent averages across redundant baselines after squaring the visibilities, meaning it is fair to assume the noise in our power spectrum is Gaussian-distributed. However, in order to collapse our data along the LST axis to form Figure 4.26 we took the absolute value of the real-component of the power spectrum before averaging. The absolute value operation transforms the noise from a Normally-distributed, mean-zero random variable into a random variable drawn from a half-Normal distribution, which is no longer mean-zero and has an expectation value of $\sigma\sqrt{\frac{2}{\pi}}$. Recall from Equation 4.23 that P_N tells us the expected RMS of the real (or imaginary) component of the complex power spectrum due to thermal noise. Therefore, the act of taking the absolute value of the real-component of the power spectra and averaging across LST means we need to multiply our final P_N estimate by a factor of $\sqrt{2/\pi}$, which is what is actually plotted in Figure 4.26 as the black-dashed line.

4.7 Physical Models for the Observed Cross Coupling

Here we explore the feasibility for some simple physical models as the origin of the "high-delay" cross coupling systematics investigated in subsection 4.5.2. In summary, we cannot find a single model that explains all of the observed behavior of the systematics, but we can tentatively rule out some simplistic models. In what follows, we adopt the mathematical conventions in Section 2 of Kern et al. (2019) when discussing voltage spectra, visibilities and coupling coefficients. Specifically, for two antennas 1 & 2 with intrinsic voltage spectra v_1 and v_2, we can write the voltage of antenna 1 corrupted by a cable reflection as

$$v_1' = v_1(1 + \epsilon_{11}) \tag{4.24}$$

where ϵ_{11} is the cable reflection coefficient, and we can write the voltage of antenna 1 corrupted by cross coupling from antenna 2 as

$$v_1' = v_1 + \epsilon_{21} v_2 \tag{4.25}$$

where ϵ_{21} is the cross coupling of antenna 2's voltage into antenna 1's voltage. If the uncorrupted cross-correlation visibility is $V_{12} = v_1 v_2^*$, then the visibility corrupted by a reflection from antenna 1 can be written as

$$V_{12}' = v_1' v_2^* = v_1 v_2^* (1 + \epsilon_{11}), \tag{4.26}$$

and the visibility corrupted by cross coupling can be written as

$$V_{12}' = v_1' v_2^* = v_1 v_2^* + \epsilon_{21} v_2 v_2^*. \tag{4.27}$$

4.7.1 A Noise Source in the Field

A cross coupling-like signal can be generated by a stable noise source in the field, which will not fringe over time. We can rule this out as the systematic mechanism simply based on the fact that we observe cross coupling systematics on short baselines at delays of $\geq$ 1000 ns: any (unreflected) source situated in the field will have a maximum achievable delay corresponding to the baseline's geometric horizon, which for short baselines is from 50 - 150 ns.

4.7.2 Mutual Coupling Boosted by Cable Reflections

One way to get cross coupling at high delays is to take cross coupling at low delays (e.g. mutual coupling) and boost it to high delays via a cable reflection. If antenna 1 observes cross coupling from antenna 2 that then travels down and gets reflected in the cables of antenna 1, we can write the final measured visibility as

$$V_{12}' = (v_1 + \epsilon_{21} v_2)(1 + \epsilon_{11}) v_2^* = v_1 v_2^* + v_1 v_2^* \epsilon_{11} + \epsilon_{21} v_2 v_2^* + \epsilon_{21} v_2 v_2^* \epsilon_{11}. \tag{4.28}$$

On the RHS of Equation 4.28, we recognize the first term as the uncorrupted visibility, the second term as the cable-reflected visibility, the third term as the first-order cross coupling systematic at low delay, and the last term as the cross coupling systematic boosted to high delay. What we find is that the systematic can only be boosted to specific delays, determined by the product $\epsilon_{21}\epsilon_{11}$. What we see in the data (specifically the right side of Figure 4.20) are cross coupling systematics at delays that are not consistent with this expectation. Furthermore, the high-delay systematics do not look like a shifted version of the low-delay systematics, which is also a prediction of this model.

4.7.3 A Broadcasting Antenna

This model is a hybrid of the first two models, and states that a single antenna, say antenna 3, receives sky signal that traverses down its signal chain, is reflected back up one of its cables, is *re-broadcasted out into the field* and then picked-up by neighboring antennas, mimicking a stable noise source in the field that has acquired a large delay lag due to the cable reflection in antenna 3's signal chain. We can write the visibility between antenna 1 and 2 in the presence of this signal as

$$V'_{12} = (v_1 + \epsilon_{31}\epsilon_{33}v_3)(v_2 + \epsilon_{32}\epsilon_{33}v_3)^* \tag{4.29}$$

$$= v_1 v_2^* + v_1\epsilon_{32}^*\epsilon_{33}^* v_3^* + \epsilon_{31}\epsilon_{33}v_3 v_2^* + \epsilon_{31}\epsilon_{33}v_3\epsilon_{32}^*\epsilon_{33}^* v_3^*. \tag{4.30}$$

We recognize the first term on the RHS as the uncorrupted visibility, the fourth term as a standard cross-coupling term (due to the auto-correlation nature of $v_3 v_3^*$) that has had its large delay canceled out due to $\epsilon_{33}\epsilon_{33}^*$ and thus does not appear at high delays. Only the second and third terms will appear at high delays, but we can see that these terms are actually fringing terms because they contain products like $v_1 v_3^*$ rather than $v_3 v_3^*$ and thus will not appear centered at a fringe rate of 0 mHz, as we observe in the data.

4.7.4 Summary

While we have tentatively ruled out a few simple physical models, we still cannot point to a single mode that seems to explain the wide variety of behavior observed in the high-delay systematics. What we can say is that the high-delay $f \sim 0$ mHz terms seem to be physically disconnected from the $f \sim 0$ mHz terms at low delays (i.e. at each baseline's geometric horizon). Regardless of its origin, we do know that the high-delay features do not look like an EoR signal, and can therefore be filtered out of the data. Work is currently underway to assess whether these systematics appear in the upgraded HERA Phase II system, and if so what can be done in the field to mitigate their presence in the field.

Chapter 5

Power Spectrum Analysis on Deep HERA Integrations

This chapter focuses on the power spectrum analysis of HERA's Second Internal Data Release (IDR2), which to-date is the most scrutinized dataset from HERA (Dillon 2017). Previous chapters have demonstrated techniques with a small subset of IDR2 data, while the chapter discusses an analysis across the entire data release. Some of this material draws from the currently in-progress HERA power spectrum analysis on the IDR2 dataset (HERA Collaboration 2020 in prep).

5.1 Data Processing

The data discussed in this section is similar to those discussed in chapter 3 and chapter 4, but span a wider range in observing nights. Specifically, this section looks at an 18-night dataset observed from December 10th – 28th, 2017. The instrument and correlator specifications are the same as those outlined in Table 3.1. In total, 13 out of 52 total antennas are flagged in the data reduction step, leaving 39 functional antennas for downstream analysis. Observations span roughly 12 hours each night, from sunset to sunrise local South African time. Over the course of 18 consecutive nights, this leads to a stripe-like sky coverage spanning 0 to 12 hours in right ascension Figure 5.1. This coincides with a patch of the sky largely devoid of diffuse galactic emission ($\alpha < 4$ hours), which is our nominal science field. In this section, we briefly summarize the data reduction pipeline applied to the raw data, parts of which have already been discussed in detail in earlier chapters.

5.1.1 Calibration

For this analysis, we solve for the direction-independent, antenna-based gains for the XX and YY visibility polarizations. To calibrate the data we first need to identify faulty antennas, which can arise due to problems with the feed or signal chain of an antenna and is fairly apparent by looking at the raw data. To capture this, we compute per-antenna metrics from the raw data to perform initial

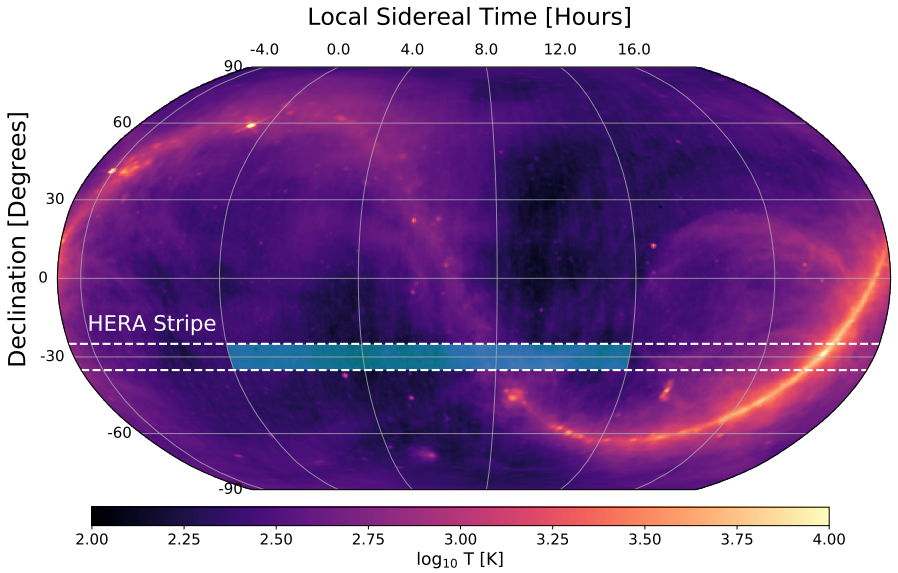

Figure 5.1: The Global Sky Model at 150 MHz (de Oliveira-Costa et al. 2008) showing the bright diffuse foregrounds from the galaxy. HERA observes a narrow stripe of sky centered at $\delta = -30.7°$ with a primary beam FWHM of 10°. The data presented here spans a range of LSTs from 0 to 12 hours (blue shaded).

antenna flagging, specifically looking at the average visibility amplitude for all baselines connected to each antenna. Any antennas that show consistent outlier behavior from the mean (greater than 5σ) are flagged. This eliminates 5 of the 52 antennas in the array. Further antenna cuts are made in the process of calibration, which we discuss next.

Our calibration approach uses redundant-baseline calibration to solve for the relative gains without needing a model of the sky. This is performed independently for every 10.7 second integration for each night in the data release (Dillon et al. 2020). A per-antenna χ^2 metric is computed (Dillon et al. 2020), and outlier antennas at the $> 4\sigma$ threshold are flagged. The calibration is then repeated (withholding the newly flagged antennas), the per-antenna χ^2 re-computed and outlier antennas flagged until no new outlier antennas appear. This process flags another 8 antennas from the dataset. As discussed in chapter 3, redundant calibration still leaves array-wide components of the gain unconstrained, which can be filled-in with a partial absolute calibration. The model visibilities for absolute calibration are constructed with the same methodology presented in section 3.2: we choose a single field, calibrate the array at that time and transfer the calibration to all other times in that observing night. We do this for three nights in the season, using different calibration fields for each night, and smoothly stitch the visibilities together from each night to create a set of calibrated model visibilities that spans nearly 24-hours in duration. The partial absolute calibration is then performed for each 10.7-second integration for each night in the data release.

