



MIT Open Access Articles

First Results from HERA Phase I: Upper Limits on the Epoch of Reionization 21 cm Power Spectrum

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation	Abdurashidova, Zara, Aguirre, James E., Alexander, Paul, Ali, Zaki S., Balfour, Yanga et al. 2022. "First Results from HERA Phase I: Upper Limits on the Epoch of Reionization 21 cm Power Spectrum." The Astrophysical Journal, 925 (2).
As Published	10.3847/1538-4357/ac1c78
Publisher	American Astronomical Society
Version	Final published version
Citable link	https://hdl.handle.net/1721.1/142049
Terms of Use	Creative Commons Attribution 4.0 International license
Detailed Terms	https://creativecommons.org/licenses/by/4.0/



First Results from HERA Phase I: Upper Limits on the Epoch of Reionization 21 cm Power Spectrum

Zara Abdurashidova¹, James E. Aguirre² , Paul Alexander³, Zaki S. Ali¹, Yanga Balfour⁴, Adam P. Beardsley^{5,6,24} , Gianni Bernardi^{4,7,8} , Tashalee S. Billings², Judd D. Bowman⁶ , Richard F. Bradley⁹, Philip Bull^{10,11} , Jacob Burba¹², Steve Carey³, Chris L. Carilli¹³ , Carina Cheng¹, David R. DeBoer¹ , Matt Dexter¹, Eloy de Lera Acedo³, Taylor Dibblee-Barkman¹⁴, Joshua S. Dillon^{1,24} , John Ely³, Aaron Ewall-Wice¹⁵ , Nicolas Fagnoni³, Randall Fritz⁴, Steven R. Furlanetto¹⁶ , Kingsley Gale-Sides³, Brian Glendenning¹³, Deepthi Gorthi¹ , Bradley Greig¹⁷ , Jasper Grobbelaar⁴, Ziyaad Haldy⁴, Bryna J. Hazelton^{18,19} , Jacqueline N. Hewitt¹⁵, Jack Hickish¹, Daniel C. Jacobs⁶ , Austin Julius⁴, Nicholas S. Kern^{1,15} , Joshua Kerrigan¹² , Piyanat Kittiwisit²⁰ , Saul A. Kohn² , Matthew Kolopanis⁶ , Adam Lanman¹², Paul La Plante^{1,2}, Telalo Lekalake⁴, David Lewis⁶, Adrian Liu¹⁴ , David MacMahon¹, Lourence Malan⁴, Cresshim Malgas⁴, Matthys Maree⁴, Zachary E. Martinot², Eunice Matsetela⁴, Andrei Mesinger²¹ , Mathakane Molewa⁴, Miguel F. Morales¹⁸ , Tshegofalang Mosiane⁴, Steven G. Murray⁶ , Abraham R. Neben¹⁵ , Bojan Nikolic³, Chuneeta D. Nunhokee¹ , Aaron R. Parsons¹, Nipanjana Patra¹ , Robert Pascua^{1,14}, Samantha Pieterse⁴, Jonathan C. Pober¹² , Nima Razavi-Ghods³, Jon Ringuette¹⁸, James Robnett¹³, Kathryn Rosie⁴, Peter Sims¹² , Saurabh Singh¹⁴ , Craig Smith⁴, Angelo Syce⁴, Nithyanandan Thyagarajan^{6,13,25} , Peter K. G. Williams^{22,23} , and Haoxuan Zheng¹⁵

The HERA Collaboration

¹ Department of Astronomy, University of California, Berkeley, CA, USA; nkern@mit.edu

² Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, USA

³ Cavendish Astrophysics, University of Cambridge, Cambridge, UK

⁴ SKA-SA, Cape Town, South Africa

⁵ Department of Physics, Winona State University, Winona, MN, USA

⁶ School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA

⁷ Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa

⁸ INAF-Istituto di Radioastronomia, via Gobetti 101, I-40129 Bologna, Italy

⁹ National Radio Astronomy Observatory, Charlottesville, VA, USA

¹⁰ School of Physics & Astronomy, Queen Mary University of London, London, UK

¹¹ Department of Physics and Astronomy, University of Western Cape, Cape Town 7535, South Africa

¹² Department of Physics, Brown University, Providence, RI, USA

¹³ National Radio Astronomy Observatory, Socorro, NM, USA

¹⁴ Department of Physics and McGill Space Institute, McGill University, 3600 University Street, Montreal, QC H3A 2T8, Canada

¹⁵ Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA

¹⁶ Department of Physics and Astronomy, University of California, Los Angeles, CA, USA

¹⁷ School of Physics, University of Melbourne, Parkville, VIC 3010, Australia

¹⁸ Department of Physics, University of Washington, Seattle, WA, USA

¹⁹ eScience Institute, University of Washington, Seattle, WA, USA

²⁰ School of Chemistry and Physics, University of KwaZulu-Natal, Westville Campus, Durban, South Africa

²¹ Scuola Normale Superiore, I-56126 Pisa, PI, Italy

²² Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA, USA

²³ American Astronomical Society, Washington, DC, USA

Received 2021 February 28; revised 2021 June 11; accepted 2021 August 1; published 2022 February 7

Abstract

We report upper limits on the Epoch of Reionization 21 cm power spectrum at redshifts 7.9 and 10.4 with 18 nights of data (~ 36 hr of integration) from Phase I of the Hydrogen Epoch of Reionization Array (HERA). The Phase I data show evidence for systematics that can be largely suppressed with systematic models down to a dynamic range of $\sim 10^9$ with respect to the peak foreground power. This yields a 95% confidence upper limit on the 21 cm power spectrum of $\Delta_{21}^2 \leq (30.76)^2 \text{ mK}^2$ at $k = 0.192 \text{ h Mpc}^{-1}$ at $z = 7.9$, and also $\Delta_{21}^2 \leq (95.74)^2 \text{ mK}^2$ at $k = 0.256 \text{ h Mpc}^{-1}$ at $z = 10.4$. At $z = 7.9$, these limits are the most sensitive to date by over an order of magnitude. While we find evidence for residual systematics at low line-of-sight Fourier k_{\parallel} modes, at high k_{\parallel} modes we find our data to be largely consistent with thermal noise, an indicator that the system could benefit from deeper integrations. The observed systematics could be due to radio frequency interference, cable subreflections, or residual instrumental cross-coupling, and warrant further study. This analysis emphasizes algorithms that have minimal inherent signal loss, although we do perform a careful accounting in a companion paper of the small forms

²⁴ NSF Astronomy and Astrophysics Postdoctoral Fellow.

²⁵ Jansky Fellow of the National Radio Astronomy Observatory.



Content from this work may be used under the terms of the Creative Commons Attribution 4.0 license. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

of loss or bias associated with the pipeline. Overall, these results are a promising first step in the development of a tuned, instrument-specific analysis pipeline for HERA, particularly as Phase II construction is completed en route to reaching the full sensitivity of the experiment.

Unified Astronomy Thesaurus concepts: [Reionization \(1383\)](#); [Cosmology \(343\)](#); [Astronomy data analysis \(1858\)](#)

1. Introduction

The 21 cm line of neutral hydrogen has emerged in the past few decades as a theoretically powerful probe of cosmology and astrophysics by tracing the growth of structure across cosmic time (Hogan & Rees 1979; Madau et al. 1997). At high redshift ($z > 6$), the 21 cm line is a tomographic probe of the intergalactic medium (IGM) and allows us to directly trace its ionization, density, and temperature state. This is particularly important for understanding the Cosmic Dawn and the subsequent Epoch of Reionization (EoR), when radiative feedback from the first generation of stars and galaxies heated and ionized the IGM over the course of baryonic structure growth at redshifts $40 < z < 6$. Understanding these two milestones will give us a much better understanding of the formation of the first luminous sources, their environments, and the growth of large-scale structure.

Our current understanding of the timing of the EoR is largely based on measuring the absorption and scattering of the IGM against background sources. These can be either (i) astrophysical sources, such as galaxies and quasars whose Lyman series transmission is sensitive to the ionization state of the IGM; or (ii) the Cosmic Microwave Background (CMB), whose fluctuations are sensitive to the integrated column density of free electrons. Recent analyses of both (i) and (ii) point toward an EoR that is evolving rapidly between $5.5 < z < 7$. In particular, IGM damping wing absorption in QSO spectra (e.g., Mesinger & Haiman 2004; Bolton et al. 2011; Greig et al. 2017; Davies et al. 2018) and the rapid decline of Ly α emitting galaxies (e.g., Stark et al. 2010; Schenker et al. 2012; Jensen et al. 2013; Caruana et al. 2014; Pentericci et al. 2014; Mesinger et al. 2015; Mason et al. 2018, 2019) suggest that the EoR is ongoing at $z \sim 7$, albeit with significant systematic uncertainties. The Ly α forest at $z \sim 6$ suggests that reionization has mostly completed by then (McGreer et al. 2015). However, the sizable sightline-to-sightline scatter in the forest transmission (Becker et al. 2015; Bosman et al. 2018) seems to require the final, overlap stages of the EoR to extend down to $z \sim 5.5$ (Kulkarni et al. 2019; Keating et al. 2020; Choudhury et al. 2021; Qin et al. 2021). This is broadly consistent with the CMB constraints on the electron scattering optical depth (Planck Collaboration 2020) that supports a relatively late EoR with a midpoint around $z \sim 7$ (e.g., Douspis et al. 2015; Mitra et al. 2015; Greig & Mesinger 2017; Millea & Bouchet 2018; Qin et al. 2020b), and it is also consistent with the high- z galaxy luminosity functions, given reasonable assumptions about their properties (Robertson et al. 2015; Price et al. 2016; Gorce et al. 2018; Hazra et al. 2020; Park et al. 2019; Qin et al. 2020a). However, much remains to be understood about the EoR, such as when it began, the properties of the astrophysical sources that drove it, and what impact it had on future generations of star-forming galaxies. Meanwhile, considerably less is understood about the Cosmic Dawn, including when the first Population III stars formed, what their physical and spectral properties were, what their impact was in heating and enriching the IGM, and how this ultimately enabled the emergence of Population II stars and modern galaxies as we understand them.

Low-frequency radio experiments are aiming to directly measure the high redshift 21 cm signal and in doing so place constraints on the timing and duration of the EoR to understand the underlying physical processes driving the heating and reionization of the IGM (for reviews, see Ciardi & Ferrara 2005; Furlanetto et al. 2006; Morales & Wyithe 2010; Pritchard & Loeb 2012; Mesinger 2016; Liu & Shaw 2020). Prior and ongoing interferometric experiments include the Donald C. Backer Precision Array for Probing the Epoch of Reionization (PAPER; Parsons et al. 2010; Cheng et al. 2018; Kolopanis et al. 2019), the Murchison Widefield Array (MWA; Tingay et al. 2013; Dillon et al. 2014, 2015; Beardsley et al. 2016; Ewall-Wice et al. 2016b; Li et al. 2019; Barry et al. 2019b; Trott et al. 2020), the Low Frequency Array (LOFAR; van Haarlem et al. 2013; Patil et al. 2017; Gehlot et al. 2019; Mertens et al. 2020), the Giant Metre Wave Radio Telescope (GMRT; Paciga et al. 2013), and the Long Wavelength Array (Eastwood et al. 2019), which have put increasingly stringent constraints on the power spectrum over the past decade. At the same time, total-power measurements of the sky-averaged signal (or global signal) have similarly put upper limits on the monopole component of the 21 cm signal (Bernardi et al. 2016; Monsalve et al. 2017; Singh et al. 2017), with a tentative first-detection of the 21 cm global signal at $z \approx 17$ from the Experiment to Detect the Global EoR Signature (Bowman et al. 2018), with a robust discussion of whether this signature may be due to instrumental systematics (Hills et al. 2018; Bradley et al. 2019; Singh & Subrahmanyan 2019; Sims & Pober 2020; Mahesh et al. 2021).

The 21 cm power spectrum at the EoR and Cosmic Dawn contains a wealth of statistical information that can be used as both a cosmological and astrophysical probe (Mao et al. 2008; Patil et al. 2014; Pober et al. 2014; Liu & Parsons 2016; Greig et al. 2016; Ewall-Wice et al. 2016c; Kern et al. 2017). While radio interferometric experiments have indeed made substantial progress over the past decade, a first power spectrum detection has yet to be made. As second-generation 21 cm experiments are designed and built, such as the Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017) and the Square Kilometre Array (SKA; Koopmans et al. 2015), the understanding and control of instrumental systematics will be the crucial factor in enabling their ultimate success in making a first robust detection of the 21 cm power spectrum.

HERA is an interferometric array of fixed, zenith pointing dishes located in the Karoo desert, South Africa. The dishes are 14 meters in diameter and packed hexagonally into a nearly continuously covered core 300 m across. The array is being built in a series of phases with simultaneous construction and observing. Phase I used the feeds and correlator from the PAPER experiment (Thyagarajan et al. 2016; Ewall-Wice et al. 2016a; Patra et al. 2018; Fagnoni et al. 2021a), while Phase II will use a new feed as well as analog and digital systems (Fagnoni et al. 2021b). The data reported here come from the second internal data release (IDR2) taken from the first Phase I observing season, which commenced with roughly 50 antennas.

A series of recent papers describe the Phase I analysis pipeline in detail, which includes redundant calibration (Dillon et al. 2020), absolute calibration (Kern et al. 2020a), systematic modeling (Kern et al. 2019, 2020b), power spectrum estimation and error propagation (Tan et al. 2021), and pipeline validation (Aguirre et al. 2022). Complementary studies on the data set discussed in this work also include foreground modeling (Ghosh et al. 2020), imaging (Carilli et al. 2018), power spectrum analysis of the bispectrum phase (Thyagarajan et al. 2020), antenna primary beam characterization (Nunhokee et al. 2017), and electromagnetic modeling of the front-end signal chain (Fagnoni et al. 2021a). Here, we give an overview of the full analysis pipeline, discuss the criteria used for data selection, present the power spectrum limits, and describe statistical tests used to characterize the performance of the system and our analysis techniques.

In this paper, we report the first upper limits on the 21 cm power spectrum from HERA Phase I at redshifts 7.9 and 10.4 with 18 nights of observations. These observations were made with only a fraction of the array built (~ 50 out of 350 antennas), as it was under active construction at the time. Our analysis shows that the data are largely consistent with the expected thermal noise level at large line-of-sight Fourier k modes, achieving a dynamic range with respect to the peak foreground emission of 10^9 in power. However, at low Fourier k modes we see evidence for residual, low-level systematics that begin to exceed the thermal noise. Nonetheless, the limits presented here are to date the most sensitive at $z \sim 8$ by over an order of magnitude. A companion astrophysical interpretation paper shows that, under standard galaxy astrophysics, these limits disfavor cold reionization scenarios, where the IGM temperature is not substantially heated above its adiabatically cooled limit (HERA Collaboration 2021, in preparation).

The paper is organized as follows. In Section 2 we give a summary of the instrument and the observations analyzed in this work. In Section 3 we discuss the data reduction pipeline. In Section 4 we discuss our power spectrum estimation pipeline and present our integrated power spectrum limits. In Section 5 we discuss a suite of statistical tests used to assess the quality and stability of the final data products. Finally, in Section 6 we summarize our results.

2. Observations

Here, we discuss the observational parameters used for this analysis, as well as the state of the HERA instrument when these observations were made. As noted, the data discussed in this work were taken with the Phase I instrument, which was a hybrid HERA/PAPER system. Phase I re-purposed the radio frequency (RF) signal chains and correlator from PAPER and attached them to new HERA dishes. The antenna consists of a 14 meter dish with a cross-dipole feed at its focal point measuring two linear polarizations (DeBoer et al. 2017). At 150 MHz the beam has a full width at half maximum (FWHM) of $\sim 10^\circ$ (Fagnoni et al. 2021a). Not all PAPER front ends could be salvaged and as a stopgap new signal chain components—feed baluns and post-amplifiers—were manufactured to be backward compatible with the $75\ \Omega$ cables carried over from PAPER.

Kern et al. (2020b) characterized the performance of the new and old signal chains for the Phase I system, finding spectral structure across a range of delays that cover the EoR window, which they attribute partially to cable reflections. They also

Table 1
HERA Phase I Array and Correlator Parameters

Array Parameters		
Array coordinates (Lat, Lon)	$-30^\circ 7\ \text{S}, 21^\circ 4\ \text{E}$	
Total antenna number	52	
Unflagged antenna number	39	
Min. baseline length	14.6	meters
Max. baseline length	140	meters
Dish diameter	14	meters
Correlator Parameters		
Min. frequency	100	MHz
Max. frequency	200	MHz
Number of channels	1024	
Channel width	97.66	kHz
Integration time	10.7	sec
Nighttime recording duration	12	hr
Total data duration	18	nights

present methods for modeling and suppressing these systematics in the data. Fagnoni et al. (2021a) also performed electromagnetic simulations of the Phase I signal chains, finding a broad range of structure induced by both instrument cross-coupling and cable reflections, as well as dish reflections.

The HERA Phase I correlator, re-used from the PAPER experiment, employs an FX architecture, which was housed in a radio frequency interference (RFI)–shielded container in the field. In the F-engine, analog-to-digital units on a ROACH2 board (Parsons & Backer 2009) digitize each antenna’s linear polarization voltage stream, which are then fed to a field programmable gate array (FPGA) that Fourier transforms the signal into voltage spectra with a 100 MHz bandwidth from 100–200 MHz. This is done across 1024 channels, leading to a spectral resolution of 97.66 kHz. The spectra are then sent to a Graphics Processing Unit–based X engine that correlates all $N(N-1)/2$ cross antenna-polarization pairs and all N auto antenna-polarization pairs, where N is the number of feeds. The correlator then integrates the data for 10.7 s before writing them to disk. Finally, the data are chunked into 10 minute files and sent to the on-site Karoo Array Processing Center for storage, and eventually transported to the National Radio Astronomy Observatory Array Operations Center in New Mexico, USA, via internet connection. Other observational parameters are summarized in Table 1.

The HERA Phase I observing season ran from 2017 October to 2018 April while the array was under active construction. Observations were taken throughout the South African summer at night when the Galactic center and Sun are below the horizon, while active construction and commissioning proceeded during the day. A major focus of development was building an online data quality assessment pipeline that reported on radio interference, calibration quality, and other metrics. These metrics, some of which we discuss in the next section, were used to select a subset of data taken during a relatively stable epoch in the observing season when construction activities were at a minimum and data quality was consistently high. This resulted in an 18 day data set that is the basis of this work, referred to as the second internal data release (IDR2). These observations spanned 2017 December 10–28, (Julian dates 2,458,098–2,458,116). For each of these days, observations were made for 12 hr per night, of which roughly

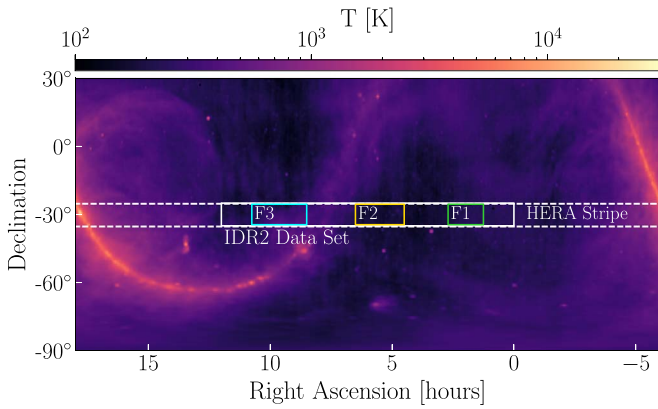


Figure 1. The Global Sky Model at 150 MHz (de Oliveira-Costa et al. 2008) showing the dominant diffuse foregrounds from the Galaxy. Being a non-tracking, zenith pointed array, HERA’s field of view is centered at $\delta = -30^\circ 7$ and in theory spans a full 360° range in R.A. With a primary beam FWHM of $\sim 10^\circ$ (at 150 MHz), HERA’s sensitivity is primarily focused on a narrow strip of sky. The data reported here span a range of right ascensions from 0 to 12 hr (IDR2), although only three specific fields (colored boxes) largely devoid of bright foregrounds are used for power spectrum estimation.

10 hr were used for science data when the Sun was below the horizon.

Together, these observations span a local sidereal time (LST) of 0 to 12 hr (Figure 1), which overlaps with a radio cold patch at 4 hr LST, as well as the Galactic plane at 8 hr LST for HERA’s latitude of $-30^\circ 7$ South. In total, this work considers an 18 night data set (for a total of 180 hr), however, as discussed in Section 3, after additional rounds of data cuts we are left with three fields, each with ~ 36 hr of integration for power spectrum analysis. The area of the sky observed in the IDR2 data set is highlighted in Figure 1, which shows the diffuse foreground emission from the Galaxy and also identifies the three main fields used for power spectrum analysis (colored boxes). The edges of the data set are not included due to reduced sensitivity of the nightly averaged data products caused by LST drift of the observations from night to night.

At the time of observations, the array had 52 fully deployed antennas arranged in a close-packed fashion, which constitutes only a fraction of the full 350 that will eventually be built, and corresponds to a corner of the full array (Figure 2). However, as we will discuss next, only 39 of these antennas were deemed science quality and were used for data analysis.

3. Data Reduction

Our analysis pipeline is summarized in Figure 3, which highlights the data reduction pipeline (green), the power spectrum pipeline (red), the data products that are input to and output by the pipeline (blue), and the steps that are tested in the companion validation analysis (dashed; Aguirre et al. 2022). Here, we will focus only detailing the data reduction pipeline, which we do in order of their appearance in the pipeline. The primary role of the reduction pipeline is to identify and flag faulty data, solve for the complex gain solutions imparted by the instrument and invert them, average the visibilities coherently across nights, and treat the data for known systematics. Note that, unlike other works, we do not perform any explicit foreground subtraction or filtering in this analysis, although future work may benefit from such techniques (e.g., Ewall-Wice et al. 2021).

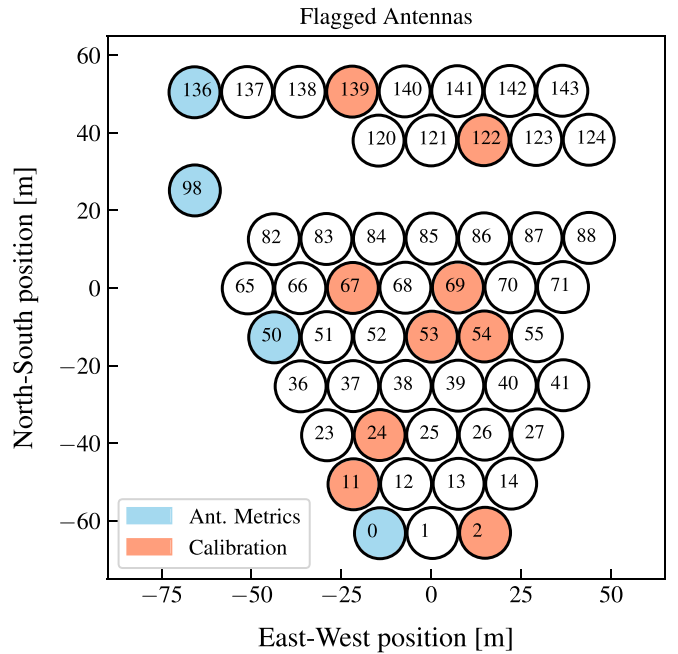


Figure 2. The HERA Phase I array layout. Antennas flagged by the antenna metrics stage are shown in blue, while those flagged during calibration are shown in red, leaving a total of 39 unflagged antennas. For details on this process, see Dillon et al. (2020).

