The Astrophysical Journal, 998:33 (44pp), 2026 February 10 https://doi.org/10.3847/1538-4357/ae2d54

First Results from HERA Phase II

Zuhra Abdurashidova[1], Tyrone Adams[2], James E. Aguirre[3], Rushelle Baartman[2], Rennan Barkana[4], Lindsay M. Berkhout[5,6], Gianni Bernardi[2,7,8], Tashalee S. Billings[3], Bruno B. Bizarria[9,10], Judd D. Bowman[5], Daniela Breitman[11], Philip Bull[9,12], Jacob Burba[9], Ruby Byrne[13], Steven Carey[14], Rajorshi Sushovan Chandra[15], Kai-Feng Chen[16,17], Samir Choudhuri[18], Tyler Cox[1], David R. DeBoer[19], Eloy de Lera Acedo[14,20], Matt Dexter[19], Jiten Dhandha[20,21], Joshua S. Dillon[1,19], Scott Dynes[16], Nico Eksteen[2], John Ely[14], Aaron Ewall-Wice[1,22], Nicolas Fagnoni[14], Anastasia Fialkov[20,21], Steven R. Furlanetto[23], Kingsley Gale-Sides[14], Hugh Garsden[9], Adelie Gorce[24], Deepthi Gorthi[1], Ziyaad Halday[2], Bryna J. Hazelton[13,25], Jacqueline N. Hewitt[16,17], Jack Hickish[19], Tian Huang[14], Daniel C. Jacobs[5], Alec Josaitis[14], Nicholas S. Kern[16,26], Joshua Kerrigan[27], Piyanat Kittiwisit[2,12], Matthew Kolopanis[5], Adam Lanman[27], Paul La Plante[28,29], Adrian Liu[1,6], Yin-Zhe Ma[30], David H. E. MacMahon[19], Lourence Malan[2], Cresshim Malgas[2], Keith Malgas[2], Bradley Marero[2], Zachary E. Martinot[3], Lisa McBride[6,24], Andrei Mesinger[11,31], Jordan Mirocha[32,33], Nicel Mohamed-Hinds[13], Mathakane Molewa[2], Miguel F. Morales[13], Julian B. Muñoz[34], Steven G. Murray[11], Bojan Nikolic[14], Hans Nuwegeld[2], Aaron R. Parsons[1,19], Robert Pascua[1,6,35,36], Nipanjana Patra[1,37], Simon Pochinda[14,20], Yuxiang Qin[38], Eleanor Rath[16,17], Nima Razavi-Ghods[14], Daniel Riley[16], Kathryn Rosie[2,39], Mario G. Santos[2,12], Saurabh Singh[15], Dara Storer[13], Hilton Swarts[2], Jianrong Tan[3], Emilie Thélie[34], Pieter van Wyngaarden[2], Michael J. Wilensky[6,40], Peter K. G. Williams[41,42], and Haoxuan Zheng[17]

(The HERA Collaboration)

[1] Department of Astronomy, University of California, Berkeley, CA, USA
[2] South African Radio Astronomy Observatory, Black River Park, 2 Fir Street, Observatory, Cape Town, 7925, South Africa
[3] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, USA
[4] School of Physics and Astronomy, Tel-Aviv University, Tel-Aviv, 69978, Israel
[5] School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA
[6] Department of Physics and Trottier Space Institute, McGill University, 3600 University Street, Montreal, QC H3A 2T8, Canada
[7] INAF-Istituto di Radioastronomia, via Gobetti 101, 40129 Bologna, Italy
[8] Department of Physics and Electronics, Rhodes University, P.O. Box 94, Grahamstown, 6140, South Africa
[9] Jodrell Bank Centre for Astrophysics, University of Manchester, Manchester, M13 9PL, UK
[10] Astrophysics division—INPE, Instituto Nacional de Pesquisas Espaciais, São José dos Campos—SP, Brazil
[11] Scuola Normale Superiore, 56126, Pisa, PI, Italy
[12] Department of Physics and Astronomy, University of Western Cape, Cape Town, 7535, South Africa
[13] Department of Physics, University of Washington, Seattle, WA, USA
[14] Cavendish Astrophysics, University of Cambridge, Cambridge, UK
[15] Raman Research Institute, Bangalore, India
[16] MIT Kavli Institute, Massachusetts Institute of Technology, Cambridge, MA, USA
[17] Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA
[18] Centre for Strings, Gravitation and Cosmology, Department of Physics, Indian Institute of Technology Madras, Chennai 600036, India
[19] Radio Astronomy Lab, University of California, Berkeley, CA, USA
[20] Kavli Institute for Cosmology, Madingley Road, Cambridge CB30HA, UK
[21] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB30HA, UK
[22] Department of Physics, University of California, Berkeley, CA, USA
[23] Department of Physics and Astronomy, University of California, Los Angeles, CA, USA
[24] Institut d'Astrophysique Spatiale, CNRS, Université Paris-Saclay, 91405 Orsay, France
[25] eScience Institute, University of Washington, Seattle, WA, USA
[26] Department of Physics, University of Michigan, Randall Lab, 450 Church St., Ann Arbor, MI 48109, USA
[27] Department of Physics, Brown University, Providence, RI, USA
[28] Department of Computer Science, University of Nevada, Las Vegas, NV 89154, USA
[29] Nevada Center for Astrophysics, University of Nevada, Las Vegas, NV 89154, USA
[30] Department of Physics, Stellenbosch University, Matieland, Western Cape, 7602, South Africa
[31] Department of Physics and Astronomy "Ettore Majorana", University of Catania, Via Santa Sofia 64, 95123 Catania, Italy
[32] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[33] California Institute of Technology, 1200 E. California Boulevard, Pasadena, CA 91125, USA
[34] Department of Astronomy, The University of Texas at Austin, 2515 Speedway, Stop C1400, Austin, TX 78712, USA
[35] Dunlap Institute for Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, ON, M5S 3H4, Canada
[36] Perimeter Institute for Theoretical Physics, 31 Caroline Street, North Waterloo, ON, N2L 2Y5, Canada
[37] International Centre for Radio Astronomy Research, Curtin University, Bentley WA 6102, Australia
[38] Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT, Australia
[39] South African Astronomical Observatory, 1 Observatory Road, Observatory, Cape Town, 7925, South Africa
[40] CITA National Fellow
[41] Center for Astrophysics, Harvard & Smithsonian, Cambridge, MA, USA
[42] American Astronomical Society, Washington, DC, USA

Received 2025 August 19; revised 2025 November 26; accepted 2025 December 14; published 2026 February 2

Abstract

We report the first upper limits on the power spectrum of 21 cm fluctuations during the Epoch of Reionization and Cosmic Dawn from Phase II of the Hydrogen Epoch of Reionization Array (HERA) experiment. HERA Phase II constitutes several significant improvements in the signal chain compared to Phase I, most notably resulting in expanded frequency bandwidth, from 50–250 MHz. In these first upper limits, we investigate a small two-week subset of the available Phase II observations, with a focus on identifying new systematic characteristics of the instrument, and establishing an analysis pipeline to account for them. We report 2σ upper limits in eight spectral bands, spanning $5.6 \leqslant z \leqslant 24.4$ that are consistent with thermal noise at the 2σ level for $k \gtrsim 0.6$–$0.9\,h\,\mathrm{Mpc}^{-1}$ (band dependent). Our tightest limit during Cosmic Dawn ($z > 12$) is $1.13 \times 10^6\,\mathrm{mK}^2$ at ($k = 0.55\,h\,\mathrm{Mpc}^{-1}$, $z = 16.78$), and during the EoR ($5.5 < z < 12$), it is $1.78 \times 10^3\,\mathrm{mK}^2$ at ($k = 0.70\,h\,\mathrm{Mpc}^{-1}$, $z = 7.05$). We find that mutual coupling has become our dominant systematic, leaking foreground power that strongly contaminates the low-k modes, resulting in the loss of modes from $k = 0.35$ to 0.55 compared to Phase I data.

Unified Astronomy Thesaurus concepts: Reionization (1383); Galaxy formation (595); Radio astronomy (1338); Astronomy data analysis (1858); Radio interferometers (1345)

1. Introduction

Observations of the redshifted 21 cm spectral line of neutral hydrogen (H I) have great potential to inform us about key physical processes throughout cosmic history: from the Dark Ages (A. Loeb 2006) and Cosmic Dawn (S. R. Furlanetto et al. 2006), through the Epoch of Reionization (A. Mesinger 2016) and into the post-reionization Universe (M. Amiri et al. 2023). In the pre-reionization Universe, the line is sensitive to both cosmological and astrophysical processes, including the assembly of cosmic structure, the birth of the first stars, and the growth of the first galaxies, through its dependence on the local density and thermal conditions of the intergalactic medium (IGM; S. R. Furlanetto et al. 2006; A. Mesinger et al. 2010). By measuring angular and line-of-sight fluctuations in the 21 cm brightness temperature with respect to the cosmic microwave background (CMB), we gain access to information about the underlying sources of heating and ionization, and may be able to precisely constrain properties of these sources (Y. Mao et al. 2008; A. H. Patil et al. 2014; J. C. Pober et al. 2014; B. Greig & A. Mesinger 2015; A. Liu & A. R. Parsons 2016; N. S. Kern et al. 2017; J. B. Muñoz et al. 2018). For reviews of the physics of the 21 cm line and its applications, see, e.g., B. Ciardi & A. Ferrara (2005), S. R. Furlanetto et al. (2006), M. F. Morales & J. S. B. Wyithe (2010), J. R. Pritchard & A. Loeb (2012), A. Mesinger (2016), and A. Liu & J. R. Shaw (2020).

Detection of the cosmological redshifted 21 cm line is possible with low-frequency radio telescopes operating at frequencies between 10 and 1420 MHz. In particular, the frequency range 47–234 MHz corresponds to $z = 5.1$–29.3, encompassing Cosmic Dawn and Reionization. Several experiments—past, present, and upcoming—have been built to observe the 21 cm signal from the Epoch of Reionization and beyond over the past two decades. These experiments can be categorized into two types. Interferometers measure the spatial fluctuations of the 21 cm signal, commonly summarized by the spherically averaged *power spectrum* $\Delta^2_{21}(k)$. Such experiments have placed increasingly stringent upper limits on this quantity. This includes completed experiments, such as the Giant Meter Wave Radio Telescope (G. Paciga et al. 2011) and the Donald C. Backer Precision Array for Probing the EoR (PAPER; A. R. Parsons et al. 2010, 2014; C. Cheng et al. 2018; M. Kolopanis et al. 2019). It also includes several ongoing experiments, such as the Low Frequency Array (LOFAR; M. P. van Haarlem et al. 2013; A. H. Patil et al. 2017; B. K. Gehlot et al. 2018; F. G. Mertens et al. 2020, 2025; A. Acharya et al. 2024), the Murchison Widefield Array (MWA; S. J. Tingay et al. 2013; J. S. Dillon et al. 2014, 2015; A. P. Beardsley et al. 2016; A. Ewall-Wice et al. 2016; D. C. Jacobs et al. 2016; N. Barry et al. 2019; W. Li et al. 2019; C. M. Trott et al. 2020; M. Rahimi et al. 2021; S. Yoshiura et al. 2021; M. Kolopanis et al. 2023; M. J. Wilensky et al. 2023; C. D. Nunhokee et al. 2025), the Long-Wavelength Array (LWA; M. W. Eastwood et al. 2019; H. Garsden et al. 2021), the Upgraded Giant Metrewave Ratio Telescope (Y. Gupta et al. 2017; S. Chatterjee & S. Bharadwaj 2019), the Nancay Upgrading LOFAR experiment (P. Zarka et al. 2012; S. Munshi et al. 2024; S. Munshi et al. 2025a), and the Hydrogen Epoch of Reionization Array (HERA; D. R. DeBoer et al. 2017; Z. Abdurashidova et al. 2022; T. H. C. Z. Abdurashidova et al. 2023). Future experiments include the Square Kilometer Array (P. Dewdney et al. 2016) and LOFAR 2.0 (E. Orrú et al. 2024).

In addition to these interferometric experiments, a number of single-antenna experiments that measure the sky-averaged 21 cm temperature as a function of frequency/redshift (the "global signal") have been constructed. Notably, the Experiment to Detect the Global Eor Signature (EDGES; A. E. E. Rogers & J. D. Bowman 2012) has claimed a detection of 21 cm absorption during Cosmic Dawn (J. D. Bowman et al. 2018), though the amplitude and profile of the absorption feature are difficult to account for under standard astrophysical and cosmological scenarios, requiring either cooling of the IGM below the adiabatic limit (R. Barkana et al. 2018; J. B. Muñoz & A. Loeb 2018), a high-redshift radio background in excess of the CMB (A. Ewall-Wice et al. 2018; C. Feng & G. Holder 2018; A. Fialkov & R. Barkana 2019; J. Mirocha & S. R. Furlanetto 2019), or an unmodeled systematic contaminating the data analysis (R. Hills et al. 2018; R. F. Bradley et al. 2019; S. Singh & R. Subrahmanyan 2019; P. H. Sims & J. C. Pober 2020; S. G. Murray et al. 2022; J. Cang et al. 2024). Other experiments that have published global signal data include the Large-Aperture Experiment to Detect the Dark Ages (D. C. Price et al. 2018), the Broadband Instrument for Global Hydrogen Reionization (M. Sokolowski et al. 2015), and the Shaped Antenna measurement of the background RAdio Spectrum (SARAS; N. Patra et al. 2015; S. Singh et al. 2017). Notably, the third-generation SARAS instrument recently published a *nondetection* of the EDGES absorption feature during Cosmic Dawn, with a confidence level of 2σ (H. T. J. Bevins et al. 2022; S. Singh et al. 2022). Upcoming

experiments of this variety include the Radio Experiment for the Analysis of Cosmic Hydrogen (E. de Lera Acedo et al. 2022), the Remote H I eNvironment Observer (P. Bull et al. 2024), the Mapper of the IGM Spin Temperature (R. A. Monsalve et al. 2024), and Probing Radio Intensity at High-Z from Marion (L. Philip et al. 2019).

In this paper, we present the first power spectrum upper limits from Phase II of the HERA experiment. HERA Phase I (D. R. DeBoer et al. 2017) operated with a subset of the full complement of antennas (<60), and used repurposed dipoles from the PAPER experiment, sensitive between 100 and 200 MHz. Phase I observations culminated with an observing season in 2017–2018 resulting in two successive power spectrum upper limits; first from a two-week subset of the data (HERA Collaboration 2022a; hereafter H22a), and finally from the full 94-night dataset (HERA Collaboration 2023; hereafter H23). In the meantime, Phase II has been steadily rolling out. Phase II (L. M. Berkhout et al. 2024) includes several upgrades: a new signal chain, correlator, and upgraded Vivaldi feeds that together extend the bandwidth out to 47–234 MHz. The redshifts newly observable in Phase II are of particular interest; they now cover the absorption feature reported by EDGES (J. D. Bowman et al. 2018) as well the tail end of reionization ($z = 5-6$), which has received considerable interest lately, with several studies based on the Lyα forest of quasi-stellar objects (QSOs) placing the end of reionization at $z \sim 5.3$–5.5 (S. E. I. Bosman et al. 2022; Y. Zhu et al. 2022; Y. Qin et al. 2024). In addition to the upgraded system components, Phase II has added more antennas, increasing the instantaneous sensitivity of the telescope. Here, we present an analysis of a 2 week subset of data observed in 2022 October, during which 180 dual-polarized antennas were online. For more details on the HERA Phase II system design, see L. M. Berkhout et al. (2024).

The greatest challenge to making a detection of 21 cm fluctuations in the pre-reionization Universe is the confluence of spectrally structured systematics and bright foregrounds that outshine the signal by up to 4 orders of magnitude. Since the bright foregrounds are most readily distinguished from the 21 cm signal by their different spectral behavior—with foregrounds intrinsically spectrally smooth and the 21 cm signal possessing spectral structure on all scales—it is important to design and calibrate 21 cm experiments to minimize the structure they imprint on the foregrounds. There are many potential sources of such structure, for example, environmental factors such as ionospheric refraction and radio frequency interference (RFI; M. J. Wilensky et al. 2020, 2023; B. K. Gehlot et al. 2024; S. Munshi et al. 2025b), signal-chain effects such as cable reflections (A. P. Beardsley et al. 2016) and crosstalk, errors in analysis such as primary-beam nonredundancy causing miscalibration (N. Orosz et al. 2019; H. Kim et al. 2022, 2023) or incomplete sky models affecting absolute calibration (N. Barry et al. 2016; R. Byrne et al. 2019), and structural concerns such as incomplete uv coverage (A. Liu et al. 2014; S. G. Murray & C. M. Trott 2018). One of the motivations of this paper is to uncover and quantify the systematics present in updated Phase II system (especially those that are new or more prominent in comparison to Phase I), and to demonstrate that they can be appropriately modeled and/or mitigated. One particular systematic that has emerged as the most detrimental and difficult to handle in Phase II is that of mutual coupling (MC): the reflection or re-emission of sky signal by one antenna into another (A. Camps et al. 1998; N. S. Kern et al. 2019, 2020a; A. T. Josaitis et al. 2021; Q. Gueuning et al. 2022; N. Charles et al. 2024; R. Pascua et al. 2024; O. S. D. O'Hara et al. 2025; E. Rath et al. 2025). MC is present at higher levels in Phase II than was seen in Phase I data. This is attributable to the increased sensitivity of the new wideband Vivaldi feeds toward the horizon and their suspension above their dish without the shielding of the surrounding cavity used for the Phase I dipoles, which created unwanted reflections within the dish. Currently, the feeds are in full view of neighboring antennas—both their feeds and the dishes themselves. While we have worked to develop tools to understand and mitigate MC in our analysis pipeline (N. S. Kern et al. 2019, 2020b; A. T. Josaitis et al. 2021; N. Charles et al. 2024; R. Pascua et al. 2024; E. Rath et al. 2025), it remains our dominant systematic in the crucial k-modes just beyond those dominated by intrinsic foregrounds. We will discuss this systematic at some length throughout this work.

To date, the limits presented in H23 remain the deepest limits of the 21 cm power spectrum from any experiment at redshifts 8 and 10. These limits were made with 94 nights of observing (of 12 hr each) with a maximum of 41 unflagged antennas per night. The limits, specifically $\Delta^2_{21}(k = 0.34\,h\ \mathrm{Mpc}^{-1}) \leqslant 457\ \mathrm{mK}^2$ at $z = 7.9$ and $\Delta^2_{21}(k = 0.36\,h\ \mathrm{Mpc}^{-1}) \leqslant 3496\ \mathrm{mK}^2$ at $z = 10.4$, are consistent with thermal noise over a wide range of k, indicating that further integration should yield corresponding improvement. Notably, both of the previous HERA upper limits were supported by extensive simulation and statistical validation (J. Tan et al. 2021; J. E. Aguirre et al. 2022), which cataloged small sources of signal loss that were corrected in the final limits. This analysis in this paper extends these supporting validation tests with updated and expanded simulations that cover the larger set of antennas in the Phase II dataset.

Although previous HERA results were only upper limits, a range of inferences with different physical models were unanimous in concluding that these limits rule out the class of "cold reionization" models (HERA Collaboration 2022b, hereafter H22b; strengthened by H23).

Cold reionization is the regime in which large-scale reionization occurs without significant heating of the IGM (due to, for example, inefficient production of X-rays from high-mass X-ray binaries). Cold reionization is the scenario that naturally produces the largest amplitude Δ^2_{21} after reionization begins ($z \sim 6$–12; A. Mesinger et al. 2014; J. C. Pober et al. 2015). Prior to being ruled out in H23, cold reionization scenarios were already disfavored by prior limits set by both LOFAR (R. Ghara et al. 2020; B. Greig et al. 2021a; R. Ghara et al. 2025) and the MWA (R. Ghara et al. 2021; B. Greig et al. 2021b). Beyond ruling out cold reionization, H23 constrained the properties of early high-mass X-ray binaries (HMXBs); under the assumption that these systems dominate the heating of the IGM before and during reionization, their soft-band X-ray luminosity per star formation rate ($L_{X<2\mathrm{keV}}/\mathrm{SFR}$) was inferred to be higher than that of local HMXBs at more than 3σ, and instead consistent with a population of low-metallicity HMXBs (T. Fragos et al. 2013; P. Madau & T. Fragos 2017; H. D. Kaur et al. 2022). However, this conclusion is conditional on the astrophysical model employed. For example, H. Lazare et al. (2024) showed that if star formation is very efficient inside the first, molecularly cooled galaxies, the constraints on $L_{X<2\mathrm{keV}}/\mathrm{SFR}$ could weaken considerably, though this depends strongly on the unknown

properties of such Population III dominated galaxies (e.g., H. Lazare et al. 2024; D. Breitman et al. 2025, in preparation).

Meanwhile, over the past several years, our understanding of the $z > 5$ Universe has continued to improve thanks to measurements beyond the 21 cm line. For reionization, the detection of high-H I opacity regions in quasar spectra at $z \lesssim 5$ has provided increasingly strong evidence that the tail end of reionization stretches to between $z \sim 5$ and 5.5 (G. D. Becker et al. 2021; T. R. Choudhury et al. 2021; S. E. I. Bosman et al. 2022; Y. Zhu et al. 2022; P. Gaikwad et al. 2023; Y. Qin et al. 2024). But the James Webb Space Telescope (JWST; J. P. Gardner et al. 2006) observations have shown that star formation is substantially more common at $z \gtrsim 10$ than expected from models calibrated to Hubble Space Telescope (HST) observations (e.g., G. C. K. Leung et al. 2023; C. T. Donnan et al. 2024; S. L. Finkelstein et al. 2024) and that the observed galaxy population is extremely blue (M. W. Topping et al. 2022; F. Cullen et al. 2023). These observations suggest that $z \gtrsim 10$ galaxies might be more efficient at emitting ionizing photons than previously expected (V. Gelli et al. 2024; J. B. Muñoz et al. 2024; I. Nikolić et al. 2024), which could result in an earlier start to reionization. In such a scenario, recombinations inside IGM clumps would be essential in extending the later stages of the EoR in order to match the Lyα forest observations (F. B. Davies et al. 2021; Y. Qin et al. 2024). Meanwhile, Lyα line emission from galaxies—which should be absorbed by the IGM once it is substantially neutral—has proven to be surprisingly common even at very high redshifts (e.g., A. J. Bunker et al. 2023; J. Witstok et al. 2025). These varied observations have raised even more questions about the reionization process.

JWST has also raised new questions about the importance of active galactic nuclei (AGNs) to the Cosmic Dawn era. Signatures of accreting supermassive black holes have been found in a surprisingly large number of high-z sources, including UV-luminous galaxies (A. J. Bunker et al. 2023; A. D. Goulding et al. 2023; M. Castellano et al. 2024; R. Maiolino et al. 2024) and the so-called "little red dots," compact sources with broad emission lines that may be due to gas surrounding accreting black holes (e.g., J. E. Greene et al. 2024; J. Matthee et al. 2024; but see also G. C. K. Leung et al. 2024 for other explanations). The prevalence of these sources (accounting for up to ~10% of the galaxy population) suggests that accreting black holes may have played a larger role in the early Universe than previously expected, though this depends on the (heretofore poorly understood) nature of the central sources (e.g., G. C. K. Leung et al. 2024; R. Maiolino et al. 2025). An additional population of accreting black holes that are X-ray luminous may affect the early thermal history of the IGM, which will be tested by HERA and other 21 cm experiments.

This paper is organized as follows. In Section 2 we describe the dataset analyzed in this work, including the instrumental configuration and data selection. In Section 3 we detail the analysis pipeline we have developed, highlighting key differences compared to our previous analyses of Phase I data (H22a; H23). In Section 4 we demonstrate the validity of our pipeline by applying it to detailed end-to-end mock simulations, as well as performing statistical tests designed to expose biases in our power spectrum estimates. In Section 5 we present the main results of the paper: both cylindrically averaged spectra and spherically averaged upper limits. In Section 6 we explore the impact of these new limits on our understanding of astrophysics during Cosmic Dawn, and finally in Section 7, we conclude with a summary and prospectus for future HERA analyses, given the large amount of data already taken.

Throughout this work, we adopt the cosmology of P. A. R. Ade et al. (2016), namely $\Omega_\Lambda = 0.68440$, $\Omega_b = 0.04911$, $\Omega_c = 0.26442$, and $h = 0.6727$.

2. Observations

2.1. The HERA Phase II Telescope

The HERA telescope is located at the SARAO site in the Karoo desert in South Africa. It consists of 350 wire-mesh parabolic dishes, each of which has a diameter of ~14 m and contains a suspended feed (D. R. DeBoer et al. 2017). The dishes are arranged in a tightly packed hexagon (J. S. Dillon & A. R. Parsons 2016) spaced 14.6 m apart (center to center), with the dishes almost touching each other. Note that 320 antennas form a core with a maximum inter-antenna distance of 292 m, along with two concentric layers of outriggers that increase the maximum baseline length to 876 m.

This array layout maximizes sensitivity to a small number of (mostly large-scale) k-modes, making HERA an extremely sensitive telescope for these cosmologically relevant modes, as well as enabling the redundant-calibration approach discussed in Section 3. Figure 1 shows HERA's antenna layout, comparing the antennas used for our previous limits to those used here.

Drift scanning from its location in South Africa, HERA observes a stripe across the $-30°$ decl. This stripe crosses the Galactic center, rising to a galactic decl. of $-80°$ (see Figure 2). Observing constraints include minimizing foreground power from the Galaxy, bright point sources, and the Sun. This results in an optimal season spanning the southern summer, generally from September through April of the following calendar year.

During a season, the number of antennas that are online is reasonably constant—generally upgrades and maintenance are most active during the off-season. In this paper, we report a small subset of a single observing season, 2022–2023.

The most important development between HERA's first two reported upper limits and those presented here is the system change from Phase I to Phase II. The Phase I system covered 100−200 MHz (z = 6–13), and reused several components from its predecessor PAPER (A. R. Parsons et al. 2010), including the dipole feeds, cables, signal processing boards (A. Parsons et al. 2006), xGPU correlator (A. Parsons et al. 2008; M. Clark et al. 2013), and some analog components. The new Phase II instrument encompasses several major upgrades aimed at increasing bandwidth and reducing systematics. Bandwidth was increased by changing the feed and upgrading the digitizers. The PAPER feeds were replaced with broader-band Vivaldi-style feeds (N. Fagnoni et al. 2021a), extending the usable frequency range on both ends, to 47−234 MHz ($z \sim$ 5–27). On the low-band end, this covers the redshift range in which Cosmic Dawn is expected to occur (J. R. Pritchard & A. Loeb 2012; J. D. Bowman et al. 2018), while on the upper end, this covers the tail end of reionization, which may extend all the way to $z \sim 5.3$ (S. E. I. Bosman et al. 2022). The analog system, signal processing boards (J. Hickish et al. 2016), and correlator were also upgraded to match the extended frequency

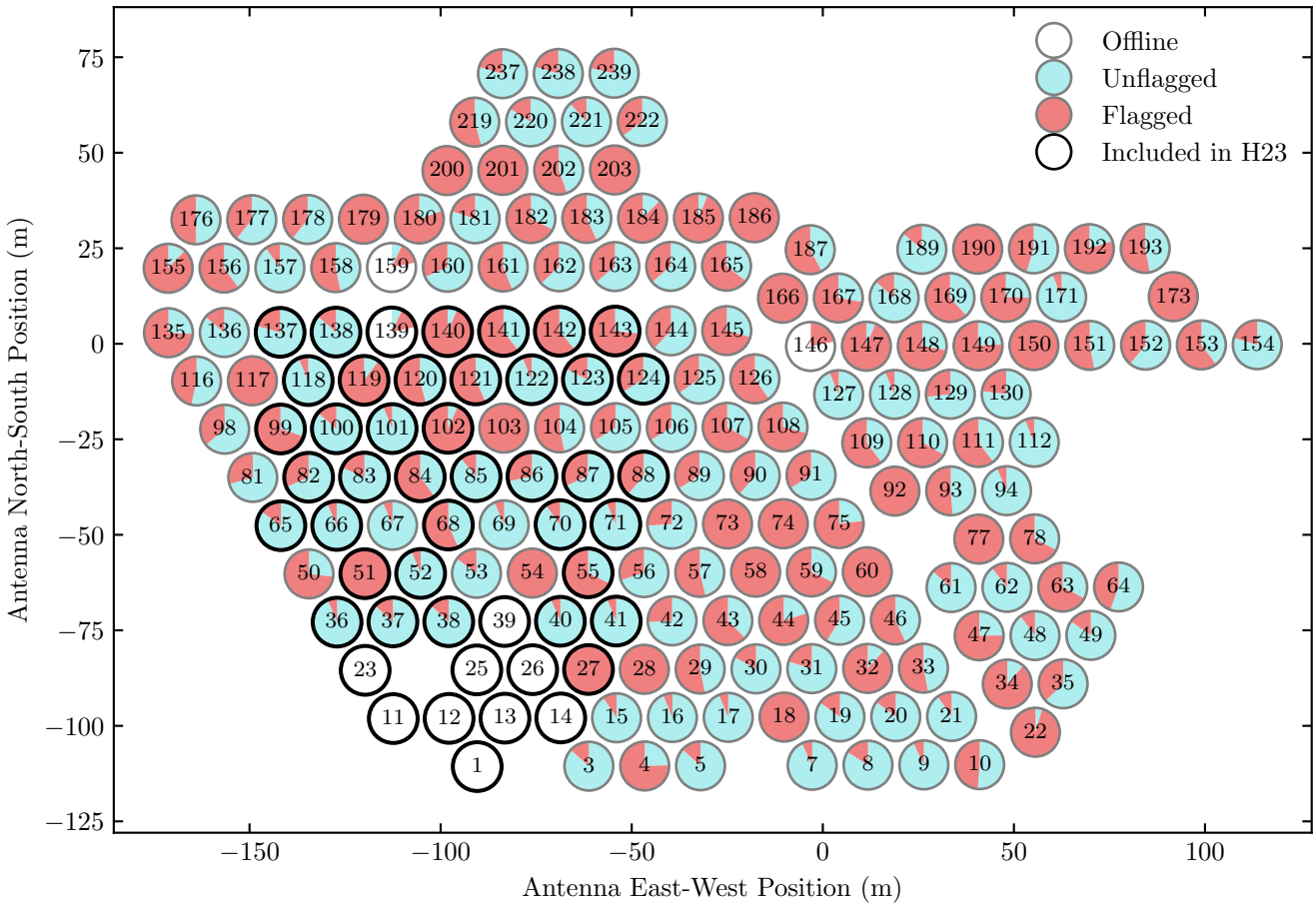

Figure 1. Map of antennas used at least once within this dataset, with inner pie-charts indicating the fraction of time (i) observed without flagging (blue), (ii) flagged (red), and (iii) offline (white). Antennas highlighted with a thick black border indicate those used in H23. Two nights completely flagged due to lightning (see Table 1) are not considered here. The indicated flagging amounts include data quality checks (see Section 3.3) and per-antenna flag synthesis (see Section 3.4.1), but not channel-based flags from, e.g., RFI. The displayed fractions for each antenna average across both feed polarizations. The total time observed across the dataset is 136.3 hr. Over 120 antennas observed usable data during this dataset, compared to the 48 in H23, corresponding naively to a ∼5−10-fold increase in sensitivity.

range of the feeds and accommodate a larger number of antennas (L. M. Berkhout et al. 2024). Finally, the 100 m analog cables were replaced with an RF over fiber (RFoF) system. Use of fiber reduced re-radiation from the coaxial cables and allowed the cables to be made 500 m long. This increased length moved reflections to delays corresponding to k modes of reduced cosmological significance. For this reason, cable reflections are no longer fit for as a part of calibration as was necessary in H22a and H23 (see N. S. Kern et al. 2019, 2020b for the methodology).

2.2. Data Selection

HERA records continuously through the observing season from sunset to sunrise for an average of 12 hr per night. The data used in this paper are a 14-night subset of the 2022–2023 observing season, starting on 2022 October 8–9 (JD 2459861) and ending on 2022 October 23–24 (JD 2459876). Of this contiguous set of nights, two were excluded from the analysis (2459865 and 2459875) due to a high prevalence of strong broadband RFI throughout the nights, which we attribute to lightning storms W. Heiligstein & D. Jacobs (2023). Within the remaining 14 nights, several more hours were flagged for the entire array due to similar broadband RFI. The quality metrics used for excluding observing times were determined based only on the autocorrelations—a small subset of the full data—prior to any of the analysis steps described in Section 3. Table 1 lists the nights and LSTs that were affected. In total, not including the two full nights that were excluded, ∼7.5 observing hr or 5.2% of the remaining 144 hr were flagged due to this broadband RFI.