The data are then flagged for radio frequency interference (RFI) and other anomalies. The flagging procedure using a median filter to identify sharp outliers in the visibilities (Kerrigan et al. 2019). This is combined with a series of data metrics for each time and frequency pixel that come

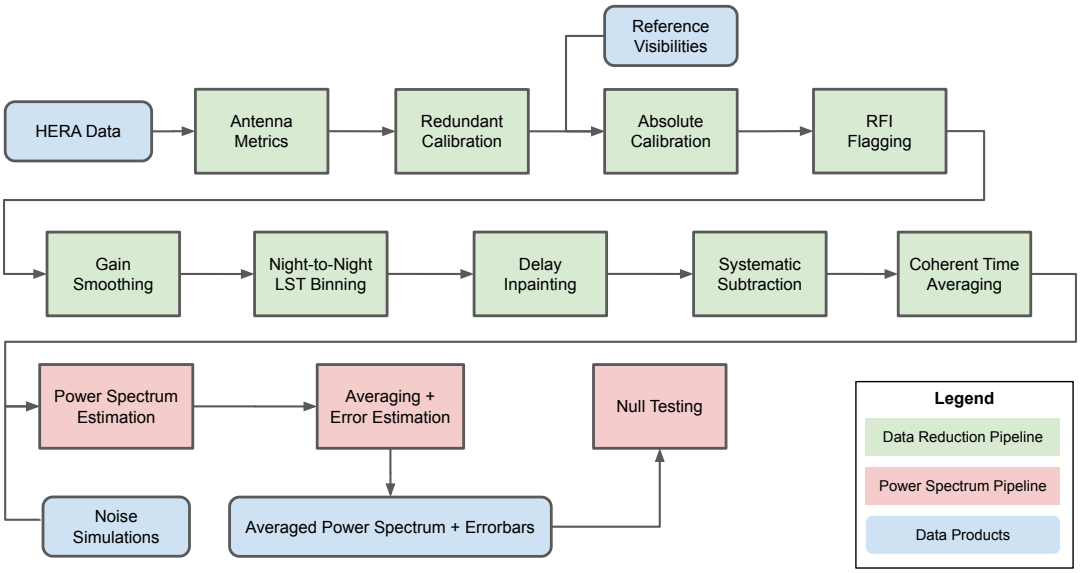

Figure 5.2: A diagram of the reduction, power spectrum, and validation pipelines, starting with raw HERA data and ending with averaged power spectra. The blue boxes represent data products, while the green and red boxes represent steps in the reduction and power spectrum pipelines, respectively. Boxes with dashed borders represent elements covered by the HERA validation pipeline, which is currently a work in progress and is not discussed in detail here. Noise simulations are generated and fed through the power spectrum pipeline for diagnostic purposes when evaluating null tests.

from the calibration, including the reduced χ^2 of the redundant and absolute calibration steps and their gain solutions. Frequencies and times that are substantially different from the model visibilities or from their neighboring data points are likely outliers due to RFI, and are generally caught by these metrics. Each data product forms a metric, which is normalized by the noise of the metric by de-trending it and dividing by a robust measure of its standard deviation. These synthesized metrics are flagged when any individual time-frequency pixel exceeds a 5σ threshold. Afterward, an additional watershed iteration is performed on the metrics which flags at a 2σ threshold (Kerrigan et al. 2019).

Lastly, we smooth the gains to eliminate excess temporal and frequency structure accrued by errors in the calibration process. Motivated by our understanding of the intrinsic drift in the gains, and by our understanding of the spectral scales at which our gains are reliable, we use a Fourier filter to smooth the gains on a 6-hour time scale and on a 10 MHz frequency scale.

5.1.2 Night-to-Night Binning

With each of the 16 nights of the data release calibrated indepdently, we next seek to coherently average the visibilities across nights. We do this by constructing a single fixed grid in local sidereal time (LST) from 0 - 24 hours at a cadence of 21.4 seconds. For each LST bin, we collect all of

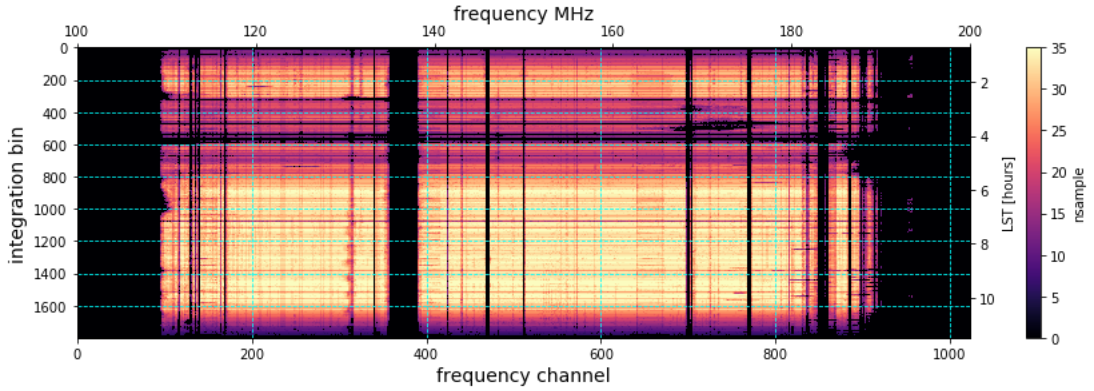

Figure 5.3: A colormap of the Nsample count after LST binning for each frequency and integration bin in the data. Persistent and strong RFI (e.g. ORBCOMM at 138 MHz) are entirely flagged leading to Nsample counts of zero. Based on this map, all times that have a frequency-averaged Nsample less than 5 are flagged due to suspicious behavior. Likewise, frequency channels with a time-averaged Nsample less than 5 are also flagged.

the data from each night in the data release that falls within the bin bounds and place them into the bin. To account for the slight drift in the pointing center from night-to-night (due to sidereal drift), we rephase the pointing center of each observation to a common zenith pointing. We then perform another round of outlier clipping to reject possible low-level RFI and other spurious anomalies in the data before averaging. We do this with a median absolute deviation (MAD) based sigma clipping, which computes the robust standard deviation of the bin

$$\sigma_{\rm mad} = 1.482 \times {\rm med}|\boldsymbol{x} - {\rm med}(\boldsymbol{x})|, \tag{5.1}$$

where $\boldsymbol{x}$ is the vector containing all nightly data for a single LST bin, and 1.482 is a correction factor that makes the metric comparable to the standard deviation in the case of Gaussian-distributed data. All data points in the bin that fall outside of $5\sigma_{\rm mad}$ of the sample median are flagged before taking a uniform average of the bin. This is repeated for all baselines, times and frequencies throughout the dataset before the bins are averaged.

If the entire dataset was free of flags, we would expect each final LST bin to contain 36 samples in its average, which we refer to as its Nsample count.[1] In reality, many of the LST bins contain very low Nsample counts due to data that is flagged across all or many nights at a specific time and/or frequency. Figure 5.3 shows an example Nsample count as a function of frequency and LST after night-to-night binning of the IDR2 dataset. Persistent narrowband RFI at expected frequencies means these channels are completely devoid of samples, and are therefore entirely flagged. Likewise at certain LSTs there are strong point sources that transit nulls in the primary beam that degrade our calibration and cause our RFI flagger to more aggressively flag the data at those times (e.g. near the Fornax A transit near α ~3 hours). However, there are some times and frequencies where

[1] 18 nights sampled at a raw cadence of 10.7 seconds means two samples per night are collected in each 21.4 second LST bin.

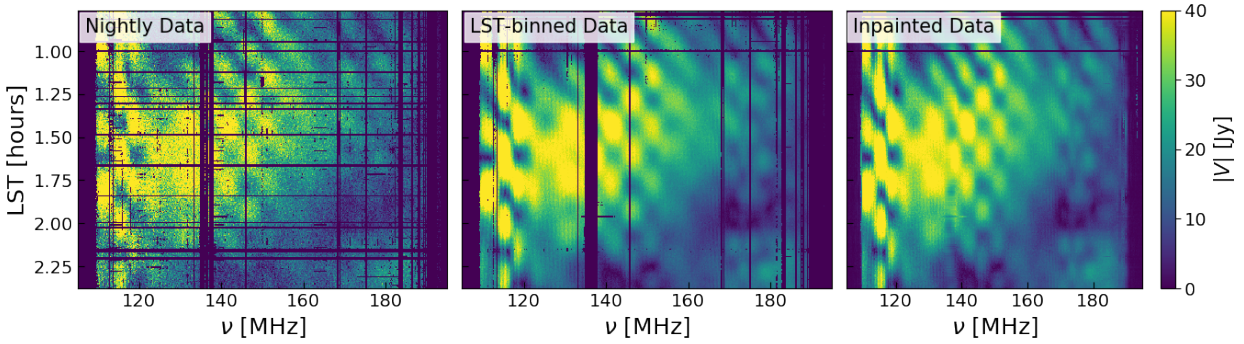

Figure 5.4: A calibrated HERA visibility before night-to-night LST binning (left), after LST binning (center), and after frequency-based inpainting (right). While inpainting reduces the final flag occupancy of the data, its performance is best when applied to narrowband RFI events, and is therefore not currently used as a practical remedy for the wideband RFI events such as ORBCOMM at 138 MHz.

not all, but the majority of the data are flagged, resulting in a small but non-zero Nsample. This is highly suspicious behavior, and likely suggests that those data should be flagged: we therefore flag all times and frequencies with an average Nsample count that dips below 5 samples.

5.1.3 Data Inpainting

Flagged data present a challenge for power spectrum analyses, as the sharp discontinuities they introduce in their weighting function create spectral ringing in the Fourier domain. Because the EoR signal is buried under foreground emission that is bright but locally compact in Fourier space, complex flagging patterns in the data can severely hamper on foreground-avoidance strategies. Even for foreground avoidance techniques, non-trivial flagging patterns can create low-level discontinuities in coherently averaged data that leads to similar spectral ringing systematics (Offringa et al. 2019). A common solution to missing or nulled data is to "inpaint" the missing data with an estimate of its value based on nearby data points. Various techniques have been developed for this process, and has been used frequently in the context of angular power spectrum estimation and map making of the Cosmic Microwave Background (Bucher & Louis 2012). Another example is the Hogbom CLEAN algorithm used widely in interferometric imaging (Högbom 1974). In this work, we use a frequency-based 1D CLEAN algorithm to inpaint flagged channels for each baseline independently (Parsons & Backer 2009).