Note that all of the important analysis packages can be found in the publicly accessible HERA-Team²⁶ software repository, including the `hera_cal`, `hera_pspec`, `hera_qm`, and `hera_opm` packages used extensively in this analysis.

3.1. Antenna Metrics

Raw data from the correlator are first sent through an antenna metrics stage where faulty antennas are identified and flagged. This is an important pre-calibration step to prevent obviously malfunctioning antennas from adversely affecting the calibration solutions. Metrics are calculated on every 10 minute file for each of the dual-polarization feeds of every antenna, which gives us a sense for how antennas and their linear polarization feeds are behaving throughout any given night.

While a few different metrics were trialed, one metric was found to be the most useful and was the primary metric used to flag faulty antennas. This metric was the mean visibility amplitude for each antenna, which we define for a given antenna i as

$$M_i = \frac{\sum_{j \neq i, \nu, t} |V_{ij}(\nu, t)|}{N - 1}, \quad (1)$$

where $V_{ij}(\nu, t)$ is the visibility between antennas i and j at frequency ν and time t , N is the number of antennas in the array, and the sum is taken over all antennas $j \neq i$, times, and frequencies in the observation. For notational brevity, we will drop the explicit frequency and time dependence of V_{ij} . This metric measures when an approximate estimate of the antenna gain is low compared to most other antennas, which can occur, for example, when the signal chain is not connected to the feed or a gain stage has lost power. It was also used to help determine when an antenna was cross-polarized, or when its

²⁶ <https://github.com/hera-team>

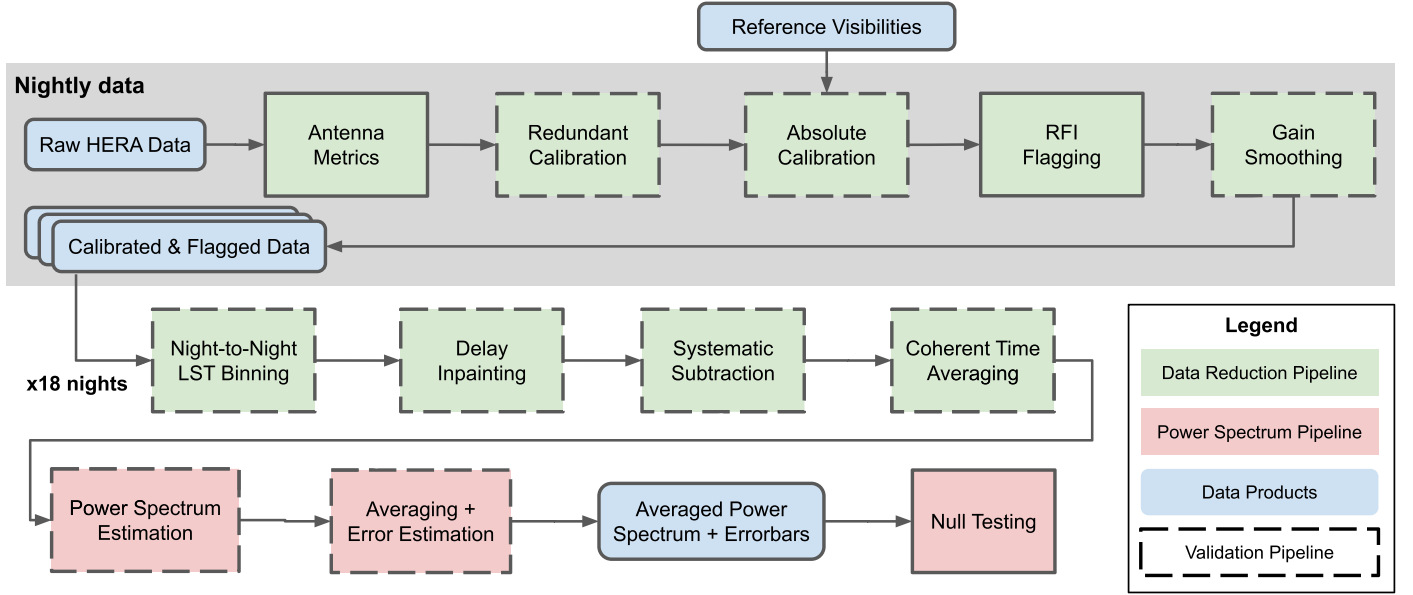


Figure 3. A diagram of the data reduction and power spectrum estimation pipelines, starting with raw HERA data and ending with the averaged power spectrum and its associated null tests. The blue boxes represent data products, while the green and red boxes represent steps in the data reduction and power spectrum pipelines, respectively. Boxes with dashed borders represent elements tested by the validation pipeline (Aguirre et al. 2022). Noise simulations are generated after the data reduction pipeline and fed through the power spectrum pipeline for diagnostic purposes when evaluating null tests.

east and west polarization signal chains were accidentally interchanged in the cable connections downstream.

Because the metric is computed before calibration, and thus uses raw data, we cannot make an absolute cut based on the expected noise level of the visibilities in Janskys. Instead, we estimate the standard deviation of the metric across antennas and compute a Z-score, which is defined as the deviation of any given point from its sample mean in units of the sample standard deviation. To account for outliers, we use the more robust, modified Z-score, defined as

$$Z_i^{\text{mod}} = \frac{x_i - \text{med}(x)}{\sigma^{\text{mad}}} \quad (2)$$

$$\sigma^{\text{mad}} = 1.482 \times \text{med}|x - \text{med}(x)|, \quad (3)$$

where x is the mean visibility amplitude metric defined above.²⁷ The median absolute deviation (σ^{mad}) is a robust estimator of the standard deviation but requires a correction factor of 1.482 to reproduce the standard deviation in the case of white Gaussian noise (Rousseeuw & Croux 1993).

It is generally clear from these metrics which antennas are poorly behaving; nonetheless, we flag antennas that deviate from the sample mean beyond a 5σ level. Antennas that were repeatedly flagged by the antenna metrics stage for nights throughout the data set are flagged outright for all nights. If either of an antenna’s polarizations is thus flagged, we flag the entire antenna (i.e., we flag both linear polarizations).

The faulty antennas caught at the antenna metrics stage are really only the very worst offenders, shown in blue in Figure 2. Other subnominal antennas are flagged subsequently in the process of calibration, which we describe next.

²⁷ Using the modified Z-score still relies on the majority of the array being healthy in order to detect outlier antennas. When this is not true, for example during correlator errors or partial power outage, it is very obvious from visual inspection of the data products.

3.2. Redundant Baseline Calibration

One of the foremost challenges for 21 cm cosmology is the task of precision instrumental gain calibration. In general, the measured visibility V_{ij}^{obs} between antennas i and j is related to the true, uncorrupted visibility V_{ij}^{true} via a product of each antenna’s gain, g_i ,

$$V_{ij}^{\text{obs}} = g_i g_j^* V_{ij}^{\text{true}} + n_{ij}, \quad (4)$$

where n_{ij} is the thermal noise in the measurement (Hamaker et al. 1996; Smirnov 2011). Note that we solve Equation (4) for both linear polarizations independently (EE and NN), which are also implicitly a function of time and frequency. This representation ignores intrafeed D terms as well as direction-dependent gain calibration, which we do not solve for in this analysis (a simulated primary beam model is used for sky-based calibration discussed in Section 3.3).

HERA’s redundant array configuration opens the possibility of calibrating the relative gains of antennas using internal degrees of freedom among measurements of groups of redundant baselines (Dillon & Parsons 2016). Redundant baseline calibration uses the principle that every redundant baseline should measure the same visibility; in practice, they do not due to antenna-based gain terms, which allows one to constrain the relative gain of each antenna (Wieringa 1992; Liu et al. 2010). Specifically, redundant baseline calibration seeks to minimize a χ^2 written as

$$\chi^2 = \sum_{i < j} \frac{|V_{ij}^{\text{obs}} - g_i g_j^* V_{i-j}^{\text{sol}}|^2}{\sigma_{ij}^2}, \quad (5)$$

where V_{i-j}^{sol} is the visibility solution for the redundant baseline type with the same vector separation as V_{ij} . The key distinction in redundant baseline calibration highlighted here is that the visibility solutions are left as free parameters to be solved for in the process of calibration along with g_i and g_j . For highly

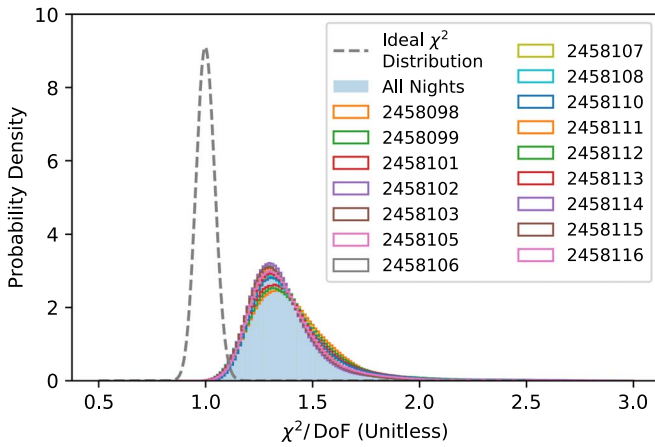


Figure 4. Redundant calibration attempts to minimize χ^2 , defined in Equation (5), using a model that assumes that redundant baselines observe the same true visibilities up to their antenna-dependent gains and the thermal noise. Were this true, we expect that χ^2 normalized by the number of degrees of freedom (DoF) in the model (approximately 520 in this analysis, see Dillon et al. (2020) for a precise quantification) would have a mean of one and follow a χ^2 distribution. Here, we show the distribution of χ^2/DoF over polarizations and unflagged (see Section 3.4) times and channels for different nights in our analysis. While consistent from night to night, the overall mean of 1.389 is clearly inconsistent with one, indicating persistent nonredundancy at levels roughly 20% in excess of the thermal noise. Figure reproduced from Dillon et al. (2020) with permission.

redundant configurations like HERA, this leads to an over-constrained system of equations that can be solved to yield estimates of the gains in addition to the model visibilities.

Due to inherent degeneracies in the system of equations, redundant baseline calibration cannot solve for the overall amplitude and directional phasing (the “Tip-Tilt” phasor) of the antenna-based gains (Liu et al. 2010; Zheng et al. 2014; Dillon et al. 2018; Li et al. 2018; Byrne et al. 2019). Consequently, these must be solved for independently and added to the redundant calibration gains, commonly referred to as absolute calibration.

A complete description of the redundant baseline calibration of HERA appears in Dillon et al. (2020), including a quantitative assessment of the redundancy of HERA via, e.g., the deviation of χ^2 for the expected value in a perfectly redundant array. For more detail on the series of algorithms with which we minimize χ^2 (Equation (5)), see Dillon et al. (2020). Furthermore, the subsequent process of solving for the degenerate modes (i.e., absolute calibration) is described in Kern et al. (2020a).

In Figure 4 we show the results of redundant baseline χ^2 minimization for different nights in our analysis.

Our results show clear evidence for nonredundancy in excess of the thermal noise at the $\sim 20\%$ level in the visibility. Dillon et al. (2020) attributes much of this excess to small deviations in the placement of dishes and to minor antenna-to-antenna variation of the primary beam. Some of it is also likely attributable to baseline-dependent systematics (Kern et al. 2020b) that break the assumption in Equation (4). As discussed in Dillon et al. (2020), the statistic is not consistent with one, meaning there are extra sources of variance in the data beyond noise that cannot be constrained by antenna-based calibration. We assess the effect of nonredundancy on our final power spectrum limits in Section 3.11.

3.3. Absolute Calibration

To finish calibration, we need to solve for the degeneracies of redundant baseline calibration by referencing to the sky. One way to do this is to compare a model of the true visibilities to redundantly calibrated visibilities and varying the degenerate parameters to make the two match. In this work, we use a series of previously calibrated visibilities over the full LST range as our set of reference (or model) visibilities that we use to extract the degenerate components of the calibration gains.

We build our reference visibilities following the sky-based calibration methodology described in Kern et al. (2020a), which uses the Common Astronomy Software Applications (CASA; McMullin et al. 2007) software to construct visibility models at specific fields where we expect to be able to model the dominant contributions of the flux on intermediate and long baselines. Specifically, we select three fields transiting our field of view (FoV) at 2.0, 5.2, and 14.4 hr LST. These fields are chosen to have minimal diffuse foregrounds within the FoV, and have a bright point source from the MWA GLEAM catalog (Hurley-Walker et al. 2017) near the field’s zenith pointing, which is used to confirm the absolute flux density scale of the calibrated data.

For each of the three fields we pick a different night in the Phase I observing season, selected such that the field transits roughly halfway through the nighttime observations. We then select out five minutes of observations centered at the field transit, average the data and perform calibration using standard CASA routines (e.g., gaincal and bandpass). These gains are then transferred to every integration for that observing night, leaving us with three nights of calibrated visibilities, each of which span a slightly different range of LSTs. Drift in the gain amplitude throughout the nighttime observation is not corrected for, but is expected to be at the $\sim 3\%$ level (Kern et al. 2020a).

Finally, these three sets of visibilities are averaged onto a common LST grid spanning nearly 20 hr of LST, which yields our full set of reference (or model) visibilities that we use for absolute calibration. After averaging, the visibilities are passed through a Fourier low-pass filter across frequency, which filters out all signals with delays beyond the baseline horizon delay plus an additional 50 nanoseconds. This suppresses noise in high delay structure that is not associated with foregrounds, and also fills in flagged frequency channels with a best guess of the foreground structure using a delay-domain deconvolution (similar to the inpainting algorithm described later; Parsons & Backer 2009; Kern et al. 2020a).

Figure 5 shows CLEANed, multifrequency synthesis images (across 135–165 MHz) of the three fields we use for constructing the reference visibilities, demonstrating the point-source distribution in each field. It also shows the Julian date of the night used to calibrated each field as well as the range in LSTs of each of the nights (green, orange, and blue bars). Stacking these visibilities onto a common LST grid gives us our final “Full Model” (purple bar) that we use in performing absolute calibration of the nightly data in the Phase I data set considered here. It should be noted that the gains solved for are inherently direction-independent gains, with no cross-feed D terms.

3.4. Radio Frequency Interference Flagging

The Karoo Astronomy Reserve is an RFI quiet zone, and as such much of the frequency band between 120 to 180 MHz is

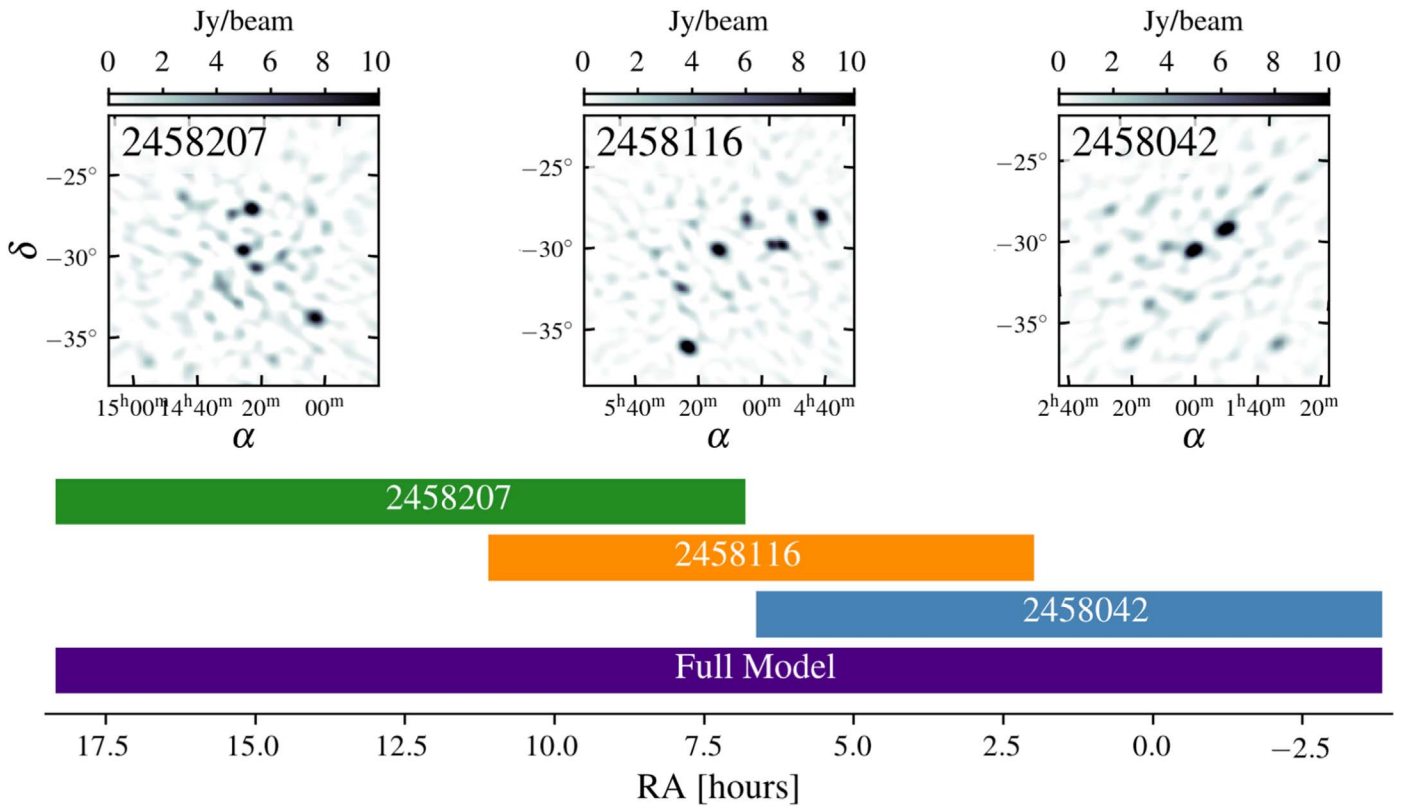


Figure 5. Construction of the absolute calibration reference visibilities (purple) from a set of three nights throughout the observing season. Each night is calibrated at a single field that transits near midway through the night. The calibration is transferred to all integrations from that observing night, and the visibilities are then LST binned onto a common grid to form the full model spanning nearly 22 hr in LST. Drift in the gain amplitude through each night is not corrected for, but is expected to drift at the $\sim 3\%$ level (Kern et al. 2020a).

relatively clean of interference. Nonetheless, we still see narrowband RFI due to satellite and terrestrial transmitters, as well as rare occurrences of wideband RFI above 180 MHz. Our RFI detection algorithm relies on the detection of local outliers, using the distribution of surrounding data points in time and frequency to distinguish RFI from thermal noise fluctuations. Our pipeline uses a number of data products to identify not just interference but also general problems with the array, such as correlator malfunctions. As its input it takes raw HERA data, the gain solutions from redundant baseline calibration and the redundant visibility solutions, its χ^2 distributions, the absolute calibration gains, and their χ^2 distributions.

For each data product (except the raw visibilities) a corresponding modified Z-score metric is computed, Z^{mod} , which is formed by median-subtracting the data product with a median filter and then normalizing it by an empirical estimate of its noise. Performing the median filter across frequency and time helps to identify occasional sources of wideband RFI from broadcast television that can be as wide as 8 MHz (Wilensky 2020). These metrics are averaged in quadrature across all baselines, antennas, and polarizations to increase the sensitivity to low-level contamination, resulting in a single metric as a function of time and frequency for each data product.

Next, we flag these metric waterfalls with an initial threshold of $Z^{\text{mod}} \geq 5$, followed by a watershed algorithm where any time and frequency pixel adjacent to a flagged pixel is itself flagged if it exceeds a lower threshold of $Z^{\text{mod}} \geq 2$ (Kerrigan et al. 2018). We found that different types of contaminants will

manifest more brightly in certain metrics, so each metric is flagged individually.

Having removed the brightest sources of RFI, we now apply the resultant flags and repeat the flagging procedure; however, this time we compute the standard Z-score metric using a mean filter rather than a median filter. The mean filter is considerably faster, so at this stage we are also able to include the raw visibility data as an additional metric. Just as before, we compute these metrics, average them in quadrature, and apply the same watershed flagging routine on the averaged metrics. The result is a single waterfall of flags for all baselines. This entire routine is performed on every 10 minute file over the course of a nightly observation.

Our last step is to examine each metric waterfall over the entire night and use medians to collapse them to a single time series or spectrum. These are again flagged by computing a modified Z-score for each channel or integration, flagging any above 7 or any above 3 that are adjacent to a previously flagged channel or integration. This step helps us find global outliers in time that may not be obvious from the analysis of a signal file and to find channels that are so frequently occupied by RFI that it is safer to assume they are always RFI-contaminated.