Figure 2 illustrates the LST coverage of this dataset, comparing it to the Phase I dataset reported in H23. The background in this figure represents the foreground intensity, here computed with the `pygdsm`[43] software. Since HERA is a fixed, zenith-pointing instrument, its observational footprint is confined to a stripe of constant decl., here demarcated by the white dashed lines—limited by the FWHM of the primary beam, which, at the lowest operable frequency, is (∼10°). However, HERA receives emission well beyond this stripe: the primary beam is illustrated in Figure 2 by the gray contours, which denote the 1% (dashed) and 0.2% (dotted) sensitivities, respectively, while the sensitivity at the horizon is estimated to be ∼0.1% (N. Fagnoni et al. 2021a). Sufficiently bright sources anywhere above the horizon significantly affect measured visibilities. The inlaid gray histogram indicates the number of independent observations at each LST, where independent observations are drawn from different nights, antennas, polarizations, and integrations within an LST bin. As can be seen, the dataset we present here covers LSTs between

[43] https://github.com/telegraphic/pygdsm

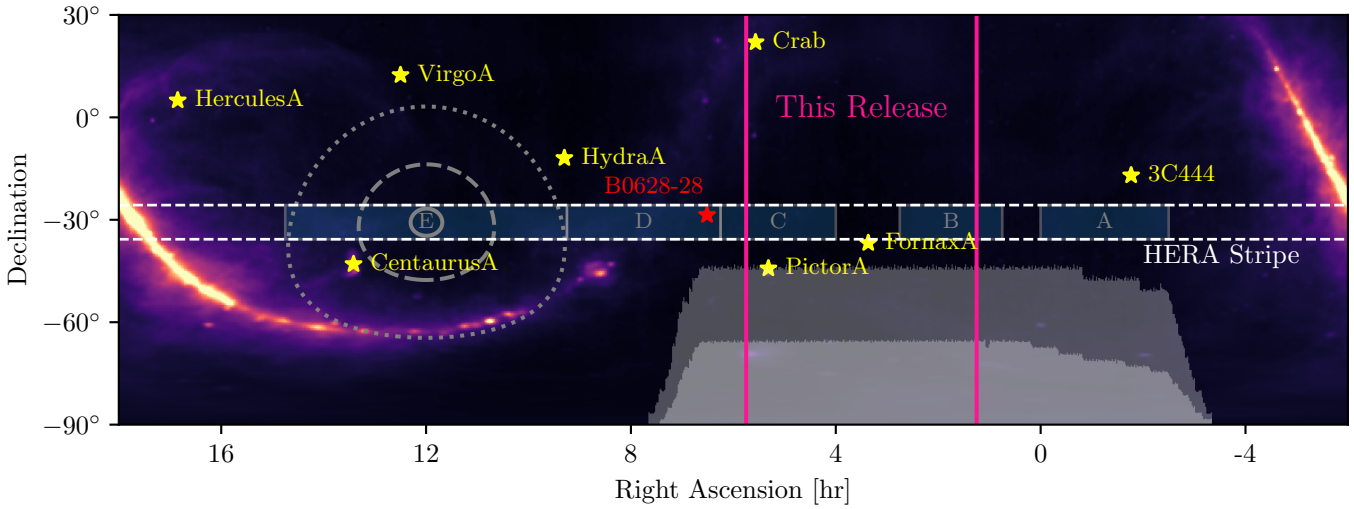

Figure 2. The sky coverage of this dataset, compared to previous HERA limits. The background shows the diffuse sky model of A. De Oliveira-Costa et al. (2008) computed with `pygdsm` (D. C. Price 2016) with an overlay of the "HERA stripe" indicating the region of sky HERA can observe. Inset transparent-blue regions indicate the fields included as part of the H23 upper limits. In this analysis, we include just one field, from LST 1.25–5.75 hr, indicated by the pink vertical lines. Note that while we denote HERA's field of view by the dashed white stripes, corresponding to the FWHM of the primary beam (at 150 MHz), the beam has nonnegligible sensitivity out to large zenith-angles, which we illustrate with the gray contours. The dashed (dotted) contour corresponds to the 1% (0.2%) sensitivity of the primary beam at 150 MHz. The beam has ∼0.1% sensitivity down to the horizon. The inset gray histograms indicate the LST-coverage of the dataset we describe here. The taller background shaded region has a height proportional to the number of antenna-feed-integrations a particular LST was observed (i.e., the number of integrations observed by any antenna on any feed-polarization at that LST), while the shorter foreground histogram is the resulting coverage after antenna flagging (see Figure 1). For reference, the maximum of the background histogram is 65,394 antenna-polarization-integrations (10 s each) within LST bins of 270.5 s (see Section 3.9.1), while the maximum number of *unflagged* observations is just over half this value.

Table 1
Nights and LSTS for Which the Entire Observation Was Flagged due to Excess Broadband RFI, Believed to Be Attributable to Lightning Storms

Night (JD)	LSTs Flagged (hr)
2459863	20^h46^m–23^h39^m
2459865	ALL
2459869	21^h10^m—0^h10^m, 7^h05^m–7^h12^m
2459872	21^h22^m–22^h20^m
2459875	ALL
2459876	21^h38^m–21^h52^m, 7^h26^m–7^h39^m

Note. In general, partially flagged nights were only flagged at the beginning or end of the night (or both) so as not to introduce large flagging gaps in an individual night.

∼21 and 7 hr, which includes Fields A, B, C, and part of D from H23. In this work, we do not split the observations into separate fields as in previous works, rather reporting the power spectrum upper limits derived from the full range of 1.25–5.75 hr, which have maximum coverage, have relatively low foreground amplitude, and avoid problematic sources such as the setting of the Galactic center and the bright, high-rotation-measure pulsar B0628-28.

Figure 1 shows the antennas that were online during the observations included in this analysis, as well as the relative fraction of time they were unflagged (blue), flagged (red), or offline (white). In this figure, the flags include the full per-antenna flags, but neglect frequency-dependent flags (e.g., due to RFI; see Section 3.4.1). The total for each antenna includes both polarization feeds. Some antennas (e.g., 139, 159) were brought online mid-way through the 2 week observing period. Note that certain antennas (e.g., 186), though technically online, were always flagged. Figure 1 also shows the antennas that were included in H23 with darker outlines, highlighting the increased number of antennas in this dataset. In summary, 180–183 antennas were online for each night in this dataset, of which 138 were unflagged for some portion of the dataset. The shortest baselines are 14.6 m in length, and the longest (unflagged) baseline in the dataset is ∼280 m in the east–west direction. We describe our antenna-based flagging metrics in more detail in Section 3.4, and summarize the details of the data selection in Table 2.

This dataset represents a small fraction, roughly 10%, of HERA's 2022–2023 observing season. It was not chosen based primarily on its quality; rather, it was simply the first set of 14 nights observed that passed basic full-night quality checks (to avoid lightning storms). During the later part of the season, the addition of new antennas and repairs improved the data quality somewhat, and fall weather brought fewer lightning storms. Compared to the dataset reported here, the full season represents a 10-fold increase in the number of observed nights/hours that remain after autocorrelation-based flagging, as well as a modest increase in the average number of antennas that survive quality checks. In total, the 2022–2023 season has just over 1300 hr of unflagged data, with an average of 140 unflagged antennas, while the dataset analyzed here contains 129 hr of data with an average of 131 unflagged antennas.

3. Analysis Pipeline

The data analysis pipeline consists of several stages of data flagging, calibration, filtering, and averaging. Almost all stages of analysis have been upgraded or modified with respect to H23. In this section, we detail the full pipeline, highlighting the main modifications compared to H23. For the reader already familiar with previous HERA analyses (H22a; H23), we provide a compact overview of the main differences introduced in this work following the full description in Section 3.11.

Table 2
Observation Characteristics for HERA's 2022–2023 Observing Season

	Here	Full Season
Array Location	−30.72°S, 21.43°E	
Avg. Antennas Connected	180	195
Avg. Antennas Unflagged	131	140
Shortest Baseline	14.6 m	14.6 m
Longest Avail. Baseline	280 m	755 m
Minimum Frequency	46.92 MHz	
Maximum Frequency	234.30 MHz	
Channels	1536	
Channel Width, $\Delta\nu$	122 kHz	
Integration Time, Δt	9.6 s	
Per-Night Obs.	12 hr	
Total Nights Used	14	147
Unflagged Obs. Hours	129	1312
Raw Data Volume (Night)	2.7 TB	2.7 TB
Raw Data Volume (Full)	36 TB	∼350 TB

3.1. Definition of Terms

We first define some notation and nomenclature that recurs throughout the upcoming analysis discussion.

The basic measurement of the interferometer is a visibility V_{ij}^{pq} correlating feed p on antenna i with q on j. The correlator integrates the instantaneous visibility over channel width $\Delta\nu$ and over time Δt (see Table 2). Colloquially known as "integrations," these visibilities measured between antennas in the HERA hexagonal grid offer repeated or "redundant" measurements of the same correlation.[44] These redundant measurements of the sky all occupy the same point in uv space and are nominally only different up to thermal noise. Often one examines a particular "redundant group" $\mathcal{G}_{ij} = \{(i, j), (k, l),\dots\}$ of $|\mathcal{G}_{ij}|$ baselines that all occupy the same uv point or "unique baseline" vector. We uniquely label the baseline group by one of its elements (here ij). The number of unique baselines in the dataset is $N_{\rm ubl}$ (this can change over time as antennas are added/removed or flagged). Visibilities from redundant baselines are eventually averaged to a single "redundantly averaged" visibility. Once this averaging is done, it is often practical to label the redundantly averaged visibility according to the "key" baseline, so that $V_{ij}^{pq} \equiv V_{\mathcal{G}_{ij}}^{pq}$. The context will make clear when a visibility is a redundant average.

Throughout the pipeline, we often make use of the "expected" variance of a particular visibility. Whenever we reference the "expected variance" of a cross-correlation, we are referring to that given by the radiometer equation, where the system temperature is estimated as the *calibrated autocorrelation*[45]:

$$\mathrm{Var}(V_{ij}^{pq}) = \frac{|V_{ii}^{pp}||V_{jj}^{qq}|}{\Delta t \Delta\nu N_{\rm samples}}. \quad (1)$$

Here, $N_{\rm samples}$ is the number of integrations or baselines that have been coherently averaged together into V_{ij}^{pq} (e.g., if a redundantly averaged visibility from group $\mathcal{G}$ is under consideration, the number of samples includes $|\mathcal{G}|$). Very often, since well-calibrated autocorrelations should be nearly equal across antennas, we use the *average* autocorrelation across antennas for a given polarization, $\overline{V^{pp}}$, instead of each V_{ii}^{pp} individually.

3.2. Overview of Pipeline

The analysis approach taken here is to estimate the power spectrum at each uv point and then average these power spectra to get a single result. Averaging in this "incoherent" way requires phase agreement between redundant baselines and correctable stability in time and frequency but not between all baselines as would be required in an imaging approach (M. F. Morales et al. 2019). The remaining challenge, still significant, is to average coherently across redundant baselines, short timescales, and many nights and then average power spectra over sidereal times. The coherent-average pipeline used here uses filters and statistical tests across time, frequency, and night to solve for antenna gains, mask interference, and check for system failures. The order of operations and choice of algorithms is constrained by the amount of data that can be held in memory, the speed of data access patterns, and similar matters. The 2 week dataset processed here, amounting to 36 TB of data, ran in 3 days on a 16 node cluster. Here, we report the algorithmic approach and give some detail into the also-important technical organization behind this processing.

The data analysis pipeline is illustrated in Figures 3 and 4. We have broken up the pipeline into two flowcharts, with Figure 3 showing the calibration and quality checks performed separately on each nights' data, and Figure 4 depicting the combination of the nights and the subsequent power spectrum estimation. Both charts flow in general from top to bottom, and where applicable from left to right. In each, data products are represented by colored rectangles, with colors indicating the *type* of data (indicated in the legend), and the data dimensionality indicated with text, as well as the 'stacked' representation.[46]

Observations are recorded on site and saved to UVH5 files[47] that each include all 1536 frequency channels, four linear polarizations, $N_{\rm bl}$ baselines, and two 9.6 s time integrations. Autocorrelation visibilities are extracted from the datafiles and saved independently, to allow for fast access during the analysis. These two raw datasets are represented as colored boxes at the top of Figure 3.

Each major step of the pipeline is implemented as a parameterized Jupyter Notebook (T. Kluyver et al. 2016), with formatted tables and figures within the notebooks making a self-documenting visual record of the analysis performed. All of the notebooks produced by this analysis, as well as the final upper limits, covariances, and window functions, are publicly available on Zenodo: DOI: 10.5281/zenodo.17676032. The underlying algorithms for quality metrics, calibration, averaging, and power spectrum estimation are defined in a suite of libraries maintained by the HERA Collaboration at https://github.com/HERA-Team/.[48] In the flowcharts, each Jupyter

[44] In practice, we define two baselines $\boldsymbol{b}_{ij}$ and $\boldsymbol{b}_{kl}$ as redundant if $|b_{ij}^{p} - b_{kl}^{p}| < 2$ m $\forall\, p \in \{x, y, z\}$.

[45] Any time the noise is estimated during the pipeline, the most up-to-date calibration solutions consistent with that level of processing are applied.

[46] The axis over which the data are 'stacked' in the flowcharts corresponds to the axis over which the data is broken into separate files on disk.

[47] https://github.com/RadioAstronomySoftwareGroup/pyuvdata/blob/main/docs/references/uvh5_memo.pdf

[48] Of particular note are the repos `hera-qm`, `hera-cal` and `hera_pspec`.

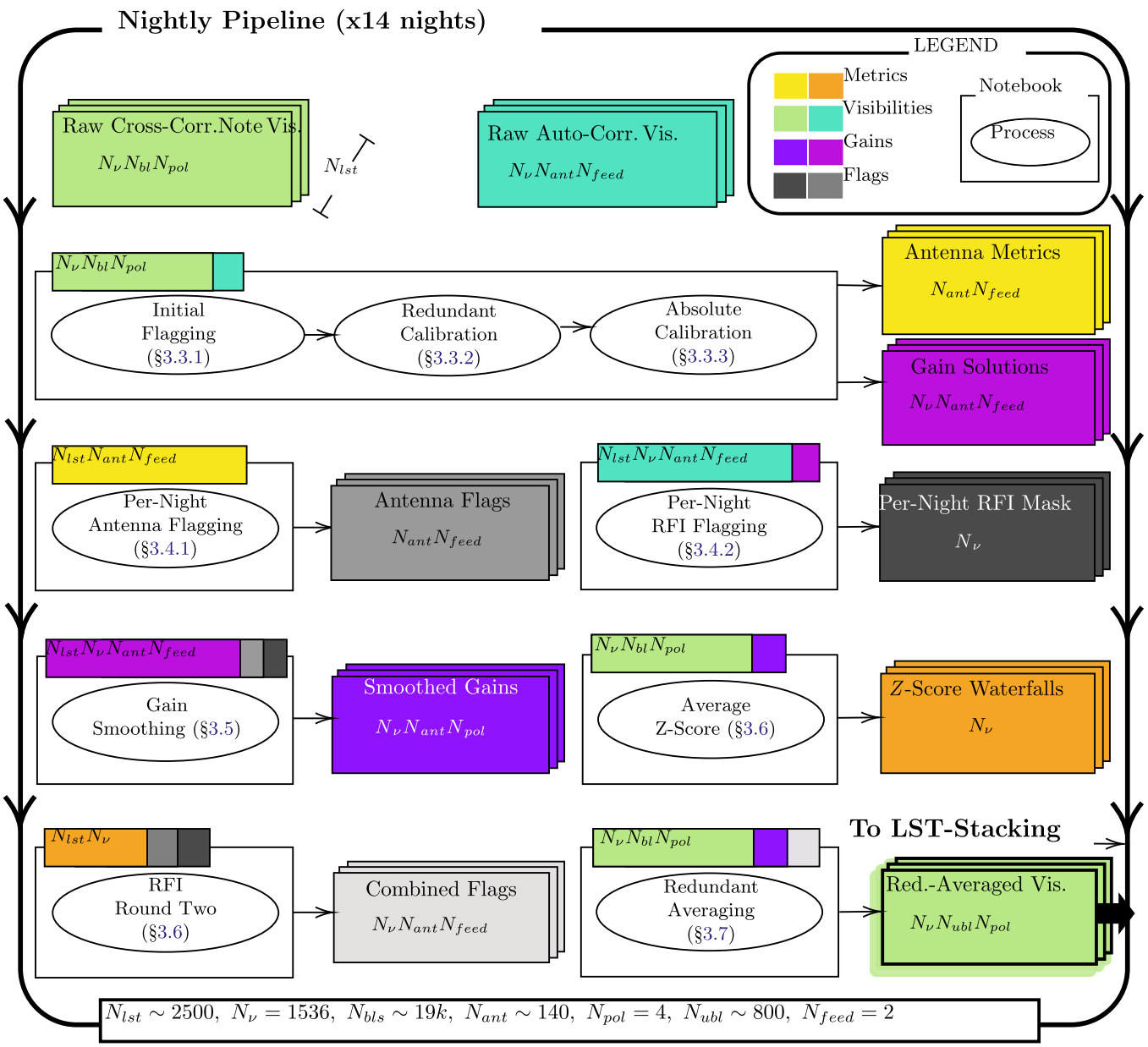

Figure 3. The HERA per-night analysis pipeline (leading to the power spectrum pipeline in Figure 4). As indicated in the title, this pipeline is run separately for each night of observation. The general flow of the analysis is from top-to-bottom and left-to-right. Large colored rectangles represent data products, with the dimensionality indicated as text in each rectangle, and the stacks representing the fact that the products are stored as multiple files over time (LST). Ovals with inset text represent analysis steps (with references to sections of this paper in which they are described), and rectangles surrounding one or more ovals represent single computational "tasks" generating Jupyter notebooks for inspection. Small colored labels inset onto the top-left of each task are the input data required for the task, with color specifying the data type (corresponding to a previous data product) and text specifying the cut of data considered simultaneously.

notebook is represented as a rectangle, within which may be several processing steps, represented by ellipses with titles and references to sections in the paper in which we describe the process. Each notebook takes input data, which are denoted by colored tabs in the top-left corner of each notebook, where the color defines which data product is being input (multiple colors indicates multiple inputs). Each notebook also produces a set of outputs that are connected by arrows.

One of the most important aspects of the analysis pipeline to keep in mind is that most of the tasks cannot see all of the data at one time, due to its size. The HERA analysis pipeline proceeds in stages by slicing data along different axes, aiming to make increasingly sensitive assessments within the bounds of the available computing and memory resources. We represent these choices in Figures 3 and 4 by defining the shape of the input data slices in the colored "input" tabs for each process, which are in general different to the size/shape of the data product on disk. Furthermore, by considering condensed data products that are either independent of frequency (e.g., Antenna Metrics) or antenna-dependent instead of baseline-dependent (e.g., Gain Solutions), we are able to simultaneously consider the full set of LSTs for a particular night (e.g., the Smoothed Gains), and then incorporate this full-night information back into the raw data (e.g., Redundant Averaging) without needing to read the full set of raw data simultaneously.

The pipeline can be split into three major conceptual sections spanning the two flowcharts: calibration, redundant

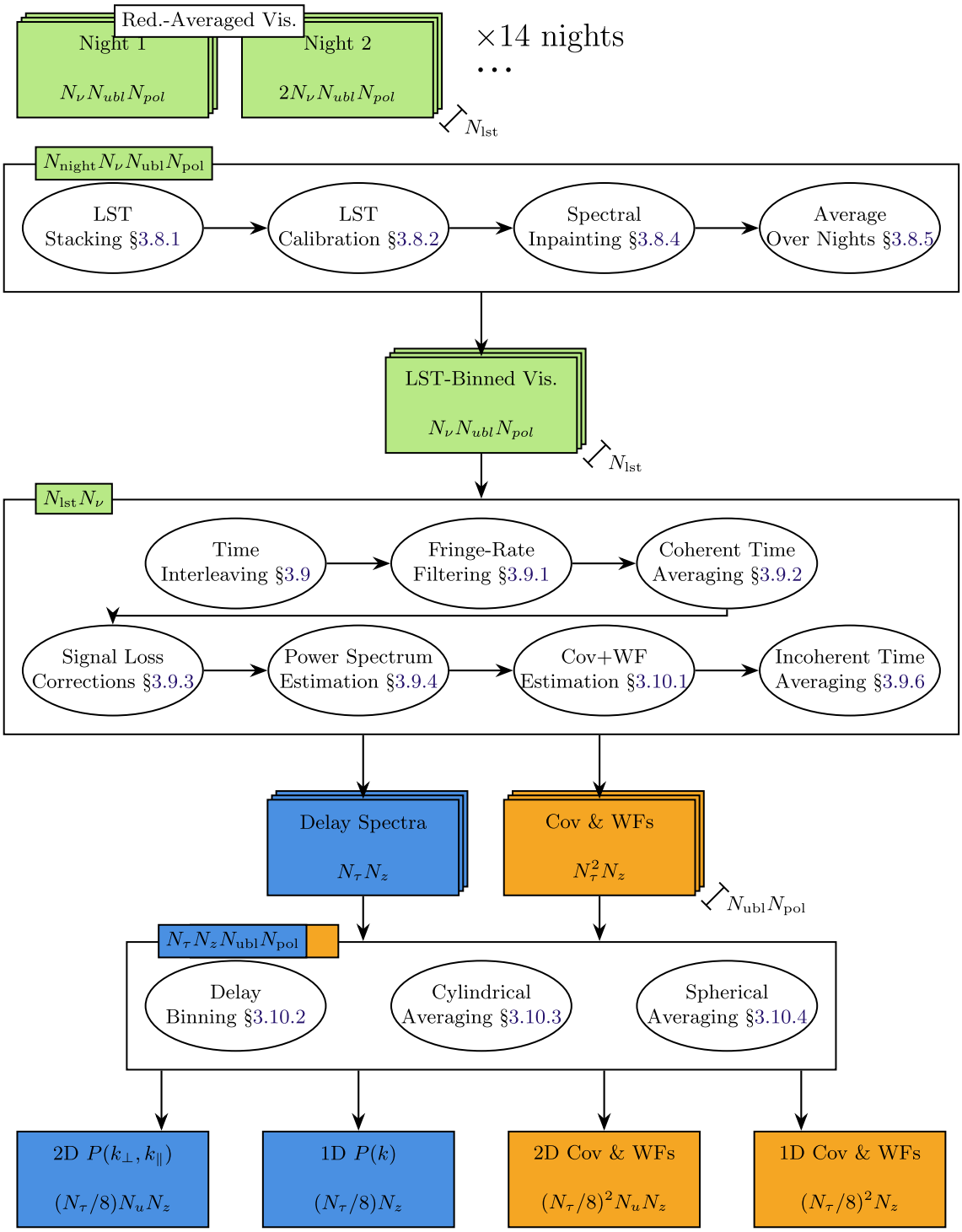

Figure 4. The HERA LST-stacking power-spectrum estimation pipeline. The visual language has the same meaning as in Figure 3, with the addition of blue data products representing power spectra, and orange data products representing power spectrum statistics (covariances and window functions). In this pipeline, the size of each stack is indicated at its bottom-right corner. Additionally, in contrast to Figure 3, this part of the pipeline is run only once, combining all nights in the first processing step.

averaging and LST averaging, and finally, power spectrum estimation. Each step includes additional calibration, filtering, flagging, and statistical tests tuned to the relevant level of averaging.

The first section—ending after "RFI Round Two" near the bottom of Figure 3—is focused on calibrating and flagging the raw data. The data itself is not modified at all in this section (note the lack of green visibility outputs up to this point); rather, we ultimately derive two new outputs: the frequency- and time-smoothed per-antenna gain solutions (purple; Sections 3.3.2, 3.3.3, and 3.5) and a combined set of flags incorporating both antenna-based (Section 3.4.1) and frequency-dependent (Sections 3.3.1, 3.4.2, 3.6) masks (light gray).

The second section—starting with Redundant Averaging and ending after LST-binning—averages over axes in which the data are assumed to be redundant: baseline groups and

nights. This results in two new intermediate visibility data products: (i) calibrated, flagged, and redundantly averaged visibilities, which are compressed by a factor $N_{\rm bl}/N_{\rm ubl} \sim 10$ (see Section 3.7), and (ii) LST-binned visibilities, in which the data from each night at the same sidereal time and baseline are averaged together, further compressing the data by a factor of $N_{\rm nights} = 14$ (see Section 3.8). The LST-binned dataset is approximately 350 GB.

The third and final section is power spectrum estimation. Here, we examine individual baselines independently rather than LSTs, allowing us to consider the full range of LST, polarization, and frequency for each baseline simultaneously. We interleave the dataset in time to produce four noise-independent subsets of data, for each of which we perform systematics mitigation through fringe-rate filtering (Section 3.9.1), average the visibilities within LST bins of 270 s (Section 3.9.2), compute signal-loss corrections due to these filters (Section 3.9.3), compute cross-correlation delay spectra (Section 3.9.4) and their error bars (Section 3.9.5) between the interleaved subsets, and incoherently average the spectra and covariances over the LST bins (Section 3.9.6). Finally, we combine all of the per-baseline delay spectra, average them within larger delay bins to reduce correlations between neighboring delays (Section 3.10.2), average over baseline orientations to produce cylindrical power spectra (Section 3.10.3), and finally, average within regular $|k|$ -bins to produce the final spherically averaged power spectrum estimates with their covariances and window functions (Section 3.10.4). In the following subsections, we describe these steps in more detail.

3.3. Calibration and Flagging

The first major step of the analysis pipeline is for each integration—spanning all baselines, channels, and polarizations—to be separately flagged and calibrated. Two types of flags are generated: per-antenna flags,[49] which are for all channels, and per-channel flags, which are for all antennas—typically for RFI. These flags and calibration solutions are later modified (i.e., smoothed or otherwise harmonized) in the context of full nights of gains and flags.

First, the data are flagged using the autocorrelations. The data that survives flagging is then calibrated; first we perform *redundant* calibration to achieve relative calibration of all antennas, and the remaining degeneracies are then fixed by comparing to a simulated sky model. We present each of these processes in more detail in the following subsections.

3.3.1. Initial, Per-integration Flagging

Interference and malfunctioning antennas can introduce outliers that do not integrate down. For this reason, each calibration or averaging step is preceded by a flagging step in which outliers are identified and masked. In the first of these steps, each time, frequency, and antenna are individually evaluated. In later steps, averaged products will be used to find smaller outliers.

Common factors causing outliers visible in individual data points are: RFI (generated by, e.g., terrestrial FM stations), nearby lightning storms, and a malfunctioning signal chain for a particular antenna or node.

For each integration,[50] flags are generated on a per-antenna and per-channel basis. For some metrics, there is an interaction between these modes; for instance, one metric for an antenna's fitness is whether the spectrum of the autocorrelation is similar to the average over all antennas. Anomalous spectral structure might indicate any number of underlying issues in the system. However, such structure might also be a manifestation of RFI, which can show up more strongly on some antennas than others, and should be flagged on a per-channel basis instead (potentially causing that antenna to pass antenna checks once the problematic channels are flagged).

It should also be noted that unlike the `auto_metrics` presented in D. Storer et al. (2022), these per-antenna cuts are mostly[51] based on absolute metrics, as opposed to their being statistical outliers relative to the rest of the array.

First, we flag entire antennas that are dead (many zero-valued visibilities), have low correlations (D. Storer et al. 2022) indicative of clock distribution issues (L. M. Berkhout et al. 2024), are cross-polarized, exhibit digital packet loss, or have anomalous autocorrelation power or slope relative to the rest of the antennas. These properties are unlikely to be contaminated by per-channel effects. We detail the precise metrics used for these cuts in Appendix A.

Our three remaining metrics for determining per-antenna flags are more likely to interact with the per-channel flags. To account for this, our strategy is to first determine an initial set of per-channel flags (we refer to this as an "RFI mask," though it may flag more than just RFI), and then to iteratively refine both this mask and the set of per-antenna flags until these sets converge. Using this strategy, we flag antennas that we deem to have broken X-engines (each X-engine correlates all antenna pairs within a block of 96 channels), excess RFI compared to other antennas, or an anomalous autocorrelation spectral shape. We detail these three metrics, and how they interact with the RFI mask through an iterative refinement, in Appendix B. Ultimately, we carry through both these per-antenna flags and the refined RFI mask to the calibration as described in the next two subsections.

3.3.2. Redundant Calibration

Redundant calibration simultaneously solves for each complex antenna gain, $g_i(\nu)$ as well as a model for each *unique* visibility $\hat{V}_{ij}$, through chi^2-minimization of the residual $V_{ij}^{\rm meas} - g_i g_j^\star \hat{V}_{ij}$. In HERA's hexagonal layout, most baselines are sampled many times, which makes gain solutions well determined. This is one of the secondary benefits of HERA's layout design (J. S. Dillon & A. R. Parsons 2016). However, the method has limitations and challenges. It assumes identical antenna elements situated on a perfect grid. It also cannot solve for a handful of degrees of freedom in calibration, requiring a sky model.

[49] Note that these are *not* per-baseline flags: we find that most unrecoverable systematics affect individual antennas (and all baselines of which they are a part) rather than baselines, and the convenience of carrying around per-antenna flags is worth the small amount of over-flagging arising from baseline-dependent systematics. Also note that these flags are applied independently for the two instrumental polarizations of each antenna.

[50] For practical reasons, we in fact read in files with *two* integrations each, and process these two integrations together. Conceptually, this makes almost no difference to the analysis (each integration is still treated independently); however, there is a small caveat that the initial per-antenna flags are generated *per-file* (i.e., for two integrations) rather than per-integration. This has very minimal impact on the results.

[51] With the exception of the comparison of the auto-spectra to the mean just mentioned.

The formalism of solving the nonlinear optimization problem of simultaneously finding gains and visibility solutions is described in A. Liu et al. (2010) and J. S. Dillon et al. (2020). The particular approach used for Phase I is described in H22a and H23. Small changes have been made for Phase II to improve efficiency.

In Phase I, redundant calibration proceeded in three steps: (i) `firstcal`, which found an approximate solution for a per-antenna phase and delay, (ii) `logcal` in which the full solutions for the gains, including the amplitude, are determined in an approximate but biased way (A. Liu et al. 2010), and (iii) `omnical` in which the χ^2 is minimized with fixed-point iteration, starting from the previous solutions (J. S. Dillon et al. 2020; H. Zheng et al. 2014). In Phase II, we completely eliminate the second step, `logcal`, and instead simply average delay-calibrated visibilities from different baselines together to get initial solutions for iterative `omnical`. This is allowed to iterate for at least 100 cycles.

However, if any new antennas are rejected in any iteration for having a χ^2 per antenna beyond our threshold of 3 (antennas consistent with noise should have χ^2 per antenna of 1; see J. S. Dillon et al. (2020) for details), we discard the worst antennas and perform another 50 iterations, iterating until all unflagged antennas are below the threshold.[52] The χ^2 of each antenna—including those flagged during the calibration process—is propagated to the following steps to be considered for exclusion from the rest of the pipeline.

3.3.3. Absolute Calibration

"Absolute calibration" refers to the process of fixing the average gain and phase gradient (or "tip-tilt") of the array to the true sky (N. S. Kern et al. 2020a), after having determined the relative gains via redundant calibration.

In H23, the absolute calibration model was based on sky-calibrated HERA data taken from three LST ranges that contained catalog sources and had fields suitable for self-calibration. Phase II datasets span a larger range of frequencies, in particular the ∼50–100 MHz low-band. With a larger span of data available, a more consistent approach was desirable, and to achieve this, we use a detailed model of the southern sky. This "Southern Sky Model" (SSM; Z. E. Martinot 2022) was propagated through the `RIMEz` visibility simulator (Z. E. Martinot 2022)[53] using the full HERA array layout and antenna beam (N. Fagnoni et al. 2021a). In particular, this sky model is anchored by the diffuse sky maps of M. Remazeilles et al. (2015) at 408 MHz and A. E. Guzmán et al. (2011) at 45 MHz, with a spectral index that varies on scales of 5°. The overall flux scale is calibrated with the EDGES measurements presented in R. A. Monsalve et al. (2021). Point sources from the GLEAM catalog (N. Hurley-Walker et al. 2016) are merged with the diffuse maps using a spherical harmonic formalism to correctly represent the power already present in the diffuse maps.

In this analysis, the flux scale is set using the autocorrelations from the SSM simulation, including an estimate of the receiver temperature spectrum, implemented as a cubic spline interpolation of the data from the laboratory measurements

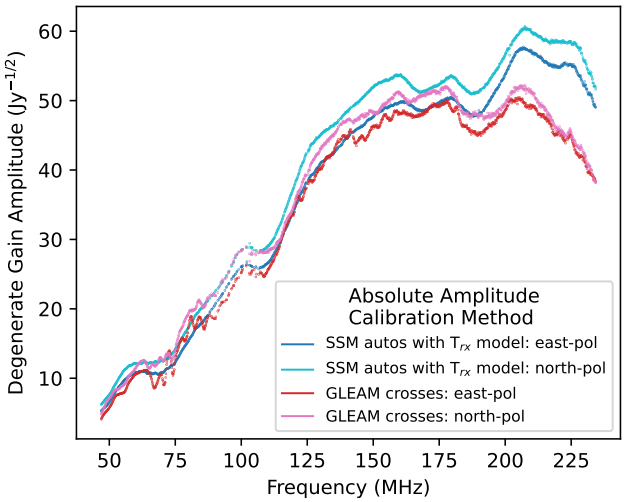

Figure 5. Comparison of the average absolute calibration vs. frequency using the SSM from this work and the autocorrelations and using a sky-based calibration from GLEAM and the cross-correlations. This comparison is performed for a single integration at JD = 2459861.4782854, which corresponds to an LST of 2.0789 hr. At this LST, the field of view is dominated by two point sources near the center, making it one of HERA's best calibrator fields (N. S. Kern et al. 2020a). Despite that, diffuse galactic emission is still quite important, and because the simulation using GLEAM lacks a diffuse model, only long baselines (60 m$<|\boldsymbol{b}|<$ 140 m) are used when performing amplitude calibration with cross-correlations. The agreement is quite good, but the SSM calibration exhibits less spurious spectral structure and is stable across LST, whereas the GLEAM calibration fails in regions with significant diffuse emission or gaps in the catalog.

performed in Section III-A of N. Fagnoni et al. (2021a). Using autocorrelations instead of cross-correlations also naturally avoids bias in the flux scale due to intrinsic thermal noise (e.g., J. E. Aguirre et al. 2022). The flux scale bias is roughly inversely proportional to the signal-to-noise ratio (SNR) of the data used. Cross-correlations have a characteristic SNR of ∼10, which produce biases of the order of 0.1, but autocorrelations achieve an instantaneous SNR of ∼1000 so that the order of the thermal noise bias becomes a negligible 10^{-3}. The agreement in the absolute flux scale between this model and that achieved by a calibration similar to N. S. Kern et al. (2020a) is shown in Figure 5.

One parameter degenerate in the redundant calibration is tip-tilt phase pointing, which must be fit with the sky model as a spatial gradient in the phase across the array aperture. In Phase I, we solved for the phase gradient using a linearized approximation, valid when the model errors are small—as was appropriate for a data-derived model. However the model used here—which covers a much wider range of frequencies and LSTs—was found to occasionally contain phase errors of more than π radians for some baselines. This necessitated a new algorithm for a more exact and robust solution of the nonlinear least-squares problem to estimate these phase-gradient degeneracies. Our new method will be explained in detail in an upcoming paper (Z. Martinot & J. S. Dillon 2025, in preparation).

While the revised calibration does still have a very small number times and frequencies that end up miscalibrated due to an insufficiently accurate sky model,[54] there are significantly fewer than with the previous absolute phase calibration algorithm. These failures are rare enough and sufficiently

[52] Additionally, outrigger antennas are excluded from redundant calibration, both for computational speed and because we have not yet verified that they can be well calibrated and brought into the analysis presented here.

[53] https://github.com/zacharymartinot/RIMEz

[54] These cases are visually obvious in plots of the phase of the gain solutions as a function of frequency and time, appearing as sharp discontinuities.

compact in time and frequency that smoothing of calibration solutions (see Section 3.5) provides sufficient mitigation.