Figure 5.4 shows a visibility waterfall before LST binning (left), after LST binning (center) and after inpainting (right). From pre to post LST binning, we see a reduction in the overall flag occupancy due to different nights having different flagging masks, in addition to the suppression of the thermal noise. Inpainting is performed across the near entirety of the bandwidth, from 110 – 190 MHz, and is able to fill-in flagged data due to both narrow-band and wide-band RFI events. Because inpainting is applied across frequency for each time integration independently, integrations that are flagged across all frequencies cannot be inpainted and remain flagged. The

accuracy of inpainting is dependent on both the algorithm used and the relative amount of flags in the data and their distribution. One can see, for example, defects in the CLEAN inpainted data in Figure 5.4 near the broad ORBCOMM band at 138 MHz: due to the sheer width of the flagged region across frequency, the CLEAN algorithm has trouble accurately determining the CLEAN model (Ewall-Wice et al. 2020). We therefore only use inpainted data in our power spectrum analysis over narrowband RFI events, such as those near 170 MHz.

5.1.4 Instrumental Systematic Modeling

Instrumental systematics in the HERA Phase I system are discussed in detail in chapter 4, and include cable reflections between the 20-meter and 150-meter coaxial cables in the front end signal chain, as well as a overall offset in the visibilities produced by spurious cross-coupling within the front end system. In addition to the two reflection components produced by a reflection at the end of the two cables, Kern et al. (2020b) show preliminary evidence for a large number of sub-reflections occurring throughout the length of the cables, generating a range of reflection terms spanning small and large time lags. Cable sub-reflections are also observed in direct electromagnetic measurements of HERA's front end system (Fagnoni et al. 2019), although these controlled lab measurements exhibit reflection amplitudes about an order of magnitude lower than what is observed in the real system.

The methodology presented in Kern et al. (2020b) shows how one can go about calibrating out multiple, overlapping reflection terms iteratively. Without knowing exactly how many sub-reflections are present in the data, we run this algorithm blindly on the visibility auto-correlations for 25 iterations, restricting the delay range to between 100 and 1500 nanoseconds, which in practice suppresses the sub-reflections in the auto-correlations down to a dynamic range of $\sim 10^{-4}$ in visibility power. These reflection terms are collected into a single gain term for each antenna and the data are re-calibrated using standard antenna-based calibration. While this process opens a larger amount of spectral degrees of freedom into the calibration, reflection calibration is actually quite restricted in the amount of information it can remove in the cross-correlation visibilities, not only because they are direction independent antenna-based gains but because in practice they are solved for using only the auto-correlation visibilities, and cannot be overfit by the cross-correlations (Kern et al. 2019). The reflection parameters are solved on a per-integration basis on the LST binned data, and the resultant gains are smoothed on a 6-hour timescale. While the reflections are slowly variable in time, one could in principle opt to solve for them on the pre-LST binned data. However, we currently suspect that the limiting factor in reflection calibration is its frequency evolution, rather than the time evolution (Kern et al. 2020b).

After reflection calibration, the visibilities are passed through a filter to remove the cross-coupling (or crosstalk) common-mode systematic observed in the data (see subsection 4.3.2). For baselines that have a projected East-West separation of larger than 14 meters, this filter can subtract most of the cross-coupling systematics without attenuating the EoR signal (a signal loss tolerance of less than 1% in power).

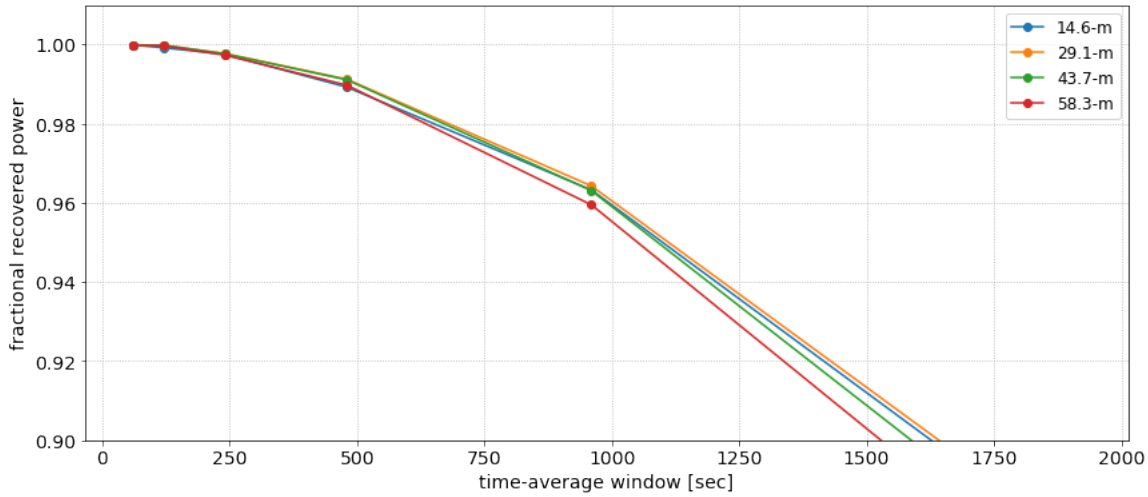

Figure 5.5: Simulated EoR power attenuation in HERA visibilities after coherent time integration for a variety of short and intermediate length baselines (14.6 – 58.3 meters). We see non-negligible signal loss (> 1%) for averaging windows of 500 seconds or longer, which informs our maximum coherent integration timescale.

5.1.5 Time Averaging

Being a transit telescope observing in drift-scan mode, HERA observes different patches of the sky at different time integrations throughout a nightly observation governed by the windowing of the sky temperature by the primary beam response. Integrations separated closely in time, however, are highly correlated, and with the appropriate phasing of the pointing center can be coherently summed without significant loss of sky signal. The time range over which data can be coherently summed is governed partly by the size of the primary beam on the sky (a wider primary beam means a patch of sky takes longer to transit the field-of-view), as well as the kind of emission one is sensitive to and the kind of re-phasing applied to the data. Ignoring the primary beam response, one can easily rephase the visibility response to a point source on the sky from one integration to the next. For diffuse emission, however, this breaks down, as the rephasing needed for one part of the sky is not identical to that needed by another part of the sky. Visibility rephasing is done by applying a phasor to the visibilities to move the pointing center to a new location on the sky (e.g. equation 21 of Zhang et al. (2018)).

To quantify how long we can integrate HERA visibilities without inducing non-negligible attenuation of a diffuse EoR-like signal, we run HERA simulations of an EoR sky with a realistic model of the HERA primary beam (see section 4.2 for details). Figure 5.5 shows the results of taking the averaged power spectrum of this EoR-only simulation having coherently time-averaged the visibilities on increasingly longer timescales. As expected, short timescales produce virtually negligible signal attenuation for all the baselines sampled. With integration windows longer than 500 seconds, we see the power spectrum attenuation of greater than 1%, which sets our threshold for the maximum coherent average timescale. Longer baselines begin to show larger attenuation

for wider time averaging windows due to the fact that the pointing center rephasing of the fringes against a diffuse sky becomes more approximate for longer baselines. We do not expect these results to be sensitive to the adopted EoR model, but to test this we repeat these simulations with a $P(k) \sim k^{-2}$ EoR model and find it to agree with the flat $P(k)$ results. After coherent time integration, we are left with ~ 150 independent drift-scan time bins.

5.2 Power Spectrum Analysis

In this section we give a brief overview of the delay spectrum estimator and present deep power spectra from the IDR2 dataset. Throughout this section, we adopt a ΛCDM cosmology (Planck Collaboration et al. 2016) with $\Omega_\Lambda = 0.6844$, $\Omega_b = 0.04911$, $\Omega_c = 0.26442$, and $H_0 = 67.27\,\mathrm{km/s/Mpc}$.

5.2.1 Estimating the Visibility-Based Delay Spectrum

The power spectrum encapsulates the variance of the sky power across spatial Fourier modes, and is a useful summary statistic for constraining astrophysical and cosmological models.. Ideally, one would measure the power spectrum from image cubes of the sky made from the data. However, one can also make an approximate estimate of the power spectrum directly from the visibilities, which is advantageous for a redundantly-spaced array like HERA. This is known as the delay approximation and is a valid for shorter baseline lengths, like the ones used in HERA (Parsons et al. 2012b; Liu et al. 2014a). We also used this estimator in section 3.5 and section 4.6.

To summarize, the power spectrum under the delay approximation (or the delay spectrum) is simply related to the square of the visibilities Fourier transformed across frequency,

$$P_{21}(k_\perp, k_\parallel) = \frac{X^2 Y}{\Omega_{pp} B} \langle \widetilde{V}(u,\tau) \widetilde{V}^*(u,\tau) \rangle, \tag{5.2}$$

where X and Y are scalars mapping sky angles and frequency to cosmological distances, Ω_{pp} is the integral of the squared primary beam response across the sky, B is the Fourier transform bandwidth, and $\langle \dots \rangle$ denote an ensemble average, with the power spectrum P_{21} having units $\mathrm{mK}^2\ h^{-3}\ \mathrm{Mpc}^3$. $\widetilde{V}$ is the Fourier-transformed visibilities, having optionally weighted the transform with a taper (or apodization) function to minimize Fourier space sidelobes. The spatial Fourier modes are related to angles on the sky and frequency separations as

$$k_\perp = \frac{2\pi\tau}{X} \tag{5.3}$$

$$k_\parallel = \frac{2\pi}{Y} \frac{b}{\lambda}, \tag{5.4}$$

where $X = c(1+z)^2 \nu_{21}^{-1} H(z)^{-1}$, $Y = D(z)$, $\nu_{21} = 1.420$ GHz, $H(z)$ is the Hubble parameter, and $D(z)$ is the transverse comoving distance (Parsons et al. 2014).

In this work we do not form the optimal quadratic estimate of the power spectrum (Tegmark 1997; Liu & Tegmark 2011), which is achieved by first weighting the data by the inverse of its covariance matrix. Recent works have highlighted certain subtleties of empirically estimated covariance matrices and their ability to induce signal loss if not treated properly. As a preliminary study, we therefore use a suboptimal uniform weighting of the data in our estimate; however, we do apply a Blackman-Harris tapering function (Blackman & Tukey 1958) across the bandwidth before taking the Fourier transform, which limits the Fourier sidelobe ringing of the foregrounds. We also do not perform any kind of foreground subtraction or removal (which can is in practice achieved via inverse covariance weighting), as we pursue a foreground avoidance approach to measuring the 21 cm power spectrum. To eliminate the noise bias when cross-multiplying the visibilities, we stagger the two visibilities such that we from the cross-multiplication of adjacent time bins in LST, which measure effectively the same sky signal but have different noise realizations. We do this for all baseline pairs within each redundant baseline type.