The strongest forms of identified RFI are persistent, narrowband emitters, which can be clearly seen in the calibration gain solutions (e.g., Figure 6). However, the final flagging mask applied to the nightly data excises both narrow and wideband features (e.g., the left panel of Figure 8). The contaminating features in the data include FM radio ($\nu < 111$ MHz), ORBCOMM satellite communication ($\nu = 137$ MHz),

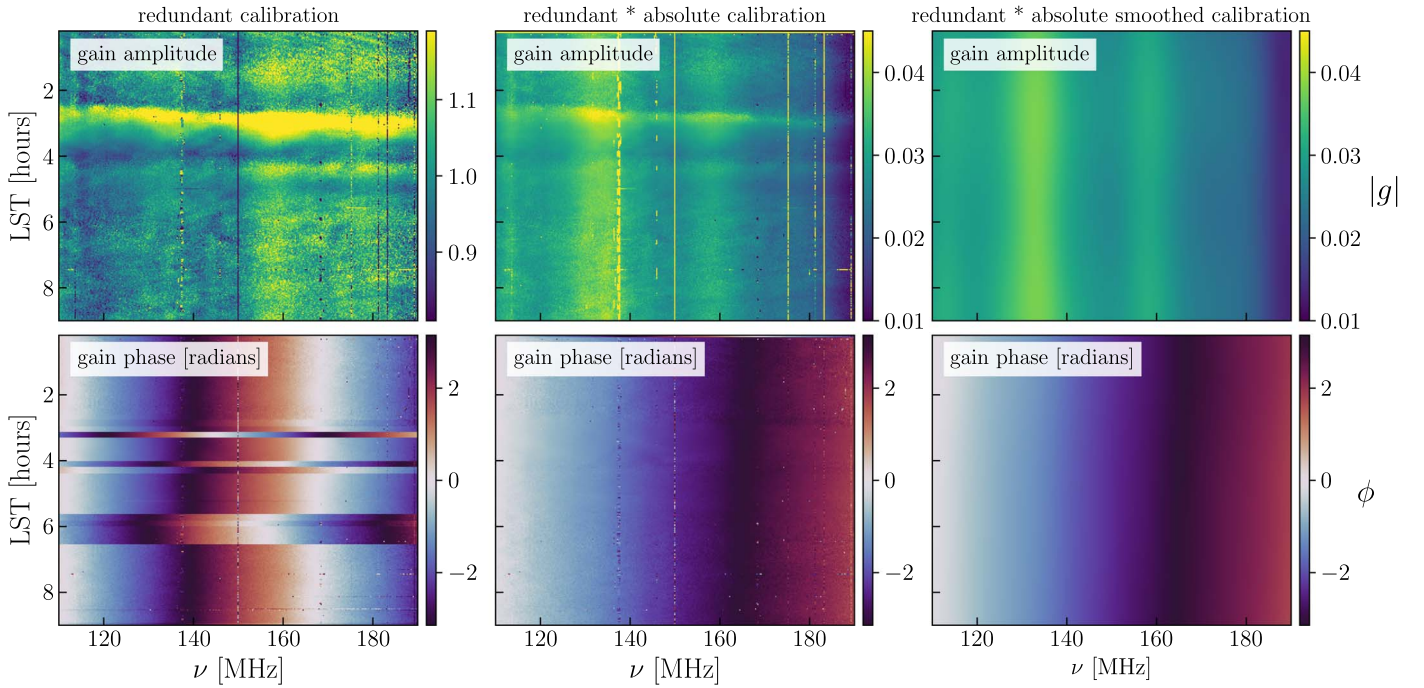


Figure 6. The progression of a single antenna-based gain through the calibration process of a single night, showing the gain amplitude (top panels) and phase (bottom panels). Redundant calibration (left) solves for the relative gain but is susceptible to phase jumps at certain times and requires certain degeneracies, like the overall gain amplitude, to be filled in. Absolute calibration (center) solves both of these problems by pinning the degenerate components and removing the phase jumps. Lastly, the final gains are smoothed across frequency and time (right) to limit excess structure (Dillon et al. 2020; Kern et al. 2020a)—for example, amplitude fluctuations caused by the passage of Fornax A around 3.4 hr seen in the top left and middle panels. Note that the process of gain smoothing is applied to the center column, not to the first column where phase jumps are apparent.

and broadcast television ($\nu > 174$ MHz), as well as intermittent correlator integration failures (wideband).

In addition to flagging the data for RFI and correlator failures, we also flag 50 channels on either edge of the frequency band where the bandpass filter falls off steeply, as well as all integrations where the Sun is at or above the horizon due to issues it creates in our calibration. Lastly, after LST binning (described below), the averaged visibilities are manually inspected for low-level narrowband RFI that was missed by the nightly RFI excision. In this process we find and flag a handful of frequency channels (for all baselines and times), which amounts to $<11\%$ of the total data.

On average, we flag roughly 15% of the band in our RFI flagging step. This estimate excludes our routine flagging of the band edges due to the roll-off of the bandpass filter (~ 10 MHz on either side of the band), as well as the complete flagging of certain times due to problems with correlator failures. Through visual inspection, we see areas where our pipeline is likely overflagging the data. This suggests that our flagging strategy is somewhat more aggressive and could be improved to reduce the false-positive rate. Future work will investigate how more precise RFI excision techniques, like the matched-filter SSINS package (Wilensky et al. 2019), can improve the overall RFI identification performance.

3.5. Gain Smoothing

While the goal of gain calibration is to remove spectral structure introduced by the instrument, it is a double-edged sword and can also impart spurious spectral structure. Many sources of uncertainty come into play when calibrating a real instrument, including baseline-to-baseline nonredundancies (Orosz et al. 2019), incomplete sky flux density models

(Byrne et al. 2019), and various baseline-dependent instrumental systematics. A number of techniques have been proposed to mitigate this effect, including the consensus optimization technique (Yatawatta 2015, 2016) used in LOFAR power spectra (e.g., Patil et al. 2017 and Mertens et al. 2020) and fitting low-order polynomials to averaged gains, a technique employed with the MWA (e.g., Barry et al. 2019a, 2019b).

Kern et al. (2020a) discussed some of these effects for HERA and found their impact to be most severe for $|\tau| > 100$ ns. They introduce a Fourier filtering procedure for low-pass filtering the gains after calibrating out a per-antenna delay, which mitigates the impact these spectrally dependent gain errors have on the data. This filter is a two-dimensional frequency and time iterative deconvolution algorithm that both acts as a low-pass filter and also fills in missing calibration solutions due to RFI flagging. It is conceptually similar to the Hogbom CLEAN algorithm (Högbom 1974).

Dillon et al. (2020) showed that per-integration redundant baseline calibration exhibits time-dependent structure that is LST-locked and argued that this variation was due to bright sources moving through direction-dependent nonredundancy. Examining the temporal power spectra of gain solutions after dividing out an LST-locked nightly average revealed no evidence for intrinsic gain fluctuations on timescales shorter than 6 hr. Likewise, Kern et al. (2020a) show that the gains drift in amplitude with changes in the ambient temperature, which varies slowly over the course of the 10 hr nightly observation. Therefore, we also use the 2D low-pass filter to remove temporal structure in the gain solutions on all timescales shorter than 6 hr. This substantially mitigates sky-dependent effects, since the filtering timescale is much longer

than the beam-crossing time—approximately 40 minutes at 150 MHz, given a 10° FWHM (Neben et al. 2016).

Figure 6 shows the progression of the antenna gains from redundant baseline calibration (left), absolute calibration (center), and gain smoothing (right). Redundant baseline calibration is prone to jumps in the degenerate subspace of calibration solutions, particularly in phase. These are solved by absolute calibration, which fills in the degenerate components, thus mitigating the phase discontinuities in time introduced by redundant calibration. Next the fast and localized spectral and time-dependent features seen in the redundant and absolute gains are filtered out by the smoothing step. Gain filtering occurs after RFI identification, meaning we can input our flag masks into the Fourier filtering algorithm to deconvolve them out and fill in the missing pixels with a model of the smoothly varying components.²⁸ This is reflected in the sharp spikes in the gains in RFI-dominated pixels that are smoothed over after gain filtering (Figure 6). Gain smoothing is performed independently for each antenna and linear polarization.

We are left with a set of gains spanning all frequencies and times of our data on a nightly basis that has been filtered from the initial solution at each frequency and time sample to a restricted number of degrees of freedom for each antenna and linear polarization. The total bandwidth of our data leads to a delay resolution of $\Delta\tau = 10$ ns meaning we keep 21 spectral degrees of freedom, and the nightly observation duration is 10 hr, leading to 3 temporal degrees of freedom, for a total of ~ 63 degrees of freedom in each antenna-polarization gain.²⁹

3.6. LST Binning

Having calibrated each night independently, we next coherently average them at fixed LST. Due to sidereal drift night to night, each observing night has a slightly offset LST grid, which must be accounted before LST binning. To do this, we establish a single LST grid spanning 0 to 24 hr at a cadence of 21.4 s, which is double the native correlator integration time and helps to increase the sample count in the subsequent outlier rejection routine. For each of these LST bins, we take data from all nights that fall within the bounds of this bin and rephase their pointing centers to the center of the LST bin. We then perform another round of outlier rejection to help flag missed RFI and other problems with the data that do not repeat night to night at the same LST (as the cosmological signal should). We do this by looking at the distribution of visibilities to be binned together for every particular LST, frequency, and baseline combination and rejecting individual samples across the nights that have a modified Z-score (as defined in Equation (2)) greater than 4σ . We set the threshold at 4σ such that the clipping has an effectively negligible impact on the Gaussian noise distribution, yet still rejects strong outliers that were for whatever reason not caught in our initially flagging round. Clipping is performed on the real and imaginary component of the visibility separately, but if either are clipped in the process we flag the pixel entirely. Indeed, in practice we find that the sigma clipping routine tends to catch clear broadband artifacts

²⁸ While we can infer gain solutions for times and frequencies that are flagged, we only do this for the purpose of finding a smooth gain model. Data that are flagged remain flagged.

²⁹ The gains are complex quantities and our filter extends from $-100 < \tau < 100$ ns, giving us 21 independent Fourier modes: 10 on each side of the $\tau = 0$ ns mode. Likewise, we keep the temporal modes corresponding to the -1 st, 0th, and 1st Fourier modes.

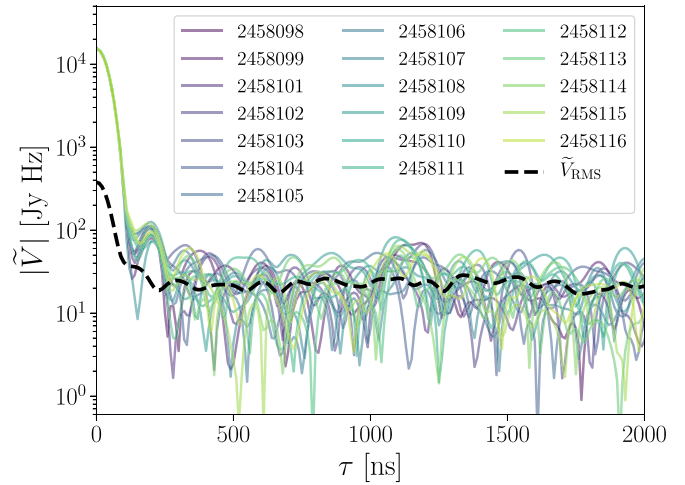


Figure 7. Delay-transformed visibilities from each night after calibration but before LST binning, showing the consistency of the calibrated visibilities from night to night for a fixed baseline and LST. We also plot the rms of the complex delay-transformed visibilities across nights (black dashed), showing excess variability above the noise at low delays (at $\sim 3\%$), indicative of night-to-night variation in calibration errors at the $\sim 1.5\%$ level.

due to correlator failures that were not initially caught. For the ~ 18 nights of nightly data sampled at a cadence of 10.7 s, each LST bin contains on average 36 samples as input to this sigma clipping routine. All data in the bin that were not originally flagged or flagged during the sigma clipping routine are then uniformly averaged onto a single LST grid.

A key requirement for LST binning is that each night is calibrated accurately, otherwise the averaged data will decohere. We assess this by looking at the variability of the calibrated data across all nights in the data set. Figure 7 shows the delay-transformed visibilities across each night from a single baseline and at a fixed LST. The dashed black line shows the root-mean-square (rms) of the visibilities across nights. At high delays we see the rms is consistent with the apparent noise level of the nightly data; however, at low delays the rms rises to a level that is $\sim 3\%$ of the peak foreground power. The most likely origin of this is the night-to-night variability of the gain calibration, which would put the gain errors on the $\sim 1.5\%$ level. This is consistent with the fidelity of the Phase I calibration pipeline assessed in Aguirre et al. (2022).

3.7. Data Inpainting

Flagging masks due to RFI and other instrumental problems are generally fairly complex, with a nontrivial frequency, time, and antenna dependence. Flagging masks presents a challenge for power spectral analyses that make measurements in the Fourier domain as they induce strong sidelobes in nonperiodic signals. This is especially troublesome for 21 cm measurements, which require high dynamic range signal separation against bright foregrounds in the Fourier domain. A related effect is produced when time averaging masked visibility data with differing frequency flags as a function to time. The resultant spectrum contains a complex weight structure as a function of frequency that can manifest as sidelobe structure of bright foregrounds (Offringa et al. 2019).

A common solution to the problem of masked or missing data is to inpaint the masked data with a best guess of the signal that should have been measured in that bin. A wide variety of algorithms have been developed to solve this kind of problem;

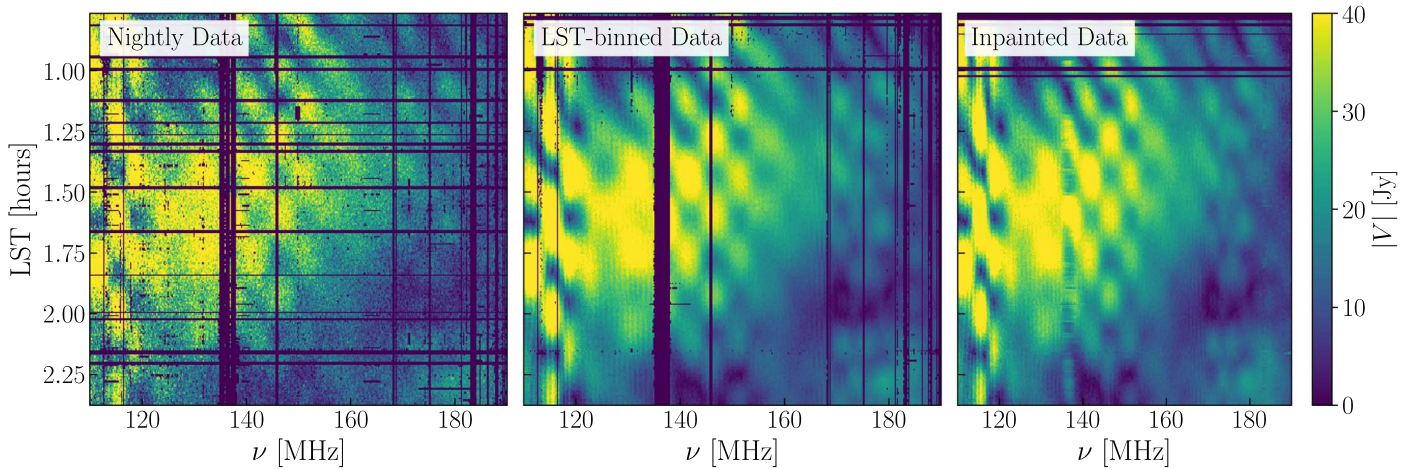


Figure 8. A progression of a visibility waterfall over a small time range through LST binning and flag inpainting, showing the reduction in flags (masked regions). Note the marked reduction in flags after LST binning, due to different nights having independent flagging masks. Data inpainting reconstructs a best guess of the data in flagged regions, but demonstrates poor performance over wideband flags like the ORBCOMM feature at 138 MHz. Because it operates on each integration independently, integrations that were fully flagged remain flagged.

the CLEAN algorithm (Högbom 1974), for example, is one way to do this, and is in fact how we addressed the problem of masked data in our gain smoothing procedure of Section 3.5. In this work, we use a similar delay-based iterative deconvolution algorithm to fill in the missing data in the LST-binned data set (Parsons & Backer 2009; Ali et al. 2015; Kerrigan et al. 2018). The deconvolution is performed across nearly the full bandwidth, spanning 115–185 MHz, and is performed on a per-integration basis down to the thermal noise of each visibility. The algorithm finds model components in delay space across a window spanning $-2000 < \tau < 2000$, which is chosen to encapsulate all of the strong features seen in the data across all baselines.

Figure 8 shows an example of the LST-binned data products and the results of delay deconvolution-based inpainting. Inpainting only affects the pixels where the data are masked, and because it is done only across frequency, integrations that are fully flagged remain fully flagged. Aguirre et al. (2022) study the accuracy of the inpainting algorithm described here, showing that its performance is degraded for widechannel gaps, like the 138 MHz ORBCOMM features (this is furthermore seen in our spectral window null test; Section 5.4).

3.8. Systematic Modeling

Kern et al. (2020b) and Fagnoni et al. (2021a) outline the major systematics seen in the HERA Phase I system, which are attributed to cable reflections in the 150 meter coaxial cable between the feed and the node and in the 20 meter cable between the node and the digitization stage, in addition to signal chain cross-coupling systematics (e.g., mutual coupling). Fagnoni et al. (2021a) use electromagnetic simulations of the HERA front end to show that mutual coupling (i.e., dish-to-dish communication) can appear as a decaying shoulder in the auto-correlation delay spectrum from 100–500 ns. They also predict the presence of a cable reflection at the termination of the coaxial cables, as well as a series of complicated subreflections in the 150 meter cable spanning 100–1200 ns that can be unresolved by the native delay resolution of the data. Inspection of the data by Kern et al. (2020b) affirm the presence of a shoulder in the auto-correlation delay spectrum,

however, its exact origin is not quite clear. Kern et al. (2020b) speculate that the shoulder could in fact be due to the cable subreflections and not mutual coupling, in part because it can be effectively modeled and suppressed via a reflection calibration procedure, which is not to be expected from a highly direction-dependent systematic like mutual coupling. Future work will seek to more clearly understand these systematics via improved electromagnetic beam modeling and visibility simulations, which explicitly include mutual coupling effects.

To deal with the subreflections, we employ the same reflection fitting algorithm on all antenna auto-correlations from the LST-binned data, iteratively solving for 25 individual reflection terms within a delay range of 150–1500 ns. These parameters were chosen manually after visual inspection of the residual visibilities. All reflection systematics are modeled at a 21.4 s cadence and then smoothed in time with a similar gain smoothing procedure employed in the first round of calibration.

The most prominent source of systematics observed in Phase I data is the presence of high delay features in the visibilities that are nearly constant in time, thus occupying the fringe-rate 0 Hz mode, and roughly symmetric at negative and positive delays (Kern et al. 2020b). For baselines with a projected east–west separation ≥ 14 meters, this systematic can be effectively filtered out from the data without significantly reducing the measured EoR power (Kern et al. 2019). However, this does lead to a small amount of signal loss on the cosmological signal that we correct for in the power spectra (Section 3.11). Overall, the combination of reflection calibration and cross-coupling filtering leads to roughly two orders of magnitude of suppression in instrumental systematics spanning $0.1 < k < 0.8 h \text{ Mpc}^{-1}$. If not enacted properly, this step can lead to nonnegligible amounts of cosmological signal loss, and is therefore vetted in our validation pipeline to ensure it performs as we expect it to (Aguirre et al. 2022).

3.8.1. Faraday-rotated Foreground Emission

After removing instrumental systematics from the data, we found low-level excesses in the power spectrum at high delays (beyond the foreground horizon) that transited on a

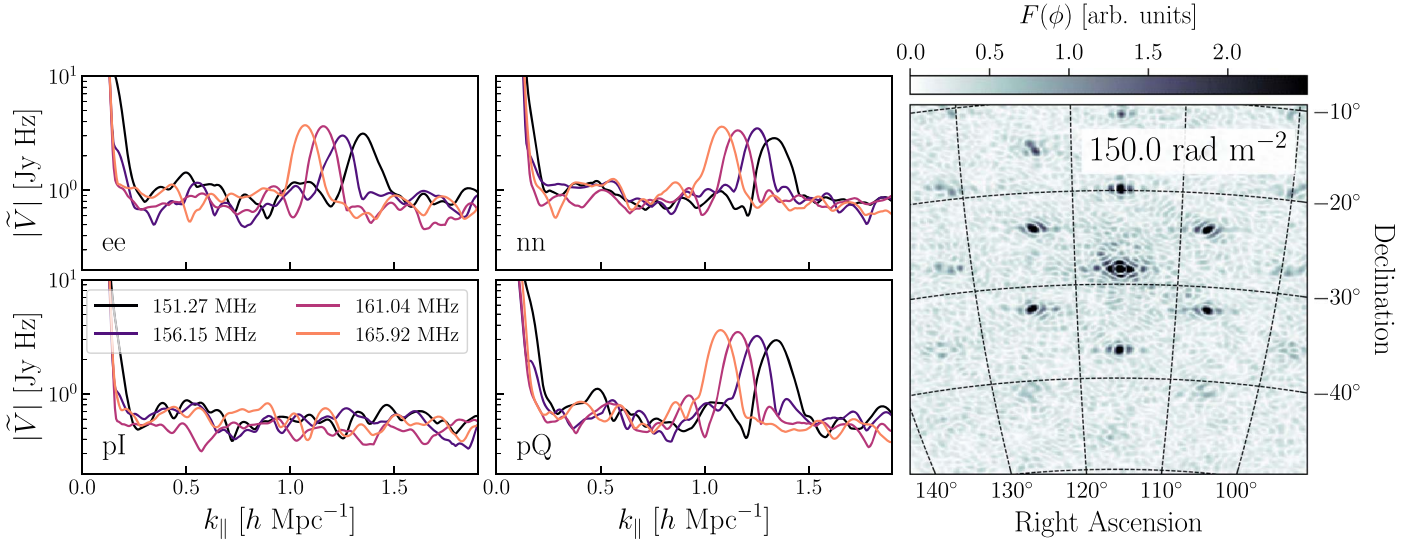


Figure 9. Faraday-rotated foreground emission from a pulsar. The four-panel plot shows the delay transform of the LST-binned data at $\alpha = 115^\circ$ for two instrumental linear polarizations (top panels) and two pseudo-Stokes polarizations (bottom panels), displaying the high k features that appear to drop out in the pseudo-Stokes I visibility. The different colors represent four different spectral windows for the delay transform with a fixed bandwidth but with increasing central frequency, showing the frequency dependence of the systematic that is suggestive of Faraday rotation. The right plot shows a dirty rotation measure synthesis image from 150–170 MHz, clearly demonstrating a strong point source with a high rotation measure at image center. The artifacts around the central source are the grating lobes of the point-spread function.

beam-crossing timescale. These excesses decreased their characteristic delay at increasing frequency, suggestive of Faraday rotation. We investigated these systematics both in instrumental and pseudo-Stokes polarization visibilities, the latter of which are simply defined as a linear combination of the instrumental polarization visibilities as

$$\begin{pmatrix} V_{pI} \\ V_{pQ} \\ V_{pU} \\ V_{pV} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & i & -i & 0 \end{pmatrix} \begin{pmatrix} V_{EE} \\ V_{EN} \\ V_{NE} \\ V_{NN} \end{pmatrix}, \quad (6)$$

where the east–east (or EE) instrumental polarization is an east-facing feed correlated with another east-facing feed, and likewise for the north–north (or NN) polarization.

We are able to identify these systematics as Faraday-rotated foreground emission through a few pieces of evidence. First, the characteristic period of the frequency oscillations (i.e., the delay or k_{\parallel} mode of the systematic) decreases at higher frequencies. This is shown in Figure 9, where the four-panel plot shows the delay transform of the instrumental and pseudo-Stokes visibilities over four different spectral windows with increasing central frequency (but fixed bandwidth of 10 MHz), showing a notable decrease in the k_{\parallel} mode of the systematic at higher frequencies. In probing the relationship between the systematic’s peak k_{\parallel} and the central frequency of the spectral window, we find a tight relationship that follows the theoretical expectation of Faraday-rotated foregrounds found in Equation (3) of Moore et al. (2017; see also Jelić et al. 2010; Asad et al. 2015). Finally, we can perform a direct rotation measure synthesis of the data to look for known sources of foreground emission with high rotation measures in the field of view. Note that we make the assumption here that the inferred rotation measure is equal to the Faraday depth, which amounts to assuming that the rotation measure is locally compact along the line of sight. We do this by forming a spectral cube of dirty Stokes Q and U images and then take their rotation measure

synthesis for each pixel on the sky. This yields the Faraday depth distribution, written as

$$F(\phi) = \frac{1}{\pi} \int [Q(\lambda^2) + iU(\lambda^2)] e^{-2i\phi\lambda^2} d\lambda^2, \quad (7)$$

where Q and U are the Stokes Q and U maps, λ is the observing wavelength, and ϕ is the Faraday depth in radians meter $^{-2}$ (Brentjens & de Bruyn 2005; Kim et al. 2016). Note we do not actually form the true Stokes Q and U maps, but simply image the pseudo-Stokes Q and U visibilities, which is a good approximation of the Stokes parameters near the center of the FoV. Doing this and scanning through ϕ reveals a bright point source at the center of the field (right panel of Figure 9), which aligns exactly in location and in the inferred rotation measure with a known pulsar (PSR J0742-2822; Lenc et al. 2017). There are few other cases in the data where we see a high delay excess and can connect it to a known pulsar in the Faraday depth maps; however, they are seen at fairly low signal-to-noise ratio.