3.4. Per-night Flagging Synthesis

In the previous step, the flags were determined considering each integration independently. Many of the issues triggering flags or poor calibration solutions, such as failed signal chains, are expected to be stable with time and so better identified when more time is considered simultaneously.

Here, we consider the antenna metrics (upon which the initial per-integration flags were based) for a full night simultaneously. In doing so, we produce updated sets of flags—both per-antenna/integration flags, and per-channel/integration flags (RFI mask). This synthesis uses additional information, such as calibration χ^2, where necessary.

3.4.1. Per-antenna Flag Synthesis

In this synthesis, a list of antennas to ignore at each integration is generated for each night of data using the autocorrelations, calibration χ^2, time-dependent antenna flags, and time- and frequency-dependent interference mask. We first re-apply all per-antenna flags previously obtained per time, as well as flag all antennas at times for which the Sun is above the horizon.

Following this, a procedure is needed to combine flags across time with protocols to handle gaps. For example, we found integrations that are not flagged, but are surrounded by integrations that are flagged. Such cases are suggestive of a low signal-to-noise situation where some integrations may fall just below the threshold for flagging, but due to their proximity to artifacts flagged at higher confidence, they are likely still affected. Consistently harmonizing this information is the aim of the "smoothed metric flagging" algorithm, described in more detail in Appendix C, which smooths antenna metrics and χ^2 from redundant calibration on a 10 minute timescale, and flags the smoothed metrics according to a pre-defined threshold. In this way, previously unflagged data can be flagged if in the vicinity of strong outliers. Furthermore, to avoid large gaps in the middle of a night (which can negatively affect fringe-rate filtering further on in the pipeline), we also flag all integrations either before or after gaps larger than $\sim$10 minutes (whichever is smaller) within a night. Antennas that are flagged for more than 50% of the integrations on a given night are flagged for the entire night. Antennas flagged in the initial, per-integration flagging are never unflagged as a result of this process.

3.4.2. Initial Per-night RFI Mask

In parallel to the antenna flagging described above, a per-night "RFI mask" must also be synthesized. This is an ($N_{\rm integrations}$, $N_{\rm freq}$)-shaped mask that applies to all antennas.

First, a list of good antennas, channels, and times is assembled using the criteria developed in Section 3.3.1, as well as requiring the Sun to be below the horizon, and omitting the FM band (87.5–108 MHz). For each antenna, we then perform a 2D high-pass filter on the autocorrelations (as a function of time and frequency) using the discrete prolate spheroidal sequence (DPSS) basis (D. Slepian 1978; A. Ewall-Wice et al. 2021). This basis set and its parameters are presented at length in A. Ewall-Wice et al. (2021), with shorter pedagogical guides appearing in P. Bull (2024) and K.-F. Chen et al. (2025) and applications to data in R. Pascua et al. (2024) and T. A. Cox et al. (2024). Here, the half-width in the frequency axis is set to 200 ns (corresponding to a smoothing scale of $\sim$5 MHz), and on the time axis is 2.2 mHz (corresponding to a smoothing scale of $\sim$ 450 s). In this filtering, we consider only the central window for the night outside of which all integrations are flagged (e.g., due to solar elevation). The weights used for determining the DPSS coefficients are the estimated thermal variance of the autocorrelations, given by the mean autocorrelation (over all candidate antennas). After filtering, the per-antenna waterfall produced is divided by the expected thermal noise level (see Equation (1)) to form a Z-score, i.e., a metric that should be close to normally distributed. Antennas are only used for RFI flagging when their rms Z-score over time, and frequency is $<$1.2, or if they are in the best-performing (i.e., most stable) quartile of all antennas.

Having obtained a set of "good" antennas, we obtain a new average autocorrelation and associated estimate of thermal noise, as well as the Z-scores as computed above, but for the *mean* autocorrelation over the set of "good" antennas. With this Z-score waterfall in hand, we flag any pixel of the waterfall with $Z > 5$, as well as neighboring pixels with $Z > 4$.

We then iteratively harmonize the flags over full channels and integrations. Here, we compute the mean over *unflagged* Z-scores over each axis, and for the axis in which the highest mean exists, we flag all elements whose mean is both higher than the highest mean in the other axis and higher than a threshold of 1.5. We repeat this process until all flagged means over both axes are below 1.5. This iterative process allows us to eliminate the most poorly behaving channels and integrations without counting, for example, an extreme outlier in an otherwise poor channel as evidence of a poor integration (and vice versa).

We then use the flag waterfall we just obtained to determine a new 2D DPSS filter, and repeat the process afresh. We do this only for two rounds.

3.5. Per-night Calibration Smoothing

All of the calibration steps discussed so far (redundant and absolute calibration) are performed independently per integration and channel. This renders the calibration solutions susceptible to both temporal and spectral fluctuations and can result in gain errors, $\hat{g}_i/g_i$, that are spectrally and temporally structured. While all forms of error in the estimated gains are undesirable, those that are spectrally structured are particularly egregious, causing foreground power to leak from the low-delay wedge into the high-delay 21 cm window. Since apparent spectral structure in gain solutions is likely due to the impact of nonredundancy on redundant calibration (N. Orosz et al. 2019) or absolute calibration (R. Byrne et al. 2019), we take a first-do-no-harm approach. We thus impose a strong smoothness prior on our calibration solution, relying on the stability and spectral smoothness of the instrument's response. Thus, any rapid fluctuations in the estimated gains are considered to be spurious. We thus smooth the gains over both time and frequency. This process has remained essentially the same as previous HERA limits, but can be summarized as follows.

We first choose a reference antenna (one for each night)—the antenna that is flagged for the fewest integrations across the night—and rephase all estimated gains such that the

reference antenna has a phase of zero:

$$g_j(t, \nu) \longrightarrow g_j(t, \nu)\exp\left[-i\phi_{\rm ref}(t, \nu)\right]. \quad (2)$$

We then smooth each antenna's gain simultaneously over time and frequency using a DPSS model with a half-width in frequency of 100 ns (corresponding to a smoothing scale of $\sim$10 MHz), and in time of 1.65 μHz (corresponding to a smoothing scale of $\sim 6 \times 10^5$ s). Though this is longer than a day, the way that DPSS filters are constructed means that the gain solution admits a few temporal modes.

Importantly, it sometimes occurs that during the course of a night, a particular antenna's gain-phase is flipped (rotated by 180°) with respect to its phase on the first (unflagged) integration of the night.[55] In this case, we detect the integrations in which this occurs, and flip all subsequent integrations (until the phase reverts back, if indeed this occurs) before fitting the DPSS model, and flip them back again after smoothing. We also flag the integrations in which any flips occur, since the flip might happen at some point during the integration.

3.6. Deeper RFI Flagging

We find that the detection of RFI in autocorrelations averaged over antennas can miss low-level RFI that is noticeable in cross-correlations after further averaging to increase signal-to-noise. Hence, we perform a check for this low-level RFI as follows.

We first consider each integration separately, computing a single array-averaged Z-score spectrum for each. We then combine the Z-scores for all integrations on a given night to perform harmonized outlier-detection.

Considering a single integration (all baselines, channels, and polarizations), we apply the smoothed calibration solutions and flags, and then redundantly average, as discussed in the next subsection (Section 3.7), yielding $N_{\rm ubl}N_{\rm pol}$ averaged visibility spectra. We then estimate the thermal noise on each unique baseline group using Equation 1 with $N_{\rm samples} = N_{\rm bl}^{\mathcal{G}}$, by which we divide each redundantly averaged visibility yielding an estimated SNR. We then down-select the baseline groups that are carried through to compute the final metric; we include only cross-correlation baselines for which (i) the median $N_{\rm bl}^{\mathcal{G}}$ (over t and ν) is at least 15% of the maximum $N_{\rm bl}^{\mathcal{G}}$ for any baseline group (in practice, the autocorrelations), and (ii) the baseline delay, $\boldsymbol{b}_{ij}/c$, is smaller than the high-pass delay-filter threshold, 750 ns. This yields $N_{\rm filt} < N_{\rm ubl}$ baseline groups satisfying the criteria.

To the SNR of each of these baselines, we fit 1D DPSS models over the spectral axis—one for each unique baseline and integration, fit independently to frequency bands below and above FM—and subtract the model from the SNR. As mentioned, this DPSS model has a half width of 750 ns—well outside the expected FG-dominated wedge, and including dominant systematics such as MC. If all of the foreground signal is concealed at delays below 750 ns, the remaining delay-filtered SNRs are expected to be foreground- and systematic-free, and should thus be normally distributed, i.e., they can be considered Z-scores.

[55] We attribute these flips to the rare accidental triggering of the Walsh switching of HERA feeds—a capacity built into the hardware but not implemented for these observations (L. M. Berkhout et al. 2024).

There is a small caveat here. There is a well-known feature of model fitting in which data toward the edge of the range (or data close to large gaps) are over-fit due to their not being constrained on one side. This leads to a systematic underestimate of $|Z|$ close to the edges and large flagging gaps (such as the FM band). This can be analytically corrected by dividing the spectrum of $Z_t(\nu_i)$ at each integration t by L, where:

$$L_i^2 = \frac{\pi}{4}(1 - h_{ii}). \quad (3)$$

Here, h_{ii} is called the "leverage," and is simply ith diagonal element of the ortho-projection (or "hat") matrix:

$$H = Z_t(X^T N X)^{-1} X^T N, \quad (4)$$

where N is the diagonal matrix whose diagonal elements are given by $N_{\rm bl}^g(t)$, and X is the DPSS design matrix.

Following the correction of the set of Z by this leverage-correction factor, we compute a modified mean-Z via the following:

$$\bar{Z} = (\langle |Z| \rangle - 1)\sqrt{\frac{\pi N_{\rm ubl,filt}}{4 - \pi}}, \quad (5)$$

where the average over the absolute Z-scores is performed over the $N_{\rm filt}$ unique baselines satisfying our selection criteria defined above. This new, averaged quantity has an expected mean of zero and variance of unity, though it is not normally distributed.

In a very similar fashion to the flagging performed on the antenna-averaged Z-score waterfall in Section 3.4.2, we combine the per-integration Z-scores over a full night, and perform similar checks but with reduced thresholds (see Appendix D for details).

We show an example of this Z-score waterfall in the top panel of Figure 6, which highlights a small $\sim$40 MHz by 2 hr window. Pre-existing flags (see Section 3.4.2) are shown in white, while the Z-score computed in the manner described above is shown on the color-scale from blue to red, clipped at 5σ. Several key features are immediately apparent: there are a number of small "blobs" of high Z-score, generally close to channels that were pre-flagged. There are also channels that have intermittent outliers across time. The new flags determined by this procedure are depicted in the bottom panel of Figure 6 as the pixels overlaid in orange. Note that very often, the resulting flags apply to entire channels for all times, though there are also instances of flags that affect a small region of time and frequency.

Finally, as an even deeper probe of RFI in particular channels, we take an average of $\bar{Z}$ over integrations, leaving a single spectrum of Z-scores, which we further normalize by dividing by the square root of the number of unflagged integrations, and subtracting a DPSS model with half-width of 250 ns, corrected for the leverage as above. We denote this quantity by Q. We then flag any channels where $Q > \max(4, \max(Q)/1.5)$, and repeat the averaging and DPSS-fitting until no new flags are found. In this way, large outliers are less likely to cause additional channels to be flagged due to their influence on the filter.

3.7. Redundant Averaging

A major distinction between this work and previous HERA limits is that here we averaged the visibilities of all redundant

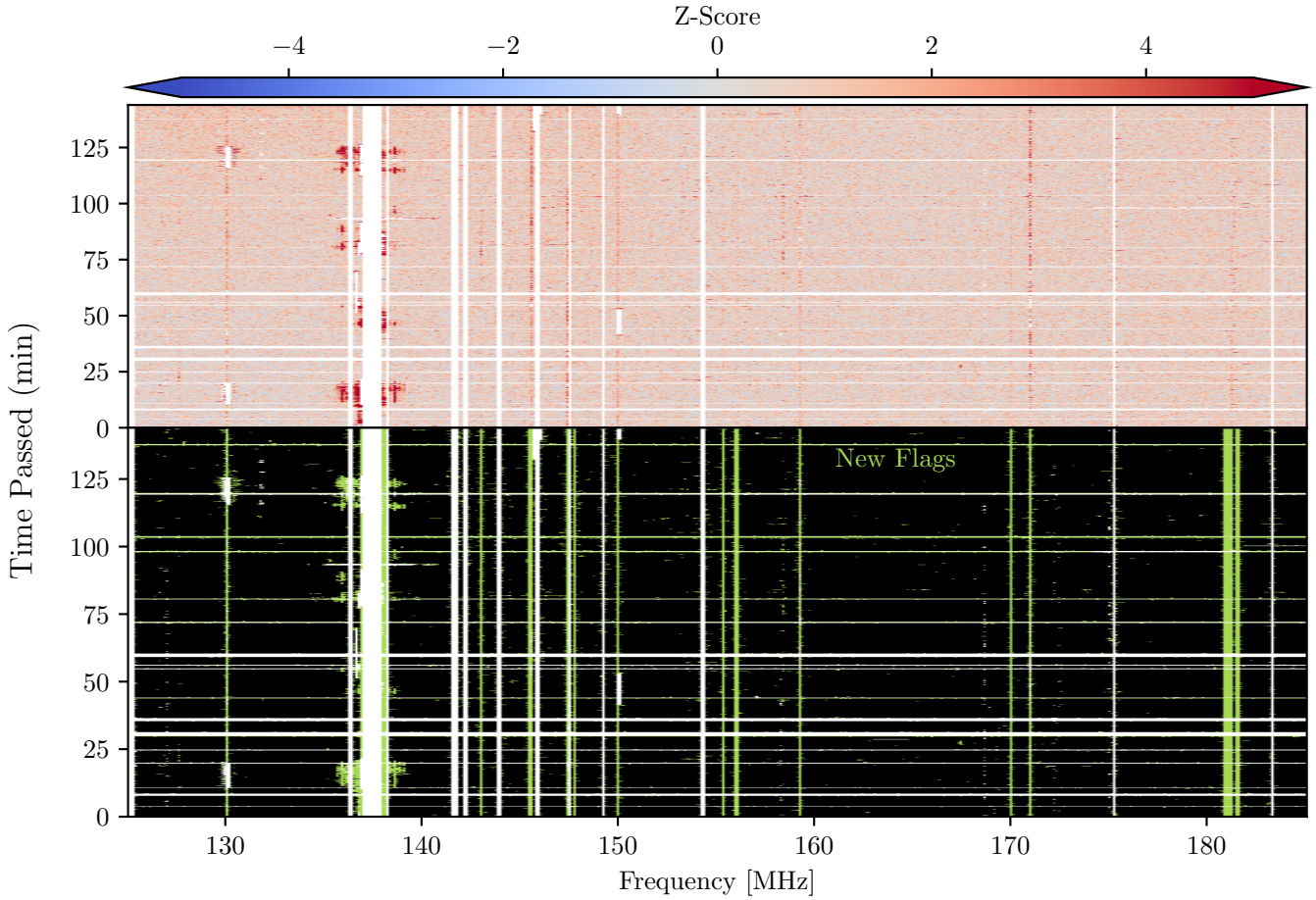

Figure 6. An illustration of per-night RFI flagging, as described in Sections 3.4.2 and 3.6. In the top panel, we show the mean Z-score computed over "good" baselines for night 2459861. Notice that there are clumps of unflagged high-Z, especially surrounding regions flagged for other reasons, e.g., at ∼137 MHz. The synthesis RFI algorithm iteratively flags individual pixels and either the worst offending entire channels (vertically) or entire integrations (horizontally). The culmination of these flags is presented in the bottom panel, where new flags from this algorithm are shown in green.

baselines prior to forming power spectra. This is a significant analysis choice, making the downstream power spectrum estimation significantly more simplified and efficient. It also means that the final power spectra include contributions from all baseline pairs within each redundant group, *including the auto-pairs*, where a baseline is multiplied by itself. This is expected to reduce the occurrence of negative-valued systematics (e.g., M. Kolopanis et al. 2023; M. F. Morales et al. 2023). However, it also introduces an increased risk of signal loss, as imperfections in redundancy between the baselines in a particular group can cause decoherence when averaging their complex visibilities. We investigate the level of this decoherence in Section 4.3.2, finding that it is of the order a few percent (see Table 4).

This step is the first in which a new *visibility* data product is created. In previous steps, the raw data is read and used to create calibration solutions, and those per-antenna calibration solutions are manipulated in various ways. In this step, the data are read, the calibration solutions are applied, and a new redundantly averaged dataset is produced. This is relatively efficient, as the output product is $N_{\rm bl}/N_{\rm ubl} \sim 10$ times smaller than the raw data.

The redundant baseline averaging process itself is quite simple. For a particular baseline group $\mathcal{G}$, we calibrate each visibility at each time and frequency independently,[56] and perform a simple masked average:

$$V_{\mathcal{G}}^{pq} = n_{\mathcal{G}}^{-1} \sum_{(a,b)\in\mathcal{G}} g_a^{\,p} g_b^{q*} \xi_a \xi_b V_{ab}^{pq}, \tag{6}$$

where $\xi \in \{0, 1\}$ are per-antenna, per-channel, per-time flags, and we propagate the number of accumulated samples as

$$n_{\mathcal{G}} = \sum_{(a,b)\in\mathcal{G}} \xi_a \xi_b \leqslant |\mathcal{G}|. \tag{7}$$

Importantly, note that $n_{\mathcal{G}}$, which is a per-baseline (time, frequency)-waterfall, is by construction uniform over the frequency axis for any particular baseline group, up to flagging. This is because the only per-channel flags we set are in the nightly synthesized RFI mask, which is a single mask *for all antennas* on a particular night. Thus, while integrations may be flagged for some antennas and not others, channels are always either flagged or unflagged for all antennas. Thus, for a particular integration t, n_b has a constant value for all channels, except for flagged channels in which it is zero. This is an important point to which we will return when discussing LST-stacking and inpainting in Section 3.8.

We note that, beyond the quality metrics and flagging already discussed, no further quality checks are performed at

[56] We omit dependence on t and ν in the expressions for simplicity.

this stage. We aim to include such checks, which are uniquely possible, when the calibrated visibilities from supposedly redundant baselines are considered together, in future analyses.

3.8. LST-stacking and Averaging

Thus far, we have only considered data from a single night, and have not jointly considered data from multiple nights. At the same LST each night, we expect to observe the same sky, so that visibilities within the same unique baseline group are random draws from the same distribution. This affords a few opportunities. First, it allows us to jointly compare the calibrated visibilities and adjust the gain solution degeneracies to optimally align them over nights. Second, it allows us to jointly infer the visibility values that are flagged, with more information at hand than we can use on a single night. Third, it allows for a deeper level of outlier identification as we compare nights with each other, and finally, it allows us to average over the nights to increase SNR.

3.8.1. LST Stacking

We refer to the process of identifying and loading the data within an LST bin as "LST stacking." This is primarily a bookkeeping problem, which we tackle in the following way.

We first identify all unique baseline groups present on any night in the dataset. We then generate a grid of LSTs that are evenly spaced between zero and 24 hr, with a spacing of as close to the integration time of the observations as possible while dividing the total 24 hr evenly (9.6 s). The visibilities for all nights, baselines, and polarizations within an LST bin are gathered in parallel over LST bins. Finally, the visibilities are rephased so that phase center is at the R.A. and decl. corresponding to zenith at the center of the LST bin:

$$V_{ij}^{pq} \rightarrow V_{ij}^{pq} \exp(-2\pi i \nu \tau'), \tag{8}$$

where

$$\tau' = (\boldsymbol{R}\hat{z} - \hat{z}) \cdot \boldsymbol{b}_{ij}/c, \tag{9}$$

and $\boldsymbol{R}$ is the rotation matrix that rotates a unit vector toward zenith, $\hat{z}$, at the LST of the observation to a unit vector toward zenith at the central LST of the bin.

3.8.2. LST Calibration

Having stacked the data within an LST bin, we implement a final calibration refinement, `lstcal`, to address systematic variance observed across nights. Initial analysis of the dataset (J. S. Dillon & S. Murray 2023) revealed that for a particular baseline at a particular LST and channel, the variance measured over nights exceeded the expected variance (computed using Equation (1) where the autocorrelations are averaged over both antennas and nights). We hypothesize that this "excess variance" is at least partially a result of day-to-day variations in the per-antenna calibration solutions that manifests as a coherent systematic error in array-wide calibration relative to the average across nights.

The `lstcal` algorithm addresses this issue by comparing each night's visibilities to the average across all nights within a given LST bin,

$$V_{\mathcal{G}}^{\mathrm{avg}} = \left(\sum_{k}^{N_{\mathrm{nights}}} V_{\mathcal{G}}^{k} N_{\mathrm{bl}}^{\mathcal{G},k} \right) \left(\sum_{k}^{N_{\mathrm{nights}}} N_{\mathrm{bl}}^{\mathcal{G},k} \right)^{-1}, \tag{10}$$

computing a frequency- and polarization-dependent gain that brings each night's visibilities into better agreement with the average. We model the per-night gain correction, $G_k(\nu)$, as a per-array quantity defined by the absolute calibration degrees of freedom. For each night k and polarization, this correction consists of a per-frequency amplitude, $A_k(\nu)$, and two phase gradients ($\Phi_{x,k}(\nu)$, $\Phi_{y,k}(\nu)$). This correction is applied to each redundant group's visibility based on that group's representative baseline vector (b_x, b_y)

$$G_k(\nu) = A_k(\nu) \exp\left(i[\Phi_{x,k}(\nu) b_x + \Phi_{y,k}(\nu) b_y]\right). \tag{11}$$

We then solve for the parameters $A_k(\nu)$, $\Phi_{x,k}(\nu)$, and $\Phi_{y,k}(\nu)$ for each night k by minimizing the difference between $V_{\mathcal{G}}^{k}(k)$ and $G_k(\nu) V_{\mathcal{G}}^{\mathrm{avg}}(k)$ across all baseline groups $\mathcal{G}$.

Given that each night's visibility data have already been redundantly averaged prior to LST-stacking, we restrict `lstcal` to these absolute calibration degrees of freedom, and do not attempt to correct the per-antenna gains. In order to prevent `lstcal` from introducing spurious spectral structure into the calibrated visibilities, we smooth the gain solutions over frequency using a DPSS model with a smoothing scale of 10 MHz before applying them to the data.

Figure 7 shows an example of the effects of LST-calibration for two particular baselines at R.A. = 2.908 hr. Notice how for both baselines, the spread of the data after LST-calibration (right panels) is reduced compared to the data that has not been LST-calibrated (left panels). To evaluate the stability of the `lstcal` solutions, we also examined the per-night gain solutions across all nights used in this analysis. The corrections did not show coherent temporal trends, and instead appear consistent with random fluctuations expected from nightly calibration variance.

3.8.3. Choice of Spectral Windows

Ultimately, our power spectra will be estimated within discrete spectral windows (or bands). While the spectral windows are primarily intended for power spectrum estimation, it is beneficial to choose them at this stage of processing, to enable more granular flagging after inpainting (see next subsection).

We choose eight spectral windows (in contrast to the single band in H22a and the two bands in H23), ranging from 50–231.1 MHz. Details of the bands are presented in Table 3, and their broad characteristics are illustrated in Figure 8. The spectral windows were selected using a few criteria; each was required to be <15 MHz in width to alleviate the lightcone effect (K. K. Datta et al. 2012, 2014; R. Ghara et al. 2015; B. Greig & A. Mesinger 2018; M. Blamart & A. Liu 2025), and an attempt was made to place wide gaps (more than a few channels) with consistently high flagging fractions between spectral windows. These choices were made since strongly flagged channels toward the center of power-spectrum bands have previously been shown to leak foregrounds to high delays (J. E. Aguirre et al. 2022; K.-F. Chen et al. 2025), even with spectral inpainting applied. Further, bands were picked so as to concentrate channels with relatively high levels of flagging into as few bands as possible.

Note that in Figure 8, two metrics of flagging are included. The gray crosses indicate the RFI occupancy of each channel, i.e., the fraction of antennas, times, and feeds that are flagged after the "RFI Round Two" processing step (Section 3.6). The

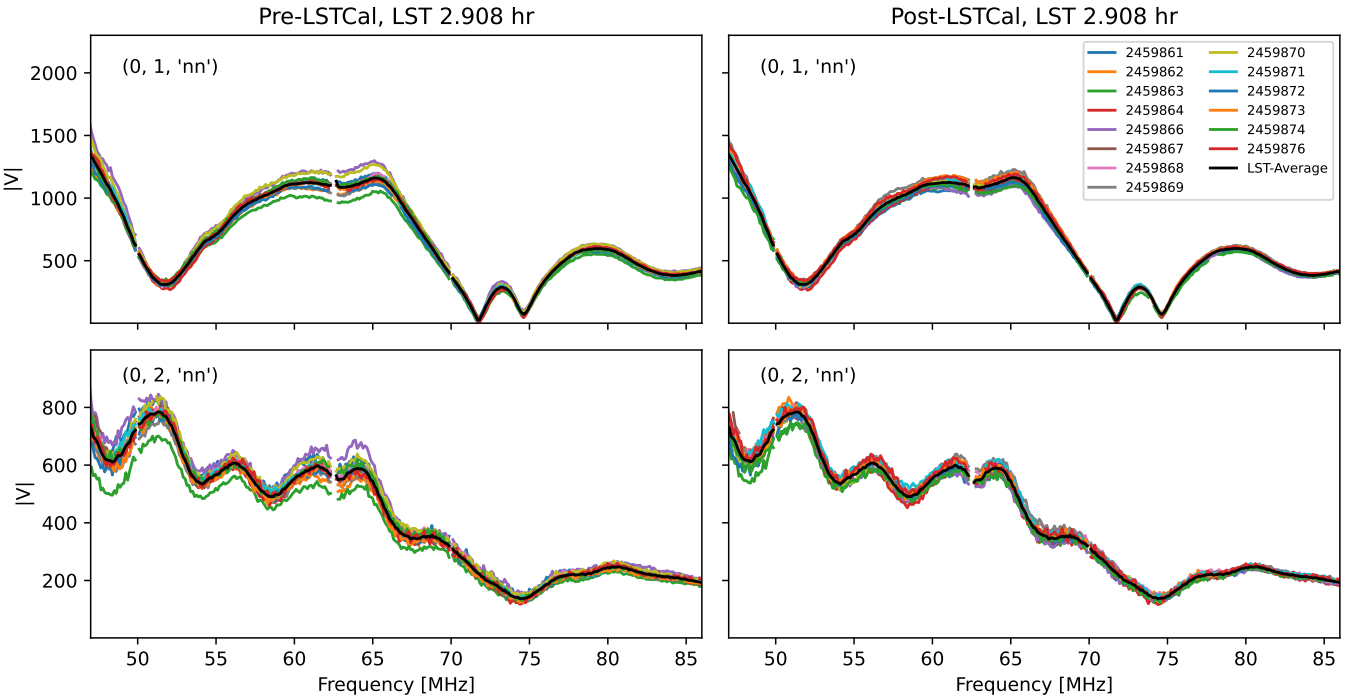

Figure 7. An example of the effects of LST-calibration, described in Section 3.8.2. The left panels shows the magnitude of the visibilities measured by the 14.6m east–west baseline group (top) and 29.2m east–west group (bottom), on the 14 nights of observation (colored), with their average shown in black. These panels are *before* LST-calibration is applied, while the right panels are the same data after LST-calibration is applied. Notice how the spread between nights is significantly reduced, while the overall structure remains the same.

Table 3
Spectral Window Definitions for the Power Spectra Presented in This Paper

Channels	$N_{\rm chans}$	Freq. Range (MHz)	$\Delta\nu$ (MHz)	z Range	Center z	Δz	Used?
27–126	99	50.2–62.2	12.1	21.82–27.3	24.6	5.5	✓
135–218	83	63.3–73.5	10.1	18.33–21.4	19.9	3.1	✓
227–316	89	74.6–85.4	10.9	15.63–18.0	16.8	2.4	✓
501–567	66	108.0–116.1	8.1	11.24–12.1	11.7	0.9	✗
577–635	58	117.3–124.4	7.1	10.42–11.1	10.8	0.7	✓
643–732	89	125.4–136.2	10.9	9.43–10.3	9.9	0.9	✓
749–830	81	138.3–148.2	9.9	8.59–9.3	8.9	0.7	✗
846–920	74	150.1–159.2	9.0	7.92–8.5	8.2	0.5	✗
921–1008	87	159.3–169.9	10.6	7.36–7.9	7.6	0.6	✓
1024–1100	76	171.9–181.1	9.3	6.84–7.3	7.1	0.4	✓
1102–1225	123	181.4–196.4	15.0	6.23–6.8	6.5	0.6	✗
1242–1323	81	198.5–208.4	9.9	5.82–6.2	6.0	0.3	✗
1355–1423	68	212.3–220.6	8.3	5.44–5.7	5.6	0.3	✓
1454–1509	55	224.3–231.1	6.7	5.15–5.3	5.2	0.2	✗

Note. Channel ranges are noninclusive on the upper end. While we define 14 bands, we only report upper limits in eight of the bands, due to high levels of flagging in the other six (see Figure 8).

spectral windows were chosen based on this statistic. Conversely, the black and red dots indicate the total number of samples (baselines, times, and pols) that go into the power spectrum estimates. These are noticeably disjoint at the edges of the spectral windows, which is an artifact of flagging choices we make after inpainting—with knowledge of the bands—described in Section 3.8.4. Thus, a low overall number of samples within a band (e.g., the band above 131) does *not* necessarily indicate a high flagging fraction due to, e.g., RFI, but instead that there were a higher number of large flagging gaps in that window.

After initially specifying 14 spectral windows, we finally chose the eight most well-behaved bands in which we report upper limits. The bands that are not ultimately used are marked in Table 3 and shown as gray in Figure 8. Of the six abandoned spectral windows, four of them exhibit very high overall levels of flagging, and the other two (just below bands 165 and 216) have a large variability in flagging fraction within the band. We found that these bands exhibit weak but noticeable artifacts in histograms of high-delay signal-to-noise for power spectra with low levels of incoherent averaging, most likely arising from inpainting imperfections.

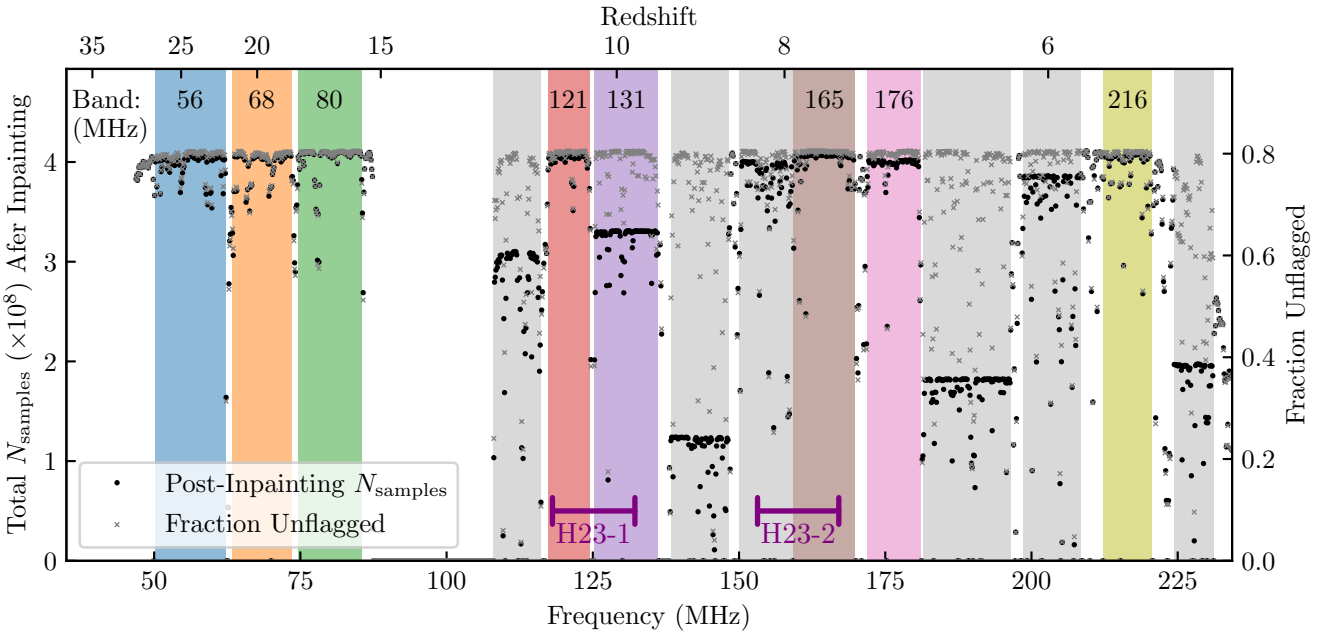

Figure 8. Spectral windows defined and used for upper limits in this paper. Gray crosses indicate the (inverse) RFI occupancy in each channel, i.e., the result of integrating the 2D RFI masks over all antennas, feeds, and times. Black and red dots indicate the total number of unflagged samples (baseline-time-pols) for each frequency channel used in power spectra, i.e., also including flags that omit entire spectral windows if there are intrinsic flag gaps of sufficient width (see Section 3.8.4). Colored regions indicate the chosen spectral windows (bands) used for reporting power spectra in this paper, labeled by the nearest integer of the central channel in each band in MHz. Gray bands are bands that were defined for analysis but not ultimately used for reporting power spectra due to their high flagging fraction. Red dots indicate channels not included in any band. The two spectral windows used in H23 are indicated in purple.

3.8.4. Inpainting

While flagging data affected by systematics such as RFI is essential, the gaps left in the data by these flags present their own challenges. These challenges primarily stem from the fact that if there is systematic variance between the visibilities on $N_{\rm nights}$ different nights (e.g., one night has consistently larger amplitude visibilities for a particular baseline compared to the other nights, while being internally spectrally smooth), then when averaging them together, if the averaging weights are nonuniform across frequency for any particular night (due to flagging), then the resulting average incurs spectral structure.

This effect is well known and studied, with a common solution being to "inpaint" a best guess of the missing data in order to preserve spectral smoothness. For example, H22a used an iterative delay-based convolution filter to inpaint missing data after averaging the raw data over nights, a technique that was numerically validated in J. E. Aguirre et al. (2022). Other basis sets for inpainting have been considered, including DPSS, least-squares spectral analysis, Gaussian process regression, and convolutional neural networks with varying degrees of success in different regimes (see M. Pagano et al. 2023 for a comparative analysis).