One can further analytically estimate the thermal noise floor of the power spectra (or the power spectrum amplitude in the limit of only thermal noise in the visibilities) as

$$P_N = \frac{X^2 Y \Omega_{\rm eff} T_{\rm sys}^2}{t_{\rm int} N_{\rm coherent} \sqrt{2 N_{\rm incoherent}}}, \tag{5.5}$$

where $t_{\rm int}$ is the integration time of the data, $N_{\rm coherent}$ is the number of coherent averages of the data (i.e. visibility averages) and $N_{\rm incoherent}$ is the number of incoherent averages (i.e. power spectrum averages). $\Omega_{\rm eff}$ is the effective beam area, defined as $\Omega_{\rm eff} = \Omega_p^2/\Omega_{pp}$, where Ω_{pp} is the sky integral of the squared primary beam (Pober et al. 2013b; Parsons et al. 2014; Cheng et al. 2018). Accurate estimates of the system temperature $T_{\rm sys}^{ij}$ for a baseline between antennas i and j can be made directly from the calibrated auto-correlation visibilities as

$$\sqrt{V_{ii} V_{jj}} = \frac{2 k_b \nu^2}{c^2 \Omega_p} T_{\rm sys}^{ij}. \tag{5.6}$$

We compute the thermal uncertainty for all cross-baseline power spectrum at each the remaining time bins using Equation 5.6 and Equation 5.5.

5.2.2 Redundant Averaging

HERA's highly redundant array layout yields a large number of baselines that measure the same cosmological modes on the sky, but have independent noise realizations. Forming cross-power spectra between all redundant baseline pairs (rather than just cross-multiplying a baseline with itself) yields an additional $1/\sqrt{N_{\rm baselines}}$ in noise suppression, and is what makes HERA a sensitive probe of 21 cm power spectrum. However, in reality, the baselines in a redundant group are not perfectly redundant, due to either position errors (i.e. the baseline vectors are not exactly identical), primary beam differences, or calibration uncertainties. Forming a cross-power spectrum between non-identical baseline can lead to slight signal loss of the EoR power on the sky, and should be quantified to build confidence in a power spectrum analysis. HERA's non-redundancy

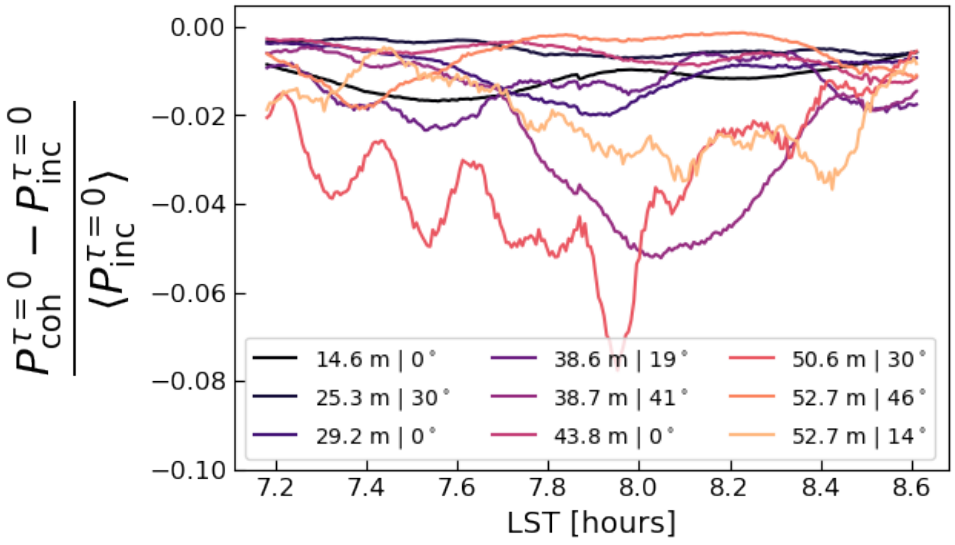

Figure 5.6: Redundancy decoherence test for 9 redundant groups, marked by the baseline length and angle in local array ENU coordinates. This plots the difference of the coherently and incoherently averaged power spectrum normalized by the time-average of the latter. We show the metric for the $\tau = 0$ Fourier mode at LSTs when bright, diffuse foregrounds fill the primary beam. On average, we see roughly 1-2% power loss, suggesting that while our final set of redundant visibilities are not perfect, they are redundant enough to retain the vast majority of sky power in the main lobe of the primary beam when forming baseline-to-baseline cross power spectra.

can be measured in a number of ways, for example, Dillon et al. (2020) showed that the reduced χ^2 after redundant calibration (which assumes perfect redundancy) is not consistent with thermal noise, indicating a level of non-redundancy in HERA baselines above the noise. However, this is not easily relatable to the signal loss induced when forming cross-baseline power spectra, as one would expect certain kinds of non-redundancies (like pure amplitude errors) to generate less severe signal loss when forming a cross-power spectrum than others. In particular, phase errors between baselines will clearly lead to signal loss due to the decoherence of the sky signal upon cross multiplication. We devise a metric to test for this in order to provide a rough quantification of the signal loss induced by cross multiplying baselines within a redundant set.

Given a single redundant baseline group consisting of M Fourier-transformed visibilities, $\widetilde{V}_1, \widetilde{V}_2, \ldots, \widetilde{V}_M$, we define an "incoherent average" of their power spectra as

$$P_{\rm inc} = \frac{1}{N}\left(|\widetilde{V}_1|^2 + |\widetilde{V}_2|^2 + |\widetilde{V}_3|^2 + \ldots\right) \tag{5.7}$$

where N is the number of antennas, and a "coherent average" as

$$P_{\rm coh} = \frac{1}{C}\left(\widetilde{V}_1\widetilde{V}_2^* + \widetilde{V}_1\widetilde{V}_3^* + \widetilde{V}_2\widetilde{V}_3^* + \ldots\right). \tag{5.8}$$

where C is the number of antenna-antenna pairs. If each visibility in the redundant set contains a random phase error from the true visibility, such that

$$\widetilde{V}_1 = e^{i\phi_1}\widetilde{V}_{\rm true};\ \widetilde{V}_2 = e^{i\phi_2}\widetilde{V}_{\rm true};\ \widetilde{V}_3 = e^{i\phi_3}\widetilde{V}_{\rm true};\ \ldots \tag{5.9}$$

then we can see that the incoherent average of three baselines is still an unbiased estimate of the true power spectrum

$$P_{\rm inc} = \frac{1}{3}\left(|\widetilde{V}_1|^2 + |\widetilde{V}_2|^2 + |\widetilde{V}_3|^2\right) = |\widetilde{V}_{\rm true}|^2. \tag{5.10}$$

Whereas for the coherent average, we see that it yields

$$P_{\rm coh} = \frac{1}{3}(\widetilde{V}_1\widetilde{V}_2^* + \widetilde{V}_1\widetilde{V}_3^* + \widetilde{V}_2\widetilde{V}_3^*) = |\widetilde{V}_{\rm true}|^2\frac{1}{3}(e^{i(\phi_1-\phi_2)} + e^{i(\phi_1-\phi_3)} + e^{i(\phi_2-\phi_3)}). \tag{5.11}$$

While the amplitude of the phase term is one, its real component is always less than or equal to one, meaning we can conclude that

$$\frac{{\rm Re}(P_{\rm coh})}{{\rm Re}(P_{\rm inc})} < 1 \tag{5.12}$$

assuming $\phi_1 \neq \phi_2 \neq \phi_3$, and thus we have induced some amount of signal loss in the power spectrum by taking a coherent average in the face of non-redundant phase errors. The ratio itself, though, is a metric for the amount of signal loss and can help us assess if it is negligible or not.

Using the data as a proxy of this metric can be complicated due to the fact that the power spectrum can go to zero when sky signals destructively interfere in the visibility, which can cause Equation 5.12 to spike spuriously. To mitigate this, we form the fractional ratio of the coherent and incoherent power spectrum with respect to the time-averaged incoherent power spectrum, or

$$R = \frac{P_{\rm coh} - P_{\rm inc}}{\langle P_{\rm inc}\rangle}, \tag{5.13}$$

where $\langle\rangle$ denotes a time-average, not an ensemble average. Of course, we cannot probe this metric for the actual EoR signal in the data because our data are not currently sensitive enough to detect it and it is not yet clear how to disentangle it from the foreground signal, but we can probe this metric for the foreground signal with a high SNR. The foreground is also a temperature field that spans the sky like the EoR, so it stands to reason that non-redundancy of the observed foregrounds may give us a handle on the non-redundancy of the EoR signal in the data. The complication here is that this is not true at all delay modes in the visibilities: smooth spectrum foregrounds are boosted to higher delays when observed at larger zenith angles (Parsons et al. 2012a), which is actually just a re-statement of the foreground wedge phenomenon. The vast majority of the EoR signal enters through the main-lobe of the primary beam, so when determining foreground signal loss due to non-redundancy, we should attempt to localize its effect to main-lobe non-redundancy. We can do this with the foregrounds by looking at the $\tau = 0$ ns delay bin, which is the Fourier mode where smooth spectrum foregrounds in the main-lobe are mapped to in the visibilities. Figure 5.6 shows

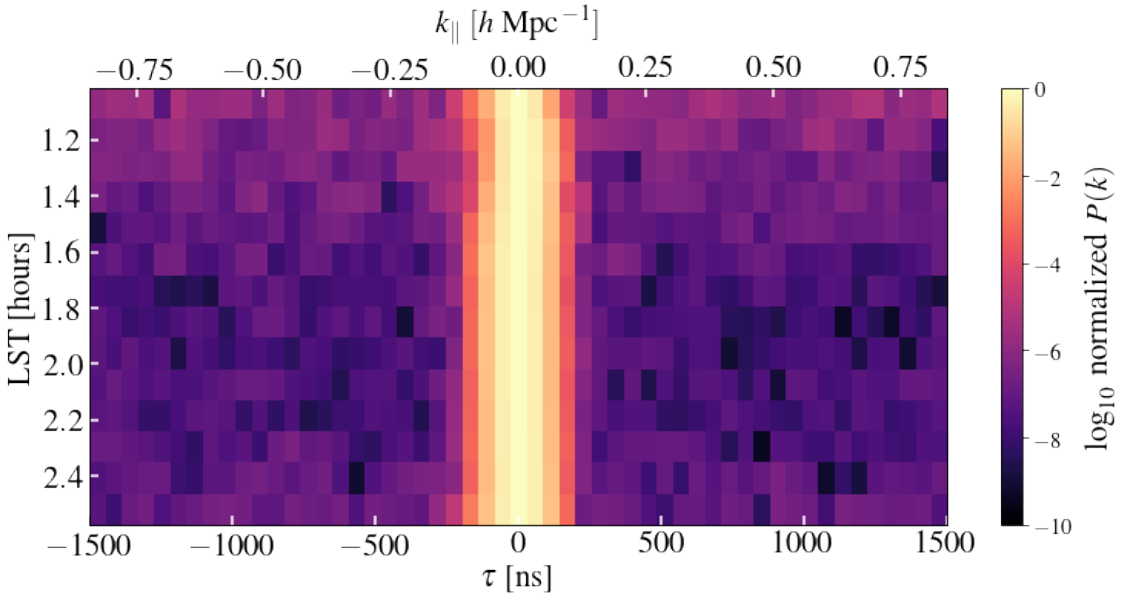

Figure 5.7: A drift-scan power spectrum after redundant averaging for a single HERA baseline group. The colorscale is normalized at each time integration to the peak foreground power at $\tau = 0$ ns. Containment of the foreground power to low delays is stable across the LST range.