While these systematics appear above the noise in the EE, NN, pseudo-Stokes Q and U visibilities, they do not appear appreciably in the pseudo-Stokes I visibility (Figure 9). While this is expected, it is also true that intrinsic $Q \rightarrow I$ leakage from primary beam asymmetry between feeds can cause these systematics to appear at suppressed levels in the Stokes I power spectrum (Moore et al. 2013; Asad et al. 2016, 2018). Feed and LST jackknife tests could help to determine if future, deeper data sets are beginning to detect these systematics in Stokes I .

3.9. Coherent Time Averaging

While HERA observations are taken in drift-scan mode, integrations taken at nearby LSTs can be coherently averaged if they are phased to a common pointing center (Parsons et al. 2016). However, summing integrations that are separated too much in time will begin to decohere the averaged signal due to drift in the primary beam weighting, in addition to the inability

to properly re-align the fringes across the entire sky with a single fringe-stopping procedure. We evaluated the amount of loss induced by time averaging by generating an ensemble set of mock EoR visibilities and averaging them with increasingly large time windows. We settled on a maximum time window of 428 s, which leads to an average of $\sim 1\%$ decoherence of EoR power that is scale independent. In Aguirre et al. (2022), this test is repeated with a different kind of EoR model and they report a similar finding of $\sim 1\%$ decoherence in power. Recall that our power spectrum estimator cross multiplies adjacent time bins to avoid a noise bias. In order to ensure that data separated by more than 428 s are not cross-correlated, we actually halve the coherent averaging time window to 214 s.

3.10. Decoherence due to Nonredundancies

HERA's array layout has a large number of instantaneously redundant baselines. Averaging the visibilities directly rather than their power spectra provides an extra boost in sensitivity by a factor of $\sqrt{N_{\text{baselines}}}$, and is a key factor in HERA's overall sensitivity. The instrument is not perfectly redundant however—variations in baseline length and orientation, antenna primary beams, and even the calibration itself are expected at some level (Orosz et al. 2019). Deviations from perfect redundancy will cause the signal to decohere to some extent under redundant averaging, and if these deviations are large enough, some degree of signal loss will be sustained.

Here, we develop an empirical metric to quantify the amount of signal loss incurred due to decoherence when averaging slightly nonredundant visibilities. The impact of nonredundancy-based decoherence can be estimated by comparing the foreground power spectrum amplitude derived using coherent redundant averaging compared to incoherent redundant averaging. The incoherently averaged power spectrum, P_{inc} , is constructed by first estimating power spectra for each baseline pair in the redundant group and then averaging them together such that

$$P_{\text{inc}} = \frac{1}{N}(|\tilde{V}_1|^2 + |\tilde{V}_2|^2 + |\tilde{V}_3|^2 + \dots), \quad (8)$$

where V_1, V_2, \dots indexes individual baselines in a group, \tilde{V} is the Fourier transformed visibility, and N is the number of baselines in a group. Since phase information is discarded by the initial power spectrum estimation step, phase errors cannot combine destructively when the individual spectra are subsequently averaged. Thus, P_{inc} should be less susceptible to phase nonredundancies (e.g., caused by miscalibration or primary beam variations).

A cross-coherent power spectrum is constructed by cross multiplying all redundant baseline pairs and then taking their average, which does allow phase errors to combine destructively. Note that this is very similar to averaging the redundant visibilities and then cross multiplying them, except for the explicit exclusion of baseline terms paired with themselves. In other words, we form the cross-coherent power spectrum as

$$P_{\text{coh}} = \frac{1}{M}(\tilde{V}_1\tilde{V}_2^* + \tilde{V}_1\tilde{V}_3^* + \tilde{V}_2\tilde{V}_3^* + \dots), \quad (9)$$

where M is the number of baseline pairs in a group. In the ideal limit of perfect redundancy there should be no decoherence effect as each baseline is an exact replication of the signal. Thus the coherent and incoherent spectra should measure the

same value, with the exception of the different integrated thermal noise level that, as we explain later, is a negligible difference for the delay modes of interest. Note that this is not the estimator used to estimate the 21 cm power spectrum is later sections; this is only used as an approximate estimator for the purposes of assessing signal loss.

To measure the amount of decoherence induced by the coherent average we compute the fractional loss in power,

$$\Delta\chi(t, \tau) = \frac{P_{\text{coh}}(t, \tau) - P_{\text{inc}}(t, \tau)}{\langle P_{\text{inc}}(t, \tau) \rangle}, \quad (10)$$

where t denotes LST, τ is delay, and $\langle \dots \rangle$ denotes a time (LST) average. The time average is necessary to provide a stable baseline that limits spikes in the ratio when P_{inc} goes through a null. Note that the delay of the visibility, τ , is simply the direct Fourier dual to frequency in the Fourier transform, with units of 1 s^{-1} .

What we are interested in is determining how the EoR signal is suppressed in the coherently averaged power spectrum due to nonredundant decoherence. Being an isotropic cosmological signal, the EoR field is statistically invariant across the sky, meaning our sensitivity to it comes predominately from the main lobe of the primary beam. In order to use the measured foreground emission, which is not statistically isotropic, as a proxy for the amount of EoR signal loss, we will sharpen the metric in two ways. First, we only inspect the $\tau = 0$ ns mode, which isolates smooth spectrum foreground emission to a strip intersecting the main field of view of the primary beam. And second, we choose to evaluate the metric at times where diffuse foregrounds fill the main lobe of the primary beam, thus upweighting the parts of the beam where we expect the EoR signal to be most sensitively measured. For HERA, this occurs most prominently at the transit of the Galactic anticenter. The behavior of this metric has been explored in an accompanying paper (Choudhuri et al. 2021). In this study, numerical simulations of a small HERA-like array were performed with different types of primary beam nonredundancies and calibration nonredundancies. The $\Delta\chi$ foreground metric was found to accurately reflect the degree of nonredundancy, signal loss, and modulation effects in the recovered EoR power spectrum. Its behavior in the presence of baseline nonredundancies was not studied.

Figure 10 shows the $\Delta\chi$ metric at $\tau = 0$ ns for several redundant baseline groups. We see an average of $\sim 1\%$ power loss, suggesting that the real nonredundancies between baselines within a redundant group do not significantly decohere the sky signal. The worst case is the $b = 50.6$ m baseline, which sees an average decoherence of about 3%. Inspecting the visibilities of this baseline we see that this is due in part because this baseline experiences a particularly strong null in power compared to the other baselines, which systematically brings down $\langle P_{\text{inc}} \rangle$ in the denominator of the metric, causing the overall ratio to slightly inflate.

3.11. Other Forms of Signal Loss in the Pipeline

Here, we tabulate all of the identified forms of systematic signal loss that arise from the analysis pipeline. Recall that the philosophy of the analysis employed here is one that is generally meant to minimize the inherent loss of the analysis pipeline; therefore, corrections to the loss that have been

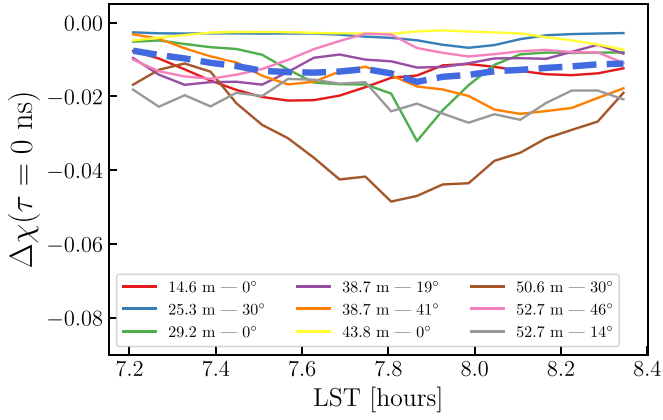


Figure 10. Redundancy decoherence test for nine redundant groups, marked by the baseline length and angle in local array XYZ coordinates (blue dashed is their average). This plots the difference of the coherently and incoherently averaged power spectrum normalized by the time average of the latter. We show the metric for the $\tau = 0$ Fourier mode at LSTs when bright, diffuse foregrounds fill the primary beam. On average, we see a roughly 1% power loss, suggesting that while our final 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.

identified are generally a small fraction of the total measured power.

Some forms of loss are entirely expected, like from the time averaging of drift-scan visibilities, and other forms were not expected but found in the process of pipeline validation (Aguirre et al. 2022), the most notable being the absolute calibration bias. This bias was due to the logarithmic linearization of the gain amplitude parameter in the process of absolute gain calibration, which when employed on visibilities with low signal-to-noise ratios can result in a bias in the recovered amplitude (Boonstra & van der Veen 2003). Because we calibrate HERA’s drift-scan observations every 10.7 s, there are some fields where this bias can be nonnegligible and reaches into the percent level on the gains. After gain smoothing, this bias is shown to be largely time independent but does have a slight frequency dependence (Aguirre et al. 2022). We estimate this bias as a single number for each of the two power spectrum bands (Section 4.3) by making multifrequency synthesis images of the calibrated and LST-binned visibilities and comparing them to images of the initial absolute flux density model. The ratio of a zenith-located point source in the data and model is used as the correction factor.

All forms of loss measured here are EoR model independent and scale independent (i.e., flat across k). In some cases we have physically motivated reasons to expect this (e.g., an overall bias in absolute calibration is scale and EoR model independent) and in other cases we have explicitly tested this (as is the case for coherent time averaging and the nonredundancy decoherence discussed above). These are tabulated in Table 2 and are each corrected for at the end stage by correcting the power spectra.

4. HERA Phase I Power Spectrum Limits

In this section we outline our power spectrum estimator, giving a brief overview of quadratic estimators (QEs) and our power spectrum pipeline, and present our derived power spectrum limits from the analysis discussed in this work. Our power spectrum

Table 2
Systematic Loss in Analysis Pipeline

Analysis Step	Fractional Power Loss
Absolute Calibration	9% (8%)
Cross-coupling Filtering	3% (1%)
LST Time Averaging	1% (1%)
Redundant Averaging	1% (1%)

Note. Percentage loss in power for Band 1 (Band 2), which are corrected for after forming the power spectrum. These figures are partially derived from the validation analysis (Aguirre et al. 2022) and from this analysis (Figure 10).

estimator code base, `hera_pspec`,³⁰ is publicly available software. Throughout this work, we adopt a Λ CDM cosmology (Planck Collaboration et al. 2016) with $\Omega_\Lambda = 0.6844$, $\Omega_b = 0.04911$, $\Omega_c = 0.26442$, and $H_0 = 67.27 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

4.1. The Quadratic Estimator Formalism

The 21 cm power spectrum is the square of the three-dimensional spatial Fourier transform of the 21 cm brightness temperature field. Given that in the flat-sky limit the interferometric visibility is the 2D transverse spatial Fourier transform of the temperature field, we can estimate the 3D 21 cm power spectrum by taking the Fourier transform across frequency,

$$\tilde{V}(\tau) = \int V(\nu) e^{2\pi i \nu \tau} d\nu, \quad (11)$$

where the delay domain (τ) is the Fourier dual of frequency. Note that delay modes are not a direct mapping to the line-of-sight k_\parallel wavevector; however, for short baselines they are nearly the same and approximating τ modes as k_\parallel modes is known as the delay approximation (Parsons et al. 2012; Liu et al. 2014a). This is written as

$$\hat{P}(k_\perp, k_\parallel) = \frac{X^2 Y}{\Omega_{pp} B} \tilde{V}_1(u, \tau) \tilde{V}_2^*(u, \tau), \quad (12)$$

where \hat{P} is our estimate of the power spectrum, X and Y are scalars mapping sky angles and frequency to cosmological distances, respectively, Ω_{pp} is the sky integral of the squared primary beam response, and B is the Fourier transform bandwidth. Furthermore, the subscript on our visibilities, \tilde{V}_1 and \tilde{V}_2 , indicates that we are cross multiplying visibilities with independent noise realizations, meaning that the estimated power spectrum has no noise bias. As written, the power spectrum P has units $\text{mK}^2 h^{-3} \text{ Mpc}^3$. The cosmological wavevectors are related to the natural telescope units of delay and baseline length (Parsons et al. 2014) via

$$k_\parallel = \frac{2\pi\tau}{X} \quad (13)$$

$$k_\perp = \frac{2\pi}{Y} \frac{b}{\lambda}, \quad (14)$$

where $X = c(1+z)^2 \nu_{21}^{-1} H(z)^{-1}$, $Y = D(z)$, $\nu_{21} = 1.420 \text{ GHz}$, $H(z)$ is the Hubble parameter, and $D(z)$ is the transverse

³⁰ https://github.com/HERA-Team/hera_pspec

comoving distance (Hogg 1999; Condon & Matthews 2018). We can further define the cosmologically dimensionless power spectrum,

$$\Delta^2(k) = P(k) \frac{k^3}{2\pi^2}, \quad (15)$$

which has units of mK^2 .

The “Fourier transform and square” delay spectrum estimator can also be cast into the more general form of a QE,

$$\hat{q}_\alpha = \frac{1}{2} \mathbf{x}_1^\dagger \mathbf{R}^\dagger \mathbf{Q}_\alpha \mathbf{R} \mathbf{x}_2, \quad (16)$$

where \hat{q}_α is the unnormalized estimate of the α bandpower, \mathbf{x}_1 and \mathbf{x}_2 are the visibility vectors we will cross-correlate, \mathbf{R} is any weighting matrix we may want to apply to the data (e.g., the inverse covariance in the optimal QE case), and \mathbf{Q}_α is the mapping from data space to power spectrum space. In general, this mapping is given by the derivative of the covariance with respect to bandpower α , i.e., $\mathbf{Q}_\alpha = \mathbf{C}_{,\alpha} = \frac{\partial \mathbf{C}}{\partial p_\alpha}$. We use \mathbf{C} to mean the true covariance matrix of the visibilities, which in general is never known exactly. Estimating the covariance empirically is a subtle process, and if not done accurately risks biasing the power spectrum estimate (Ali et al. 2018; Cheng et al. 2018). The lack of access to the true covariance also makes accurate error estimation tricky, as we will discuss in more detail later in this section. For our purposes, we simply define $\mathbf{Q}_\alpha = \mathbf{c}_\alpha^\dagger \mathbf{c}_\alpha$, where \mathbf{c}_α is a discrete Fourier transform operator that takes the (weighted) data to delay space. We denote our data vectors as \mathbf{x}_1 and \mathbf{x}_2 because in this work we opt to form cross power spectra between data vectors with independent noise realizations, which eliminates the noise bias term that can be difficult to estimate precisely (Dillon et al. 2014; Ali et al. 2015; Pober et al. 2016).

To normalize our power spectrum estimate we gather our unnormalized bandpowers into a vector and form the quantity,

$$\hat{\mathbf{p}} = \mathbf{M} \hat{\mathbf{q}}, \quad (17)$$

where \mathbf{M} is the normalization matrix and $\hat{\mathbf{p}}$ is the normalized bandpower estimates. Our estimated power spectra are related to the true bandpowers by the window function matrix,

$$\hat{\mathbf{p}} = \mathbf{W} \mathbf{p}, \quad (18)$$

where \mathbf{p} is a vector holding the true bandpowers. While \mathbf{p} is never known a priori, we can use the window function matrix to appropriately map a theoretical EoR power spectrum to the basis of the measured power spectrum. Relating this to our normalization matrix we have

$$\mathbf{W} = \mathbf{M} \mathbf{H}, \quad (19)$$

where \mathbf{H} is the response matrix of the bandpowers, defined as

$$H_{\alpha\beta} = \frac{1}{2} \text{tr}(\mathbf{R}^\dagger \mathbf{Q}_\alpha \mathbf{R} \mathbf{Q}_\beta). \quad (20)$$

In the case where $\mathbf{R} = \mathbf{C}^{-1}$, our estimator becomes the optimal estimator, and \mathbf{H} becomes \mathbf{F} , the bandpower Fisher matrix (Tegmark 1997).

The last component of our power spectrum formalism is the bandpower covariance matrix,

$$\Sigma = \text{Cov}[\hat{\mathbf{p}}] = \langle \hat{\mathbf{p}} \hat{\mathbf{p}}^\dagger \rangle - \langle \hat{\mathbf{p}} \rangle \langle \hat{\mathbf{p}} \rangle^\dagger. \quad (21)$$

Defining $\mathbf{E}^\alpha = \frac{1}{2} \sum_\beta M_{\alpha\beta} \mathbf{R}^\dagger \mathbf{Q}_\beta \mathbf{R}$ such that $\hat{p}_\alpha = \mathbf{x}_1^\dagger \mathbf{E}^\alpha \mathbf{x}_2$ and substituting this in, we get that the power spectrum covariance is

$$\Sigma_{\alpha\beta} = \text{tr}[\mathbf{C} \mathbf{E}^\alpha \mathbf{C} \mathbf{E}^\beta] + \text{tr}[\mathbf{S} \mathbf{E}^\alpha \mathbf{S} \mathbf{E}^\beta], \quad (22)$$

where the total covariance, $\mathbf{C} = \langle \mathbf{x} \mathbf{x}^\dagger \rangle = \mathbf{S} + \mathbf{N}$, can be written as a sum of the sky signal and noise covariance (Dillon et al. 2014). While the noise covariance of our data is straightforward to quantify, the signal terms are considerably harder. We will come back to how we estimate the bandpower errors in practice.

An unbiased estimate of the power spectrum is made by building a window function matrix whose rows sum to unity, giving the analyst some freedom on exactly how to construct \mathbf{M} , with statistical implications for $\hat{\mathbf{p}}$ and Σ (Tegmark 1997; Liu & Tegmark 2011; Ali et al. 2015). Choosing $\mathbf{M} = \mathbf{H}^{-1}$, for example, yields $\mathbf{W} = \mathbf{I}$, meaning the bandpowers are independent; however, we also see that the bandpower errors become quite large and correlated. In this work we choose arguably the simplest approach, which is to set \mathbf{M} to a diagonal matrix. This minimizes the resultant error bars at the expense of measurements that overlap in Fourier space and also have slightly correlated errors. This returns us to the simple delay spectrum estimator, but framed in the machinery of a QE, with the diagonal of \mathbf{M} given by the coefficients of Equation (12).

Our QE is suboptimal in the sense that we do not weight our data by the inverse covariance matrix; however, this also safeguards our estimator against concerns about signal loss from empirical inverse covariance weighting (Cheng et al. 2018) and the distortion of the window functions at low k modes due to complicated data weighting (Liu et al. 2014b; Kern & Liu 2021). In the meantime, we defer optimal inverse covariance weighted power spectrum estimation to future work. Nonetheless, we still need to account for the fact that, without foreground removal, there exists a large dynamic range between the foreground signal at low k modes and the EoR and noise signal at higher k modes. To limit foreground spectral leakage in the discrete Fourier transform, we apply a tapering (or apodization) function along the diagonal of \mathbf{R} . We use a Blackman–Harris function (Blackman & Tukey 1958), which achieves 50 decibels of sidelobe suppression in Fourier space.

4.2. Error Bar Methodology

Tan et al. (2021) discuss different kinds of error bars one can use to quantify uncertainty on the measured bandpowers. In summary, they find that all of the error bar methodologies explored are in general agreement with each other and with the true sampling distribution of the bandpowers in the ensemble limit; however, they suggest two specific error bars that: (1) encapsulate the full thermal noise contribution to the bandpower uncertainty, and (2) can be computed relatively straightforwardly without requiring a model of the signal covariance. We will briefly summarize these here.

At minimum, one needs to encapsulate the fundamental thermal noise uncertainty of the fully averaged power spectrum, which is the root-mean-square (rms) of the bandpowers in the limit that they are noise dominated. However, as Tan et al. (2021) point out, there is an additional source of noise variance on the bandpowers in the limit of a

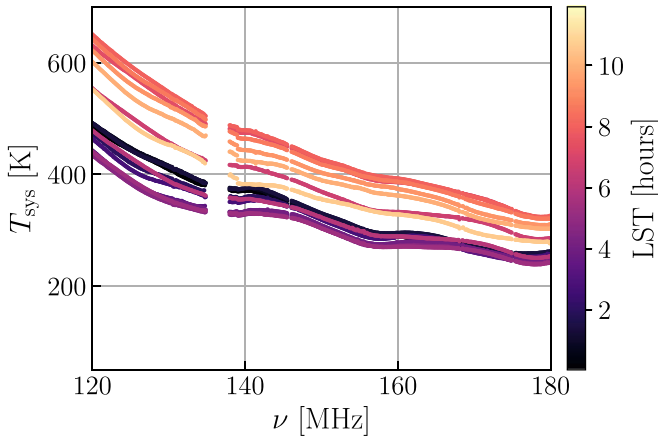


Figure 11. A frequency and LST dependent model for the system temperature of a particular antenna that is used to compute power spectrum error bars and also to generate realistic thermal noise realizations of the data.

nonnegligible coherent signal in the data.³¹ For weak, marginally detectable signals the additional variance is a small correction; however, when there is a significant detection of the signal this correction dominates the thermal noise uncertainty on the bandpowers, and thus should be accounted for.

An analytic expression for the root-mean-square (rms) of noise-limited power spectra after a given amount of averaging can be written as

$$P_N = \frac{X^2 Y \Omega_{\text{eff}} T_{\text{sys}}^2}{t_{\text{int}} N_{\text{coherent}} \sqrt{2 N_{\text{incoherent}}}}, \quad (23)$$

where t_{int} is the integration time of the data, N_{coherent} is the number of coherent averages of the data (i.e., visibility averages), $N_{\text{incoherent}}$ is the number of incoherent averages (i.e., averaging after forming the power spectra), and T_{sys} is the system temperature of the signal chain, which is the sum of the sky and receiver temperatures. Lastly, Ω_{eff} is the effective beam area, defined as $\Omega_{\text{eff}} = \Omega_p^2 / \Omega_{pp}$, where Ω_p is the sky integral of the primary beam and Ω_{pp} is the sky integral of the squared beam (Pober et al. 2013a; Parsons et al. 2014; Cheng et al. 2018). A frequency and LST dependent model for T_{sys} is derived from the auto-correlation visibilities of each antenna to produce a baseline, time, and frequency dependent noise model (e.g., Dillon et al. 2020; Kern et al. 2020a; Tan et al. 2021). Figure 11 shows that HERA measures a system temperature in the range of 200–400 K at 160 MHz depending on the LST. Note that the T_{sys} models are used not only to compute the error bars on the power spectrum, but also to generate realistic noise simulations of the data that are used for null testing.