One problem with the previous applications of inpainting to HERA data is that it was applied *after* averaging of the data over nights; that is, it only filled in channels for which no data was available on *any* night. While this mitigates the problems encountered when Fourier transforming the data over the frequency axis to produce power spectra, it neglects the problem of induced spectral structure inherent in the averaging of nonstationary data with spectrally structured weights. This problem was explored at length in K.-F. Chen et al. (2025).

A simple way to avoid this is to inpaint the calibrated visibilities *before* averaging over nights. In this work, we adopted the DPSS basis functions to interpolate our data. We determine DPSS basis coefficients per-night, per-antenna-pol and per-LST, such that each is interpolating a 1D function of frequency. An important consideration is the half-width of the basis functions, i.e., the delay-scale out to which to fit spectral structure. The aim is to include all relevant foreground and systematic effects, but not to exceed these scales by much, as we do not wish to inpaint the scales at which we hope to measure the cosmic signal. With this in mind, we choose to use a half-width of

$$\tau_{\rm hw} = \max(500\ {\rm ns},\ |\boldsymbol{b}|/c), \tag{12}$$

which covers all sky-based sources out to the horizon, with a minimum buffer of 500 ns that covers the bulk of known systematic effects, such as MC. The inpainted data takes the value of the true measured data when the data is unflagged, and the value of the smooth inpaint solution when it is flagged.

Unfortunately, inpainting out to such high delays means that the behavior of the inpaint solutions within moderate-to-large flag gaps can be poorly constrained. Naively, we are solving for scales down to $\tau_{\rm hw}^{-1} \approx 2$ MHz, which means that gaps in the spectrum around this size (or larger) can cause poor behavior of the solutions within the gaps. While the solutions are constrained to only contain power out to the horizon delay, the final inpainted data product may have sharp transitions between flagged and unflagged regions when the inpaint model is not well constrained, resulting in power spilling out to much higher delay. We thus perform some checks to ensure that poorly behaved inpainted data is flagged before averaging over nights.

These checks center around identifying regions of the spectrum that are highly flagged over a wide enough region such that the solutions cannot be trusted toward the center of the region, far away from unflagged data. In practice, we

define a maximum flagged gap size (in units of channels) as

$$s_{\max} = \lfloor f_{\rm gap}\, \tau_{\rm nw}/\Delta_n u \rfloor, \tag{13}$$

where $f_{\rm gap} = 1$ is a tunable scaling factor. We then convolve the binary flags $\xi_\nu \in \{0, 1\}$ with a triangular filter of size $2s_{\max} - 1$. The resulting convolved array, $f_{\rm cnv}$, represents a 'flag density' of the surrounding channels, between zero and unity. We then create a new flag array by thresholding the flag densities at $f_{\rm crit} = 0.4$:

$$\xi_{\rm inp} = \begin{cases} 0 & \xi_\nu = 0 \text{ or } f_{\rm cnv} > 0.4 \\ 1 & \text{otherwise.} \end{cases} \tag{14}$$

Any region with more than $s_{\max}$ contiguous flags in $\xi_{\rm inp}$ is then identified as needing to be flagged.

Since we require spectrally uniform flags before averaging (which is the entire point of inpainting in the first place), flagging here must be for an entire spectral window per-integration. That is, if any such large contiguous gaps exist, any spectral windows overlapping with the gaps must be discarded for that integration. This is the motivation for choosing the spectral windows prior to inpainting—it allows us to selectively flag per spectral window rather than discarding the entire spectrum.

We flag any window that overlaps with a contiguously flagged region of sufficient size, as described above. However, due to the fact that we use a frequency taper when performing power spectrum estimation (see Section 3.9), we allow flag gaps to overlap with the outer two channels of each band without consequence.[57]

Overall, these choices of spectral windows and our procedure for flagging large gaps results in ∼10%–12% more of the data being flagged, but our tests indicate that these choices are conservative, i.e., they lead to no poorly constrained solutions inside the flagged gaps.

3.8.5. Averaging over Nights

After inpainting the flagged data and determining the post-inpainting flags, we are ready to average over nights. This process is reasonably simple. For each unique baseline group, channel, and LST bin, we take the weighted average,

$$\bar{V}^{pq}_{\mathcal{G}} = M^{-1} \sum_k^{N_{\rm nights}} \xi^{\mathcal{G}}_{{\rm inp},k} N^{\mathcal{G}}_{\rm bl} V^{pq}_{\mathcal{G},k}, \tag{15}$$

with

$$M = \sum_k^{N_{\rm nights}} \xi_{{\rm inp},k} N^{\mathcal{G}}_{{\rm bl},k}. \tag{16}$$

Recall that $N^{\mathcal{G}}_{\rm bl}$ is binary: its value is either some positive constant or zero. However, wherever it is zero, we have inputted data via inpainting, and we treat it as being uniform in frequency, taking its nonzero value. Furthermore, $\xi_{\rm inp}$ is either all zero or all one within a particular spectral window. This means that the combined weighting function, $\xi_{\rm inp} N^{\mathcal{G}}_{\rm bl}$, is uniform within each spectral window.

However, in computing the effective number of samples that is propagated to power spectrum and covariance estimation, we use

$$N = \sum_k^{N_{\rm nights}} \xi_k \xi_{\rm inp} N^{\mathcal{G}}_{\rm bl}, \tag{17}$$

where here ξ_k are the pre-inpainted flags (i.e., the flags resulting from quality metrics, not the flags resulting from identifying large gaps after inpainting). Thus, N is not in general spectrally uniform within bands.

3.9. Per-baseline Systematics Mitigation and Power Spectrum Estimation

After averaging the data over nights within narrow LST bins, we begin the process of mitigation of instrumental systematics, and eventually power-spectrum estimation, on a per-baseline basis. In practice, we rearrange the data so that we read the visibilities across all LSTs for a particular baseline, independently identifying systematics for each baseline in parallel.

Before any further analysis, we first remove LSTs that have fewer than 20% integrated samples compared to the maximum samples over all LSTs (for a particular baseline group). This typically removes the first and last few LSTs in the dataset, which may disproportionately affect any time-based filters used for mitigating systematics.

Our power spectrum estimation method (which we will describe below) does not allow for nonuniform weights over frequency channels (but does allow for nonuniform weights over LSTs). To avoid data weights with frequency discontinuities, we use inverse noise variance weights that require the number of samples to be constant within each spectral window. To do this, we simply average the number of samples within each window (per-baseline and LST). This means that channels with fewer samples (due to flagging and inpainting) bring down the average across the window for that particular LST. This does not affect power spectrum estimation, in the sense that our algorithm already ignores relative weightings between channels. However, it does impact the estimate of the error bars in a small way. We find that this effect is small, with the power spectrum estimates at high delay being consistent with the predicted thermal noise computed with this approximation (see Figure 16), and we leave the handling of nonuniform spectral weights to future work.

Following this, to avoid a noise-power bias in the power spectrum estimate, which arises when cross-multiplying identical (or correlated) visibilities, we split the data into four "interleaved" datasets over the LST axis. That is, we construct four nonoverlapping datasets from the data at hand, V_t (where t indexes the LST bin):

$$\mathcal{D}_i = \{V_i, V_{i+4}, V_{i+8}, ...\}, \quad i \in \{0, 1, 2, 3\}. \tag{18}$$

Each "interleave" spans the full LST range, at a cadence of $4\Delta t$, where $\Delta t \approx 9.6$ s is the LST-bin width. The following procedures for mitigation of systematics via fringe-rate filtering are applied independently to each interleave, in order to prevent introduction of correlated noise between the sets.

In previous HERA upper limits, we have avoided this noise bias by cross-correlating different redundant baselines (i.e., cross-multiplying the visibilities V_{ab} and V_{cd}, where both ab and cd baselines are in the same redundant group) and ignoring "auto-baseline" pairs (i.e., $V_{ab} V^*_{ab}$). While this effectively

[57] This helps some bands that are adjoined to regions of continuously high flagging fraction to not be flagged every time their first or last channel is flagged.

avoids the noise bias, it allows for the introduction of *negative* systematics when different nominally redundant baselines have systematics whose phases are different, as first observed by M. Kolopanis et al. (2023) and discussed in M. F. Morales et al. (2023). Since the probability of phase discontinuities between adjacent times is far smaller than between nominally redundant baselines at the *same* time, this risk is significantly reduced by cross-multiplying adjacent times, and using *all* baseline pairs within a redundant group (including the auto-baseline pairs).

3.9.1. Mitigating Instrumental Coupling Systematics via Time Domain Filtering

Two of the major systematics present in HERA's Phase I data were cable reflections and over-the-air coupling (N. S. Kern et al. 2019, 2020a; J. E. Aguirre et al. 2022; H22a; H23). In Phase II, the cable reflections have been mitigated at the instrument level by switching to RFoF cables with a sufficient length to push residual reflections outside the delays of cosmological interest ($\gtrsim$2700 ns). However, in Phase II, we have found an increased level of MC, which we define as the reflection or re-emission of sky signal from one antenna (whether from the feed or glinting off the dish) into surrounding antennas. The cause of this increased amplitude of MC in Phase II is likely the increased sensitivity of the new Vivaldi feeds at low elevation angles (in other words, the Vivaldi feeds have a larger vertical cross section than their PAPER-style predecessors), in addition to the removal of a cage that surrounded the Phase I feeds, which had the undesirable effect of increased dish-to-feed reflections.

N. S. Kern et al. (2019, 2020a) developed a semianalytic, re-radiative coupling model that produced a good phenomenological match to the observed Phase I systematics, which only necessitated a two-element coupling model to achieve noise-limited suppression. Building on this, A. T. Josaitis et al. (2021) and E. Rath et al. (2025, hereafter R&P25) extended the re-radiative MC model to multi-element terms that are needed to model the more complex Phase II coupling systematics. In fringe rate versus delay space, where fringe-rate is the Fourier dual of observing time (measured in mHz) and delay is the Fourier dual of observing frequency (measured in nanoseconds), the effect of MC can be characterized geometrically. Since the mutually coupled signal for baseline ij can be cast as the addition of down-weighted and delayed copies of all other baselines that include i or j, we expect that in general, short baselines that probe low fringe rates will correspondingly exhibit short delays, and vice versa. For HERA's array geometry, this results in a characteristic "X" shape of enhanced power when viewed in fringe-rate versus delay space, with the center of the "X" at a delay of zero, and a fringe-rate corresponding to the east–west projection of the baseline ij:

$$f_{r,\mathrm{peak}} \approx -0.85 w_{\oplus}(b_{ij,\mathrm{EW}}/\lambda), \tag{19}$$

with $w_{\oplus}$ the angular velocity of Earth's rotation.

This geometric picture, which is laid out in detail in R&P25, is borne out by this dataset, as can be seen in Figure 9. In this figure, the left panel shows the amplitude of the visibilities for the 29.2 m east–west oriented baseline group, after averaging over both redundant baselines and nights, in the fringe-rate versus delay space. Immediately evident is a bright central core at zero delay and a fringe-rate of -1 mHz, which represents the bulk of the spectrally smooth foreground emission as observed by this (short) baseline. At a fringe-rate of zero, there is a horizontal "bar" that captures the nonrotating components of the observed power—in particular the coupling of autocorrelations into cross-correlations (N. S. Kern et al. 2019), as well as horizon-based effects such as the "pitchfork" (N. Thyagarajan et al. 2015) and also MC from north–south oriented baselines.

Conversely, at zero delay there is a tall vertical bar that can be attributed to bright sources crossing the horizon. Such sources are intrinsically smooth (and therefore appear at low delay), but the sharp transition from below to above the horizon spills power over all fringe rates. This has been confirmed by a lack of such a feature in simulations in which no bright sources cross the horizon. Finally, there is an "X"-like feature, crossing at the center of the foreground blob, that corresponds to the bulk of MC, as we have discussed. This MC signal appears to be at an amplitude of $\sim$ 1% of the foreground amplitude (in visibility units), nominally a factor of $\sim$100$\times$ the cosmic signal amplitude.

Unfortunately, while our semianalytical models of the MC confirm our general picture of the origin of this excess power, they are not detailed enough to enable a subtraction of the systematic at the visibility-level. While this remains a future goal, in this work we rely on the mitigation techniques laid out in R&P25 and R. Pascua et al. (2024). In particular, we use a pair of filters applied in fringe-rate space to mitigate both the crosstalk and MC.

Since autocorrelations coupled into cross-correlations present as excess power at zero fringe-rate, we mitigate this by applying a "notch" filter that removes power for fringe-rate modes within $\sim$10 μHz (a few bins) of zero.

We additionally apply a "main-lobe" filter, as discussed in R&P25, to mitigate the effects of MC. We design the main-lobe filter to nominally retain 90% of the 21 cm signal, implemented as a DPSS filter, according to the prescription in R. Pascua et al. (2024). In practice, the signal loss is slightly lower than the nominal 10% target (which we correct for; see Section 3.9.3). Figure 9 illustrates this filter by overlaying a transparent screen over the fringe rates that are filtered out for this particular baseline (29 m east–west).[58] The filter retains the bulk of the sky signal (the bright-red regions toward the center) while eliminating much of the characteristic X pattern of the MC.

3.9.2. Coherent Time Averaging

Next we coherently average the visibilities from each interleave within larger LST bins of $\sim$270 s. This range is intended to be as close to 300 s as possible while maintaining an integral number of LST bins in each time average. When performing this average, we first rephase each visibility (see Equation 8) within each interleave to a common central LST (the mean of the LSTs in *all* interleaves for that bin). Then we convert the data in each interleave into pseudo-Stokes

[58] Note that for this baseline, the main-lobe filter overlaps with the notch filter, but this is not true in general, particularly for predominantly north–south baseline orientations.

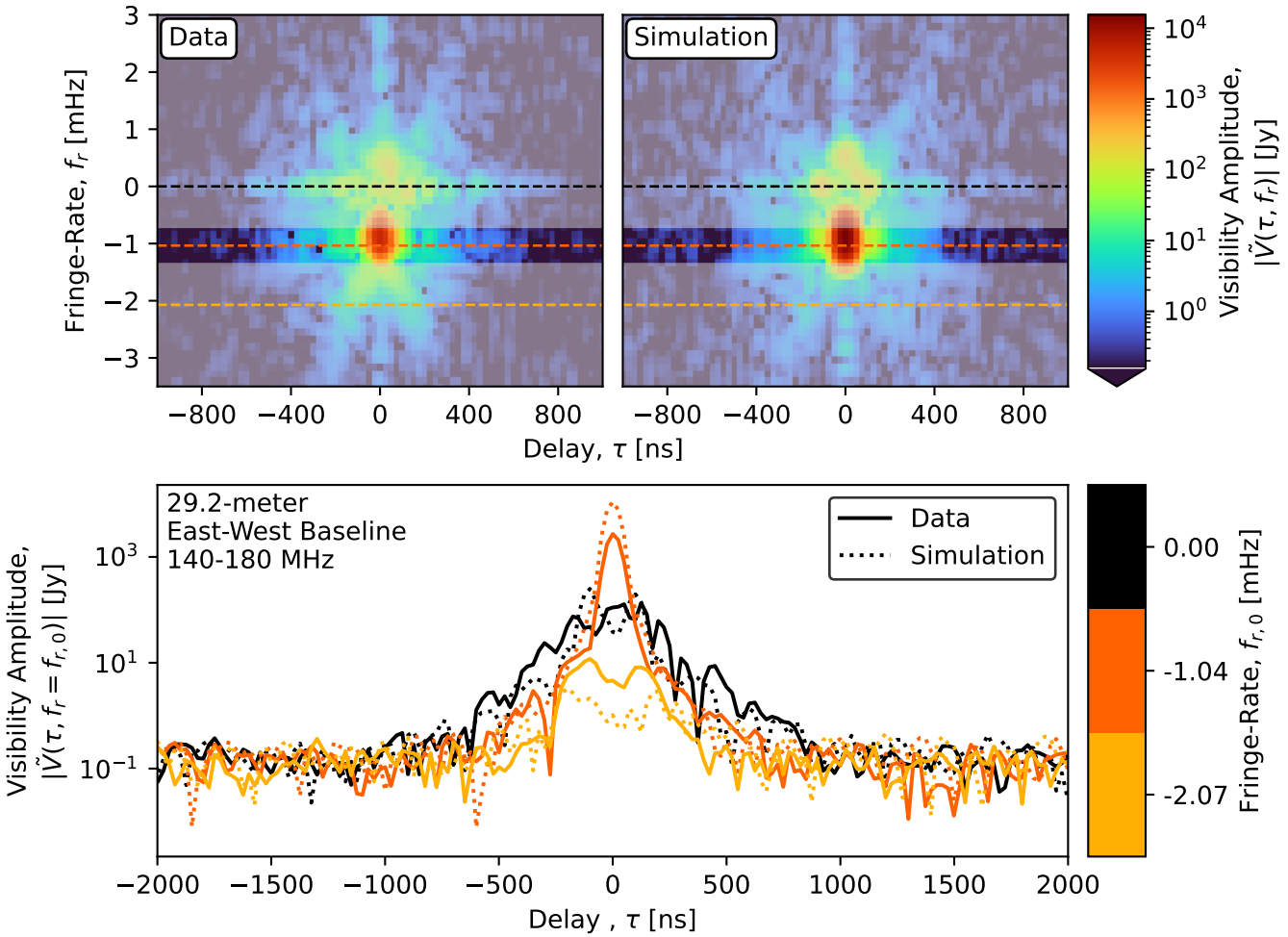

Figure 9. Comparison of real data to validation simulation for a single 29.2 m east–west baseline. The top panels display data on the left and simulation on the right. Each panel has delay on the x-axis and fringe-rate on the y-axis, with colors representing visibility amplitude. Three horizontal dashed lines are shown, at fringe rates of zero (black), ~ -1 mHz (dark orange, representing the peak fringe rate of sky-locked power for this baseline, see Equation (19)) and ~ -2 mHz (yellow, representing non-sky-like power). Semiopaque white overlay represents fringe rates that are filtered out using the "main-lobe" filter of Section 3.9.1. Note that the fringe rates corresponding to the main-lobe (transparent horizontal window in this plot) shift up and down depending on the east–west projected length of the baseline and the frequency of observation. The bottom panel focuses on cross sections of the top panels at the three dashed lines. This illustrates that while there is some disagreement between the mean foreground amplitudes of simulation and data, the overall shape in fringe-rate/delay space is remarkably consistent, and that the noise level (high-τ amplitude) is well matched.

representations, from their native instrumental polarizations:[59]

$$\begin{pmatrix} I \\ Q \\ U \\ V \end{pmatrix} = \begin{pmatrix} V_{xx} + V_{yy} \\ V_{xx} - V_{yy} \\ V_{xy} + V_{yx} \\ i[V_{yx} - V_{xy}] \end{pmatrix}. \tag{20}$$

Here we note that throughout the analysis we use the "sum" convention for converting instrumental polarization to pseudo-Stokes, as denoted in the above equation, instead of the "average" convention, which places a factor of 0.5 on the RHS. Absolute calibration is, by necessity, performed with the same convention.

[59] We use the convention in which the total intensity is the sum of polarizations, rather than their average (hence, Equation (20) does not have a factor of 0.5). This is applied consistently during absolutely calibration.

3.9.3. Signal-loss Estimates

The fringe-rate filters and coherent time average both induce some level of attenuation of the cosmic 21 cm signal (A. R. Parsons et al. 2016). In order to account for this attenuation when reporting our limits on the 21 cm power spectrum, we compute correction factors based on the signal loss expected for the filters and time averaging applied to the data. In H22a, these were computed via Monte Carlo trials against mock data with a known signal amplitude. In this work, we follow the procedure laid out in R. Pascua et al. (2024), who constructed an effective "filter transfer matrix" that represents the cumulative effect of fringe-rate filtering and coherently averaging the data, which we then use to compute the expected signal loss for each baseline.

While the main-lobe fringe-rate filter (designed to mitigate MC) is designed to attenuate $<10\%$ of the 21 cm signal power, this target is computed using a simple top-hat filter in fringe-rate space. In practice, we use a DPSS filter, which always extends to some extent farther than the nominal fringe-rate window, resulting in lower than 10% loss. However, for

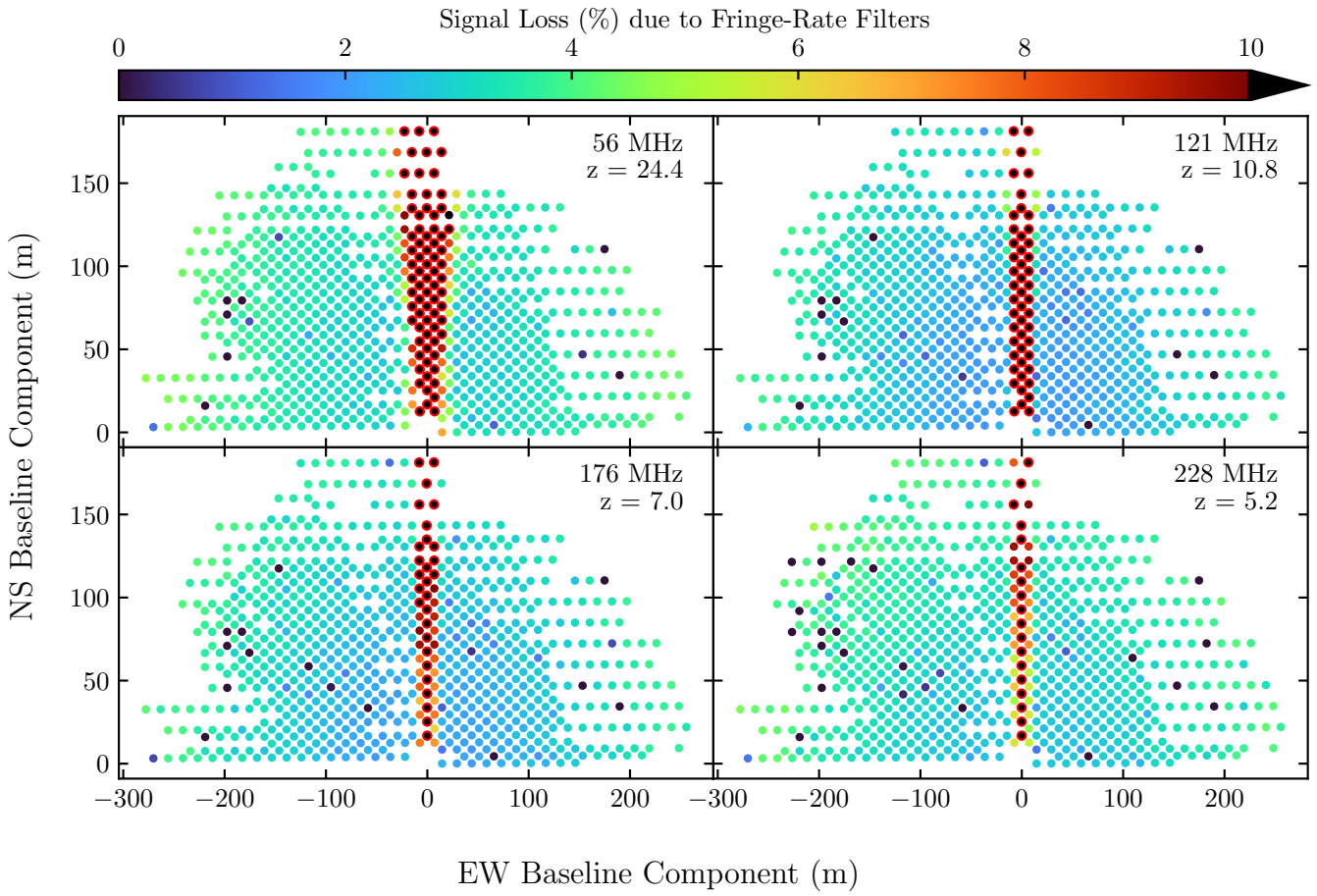

Figure 10. Estimated signal loss from the combined notch and main-lobe fringe-rate filters described in Section 3.9.3, as a function of baseline vector. Circles with red outlines indicate baselines whose estimate loss exceeds the threshold of 10% beyond which the baseline is omitted from further analysis. Each panel is a different spectral band, with low frequencies tending to induce more signal loss. Signal loss tends to be highest for north–south baseline orientations, and also grows modestly with increasing baseline length.

baselines with short east–west projections the main-lobe fringe rates are close to zero, overlapping with the notch filter. This increases the combined signal loss from both filters, sometimes exceeding the target 10%.

The power spectrum estimate for each baseline is corrected for its expected signal loss via

$$P_{ij} \rightarrow P_{ij}/(1 - L_{ij}), \tag{21}$$

where the loss is baseline- and spectral-window-dependent, but LST-independent. Ultimately, since the estimates of the signal loss depend on simplified sky and beam models, we conservatively throw away baselines within a spectral window if the estimated signal loss is >10%. Figure 10 shows the estimate per-baseline signal loss from the two fringe-rate filters, with baselines outlined in red excluded from the final power spectrum estimates. Excluded baselines tend to have short east–west projections, as discussed.

3.9.4. Power Spectrum Estimation

Our methodology for estimating power spectra from the LST-binned, coherently averaged visibilities we have now obtained (still in separate interleaves) is essentially the same as that used in both H22a and H23, and we do not repeat it here.

The only substantive difference in this analysis is that instead of forming cross-spectra between different baselines within the same redundant group, we form them between different interleaves for the same redundantly averaged unique baseline. Furthermore, to normalize the power spectrum, we use the updated model of the Phase II Vivaldi primary beam computed in N. Fagnoni et al. (2021a). As in H23, we use a Blackman–Harris spectral taper when Fourier transforming the visibilities.

We note that, similarly to our previous upper limits, we do not use empirical data covariances (including the number of integrated samples we have thus far propagated) to weight the power spectrum estimate. Doing so is subtle and requires great care not to introduce signal loss (Z. S. Ali et al. 2015; C. Cheng et al. 2018; M. Kolopanis et al. 2019), and we defer this to future work.

3.9.5. Error Bar Estimation

Nevertheless, we do use the number of samples integrated into each averaged datum in order to predict the errors on the estimated power spectra arising from thermal noise. Here we follow the same formalism presented in J. Tan et al. (2021) and outlined in H22a, with one small correction. We calculate P_N (the expected standard deviation from pure thermal noise) for a given baseline type, interleave pair, and spectral window as

$$P_N = \frac{X^2 Y \Omega_{\rm eff} T_{\rm sys}^2}{\chi_{\rm coh}\, \Delta t N_{\rm cohererent} \sqrt{2 N_{\rm incoherent}}}, \tag{22}$$

which is the same as Equation (23) of H22a, but with an extra factor $\chi_{\rm coh}$, which we will define shortly. The system temperature, $T_{\rm sys}$, is estimated using the antenna-averaged autocorrelation, averaged over the band with a Blackman–Harris weighting. $N_{\rm coherent}$ is the band-averaged number of samples averaged into the data, from different nights and baselines within the redundant group (i.e., N given by Equation (17), but averaged within the spectral window). At this stage, $N_{\rm incoherent} = 1$.

In Equation (22), the factor $\chi_{\rm coh}$ is a correction factor that accounts for the fact that neighboring LSTs within a single interleaved stream have become highly correlated by the fringe-rate filter applied for mitigating systematics. One can think of this factor as being, roughly speaking, the ratio of the timescale of the fringe-rate filter to the timescale of coherent averaging applied. More precisely, the factor is computed per spectral window and baseline group as follows.

First, the visibility noise variance for a particular baseline group can be estimated from the autocorrelations using Equation (1), where $N_{\rm samples}$ is the sum of the number of unflagged baselines in the baseline group for each night of observation at the LST in question. Since the baselines have been redundantly averaged, we construct the noise variance estimates from the LST-stacked, redundantly averaged autocorrelation measurement. We can pack these variances into a diagonal noise covariance matrix for each frequency, baseline group and interleave, k, which we denote $\boldsymbol{N}(\nu, k)$ (omitting baseline subscripts for brevity[60]).

The fringe-rate filter, rephasing, and coherent averaging steps are all linear operations on the visibilities. This means we can compose the operations into one linear operator, which we denote as $\boldsymbol{T}(\nu, k)$, which is diagonal in frequency for a given spectral window and operates independently per interleave. $\boldsymbol{N}$ is a diagonal square matrix of size (N_T, N_T) for each ν, k, and baseline group, where N_T is the number of times in one interleave at the LST-binning resolution. $\boldsymbol{T}$ is a rectangular matrix of size (N_t, N_T), where N_t is the number of times in an interleave after fringe-rate filtering and coherent averaging. Note that $N_t \neq N_{\rm incoherent}$ because some LSTs are not used in the incoherent average. The noise covariance matrix can be propagated through these processing steps as

$$\boldsymbol{N}'(\nu, k) = \boldsymbol{T}(\nu, k)\boldsymbol{N}(\nu, k)\boldsymbol{T}^{\dagger}(\nu, k). \tag{23}$$

From this, we can define a correlation matrix at each frequency and interleave elementwise as

$$\rho_{tt'}(\nu, k) = \frac{N'_{tt'}(\nu, k)}{\sqrt{N'_{tt}(\nu, k)N'_{t't'}(\nu, k)}}. \tag{24}$$

The correction factor at a given frequency and interleave is then defined as

$$\chi_{\rm coh}(\nu, k) = \frac{1}{N_{\rm t}}\sum_{tt'}|\rho_{tt'}(\nu, k)|^2. \tag{25}$$

This factor can be thought of as the ratio of total number of coherently averaged times to the "effective number of independent times" in the sense that it can be used to preserve the variance of the χ^2 statistic without explicitly modeling the correlations.[61] It turns out that $\chi_{\rm coh}(\nu, k)$ is very smooth as a function of frequency and varies little across different interleaves. So, to save on computational cost, we calculate this for every tenth frequency channel in the spectral window and average over the values computed for each of these frequencies and across interleaves. The correction factor is then applied to the noise variance, P_N. However, because these power spectra still have noise that is correlated in time, an additional correction factor will need to be applied after incoherent time averaging, as we discuss below.

3.9.6. Incoherent Power Spectrum Averaging

Finally, we average the power spectrum estimates over both LSTs and interleaves for each baseline.

First, we average over LSTs between 1.25 and 5.75 hr. As described in Section 2.2, we choose this range of LST because it is well covered over all of the nights in our dataset and avoids some strong high-rotation-measure sources entering our main beam (e.g., B0628-28; see Figure 2). LSTs outside this range were found to exhibit spurious high-delay spectral structure, either due to polarization leakage from these sources, or perhaps strong MC from bright sources in primary-beam side-lobes. For these limits, we conservatively choose to ignore these LSTs, though in the future we will seek to mitigate the structure in the extended LST range.

When incoherently averaging (whether over LSTs or other axes), we always weight each sample by its inverse noise variance, P_N^{-2}. We furthermore update P_N in the resulting average accordingly:

$$P_N^{\rm avg} = \sqrt{\frac{\chi_{\rm coh}}{\sum P_N^{-2}}}. \tag{26}$$

The factor of $\chi_{\rm coh}$ from Equation (25) accounts for the fact that power spectra still have correlated noise in time (due to the fringe-rate filter), and it is only applied to the incoherent average over LSTs. For averages over baseline or interleaved-pairs, it is taken to be unity.

3.9.7. Refined Error-bar Estimation

Due to cross-terms between the signal (including foregrounds) and noise, the variance of the measurement includes a term proportional to the signal power, along with the pure-noise power quantified by Equation (22). As was shown in J. Tan et al. (2021), and used in H23, the variance of $\hat{p}$, excluding cosmic variance, can be written

$$\mathrm{Var}(\hat{p}) \approx \tilde{P}_{\rm SN}^2 = \sqrt{2}P_sP_N + P_N^2 \tag{27}$$

$$\approx \sqrt{2}\hat{P}_sP_N + P_N^2 - P_N^2/\sqrt{\pi}, \tag{28}$$

where P_s is the signal power (including foregrounds), and $\hat{P}_s$ is an estimate of that power from the same noisy data that goes into $\hat{p}$. J. Tan et al. (2021) showed that subtraction of $P_N^2/\sqrt{\pi}$ accounts for double-counting of this noise power, yielding unbiased estimates of the variance for power spectra that have been incoherently averaged. In this work, to reduce this double-counting, we calculate P_s for interleave k by using *all*

[60] The calculation of $\chi_{\rm coh}$ does not mix frequency channels or interleaves until the very last stage, so we keep these as functional dependencies throughout this derivation. However, baseline groups are *never* mixed, and so we omit these for notational clarity.

[61] https://reionization.org/manual_uploads/HERA132_chi_square_with_correlated_random_variables.pdf

other interleaves, i.e.,

$$P_{s,k} = \frac{1}{N_{\rm intpairs} - 1} \sum_{k' \neq k} \hat{p}_{k'}, \tag{29}$$

where $\hat{p}$ is the LST-averaged power spectrum. This reduces the double-counting correction by a factor of $N_{\rm intpairs} - 1$, where the number of interleave pairs is $N_{\rm intpairs} = N_{\rm interleaves}(N_{\rm interleaves} - 1)/2 = 6$. To de-bias the estimate of the variance at high delays, where the signal is subdominant to the noise, we therefore subtract a modified correction term:

$$\mathrm{Var}(\hat{p}) \approx \tilde{P}^2_{\rm SN} = \sqrt{2}\,\hat{P}_s P_N + P_N^2 - \frac{P_N^2}{\sqrt{\pi}\,(N_{\rm intpairs} - 1)}. \tag{30}$$

We use this estimator for our error bars throughout the rest of the paper.

Finally, we incoherently average over the $N_{\rm intpairs}$ interleave pairs, again weighting by P_N^2. We find that histograms of the power normalized by P_N for $\tau > 1000$ ns on any particular baseline are very well described by a standard normal distribution, as expected.

3.10. Cylindrical and Spherical Averaging

Our ultimate products from this analysis are cylindrically and spherically averaged power spectra. Similarly to the analysis of previous limits, we focus on the spherically averaged spectra and use the cylindrically averaged spectra predominantly as a diagnostic of systematics. However, we note that for data more sensitive than that reported here, it will become more important to consider the cylindrical PS when performing inference due to anisotropies caused by redshift space distortions (D. Breitman et al. 2025, in preparation).