this metric for the most sensitive redundant baseline groups in the HERA Phase I array. We choose a range of LSTs where the FoV is filled with diffuse emission (from the galactic anti-center), which is the most representative of a diffuse EoR signal Each redundant group is separated based on the baseline length and its orientation in the local East-North-Up (ENU) array coordinates. While some groups show stronger signal loss than others at specific times (possibly due to point sources transiting primary beam sidelobes), we see that on average, the cross-multiplication of power spectra within redundant groups leads to percent-level signal loss for the Phase I array, which is well within the limiting error budget set by the absolute flux calibration of $\sim 10\%$.

5.2.3 Averaged Power Spectra

In this section we explore the power spectrum of deep HERA integrations, quantifying the dynamic range achieved between the peak foreground power and the noise / systematic floor of the power spectrum. First, we form a pseudo-Stokes I visibility as the average of the XX and YY dipole polarization visibilities, or

$$V_I = (V_{XX} + V_{YY})/2, \tag{5.14}$$

conforming to recent convention in the literature (Smirnov 2011). In the limit that there are no direction-dependent effects leftover after calibration, the pseudo-Stokes I visibility is equivalent to the spatial Fourier transform of the true Stokes I parameter on the sky. In the main-lobe of

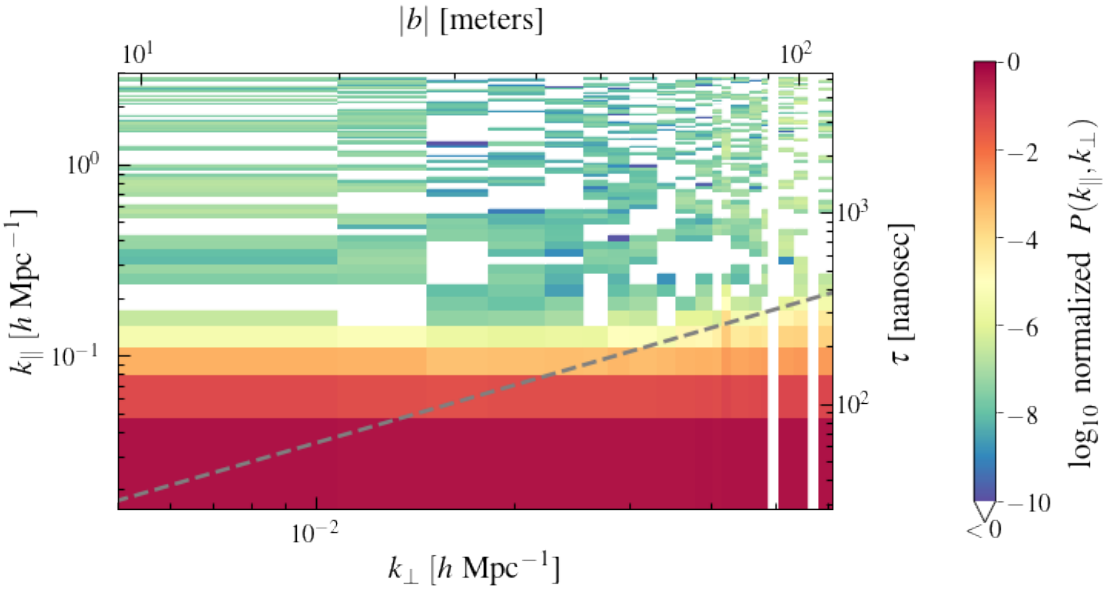

Figure 5.8: A cylindrically averaged power spectrum, normalized to $k_\parallel = 0$ h Mpc^{-1} at each $k_\perp$ bin. Minimal foreground leakage is observed outside of the horizon delay (grey dashed) except for on the shortest few baselines, where a foreground floor is reached. This could be due in part to residual calibration uncertainties, low-level RFI, or residual cable reflection and/or cross-coupling systematics. Outside of the foreground dominated regions the data show noise-like behavior, oscillating between positive and negative power (negative power indicated by white pixels).

the primary beam, where we are most sensitive to the EoR signal, direction-dependent effects are minimal, and thus power spectra formed from pseudo-Stokes visibilities are a good measure of the Stokes I power spectrum.

Power spectra are then formed across a 20 MHz bandwidth in a spectral window that has the least amount of RFI and flag occupancy, centered at 158.6 MHz (z = 8.0). However, having applied a Blackman-Harris tapering function, this translates to an effective bandwidth of $\sim$ 10 MHz, which has a cosmological line-of-sight evolution of $\Delta z \sim 0.56$. We cross-multiply all baselines within a redundant set and drop all baselines cross multiplied with themselves, as antenna and baseline-based systematics are generally more prominent in the latter. We focus on a section of the data that coincides with the coldest region of foregrounds that transited HERA's FoV during the observing season. As shown in Figure 5.1, this corresponds to an LST (or, equivalently, right ascension) range of $1 - 2.5$ hours, which leaves out data contaminated by the rise of the bright radio galaxy Fornax A in the beam and its transit of the FoV at a right ascension of 3.3 hours.

Inspecting the power spectra as a function of the remaining time bins (Figure 5.7) shows containment of foregrounds power and leakage to low delays, which is normalized by the peak foreground power at $\tau = 0$ ns at each time. This suggests that noise-limited power spectrum measurements may be possible for $k \sim 0.2$ h Mpc^{-1} modes. The evolution of the background in Figure 5.7 is reflective of the evolving signal-to-noise ratio of the foreground power across the LST

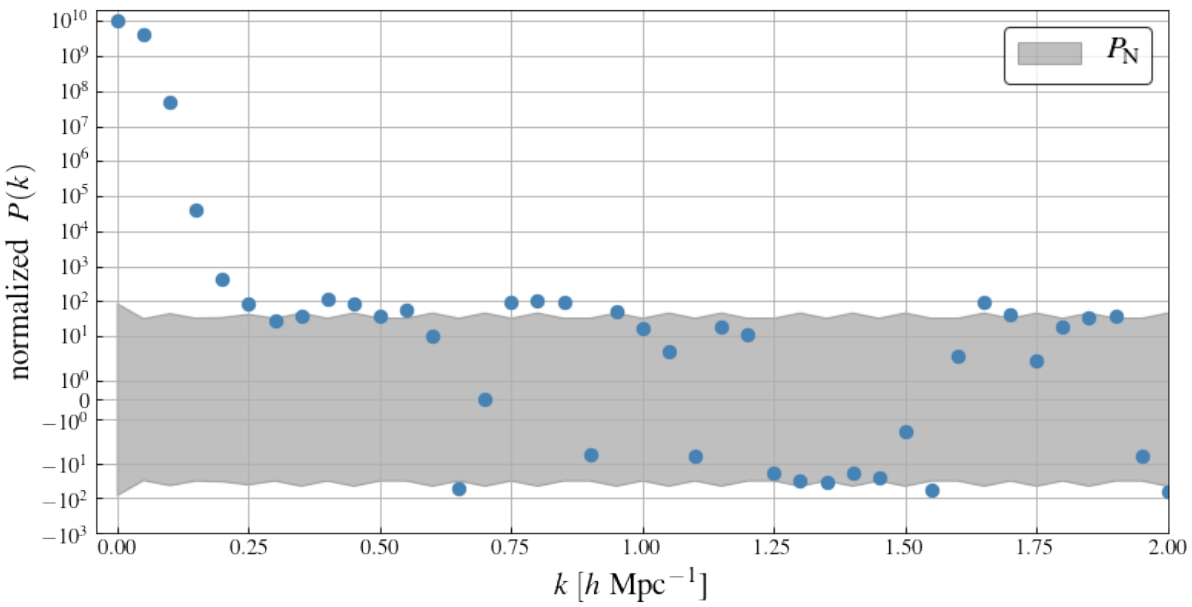

Figure 5.9: The spherically averaged power spectrum, normalized to 10^{10} at $k = 0$ h Mpc^{-1} to demonstrate the $> 10^8$ in dynamic range achieved for $k \geq 0.25$ h Mpc^{-1} modes with respect to the peak foreground power. The grey shaded region denotes the 1σ thermal noise floor P_N value. Bandpowers that are noise-dominated should fluctuate positive and negative, which is seen at intermediate and high k. Low-level systematics for $0.25 < k < 0.5$ h Mpc^{-1} may be marginally detected. Due to the Blackman-Harris tapering, we expect neighboring points in k to be fairly correlated.

range.