The second error bar that Tan et al. (2021) investigate accounts for the additional sources of noise in the power spectrum that arise in the presence of a signal, also derived in Kolopanis et al. (2019). These extra terms come simply from the signal–noise cross terms when squaring the visibilities to form the power spectrum. This error bar is written as

$$P_{\text{SN}} = \sqrt{\sqrt{2} P_S P_N + P_N^2}, \quad (24)$$

³¹ Note that this is entirely separate from cosmic variance uncertainty, and also by signal we mean any coherent signal in the data, be it the EoR, foregrounds, or systematics.

where P_S is the power spectrum of the signal in the data. Note that, like P_N , P_{SN} can be thought of as the 1σ uncertainty on the measured bandpower. What is readily apparent is that to compute this quantity we need the power spectrum of the signal, which we do not have any more than we have the covariance of the signal. In that case, we can estimate it directly from the data, using \hat{P} in its place, which Tan et al. (2021) show is a good approximation. However, for noise-limited bandpowers where $P_S \ll P_N$ we see that by substituting P_S for \hat{P} we are actually overestimating the error bar (in a sense we are double counting the noise–noise contribution). Nevertheless, Tan et al. (2021) show that this can be corrected by constructing a modified error bar estimator

$$\tilde{P}_{\text{SN}} = \sqrt{\sqrt{2} P_S P_N + P_N^2} - (\sqrt{1/\sqrt{\pi}} + 1 - 1) P_N. \quad (25)$$

One downside to Equations (23), (24), and (25) is that they make certain simplifying assumptions. Mainly, they assume that the bandpowers are uncorrelated. In other words, they only account for the diagonal of the bandpower covariance. However, we expect the off-diagonal components to possibly have nonnegligible covariance (specifically due to the apodization function) and it should be accounted for. Tan et al. (2021) show how the error bar estimators presented above can also be derived by propagating a combination of the data and the frequency–frequency noise covariance through the QE, which conveniently yields the full bandpower covariance matrix, not just its diagonal. Therefore, to quantify the off-diagonal components in the bandpower covariance matrix, we use the frequency–frequency noise covariance of the data propagated through the QE formalism to Σ .

When quoting upper limits on the power spectrum or presenting power spectrum measurements, we will exclusively use the modified \tilde{P}_{SN} error bar as the 1σ uncertainty, which is also the 68% confidence interval given that we expect the bandpowers to be Gaussian distributed (see Tan et al. 2021). We will also generally plot the P_N limit as a reference for whether the data are consistent with the thermal noise floor. Lastly, we also use P_N and the bandpower noise covariance matrix when evaluating statistical null tests in Section 5.

4.3. Forming Power Spectra

We form power spectra from the pseudo-Stokes I visibilities (Equation (6)) after LST binning, systematics treatment, and coherent time averaging. We choose two spectral windows over which to form power spectra spanning 117.1–132.6 MHz and 150.3–167.8 MHz, corresponding to redshifts of 10.4 and 7.9, respectively. We refer to these as Band 1 and Band 2, shown in Figure 12, which were selected based on their relatively low flag occupancy of a few percent. The shaded curves show the extent of the Blackman–Harris tapering function applied along the spectral window of each band. Evolution of the cosmological signal will begin to affect the power spectrum over large line-of-sight bandwidths, also known as lightcone effects. Although this effect is technically EoR model dependent, it has been found that for $\Delta\nu < 10$ MHz, lightcone effects on the power spectrum are kept below 10% for $k < 0.5 h \text{ Mpc}^{-1}$ and $7 < z < 10$ (Datta et al. 2014). Note that because we apply a Blackman–Harris tapering function across each of our spectral windows, their

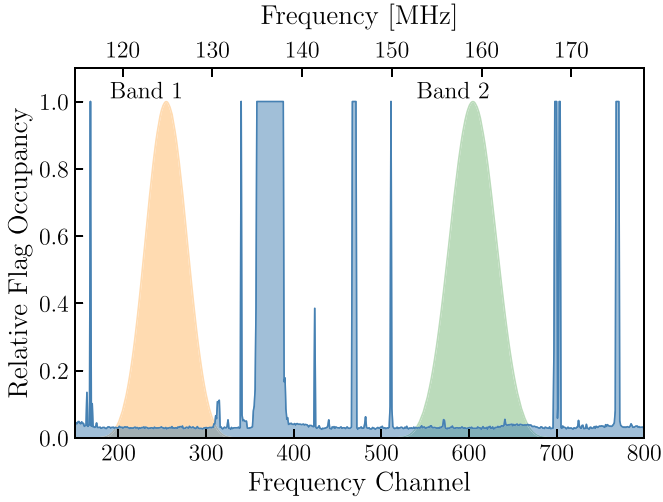


Figure 12. The average flag occupancy of each frequency channel across LST (blue shaded), showing the two spectral windows used in the power spectrum analysis. The orange and green shaded regions show the tapering function applied across each spectral window for Band 1 and Band 2, which have central frequencies of 124.8 and 159.0 MHz, corresponding to redshifts of 10.4 and 7.9, respectively.

effective bandwidths are 7.75 and 8.75 MHz for Band 1 and Band 2, respectively.

To form power spectra following Equation (16), we cross multiply adjacent time integrations (Pober et al. 2013b) separated by 214 s after time averaging between all baselines in a redundant set. A redundant set is defined by all physical baselines that share a common length and orientation. Note that we compute the power spectrum of all baseline permutations (i.e., we compute $V_1 \times V_2$ and $V_2 \times V_1$, where V_1 and V_2 are visibilities in the same redundant set); however, we do not compute the power spectrum of a baseline paired with itself (i.e., $V_1 \times V_1$), which reduces the impact of baseline-dependent systematics. This achieves nearly the full sensitivity of a coherent visibility average across redundant baselines, thus the need to account for signal loss from a coherent redundant average (Section 3.11).

Next, we motivate the choice of the three unique “fields” in LST over which we form power spectra. Tracking arrays like the GMRT, the MWA, and LOFAR can pick out specific parts of the sky where foregrounds emission is minimal (Ghosh et al. 2012; Mertens et al. 2020; Trott et al. 2020). However, HERA is not afforded this luxury because it is a static array that observes a uniform stripe across the sky. Nonetheless, we can pick out certain LSTs where foreground emission is relatively less bright. This is important because we know that systematics like RFI and residual instrumental gains act to leak foregrounds modes in the power spectrum to higher k_{\parallel} , and thus partially contaminate the EoR window. Furthermore, as noted in Section 3, we do not perform any kind of foreground subtraction in this work. Therefore, picking LSTs that avoid the brightest foregrounds can help to minimize the impact of residual systematics. Specifically, we pick three fields, each spanning ~ 2 hr in LST that avoid the extended radio galaxy Fornax A at an R.A. of 3.36^{h} and the Galactic anticenter at $\sim 7.5^{\text{h}}$ (Figure 1). They also give a 0.5 hr buffer at the ends of the total LST range to limit edge effects in our fringe-rate filtering. The chosen fields span an LST range of 1.25–2.7 hr, 4.5–6.5 hr, and 8.5–10.75 hr for fields 1, 2, and 3, respectively.

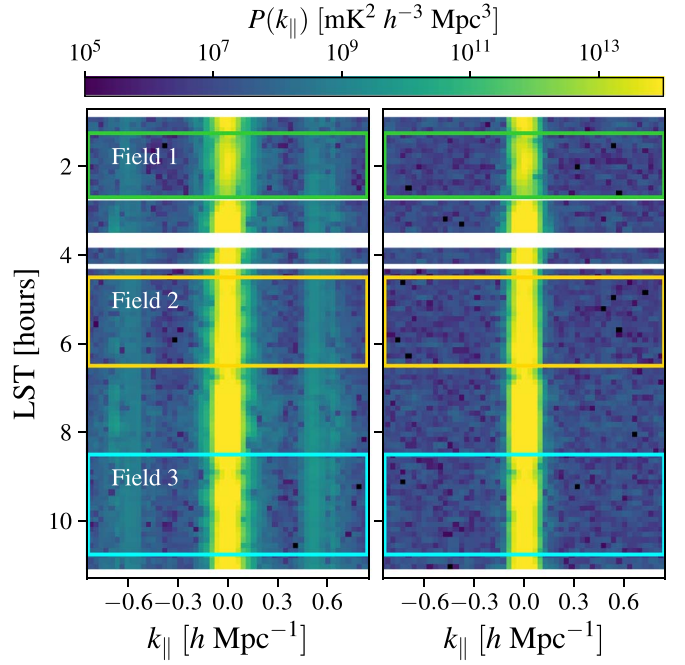


Figure 13. A waterfall of a redundantly averaged power spectrum for a single baseline group without systematic treatment (left) and with systematic treatment (right). The LST cuts for each of the three fields are shown in blue, orange, and green, which are chosen to avoid the transiting of bright foreground sources like Fornax A at 3.36^{h} , the Galactic anticenter at $\sim 7.5^{\text{h}}$, and to give a buffer of ~ 30 minutes from both the beginning and end of the full LST range.

Figure 13 shows a redundantly averaged power spectrum waterfall for a single redundant set with and without systematics treatment (Section 3.8) over the full LST range, with the three fields marked. Based on the amplitude of the $k_{\parallel} = 0 \text{ h Mpc}^{-1}$ mode, we can see that Field 1 has the least amount of overall foreground power, making it an ideal candidate for power spectrum analysis. Figure 13 also demonstrates the ~ 2 orders of magnitude of systematic suppression achieved by our pipeline, which would otherwise contaminate a large portion of the EoR window modes.

Next, we incoherently average the power spectra over the redundant baseline axis and across the remaining time bins in each field. This is equivalent to a cylindrical binning in \mathbf{k} space onto a k_{\parallel} and k_{\perp} plane. All averaging is weighted by the squared inverse of the computed thermal noise floor, P_N , which is both a function of baseline pair and time but is independent of k_{\parallel} . Simultaneously, we also propagate the bandpower covariance and the \tilde{P}_{SN} error bar through the averaging. This leaves us with a 2D power spectrum for each band and field, shown in Figure 14 for Band 1 and Figure 15 for Band 2. In each figure the top panels show the real component of the power spectrum, with negative measured bandpowers shown in white, while the bottom panels show the ratio of the power spectrum with the thermal noise floor. Even though the power spectrum of a coherent signal is nonnegative by construction, the measured power spectrum can fluctuate negative due to the fact that we have cross-correlated data with independent noise realizations, which is a mean-zero process with nonzero variance.

The root-mean-square of the 2D power spectra divided by their thermal noise floors for $k_{\parallel} > 0.2 \text{ h Mpc}^{-1}$ is consistent with one to within $\sim 3\%$, which means that (1) we are able to

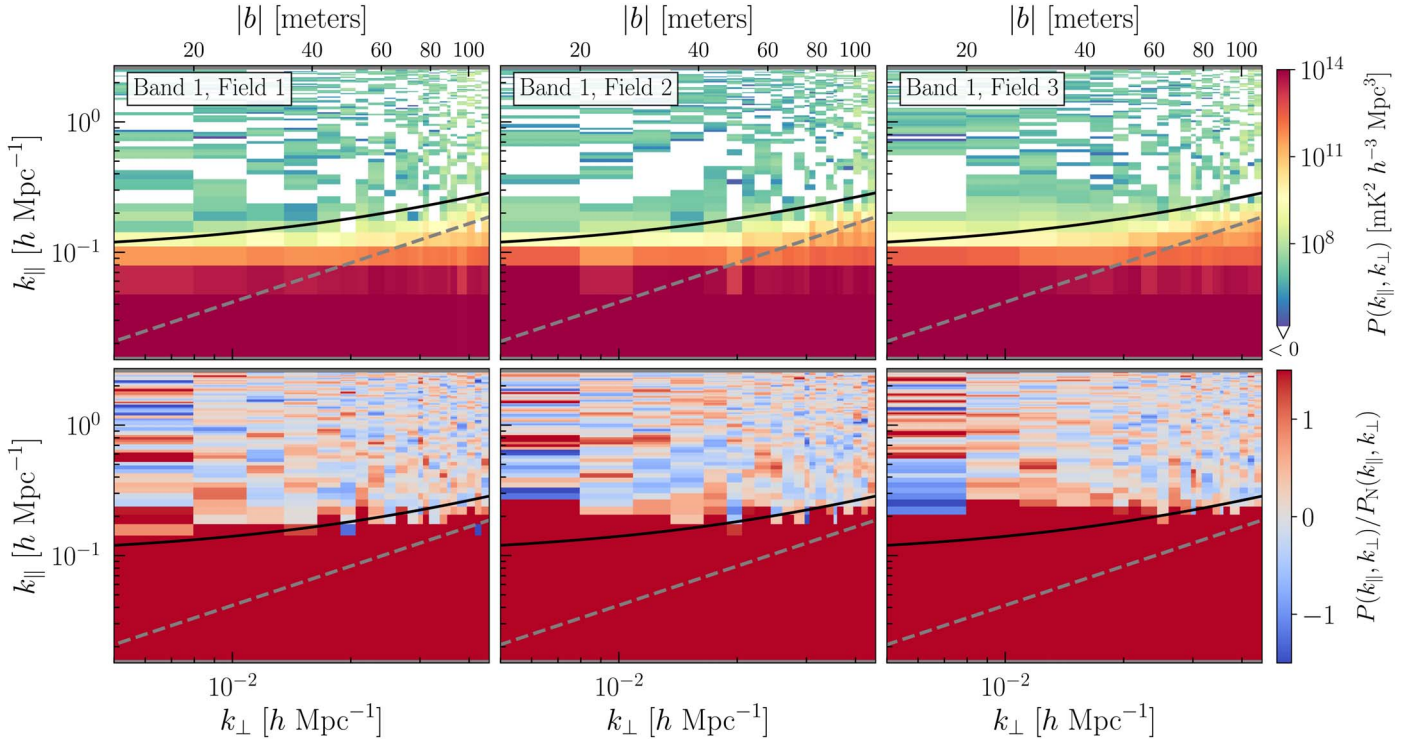


Figure 14. The 2D power spectra $P(k_{\parallel}, k_{\perp})$ for Band 1 ($z = 10.4$) at each field (top panels), and their ratio with the 1σ thermal noise floor (bottom panels). We also plot the horizon wedge (dashed) and the horizon buffer used when spherically binning (solid).

estimate the noise variance in the real data at a precision of a few percent, and (2), as we will also demonstrate later, the data at high k modes are largely noise dominated. The dashed line plots the foreground horizon limit, which is the theoretical maximum k_{\parallel} for smooth spectrum foregrounds, and the solid line plots the horizon limit plus a buffer of 200 nanoseconds ($\Delta k_{\parallel} = 0.11 h \text{ Mpc}^{-1}$ for $z = 7.9$ and $\Delta k_{\parallel} = 0.10 h \text{ Mpc}^{-1}$ for $z = 10.4$). As we will discuss below, this cut at the horizon limit plus a buffer helps to reduce foreground contamination in the final spherically averaged power spectrum. The exact buffer was chosen to mitigate foreground leakage beyond the horizon limit (particularly at large k_{\perp} modes) while keeping some nonzero weight at the lower k_{\parallel} modes, even though there is some observed leakage beyond this buffer at low k_{\perp} (Figures 14 and 15).

The fact that the data are in rough agreement with the expected noise at high k_{\parallel} , especially given that we have performed no foreground removal, is a testament to both the instrument design as well as the data reduction algorithms, which have been able to effectively contain foreground emission to the lowest k_{\parallel} modes within the foreground wedge. Figure 16 shows this more clearly, showing the spherically averaged power spectrum (discussed below) without enacting a low k_{\parallel} cut and normalizing by the peak foreground power at $k = 0 h \text{ Mpc}^{-1}$. This shows that the full process of measuring the sky brightness, including the instrument itself and our analysis pipeline, achieves a dynamic range of 10^9 in power between the peak foreground emission and the thermal noise floor at high k , which is a necessary but not quite sufficient criterion for an eventual 21 cm signal detection. Note that a complementary analysis of Phase I data using the bispectrum phase also achieved a dynamic range of 10^8 between peak foreground power and the recovered noise floor (Thyagarajan et al. 2020), but used less data than what is presented here.

However, while Figures 14 and 15 show decent agreement with the noise floor at high k_{\parallel} (a point we come back to in Section 5), we also clearly see evidence of foreground emission at k_{\parallel} beyond that determined by the horizon delay (termed suprahorizon emission or foreground leakage), particularly at low k_{\perp} . Some of this is simply due to the frequency tapering function’s footprint in k_{\parallel} space, which pushes $k_{\parallel} \approx 0$ foreground emission to higher k_{\parallel} ; however, that cannot explain the full extent of the suprahorizon emission. Kern et al. (2020b) speculate that this could be due to residual instrumental cross-coupling, which is both generally most prominent and hardest to filter off at the low k_{\perp} modes.

Looking at Band 2, Field 2, we also see evidence for a slight excess beyond the foreground horizon and buffer. While it is unclear exactly what this is due to, one possibility is that it is a low-level spectral artifact (possibly unflagged RFI) that exhibits a baseline dependence, which has been seen before at a low-level in Phase I data. Future, more comprehensive work detailing low-level spectral artifacts in the fully averaged data, for example with the SSINS algorithm (Wilensky et al. 2019), may help us better understand these features.

4.4. Integrated Limits on the Power Spectrum

Next, we spherically average the data by binning the 2D power spectra in bins of constant $|k|$, assuming statistical isotropy of the cosmological signal, and again weighting the average by the squared inverse of $P_N(k_{\parallel}, k_{\perp})$ in each pixel. We compute the spherically averaged power spectrum and propagate our measures of uncertainty using Equations (33)–(38) of Dillon et al. (2014). Recall that we give zero weight to all pixels with $k_{\parallel} < k_{\text{horizon}} + \Delta k_{\parallel}$, where Δk_{\parallel} is the foreground horizon buffer. This leaves us with three spherically averaged power spectra at each field for Band 1 and Band 2, which we

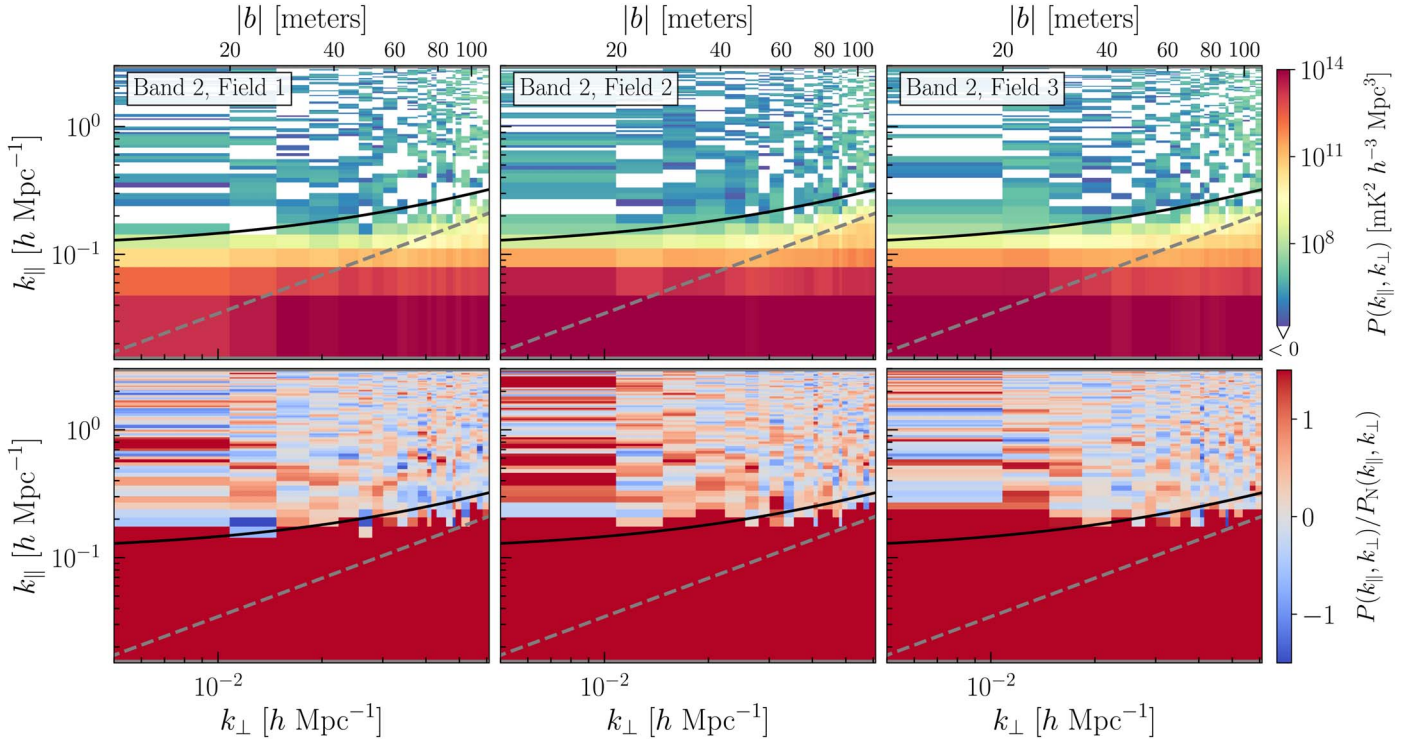


Figure 15. The same 2D power spectra as Figure 14 but for Band 2 ($z = 7.9$).

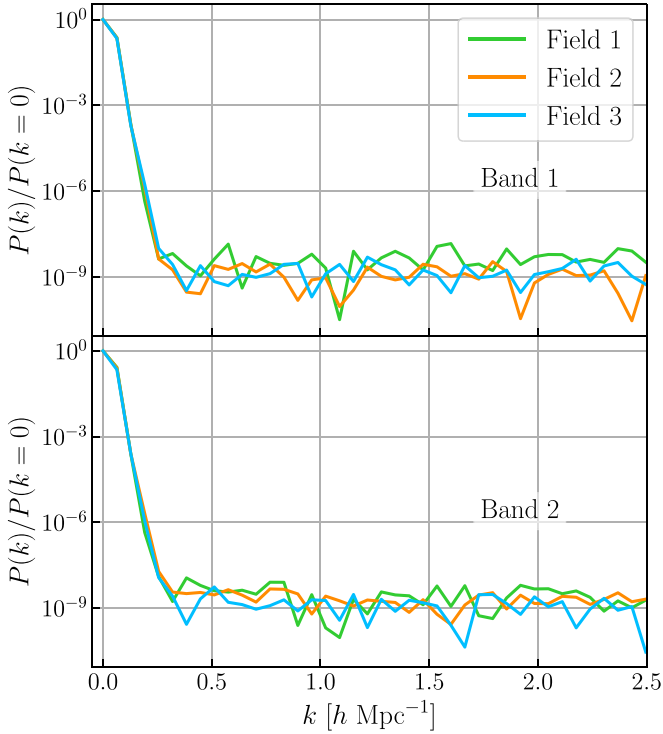


Figure 16. The amplitude of the spherically binned power spectrum without enacting a minimum k_{\parallel} cut in the binning, normalized to the foreground power at $k = 0 \text{ h Mpc}^{-1}$. This demonstrates that our analysis pipeline achieves a dynamic range of $\sim 10^9$ with respect to the peak foreground power.

show in Figure 17. The vertical error bars are the $2\sigma P_{\text{SN}}$ error bar discussed in Section 4.2 and Tan et al. (2021), and the horizontal error bars are the 16th and 84th percentiles of the window functions. We also plot the theoretical 1σ thermal noise floor (dashed), which is simply the P_N error bar discussed

in Section 4.2. Bandpowers that are measured as negative are set to zero. While coherent sky signal manifests as a purely positive signal in the power spectrum, thermal noise and other uncorrelated systematics are mean-zero with nonzero variance, which can manifest as negative values in the power spectrum. Note that in the limit that the data are nearly noise dominated, which is true for most of the data shown in Figure 17, the 1σ error bar is the same as the $1\sigma P_N$ dashed line.