3.10.1. Error Covariance and Window Functions

Thus far we have used only the variance, P_N^2, to weight incoherent averages, treating correlations between the averaged samples in an approximate way.[62] While this has been justifiable for averaging over LSTs and interleaves, we will soon be averaging over delays, which are known to be highly correlated due to the application of a Blackman–Harris frequency taper.

Here, we describe how we model the covariance of our data, as well as the window functions that relate the true underlying power spectrum to the measurements. In principle, the data is correlated between LSTs, baselines, and delays, resulting in very large and computationally demanding covariance matrices. However, we have already treated the correlation between LSTs approximately via the coherent-average correction factor $\chi_{\rm coh}$, and the correlations between baseline types are negligible. In that case, the covariance of the LST-averaged data can be defined as an $N_\tau \times N_\tau$-matrix per baseline and spectral window, which greatly reduces computational complexity.

The true data covariance depends in a nontrivial way on the data-inpainting process (see Section 3.8.4), which fills flagged channels using information from surrounding channels (K.-F. Chen et al. 2025), correlating neighboring delays in the power spectra. Nevertheless, we find that the effects of inpainting on the covariance are negligible; not only are our spectral windows specifically chosen to minimize the amount of inpainting required, but the integration times with large gaps are completely flagged. Moreover, the main application of the error covariance in this work is to help us correctly combine neighboring delay bins to reduce residual correlations (see Section 3.10.2). In this regime, the correlation is dominated by the frequency taper.

Given these considerations, in this work we choose to model the power spectrum covariance simply using the frequency taper:

$$\mathbf{C}_{\tau\tau'} = P_{\rm SN}(\tau) P_{\rm SN}(\tau') |\mathcal{F}_\nu\{\mathcal{W}^2_{\rm BH}\}|^2(\tau - \tau'), \tag{31}$$

where $\mathcal{F}_\nu$ denotes the Fourier transform over frequency. We found that this approximation is accurate to within a few percent for the four delay–delay bins closest to the diagonal.

To understand the relationship of the true underlying power spectrum to the measurements, we also need to compute the window functions. Exact window functions in the absence of data defects such as flags were derived in A. Gorce et al. (2023). The generalization of these window functions to the case in which flagged gaps are inpainted per-night, as we have done in this analysis, is presented in K.-F. Chen et al. (2025). They found that for the flagging patterns characteristic of this dataset, accounting for per-night inpainting results in differences of <1% in the FWHM of the window functions. We ignore this small effect for the sake of computational performance, using the formalism of A. Gorce et al. (2023).

Figure 11 shows an example of the window functions of this dataset, for the lowest-frequency band ($z = 24.4$), after delay binning and spherical averaging (which are described in the following subsections). In Figure 11 we explicitly show the window function values at the k-modes at which they are calculated as circle markers. Importantly, since we estimate the window functions at the same k-resolution as the power

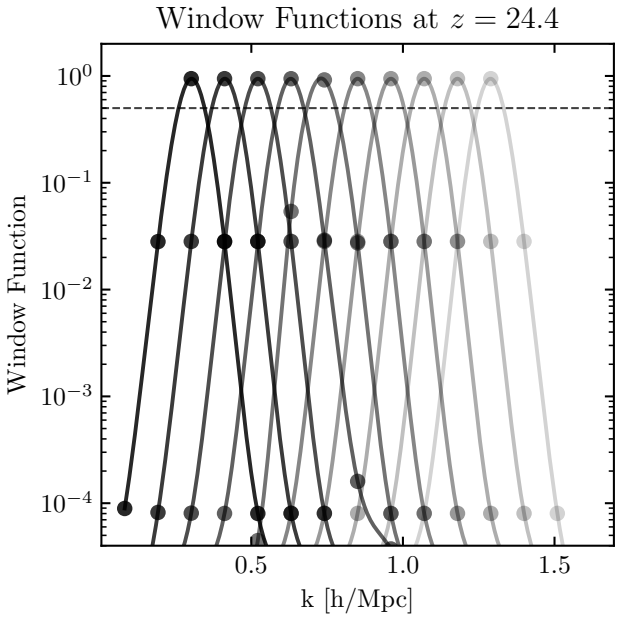

Figure 11. An example set of window functions at $z = 24.4$. The window functions are computed only sparsely, at the circular markers, and are interpolated at third order in log-space. The window functions are largely symmetrical down to the 0.01% level, and are generally ∼3% at the neighboring k-mode.

[62] That is, we ignored correlations between interleaves, and treated correlations between LSTs within a single interleave via the "effective" correction, Equation (25).

spectra, they are rather sparse, dropping to 10^{-4} of the maximum within two k-bins. When computing the FWHM (which we use only to indicate horizontal error bars in Figure 16, and for no analysis purposes), we interpolate with a cubic spline in log-space (as visualized in Figure 11), which is a good approximation to the analytic form of the window function as presented in A. Gorce et al. (2023).

3.10.2. Delay Binning

Our upper-limit likelihood, as presented in Section 6.2, assumes each k-mode to be uncorrelated. As described in Section 3.10.1, for a particular baseline, the power spectrum estimate between two $k_{||}$ modes is in general correlated, with a correlation length of $\sim$4 modes. We reduce correlations between neighboring k-modes by averaging within wider $k_{||}$-bins, each containing four modes, using a weighting kernel $\mathcal{K} = \{0.02, 0.48, 0.48, 0.02\}$. This particular kernel was chosen such that the resulting correlations between neighboring (averaged) k-bins are $<1\%$.[63] The $\tau = 0$ mode is left unaveraged, so that the first bin corresponds to the first four positive delay modes, and the positive and negative delays are kept symmetrical.

The delay–delay covariance is also binned appropriately:

$$\boldsymbol{C}'_{mn} = \sum_{1 \leqslant i,j \leqslant 4} \mathcal{K}_i \mathcal{K}_j \boldsymbol{C}_{4m+i,4n+j}, \tag{32}$$

resulting in re-binned error bars $P_{\rm SN}$ derived from the diagonal of the binned covariance, and re-binned window functions:

$$W_{k_\perp,k_\parallel,\tau_m,\boldsymbol{u}} = \sum_{1 \leqslant i \leqslant 4} \mathcal{K}_i W_{k_\perp,k_\parallel,\tau}. \tag{33}$$

3.10.3. Cylindrical Averaging

Given the per-baseline delay spectra $\hat{p}_{ij}$, the cylindrically averaged power is given by first averaging together the negative and positive delays, and then averaging together all baselines of the same length (within a 1 m bin). Each average is weighted by the noise variance P_N^2 (which itself is averaged in the same way). The resulting 2D power spectrum has irregularly spaced bins of $k_\perp$.

3.10.4. Spherical Averaging

Finally, we perform spherical averaging into bins of $k \equiv |\boldsymbol{k}|$ of width $4\Delta k_{||}$ (where $\Delta k_{||}$ is the natural bin width of each delay bin prior to delay averaging), starting at $3\Delta k_{||}$ and ending at $k = 2.5\ h^{-1}$ Mpc. These bins are designed such that the range of delays covered by each delay-averaged delay bin across all baselines is predominantly kept within a single spherical bin.

3.11. Summary of Important Analysis Differences

While many of the core techniques have remained the same between HERA's first two limits and this work, a number of techniques and decisions have changed. Some of these stem from the differing instrumental characteristics of the Phase II instrument, while others reflect improvements in our understanding of systematics, or the implementation of their mitigation. Here, we list all major changes and improvements in the pipeline in a compact way, for easy reference.

Absolute Calibration. In previous limits, we performed absolute calibration using pre-calibrated visibilities observed during quiet skies. Here, we use simulated visibilities, based on a tailor-made sky model including both diffuse and compact sources, allowing us to cover more LSTs. Our solver has also received a significant upgrade, making it far more robust to large differences in phase between the data and the model when estimating phase-gradient degeneracies.

Deeper RFI Flagging. For these limits, our RFI flagging is a deeper, multistage process that begins with rough flags on single integrations, and progresses to deeper cuts based on full nights of data, averaged over baselines. This allows us to more precisely flag frequency-dependent artifacts that would otherwise go undetected.

Coherent Redundant Averaging. In previous limits, we conservatively did not average nominally redundant visibilities, instead incoherently averaging the data from these baselines after power spectrum estimation. Here, we perform coherent redundant averaging, resulting in a considerable increase in efficiency of the downstream pipeline.

LST-calibration. For the first time, in this work we utilize our prior that night-to-night observations at the same LST should be consistent up to thermal noise and calibration errors to perform limited gain corrections informed by the full set of nights. This increases consistency between nightly visibilities, minimizing the spectral structure induced in the average over nights due to frequency- and night-dependent flags.

Per-night inpainting. In previous limits, we performed spectral inpainting only after averaging visibilities over nights. This was done to enable our power spectrum estimator, which is currently unable to handle nonuniform spectral weighting. However, ill effects from nonuniform weighting arise whenever averaging is done, long before power spectrum estimation. In this work, we inpaint each night independently, before averaging them together, to mitigate this spectral leakage.

Fringe-rate Filtering. In past HERA limits, signal-chain systematics such as cable reflections and over-the-air crosstalk have been mitigated by fitting models in delay space using a CLEAN-like algorithm (J. A. Högbom 1974; N. S. Kern et al. 2020b). This has the notable drawback that the cleaning is a nonlinear transform, so error bars and window functions are difficult to propagate. Here, we utilize a linear DPSS model in 2D fringe-rate/delay space to mitigate such systematics, including MC, which has an increased amplitude in this data.

Time-interleaving. Forming power spectra by cross-multiplying the same visibilities leads to noise bias. In previous upper limits, we have avoided this by cross-multiplying different redundant baselines. Here, since we coherently average visibilities within redundant groups, we instead use the approximate redundancy between adjacent 10 s integrations to form our cross-pairs. This also reduces the risk of incurring negative systematics due to decorrelation of systematic phases between nominally redundant baselines, as observed in M. Kolopanis et al. (2023) and described in M. F. Morales et al. (2023).

Delay binning. For past HERA limits, we have formed spherically averaged power spectra in which the spherical bins had a width of about two native bins of $k_{||}$ under the delay approximation. Since adjacent spherical bins were correlated —a property that our so-called 'upper-limit likelihood' cannot

[63] We find that up to 10% residual correlation does not affect the posteriors derived from our 'upper-limit' likelihood, which assumes zero correlation. We discuss this at greater length in Section 6.2.

handle—we then removed every second k-bin. In this dataset, we instead bin delays in the cylindrically averaged spectra with a weighting kernel specifically designed to suppress correlations between adjacent bins to below 1%. We then spherically average the resulting binned spectra.

4. Validation and Statistical Tests

4.1. Validation of the Analysis Pipeline

Similarly to previous HERA upper limits, we performed detailed checks of our analysis pipeline for this work via realistic simulations (for similar validation methodologies employed for other 21 cm experiments; see, e.g., J. L. B. Line et al. 2024, 2025). Much of the framework used for this updated validation effort remains consistent with our original framework, which was detailed in J. E. Aguirre et al. (2022). However, there are a number of components that have been updated for this analysis, and we highlight these components before presenting the results of these detailed checks in Section 4.1.1.

4.1.1. Pipeline Validation Methodology

To validate our analysis pipeline, we produced a realistic simulation of our data, closely matching the detailed observational characteristics of the entire dataset. In broad terms, this simulation includes full-sky models of unresolved point sources, diffuse Galactic emission, and the cosmic 21 cm signal, passed through an accurate visibility simulator including the full 350-element hexagonal core of the HERA array coupled with an electromagnetic simulation of the Phase II primary beam (N. Fagnoni et al. 2021a). We produced these "ideal" simulations for the exact same channels, polarizations, and time stamps as the full dataset, and added instrumental systematics and noise to them: for example, a bandpass shape and MC (see below). We then passed this mock data through almost the exact same analysis pipeline as the real data. Given that the systematic effects included in the simulation conform to the assumptions of the analysis pipeline, the purpose of this effort was not to investigate the optimality of our analysis choices, but rather to check for subtle inconsistencies in the analysis that produce artifacts in the end result.

In comparison to J. E. Aguirre et al. (2022), there are a couple of novel aspects to the simulations performed for this work. The first is that the "ideal" simulations incorporate all 350 of HERA's antennas, whereas in our previous simulations, we produced only the antennas that appeared in the Phase I datasets. Simulating all 350 antennas means that our simulations can be adapted for future datasets for which an increasing number of antennas will be online. However, this came with a significant computational cost, since naively the compute scales as $N_{\rm ant}^2$. To alleviate this cost, we used the new `fftvis` simulator (T. A. Cox et al. 2025) instead of the `matvis` simulator (P. Kittiwisit et al. 2025) used in previous limits. This simulator uses fast nonuniform fast Fourier transform (FFT) algorithms as implemented in the `finufft`[64] code (A. H. Barnett et al. 2019) to compute the radio interferometer measurement equation (RIME; O. M. Smirnov 2011), leveraging the $N \log N$ scaling of FFTs to achieve high performance with arbitrary precision. For the simulation configurations that we required for this work (350 coplanar antennas, and sky models with a number of sources equivalent to a HEALPix map; K. M. Górski et al. 2005; with $N_{\rm side} = 1024$), we found that `fftvis` on CPU completed in a comparable amount of walltime as `matvis` on GPU.[65] Given that CPUs are vastly more available than GPUs at this time, this represented a significant increase in efficiency, allowing us to produce three full simulations (point sources, diffuse Galactic emission, and 21 cm signal) each with 12.5 million sources, 350 antennas, 1536 frequency channels, and 17,280 time stamps. We refer the reader to T. A. Cox et al. (2025) for further details on the `fftvis` simulator, and its strengths and limitations.

Our sky models consist of three components. The diffuse Galactic sky is simulated using the `pygdsm` software, using the GSM sky model (A. De Oliveira-Costa et al. 2008). The native resolution of the GSM sky model is $N_{\rm side} = 512$; we smooth this map with a Gaussian filter of size 1° to avoid aliasing, before up-sampling to $N_{\rm side} = 1024$. The point sources have two components: first, as in J. E. Aguirre et al. (2022), we include the so-called "A-team" sources—10 of the brightest compact sources on the sky (Centaurus A, Hydra A, Pictor A, Hercules A, Virgo A, Crab, Cygnus A, Cass A Fornax A, and 3C44) with a maximum flux density of 11,900 Jy and minimum flux density of 60 Jy at 200 MHz. The positions, fluxes, and spectral indices of these sources are taken from the GLEAM survey (N. Hurley-Walker et al. 2016), with the exception of Fornax A, whose position is taken from its host galaxy, and whose flux density is taken from B. McKinley et al. (2015). Second, we add 12.5 million randomly drawn sources from the flux-density distribution reported in T. M. O. Franzen et al. (2019), using GLEAM data. These sources are uniformly placed on the sky. This model *statistically* corresponds to the true sky, but is clearly not representative of the *realization* of the true sky. This does not affect our fundamental purpose for these simulations: to validate the analysis pipeline for 21 cm power spectrum estimation. Finally, our 21 cm signal model is a Gaussian random field following a power-law power spectrum with spectral index of -2.7. We compute the redshift-dependent field as HEALPix maps using the `redshifted_gaussian_fields`[66] code (Z. E. Martinot 2022). The amplitude of the 21 cm power spectrum is tuned such that, given the known noise levels of the data, we expect it to be subdominant to foregrounds at low-k, subdominant to thermal noise at high-k, but dominant at intermediate k ($\sim$0.5–1 $h\,{\rm Mpc}^{-1}$). While this is clearly unrealistic, it should not affect the analysis pipeline, and offers the chance to understand how the pipeline behaves in a detection scenario.

After producing the ideal simulations, we interpolated the visibilities to the precise time stamps of the observed dataset (individually for each of the 14 nights). Following this, we added thermal noise with an amplitude given by the radiometer equation with $T_{\rm sys} = T_{\rm FG} + T_{\rm rcv}$, where $T_{\rm FG}$ was equivalent to the simulated autocorrelations, and $T_{\rm rcv}$ was set to a frequency- and time-independent value of 100 K.

An important new addition to these simulations is MC. We used the first-order semianalytical model of MC first defined in A. T. Josaitis et al. (2021) and extended in R&P25 to inject this systematic. In this model, the first-order coupled visibilities $\boldsymbol{V}^{(1)}$ are related to the uncoupled visibilities $\boldsymbol{V}^{(0)}$

[64] https://github.com/flatironinstitute/finufft

[65] These simulations and scaling tests were performed on the Bridges-2 Regular Memory HPC at the Pittsburgh Supercomputing Center, via the ACCESS supercomputing scheme (T. J. Boerner et al. 2023).

[66] https://github.com/zacharymartinot/redshifted_gaussian_fields

through a traceless coupling matrix $\boldsymbol{X}$ via

$$\boldsymbol{V}^{(1)} = \boldsymbol{V}^{(0)} + \boldsymbol{X}\boldsymbol{V}^{(0)} + (\boldsymbol{X}\boldsymbol{V}^{(0)})^{\dagger}. \quad (34)$$

In each term, the rows and columns of the matrix index over antenna-polarization pairs, so this provides a fully polarized description of the coupling to first-order in the coupling coefficients. Each term in the coupling matrix is computed analytically from a model of the uncoupled beam, a model of the reflection coefficient at the feed-load interface, and information about the antenna positions and sampled frequencies, as described in R&P25. When creating mock data for a particular night, we first apply MC using all of the baselines formed with the antennas that were taking data on that night. After applying MC, we then perform a down-select to only keep antennas that were not flagged for the entirety of the night. The bandpass we applied to the ideal data was derived from a fit to lab measurements of the signal chain bandpass.

With these systematics applied, our mock data are a good approximation of the real data. This is illustrated in Figure 9, which compares the mock data to real data, represented in fringe rate versus delay. In the top panels, both data and simulation are displayed as amplitude waterfalls in fringe-rate/delay, for a 29.2 m baseline of east–west orientation (two units east–west in the HERA hexagon), and east–east polarization.

Following J. E. Aguirre et al. (2022), in this work we did not simulate RFI. RFI is an incredibly complex systematic whose joint statistical distribution over the various axes of the data (frequency, time, baseline, etc.) is very poorly understood. Instead of attempting to simulate this systematic, we instead opt to inject the flags obtained via quality metrics on the true dataset into the mock data. This allows us to robustly test the impact of flagging gaps on our analysis, but means we cannot explore the impact of residual unflagged RFI in the data.

Ultimately, we passed the mock dataset through *almost* the same analysis pipeline as the true data. The only differences in the pipelines centered on the injection of flags that we just described. While we ran all of the flagging algorithms on the mock data, we did not expect them to detect any significant outliers—and this is indeed what we found. After the flagging steps, we copied the true data flags into the mock data, and continued the analysis and power spectrum pipeline identically for both true and mock data.

The resulting spherically averaged power spectrum estimates for the mock data are shown in Figure 12, in comparison to both the analytic input power spectrum, and also a power spectrum estimate in which the mock data consisted of purely the 21 cm signal without noise or systematics (but observed through the full instrumental pipeline). The 21 cm realization is generally in very good agreement with the input analytic spectrum, with cosmic variance at the largest scales causing some deviations (see black crosses in lower panel for ratios of 21 cm-only realization to input theory). We recall also that the 21 cm signal has been boosted to enable testing different regimes: at low-k, the foregrounds dominate (resulting in noticeable excess power with respect to the 21 cm signal), at high-k, the noise dominates (indicated by the dashed black line), but at $k \sim 0.5$–$1\,h\,\mathrm{Mpc}^{-1}$, the boosted 21 cm signal is marginally detectable. The bottom panel of each subplot in Figure 12 shows the ratio of the power spectrum estimates from the full mock data to the analytic form input into the simulations (corrected for expected aliasing and approximate instrumental window functions[67]). Here, blue symbols indicate estimates within 2σ of the analytic input and orange the converse. Furthermore, triangles indicate points that are consistent with zero at the 2σ level (these are marked at the amplitude of the 2σ upper limit) while circles represent the converse, i.e., mock 2σ detections.

We expect that, due to spectral structure from systematics such as MC, and our imperfect demarcation of foreground-dominated modes when performing spherical power spectrum averaging, low-k modes may have some detections dominated by foregrounds. This is indeed what is seen in the lowest five bands (highest redshifts), where the first 1–3 k-modes are orange circles. While such biased detections are problematic if interpreted as detections of 21 cm signal, in this paper—as for previous limits—we adopt a conservative likelihood that essentially treats all estimates as upper limits regardless of their thermal noise. In this case, we are most concerned with orange points whose ratio to the 21 cm-only realization is less than unity—i.e., points that display statistically significant signal loss. No such points are evident in our validation tests. This gives us confidence that our analysis techniques are accurate.

4.2. Mutual Coupling as the Dominant Systematic

We have claimed several times that MC is the dominant residual systematic in our spherically averaged power spectra at medium delays of $300\,\mathrm{ns} < \tau < 600\,\mathrm{ns}$, and that this effect has been enhanced significantly in Phase II compared to Phase I.

That MC is an important component of the Phase II data is visually evident in Figure 9, where a simulation of data with MC included via the first-order model of R&P25 is compared to real data, and the signature of MC is clearly visible in both. However, this does not demonstrate that this effect is *new* in Phase II. Neither does it necessitate that MC is an important effect in the final spherically averaged power spectrum estimates, since it should be at least partially mitigated by the fringe-rate filter illustrated in that very figure.

To establish that MC is indeed expected to be the dominant systematic at scales just outside the foreground wedge, we performed like-to-like simulation comparisons between Phase I and Phase II with and without MC. We show the results in Figure 13. In this plot, the simulations include only diffuse foregrounds in the sky model as well as thermal noise, and we show just one spectral window, centered at $z = 9.9$. MC, is modeled using the first-order coupling model of R&P25, which depends crucially on the sensitivity of the primary beam at the horizon, as well as the layout of the active antennas. Phase I simulations include the antennas active during Phase I measurements (e.g., H23), and use a model of the Phase I primary beam simulated with CST (N. Fagnoni et al. 2021a). Phase II simulations include all antennas active in this analysis and the primary beam simulation discussed in Section 4.1. While the MC is computed using the set of antennas just described, the power spectra are estimated using a subset of

[67] The effect of the window functions is most evident at k's inside the wedge, where the asymmetry in the window functions there (see Figure 11) moves power to higher k. There are two (effectively competing) effects at higher k: the aliasing of intrinsic power above the Nyquist sampling in frequency to lower k (discussed in J. E. Aguirre et al. 2022), and the behavior of the correlator, which integrates over the power in a frequency channel, rather than sampling it at the midpoint, as had been calculated previously. All of these effects are small in the k-range shown here.

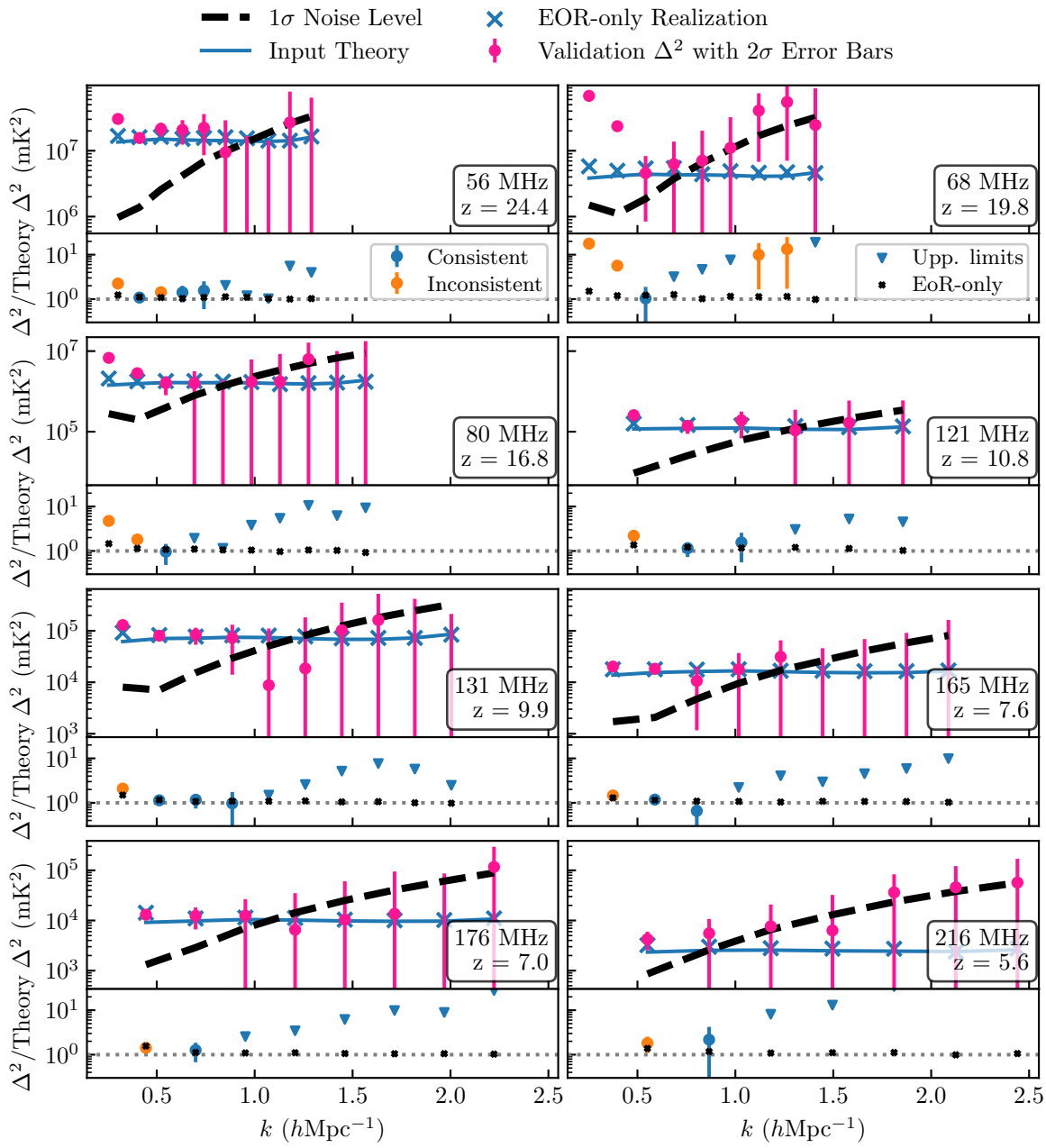

Figure 12. Comparison of estimated power spectra from end-to-end validation simulations (pink circles with error bars) to input analytic 21 cm power spectrum (blue solid line) and EoR-only realization based on this analytic spectrum (blue crosses). The black dashed line is the estimated 1σ thermal noise level, computed using the measured autocorrelations and number of samples integrated together. Each subplot is a different spectral window, representing a different redshift. The bottom panel of each subplot shows the ratio of the power spectrum estimate to the input theory. In this panel, blue points indicate estimates that are consistent with the truth to within 2σ, while orange points are inconsistent at this same significance. Triangles represent 2σ upper limits (i.e., points whose estimate is consistent with zero to within 2σ) while circles represent the converse (detections of nonzero power at 2σ). While all orange points indicate the presence of systematics beyond the noise level, and are therefore of some concern, our primary concern here is orange points with a ratio of less than unity, as these indicate significant signal loss. No such points exist in the validation dataset.

baselines that is the same in all cases, in order to maintain the same signal-to-noise. The data is passed through a fringe-rate filter, as described in Section 3.9.1 to mitigate MC in the spherical power spectrum estimate.

Comparison of the dark- and light-pink hexagons at low k shows that adding MC to the visibilities results in significant extra power. Since the only difference between these points is that one is uncoupled and the other is coupled, we can reasonably conclude that—despite mitigation by fringe-rate filtering—MC is the dominant cause of excess power just outside the wedge.[68] The similarity between the shape of the excess power at lower k in Figure 13 and that seen in the real data (Figure 16) suggests that MC is a likely culprit for this excess power in the data.

Conversely, comparison of the dark-pink hexagons to the dark-blue triangles reveals that—at least insofar as the first-order model of MC reflects reality—the magnitude of MC is significantly increased in Phase II with respect to Phase I. The

[68] This fact is not as readily evident in Figure 12, since the low-k modes are often dominated by the artificially enhanced 21 cm power.

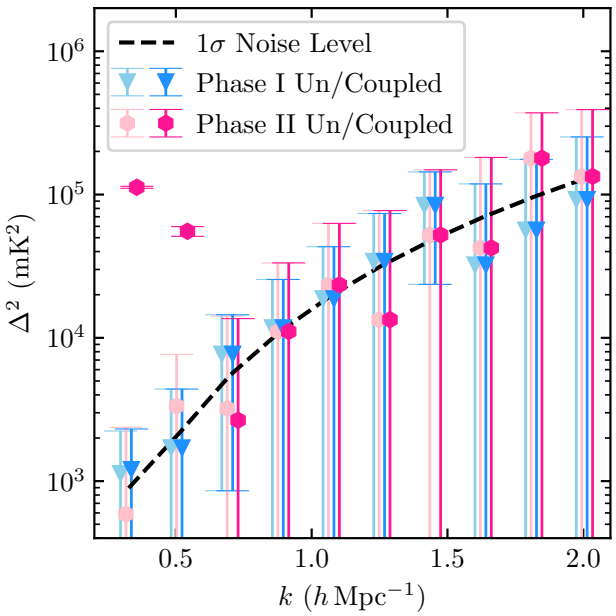

Figure 13. Simulated power spectra demonstrating the difference in mutual coupling between Phase I (blue) and Phase II (pink). Data selection parameters have been adjusted to provide datasets with comparable noise levels (dashed black line): details are described in the text. Mutual coupling (dark points) is simulated using the re-radiative model (A. T. Josaitis et al. 2021) with the amplitude set by the antenna horizon gain, which is much higher for the Phase II Vivaldi antenna. The data is passed through a fringe-rate filter, as described in Section 3.9.1 to mitigate mutual coupling in the spherical power spectrum estimate. The larger coupling results in the excess seen at lower k, which is morphologically similar to the actual result in Figure 16.

reason for this increase, according to the first-order MC model, is the increased sensitivity of the Phase II antennas to the horizon.

4.3. Statistical Tests of Power Spectra

As in previous data releases, we perform a series of tests to check that the statistical behavior of the power spectrum estimates is consistent with the 'null hypothesis' interpretation that random noise dominates our measurements far outside the foreground wedge. If the null hypothesis was rejected in a significant way by any of our tests, and given that a cosmologically viable 21 cm power spectrum detection is expected to be at least an order of magnitude below our present sensitivity level, we would have cause to suspect that systematic contamination remained in our data. Since we marginalize over large positive systematic effects for our cosmological analysis, the presence of systematics in this analysis is less problematic than in a detection scenario, where the precise value of the power spectrum is important, and more detailed systematic models would be necessary. In this context, these null tests mainly act to point us to portions of the analysis that may show what these systematic effects are, so that we can better remove or model them in future analyses where it is more critical to do so.

At the same time, we do note that our validation tests (e.g., Section 4.1) indicate that no *modeled* systematics dominate beyond medium k (though several known effects were omitted from the simulations, including low-level residual RFI, sky model errors, and nonsmooth bandpasses). Positive systematics that may contribute at lower k include MC, low-level RFI, and inpainting residuals.

4.3.1. Test of Sensitivity to Choice of Spectral Windows

In Section 3.8.3, we present a set of spectral windows chosen to minimize spectral structure introduced by high flagging fractions. To ensure that our results are not finely tuned by these choices, we conduct a statistical test to examine the robustness of our power spectra against variations in the definition of spectral windows.

This test is carried out as follows. For each spectral window and the visibility data that have gone through all of the analysis steps until Section 3.9.4, we take two subsets of the frequency channels, one with the upper 10 channels removed and the other with the lower 10 channels removed. We then form two power spectrum estimators and take the difference. As the foreground emission is chromatic, the difference is not noise-like in the foreground-dominant regime. We therefore examine only the statistics of the power spectrum difference for $|\tau| > 1500$ ns. To do this, we generate 100 realizations of noise-only visibility data following Equation 1, using $N_{\rm samples}$ from the LST-binned data (Section 3.8.5). These noise simulations undergo the exact same analysis pipeline as our data and are used to estimate the expected behaviors of the difference between the two sub-band power spectrum estimators at high delays.

Between LSTs of 1.25 and 5.75 hr, for the eight spectral windows, we find that 90.90%, 93.27%, 95.55%, 89.69%, 95.77%, 97.88%, 95.47%, and 94.13% of high-delay bins are within the 2σ confidence interval given by the noise-only simulations, compared to the expected 95% if they were purely noise-like. For almost all spectral windows, the agreement with the noise is very strong. We see that the deviation is slightly larger for the two spectral windows at the low-frequency end and for the spectral window centered at $z \sim 10.8$. The $z \sim 10.8$ spectral window is the narrowest among the eight windows and thus is subject to the largest statistical fluctuations because it has the least number of delay bins. Meanwhile, the evolution of foregrounds within the spectral window is larger at lower frequencies. Overall, this test indicates that our results are robust against variations in the definition of spectral windows.

4.3.2. Test of Signal Loss due to Nonredundancy

Our pipeline makes strong use of the assumed redundancy of different baselines, both for increasing sensitivity and for calibration (see Section 3.3.2). However, several physical factors can lead to baselines that have assumed redundancy to have considerable nonredundancy: for example, variations in antenna positioning, pointing offsets, and antenna primary beam variations. This typically results in baselines within a redundant group containing a distribution of phase offsets from their common mean. These phase differences lead to decoherence when "redundantly" averaging visibilities (see Section 3.7), resulting in a loss of power and therefore signal. We estimate the extent of the loss in the same way as H22a and H23, by forming the statistic

$$\hat{L}_{\rm nonred} = 1 - P^{\tau=0}_{\rm coh}/P^{\tau=0}_{\rm incoh}, \tag{35}$$

where $P_{\rm coh}$ is computed by first coherently averaging visibilities within a baseline group, and then forming power spectra (which is the approach used to generate the upper limits we report in this paper), and $P_{\rm incoh}$ is formed by first

computing power spectra for each baseline (as an auto-pair with itself), and then averaging over these baselines.

This statistic has some desirable properties under a model in which the data consists purely of an underlying signal $V_{\rm true}$, thermal noise n drawn from a complex Gaussian distribution with a standard deviation of $\sigma = |V_{\rm true}|/{\rm SNR}$, and in which each visibility with a nominally redundant group has an additional phase offset $\phi_{\rm off}$ assumed to derive from its nonredundancy:

$$V_{\rm model} = V_{\rm true}\exp(-i\phi_{\rm off}) + n. \quad (36)$$

In the limit of a wide distribution of $\phi_{\rm off}$, regardless of the SNR, the signal loss tends toward unity (i.e., 100% loss). Furthermore, for high-SNR measurements, the statistic tends toward an unbiased estimate of the true signal loss (defined as $P_{\rm coh}/|V|^2_{\rm true}$).