Next we show a cylindrically averaged power spectrum, or a power spectrum in the $k_\parallel$ and $k_\perp$ plane, having averaged the transverse k_x and k_y plane in annuli of constant $k_\perp$. In this space, the extent of the foreground emission is confined to low $k_\parallel$, with a maximum reach that increases with increasing $k_\perp$, producing the foreground wedge phenomenon (Datta et al. 2010; Morales et al. 2012; Parsons et al. 2012b; Liu et al. 2014a). Depending on the instrumental design, an individual experiment can fill the wedge in different ways, which is largely driven by the field of view of the instrument and the feed design (Thyagarajan et al. 2015). In Figure 3.17, we showed what a foreground wedge looks like for HERA Phase I, demonstrating slight amounts of foreground leakage on short baselines, and nominal foreground containment for larger baselines. In Figure 5.8, we show a more deeply integrated power spectrum that largely agrees with this picture. The observed foregrounds are well isolated to within the horizon delay (grey dashed) on long baselines, and on short baselines we see evidence for foreground leakage out to $k_\parallel \sim 0.2$ h Mpc^{-1}. Part of this is due to the intrinsic sidelobes of the Blackman-Harris tapering function pushing $k_\parallel = 0$ foreground power to higher $|k_\parallel|$, but given the quick suppression in sidelobe power from such a window, it does not seem to explain all of the observed power out to $k_\parallel \sim 0.2$ h Mpc^{-1}. The colorscale is normalized to the peak foreground emission at $k_\parallel = 0$ h Mpc^{-1} at each $k_\perp$ bin, demonstrating

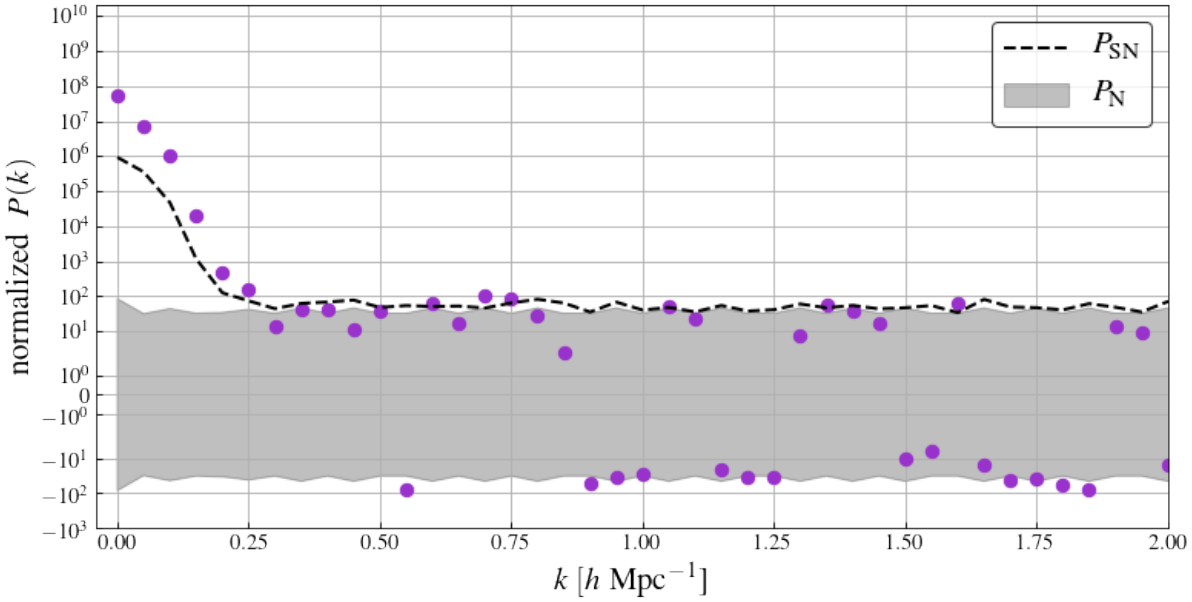

Figure 5.10: The imaginary component of the spherically averaged power spectrum from Figure 5.9. The grey shaded region shows the estimated P_N, and the black dashed line shows the estimated P_{SN}, the latter of which encapsulates additional noise in signal dominated bandpowers. The imaginary component of the power spectrum should be consistent with thermal noise at the same delay modes as the real component if foregrounds and systematics have been dealt with appropriately. For $k \geq 0.25\ h$ Mpc^{-1}, the imaginary component is consistent with the real component and the thermal noise floor estimate, implying that these k modes are consistent with the null hypothesis for this particular test. At smaller spatial wavevectors, the rise in the imaginary component begins to exceed that of P_{SN} suggesting an additional source of variance not described by thermal noise.

nearly 10^8 in dynamic range outside of the foreground wedge. Outside of the foreground dominated regions, the data seem consistent with noise, suggested by the oscillation of the power spectrum from positive to negative power (negative power indicated by white pixels).

To show this more concretely, we spherically average the power spectrum on a $|k|$ grid. We propagate the thermal noise uncertainty P_N throughout all of the averaging, using it to perform a weighted average at each stage. The result is shown in Figure 5.9, having normalized the $k = 0$ mode to 10^{10}. The shaded region marks the propagated thermal noise floor, demonstrating agreement with the data for $k > 0.25\ h$ Mpc^{-1}. Agreement with the noise is further evidenced by the fluctuation between positive and negative power, which is expected for noise-dominated data. Note that because of the Blackman-Harris tapering, we expect neighboring points in k to be fairly correlated. For $0.25 < k < 0.5\ h$ Mpc^{-1}, we may be marginally detected residual systematics. This demonstrates the dynamic range achieved throughout the analysis: residual spectral structure after in data calibration systematic modeling, and other analyses is kept below 10^8 in power. Generally, fiducial EoR levels are expected to lie at roughly $\sim 10^{10}$ in dynamic range against

the peak foreground power.[2] Recall this analysis utilized 18-days of HERA data in a 40-element configuration, spanning only a few hours in local sidereal time from each night. Provided that low-level systematics can continued to be mitigated, this result bodes well for future full-season (180 day) observations with HERA in its full 350-element configuration.

A Simple Null Test: The Imaginary Component of the Power Spectrum

Redundant baselines measure the same sky signal, and thus have visibilities that are coherent with each other. When taking their cross-multiplication to from a power spectrum (Equation 5.2), this projects the sky power into a purely real quantity. Thermal noise due to the sky temperature and the ambient temperature of the environment and instrument, however, is uncorrelated between baselines, and its effect is to generate random, mean-zero fluctuations in the real and imaginary components of the power spectrum with a standard deviation given by Equation 5.5. For bandpowers that detect a signal with SNR > 1, the noise fluctuations are amplified by the signal-noise cross terms, and thus the noise amplitude will exceed the $P_{\rm N}$ estimate. This is derived in Kolopanis et al. (2019) as $P_{SN} = \sqrt{2P_S P_N + P_N^2}$, where P_S is an estimate of the signal dominated bandpowers, which we approximate with the estimated bandpowers. Non-ideal effects like inherent baseline non-redundancies and residual gain calibration errors can cause power to be leaked from the real component into the imaginary component. This makes the imaginary component of the power spectrum a useful null test, to provide another axis by which to assess if a particular bandpower is consistent with our expectation of thermal noise.

Figure 5.10 shows the imaginary component of our fully averaged power spectrum, normalized in the same manner as before such that $\mathrm{Re}[P(k = 0)] = 10^{10}$. Similar to the real component, we see good agreement with the thermal noise floor for $k \geq 0.25\ h\ \mathrm{Mpc}^{-1}$, implying that these k modes are consistent with the null hypothesis for this particular test. For $k < 0.25\ h\ \mathrm{Mpc}^{-1}$, the imaginary component rises at a rate slightly exceeding the P_{SN} estimate, suggesting a low-level component of additional variance other than thermal noise. This is likely due to percent-level gain calibration errors that leak power from the real component into the imaginary component of the power spectrum at the same k mode, which is only visible at low k where the power in the real component is large due to astrophysical foreground emission.

[2] While this statement depends on the experimental design and the part of the sky one is pointing at, this is a rough estimate for HERA (Thyagarajan et al. 2016).

Chapter 6

Conclusion

This thesis has focused on the development of algorithms and analysis pipelines for precision data analysis of next-generation 21 cm intensity mapping surveys. Throughout this work, we have applied these techniques to early observations from the Hydrogen Epoch of Reionization Array (HERA), an up-coming 21 cm experiment that will probe the Epoch of Reionization (EoR) with unprecedented sensitivity when completed.

In Chapter 2, we addressed the problem of cosmological parameter inference, which is particularly difficult for EoR 21 cm cosmology, as the 21 cm parameter space is weakly constrained, and the 21 cm signal is also difficult to model accurately from first principles. This latter has led to the development of expensive numerical simulations to capture the interplay of detailed physics spanning many orders of magnitude in size scales; however, actually constraining a large and weakly constrained parameter space is impractical with such expensive simulations. An alternative method is to construct a surrogate model that approximates the simulation output across its parameter space, thereby accelerating parameter inference. Such an approach, known as computer emulation, has become more widely adopted within astrophysical parameter inference problems. In this work, we presented the first application of this techniques for 21 cm cosmology, demonstrating a comprehensive parameter constraint forecast for the full HERA experiment. This result was the first joint parameter forecast across cosmological, IGM photo-ionizaton, and IGM heating parameters for a Cosmic Dawn 21 cm experiment (Kern et al. 2017).

Standing in the way of the scientific potential of future 21 cm experiments is the challenge of precise, high-dynamic range signal separation: we are faced with disentangling a faint EoR signal from foreground emission (from the Milky Way and tens of thousands of extragalactic radio point sources) that is nearly 10^5 times brighter. This sets a tight requirement on the fidelity of data reduction procedures, and necessitates a careful accounting of instrumental and environmental systematics. In Chapter 3, we outlined and demonstrated a sky-based calibration pipeline for HERA using standard calibration software in combination with custom, HERA-specific reduction software (Kern et al. 2020a). A host of recent works have shown that sky-based calibration can introduce unwanted spectral structure into the data through the inevitable exclusion of faint point sources in the flux density model. We show in Kern et al. (2020a) that a related effect also occurs from the presence of instrumental coupling systematics in the data, which can be spread to other

parts of the data when performing gain calibration. We demonstrate a straightforward filtering method for mitigating the impact of these effects at intermediate and high delays, which allows for considerably deeper power spectrum sensitivity for a foreground avoidance estimator. We also used this sky-calibration pipeline to demonstrate how to combine it with redundant baseline calibration, with the former constraining the degenerate degrees of freedom leftover after redundant calibration. For HERA, we found a marginal improvement in the gain solutions with redundant calibration; however, we also find that it brings its own set of uncertainties, which we speculated originates from the non-redundancy of the per-antenna beams, particularly near the horizon where diffuse emission is accumulated more strongly than point sources.

In Chapter 4, we investigate the origin of various instrumental systematics seen in HERA data. We break these down into two types: signal chain reflections and antenna-to-antenna cross coupling. Building upon previous works, we developed a straightforward analytic model for these systematics, and used numerical simulation to understand their phenomenology in the data. We constructed models for the systematics and used ensemble sets of numerical simulations to probe their performance and understand their signal loss properties. Applying these techniques to HERA data, we find they can suppress the leading-order systematics by over two orders of magnitude in the power spectrum. However, we also find evidence for low-level and highly complex set of sub-reflection terms in the auto-correlation visibilities will likely be difficult to remove, due to their strong frequency dependence. While faint, these terms are still above fiducial EoR signal amplitudes, and will require more careful treatment if HERA Phase I observations are to ever make a detection of the 21 cm power spectrum.