Broadly, we can evaluate whether the data are consistent with thermal noise fluctuations if (1) the vertical error bars reach $\Delta^2 = 0 \text{ mK}^2$ for the majority of the data points, (2) the measured bandpowers are roughly consistent with the P_N curve, and (3) we measure a roughly equal number of positive and negative bandpowers. However, all of this is complicated by the fact that the fully averaged products reduce us to a regime of low number statistics, in addition to the fact that there will be some amount of correlation between the bandpowers in k , in part due to the frequency tapering function (the bandpowers are independent across bands and fields). With just a qualitative assessment for the time being, however, we can see general agreement of the data with the expected thermal noise level at high k modes, and increasing discrepancy at low k modes.

The discrepancy at low k is easily explained as residual foreground leakage that can be directly observed in Figures 14 and 15. The fact that this leakage is stronger for Band 1 is largely due to the fact that the foregrounds are brighter at low frequencies. Furthermore, we see that Field 1 exhibits the least amount of foreground leakage at the low k modes, which is also largely due to the fact that it sees relatively dimmer foregrounds than the other fields (Figures 1 and 13). Furthermore, we also see that Band 2, Field 2 sees a systematic excess in power compared to thermal noise expectation, which is likely due to a low-level time and frequency dependent systematic. We speculate that this could be the effect of low-

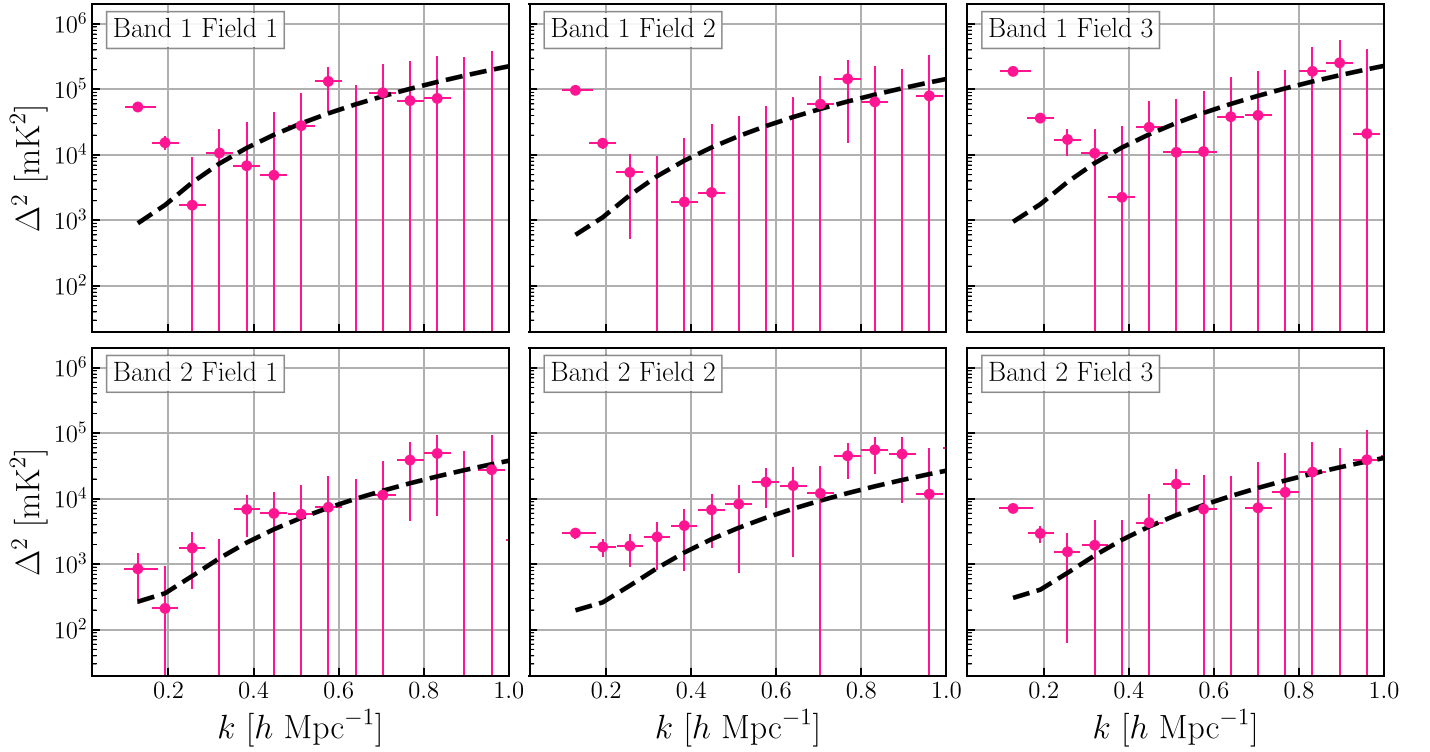


Figure 17. Spherically averaged, dimensionless power spectra for Band 1 ($z = 10.4$) and Band 2 ($z = 7.9$) at each of the three fields. The vertical error bars are 2σ , while the horizontal error bars are the 16th and 84th percentiles of the window functions. The theoretical thermal noise floor (dashed) is also shown, which represents the 1σ noise floor. Bandpowers that are measured to be negative are set to zero. The most sensitive limits come from Field 1 (for both Band 1 and Band 2), which is largely due to the fact that Field 1 has the least amount of foreground power in the field of view.

level RFI, due in part to its spectral and temporal transience; however, more comprehensive work identifying low-level RFI is required to understand this at a deeper level.

The bandpowers in Figure 17 are sampled at a cadence of $\Delta k = 0.64 \, h \, \text{Mpc}^{-1}$, which is twice the native spacing of the Fourier modes given our choice of spectral windows. This helps to reduce the correlations between neighboring bandpowers, especially given that we applied a tapering function across frequency before estimating the power spectra. This tapering impacts both the resultant bandpower window functions as well as the bandpower covariance matrix. Figure 18 shows the Band 2, Field 1 power spectrum with its associated window functions (bottom panel), which shows that they are well behaved even without decorrelation; however, they do show low-level overlap between neighboring modes, which in principle should be taken into account when comparing the data to astrophysical models. The window functions are nearly identical for all measured modes,³² and are also nearly identical between Band 1 and Band 2. The 16th and 84th percentiles of the window functions are $\Delta k = 0.031 \, h \, \text{Mpc}^{-1}$ away from their 50th percentile, which is a rough approximation of the 1σ horizontal error bar of the bandpowers.

Upper limits on the 21 cm power spectrum (at a 95% confidence level) are constructed by taking the measured dimensionless power spectrum, Δ^2 , and adding the 2σ error bar. In the case where the measured bandpower is negative, we set it to zero, which is justified by our assertion that the cosmological signal (or any coherent signal in the visibilities

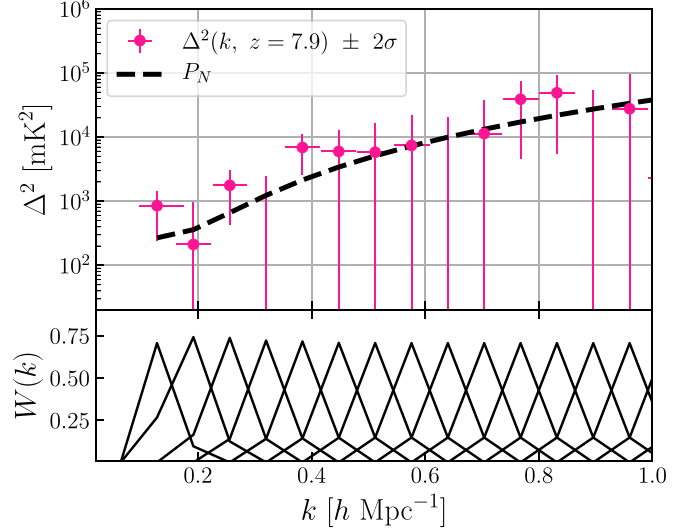


Figure 18. The spherically averaged, dimensionless power spectrum from Band 2 and Field 1, showing 2σ vertical error bars, their associated window functions (lower panel), and the theoretical thermal noise floor (dashed). The horizontal error bar is the distance of the 16th and 84th percentiles of the window function from its 50th percentile, equal to $\Delta k = 0.031 \, h \, \text{Mpc}^{-1}$ for all the bandpowers except the lowest k mode, which is slightly asymmetric due to the horizon buffer's effect on spherical binning.

for that matter) is a purely real and positive quantity in the power spectrum by definition. In this case, the upper limit simply becomes the 2σ error bar.

The measured bandpowers, their uncertainties, and the derived upper limits are tabulated in Tables 3 and 4. The most stringent upper limits at both Band 1 and 2 are achieved from

³² The lowest k mode in Figure 18 has no response below $k = 0.128 \, h \, \text{Mpc}^{-1}$ due to the horizon buffer enacted in spherical binning.

Table 3
Band 1 Power Spectra, Errors, and Upper Limits on $\Delta_{21}^2(z = 10.4)$ from Figure 17

k $h \text{ Mpc}^{-1}$	Field 1			Field 2			Field 3		
	Δ^2 (mK) ²	1σ (mK) ²	Δ_{UL}^2 (mK) ²	Δ^2 (mK) ²	1σ (mK) ²	Δ_{UL}^2 (mK) ²	Δ^2 (mK) ²	1σ (mK) ²	Δ_{UL}^2 (mK) ²
0.128	(232.14) ²	(36.90) ²	(237.93) ²	(311.36) ²	(36.84) ²	(315.69) ²	(434.21) ²	(48.03) ²	(439.49) ²
0.192	(124.01) ²	(42.26) ²	(137.66) ²	(122.81) ²	(34.88) ²	(132.34) ²	(191.00) ²	(43.14) ²	(200.50) ²
0.256	(41.34) ²	(61.07) ²	(95.74) ²	(73.63) ²	(49.50) ²	(101.60) ²	(131.11) ²	(61.15) ²	(157.06) ²
0.320	(103.46) ²	(84.89) ²	(158.49) ²	−(81.56) ²	(68.63) ²	(97.06) ²	(103.08) ²	(84.82) ²	(158.16) ²
0.384	(82.64) ²	(111.66) ²	(178.23) ²	(43.73) ²	(90.51) ²	(135.27) ²	(47.52) ²	(111.84) ²	(165.15) ²
0.448	(69.98) ²	(141.55) ²	(212.06) ²	(51.61) ²	(114.92) ²	(170.52) ²	(162.61) ²	(141.37) ²	(257.70) ²
0.512	(167.19) ²	(172.97) ²	(296.30) ²	−(195.21) ²	(139.86) ²	(197.80) ²	(104.72) ²	(172.32) ²	(265.24) ²
0.576	(364.27) ²	(206.10) ²	(466.52) ²	−(200.33) ²	(166.17) ²	(235.00) ²	(106.11) ²	(205.11) ²	(308.86) ²
0.640	−(73.05) ²	(240.38) ²	(339.95) ²	−(296.41) ²	(194.00) ²	(274.36) ²	(195.72) ²	(239.46) ²	(391.14) ²
0.704	(298.47) ²	(277.29) ²	(492.81) ²	(244.19) ²	(223.98) ²	(399.95) ²	(200.57) ²	(275.83) ²	(438.63) ²
0.768	(259.69) ²	(314.80) ²	(515.40) ²	(380.60) ²	(254.70) ²	(524.02) ²	−(264.99) ²	(314.01) ²	(444.07) ²
0.832	(270.03) ²	(354.73) ²	(569.72) ²	(254.18) ²	(286.82) ²	(478.69) ²	(434.40) ²	(354.47) ²	(663.33) ²
0.896	−(326.77) ²	(396.45) ²	(560.67) ²	−(111.95) ²	(320.48) ²	(453.23) ²	(504.58) ²	(395.78) ²	(753.58) ²
0.960	−(523.27) ²	(439.41) ²	(621.42) ²	(282.52) ²	(355.29) ²	(576.44) ²	(145.43) ²	(438.38) ²	(636.80) ²

Note. The upper limit is the measurement plus the 2σ error bar. In the process, Δ^2 is set to zero where it is measured to be negative.

Table 4
Band 2 Power Spectra, Errors, and Upper Limits on $\Delta_{21}^2(z = 7.9)$ from Figure 17

k $h \text{ Mpc}^{-1}$	Field 1			Field 2			Field 3		
	Δ^2 (mK) ²	1σ (mK) ²	Δ_{UL}^2 (mK) ²	Δ^2 (mK) ²	1σ (mK) ²	Δ_{UL}^2 (mK) ²	Δ^2 (mK) ²	1σ (mK) ²	Δ_{UL}^2 (mK) ²
0.128	(29.17) ²	(17.39) ²	(38.16) ²	(54.66) ²	(16.01) ²	(59.17) ²	(84.32) ²	(19.49) ²	(88.71) ²
0.192	(14.55) ²	(19.17) ²	(30.76) ²	(42.85) ²	(16.74) ²	(48.95) ²	(54.33) ²	(20.64) ²	(61.67) ²
0.256	(42.04) ²	(25.84) ²	(55.70) ²	(43.60) ²	(22.08) ²	(53.63) ²	(39.36) ²	(27.25) ²	(55.09) ²
0.320	−(22.72) ²	(35.07) ²	(49.60) ²	(51.11) ²	(29.87) ²	(66.31) ²	(44.19) ²	(36.84) ²	(68.31) ²
0.384	(82.98) ²	(46.17) ²	(105.59) ²	(62.34) ²	(39.28) ²	(83.50) ²	−(18.72) ²	(48.42) ²	(68.48) ²
0.448	(77.29) ²	(58.66) ²	(113.39) ²	(82.31) ²	(49.99) ²	(108.51) ²	(65.69) ²	(61.53) ²	(109.02) ²
0.512	(76.19) ²	(71.90) ²	(127.06) ²	(91.07) ²	(61.47) ²	(125.90) ²	(129.66) ²	(75.54) ²	(168.00) ²
0.576	(86.08) ²	(85.87) ²	(148.85) ²	(133.99) ²	(73.20) ²	(169.32) ²	(83.39) ²	(90.01) ²	(152.18) ²
0.640	−(108.73) ²	(100.34) ²	(141.90) ²	(126.10) ²	(85.39) ²	(174.59) ²	−(89.07) ²	(104.92) ²	(148.38) ²
0.704	(106.66) ²	(115.28) ²	(194.83) ²	(109.95) ²	(98.06) ²	(176.98) ²	(85.27) ²	(120.75) ²	(190.87) ²
0.768	(197.25) ²	(131.05) ²	(270.66) ²	(212.69) ²	(111.44) ²	(264.71) ²	(112.00) ²	(137.38) ²	(224.26) ²
0.832	(221.71) ²	(147.79) ²	(304.69) ²	(236.41) ²	(125.53) ²	(295.64) ²	(159.84) ²	(154.98) ²	(271.27) ²
0.896	−(43.63) ²	(164.64) ²	(232.84) ²	(219.42) ²	(140.29) ²	(295.82) ²	−(114.21) ²	(172.95) ²	(244.58) ²
0.960	(166.53) ²	(182.58) ²	(307.25) ²	(108.48) ²	(155.43) ²	(245.12) ²	(198.43) ²	(191.65) ²	(335.91) ²

Note. The same procedure as Table 3 is used to construct the upper limits.

Field 1, yielding an upper limit of $(95.74)^2 \text{ mK}^2$ at $z = 10.4$ and $k = 0.256 h \text{ Mpc}^{-1}$ and $(30.76)^2 \text{ mK}^2$ at $z = 7.9$ $k = 0.192 h \text{ Mpc}^{-1}$. This is the most sensitive upper limit on the EoR 21 cm power spectrum at $z \sim 8$ by over an order of magnitude, achieved with only a fraction of the full sensitivity of the HERA array. HERA’s nominal sensitivity over a year-long observing campaign could be pushed two orders of magnitude deeper, thus reaching the fiducial signal amplitudes of standard EoR models; however, low-level systematics will need to be modeled and appropriately resolved in order to reach those sensitivities.

5. Validation of Upper Limits

Current 21 cm analysis pipelines are becoming increasingly sophisticated in order to beat down a host of systematic effects and reach the level of precision necessary to detect the cosmological 21 cm signal. A key concern is that these analysis choices could cause nonnegligible amounts of signal loss, whether

from calibration (Mouri Sardarabadi & Koopmans 2019), power spectrum estimation (Cheng et al. 2018), or other steps. The extent of any possible signal loss needs to be quantified in order to build confidence that our results are not prone to signal loss. In this analysis we use a combination of three complementary approaches to build confidence in our analysis and the robustness of upper limits on the power spectrum: simulation-based checks of the pipeline, a comparative study of uncertainty quantification, and statistical null tests on data subsets

In a companion paper, Aguirre et al. (2022) present a validation of the Phase I analysis and power spectrum pipeline, using data simulations with realistic sky, instrument, and systematic models. In addition to tests on individual components of the pipelines, such as calibration, systematic modeling, and power spectrum estimation, they also present a near end-to-end pipeline test, demonstrating the pipeline’s ability to accurately recover a realistic EoR model for $k \geq 0.192 h \text{ Mpc}^{-1}$ at both of the spectral windows considered in this work. This end-to-end study is

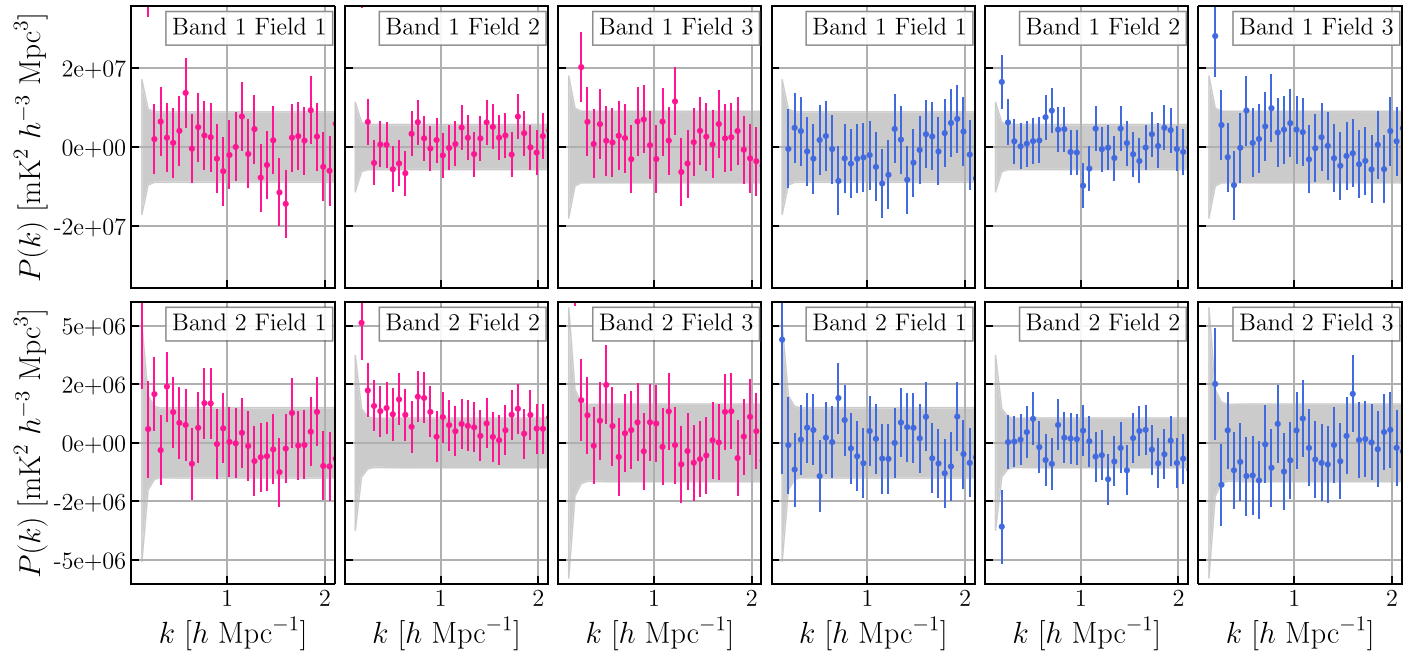


Figure 19. The real and imaginary components of the spherically averaged power spectrum $P(k)$. The six plots on the left show the real component (pink dots), while the six plots on the right show the imaginary component (blue dots). The error bars are 2σ and the shaded region is the P_N theoretical noise expectation. Significant power in the imaginary component could indicate phase calibration errors resulting from unmodeled systematics or baseline nonredundancies. We quantify the consistency of the data with noise via p -value tests, tabulated in Table 5.

particularly important, as different steps of the pipeline could potentially interact in nonlinear ways that introduce signal loss. While the instrument model used in that work are complex enough to model the most prominent systematics, there are some areas where the models can be made more sophisticated in future work to test more subtle features of the instrument, which may be necessary for validating an analysis with a putative EoR detection. In addition to testing the Phase I pipeline with simulated data sets, Aguirre et al. (2022) also show a semiblinded analysis of the simulated data with an independent power spectrum pipeline as a consistency test with the Phase I power spectrum pipeline. In this exercise, a small group of people were excluded from discussions on the construction of the simulated data and systematics, and applied an independent, simplified power spectrum pipeline to the data products to test for experimenter bias in the derived limits. Their resultant power spectra are in good agreement with the output of the Phase I pipeline, in that they show unbiased detection of the simulated EoR signal for $k \gtrsim 0.2 h \text{ Mpc}^{-1}$.

In addition to the simulated pipeline validation from Aguirre et al. (2022), Tan et al. (2021) present a comparative study of uncertainty quantification in the power spectrum pipeline. They highlight a number of techniques used in the literature for estimating uncertainty on the 21 cm power spectrum, both analytic and semi-empirical, and demonstrate that they all show relatively good agreement with each other when applied to both real and simulated HERA Phase I power spectra. Specifically, as discussed in Section 4.1, we use the P_N and P_{SN} error bars that quantify the root-mean-square (rms) of the power spectra due solely to thermal noise (P_N), and the rms due to thermal noise and its coupling to residual signal terms in the data (P_{SN}). The final error bars quoted in Tables 3 and 4 come from this latter metric.