Nevertheless, this statistic is not without its problems. The most obvious source of problems is when the data does not conform to the assumptions of the simple model in which it has the desirable properties: for example, when there are large phase differences between the visibilities, but they arise primarily due to non-21 cm sources, such as bright sources in the side-lobes of the primary beam. Unless these phase differences enter via errant gain calibration due to these sources, any signal loss resulting from such an effect would be limited to loss of *foreground* signal, rather than the diffuse all-sky 21 cm signal. In H22a and H23, this was overcome by selecting the quietest skies at which to evaluate the statistic (around the Galactic ant-center). This is not possible in this analysis, since we do not observe this field; we instead simply measure the loss for all LSTs in our full range of 1.25–5.75 hr.

Additionally, while the estimator is unbiased under this model for high-SNR data, it is not unbiased for low-SNR data,[69] generally leading to higher predictions for the loss than it should have. Low-SNR can occur either when the noise is large, or when the signal is small. The former effect was largely overcome in H22a and H23 by including only highly redundant baselines, limiting those considered for computing the signal loss to <60 m. We follow this approach in this analysis. The latter effect (low SNR due to low signal) can arise at particular LSTs where the signal drops to brief "null." This was overcome in H22a by smoothing $P_{\rm incoh}$ along the LST axis with a 1 hr window. We do the same in this work.

Finally, there is a curious effect in which, for moderate SNR and a wide range of per-baseline variance (stemming from different per-night flags on each baseline), the statistic can predict signal *gain* (when no such gain is present). The relative occurrence of this effect (under the toy model) is strongly dependent on the SNR of the data, the distribution of numbers of unflagged nights within a baseline group, and the size of the baseline group (the smaller the group, the more likely an estimate of signal gain) and is less sensitive to the distribution of phase offsets within the group. We find (relatively rare) occurrences of this signal gain both in simulations of the toy model, and in real data.

After these considerations, we ultimately compute a signal-loss estimate for each baseline group shorter than 60 m (and that also passes the criterion of having lower than 10% signal loss from the time-based filtering; see Section 3.9.1), for each polarization (*XX* and *YY*), each LST between 1.25 and 5.75 hr, and each spectral window. As in H22a and H23, our assumption is that the true signal loss on the *21 cm component* is very similar between baseline groups, polarizations, and LSTs (due to isotropy of the 21 cm signal, and the likely time-invariance of the sources of nonredundancy). Thus, we obtain a single loss per spectral window, as an estimate of the true underlying loss that is consistent across baselines, polarizations, and LSTs.

Figure 14 shows histograms of the loss for each spectral window, where the samples for each histogram come from different baseline groups, polarizations, and LSTs. These histograms exhibit a peaked structure, with long negative tails (out to high signal loss), and short positive tails (including small signal gains). We have confirmed that the long negative tail is dominated by the lowest-SNR measurements, which calls into question whether they should be given full consideration when deciding a final single loss per spectral window (given the known biases for low-SNR measurements from the toy model). On the other hand, high-SNR measurements still have a spread in their estimated signal loss that we have found to be highly correlated in LST (i.e., certain sky distributions tend to produce particular loss estimates that can differ from other skies, even if both are high-SNR). The details of the physical source of the fluctuations in the loss estimate as a function of LST and polarization are of great future interest, but in this work, we take the same approach as H22a and simply use the median of all measurements in the spectral window as our final loss estimate (shown as black vertical lines in Figure 14). The median down-weights the long negative tail, which we have reason to be skeptical of, while capturing the main information coming from the measurements.

Regardless of the particular merits of our choice of using the median, we note that the signal loss estimates in each spectral window are quite small: <6% for more than 84% of the data all cases (leftmost vertical red dashed line in each panel of Figure 14). For the current modality of upper limits, such small losses are not important for astrophysical inference, being dwarfed by other uncertainties, such as theoretical modeling uncertainties. Furthermore, the estimates are similar across spectral windows, giving confidence that the estimates are internally consistent and robust. Under these considerations, the median in each band is a justifiably sufficient estimate of the signal loss, and we report each such median in Table 4.

5. Results

5.1. Cylindrical Power Spectra

We show our estimated cylindrically averaged Stokes-*I* power spectra for each spectral window in Figure 15. In this figure, we also mark the delay corresponding to a source on the horizon (solid black), as well as our final foreground cut corresponding to this horizon line plus a buffer of 500 ns (dashed black line). In this plot, each pixel corresponds to the cylindrical average over all baselines within 1 m length bins, and over four adjacent delay bins (as described in Section 3.10).

In each spectral window, the lowest delay bin is visibly dominated by foregrounds, with a dynamic range of 6–10 orders of magnitude compared to the high-delay modes. In the highest-redshift bands, the second delay bin (and even the third, for longer baselines) is also foreground dominated. As discussed throughout this paper, we attribute the bulk of this power leakage beyond the horizon to MC. This is supported by

[69] This can easily be shown by conducting small Monte Carlo simulations with the toy model described.

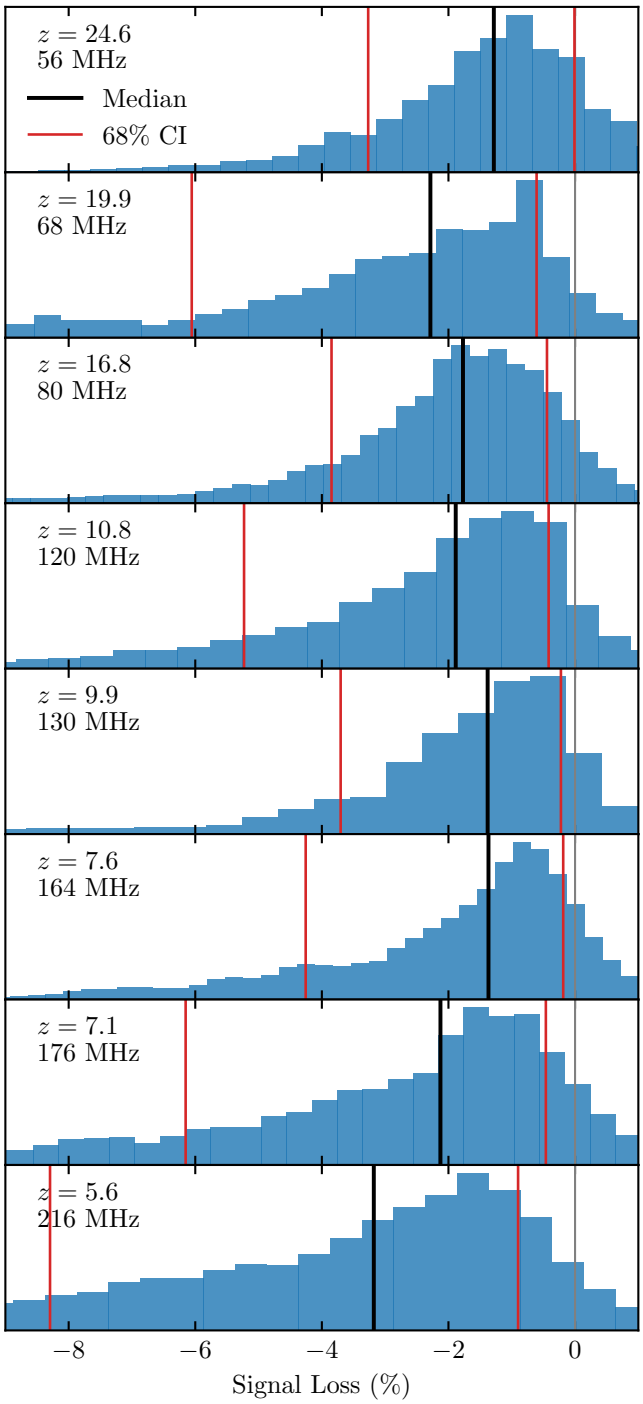

Figure 14. Histograms of estimates of signal loss incurred by nonredundancy of nominally redundant baselines. Each panel is a spectral window, and samples for each histogram come from different baseline types, LSTs, and polarizations. The median, marked as black vertical lines, is taken as the signal loss estimate. The red lines show the 68% confidence interval. A small fraction of estimates are signal *gains*, consistent with toy models (see the text).

Figure 15, which reveals a larger extent of foreground leakage at low frequencies, where MC is expected to be more severe. Furthermore, while the "foreground wedge" (A. R. Parsons et al. 2012; A. Liu et al. 2014; N. Thyagarajan et al. 2015) is certainly present, the dominant foreground structure is rather "brick"-like; that is, there is a strong contribution to foreground leakage that is roughly baseline-independent. This is consistent with predictions from our first-order model of MC, for which the strength of the coupling power at high delay for a baseline type averaged across the array is primarily determined by its orientation, rather than its length.

Table 4
Median Percent Loss per Spectral Window

z	$\bar{\nu}$ (MHz)	Median Loss (%)
24.4	56	1.2
19.8	68	2.2
16.8	80	1.7
10.8	121	1.8
9.9	131	1.3
7.6	165	1.3
7.0	177	2.1
5.6	216	3.1

Our choice of foreground cut, illustrated by the dashed black line in each band, was chosen primarily to ensure that leakage from MC is avoided, based on our observations of delay spectra of much less-averaged data.[70] Visually, Figure 15 suggests that our choice of foreground cut may be somewhat conservative for our longest baselines, though the fractional sensitivity of these baselines with respect to the shorter baselines is quite low, so this conservatism is not likely to affect our final limits to a large extent. On the other hand, we will see evidence in the spherically averaged spectra (Figure 16) that on short baselines, our foreground cut is not conservative enough, at least in the high-redshift bands (< 100 MHz). Since in this analysis we only consider upper limits, inclusion of excess foreground power in our estimates is not an issue; however, it will become very important to better understand the limits of the foreground leakage in the future when a detection might be claimed.

5.2. Upper Limits on Spherical Power Spectra

Figure 16 shows the final spherically averaged power spectra estimated using the Phase II data analyzed in this paper. In the same figure, we show the predicted noise power (black dashed line), as well as the lowest upper limits from H23 at any k, shown in the bands closest to those in which they were measured (green triangles).

The lowest k-modes at each redshift ($k \lesssim 0.5$–$0.8\,h\,\mathrm{Mpc}^{-1}$ depending on the redshift) are consistently dominated by foregrounds, which have leaked well beyond the horizon (all modes inside the horizon have been cut, as described in the previous subsection). Indeed, while the sensitivity of this dataset is comparable to our deepest Phase I upper limit (i.e., the green triangles are roughly consistent with the black lines), our previous limits were significantly lower (approximately an order of magnitude) simply because the foregrounds were not as strong on the lowest k-modes we measure here. This foreground leakage predominantly comes from MC, as we established in Section 4.2. This highlights MC as the greatest cause of sensitivity loss for HERA Phase II and, thus, the greatest and most important challenge for future progress.

Nevertheless, while low-k modes are systematics-dominated, for $k \gtrsim 0.5\,h\,\mathrm{Mpc}^{-1}$, our estimates are almost completely

[70] For example, we inpainted our data out to 500 ns to capture foreground structure from mutual coupling in Section 3.8.4, though this choice is not necessarily tied to our decision on the extent of the foreground cut in the cylindrically averaged spectra.

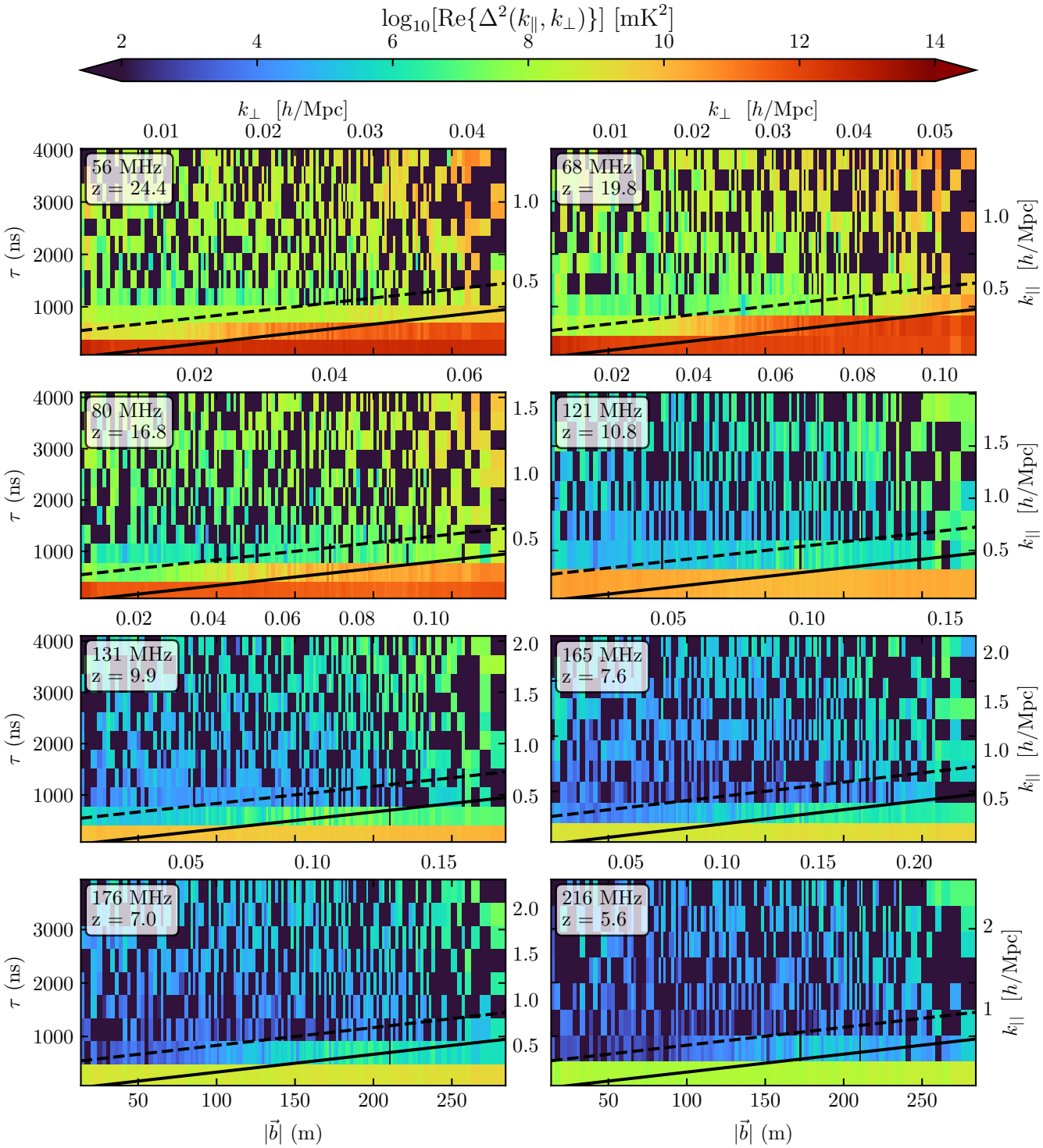

Figure 15. Real parts of the pseudo-Stokes-*I* cylindrical power spectra estimated from the data in this analysis. Each panel shows a different spectral window (redshift). The *x*-axes are linearly spaced in baseline length, identical for all spectral windows, with equivalent $k_\perp$ for our fiducial cosmology marked on the top of each. The *y*-axes are regularly spaced delays, with corresponding $k_\parallel$ marked on the right axis. The horizon line is marked in solid black, and an additional buffer of 500 ns is shown as the dashed black line. Values below 10^{-2} m K^2 (including negative values) are shown in navy. Outside the buffer, the estimates are visually noise-like, with values scattered between positive and negative. The gradual rise in power toward larger baseline lengths is indicative of higher P_N due to the decrease in the size of redundant groups at these lengths.

consistent with thermal noise in all bands, at the 2σ level. This is especially true of the bands above FM frequencies (>100 MHz), which not only have error bars that encompass $P_{\rm SN}$, but also whose estimates are generally scattered evenly about it (estimates not shown with dots are negative, but their error bars extend into the positive far enough to encompass $P_{\rm SN}$). In the sub-FM bands, there is some evidence of an overall positive bias, since most of the estimated values fall above the noise (even though their error bars generally encompass the noise level). Such a positive bias can only affect our inferences in a conservative way in this analysis, but in future work, we will investigate these small biases more thoroughly. In the end,

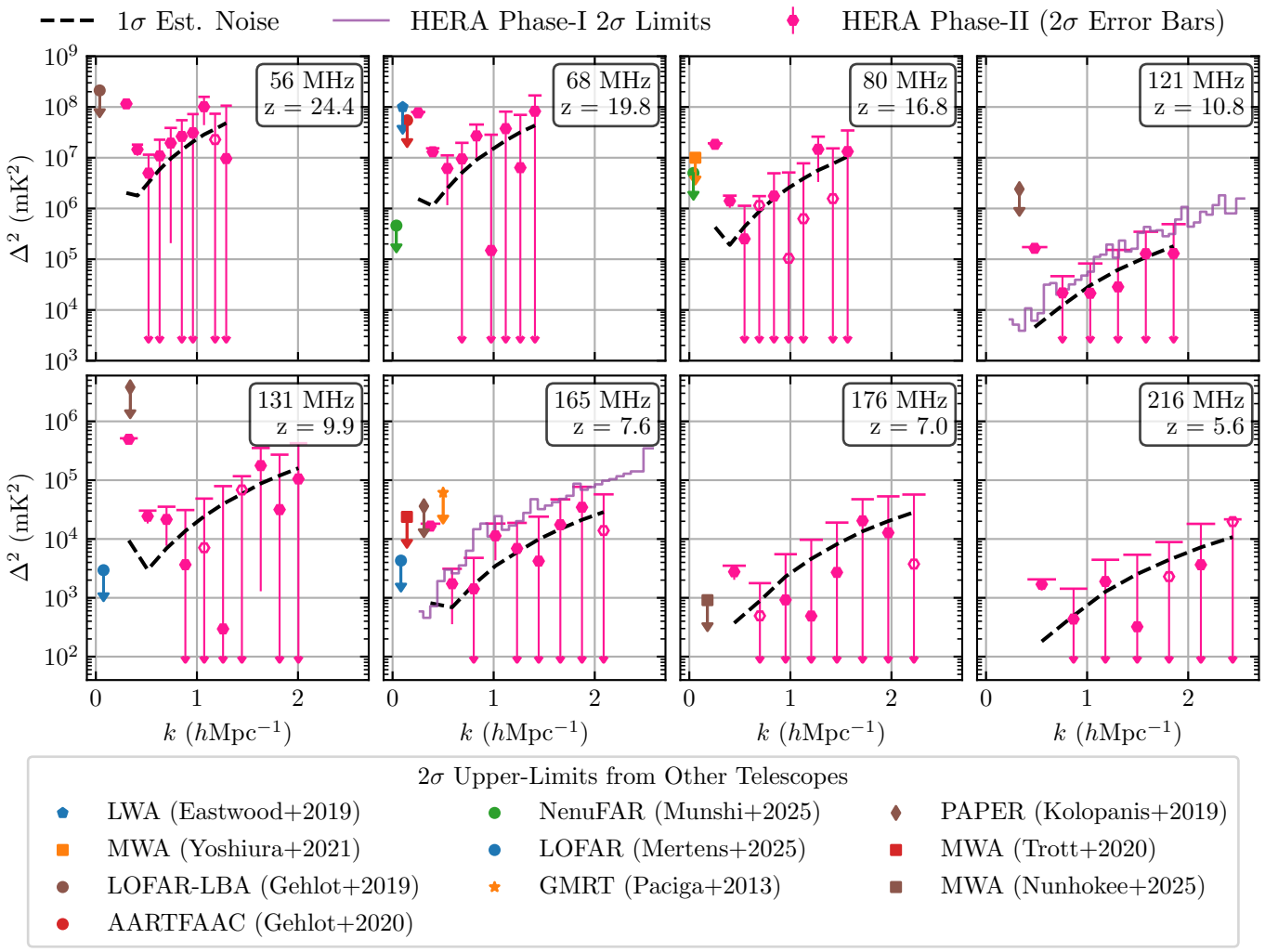

Figure 16. Spherically averaged power spectrum estimates in each spectral window (pink hexagons with error bars). Vertical error bars are 2σ estimates, where the noise is given by $P_{\rm SN}$ (see Equation (30)). The width of the top-cap of each error bar represents the FWHM of the window function (see Figure 11). Each pink marker represents the power spectrum estimate, with empty markers indicating negative values. Where error bars end with arrows, this indicates that the estimate is consistent with zero at the 2σ level. The thick black dashed line shows the thermal noise level, $P_{\rm SN}$. The purple steps in the panels at $z = 10.8$ and $z = 7.6$ denote the 2σ upper limits reported in H23. While sensitivity of this Phase II dataset is similar to the limits of H23, mutual coupling causes spectral leakage of foregrounds out to much higher k, causing these limits to be less strong. A selection of the most stringent upper limits across redshift from different experiments are also shown (G. Paciga et al. 2013; B. K. Gehlot et al. 2018, 2020; M. W. Eastwood et al. 2019; M. Kolopanis et al. 2019; C. M. Trott et al. 2020; S. Yoshiura et al. 2021; S. Munshi et al. 2024; F. G. Mertens et al. 2025; M. J. Wilensky et al. 2023; C. D. Nunhokee et al. 2025). H23 remains the deepest limits to date at any redshift.

it is not surprising that the largest bias arises at the lowest frequencies, where foregrounds and MC are the strongest.

We emphasize that on these noise-dominated scales, the sensitivity of this small 2 week dataset is similar to that of the full 100-night Phase I limits presented in H23. The source of this extra sensitivity is the primarily the increase in number of antennas. The fact that these modes are noise-dominated is promising for future analyses that can utilize a larger portion of the several-hundred nights thus far observed with Phase II.

All estimates and upper limits can be found in Table 5.[71]

6. Astrophysical Interpretation

In this work, we follow the strategy of previous HERA analyses and adopt several theoretical models to infer what our multiredshift limits tell us about cosmology and astrophysics. While our limits at $z < 10$ are less constraining than those presented in H23 (primarily due to foreground leakage from MC), we have limits from a much wider range of redshifts in this dataset. This motivates the adoption of models with mechanisms for producing very strong absorption features during Cosmic Dawn, to see whether such models—motivated by results from the EDGES experiment—are constrained by this multiredshift data. Given that the limits at $z < 10$ are less constraining than those presented in H23, we do not reconsider any reionization-driven scenarios in this work (e.g., J. Mirocha et al. 2022). Furthermore, we note that the models represented here are not exhaustive, even in the context of seminumerical/semianalytical models (e.g., M. Pagano & A. Liu 2020; H. Trac et al. 2022; A. Schneider et al. 2023; P. Hirling et al. 2024; B. Cyr et al. 2024; R. Ghara et al. 2025), but they do span a range of interesting physical scenarios.

6.1. Theoretical Models

We adopt five different theoretical models in this work, using three different simulation codes. The majority of these

[71] This data may also be downloaded in a standard YAML format from the https://github.com/EoRImaging/eor_limits repository.

models include some mechanism for producing values of the volume-averaged ratio of the spin temperature to the background radio temperature that are much lower than standard models allow. So-called standard models assume that the minimum spin temperature is achieved via strong coupling to the gas kinetic temperature, whose standard minimum value is given by adiabatic cooling. Likewise, the radio background is typically assumed to be dominated by the CMB, which yields a prediction for a *lower limit* on T_S/T_R (see the black dotted line in Figure 18). However, motivated by results from the EDGES experiment (J. D. Bowman et al. 2018), a number of nonstandard models have been proposed that can exceed this lower limit, either by reducing T_S via new sources of cooling, or by increasing the radio background via new populations of high-redshift radio sources. Given that such models are often focused on achieving such low values of T_S/T_R at redshifts of $z \gtrsim 14$ (where the EDGES experiment made its measurement), we include these models in this work to investigate if their exotic physics might be constrained by the large redshift range of Phase II measurements.

Specifically, the models we adopt are as follows:

21cmFAST $T_R \equiv T_\gamma$. This model uses the `21cmEMUv1` emulator (D. Breitman et al. 2024)[72] to emulate several observational summary statistics computed by `21cmFASTv3` (A. Mesinger et al. 2010; S. Murray et al. 2020).[73] The underlying astrophysical model is defined in J. Park et al. (2019), and is defined by nine parameters governing the relationship of UV, X-ray, and Lyα emissivity to halo mass and redshift. This model contains no nonstandard contributions to T_S/T_R, taking the usual $T_R \equiv T_\gamma$ assumption, and is equivalent to that used to infer astrophysics in H22b and H23. We assume the same priors over all parameters as in H22b and D. Breitman et al. (2024; see ranges of Figure 6 in both).

21cmFAST Mini-HaloT_R. This model uses v2 of `21cmEMU`, trained on `21cmFAST` models that include a contribution to the radio background from Population III stellar remnants hosted in molecularly cooled galaxies (or mini-halos; see Y. Qin et al. 2020a, 2021). The inhomogeneous radio background sourced by the mini-halos is presented in J. Cang et al. (2024), and the eventual sterilization of Pop III stars by Lyman–Werner and photo-heating feedback provides a physically motivated mechanism to avoid violating limits on the present-day radio background (e.g., D. J. Fixsen et al. 2011). Note that while this model contains a superset of the physical parameters defined in J. Park et al. (2019), in this work we only vary the five minihalo parameters, setting the parameters governing Population II stars to the fiducial ones of J. B. Muñoz et al. (2022). The prior over these five parameters is the same as in J. Cang et al. (2024; see T_{21} component in Table 1).

21cmSPACE UniformT_R. This model uses an emulator (described below) trained on `21cmSPACE` simulations (e.g., E. Visbal et al. 2012; A. Fialkov et al. 2012, 2013, 2014)[74] that include a spatially uniform synchrotron background, which replaces the CMB as $T_R = T_\gamma(1+z)[1 + A_{\rm rad}(\nu_{\rm obs}/78\ {\rm MHz})^{-2.6}]$ with the phenomenological parameter $A_{\rm rad}$ (A. Fialkov & R. Barkana 2019). The `21cmSPACE` simulator includes flexible Population II and Population III star formation in galaxies, contributing to Wouthuysen–Field (WF) coupling, Lyman–Werner feedback, Lyα heating and ionization, as well as flexible X-ray spectra and efficiencies. The simulations used here are described in more detail in T. Gessey-Jones et al. (2024; see their Table 2 for the nine-parameter prior space).

21cmSPACE Inhomogeneous T_R. This model is identical to the Uniform T_R model, except that the excess radio background is instead sourced by radio luminous galaxies, with an efficiency parameter $f_{\rm rad}$, and is spatially inhomogeneous via the SFR (I. Reis et al. 2020). For more details on the simulations, see S. Pochinda et al. (2024). For both the uniform and inhomogeneous T_R models, we employ an updated version of the analysis pipeline from S. Pochinda et al. (2024), built using `PyTorch` and including GPU accelerated distributed training.

Millicharged Dark Matter. This model uses the `Zeus21` analytic code (J. B. Muñoz 2023), which includes both atomic- and molecular-cooling galaxies and a flexible kinetic temperature from the Boltzmann solver `CLASS` (D. Blas et al. 2011). Here we replace the adiabatic-cooling gas temperature in `Zeus21` with a temperature consistent with cooling from millicharged dark matter (J. B. Muñoz & A. Loeb 2018; R. Barkana et al. 2018; A. Berlin et al. 2018), and include the T_k–T_S correction from S. Mittal & G. Kulkarni (2021) for very low temperatures.[75] As in the "Minihalo T_R" model, we keep the Pop II stellar parameters fixed to the fit to the HST and JWST UVLFs from J. B. Muñoz et al. (2023), except the X-ray efficiency L_{40} (in units of 10^{40} erg s^{-1} per unit SFR), and vary the Pop III star formation efficiency $f_{\star,III}$ (for more details on the code, including implementation of feedback, see H. A. G. Cruz et al. 2024). In addition, we assume a fraction of 0.5% for millicharged dark matter and vary over its mass m_χ and charge Q_χ (in units of the electron). We assume log priors on all parameters over the following ranges: $-3.3 < \log_{10}(m_\chi) < -0.7$, $-7 < \log_{10}(Q_\chi) < -2$, $-3 < \log_{10}(L_{40}) < -3$, and $-4 < \log_{10}(f_{\star,III}) < -0$.

6.2. Choice of Likelihood

Each of the theoretical models employed here computes spherically averaged power spectra, $\Delta^2_{\rm theory}(\boldsymbol{k}, \boldsymbol{z})$ for vectors of $\boldsymbol{k}$ and $\boldsymbol{z}$ defined by each theory code. These models are interpolated to the scales and redshifts of the data, and processed through the instrumental window function, to obtain theoretical residuals. We then use Bayesian inference to obtain posterior distributions for the parameters of each model, $\boldsymbol{\theta}_{\rm mdl}$ as well as functional posteriors for $\Delta^2_{\rm theory}$ and other derived quantities.

Here, as in H22b and H23, we use the so-called "upper-limit likelihood" (R. Ghara et al. 2020) when defining our posterior probability (see Equation (11) of H23). Formally, this likelihood assumes a systematic whose value is uncorrelated between power spectrum bins and may range from 0 to infinity with uniform a priori probability. It also assumes uncorrelated noise σ_i between power spectrum bins, which is not reflective of our actual noise. Our power spectrum is weighted such that neighboring bins have 1% correlation, which is different from previous analyses where we decimated in k-space to avoid modeling noise correlations.

We justify using the same likelihood (i.e., the one corresponding to uncorrelated noise) via a numerical

[72] https://github.com/21cmFAST/21cmEMU
[73] https://github.com/21cmFAST/21cmFAST
[74] https://www.cosmicdawnlab.com/21cmSPACE
[75] We conservatively do not include the additional fluctuations from the velocity dependence of the cooling in this analysis (J. B. Muñoz et al. 2018; H. Liu et al. 2019)

investigation into the marginal distribution of the systematic and noise (jointly between power spectrum bins). We detail this investigation in Appendix E, but in summary, we find that residual correlations between neighboring k-bins that are left unmodeled have a negligible effect on the estimated posterior, even when the bins are up to 90% correlated. This is mostly due to the systematic model, whose large uncertainty and independence between bins washes out the effect of noise correlations.

In producing the constraints shown in the next subsection, all models were run with two different datasets: one using only the upper limits reported here from Phase II data, and another that also adds the more stringent upper limits at $z = 7.9$ and 10.4 from H23. In all cases, each upper limit at each scale and redshift is treated independently.

In addition to this likelihood, the models using the `21cmFAST`/`21cmEMU` code use a variety of other likelihoods (each treated independently): (i) *Thomson scattering optical depth to the CMB*: As in H23, this term is a Gaussian likelihood centered around $\tau_e = 0.0569^{+0.0081}_{-0.0086}$ based on the median and 68% credible interval (CI) from the posterior obtained in Y. Qin et al. (2020b) from the re-analysis of Planck Collaboration et al. (2020) data. The CMB optical depth is an integral measure of the EoR history, disfavoring early reionization. (ii) *Lyman forest dark fraction*: As in H23, this term compares the proposed model's global neutral fraction at $z = 5.9$ with the upper bound $\bar{x}_{HI} < 0.06 \pm 0.05$ at 68% CI obtained with the model-independent QSO dark fraction statistic (introduced in A. Mesinger et al. 2010 and applied to observations in I. D. McGreer et al. 2015). The likelihood function is unity if the proposed global neutral fraction is below the upper bound at $z = 5.9$, then it decreases as a one-sided Gaussian for higher values of $\bar{x}_{HI}$. The dark fraction observations disfavor extremely late reionization. (iii) *UV luminosity functions* (LFs), which compare the model with well-established $z = 6$, 7, 8, 10 UVLFs, observed with Hubble (Bouwens et al. 2015, 2016; Oesch et al. 2018) in the magnitude range $M_{UV} \in [-20, -10]$. For the UVLFs, a Gaussian likelihood is used. UV LFs constrain how star formation is assigned to dark matter halos and, therefore, provide limits on the redshift evolution of the star formation rate density.

The inferences using `21cmEMUv1` with $T_R \equiv T_\gamma$ include all likelihood terms, while inferences using `21cmEMUv2` involve an excess radio background sourced by mini-halos omit (iii), since the UV LF data only constrains the more-massive galaxies that are held fixed in this model.

6.3. Inference Results

Since the upper limits at redshifts 8 and 10 from this work are weaker than those reported in H23, we perform inference with all five models under different combinations of data: (i) only the data from this upper limit, and (ii) this limit combined with the lowest limits from H23 (+Phase I). Naively, we expect that constraints at $z \leqslant 10$ would be dominated by the Phase I data, while the limits reported here may have some additional constraining power at higher redshift. However, given the strong correlations between redshifts for physical models of the power spectrum, strong constraints tied to lower redshifts may dominate even over direct measurements at higher redshifts that are comparatively weaker.

We find that for all models, the constraints on parameters are relatively weak, and very much consistent with the constraints presented in H23. That is, parameter constraints are dominated by the more stringent upper limits from Phase I at $z = 8$, 10. We do not show these constraints directly here, and refer the reader to H23 for further discussion of the models ruled out by HERA—typically "cold reionization" models.

Figure 17 shows the 95% confidence upper limits on the *inferred* power spectrum at $k = 0.5\,h\,\mathrm{Mpc}^{-1}$ as a function of redshift. The first thing to note from this figure is that the low-redshift limits dominate the constraints on each theoretical model. Indeed, the $z > 15$ limits are not quite within the prior region of any of these models, which is significant since four of these models have "exotic" mechanisms for producing significantly boosted power at high redshift (note how far the limits are from the prior of the standard $T_R \equiv T_\gamma$ model: > 3 orders of magnitude). Further making this point, the Millicharged DM model is only constrained by data at $z > 10$, as the model does not include reionization processes, and its constraints are by far the weakest. Because the strongest constraints at lower redshifts come from H23 (when they are included), they are completely dominant. That is, this dataset does not provide any further constraining power—under this set of models—beyond the existing limits.

In Figure 18 we show the derived posterior limits on the ratio T_S/T_R as a function of redshift, which characterizes the mean thermal evolution of the IGM. Here, it is clear from the prior (illustrated by thin dotted lines) that each model is able to produce very low values of this ratio—well in excess of the adiabatic cooling limit.[76] As a comparison point, we show the value of T_S/T_R implied by the reported global 21 cm absorption depth reported by EDGES (J. D. Bowman et al. 2018) as a green marker. All models, by construction, include this point in their prior.