Finally, in Chapter 5, we discuss the Phase I data reduction pipeline, including steps for faulty antenna identification, radio frequency interference (RFI) excision, and data averaging. We use additional numerical simulations to support our analysis choices, particularly the ones behind coherent LST averaging and redundant baseline averaging, both of which can induce unwanted signal loss of the cosmological signal if not handled properly. We furthermore expand our systematic subtraction to subtract some of the additional low-level systematics described before. Using 18 nights of Phase I data, we demonstrate that our analysis pipeline enables over 8 orders of magnitude in dynamic range between the peak foreground power at $k = 0$ and the noise floor of the data at $k = 0.25\ h\ \mathrm{Mpc}^{-3}$. Given that fiducial EoR amplitudes are expected to lie at roughly 10^{10} in dynamic range for HERA (Thyagarajan et al. 2016), this analysis bodes well for the possible future detection of the EoR with HERA. However, any putative detection will require additional end-to-end pipeline validation with mock EoR trials, in addition to more comprehensive jacknife and null testing. As an example of a simple null test, we show that the imaginary component of the power spectrum, which in theory should contain only noise, is in good agreement with the real component of the power spectrum where it meets the noise floor at $k \sim 0.25\ h\ \mathrm{Mpc}^{-1}$.

As a field, 21 cm intensity mapping experiments are increasingly moving towards more sophisticated techniques for data analysis and systematic subtraction. In this thesis, we have developed, applied, and validated a series of techniques for calibration, systematic modeling, and parameter inference. These techniques are enabling HERA to make more sensitive limits on the 21 cm power spectrum, and may eventually enable a first-detection of the 21 cm power spectrum and a constraint on the photo-ionization and heating of the IGM during Cosmic Dawn.

Bibliography

Addison, G., Huang, Y., Watts, D., et al. 2016, ApJ, 818, 132
Ali, Z., Parsons, A., Zheng, H., et al. 2015, ApJ, 809, 61
Aslanyan, G., Easther, R., & Price, L. 2015, J. Cosmology Astropart. Phys., 9, 5
Barkana, R., & Loeb, A. 2002, ApJ, 578, 1
—. 2005, ApJ, 626, 1
Barkana, R., Outmezguine, N. J., Redigol, D., & Volansky, T. 2018, Phys. Rev. D, 98, 103005
Barry, N., Beardsley, A. P., Byrne, R., et al. 2019a, PASA, 36, e026
Barry, N., Hazelton, B., Sullivan, I., Morales, M., & Pober, J. 2016, MNRAS, 461, 3135
Barry, N., Wilensky, M., Trott, C. M., et al. 2019b, ApJ, 884, 1
Beardsley, A., Hazelton, B., Sullivan, I., et al. 2016, ApJ, 833, 102
Beers, T., Flynn, K., & Gebhardt, K. 1990, AJ, 100, 32
Bernardi, G., Greenhill, L. J., Mitchell, D. A., et al. 2013, ApJ, 771, 105
Bernardi, G., Zwart, J. T. L., Price, D., et al. 2016, MNRAS, 461, 2847
Betancourt, M. 2017, arXiv e-prints, arXiv:1701.02434
Bhatnagar, S., Cornwell, T. J., Golap, K., & Uson, J. M. 2008, A&A, 487, 419
Blackman, R. B., & Tukey, J. W. 1958, Bell System Technical Journal, 37, 185
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. 2016, arXiv e-prints, arXiv:1601.00670
Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J., & Mahesh, N. 2018, Nature, 555, 67
Bowman, J. D., Cairns, I., Kaplan, D. L., et al. 2013, PASA, 30, e031
Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. 2011, Handbook of Markov Chain Monte Carlo (CRC press)
Bucher, M., & Louis, T. 2012, MNRAS, 424, 1694
Byrne, R., Morales, M. F., Hazelton, B., et al. 2019, ApJ, 875, 70
Carilli, C. L., Nikolic, B., Thyagarayan, N., & Gale-Sides, K. 2018, Radio Science, 53, 845
Chaudhari, S. C., Gupta, Y., Kumar, A., et al. 2017, Journal of Astronomical Instrumentation, 6, 1641017
Cheng, C., Parsons, A. R., Kolopanis, M., et al. 2018, ApJ, 868, 26
Cohen, A., Fialkov, A., Barkana, R., & Monsalve, R. 2019, arXiv e-prints, arXiv:1910.06274
Condon, J. J. 1997, PASP, 109, 166
Condon, J. J., Cotton, W. D., Greisen, E. W., et al. 1998, AJ, 115, 1693

Cornwell, T. J., Golap, K., & Bhatnagar, S. 2008, IEEE Journal of Selected Topics in Signal Processing, 2, 647
Cui, G., Yu, X., Iommelli, S., & Kong, L. 2016, IEEE Signal Processing Letters, 23, 1662
Das, A., Mesinger, A., Pallottini, A., Ferrara, A., & Wise, J. 2017, ArXiv e-prints, arXiv:1702.00409
Datta, A., Bowman, J., & Carilli, C. 2010, ApJ, 724, 526
de Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., et al. 2008, MNRAS, 388, 247
DeBoer, D., Parsons, A., Aguirre, J., et al. 2017, PASP, 129, 45001
Dicke, R. H., Peebles, P. J. E., Roll, P. G., & Wilkinson, D. T. 1965, ApJ, 142, 414
Dillon, J. 2017, H1C Internal Data Release 2.2, Tech. rep., Department of Astronomy, University of California, Berkeley, CA, HERA Memo #69
Dillon, J., & Parsons, A. 2016, ApJ, 826, 181
Dillon, J., Liu, A., Williams, C., et al. 2014, Phys. Rev. D, 89, 23002
Dillon, J., Neben, A., Hewitt, J., et al. 2015, Phys. Rev. D, 91, 123011
Dillon, J. S., Kohn, S. A., Parsons, A. R., et al. 2018, MNRAS, 477, 5670
Dillon, J. S., Lee, M., Ali, Z. S., et al. 2020, arXiv e-prints, arXiv:2003.08399
Eastwood, M. W., Anderson, M. M., Monroe, R. M., et al. 2018, AJ, 156, 32
—. 2019, AJ, 158, 84
Efron, B., & Tibshirani, R. 1994, An Introduction to the Bootstrap, Chapman & Hall/CRC Monographs on Statistics & Applied Probability (Taylor & Francis)
Ewall-Wice, A., Dillon, J. S., Liu, A., & Hewitt, J. 2017, MNRAS, 470, 1849
Ewall-Wice, A., Hewitt, J., Mesinger, A., et al. 2016a, MNRAS, 458, 2710
Ewall-Wice, A., Dillon, J., Hewitt, J., et al. 2016b, MNRAS, 460, 4320
Ewall-Wice, A., Bradley, R., Deboer, D., et al. 2016, ApJ, 831, 196
Ewall-Wice, A., Kern, N., Dillon, J. S., et al. 2020, arXiv e-prints, arXiv:2004.11397
Fagnoni, N., de Lera Acedo, E., DeBoer, D. R., et al. 2019, arXiv e-prints, arXiv:1908.02383
Fan, X., Strauss, M., Becker, R., et al. 2006, AJ, 132, 117
Fendt, W., & Wandelt, B. 2007, ApJ, 654, 2
Fialkov, A., Barkana, R., & Visbal, E. 2014, Nature, 506, 197
Field, G. 1958, Proc. IRE, 46, 240
Field, S., Galley, C., Hesthaven, J., Kaye, J., & Tiglio, M. 2014, Phys. Rev. X, 4, 31006
Furlanetto, S., Oh, S., & Briggs, F. 2006, Phys. Rep., 433, 181
Furlanetto, S., Zaldarriaga, M., & Hernquist, L. 2004, ApJ, 613, 1
Gehlot, B. K., Koopmans, L. V. E., de Bruyn, A. G., et al. 2018, MNRAS, 478, 1484
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Gosh, A., Mertens, F., & the HERA Collaboration. in prep.
Gramacy, R. B., & Lee, H. K. H. 2009, Technometrics, 51, 130
Greig, B., & Mesinger, A. 2015, MNRAS, 449, 4246
—. 2017a, eprint arXiv:1705.03471, arXiv:1705.03471
—. 2017b, MNRAS, 465, 4838
Greig, B., Mesinger, A., & Pober, J. 2016, MNRAS, 455, 4295
Gunn, J. E., & Peterson, B. A. 1965, ApJ, 142, 1633
Habib, S., Heitmann, K., Higdon, D., Nakhleh, C., & Williams, B. 2007, Phys. Rev. D, 76, 83503

Haiman, Z., Abel, T., & Rees, M. 2000, ApJ, 534, 11
Haiman, Z., Rees, M., & Loeb, A. 1997, ApJ, 476, 458
Hamaker, J. P., Bregman, J. D., & Sault, R. J. 1996, A&AS, 117, 137
Haynes, M. P., Giovanelli, R., Kent, B. R., et al. 2018, ApJ, 861, 49
Hazelton, B. J., Jacobs, D. C., Pober, J. C., & Beardsley, A. P. 2017, The Journal of Open Source Software, 2, 140
Heitmann, K., Higdon, D., Nakhleh, C., & Habib, S. 2006, ApJ, 646, L1
Heitmann, K., Higdon, D., White, M., et al. 2009, ApJ, 705, 156
Higdon, D., Nakhleh, C., Gattiker, J., & Williams, B. 2008, Comput. Methods Appl. Mech. Eng., 197, 2431
Högbom, J. A. 1974, A&AS, 15, 417
Hurley-Walker, N., Callingham, J. R., Hancock, P. J., et al. 2017, MNRAS, 464, 1146
Iliev, I., Mellema, G., Ahn, K., et al. 2014, MNRAS, 439, 725
Intema, H. T. 2014, in Astronomical Society of India Conference Series, Vol. 13, Astronomical Society of India Conference Series, 469
Intema, H. T., Jagannathan, P., Mooley, K. P., & Frail, D. A. 2017, A&A, 598, A78
Jacobs, D., Pober, J., Parsons, A., et al. 2015, ApJ, 801, 51
Jacobs, D. C., Parsons, A. R., Aguirre, J. E., et al. 2013, ApJ, 776, 108
Jacobs, D. C., Hazelton, B. J., Trott, C. M., et al. 2016, ApJ, 825, 114
Jennings, W. D., Watkinson, C. A., Abdalla, F. B., & McEwen, J. D. 2019, MNRAS, 483, 2907
Joseph, R. C., Trott, C. M., & Wayth, R. B. 2018, AJ, 156, 285
Kennedy, M. C., & O'Hagan, A. 2001, J. R. Stat. Soc. Ser. B (Statistical Methodol., 63, 425
Kern, N. S., Liu, A., Parsons, A. R., Mesinger, A., & Greig, B. 2017, ApJ, 848, 23
Kern, N. S., Parsons, A. R., Dillon, J. S., et al. 2019, ApJ, 884, 105
Kern, N. S., Dillon, J. S., Parsons, A. R., et al. 2020a, ApJ, 890, 122
Kern, N. S., Parsons, A. R., Dillon, J. S., et al. 2020b, ApJ, 888, 70
Kerrigan, J., La Plante, P., Kohn, S., et al. 2019, MNRAS, 488, 2605
Kohn, S., Aguirre, J., Nunhokee, C., et al. 2016, ApJ, 823, 88
Kohn, S. A., Aguirre, J. E., La Plante, P., et al. 2019, ApJ, 882, 58
Kolopanis, M., Jacobs, D. C., Cheng, C., et al. 2019, ApJ, 883, 133
Komatsu, E., Smith, K. M., Dunkley, J., et al. 2011, ApJS, 192, 18
Kuhlen, M., & Faucher-Giguère, C.-A. 2012, MNRAS, 423, 862
Lane, W. M., Cotton, W. D., van Velzen, S., et al. 2014, MNRAS, 440, 327
Lanman, A. E., & Pober, J. C. 2019, MNRAS, 487, 5840
Lanman, A. E., Pober, J. C., Kern, N. S., et al. 2020, MNRAS, 494, 3712
Large, M. I., Mills, B. Y., Little, A. G., Crawford, D. F., & Sutton, J. M. 1981, MNRAS, 194, 693
Lenc, E., Anderson, C. S., Barry, N., et al. 2017, PASA, 34, e040
Lewis, A., & Bridle, S. 2002, Phys. Rev. D, D66, 103511
Lewis, A., Challinor, A., & Lasenby, A. 2000, Astrophys. J., 538, 473
Li, W., Pober, J. C., Hazelton, B. J., et al. 2018, ApJ, 863, 170
Li, W., Pober, J. C., Barry, N., et al. 2019, arXiv e-prints, arXiv:1911.10216
Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. 2016, Systematic Biology, 66,