In the rest of this section, we discuss a series of statistical consistency tests to better understand residual systematics observed in the data. Some of these tests are formulated as true null tests, where the null hypothesis is that the spherically

averaged power spectrum is consistent with noise-only fluctuations, while others are better described as consistency checks that seek to verify a particular scaling law. Such tests can help to quantify the parts of the data that are systematic or noise limited (e.g., Kolopanis et al. 2019).

5.1. Real and Imaginary Parts of the Power Spectrum

This test seeks to assess whether the final spherically averaged power spectra are consistent with thermal noise. In the absence of residual systematics, we expect the power spectra to consist of foreground emission, EoR emission, and thermal noise. Given that we have masked the foreground horizon upon spherical binning, any excess power above the predicted thermal noise variance is likely due to residual systematics, which can take many forms. Generally, these systematics leak foreground emission to k modes beyond the foreground horizon modes, thereby contaminating the lowest k modes in the EoR window. We assess whether the data are consistent with noise by evaluating the p -value of the power spectrum χ^2 statistic for each band and field combination. The statistic is defined as

$$\chi^2 = \mathbf{d}^T \mathbf{\Sigma}^{-1} \mathbf{d}, \quad (26)$$

where \mathbf{d} is the power spectrum data vector and $\mathbf{\Sigma}$ is its noise covariance. The noise covariance is nearly diagonal, with small but nonzero neighbor-to-neighbor covariance (Figure 20). The p -value for a particular data vector is the area under the χ^2 sampling distribution that exceeds the χ^2 of the data. The χ^2 sampling distribution is derived in a Monte Carlo fashion by drawing a large number ($N = 10^6$) of random noise realizations and evaluating their χ^2 . For p -values below 0.001 we can only quote an upper limit on the p -value given the discrete number of random draws, however, this is a sufficiently small value to enable an unambiguous rejection of the null hypothesis. Note

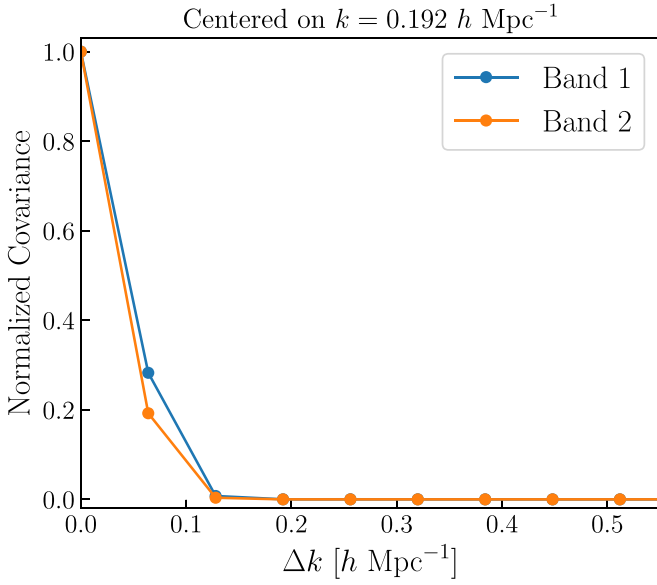


Figure 20. The normalized covariance between the $k = 0.192 \, h \, \text{Mpc}^{-1}$ mode and neighboring modes used in the statistical tests of Table 5. This shows there is a small but nonzero covariance between neighboring modes, and effectively no covariance at larger separations.

Table 5
Statistical p -value Tests on the Real and Imaginary Components of the Spherical $P(k)$; (Figure 19)

Data Selection	p -value		
	$k \geq 0.192$	$k \geq 0.5$	$k \geq 1.0$
Re[P], Band 1, Field 1	<0.001	<0.001	0.003
Im[P], Band 1, Field 1	0.605	0.477	0.345
Re[P], Band 1, Field 2	<0.001	0.044	0.203
Im[P], Band 1, Field 2	<0.001	0.002	0.011
Re[P], Band 1, Field 3	<0.001	0.107	0.050
Im[P], Band 1, Field 3	<0.001	0.358	0.523
Re[P], Band 2, Field 1	0.020	0.372	0.716
Im[P], Band 2, Field 1	0.232	0.216	0.504
Re[P], Band 2, Field 2	<0.001	<0.001	0.007
Im[P], Band 2, Field 2	<0.001	0.064	0.111
Re[P], Band 2, Field 3	<0.001	0.439	0.717
Im[P], Band 2, Field 3	0.036	0.292	0.733

Note. The null hypothesis is that the power spectrum is consistent with mean-zero fluctuations drawn from the propagated bandpower noise covariance (Figure 20). Upper limits on the p -value are quoted where it is sufficiently small. The k cuts are in $h \, \text{Mpc}^{-1}$.

that we do not enact a hard cut on the p -value to determine whether the data reject the null hypothesis, but rather use the collection of p -values for a given band and field in both the real and imaginary components to ascertain whether the data seem consistent with thermal noise (Figure 19).

Based on the estimator presented in Section 4.1, the power spectrum of a noise-only data set will be a mean-zero process in

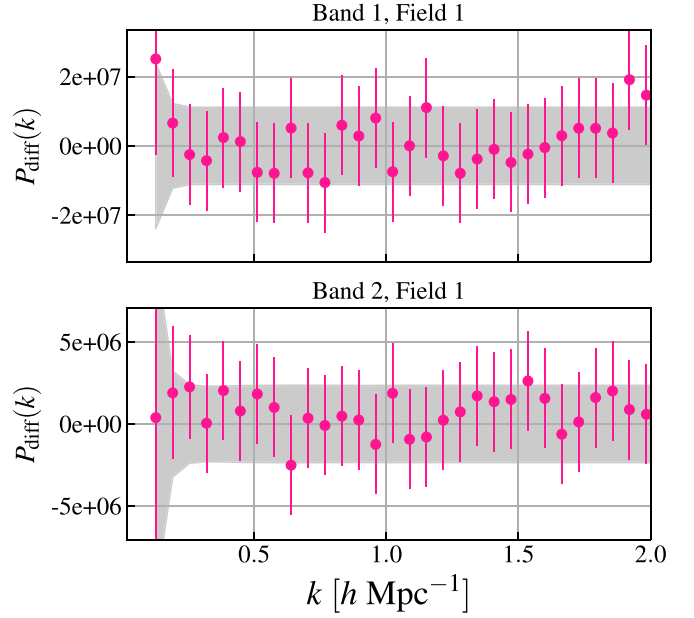


Figure 21. The real component of the power spectrum computed from the difference of interleaved nights for Field 1. Vertical error bars are 2σ , and the gray shaded region is $\pm 2P_N$. The p -values for these power spectra with k cuts of ≥ 0.192 , ≥ 0.5 , and $\geq 1.0 \, h \, \text{Mpc}^{-1}$ are 0.651, 0.491, and 0.643 for Band 1 and 0.585, 0.584, and 0.619 for Band 2, respectively. These are consistent with the null hypothesis that the data are drawn from the thermal noise covariance.

both the real and imaginary components with equal variance. Coherent signals in the data, whether they are sky signals or instrumental systematics, are confined to the real component of the power spectrum. However, systematics with phase differences between cross-multiplied visibilities can cause excess variance in the imaginary component of the power spectrum (Kolopanis et al. 2019). Figure 19 shows the real component (pink panels) and the imaginary component (blue panels) of the power spectra in $P(k)$, with their p -values tabulated in Table 5 showing the response of the p -value with an increasingly large k cut. We see that for both bands, most fields are inconsistent with the noise distribution at the low k modes, evidenced by their consistently low p -values in both the real and imaginary components. Reassuringly, the p -values for the imaginary component are generally more consistent with noise than the p -values for the real component, even when there is strong evidence for systematics, like Band 2, Field 2. The interesting exception to the rule is Band 1, Field 2, whose imaginary component is less consistent with the noise than its real component, possibly indicative of a low-level, baseline-dependent systematic. Overall, the quantification provided by the p -value tests on the real and imaginary components supports the notion that the data are largely consistent with thermal noise at high k modes, while confirming our intuition that residual systematics are beginning to be detected at the lowest k modes in the EoR window.

5.2. Night-to-night Binning Split

A check on long-term temporal stability can be made by taking the nightly data from the data set and separating them into two groups, differencing them, and then forming the power spectra of the differenced visibilities. To do this we take every other night in the 18 night data set and form a set of “even” nights and a set of “odd” nights in terms of their Julian date. We then perform the standard night-to-night LST binning described in Section 3.6 on both the even and odd sets independently. Afterwards, we pass both data sets through the

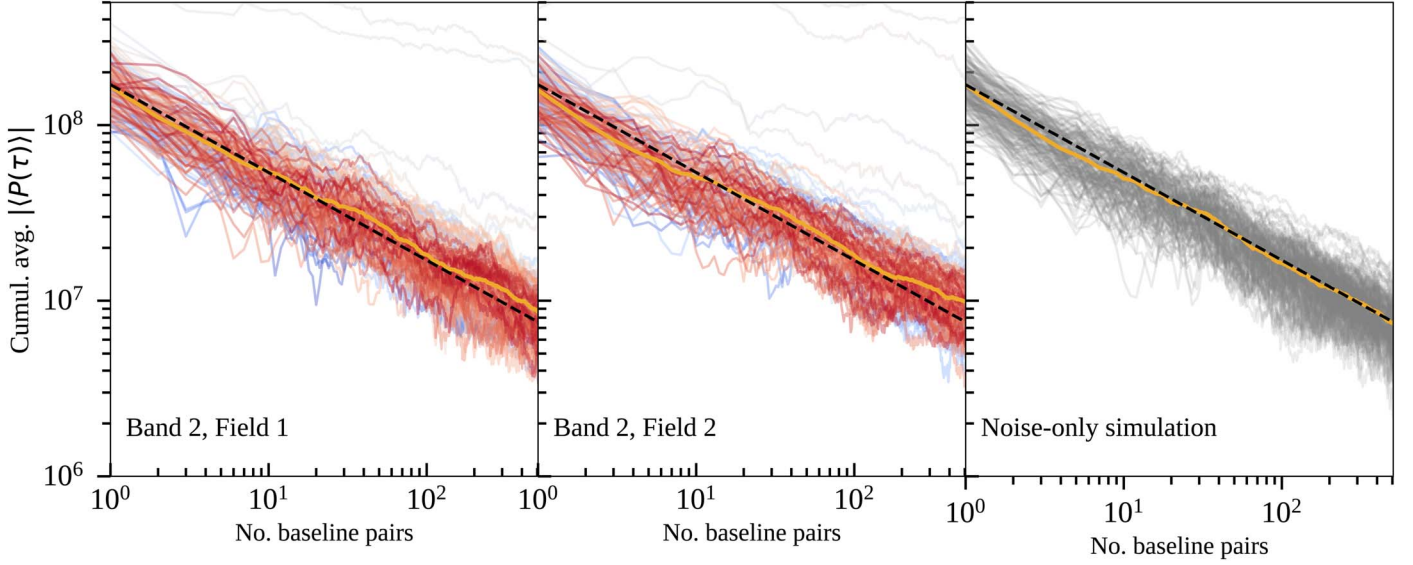


Figure 22. Cumulative incoherent averaging of the delay spectra in arbitrary units within a single redundant baseline group (all 14.6 meter, east–west baselines) for Band 2 of Field 1 (left panel), Field 2 (middle panel), and a corresponding thermal noise-only realization of Band 2, Field 2 (right panel). Each line denotes a different bandpower that has been cumulatively averaged over the baseline pairs within the redundant group. To reduce the noisiness of the statistic, the absolute value of the real part was then taken and averaged over all LST samples in the field. Red lines denote negative delays while blue lines denote positive delays. Note that low delay modes of $|\tau| \lesssim 200$ ns are not shown in the plot. The black dashed line shows a theoretical $1/\sqrt{N_{\text{baseline-pair}}}$ scaling, with the same arbitrary normalization in all panels. The thick orange line shows the mean of the bandpowers for $|\tau| > 500$ ns ($k_{\parallel} \sim 0.25$ h Mpc $^{-1}$), showing a slight detection of a systematic at deep integration levels, particularly for Band 2, Field 2.

standard systematics treatment and then difference the resultant visibilities before estimating the power spectrum. This test is sensitive to systematics that exhibit nonnegligible variation from night to night.

Figure 21 shows the real component of the differenced power spectrum for Field 1. Computing the p -values of these spectra in the same manner as in Section 5.1, we find that the spectra for both Band 1 and Band 2 are consistent with thermal noise for $k \geq 0.192$ h Mpc $^{-1}$, which was not observed for Field 1 in the undifferenced power spectrum (Table 5). This suggests that the low k residual systematics seen in the data are fairly stable from night to night.

5.3. Scaling of Bandpowers under Cumulative Averaging

If delay modes outside the foreground horizon are dominated by thermal noise, as we would expect in the absence of residual systematics or an EoR signal detection, the measured power in those modes should integrate down in a characteristic way as more time samples and baseline pairs are averaged over. In this section, we study the scaling of the measured delay spectrum bandpowers when cumulatively averaging individual baseline pairs in a redundant baseline group, which should in principle measure the same sky signal but have independent noise contributions. Note that this test is performed higher up in the power spectrum pipeline before cylindrical or spherical binning of the power spectra, which is necessary because we need a large number of independent samples for computing the cumulative average. We specifically target the east–west 14.6 meter baseline group for this test as it has the largest number of independent baselines.

The left and center panel of Figure 22 show the scaling of the delay spectrum bandpowers for Band 2, Fields 1 and 2, having cumulatively averaged over the redundant baseline pairs. Each line shows the a different bandpower, ranging from negative

delays to positive delays, while the black dashed line shows a representative $1/\sqrt{N}$ scaling, where N is the number of averaged baseline pairs. To reduce the scatter in this statistic, we take the absolute value of the real part of the bandpower and average across the remaining LST samples associated with Field 1.

While low delay modes ($|\tau| \lesssim 200$ ns) are not plotted, we see that the high delay modes are generally consistent with the expected $1/\sqrt{N}$ scaling of noise. As a demonstration, the right panel of Figure 22 shows the same procedure applied to a pure-noise simulation for Band 2, Field 2.

The thick orange line shows the average of the bandpowers across delays for $|\tau| > 500$ ns (or $|k_{\parallel}| \gtrsim 0.25$ h Mpc $^{-1}$). This more clearly shows the marginal detection of a systematic at deep integration levels, which is more pronounced for Band 2, Field 2, as we would expect given our conclusions from Figure 19. Nevertheless, this test shows that the low-level systematics do partially integrate when averaging across different baselines, as the mean of the bandpowers (i.e., the orange line) still shows a downward trend. Integrating more data will help to understand at what dynamic range these systematics hit a plateau, if at all.

5.4. Stability with Respect to Frequency Selection

Here, we seek to assess the overall containment of foreground structure within the foreground horizon as a function of frequency. To do this, we compute the delay spectrum of the data over a subband with fixed bandwidth and then iteratively shift the center of the subband through the full frequency range of the data, thereby probing for spectral discontinuities in the data at particular locations in frequency. In order to test the consistency of the measurements with noise, we simulate mock visibilities consisting of Gaussian random noise with the same sampling and flagging patterns as the real

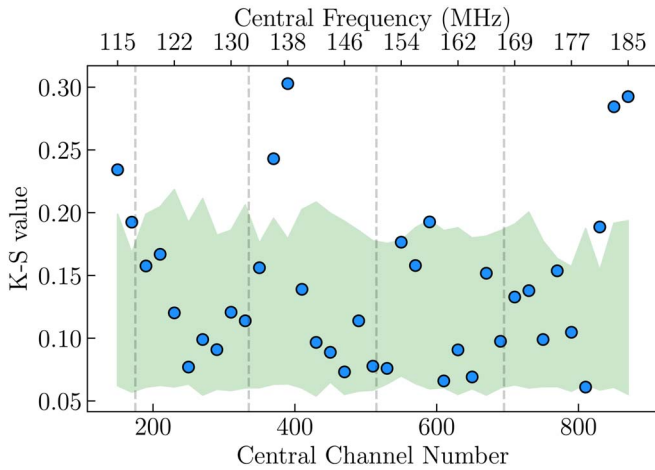


Figure 23. Blue points show the K–S value between data and noise-only simulations. The green regions are 5th–95th percentile range of the K–S values from a set of noise-only simulations. The vertical dashed lines show the spectral window chosen for the power spectrum analysis (Section 4.1).

data. As a metric, we look at the mean bandpower within a delay range of $2000 < |\tau| < 4000$ nanoseconds for both the real data and simulated noise data sets. By picking out the high delay regime, we are constructing a test that is sensitive to spurious spectral structure that would cause foreground power to leak significantly beyond the foreground horizon. Similar to Section 5.3, we perform this test with the east–west 14.6 meter redundant baseline group to enable a large sample size, and focus on the LST ranges associated with Field 1. Note that this test is performed after delay-based inpainting, where the data have been reconstructed in previously flagged channels.

We use a two-sample Kolmogorov–Smirnov (K–S) test (Hodges 1958) to examine if the underlying probability distribution of the measured bandpowers from the data and noise-only simulations are different. Again, this process is repeated as we shift the spectral window up in frequency, where the window has a width of 100 channels, and is moved up 20 channels every iteration. We compute the allowed range of the metric by taking the 5th and 95th percentiles of the two-sample K–S test, having paired two independent noise simulations, and then repeating for all pair combinations of the noise simulations and averaging the result. This interval is shown as the green shaded region in Figure 23, which also shows the two-sample K–S test between the data and the noise simulation (blue points). This shows that for most of the total bandwidth, the measured bandpowers at high delay are consistent with that of pure-noise fluctuations. The Band 1 and Band 2 spectral windows are shown as gray dashed lines.

The notable exception are the band edges, which are known to be suboptimal to due tapering effects in the process of calibration, as well as the ORBCOMM band at ~ 138 MHz, where a large swath of channels are routinely flagged, which is known to degrade the ability of the delay-based inpainting algorithm.

5.5. Impact of Systematic Treatment

This null test is aimed at examining the consistency of the bandpowers with thermal noise before and after systematic treatment, discussed in more detail in Dibblee-Barkman & Singh (2021). We compute the delay power spectra for a single 14.6 meter, east–west oriented baseline, this time spanning an

LST range of 5–5.4 hr overlapping with Field 2. We estimate the cumulative probability distribution (CDF) of the bandpowers at different delays, constructing the CDF across the 30 time bins in the LST range studied here. We then perform a one-sample K–S test between the bandpower at each delay and the expected analytic distribution of the bandpowers under the assumption that the visibilities are drawn from random Gaussian noise. Additionally, we also construct bandpower CDFs having resampled the time bins with replacement, which gives us a sense for the inherent variation of the CDF throughout the LST range. We compute the one-sample K–S test for each of the resampled CDFs, and then take their average.

The result is shown in Figure 24, where the solid line shows the K–S statistic from the un-resampled CDF as a function of delay mode and the dashed line shows the average K–S statistic from the resampled CDF, showing good agreement between the two, suggesting that variation of the K–S statistic as a function of time is minimal. We repeat this test on the data before systematics treatment (top panel), after systematics treatment (middle panel), and also compute it on a pure-noise simulation for validation purposes (bottom panel). The horizontal black line with a K–S value of ~ 0.25 is the analytically computed 95th percentile of the K–S values assuming the CDF is drawn from pure noise, which is validated by the results of the bottom panel showing the vast majority of the KS values falling below the solid black line. What we see from the top and middle panels of Figure 24 is a clear discrepancy between the data and the assumed noise distribution at intermediate delay modes before performing systematics treatment (top panel), which is not surprising given that we know those modes to be systematics dominated. After systematics treatment (middle panel), these delay modes show good agreement with the noise model, except for at very low delays where we expect intrinsic foreground emission to dominate.

6. Summary

In this work we have presented upper limits on the 21 cm power spectrum from Phase I observations of the 50 element HERA at redshifts 7.9 and 10.4. The most sensitive 2σ limits achieved are $(30.76)^2 \text{ mK}^2$ at $z = 7.9$ and $k = 0.192 \text{ h Mpc}^{-1}$ and $(95.74)^2 \text{ mK}^2$ at $z = 10.4$ and $k = 0.256 \text{ h Mpc}^{-1}$. At $z = 7.9$, the limits presented here are the most sensitive to date within the literature by roughly an order of magnitude. Along with a series of paper describing the Phase I analysis pipeline, including redundant calibration (Dillon et al. 2020), absolute calibration (Kern et al. 2020a), systematic modeling (Kern et al. 2019, 2020b), uncertainty quantification (Tan et al. 2021), and pipeline validation (Aguirre et al. 2022), this work details the Phase I analysis and power spectrum pipelines, and discusses a series of statistical tests that show that the data are largely thermal noise limited for $k \geq 0.5 \text{ h Mpc}^{-1}$, whereas at lower k modes the data show evidence for low-level systematics. We speculate that these residual systematics could be due to radio frequency interference, as well as possibly residual gain errors or residual instrumental coupling effects. Future work exploring better RFI identification, as well as more comprehensive systematic models, will help to better understand the origin of these systematics.