After constraints from HERA data (both from this work and H23), the excess radio background models (blue, orange, and red) all tend to disfavor the EDGES-inspired temperature ratio. However, we caution against over-interpreting this result, as it is highly model-prior-dependent. As mentioned previously in the context of Figure 17, almost all of the constraining power for the `21cmFAST`- and `21cmSPACE`-based models comes from H23 at $z = 8$ and 10. It is the smooth evolution of these models that yields the constraints at higher redshifts. Conversely, the `Zeus21`-based model (purple line) is *only* exposed to data at $z > 10$, and is clearly far less constrained. The black triangles in Figure 18 illustrate a less model-dependent constraint. These are lower limits obtained directly from the upper limits on Δ^2_{21} via a phenomenological argument that relates the 21 cm power to the underlying density field through a linear bias (see H23; here, we conservatively assume observations at $\mu = 1$). Since each density-driven lower limit is obtained directly from an upper limit at that redshift, it provides some measure of how far our upper limits are from providing direct model-independent constraints on excess cooling/radio background during Cosmic Dawn. We find that we require an order-of-magnitude improvement[77] at $z = 17.5$ to yield such a constraint on EDGES-motivated models.[78]

[76] In this plot, we show only the four models that have mechanisms to produce excess cooling or excess radio background, leaving out the 'standard' $T_S = T_R$ model.

[77] Precisely the improvement we might naively expect from processing the full season of data; see Table 2.

[78] "EDGES-motivated" here refers to any physical model incorporating excess cooling and/or radio backgrounds during Cosmic Dawn, and does not imply that these models are in fact *favored* by the EDGES data (e.g., J. Cang et al. 2024), only that their development was motivated by the enhanced absorption reported in J. D. Bowman et al. (2018).

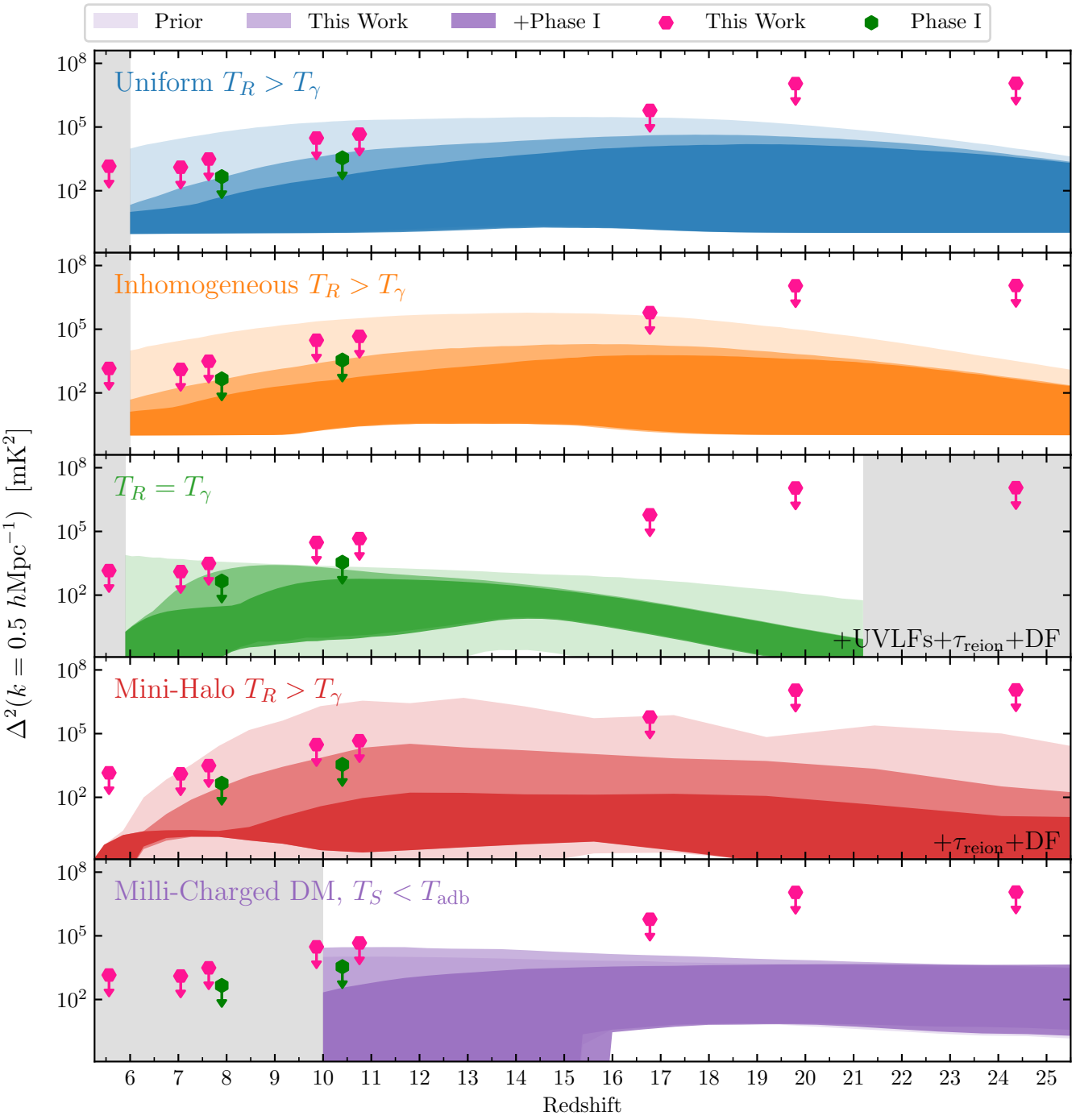

Figure 17. Shown here are 95% confidence intervals on $\Delta^2(z)$ at $k = 0.5\ h$ Mpc^{-1}. Each panel displays a different theoretical model, and different shading levels represent different combinations of data: the lightest is the model prior, the next layer includes the limits from this work (shown as pink hexagons), and the darkest layer combines these limits with those from H23 (shown as green hexagons). The `21cmFAST`-based model posteriors also include other data, as listed in the bottom-right of each panel and described in-text. The limits indicated as hexagons are the lowest limit at any k for each particular redshift (generally $k \sim 0.5\ h$ Mpc^{-1}). Not all models are defined at all redshifts where limits were placed: both `21cmSPACE` models (blue and orange) and the `21cmFAST` $T_R \equiv T_\gamma$ model only extend down to $z = 6$—an artifact of the emulator training for each model. Conversely, the `Zeus21` model (purple) only extends down to $z = 10$, as it does not yet include realistic reionization. The limits of the predictions of each model are demarcated by gray shaded regions. The upper limits, as expected, do not constrain the lower edge of the model priors, although in the case of `21cmFAST`, the lower edges are constrained by the non-21 cm data. In all cases, the low-redshift limits dominate the constraints at all redshifts, which can be seen most clearly in the `Zeus21` model, whose constraints are weakest since they do not use any low-z data.

7. Conclusions

In this work, we have presented the first science results from Phase II of the HERA telescope. Comprising just 2 weeks of observations, this analysis is aimed at understanding the unique new characteristics of the Phase II system, and developing a robust analysis pipeline that accounts for the spectral and temporal features in the data. With increased frequency coverage of 47–234 MHz, this is the largest range of redshifts simultaneously examined in this field, including Cosmic Dawn ($z > 20$) and the tail end of reionization ($z < 6$).

Our analysis pipeline, composed primarily of reproducible and scalable Jupyter notebooks, first calibrates each integration

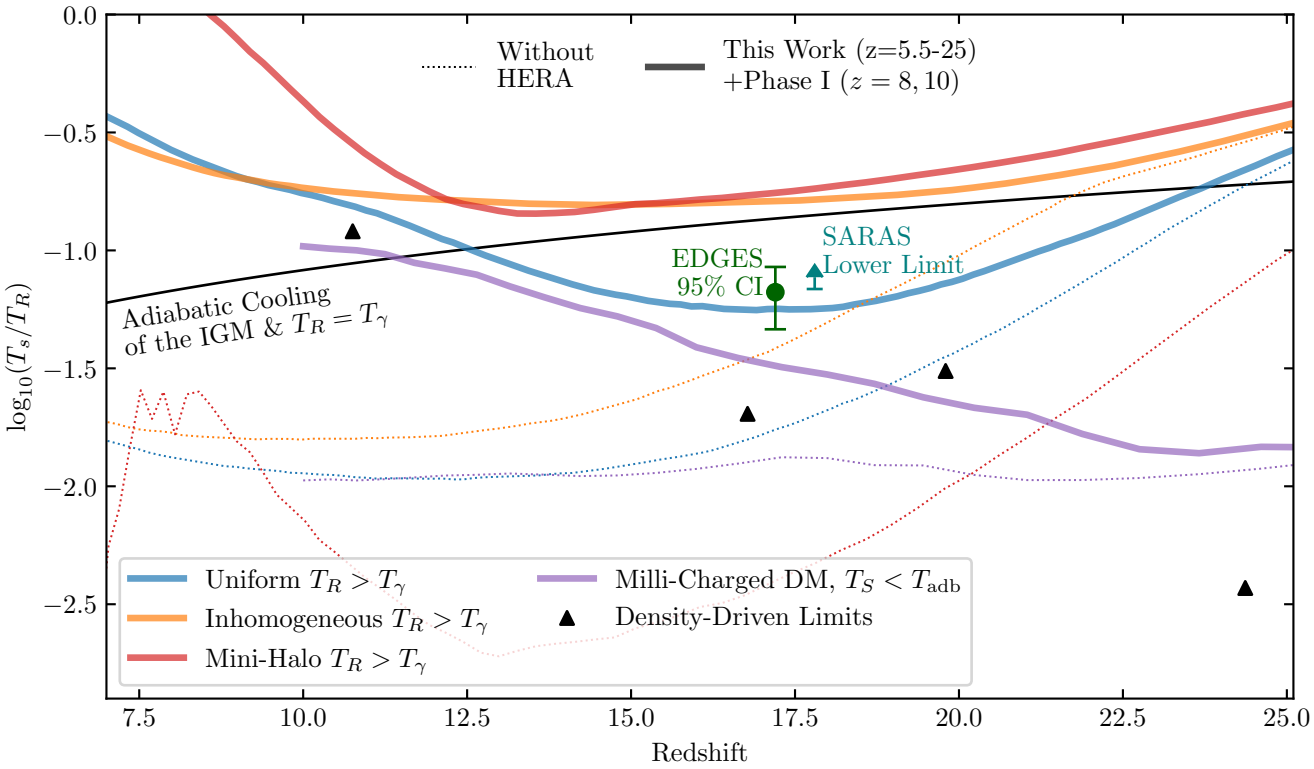

Figure 18. Shown here are inferred 95% lower limits on the ratio T_S/T_R from several theoretical models. In this figure, different colors represent different theoretical models, all of which include some mechanism for producing T_S/T_R lower than the "standard" adiabatic cooling limit with a uniform CMB radio background (black solid line). Dotted lines indicate the prior region of each model, while solid lines indicate the constraints from the combination of this data (from $z = 5.5$ to 25) with the limits from H23 (at $z = 8$, 10). Black triangles indicate lower limits derived from a more model-agnostic approach, in which 21 cm fluctuations before the onset of reionization are considered to be driven by density fluctuations (see H23). These correspond to 2σ lower limits derived directly from the data. Also shown is the inferred 95% confidence interval on T_s/T_R derived from the 21 cm absorption depth reported in J. D. Bowman et al. (2018) and, similarly, the lower limit derived in the same way from the limits on the absorption depth presented in S. Singh et al. (2022). In general, the models that include excess radio backgrounds—whose priors allow for the EDGES-derived ratio—tend to disfavor such small ratios once exposed to the HERA limits. These results are highly model-dependent, since the constraints for the models are almost entirely derived from lower-redshift measurements. The most model-agnostic limits—those from the density-driven model—are still entirely consistent with excess cooling/radio background, and will require an order-of-magnitude reduction if they are to begin to rule out the EDGES-derived ratio.

independently, then performs several quality checks based on metrics compiled over a full night's observations, removing data plagued by RFI, lightning, packet loss, low correlation, and other systematics. The calibration solutions are then smoothed over long temporal and medium spectral scales, taking advantage of our prior that the intrinsic instrument has a smooth response. Once the calibration is completed, and the most egregious defects have been flagged, we start averaging the data: first we average visibilities within nominally redundant baseline groups, and then we average visibilities at the same LST on different nights. The latter requires us to "inpaint" the gaps in the data incurred by frequency-dependent flagging (generally for RFI), which we do on a per-night, -baseline-group, -polarization, and -LST basis. We then attempt to mitigate pervasive systematics such as MC using filters applied in fringe-rate and delay space, before computing delay spectra for each baseline independently. Finally, we average these spectra (incoherently) over LST, orientation (cylindrical average), and baseline group (spherical average) to yield power spectra in a set of eight redshift bands ($z = 24.4$, 19.8, 6.8, 10.8, 9.9, 7.6, 7.0, and 5.6). For details on these analysis steps, see Section 3.

Our analysis pipeline was validated using large end-to-end mock simulations, in a similar fashion to H22a and H23, using the framework laid out in J. E. Aguirre et al. (2022) and described in Section 4. This validation procedure did not uncover any unexpected biases in the estimated 21 cm power spectrum. The input spectrum was artificially boosted in the simulations to enable a "detection" at scales of $k \sim 0.5$–$1.0\,h$ Mpc^{-1}, and the mock detection was consistent with the input across all redshifts.

We reported upper limits on the spherically averaged power spectrum, for which the sensitivity is similar to our deepest Phase I limits (see Figure 16 and Table 5), using only 14 nights of data compared to the 94 used in H23. This increase in nightly sensitivity is attributable to the increased number of active antennas in this dataset: ∼130 versus ∼40 in Phase I. For $k \gtrsim 0.5\,h$ Mpc^{-1}, our estimated power spectra are consistent with the predicted noise variance, derived from the autocorrelations and accumulated number of samples (see Equation (30)), at the 2σ level. However, on larger scales, we find that our estimates are foreground/systematics-limited. This foreground leakage is more extensive than was found in Phase I, such that even though the overall sensitivity of this data is similar to Phase I, the upper limits are less constraining since the lowest, most constraining, k-modes—which were noise-limited in Phase I—are here contaminated by foregrounds.

Nevertheless, these limits are competitive in two primary ways: they are the first to simultaneously cover such a large redshift range, and they are the most constraining at high redshifts ($z > 16$) (see Figure 17). While they do not yet yield further insight into the physics of Cosmic Dawn and the EoR beyond what is known from previous 21 cm power spectrum

upper limits, they confirm the accuracy and efficiency of HERA's analysis pipeline, which is ready to process an order-of-magnitude more data that has already been taken.

7.1. Mutual Coupling

Our primary extant systematic at $k \lesssim 0.7\,h$ Mpc^{-1}, and the cause of the foreground leakage that precludes this dataset from setting limits comparable to H23, is MC (see Section 4.2). This systematic, which occurs when radiation is reflected or re-emitted from one element of the array before being received into another antenna, is enhanced in Phase II with respect to Phase I due to the larger vertical cross section of the updated Vivaldi feed, the increased height of the feed within the dish, and the removal of cylindrical mesh guards around the feeds (see, e.g., L. M. Berkhout et al. 2024). Taken together, these effects result in an increased interaction between neighboring antennas and an enhancement of the observed MC. Though it is difficult to measure the precise extent to which the effect is enhanced in Phase II, Figure 13 suggests via approximate models that it now dominates modes up to $k \sim 0.7\,h$ Mpc^{-1} that were previously noise-dominated. This is a $\sim$1% effect at 500 ns scales, which is beyond the precision required for 21 cm science, and should be considered carefully by planned experiments with densely packed antennas.

The impact of MC arises from delayed signals coupling back into a baseline's visibility. Naturally, these delays are super-horizon scale in the cylindrical power spectrum, resulting in foreground leakage into the 21 cm window. While a first-order analytic model of this effect (A. T. Josaitis et al. 2021; R&P25) has been successful in describing the general characteristics of the systematic, it is not accurate enough to enable us to invert and remove it. Instead, we have used the insight gained from our analytic model to partially localize the leaked power in fringe-rate/delay space, and apply a simple filter in this space to remove the bulk of the systematic power. This mitigation technique is effective, but is limited in the amount of leaked power it can remove without significant loss of cosmic 21 cm power, because both MC and 21 cm signal exist within the bounds of the filter.

Given the approximate nature of the first-order coupling model, the relative amplitude of MC to cosmic signal power as a function of scale is highly uncertain. Figure 10 of R&P25 suggests a smooth decline until $k \sim 0.9\,h$ Mpc^{-1} before a steeper roll-off, although this roll-off corresponds to the light-crossing time of the array, which is the maximum delay that the *first-order* coupling model predicts, and higher-order terms may extend this range. While we have demonstrated that MC is dominant below $k \sim 0.7\,h$ Mpc^{-1}, its behavior on smaller scales—how quickly it tapers off with increasing k—remains unobserved. More sensitive observations will be able to reveal this behavior in greater detail.

A key takeaway from this work is that more detailed models of MC may yield the greatest improvement to HERA's sensitivity, even as more data is averaged. The current first-order linear model is clearly insufficient for the purposes of inverting the systematic from the data; beyond the possible importance of higher-order terms (i.e., re-reflections and the correlation of two separate reflections), it is likely that reflections from elements other than feeds may contribute to high delay structure in the visibilities (N. Fagnoni et al. 2021b). While stability of the coupling over nights is broadly expected, it has not been observationally established beyond the resilience of the systematic to data averaging on medium-to-large scales. Conversely, differences of the MC signal between otherwise redundant baselines might be leveraged to help disentangle the effect. We continue to explore these issues with embedded element antenna simulations, as well as extended analytical modeling.

Happily, our results indicate that the amplitude of foreground leakage from MC rapidly decays with delay, and so we can expect a clean signal (clean from MC, at least) at some scale, given enough thermal sensitivity. At lower delays, approximate mitigation or inversion techniques will temporarily buy sensitivity for tighter upper limits, but will need to be significantly more accurate to allow detections. Simulations of ever-increasing realism will play a crucial role in validating these models and algorithms.

7.2. Other Key Issues for Future Investigation

Our second-most prominent residual systematic is that of night-to-night gain variations that, when combined with nonuniform flagging patterns, result in spectral discontinuities in the average over nights. Inconsistency over nights can be measured as a ratio of the variance over nights (for a particular unique baseline group) to the expected variance derived from autocorrelations (see Equation (1)), a metric directly related to the mean Z^2-score over the nights (S. G. Murray 2024). We generally measure such an "excess variance" of $\sim$2, where no excess would yield unity. This inconsistency is generally spectrally smooth, evidenced by the fact that the high-delay spectra are consistent with the predicted noise, and so is reasonably associated with errors in the low-delay modes of the gain solutions that fluctuate night-to-night. While such fluctuations are not intrinsically harmful, as long as they remain in foreground-dominated delays, they can be leaked to higher-delay modes when averaging with nonuniform weights. We see evidence for this when looking at the more highly flagged spectral windows that were omitted from this analysis. In these bands, the high-delay spectra consistently have a visibly different distribution than predicted based on the autocorrelations. The correlation of this effect with the higher level of flagging in those bands suggests that the combination of night-to-night inconsistencies and imperfect inpainting is its root cause. We are investigating more sophisticated inpainting techniques to address this in the future.

Another potential issue that is closely related to the inpainting issues we have just described is that while we take great care to maintain spectral smoothness by inpainting over frequency with smooth models, we do not take the same care to maintain temporal smoothness. This is, admittedly, a secondary concern, since spectral structure is the primary distinguishable characteristic of the 21 cm signal. However, we utilize temporal structures to enable systematics mitigation through the use of fringe-rate filtering, which ultimately feeds back into spectral structure. Future more sophisticated inpainting techniques are likely to improve this issue as well.

While our DPSS-filtering-based technique for identifying low-level RFI in redundantly averaged cross-correlations is a marked improvement over previous techniques, there are at least two possible avenues for improvement. First, some moderately broadband RFI (like from TV stations that broadcast in 8 MHz allocations) may be degenerate with the filter, making it harder to find low-level RFI. Second, the

technique does not incorporate any spatial information about the source of the emitter. Instead of an incoherent average over redundant baseline groups, one could use prior information about the locations of known transmitters to coherently combine visibilities for maximum SNR. This can be done both for fixed transmitters like radio and TV stations, as well as for ones that move in predictable ways, like ORBCOMM (A. R. Neben et al. 2016) or Starlink (F. D. Vruno et al. 2023; C. G. Bassa et al. 2024) satellites.

Another systematic effect that requires further consideration is the leakage of a small number of highly polarized, high-rotation measure (RM) point sources into our pseudo-Stokes I visibilities. To estimate the unpolarized sky signal, we take a sum of the visibility response from the two orthogonal linear feed polarizations. However, due to asymmetries in the polarized beam responses, this operation introduces leakage from Stokes Q into the pseudo-Stokes I visibilities. In the presence of bright polarized sources with large RMs, this leakage imprints spectral structure that varies approximately as $\exp(2i\mathrm{RM}\lambda^2)$, contaminating high $k_\parallel$ modes and potentially biasing power spectrum estimates (D. F. Moore et al. 2017). Fortunately, this contamination is temporally localized to epochs when such sources enter the primary beam, enabling partial mitigation through careful field selection. However, this imposes constraints on the set of fields that are viable for power spectrum estimation. Catalogs such as those produced by the NRAO VLA Sky Survey (A. R. Taylor et al. 2009) and the POlarised GLEAM Survey (C. J. Riseley et al. 2020) have demonstrated that, although most sources are weakly polarized, a nonnegligible population exhibits polarization fractions and RMs sufficient to pose a risk for spectral leakage. These catalogs support identification and masking efforts, but complete mitigation will likely require forward modeling of polarized emission, incorporating both the polarized beam response and RM synthesis. For this work, we have simply avoided observations where known, bright, high-RM pulsars are in the field of view (see Figure 2). A detailed treatment of this effect is left to future work.

7.3. Future Outlook

This analysis was aimed at evaluating the updated analysis pipeline as applied to HERA Phase II data. As such, we considered only a small subset of the data taken with Phase II, starting in 2022 and still ongoing. As described in Table 2, the full 2022–2023 observing season alone contains an order of magnitude more data than presented here, and as of 2025 July we have completed an additional two full seasons of observing with even more active antennas. While our lowest k-modes are currently systematics-limited, we expect that the sensitivity of 3 full years of data (with similar levels of flagging as those encountered in this dataset[79]) at $k \sim 0.8\,h\ \mathrm{Mpc}^{-1}$—which is currently noise-limited in all bands—will be at least 700 times more sensitive (i.e., $\Delta^2_{\mathrm{UL}} \approx 2.5\ \mathrm{mK}^2$ at $z = 7$): sufficient to make detections for realistic cosmological scenarios. Thus, while further mitigation of MC remains a top priority, it is not necessarily required for an eventual detection, given the enormous raw sensitivity of the HERA array.

[79] While other instruments have reported an increase in RFI from satellite constellations such as Starlink, it is uncertain whether these will make a significant impact on our future data, as we do not currently localize our sources of RFI.

In summary, in this work, we have demonstrated that HERA Phase II is already able to achieve noise-limited limits over multiple k-bins (albeit at higher k than in Phase I) and a broad range in redshift. With roughly an order-of-magnitude more data already taken and new techniques for systematics mitigation soon to be validated and implemented, we expect continued progress in the near future toward a first detection of the highly redshifted 21 cm power spectrum.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant Nos. 1636646, 1836019, 1352519, and 1407804 as well as institutional support from the HERA collaboration partners. This research is funded in part by the Gordon and Betty Moore Foundation through grant GBMF5212 to the Massachusetts Institute of Technology. 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.

This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation PHY201142 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grant Nos. 2138259, 2138286, 2138307, 2137603, and 2138296. We acknowledge the use of the Ilifu cloud computing facility (www.ilifu.ac.za) and the support from the Inter-University Institute for Data Intensive Astronomy (IDIA; https://www.idia.ac.za).

S.G.M. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101067043. This result is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 948764; P.B., J.B., M.J.W.). B.B.B. acknowledges the funding received from FAPESP under process 2024/12902-3. N.S.K. acknowledges support from NASA through the NASA Hubble Fellowship grant No. HST-HF2-51533.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. J.M. was supported by an appointment to the NASA Postdoctoral Program at the Jet Propulsion Laboratory/California Institute of Technology, administered by Oak Ridge Associated Universities under contract with NASA. A.M. acknowledges support from the Italian Ministry of Universities and Research (MUR) through the PRIN project "Optimal inference from radio images of the epoch of reionization," and the PNRR project "Centro Nazionale di Ricerca in High Performance Computing, Big Data e Quantum Computing." A.C.L. acknowledges support from the Natural Sciences and Engineering Research Council of Canada through their Discovery Grants Program, as well as the William Dawson Scholar program at McGill University. R.P. acknowledges support from the Faculty of Arts & Science at the University of Toronto and the Dunlap Institute. The Dunlap Institute is funded through an endowment established by the David Dunlap family and the University of Toronto.

Facility: HERA.

Software: `numpy` (C. R. Harris et al. 2020), `scipy` (P. Virtanen et al. 2020), `matplotlib` (J. D. Hunter 2007), `astropy` (Astropy Collaboration et al. 2013, 2018), `jupyter`

(T. Kluyver et al. 2016), `pyuvdata` (B. J. Hazelton et al. 2017; G. Keating et al. 2025), `fftvis` (T. A. Cox et al. 2025). `hera_opm` (P. L. Plante et al. 2021), `hera_cal` (https://github.com/hera-team/hera_cal, `hera_pspec` (https://github.com/hera-team/hera_pspec

Appendix A
Antenna Metrics

We flag entire antennas with the following specific characteristics, irrespective of channel-based flags:

1. *Dead*. Antennas with visibilities whose value is zero more than half of the time are considered "dead" and flagged.
2. *Low Correlation.* Antennas with low correlation coefficients (from 0–0.2) are flagged (D. Storer et al. 2022). We found that such low correlations are often indicative of clock distribution issues that affect entire nodes (L. M. Berkhout et al. 2024).
3. *Cross-polarized*. Antennas whose correlation coefficient with other antennas is larger when correlating the other antenna is putatively of the opposite polarization are considered cross-polarized (D. Storer et al. 2022) and flagged (and reported for in-field maintenance).
4. *Packet Loss*. While our integration time for snapshot observations is $\sim$10 s, we also keep "diff" files containing even–odd differences on a much finer timescale. These can be rearranged into even and odd samples. An excess of zeros in either the evens or the odds is interpreted as a sign of packet loss. We flag antennas with more than eight zeros in either the odds or evens.[77]
5. *Anomalous Autocorrelation Power*. Antennas with a median amplitude over all channels for each autocorrelation more than 60 times the integration count (i.e., the integration time by the channel width, $\tau_{\rm int}\Delta\nu$) are flagged. These are at risk of a nonlinear ADC response. Likewise, antennas with median amplitude below 1 times the integration count are flagged for low sky-response.
6. *Anomalous Autocorrelation Slope*. Antennas whose relative absolute slopes (i.e., the absolute value of the ratio of the linear coefficient to the constant coefficient in a least-squares fit to the median-filtered spectrum) are more than 0.6 are flagged. This is generally a sign of low or unusual sky-response.

Appendix B
Iterative Algorithm for Antenna and RFI flags

In this appendix, we describe in more detail the algorithm used in the per-integration flagging and calibration (see Section 3.3) to iteratively refine the per-antenna and per-channel flags in tandem.

We first determine an initial RFI mask (per-channel flags) as follows. We subtract neighboring channels of the autocorrelations for each antenna (resulting in spectra close to white noise, if no RFI were present), and then we identify 20σ outliers, where σ is estimated via Equation (1) (using each autocorrelation to estimate its own noise). Crucially, the final product is only channel-dependent, not antenna-dependent: channels that are flagged for more than 50% of antennas are then flagged for *all* antennas, and otherwise, the channels are left *unflagged*.

Following this, a linear DPSS model is fit to each autocorrelation, respecting the flags found from the first step. Here, we use the parameters $\tau_c = 0$, $\Delta\tau = 300$ ns, $\lambda_{\rm min} = 10^{-9}$. After subtraction of this model, the resulting residuals should be white noise if no RFI were present, and we perform a more aggressive cut at 4σ (with σ computed in the same way as above). Again, only channels with a flag fraction of more than 50% across antennas are ultimately flagged.

This rough RFI mask is used to identify cross-correlation engines (X-engines) that are systematically aberrant. Each X-engine is responsible for 96 channels. For each set of 96 channels, we take finely time-differenced data, which should be noise-like, and compute a quasi-Z-score based on its amplitude: $(|V_{\rm diff}| - \sigma\sqrt{(\pi/4)})/(\sigma\sqrt{(4-\pi)/4})$, where σ is the predicted standard deviation of the thermal noise of the observation, based on the autocorrelations. This quantity is expected to have mean zero and variance unity (though it is not Gaussian distributed), and we compute the mean of its absolute value within each set of 96 channels controlled by a particular X-engine. This mean, $|\overline{Z}|$ should be half-normally distributed, with mean $\sqrt{2}/\pi$ and variance $(1–2/\pi)/N_{\rm samples}$, where $N_{\rm samples} \sim 96$ is the number of channels in the 96-channel chunk that are *not* flagged by the previously calculated RFI mask. We re-scale $|\overline{Z}|$ to have mean zero and variance unity before counting the number of times a particular antenna has an X-engine with a rescaled mean Z-score above 10. Then, starting with the antennas that have the highest count of bad X-engines, we iteratively remove them, until we are left with no baselines with bad X-engines.

Our final two antenna-based cuts are established in conjunction with our final spectral mask—which refines our initial rough RFI mask—in a single iterative loop. This loop is as follows, and runs for *at least* three iterations, and until no new channels are flagged and no new antennas are flagged. Each operation is performed only on autocorrelations. Initially, the shape-N_ν vector of RFI flags $\xi_\nu \in (0, 1)$ are those obtained from the "rough" RFI mask described above.

1. Calculate $Z_{ii} = (P_{ii} - M)_{ii}/(\sqrt{2}M_{ii}/\sqrt{\tau\Delta\nu})$ for each autocorrelation power P_{ii}, where M_{ii} is a DPSS model of the power over frequency, with $\tau_c = 0$, $\Delta\tau = 300$ ns, $\lambda_{\rm min} = 10^{-9}$.
2. Compute the rms, $\sigma_{Z_{ii}} = \sqrt{\sum_\nu \xi_\nu Z_{ii}^2/N_\xi}$ for each auto-power, where N_ξ is the number of unflagged channels, $\sum_\nu \xi_\nu$ (the same for all antennas).
3. Choose which autocorrelations to use for estimating the RFI, starting from a candidate pool of all those that have not yet been flagged (either by previous checks, or previous iterations of this loop), then:
   a. If this is the *first* iteration: use the half of the candidate pool that has the lowest rms.
   b. If this is the *second* iteration: use all candidate antennas whose RFI classification is "good" (see below) unless this is less than half of the candidate pool, in which case use the half of the candidate pool that has the lowest rms.

[77] Note that while packet loss occurs on a per-baseline basis, we find that it generally affects a few antennas disproportionately because of the ordering of the visibility information. While maintaining per-antenna flags here does result in over-flagging, the utility of carrying forward per-antenna instead of per-baseline flags at this stage is considered worth the small excess data loss.

c. If this is the *third* or higher iteration: use all candidate antennas whose RFI classification is not "bad" (see below).

4. New channel-based flags, ξ_ν, are computed exactly in the same way as the "rough" RFI flags above, except that instead of *each* autocorrelation being tested (and then the final flags being combined via thresholding over all antennas), only the mean autocorrelation over the antennas in the set selected above are used.
5. These flags are used to identify outliers in *spectral shape*. Here, we re-scale each RFI-flagged autocorrelation to be mean unity, and then find the mean rescaled autocorrelation across all antennas (per polarization) that have not thus far been flagged. We flag antennas whose rms difference with respect to the mean is >20%.

Appendix C
Temporally Smoothed Metric Flagging

In Section 3.4.1 we outlined an algorithm for temporally smoothing quality metrics, and re-flagging based on these smoothed metrics. Here we describe the algorithm in more detail. The algorithm is as follows:

1. Begin with a per-antenna time-series of a particular metric, m_t, and associated flags ξ_t when m_t is outside some threshold ($M_{\rm low}$, $M_{\rm high}$).
2. Convolve the metric series with a Gaussian of width $\sigma = 60$ integrations (about 10 minutes).
3. Iteratively:
   a. Partition the time-series into contiguous regions that are either completely flagged or completely unflagged, and further categorize the flagged regions as those in which the convolved metric either surpasses ("persistent") the threshold at least once or not ("nonpersistent").
   b. Re-partition the time-series such that both unflagged and nonpersistent flagged regions from the previous partitioning are merged together (type A), and persistent flagged regions remain separate (type B).
   c. For each region of type A, if the smoothed metric is outside the threshold *everywhere* in the region, flag the entire region. This is meant to flag small unflagged regions embedded in larger flagged regions, as well as regions with a high time-to-time variability of flags.
   d. Finally, if there is any flagged region of size >60 integrations, flag all times before or after the gap, whichever is smaller.
   e. If there are new flags in this iteration, keep iterating.

This procedure is separately performed for each of the autocorrelation power, shape, and slope metrics, as well as the RFI rms (see Section 3.3.1) and the χ^2 computed from redundant calibration. For the latter, the initial flags are set to *ignore* times flagged for any other reason (whereas other metrics are considered in isolation).

After separately performing this procedure for each metric, and obtaining a final set of unified flags (where any of the metrics exceeds the threshold), we once more flag regions either before or after flag gaps of more than 60 integrations (whichever is smaller). Finally, antennas that are flagged for more than 50% of the integrations on a given night are flagged entirely.

Appendix D
Iterative Algorithm for Deeper RFI Excision

In Section 3.6 we describe our method for constructing an RFI mask based on baseline-averaged per-night Z-scores derived from a high-pass delay filter. Here we give the details of the precise steps and thresholds of this procedure. We flag the following (in this order): (1) Iteratively flag the worst offending integrations and channels, in exactly the same way as the iterative procedure outlined in Section 3.4.2, but with a threshold of 1.0 instead of 1.5, and using the median over each axis instead of the mean; (2) any channels that are flagged for more than 25% of integrations; (3) any integrations that are flagged for more than 10% of channels; (4) any particular integration-channel with $\bar{Z} > 4$; (5) any integration-channel neighboring a flagged channel with $\bar{Z} > 2$; (6) repeat step (1) using the mean instead of median; (7) repeat steps (2) and (3).