e66
Liu, A., Parsons, A., & Trott, C. 2014a, Phys. Rev. D, 90, 23018
—. 2014b, Phys. Rev. D, 90, 23019
Liu, A., Pritchard, J., Allison, R., et al. 2016, Phys. Rev. D, 93, 43013
Liu, A., & Shaw, J. R. 2019, arXiv e-prints, arXiv:1907.08211
Liu, A., & Tegmark, M. 2011, Phys. Rev. D, 83, 103006
Liu, A., Tegmark, M., Morrison, S., Lutomirski, A., & Zaldarriaga, M. 2010, MNRAS, 408, 1029
Mao, Y., Tegmark, M., McQuinn, M., Zaldarriaga, M., & Zahn, O. 2008, Phys. Rev. D, 78, 23529
Martinot, Z. E., Aguirre, J. E., Kohn, S. A., & Washington, I. Q. 2018, ApJ, 869, 79
McGreer, I., Mesinger, A., & D'Odorico, V. 2015, MNRAS, 447, 499
McKay, M. D., Beckman, R. J., & Conover, W. J. 1979, Technometrics, 21, 239
McKinley, B., Yang, R., López-Caniego, M., et al. 2015, MNRAS, 446, 3478
McMullin, J. P., Waters, B., Schiebel, D., Young, W., & Golap, K. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 376, Astronomical Data Analysis Software and Systems XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell, 127
McQuinn, M., Zahn, O., Zaldarriaga, M., Hernquist, L., & Furlanetto, S. 2006, ApJ, 653, 815
Mellema, G., Iliev, I., Pen, U.-L., & Shapiro, P. 2006, MNRAS, 372, 679
Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, MNRAS, 493, 1662
Mesinger, A., Ferrara, A., & Spiegel, D. 2013, MNRAS, 431, 621
Mesinger, A., & Furlanetto, S. 2007, ApJ, 669, 663
Mesinger, A., Furlanetto, S., & Cen, R. 2011, MNRAS, 411, 955
Mineo, S., Gilfanov, M., & Sunyaev, R. 2012, MNRAS, 419, 2095
Mondal, R., Fialkov, A., Fling, C., et al. 2020, arXiv e-prints, arXiv:2004.00678
Moore, D. F., Aguirre, J. E., Kohn, S. A., et al. 2017, ApJ, 836, 154
Morales, M., Hazelton, B., Sullivan, I., & Beardsley, A. 2012, ApJ, 752, 137
Morales, M., & Wyithe, J. 2010, ARA&A, 48, 127
Morales, M. F., Beardsley, A., Pober, J., et al. 2019, MNRAS, 483, 2207
Mouri Sardarabadi, A., & Koopmans, L. V. E. 2019, MNRAS, 483, 5480
Muñoz, J. B., Dvorkin, C., & Loeb, A. 2018, Phys. Rev. Lett., 121, 121301
Neben, A. R., Bradley, R. F., Hewitt, J. N., et al. 2016, ApJ, 826, 199
Nunhokee, C. D., Bernardi, G., Kohn, S. A., et al. 2017, ApJ, 848, 47
Nunhokee, C. D., Parsons, A. R., Kern, N. S., et al. 2020, ApJ, 897, 5
Offringa, A. R., Mertens, F., & Koopmans, L. V. E. 2019, MNRAS, 484, 2866
Orosz, N., Dillon, J. S., Ewall-Wice, A., Parsons, A. R., & Thyagarajan, N. 2019, MNRAS, 487, 537
Paardekooper, J.-P., Khochfar, S., & Dalla Vecchia, C. 2015, MNRAS, 451, 2544
Paciga, G., Albert, J., Bandura, K., et al. 2013, MNRAS, 433, 639
Pacucci, F., Mesinger, A., Mineo, S., & Ferrara, A. 2014, MNRAS, 443, 678
Parsons, A., Pober, J., Aguirre, J., et al. 2012a, ApJ, 756, 165
Parsons, A., Pober, J., McQuinn, M., Jacobs, D., & Aguirre, J. 2012b, ApJ, 753, 81
Parsons, A., Liu, A., Aguirre, J., et al. 2014, ApJ, 788, 106
Parsons, A. R., & Backer, D. C. 2009, AJ, 138, 219

Parsons, A. R., Liu, A., Ali, Z. S., & Cheng, C. 2016, ApJ, 820, 51
Parsons, A. R., Backer, D. C., Foster, G. S., et al. 2010, AJ, 139, 1468
Patil, A., Yatawatta, S., Koopmans, L., et al. 2017, ApJ, 838, 65
Patra, N., Parsons, A. R., DeBoer, D. R., et al. 2018, Experimental Astronomy, 45, 177
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2012, ArXiv e-prints, arXiv:1201.0490 [cs.LG]
Pen, U.-L., Chang, T.-C., Hirata, C. M., et al. 2009, MNRAS, 399, 181
Penzias, A. A., & Wilson, R. W. 1965, ApJ, 142, 419
Petri, A., Liu, J., Haiman, Z., et al. 2015, Phys. Rev. D, 91, 103511
Planck Collaboration, Ade, P., Aghanim, N., et al. 2016, A&A, 594, A13
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2018, arXiv e-prints, arXiv:1807.06209
Pober, J., Parsons, A., Aguirre, J., et al. 2013a, ApJ, 768, L36
Pober, J., Parsons, A., DeBoer, D., et al. 2013b, AJ, 145, 65
Pober, J., Liu, A., Dillon, J., et al. 2014, ApJ, 782, 66
Pober, J. C., Parsons, A. R., Jacobs, D. C., et al. 2012, AJ, 143, 53
Pritchard, J., & Loeb, A. 2012, Reports Prog. Phys., 75, 86901
Rasmussen, C. E., & Williams, C. K. I. 2006, Gaussian processes for machine learning (MIT Press), 248
Robertson, B., Ellis, R., Furlanetto, S., & Dunlop, J. 2015, ApJ, 802, L19
Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. 1989, Statist. Sci., 4, 409
Santos, M., Ferramacho, L., Silva, M., Amblard, A., & Cooray, A. 2010, MNRAS, 406, 2421
Sault, R. J., Hamaker, J. P., & Bregman, J. D. 1996, A&AS, 117, 149
Schmit, C. J., & Pritchard, J. R. 2018, MNRAS, 475, 1213
Schneider, M., Holm, Ó., & Knox, L. 2011, ApJ, 728, 137
Scott, D., & Rees, M. 1990, MNRAS, 247, 510
Seljak, U., & Yu, B. 2019, arXiv e-prints, arXiv:1901.04454
Semelin, B., Eames, E., Bolgar, F., & Caillat, M. 2017, MNRAS, 472, 4508
Shaw, J. R., Sigurdson, K., Pen, U.-L., Stebbins, A., & Sitwell, M. 2014, ApJ, 781, 57
Singh, S., Subrahmanyan, R., Udaya Shankar, N., et al. 2017, ApJ, 845, L12
Sivia, D. S., & Skilling, J. 2006, Data Analysis - A Bayesian Tutorial, 2nd edn., Oxford Science Publications (Oxford University Press)
Smirnov, O. M. 2011, A&A, 527, A106
Sullivan, I. S., Morales, M. F., Hazelton, B. J., et al. 2012, ApJ, 759, 17
Sun, G., & Furlanetto, S. 2016, MNRAS, 460, 417
Tegmark, M. 1997, Phys. Rev. D, 55, 5895
Thompson, A. R., Moran, J. M., & Swenson, Jr., G. W. 2017, Interferometry and Synthesis in Radio Astronomy, 3rd Edition
Thyagarajan, N., Parsons, A. R., DeBoer, D. R., et al. 2016, ApJ, 825, 9
Thyagarajan, N., Udaya Shankar, N., Subrahmanyan, R., et al. 2013, ApJ, 776, 6
Thyagarajan, N., Jacobs, D., Bowman, J., et al. 2015, ApJ, 804, 14
Tingay, S. J., Goeke, R., Bowman, J. D., et al. 2013, PASA, 30, e007
Trott, C., Wayth, R., & Tingay, S. 2012, ApJ, 757, 101
Trott, C. M., Jordan, C. H., Midgley, S., et al. 2020, MNRAS, 493, 4711

van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2
Wells, W. T., Anderson, R. L., & Cell, J. W. 1962, The Annals of Mathematical Statistics, 33, 1016
Wieringa, M. H. 1992, Experimental Astronomy, 2, 203
Wise, J. H. 2019, Contemporary Physics, 60, 145
Xu, H., Wise, J., Norman, M., Ahn, K., & O'Shea, B. 2016, ApJ, 833, 84
Zahn, O., Lidz, A., McQuinn, M., et al. 2007, ApJ, 654, 12
Zahn, O., Mesinger, A., McQuinn, M., et al. 2011, MNRAS, 414, 727
Zhang, Y. G., Liu, A., & Parsons, A. R. 2018, ApJ, 852, 110
Zheng, H., Tegmark, M., Buza, V., et al. 2014, MNRAS, 445, 1084
Zheng, H., Tegmark, M., Dillon, J. S., et al. 2017, MNRAS, 464, 3486