Note that no explicit foreground subtraction or filtering was performed in the analysis. Instead, our analysis emphasized

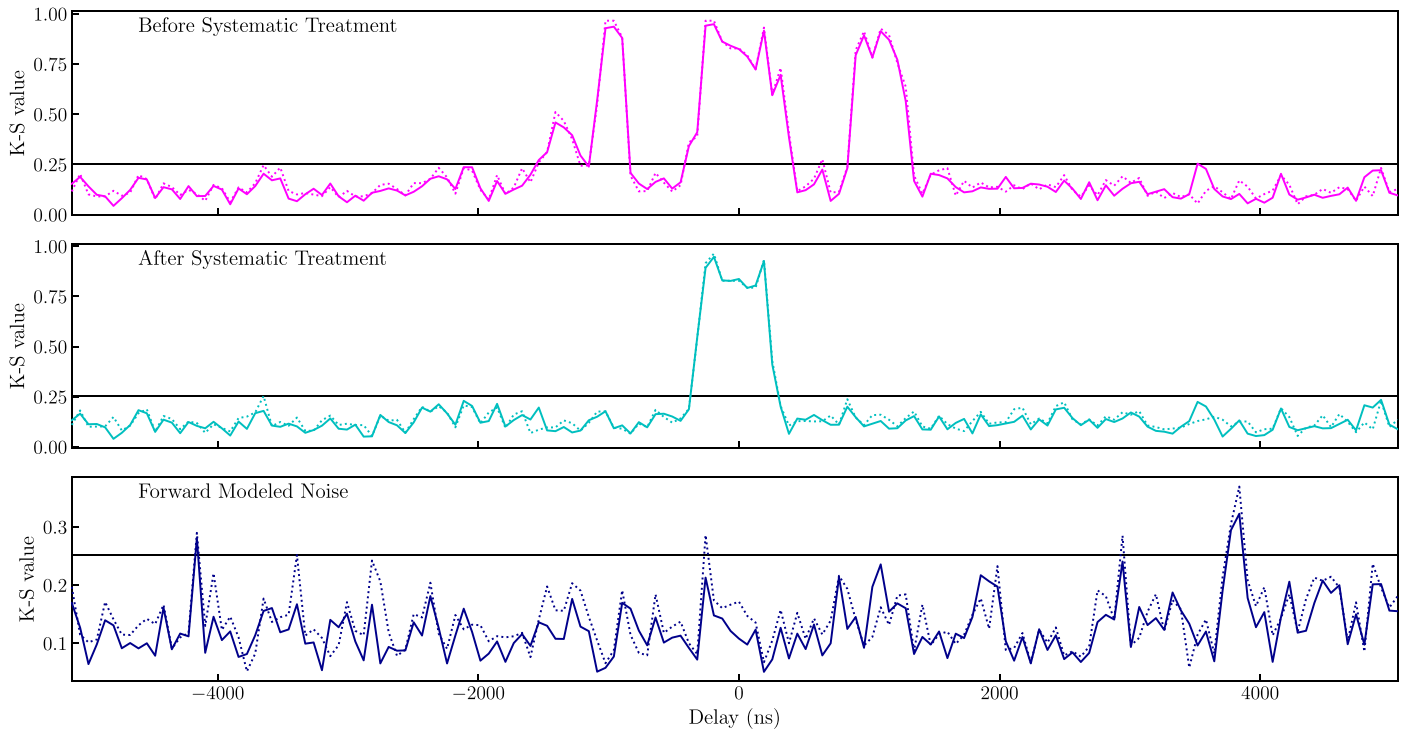


Figure 24. One-sample K–S test at different delays, comparing CDFs of the bandpowers from the data, and the mean bootstrapped CDF with that expected from Gaussian random noise. The data without systematic subtraction (top panel) show two clear regions, foreground and systematic dominated, where K–S values are clearly higher than the threshold, marked with the black horizontal line. Systematic subtraction (middle panel) continues to have high K–S values for foreground dominated region, while all other delay ranges seem consistent with thermal noise. As a reference, bottom panel shows the K–S values from the simulated data set that consists of Gaussian random noise with the same variance as that of the data. Naturally, in this case, the K–S values are consistent with the reference distribution across all delays. Solid and dashed lines represent the results from a K–S test on original bandpowers and the mean of bootstrapped bandpowers, respectively.

strong control of spectral systematics in order to keep foreground structure largely contained within the foreground horizon. Ultimately, this enabled the Phase I analysis pipeline achieved a dynamic range between the thermal noise floor and the peak foreground power of 10^9 in power. Future work will focus both on a more comprehensive analysis of low-level systematics, and the modeling and subtraction of foreground emission. Hardware upgrades in the HERA Phase II system will potentially mitigate some of the observed Phase I systematics in the field, such as cross-coupling and reflection contamination.

The overall systematic uncertainties in this analysis come predominately from the uncertainty on the absolute flux density scale that is known to an accuracy of $\sim 10\%$, which is pinned to the GLEAM catalog (Hurley-Walker et al. 2017). Furthermore, nightly drift in the gain amplitude is not corrected for in this work, which Kern et al. (2020a) estimate to be a roughly 3% effect in the gains. Signal loss arising from the analysis pipeline are explicitly derived and corrected for, but in total this amounts to less than a 15% correction of the power spectrum amplitude. Cosmic variance uncertainty on the 21 cm power spectrum limits is expected to be 5.5% (1σ) given the uv sampling of the instrument for a 2 hr integration across LST (Aguirre et al. 2022).

Going forward, future Phase I analyses may include a ~ 100 night data set that theoretically stands to improve the sensitivity of our limits here by over a factor of 5, assuming the low k systematics observed in this work can be further mitigated. Additionally, Phase II commissioning and observations have already commenced, with a new front-end receiver spanning an increased bandwidth from 50–250 MHz. Overall, the analysis

pipeline presented in this work has laid a framework for future analyses of HERA data as construction and commissioning is completed.

This material is based upon work supported by the National Science Foundation under grant Nos. 1636646 and 1836019 and institutional support from the HERA collaboration partners. This research is funded in part by the Gordon and Betty Moore Foundation. HERA is hosted by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. Parts of this research were supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. G. Bernardi acknowledges funding from the INAF PRIN-SKA 2017 project 1.05.01.88.04 (FORECaST), support from the Ministero degli Affari Esteri della Cooperazione Internazionale—Direzione Generale per la Promozione del Sistema Paese Progetto di Grande Rilevanza ZA18GR02 and the National Research Foundation of South Africa (grant No. 113121) as part of the ISARP RADIOSKY2020 Joint Research Scheme, from the Royal Society and the Newton Fund under grant NA150184 and from the National Research Foundation of South Africa (grant No. 103424). P. Bull acknowledges funding for part of this research from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 948764), and from STFC grant ST/T000341/1. E. de Lera Acedo acknowledges the funding support of the UKRI Science and Technology Facilities Council SKA grant. J.S. Dillon gratefully acknowledges the

support of the NSF AAPF award #1701536. N. Kern acknowledges support from the MIT Pappalardo fellowship. A. Liu acknowledges support from the New Frontiers in Research Fund Exploration grant program, the Canadian Institute for Advanced Research (CIFAR) Azrieli Global Scholars program, a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Discovery Launch Supplement, the Sloan Research Fellowship, and the William Dawson Scholarship at McGill. We gratefully acknowledge an anonymous referee whose feedback improved the clarity of this work.






























Software

This analysis utilized custom-built software by the HERA Collaboration (<https://github.com/hera-team>) in addition to software built by both HERA members and collaborators (<https://github.com/RadioAstronomySoftwareGroup>). This analysis also relied on publicly accessible and open-source software, including *numpy* (Harris et al. 2020), *scipy* (Virtanen et al. 2020), and *astropy* (Astropy Collaboration et al. 2018).

Appendix Data

The data from Tables 3, 4, Figures 17, and 18 are publicly available and can be accessed at <https://reionization.org/science/public-data-release-1/>.

ORCID iDs

James E. Aguirre  <https://orcid.org/0000-0002-4810-666X>
 Adam P. Beardsley  <https://orcid.org/0000-0001-9428-8233>
 Gianni Bernardi  <https://orcid.org/0000-0002-0916-7443>
 Judd D. Bowman  <https://orcid.org/0000-0002-8475-2036>
 Philip Bull  <https://orcid.org/0000-0001-5668-3101>
 Chris L. Carilli  <https://orcid.org/0000-0001-6647-3861>
 David R. DeBoer  <https://orcid.org/0000-0003-3197-2294>
 Joshua S. Dillon  <https://orcid.org/0000-0003-3336-9958>
 Aaron Ewall-Wice  <https://orcid.org/0000-0002-0086-7363>
 Steven R. Furlanetto  <https://orcid.org/0000-0002-0658-1243>
 Deepthi Gorthi  <https://orcid.org/0000-0002-0829-167X>
 Bradley Greig  <https://orcid.org/0000-0002-4085-2094>
 Bryna J. Hazelton  <https://orcid.org/0000-0001-7532-645X>
 Daniel C. Jacobs  <https://orcid.org/0000-0002-0917-2269>
 Nicholas S. Kern  <https://orcid.org/0000-0002-8211-1892>
 Joshua Kerrigan  <https://orcid.org/0000-0002-1876-272X>
 Piyanat Kittiwisit  <https://orcid.org/0000-0003-0953-313X>
 Saul A. Kohn  <https://orcid.org/0000-0001-6744-5328>
 Matthew Kolopanis  <https://orcid.org/0000-0002-2950-2974>
 Adrian Liu  <https://orcid.org/0000-0001-6876-0928>
 Andrei Mesinger  <https://orcid.org/0000-0003-3374-1772>
 Miguel F. Morales  <https://orcid.org/0000-0001-7694-4030>
 Steven G. Murray  <https://orcid.org/0000-0003-3059-3823>
 Abraham R. Neben  <https://orcid.org/0000-0001-7776-7240>
 Chuneeta D. Nunhokee  <https://orcid.org/0000-0002-5445-6586>
 Nipanjana Patra  <https://orcid.org/0000-0002-9457-1941>
 Jonathan C. Pober  <https://orcid.org/0000-0002-3492-0433>
 Peter Sims  <https://orcid.org/0000-0002-2871-0413>
 Saurabh Singh  <https://orcid.org/0000-0001-7755-902X>
 Nithyanandan Thyagarajan  <https://orcid.org/0000-0003-1602-7868>
 Peter K. G. Williams  <https://orcid.org/0000-0003-3734-3587>

References

- Aguirre, J. E., Murray, S. G., Pascua, R., et al. 2022, *ApJ*, 924, 85
 Ali, Z., Parsons, A., Zheng, H., et al. 2015, *ApJ*, 809, 61
 Ali, Z., Parsons, A., Zheng, H., et al. 2018, *ApJ*, 863, 201, (Erratum)
 Asad, K., Koopmans, L., Jelić, V., et al. 2015, *MNRAS*, 451, 3709
 Asad, K. M. B., Koopmans, L. V. E., Jelić, V., et al. 2016, *MNRAS*, 462, 4482
 Asad, K. M. B., Koopmans, L. V. E., Jelić, V., et al. 2018, *MNRAS*, 476, 3051
 Astropy Collaboration, Price-Whelan, A. M., SipHocz, B. M., et al. 2018, *AJ*, 156, 123
 Barry, N., Beardsley, A. P., Byrne, R., et al. 2019a, *PASA*, 36, e026
 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
 Becker, G. D., Bolton, J. S., Madau, P., et al. 2015, *MNRAS*, 447, 3402
 Bernardi, G., Zwart, J. T. L., Price, D., et al. 2016, *MNRAS*, 461, 2847
 Blackman, R. B., & Tukey, J. W. 1958, *BSTJ*, 37, 185
 Bolton, J. S., Haehnelt, M. G., Warren, S. J., et al. 2011, *MNRAS*, 416, L70
 Boonstra, A., & van der Veen, A. 2003, *ITSP*, 51, 25
 Bosman, S. E. I., Fan, X., Jiang, L., et al. 2018, *MNRAS*, 479, 1055
 Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J., & Mahesh, N. 2018, *Natur*, 555, 67
 Bradley, R. F., Tauscher, K., Rapetti, D., & Burns, J. O. 2019, *ApJ*, 874, 153
 Brentjens, M. A., & de Bruyn, A. G. 2005, *A&A*, 441, 1217
 Byrne, R., Morales, M. F., Hazelton, B., et al. 2019, *ApJ*, 875, 70
 Carilli, C. L., Nikolic, B., Thyagarayan, N., & Gale-Sides, K. 2018, *RaSc*, 53, 845
 Caruana, J., Bunker, A. J., Wilkins, S. M., et al. 2014, *MNRAS*, 443, 2831
 Cheng, C., Parsons, A. R., Kolopanis, M., et al. 2018, *ApJ*, 868, 26
 Choudhuri, S., Bull, P., & Garsden, H. 2021, *MNRAS*, 506, 2066
 Choudhury, T. R., Paranjape, A., & Bosman, S. E. I. 2021, *MNRAS*, 501, 5782
 Ciardi, B., & Ferrara, A. 2005, *SSRv*, 116, 625
 Condon, J. J., & Matthews, A. M. 2018, *PASP*, 130, 073001
 Datta, K. K., Jensen, H., Majumdar, S., et al. 2014, *MNRAS*, 442, 1491
 Davies, F. B., Hennawi, J. F., Bañados, E., et al. 2018, *ApJ*, 864, 142
 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
 Dibble-Barkman, T., & Singh, S. 2021, KS Testing on HERA Data 92, HERA Memo 99, (HERA: Carnarvon)
 Dillon, J., Liu, A., Williams, C., et al. 2014, *PhRvD*, 89, 23002
 Dillon, J., Neben, A., Hewitt, J., et al. 2015, *PhRvD*, 91, 123011
 Dillon, J., & Parsons, A. 2016, *ApJ*, 826, 181
 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, *MNRAS*, 499, 5840
 Douspis, M., Aghanim, N., Ilić, S., & Langer, M. 2015, *A&A*, 580, L4
 Eastwood, M. W., Anderson, M. M., Monroe, R. M., et al. 2019, *AJ*, 158, 84
 Ewall-Wice, A., Bradley, R., DeBoer, D., et al. 2016, *ApJ*, 831, 196
 Ewall-Wice, A., Dillon, J., Hewitt, J., et al. 2016b, *MNRAS*, 460, 4320
 Ewall-Wice, A., Hewitt, J., Mesinger, A., et al. 2016a, *MNRAS*, 458, 2710
 Ewall-Wice, A., Kern, N., Dillon, J. S., et al. 2021, *MNRAS*, 500, 5195
 Fagnoni, N., de Lera Acedo, E., DeBoer, D. R., et al. 2021a, *MNRAS*, 500, 1232
 Fagnoni, N., de Lera Acedo, E., Drought, N., et al. 2021b, *ITAP*, 69, 8143
 Furlanetto, S., Oh, S., & Briggs, F. 2006, *PhR*, 433, 181
 Gehlot, B. K., Mertens, F. G., Koopmans, L. V. E., et al. 2019, *MNRAS*, 488, 4271
 Ghosh, A., Mertens, F., Bernardi, G., et al. 2020, *MNRAS*, 495, 2813
 Ghosh, A., Prasad, J., Bharadwaj, S., Ali, S. S., & Chengalur, J. N. 2012, *MNRAS*, 426, 3295
 Gorce, A., Douspis, M., Aghanim, N., & Langer, M. 2018, *A&A*, 616, A113
 Greig, B., & Mesinger, A. 2017, *MNRAS*, 465, 4838
 Greig, B., Mesinger, A., Haiman, Z., & Simcoe, R. A. 2017, *MNRAS*, 466, 4239
 Greig, B., Mesinger, A., & Pober, J. 2016, *MNRAS*, 455, 4295
 Hamaker, J. P., Bregman, J. D., & Sault, R. J. 1996, *A&AS*, 117, 137
 Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Natur*, 585, 357
 Hazra, D. K., Paoletti, D., Finelli, F., & Smoot, G. F. 2020, *PhRvL*, 125, 071301
 Hills, R., Kulkarni, G., Meerburg, P. D., & Puchwein, E. 2018, *Natur*, 564, E32
 Hodges, J. L. 1958, *ArM*, 3, 469
 Hogan, C., & Rees, M. 1979, *MNRAS*, 188, 791
 Högbom, J. A. 1974, *A&AS*, 15, 417
 Hogg, D. W. 1999, *arXiv:astro-ph/9905116*
 Hurley-Walker, N., Callingham, J. R., Hancock, P. J., et al. 2017, *MNRAS*, 464, 1146

- Jelić, V., Zaroubi, S., Labropoulos, P., et al. 2010, *MNRAS*, **409**, 1647
- Jensen, H., Laursen, P., Mellema, G., et al. 2013, *MNRAS*, **428**, 1366
- Keating, L. C., Weinberger, L. H., Kulkarni, G., et al. 2020, *MNRAS*, **491**, 1736
- Kern, N. S., Dillon, J. S., Parsons, A. R., et al. 2020a, *ApJ*, **890**, 122
- Kern, N. S., & Liu, A. 2021, *MNRAS*, **501**, 1463
- 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., Parsons, A. R., Dillon, J. S., et al. 2020b, *ApJ*, **888**, 70
- Kerrigan, J. R., Pober, J. C., Ali, Z. S., et al. 2018, *ApJ*, **864**, 131
- Kim, K. S., Lilly, S. J., Miniati, F., et al. 2016, *ApJ*, **829**, 133
- Kolopanis, M., Jacobs, D. C., Cheng, C., et al. 2019, *ApJ*, **883**, 133
- Koopmans, L., Pritchard, J., Mellema, G., et al. 2015, in Proc. of Advancing Astrophysics with the Square Kilometre Array (Trieste: PoS)
- Kulkarni, G., Keating, L. C., Haehnelt, M. G., et al. 2019, *MNRAS*, **485**, L24
- Lenc, E., Anderson, C. S., Barry, N., et al. 2017, *PASA*, **34**, e040
- Li, W., Pober, J. C., Barry, N., et al. 2019, *ApJ*, **887**, 141
- Li, W., Pober, J. C., Hazelton, B. J., et al. 2018, *ApJ*, **863**, 170
- Liu, A., & Parsons, A. 2016, *MNRAS*, **457**, 1864
- Liu, A., Parsons, A., & Trott, C. 2014a, *PhRvD*, **90**, 23018
- Liu, A., Parsons, A., & Trott, C. 2014b, *PhRvD*, **90**, 23019
- Liu, A., & Shaw, J. R. 2020, *PASP*, **132**, 062001
- Liu, A., & Tegmark, M. 2011, *PhRvD*, **83**, 103006
- Liu, A., Tegmark, M., Morrison, S., Lutmirski, A., & Zaldarriaga, M. 2010, *MNRAS*, **408**, 1029
- Madau, P., Meiksin, A., & Rees, M. 1997, *ApJ*, **475**, 429
- Mahesh, N., Bowman, J. D., Mozdzen, T. J., et al. 2021, *ApJ*, **162**, 38
- Mao, Y., Tegmark, M., McQuinn, M., Zaldarriaga, M., & Zahn, O. 2008, *PhRvD*, **78**, 23529
- Mason, C. A., Fontana, A., Treu, T., et al. 2019, *MNRAS*, **485**, 3947
- Mason, C. A., Treu, T., Dijkstra, M., et al. 2018, *ApJ*, **856**, 2
- McGreer, I., Mesinger, A., & D'Odorico, V. 2015, *MNRAS*, **447**, 499
- McMullin, J. P., Waters, B., Schiebel, D., Young, W., & Golap, K. 2007, in ASP Conf. Ser. 376, Astronomical Data Analysis Software and Systems XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco, CA: ASP), **127**
- Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, *MNRAS*, **493**, 1662
- Mesinger, A. 2016, Understanding the Epoch of Cosmic Reionization: Challenges and Progress, Astrophysics and Space Science Library, Vol. 423 (Berlin: Springer), **423**
- Mesinger, A., Aykutan, A., Vanzella, E., et al. 2015, *MNRAS*, **446**, 566
- Mesinger, A., & Haiman, Z. 2004, *ApJL*, **611**, L69
- Millea, M., & Bouchet, F. 2018, *A&A*, **617**, A96
- Mitra, S., Choudhury, T. R., & Ferrara, A. 2015, *MNRAS*, **454**, L76
- Monsalve, R. A., Rogers, A. E. E., Bowman, J. D., & Mozdzen, T. J. 2017, *ApJ*, **847**, 64
- Moore, D. F., Aguirre, J. E., Kohn, S. A., et al. 2017, *ApJ*, **836**, 154
- Moore, D. F., Aguirre, J. E., Parsons, A. R., Jacobs, D. C., & Pober, J. C. 2013, *ApJ*, **769**, 154
- Morales, M., & Wyithe, J. 2010, *ARA&A*, **48**, 127
- Mouri Sardarabadi, A., & Koopmans, L. V. E. 2019, *MNRAS*, **483**, 5480
- 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
- 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
- Paciga, G., Albert, J., Bandura, K., et al. 2013, *MNRAS*, **433**, 639
- Park, J., Mesinger, A., Greig, B., & Gillet, N. 2019, *MNRAS*, **484**, 933
- Parsons, A., Liu, A., Aguirre, J., et al. 2014, *ApJ*, **788**, 106
- Parsons, A., Pober, J., McQuinn, M., Jacobs, D., & Aguirre, J. 2012, *ApJ*, **753**, 81
- Parsons, A. R., & Backer, D. C. 2009, *AJ*, **138**, 219
- Parsons, A. R., Backer, D. C., Foster, G. S., et al. 2010, *AJ*, **139**, 1468
- Parsons, A. R., Liu, A., Ali, Z. S., & Cheng, C. 2016, *ApJ*, **820**, 51
- Patil, A., Yatawatta, S., Koopmans, L., et al. 2017, *ApJ*, **838**, 65
- Patil, A. H., Zaroubi, S., Chapman, E., et al. 2014, *MNRAS*, **443**, 1113
- Patra, N., Parsons, A. R., DeBoer, D. R., et al. 2018, *ExA*, **45**, 177
- Pentericci, L., et al. 2014, *ApJ*, **793**, 113
- Planck Collaboration 2020, *A&A*, **641**, A6
- Planck Collaboration, Ade, P., Aghanim, N., et al. 2016, *A&A*, **594**, A13
- Pober, J., Hazelton, B., Beardsley, A., et al. 2016, *ApJ*, **819**, 8
- Pober, J., Liu, A., Dillon, J., et al. 2014, *ApJ*, **782**, 66
- Pober, J., Parsons, A., Aguirre, J., et al. 2013b, *ApJL*, **768**, L36
- Pober, J., Parsons, A., DeBoer, D., et al. 2013a, *AJ*, **145**, 65
- Price, L. C., Trac, H., & Cen, R. 2016, arXiv:1605.03970
- Pritchard, J., & Loeb, A. 2012, *PPh*, **75**, 86901
- Qin, Y., Mesinger, A., Bosman, S. E. I., & Viel, M. 2021, *MNRAS*, **506**, 2390
- Qin, Y., Mesinger, A., Park, J., Greig, B., & Muñoz, J. B. 2020a, *MNRAS*, **495**, 123
- Qin, Y., Poulin, V., Mesinger, A., et al. 2020b, *MNRAS*, **499**, 550
- Robertson, B., Ellis, R., Furlanetto, S., & Dunlop, J. 2015, *ApJL*, **802**, L19
- Rousseeuw, P. J., & Croux, C. 1993, *J. Am. Stat. Assoc.*, **88**, 1273
- Schenker, M. A., Stark, D. P., Ellis, R. S., et al. 2012, *ApJ*, **744**, 179
- Sims, P. H., & Pober, J. C. 2020, *MNRAS*, **492**, 22
- Singh, S., & Subrahmanyan, R. 2019, *ApJ*, **880**, 26
- Singh, S., Subrahmanyan, R., Udaya Shankar, N., et al. 2017, *ApJL*, **845**, L12
- Smirnov, O. M. 2011, *A&A*, **527**, A106
- Stark, D. P., Ellis, R. S., Chiu, K., Ouchi, M., & Bunker, A. 2010, *MNRAS*, **408**, 1628
- Tan, J., Liu, A., Kern, N. S., et al. 2021, *ApJS*, **255**, 26
- Tegmark, M. 1997, *PhRvD*, **55**, 5895
- Thyagarajan, N., Carilli, C. L., Nikolic, B., et al. 2020, *PhRvD*, **102**, 022002
- Thyagarajan, N., Parsons, A. R., DeBoer, D. R., et al. 2016, *ApJ*, **825**, 9
- Tingay, S. J., Goeke, R., Bowman, J. D., et al. 2013, *PASA*, **30**, e007
- 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
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *NatMe*, **17**, 261
- Wieringa, M. H. 1992, *ExA*, **2**, 203
- Wilensky, M. 2020, Useful Information About TV and Digital Audio RFI As Observed By HERA, HERA Memo 82, (HERA: Carnarvon)
- Wilensky, M. J., Morales, M. F., Hazelton, B. J., et al. 2019, *PASP*, **131**, 114507
- Yatawatta, S. 2015, *MNRAS*, **449**, 4506
- Yatawatta, S. 2016, arXiv:1605.09219
- Zheng, H., Tegmark, M., Buza, V., et al. 2014, *MNRAS*, **445**, 1084