Appendix E
Justification of Residual Bin-to-bin Correlations

We claimed in Section 6.2 that allowing 10% correlations between neighboring k-bins does not significantly affect the inferred posterior. Here we detail our justification for this claim.

Formally, our prior on the systematic is improper. To carry out this investigation, we instead assumed it is uniformly distributed from 0 to 100,000 mK^2, far in excess of the noise levels used in this numerical experiment (and therefore well-approximating the improper prior assumed in the likelihood). We then simulated samples from the systematic prior and different Gaussian noise distributions for two fictitious "neighboring" power spectrum bins with varying correlation coefficient: 0, 0.1, and 0.99, and standard deviation 1000 mK^2 (chosen because they are simple, round numbers at approximately the sensitivity of the $z < 8$ limits in Table 5). In each case, we used the large number of simulated draws to form a kernel density estimate of the marginal density of these contributions to the pair of power spectrum bins. Then, for several fictitious measurements and broad uniform priors from 10–1000 mk^2 on the signal power spectrum in the two bins, we normalized the density estimates so that they were properly normalized posterior probability distributions.

In summary, we derived marginal posterior distributions for the "21 cm signal" (parameterized simply as the power spectrum amplitude in two neighboring k-bins) for mock data generated with different levels of true correlation between the neighboring k-bins, but for which the likelihood model assumed no correlation. We obtained such posteriors in a range of scenarios: systematics-dominated, noise-dominated, and signal-dominated.

We found that, regardless of whether the measurements were noise-limited (e.g., 1000 mK^2), or if one (or both) were strongly systematically contaminated (e.g., 10000 mK^2), the three resulting posteriors (for the different true correlation levels) were all nearly identical—even for the case where the noise correlation between the two bins was 0.99. Variations in posterior probability density were at most ∼2%, where the strongest deviations occurred in the noise-limited case with correlation coefficient equal to 0.99. For a correlation coefficient of 0.1, deviations were subpercent in all cases. This means that posterior inferences would be nearly the same regardless of whether we fully modeled the noise correlations

in our likelihood. We believe that this occurs because of the strong uncertainty (and lack of correlation) that is reflected by our model for systematic effects, which is not particularly realistic. We suspect that more realistic treatments of the systematic effect (particularly if they do not range by many orders of magnitude and are not totally uncorrelated) may create very different posteriors in the presence of correlated noise, and would therefore demand more careful noise modeling in the astrophysical inference step.

Appendix F
Table of Power Spectrum Upper Limits

The following table contains the limits presented in Figure 16.

Table 5
The Full Set of Spherically-Averaged Upper Limits in All Redshift Bands From This Work, As Displayed in Figure 16

	k (h Mpc^{-1})	$\Delta^2(k)$ (mK^2)	1σ (mK^2)	Δ^2_{UL} (mK^2)	k (h Mpc^{-1})	$\Delta^2(k)$ (mK^2)	1σ (mK^2)	Δ^2_{UL} (mK^2)
$z = 24.37$	0.30	115,193,441	2,024,176	119,241,794	0.85	26,288,139	14,288,749	54,865,638
...	0.41	14,545,056	1,790,638	18,126,332	0.96	31,125,314	20,641,861	72,409,037
...	0.52	4,975,043	3,255,384	**11,485,811**	1.07	101,151,265	28,620,389	158,392,044
...	0.63	10,890,061	5,868,207	22,626,477	1.18	−22,797,735	36,778,435	73,556,871
...	0.74	19,442,615	9,617,651	38,677,919	1.29	9,617,674	48,273,574	106,164,823
$z = 19.80$	0.25	76,689,186	1,543,420	79,776,026	0.98	148,711	14,129,685	28,408,081
...	0.40	13,100,453	1,126,258	15,352,970	1.12	37,439,608	21,798,781	81,037,171
...	0.54	6,144,314	2,491,275	**11,126,866**	1.26	6,369,229	31,762,166	69,893,562
...	0.69	9,482,752	5,083,277	19,649,308	1.41	82,269,409	43,227,317	168,724,043
...	0.83	26,944,303	9,078,540	45,101,385	⋯	⋯	⋯	⋯
$z = 16.78$	0.25	18,392,322	429,193	19,250,710	0.98	−104,028	2,557,758	5,115,517
...	0.40	1,408,317	190,982	1,790,281	1.13	−628,410	3,880,214	7,760,429
...	0.55	252,112	439,236	**1,130,584**	1.27	14,663,521	5,668,569	26,000,660
...	0.69	−1,146,024	875,715	1,751,430	1.42	−1,565,451	7,639,973	15,279,947
...	0.84	1,759,759	1,588,757	4,937,274	1.57	13,252,746	10,600,422	34,453,590
$z = 10.76$	0.48	164,176	4,589	173,355	1.31	28,485	61,975	152,436
...	0.76	21,703	12,291	**46,285**	1.58	128,749	109,418	347,587
...	1.03	21,326	30,541	82,409	1.86	129,559	180,625	490,811
$z = 9.87$	0.33	495,706	9428	514,562	1.26	296	39,414	79,125
...	0.51	24,337	2995	**30,328**	1.44	−68,169	58,410	116,821
...	0.70	21,512	6926	35,364	1.63	176,688	87,697	352,083
...	0.89	3656	13,695	31,048	1.82	31,596	119,782	271,160
...	1.07	−7101	24,293	48,586	2.00	104,834	159,778	424,392
$z = 7.63$	0.37	16,563	811	18,187	1.45	4208	9880	23,969
...	0.59	1734	689	**3113**	1.66	17,483	14,816	47,116
...	0.80	1421	1677	4776	1.87	34,627	21,188	77,004
...	1.02	11,316	3487	18,291	2.09	−13,916	28,712	57,424
...	1.23	6891	5942	18,777	⋯	⋯	⋯	⋯
$z = 7.05$	0.44	2768	372	3513	1.46	2703	8155	19,014
...	0.70	−494	887	**1775**	1.71	20,386	13,505	47,398
...	0.95	921	2296	5514	1.97	12,768	20,132	53,032
...	1.21	490	4604	9699	2.22	−3,748	28,476	56,952
$z = 5.56$	0.55	1695	181	2058	1.81	−2295	4434	8,869
...	0.87	435	496	**1428**	2.13	3648	7218	18,086
...	1.18	1894	1269	4433	2.44	−19,753	10,793	21,586
...	1.50	322	2536	5394	⋯	⋯	⋯	...

Note. Columns are the wavenumber k, the estimated power spectrum, the 1-sigma estimated thermal noise error, and the 2-sigma upper limit. The lowest absolute limit at each redshift is shown in bold font.

ORCID iDs

James E. Aguirre https://orcid.org/0000-0002-4810-666X
Rennan Barkana https://orcid.org/0000-0002-1557-693X
Lindsay M. Berkhout https://orcid.org/0000-0002-2293-9639
Gianni Bernardi https://orcid.org/0000-0002-0916-7443
Bruno B. Bizarria https://orcid.org/0000-0001-7794-6599
Judd D. Bowman https://orcid.org/0000-0002-8475-2036
Daniela Breitman https://orcid.org/0000-0002-2349-3341
Philip Bull https://orcid.org/0000-0001-5668-3101
Jacob Burba https://orcid.org/0000-0002-8465-9341
Ruby Byrne https://orcid.org/0000-0003-4980-2736
Rajorshi Sushovan Chandra https://orcid.org/0009-0002-6260-0055
Kai-Feng Chen https://orcid.org/0000-0002-3839-0230
Samir Choudhuri https://orcid.org/0000-0002-2338-935X
Tyler Cox https://orcid.org/0009-0008-2574-3878
David R. DeBoer https://orcid.org/0000-0003-3197-2294
Eloy de Lera Acedo https://orcid.org/0000-0001-8530-6989
Jiten Dhandha https://orcid.org/0000-0002-1481-0907
Joshua S. Dillon https://orcid.org/0000-0003-3336-9958
Aaron Ewall-Wice https://orcid.org/0000-0002-0086-7363
Anastasia Fialkov https://orcid.org/0000-0002-1369-633X
Steven R. Furlanetto https://orcid.org/0000-0002-0658-1243
Hugh Garsden https://orcid.org/0009-0001-3949-9342
Adelie Gorce https://orcid.org/0000-0002-1712-737X
Deepthi Gorthi https://orcid.org/0000-0002-0829-167X
Bryna J. Hazelton https://orcid.org/0000-0001-7532-645X
Jacqueline N. Hewitt https://orcid.org/0000-0002-4117-570X
Jack Hickish https://orcid.org/0000-0003-0216-1417
Daniel C. Jacobs https://orcid.org/0000-0002-0917-2269
Alec Josaitis https://orcid.org/0000-0002-4118-6695
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
Matthew Kolopanis https://orcid.org/0000-0002-2950-2974
Adam Lanman https://orcid.org/0000-0003-2116-3573
Paul La Plante https://orcid.org/0000-0002-4693-0102
Adrian Liu https://orcid.org/0000-0001-6876-0928
Yin-Zhe Ma https://orcid.org/0000-0001-8108-0986
David H. E. MacMahon https://orcid.org/0000-0001-6950-5072
Lisa McBride https://orcid.org/0009-0008-7585-9385
Andrei Mesinger https://orcid.org/0000-0003-3374-1772
Jordan Mirocha https://orcid.org/0000-0002-8802-5581
Miguel F. Morales https://orcid.org/0000-0001-7694-4030
Julian B. Muñoz https://orcid.org/0000-0002-8984-0465
Steven G. Murray https://orcid.org/0000-0003-3059-3823
Aaron R. Parsons https://orcid.org/0000-0002-5400-8097
Robert Pascua https://orcid.org/0000-0003-0073-5528
Nipanjana Patra https://orcid.org/0000-0002-9457-1941
Simon Pochinda https://orcid.org/0009-0005-8660-0713
Yuxiang Qin https://orcid.org/0000-0002-4314-1810
Eleanor Rath https://orcid.org/0009-0008-7886-2766
Mario G. Santos https://orcid.org/0000-0003-3892-3073
Saurabh Singh https://orcid.org/0000-0001-7755-902X
Dara Storer https://orcid.org/0000-0003-4092-0103
Jianrong Tan https://orcid.org/0000-0001-6161-7037
Emilie Thélie https://orcid.org/0000-0001-8838-1394
Michael J. Wilensky https://orcid.org/0000-0001-7716-9312
Peter K. G. Williams https://orcid.org/0000-0003-3734-3587

References

Abdurashidova, T. H. C. Z., Adams, T., Aguirre, J. E., et al. 2023, ApJ, 945, 124
Abdurashidova, Z., Aguirre, J. E., Alexander, P., et al. 2022, ApJ, 925, 221
Acharya, A., Mertens, F., Ciardi, B., et al. 2024, MNRAS, 534, L30
Ade, P. A. R., Aghanim, N., Arnaud, M., et al. 2016, A&A, 594, A13
Aguirre, J. E., Murray, S. G., Pascua, R., et al. 2022, ApJ, 924, 85
Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2015, ApJ, 809, 61
Amiri, M., Bandura, K., Chen, T., et al. 2023, ApJ, 947, 16
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Barkana, R., Outmezguine, N. J., Redigol, D., & Volansky, T. 2018, PhRvD, 98, 103005
Barnett, A. H., Magland, J., & af Klinteberg, L. 2019, SJSC, 41, C479
Barry, N., Hazelton, B., Sullivan, I., Morales, M. F., & Pober, J. C. 2016, MNRAS, 461, 3135
Barry, N., Wilensky, M., Trott, C. M., et al. 2019, ApJ, 884, 1
Bassa, C. G., Vruno, F. D., Winkel, B., et al. 2024, A&A, 689, L10
Beardsley, A. P., Hazelton, B. J., Sullivan, I. S., et al. 2016, ApJ, 833, 102
Becker, G. D., D'Aloisio, A., Christenson, H. M., et al. 2021, MNRAS, 508, 1853
Berkhout, L. M., Jacobs, D. C., Abdurashidova, Z., et al. 2024, PASP, 136, 045002
Berlin, A., Hooper, D., Krnjaic, G., & McDermott, S. D. 2018, PhRvL, 121, 011102
Bevins, H. T. J., de Lera Acedo, E., Fialkov, A., et al. 2022, MNRAS, 513, 4507
Blamart, M., & Liu, A. 2025, arXiv:2505.09674
Blas, D., Lesgourgues, J., & Tram, T. 2011, JCAP, 2011, 034
Boerner, T. J., Deems, S., Furlani, T. R., Knuth, S. L., & Towns, J. 2023, in Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good, PEARC '23 (Association for Computing Machinery), 173
Bosman, S. E. I., Davies, F. B., Becker, G. D., et al. 2022, MNRAS, 514, 55
Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2015, ApJ, 803, 34
Bouwens, R. J., Oesch, P. A., Labbé, I., et al. 2016, ApJ, 830, 67
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
Breitman, D., Mesinger, A., Murray, S. G., et al. 2024, MNRAS, 527, 9833
Bull, P. 2024, A Quick Primer on the HERA DPSS Fringe-Rate Filter Implementation Memo 129, Univ. Manchester
Bull, P., El-Makadema, A., Garsden, H., et al. 2025, RASTI, 4, rzaf046
Bunker, A. J., Saxena, A., Cameron, A. J., et al. 2023, A&A, 677, A88
Byrne, R., Morales, M. F., Hazelton, B., et al. 2019, ApJ, 875, 70
Camps, A., Torres, F., Corbella, I., Bará, J., & de Paco, P. 1998, RaSc, 33, 1543
Cang, J., Mesinger, A., Murray, S. G., et al. 2025, A&A, 698, A152
Castellano, M., Napolitano, L., Fontana, A., et al. 2024, ApJ, 972, 143
Charles, N., Kern, N. S., Pascua, R., et al. 2024, MNRAS, 534, 3349
Chatterjee, S., & Bharadwaj, S. 2019, MNRAS, 483, 2269
Chen, K.-F., Wilensky, M. J., Liu, A., et al. 2025, ApJ, 979, 191
Cheng, C., Parsons, A. R., Kolopanis, M., et al. 2018, ApJ, 868, 26
Choudhury, T. R., Paranjape, A., & Bosman, S. E. I. 2021, MNRAS, 501, 5782
Ciardi, B., & Ferrara, A. 2005, Space Sci. Rev., 116, 625
Clark, M., Plante, P. L., & Greenhill, L. 2013, IJHPC, 27, 178
Cox, T. A., Parsons, A. R., Dillon, J. S., Ewall-Wice, A., & Pascua, R. 2024, MNRAS, 532, 3375
Cox, T. A., Murray, S. G., Parsons, A. R., et al. 2025, RASTI, 4, rzaf056
Cruz, H. A. G., Munoz, J. B., Sabti, N., & Kamionkowski, M. 2025, PhRvD, 111, 083503
Cullen, F., McLure, R. J., McLeod, D. J., et al. 2023, MNRAS, 520, 14
Cyr, B., Acharya, S. K., & Chluba, J. 2024, MNRAS, 534, 738
Datta, K. K., Jensen, H., Majumdar, S., et al. 2014, MNRAS, 442, 1491
Datta, K. K., Mellema, G., Mao, Y., et al. 2012, MNRAS, 424, 1877
Davies, F. B., Bosman, S. E. I., Furlanetto, S. R., Becker, G. D., & D'Aloisio, A. 2021, ApJ, 918, L35
de Lera Acedo, E., de Villiers, D. I. L., Razavi-Ghods, N., et al. 2022, NatAs, 6, 984

De Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., et al. 2008, MNRAS, 388, 247
DeBoer, D. R., Parsons, A. R., Aguirre, J. E., et al. 2017, PASP, 129, 045001
Dewdney, P., Turner, W., Braun, R., et al. 2016, SKA1 System Baseline Design v2 SKA-TEL-SKO-0000002, SKA Observatory
Dillon, J. S., Lee, M., Ali, Z. S., et al. 2020, MNRAS, 499, 5840
Dillon, J. S., Liu, A., Williams, C. L., et al. 2014, PhRvD, 89, 023002
Dillon, J. S., & Murray, S. 2023, H6C Internal Data Release 2.2 Memo 125, Univ. of California
Dillon, J. S., Neben, A. R., Hewitt, J. N., et al. 2015, PhRvD, 91, 123011
Dillon, J. S., & Parsons, A. R. 2016, ApJ, 826, 181
Donnan, C. T., McLure, R. J., Dunlop, J. S., et al. 2024, MNRAS, 533, 3222
Eastwood, M. W., Anderson, M. M., Monroe, R. M., et al. 2019, AJ, 158, 84
Ewall-Wice, A., Chang, T.-C., Lazio, J., et al. 2018, ApJ, 868, 63
Ewall-Wice, A., Dillon, J. S., Hewitt, J. N., et al. 2016, MNRAS, 460, 4320
Ewall-Wice, A., Kern, N., Dillon, J. S., et al. 2021, MNRAS, 500, 5195
Fagnoni, N., Acedo, E. d. L., Drought, N., et al. 2021a, ITAP, 69, 8143
Fagnoni, N., de Lera Acedo, E., DeBoer, D. R., et al. 2021b, MNRAS, 500, 1232
Feng, C., & Holder, G. 2018, ApJ, 858, L17
Fialkov, A., & Barkana, R. 2019, MNRAS, 486, 1763
Fialkov, A., Barkana, R., Tseliakhovich, D., & Hirata, C. M. 2012, MNRAS, 424, 1335
Fialkov, A., Barkana, R., & Visbal, E. 2014, Natur, 506, 197
Fialkov, A., Barkana, R., Visbal, E., Tseliakhovich, D., & Hirata, C. M. 2013, MNRAS, 432, 2909
Finkelstein, S. L., Leung, G. C. K., Bagley, M. B., et al. 2024, ApJ, 969, L2
Fixsen, D. J., Kogut, A., Levin, S., et al. 2011, ApJ, 734, 5
Fragos, T., Lehmer, B., Tremmel, M., et al. 2013, ApJ, 764, 41
Franzen, T. M. O., Vernstrom, T., Jackson, C. A., et al. 2019, PASA, 36, e004
Furlanetto, S. R., Peng Oh, S., & Briggs, F. H. 2006, PhR, 433, 181http://arxiv.org/pdf/astro-ph/0608032v2.pdf
Gaikwad, P., Haehnelt, M. G., Davies, F. B., et al. 2023, MNRAS, 525, 4093
Gardner, J. P., Mather, J. C., Clampin, M., et al. 2006, SSRv, 123, 485
Garsden, H., Greenhill, L., Bernardi, G., et al. 2021, MNRAS, 506, 5802
Gehlot, B. K., Koopmans, L. V. E., Brackenhoff, S. A., et al. 2024, A&A, 681, A71
Gehlot, B. K., Koopmans, L. V. E., de Bruyn, A. G., et al. 2018, MNRAS, 478, 1484
Gehlot, B. K., Mertens, F. G., Koopmans, L. V. E., et al. 2020, MNRAS, 499, 4158
Gelli, V., Mason, C., & Hayward, C. C. 2024, ApJ, 975, 192
Gessey-Jones, T., Pochinda, S., Bevins, H. T. J., et al. 2024, MNRAS, 529, 519
Ghara, R., Datta, K. K., & Choudhury, T. R. 2015, MNRAS, 453, 3143
Ghara, R., Giri, S. K., Ciardi, B., Mellema, G., & Zaroubi, S. 2021, MNRAS, 503, 4551
Ghara, R., Giri, S. K., Mellema, G., et al. 2020, MNRAS, 493, 4728
Ghara, R., Zaroubi, S., Ciardi, B., et al. 2025, A&A, 699, A109
Gorce, A., Ganjam, S., Liu, A., et al. 2023, MNRAS, 520, 375
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Goulding, A. D., Greene, J. E., Setton, D. J., et al. 2023, ApJL, 955, L24
Greene, J. E., Labbe, I., Goulding, A. D., et al. 2024, ApJ, 964, 39
Greig, B., & Mesinger, A. 2015, MNRAS, 449, 4246
Greig, B., & Mesinger, A. 2018, MNRAS, 477, 3217
Greig, B., Mesinger, A., Koopmans, L. V. E., et al. 2021a, MNRAS, 501, 1
Greig, B., Trott, C. M., Barry, N., et al. 2021b, MNRAS, 500, 5322
Gueuning, Q., Crichton, D., Sampath, A., et al. 2022, in 2022 Int. Conf. Electromagnetics in Advanced Applications (ICEAA), 288
Gupta, Y., Ajithkumar, B., Kale, H. S., et al. 2017, CSci, 113, 707
Guzmán, A. E., May, J., Alvarez, H., & Maeda, K. 2011, A&A, 525, A138
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hazelton, B. J., Jacobs, D. C., Pober, J. C., & Beardsley, A. P. 2017, JOSS, 2, 140
Heiligstein, W., & Jacobs, D. 2023, How Often Is Lightning to Blame for RFI in HERA Data? Memo 121, HERA
Hickish, J., Abdurashidova, Z., Ali, Z., et al. 2016, JAI, 5, 1641001
Hills, R., Kulkarni, G., Meerburg, P. D., & Puchwein, E. 2018, Natur, 564, E32
Hirling, P., Bianco, M., Giri, S. K., et al. 2024, A&C, 48, 100861
Högbom, J. A. 1974, A&AS, 15, 417
Hunter, J. D. 2007, CSE, 9, 90
Hurley-Walker, N., Callingham, J. R., Hancock, P. J., et al. 2016, MNRAS, 464, 1146
Jacobs, D. C., Hazelton, B. J., Trott, C. M., et al. 2016, ApJ, 825, 114
Josaitis, A. T., Ewall-Wice, A., Fagnoni, N., & de Lera Acedo, E. 2022, MNRAS, 514, 1804
Kaur, H. D., Qin, Y., Mesinger, A., et al. 2022, MNRAS, 513, 5097
Keating, G., Hazelton, B., Kolopanis, M., et al. 2025, JOSS, 10, 7482
Kern, N. S., Liu, A., Parsons, A. R., Mesinger, A., & Greig, B. 2017, ApJ, 848, 23
Kern, N. S., Dillon, J. S., Parsons, A. R., et al. 2020a, ApJ, 890, 122
Kern, N. S., Parsons, A. R., Dillon, J. S., et al. 2019, ApJ, 884, 105
Kern, N. S., Parsons, A. R., Dillon, J. S., et al. 2020b, ApJ, 888, 70
Kim, H., Kern, N. S., Hewitt, J. N., et al. 2023, ApJ, 953, 136
Kim, H., Nhan, B. D., Hewitt, J. N., et al. 2022, ApJ, 941, 207
Kittiwisit, P., Murray, S. G., Garsden, H., et al. 2025, RASTI, 4, rzaf001
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, ed. F. Loizides & B. Schmidt (IOS Press), 87
Kolopanis, M., Jacobs, D. C., Cheng, C., et al. 2019, ApJ, 883, 133
Kolopanis, M., Pober, J. C., Jacobs, D. C., & McGraw, S. 2023, MNRAS, 521, 5120
La Plante, P., Williams, P. K. G., Kolopanis, M., et al. 2021, A&C, 36, 100489
Lazare, H., Sarkar, D., & Kovetz, E. D. 2024, PhRvD, 109, 043523
Leung, G. C. K., Bagley, M. B., Finkelstein, S. L., et al. 2023, ApJ, 954, L46
Leung, G. C. K., Finkelstein, S. L., Pérez-González, P. G., et al. 2025, ApJ, 992, 26
Li, W., Pober, J. C., Barry, N., et al. 2019, ApJ, 887, 141
Line, J. L. B., Trott, C., Barry, N., Null, D., & Jordan, C. H. 2025, PASA, 42, e024
Line, J. L. B., Trott, C. M., Cook, J. H., et al. 2024, PASA, 41, e067
Liu, A., & Parsons, A. R. 2016, MNRAS, 457, 1864
Liu, A., Parsons, A. R., & Trott, C. M. 2014, PhRvD, 90, 023018
Liu, A., & Shaw, J. R. 2020, PASP, 132, 062001
Liu, A., Tegmark, M., Morrison, S., Lutomirski, A., & Zaldarriaga, M. 2010, MNRAS, 408, 1029
Liu, H., Outmezguine, N. J., Redigolo, D., & Volansky, T. 2019, PhRvD, 100, 123011
Loeb, A. 2006, SciAm, 295, 46
Madau, P., & Fragos, T. 2017, ApJ, 840, 39
Maiolino, R., Risaliti, G., Signorini, M., et al. 2025, MNRAS, 538, 1921
Maiolino, R., Scholtz, J., Curtis-Lake, E., et al. 2024, A&A, 691, A145
Mao, Y., Tegmark, M., McQuinn, M., Zaldarriaga, M., & Zahn, O. 2008, PhRvD, 78, 023529
Martinot, Z. E. 2022, PhD thesis, Univ. Pennsylvania
Matthee, J., Naidu, R. P., Brammer, G., et al. 2024, ApJ, 963, 129
McGreer, I. D., Mesinger, A., & D'Odorico, V. 2015, MNRAS, 447, 499
McKinley, B., Yang, R., López-Caniego, M., et al. 2015, MNRAS, 446, 3478
Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, MNRAS, 493, 1662
Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2025, A&A, 698, A186
Mesinger, A. 2016, ASSL, 423
Mesinger, A., Ewall-Wice, A., & Hewitt, J. 2014, MNRAS, 439, 3262
Mesinger, A., Furlanetto, S., & Cen, R. 2010, MNRAS, 411, 955
Mirocha, J., & Furlanetto, S. R. 2019, MNRAS, 483, 1980
Mirocha, J., Muñoz, J. B., Furlanetto, S. R., Liu, A., & Mesinger, A. 2022, MNRAS, 514, 2010
Mittal, S., & Kulkarni, G. 2021, MNRAS, 503, 4264
Monsalve, R. A., Rogers, A. E. E., Bowman, J. D., et al. 2021, ApJ, 908, 145
Monsalve, R. A., Altamirano, C., Bidula, V., et al. 2024, MNRAS, 530, 4125
Moore, D. F., Aguirre, J. E., Kohn, S. A., et al. 2017, ApJ, 836, 154
Morales, M. F., Beardsley, A., Pober, J., et al. 2019, MNRAS, 483, 2207
Morales, M. F., Pober, J., & Hazelton, B. J. 2023, MNRAS, 525, 2834
Morales, M. F., & Wyithe, J. S. B. 2010, ARA&A, 48, 127
Muñoz, J. B. 2023, MNRAS, 523, 2587
Muñoz, J. B., Dvorkin, C., & Loeb, A. 2018, PhRvL, 121, 121301
Muñoz, J. B., & Loeb, A. 2018, Natur, 557, 684
Muñoz, J. B., Mirocha, J., Chisholm, J., Furlanetto, S. R., & Mason, C. 2024, MNRAS, 535, L37
Muñoz, J. B., Mirocha, J., Furlanetto, S., & Sabti, N. 2023, MNRAS: Letters, 526, L47
Muñoz, J. B., Qin, Y., Mesinger, A., et al. 2022, MNRAS, 511, 3657
Munshi, S., Mertens, F. G., Chege, J. K., et al. 2025a, MNRAS, 542, 2785
Munshi, S., Mertens, F. G., Koopmans, L. V. E., et al. 2024, A&A, 681, A62
Munshi, S., Mertens, F. G., Koopmans, L. V. E., et al. 2025b, A&A, 697, A203
Murray, S., Greig, B., Mesinger, A., et al. 2020, JOSS, 5, 2582
Murray, S. G. 2024, Excess Variance of Visibilities and the Z-Squared Statistic Memo 130, HERA
Murray, S. G., Bowman, J. D., Sims, P. H., et al. 2022, MNRAS, 517, 2264

Murray, S. G., & Trott, C. M. 2018, ApJ, 869, 25
Neben, A. R., Bradley, R. F., Hewitt, J. N., et al. 2016, ApJ, 826, 199
Nikolić, I., Mesinger, A., Davies, J. E., & Prelogović, D. 2024, A&A, 692, A142
Nunhokee, C. D., Null, D., Trott, C. M., et al. 2025, ApJ, 989, 57
O'Hara, O. S. D., Gueuning, Q., de Lera Acedo, E., et al. 2025, MNRAS, 538, 31
Oesch, P. A., Bouwens, R. J., Illingworth, G. D., Labbé, I., & Stefanon, M. 2018, ApJ, 855, 105
Orosz, N., Dillon, J. S., Ewall-Wice, A., Parsons, A. R., & Thyagarajan, N. 2019, MNRAS, 487, 537
Orrú, E., Norden, M. J., Iacobelli, M., ter Veen, S., & Ahmadi, A. 2024, SPIE, 13098, 238
Paciga, G., Albert, J. G., Bandura, K., et al. 2013, MNRAS, 433, 639
Paciga, G., Chang, T.-C., Gupta, Y., et al. 2011, MNRAS, 413, 1174
Pagano, M., & Liu, A. 2020, MNRAS, 498, 373
Pagano, M., Liu, J., Liu, A., et al. 2023, MNRAS, 520, 5552
Park, J., Mesinger, A., Greig, B., & Gillet, N. 2019, MNRAS, 484, 933
Parsons, A., Backer, D., Chang, C., et al. 2006, in 2006 Fortieth Asilomar Conf. Signals, Systems and Computers (IEEE), 2031
Parsons, A., Backer, D., Siemion, A., et al. 2008, PASP, 120, 1207
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
Parsons, A. R., Liu, A., Aguirre, J. E., et al. 2014, ApJ, 788, 106
Parsons, A. R., Pober, J. C., Aguirre, J. E., et al. 2012, ApJ, 756, 165
Pascua, R., Martinot, Z. E., Liu, A., et al. 2025, ApJ, 985, 127
Patil, A. H., Yatawatta, S., Koopmans, L. V. E., et al. 2017, ApJ, 838, 65
Patil, A. H., Zaroubi, S., Chapman, E., et al. 2014, MNRAS, 443, 1113
Patra, N., Subrahmanyan, R., Sethi, S., Shankar, N. U., & Raghunathan, A. 2015, ApJ, 801, 138
Philip, L., Abdurashidova, Z., Chiang, H. C., et al. 2019, JAI, 8, 1950004
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6
Pober, J. C., Ali, Z. S., Parsons, A. R., et al. 2015, ApJ, 809, 62
Pober, J. C., Liu, A., Dillon, J. S., et al. 2014, ApJ, 782, 66
Pochinda, S., Gessey-Jones, T., Bevins, H. T. J., et al. 2024, MNRAS, 531, 1113
Price, D. C., 2016 PyGDSM: Python interface to Global Diffuse Sky Models, Astrophysics Source Code Library, ascl:1603.013
Price, D. C., Greenhill, L. J., Fialkov, A., et al. 2018, MNRAS, 478, 4193
Pritchard, J. R., & Loeb, A. 2012, RPPh, 75, 086901
Qin, Y., Mesinger, A., Greig, B., & Park, J. 2021, MNRAS, 501, 4748
Qin, Y., Mesinger, A., Park, J., Greig, B., & Muñoz, J. B. 2020a, MNRAS, 495, 123
Qin, Y., Mesinger, A., Prelogović, D., et al. 2025, PASA, 42, e049
Qin, Y., Poulin, V., Mesinger, A., et al. 2020b, MNRAS, 499, 550
Rahimi, M., Pindor, B., Line, J. L. B., et al. 2021, MNRAS, 508, 5954
Rath, E., Pascua, R., Josaitis, A. T., et al. 2025, MNRAS, 541, 1125
Reis, I., Fialkov, A., & Barkana, R. 2020, MNRAS, 499, 5993
Remazeilles, M., Dickinson, C., Banday, A. J., Bigot-Sazy, M.-A., & Ghosh, T. 2015, MNRAS, 451, 4311
Riseley, C. J., Galvin, T. J., Sobey, C., et al. 2020, PASA, 37, e029
Rogers, A. E. E., & Bowman, J. D. 2012, RaSc, 47, RS0K06
Schneider, A., Schaeffer, T., & Giri, S. K. 2023, PhRvD, 108, 043030
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
Singh, S., Nambissan, T. J., Subrahmanyan, R., et al. 2022, NatAs, 6, 607
Slepian, D. 1978, ATTTJ, 57, 1371
Smirnov, O. M. 2011, A&A, 527, A106
Sokolowski, M., Tremblay, S. E., Wayth, R. B., et al. 2015, PASA, 32, e004
Storer, D., Dillon, J. S., Jacobs, D. C., et al. 2022, RaSc, 57, e2021RS007376
Tan, J., Liu, A., Kern, N. S., et al. 2021, ApJS, 255, 26
Taylor, A. R., Stil, J. M., & Sunstrum, C. 2009, ApJ, 702, 1230
Thyagarajan, N., Jacobs, D. C., Bowman, J. D., et al. 2015, ApJ, 804, 14
Tingay, S. J., Goeke, R., Bowman, J. D., et al. 2013, PASA, 30, e007
Topping, M. W., Stark, D. P., Endsley, R., et al. 2022, ApJ, 941, 153
Trac, H., Chen, N., Holst, I., Alvarez, M. A., & Cen, R. 2022, ApJ, 927, 186
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, 53
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261
Visbal, E., Barkana, R., Fialkov, A., Tseliakhovich, D., & Hirata, C. M. 2012, Natur, 487, 70
Vruno, F. D., Winkel, B., Bassa, C. G., et al. 2023, A&A, 676, A75
Wilensky, M. J., Barry, N., Morales, M. F., Hazelton, B. J., & Byrne, R. 2020, MNRAS, 498, 265
Wilensky, M. J., Morales, M. F., Hazelton, B. J., et al. 2023, ApJ, 957, 78
Witstok, J., Jakobsen, P., Maiolino, R., et al. 2025, Natur, 639, 897
Yoshiura, S., Pindor, B., Line, J. L. B., et al. 2021, MNRAS, 505, 4775
Zarka, P., Girard, J. N., Tagger, M., & Denis, L. 2012, in SF2A-2012 Proc. Annual Meeting of the French Society of Astronomy and Astrophysics, 687
Zheng, H., Tegmark, M., Buza, V., et al. 2014, MNRAS, 445, 1084
Zhu, Y., Becker, G. D., Bosman, S. E. I., et al. 2022, ApJ, 932, 76