The Astrophysical Journal, 924:85 (26pp), 2022 January 10 https://doi.org/10.3847/1538-4357/ac32cd

OPEN ACCESS

# Validation of the HERA Phase I Epoch of Reionization 21 cm Power Spectrum Software Pipeline

James E. Aguirre[1], Steven G. Murray[2], Robert Pascua[3], Zachary E. Martinot[1], Jacob Burba[4], Joshua S. Dillon[5,23], Daniel C. Jacobs[2], Nicholas S. Kern[6], Piyanat Kittiwisit[7], Matthew Kolopanis[2], Adam Lanman[3], Adrian Liu[3], Lily Whitler[2], Zara Abdurashidova[5], Paul Alexander[8], Zaki S. Ali[5], Yanga Balfour[9], Adam P. Beardsley[2], Gianni Bernardi[9,10,11], Tashalee S. Billings[1], Judd D. Bowman[2], Richard F. Bradley[12], Philip Bull[7,13], Steve Carey[8], Chris L. Carilli[14], Carina Cheng[5], David R. DeBoer[5], Matt Dexter[5], Eloy de Lera Acedo[8], John Ely[8], Aaron Ewall-Wice[15], Nicolas Fagnoni[8], Randall Fritz[9], Steven R. Furlanetto[16], Kingsley Gale-Sides[8], Brian Glendenning[14], Deepthi Gorthi[5], Bradley Greig[17], Jasper Grobbelaar[9], Ziyaad Halday[9], Bryna J. Hazelton[18,19], Jacqueline N. Hewitt[6], Jack Hickish[5], Austin Julius[9], Joshua Kerrigan[4], Saul A. Kohn[1], Paul La Plante[1], Telalo Lekalake[9], David Lewis[2], David MacMahon[5], Lourence Malan[9], Cresshim Malgas[9], Matthys Maree[9], Eunice Matsetela[9], Andrei Mesinger[20], Mathakane Molewa[9], Miguel F. Morales[18], Tshegofalang Mosiane[9], Abraham R. Neben[6], Bojan Nikolic[8], Aaron R. Parsons[5], Nipanjana Patra[5], Samantha Pieterse[9], Jonathan C. Pober[4], Nima Razavi-Ghods[8], Jon Ringuette[18], James Robnett[14], Kathryn Rosie[9], Mario G. Santos[7,9], Peter Sims[2,3], Saurabh Singh[3], Craig Smith[9], Angelo Syce[9], Nithyanandan Thyagarajan[2,14,24], Peter K. G. Williams[21,22], and Haoxuan Zheng[6]

HERA Collaboration

[1] Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, USA;jaguirre@sas.upenn.edu
[2] School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA
[3] Department of Physics and McGill Space Institute, McGill University, 3600 University Street, Montreal, QC H3A 2T8, Canada
[4] Department of Physics, Brown University, Providence, RI, USA
[5] Department of Astronomy, University of California, Berkeley, CA, USA
[6] Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA
[7] Department of Physics and Astronomy, University of Western Cape, Cape Town, 7535, South Africa
[8] Cavendish Astrophysics, University of Cambridge, Cambridge, UK
[9] South African Radio Observatory (SARAO), 2 Fir Street, Observatory, Cape Town, 7925, South Africa
[10] Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa
[11] INAF-Istituto di Radioastronomia, via Gobetti 101, I-40129 Bologna, Italy
[12] National Radio Astronomy Observatory, Charlottesville, VA, USA
[13] School of Physics & Astronomy, Queen Mary University of London, London, UK
[14] National Radio Astronomy Observatory, Socorro, NM, USA
[15] Department of Astronomy and Physics, University of California, Berkeley, CA, USA
[16] Department of Physics and Astronomy, University of California, Los Angeles, CA, USA
[17] School of Physics, University of Melbourne, Parkville, VIC 3010, Australia
[18] Department of Physics, University of Washington, Seattle, WA, USA
[19] eScience Institute, University of Washington, Seattle, WA, USA
[20] Scuola Normale Superiore, I-56126 Pisa, PI, Italy
[21] Center for Astrophysics | Harvard & Smithsonian, Cambridge, MA, USA
[22] American Astronomical Society, Washington, DC USA

*Received 2021 April 15; revised 2021 September 28; accepted 2021 October 23; published 2022 January 14*

## Abstract

We describe the validation of the HERA Phase I software pipeline by a series of modular tests, building up to an end-to-end simulation. The philosophy of this approach is to validate the software and algorithms used in the Phase I upper-limit analysis on wholly synthetic data satisfying the assumptions of that analysis, not addressing whether the actual data meet these assumptions. We discuss the organization of this validation approach, the specific modular tests performed, and the construction of the end-to-end simulations. We explicitly discuss the limitations in scope of the current simulation effort. With mock visibility data generated from a known analytic power spectrum and a wide range of realistic instrumental effects and foregrounds, we demonstrate that the current pipeline produces power spectrum estimates that are consistent with known analytic inputs to within thermal noise levels (at the $2\sigma$ level) for $k > 0.2h\,\mathrm{Mpc}^{-1}$ for both bands and fields considered. Our input spectrum is intentionally amplified to enable a strong "detection" at $k \sim 0.2\,h\,\mathrm{Mpc}^{-1}$—at the level of $\sim 25\sigma$—with foregrounds dominating on larger scales and thermal noise dominating at smaller scales. Our pipeline is able to detect this amplified input signal after suppressing foregrounds with a dynamic range (foreground to noise ratio) of $\gtrsim 10^7$. Our validation test suite uncovered several sources of scale-independent signal loss throughout the pipeline, whose amplitude is well-

[23] NSF Astronomy and Astrophysics Postdoctoral Fellow.

[24] Jansky Fellow of the National Radio Astronomy Observatory, USA.

characterized and accounted for in the final estimates. We conclude with a discussion of the steps required for the next round of data analysis.

*Unified Astronomy Thesaurus concepts:* Reionization (1383); Astronomical simulations (1857); Astronomy data analysis (1858); H I line emission (690)

## 1. Introduction

Measurement of the highly redshifted 21 cm hyperfine transition of neutral hydrogen holds great promise as a probe of the Epoch of Reionization (EoR), as well as earlier and later epochs. Because the power spectrum of 21 cm fluctuations must be measured in the presence of foregrounds that are $\sim 10^5$ brighter (in temperature units) than the EoR signal, the level of precision required of every aspect of the analysis is extraordinarily high. Because the line-of-sight power spectrum is measured using the spectral axis, it is critically important to avoid introducing additional spectral structure in the data during the analysis, as this can contaminate the 21 cm spectrum with foreground power. This is particularly a problem for interferometric measurements, which mix spatial and spectral structure (Datta et al. 2009; Parsons et al. 2012a). Inaccuracies may also be introduced by analysis choices that affect the amplitude of the desired signal relative to other portions of the data, e.g., biased estimators of the power spectrum or overfitting of foreground models or calibration parameters. It is thus necessary to demonstrate the accuracy of the analysis both for individual steps in the analysis and for the interconnected, complicated chain of analysis from raw data to power spectrum.

A number of groups are currently seeking to detect the H I fluctuation signal from the EoR via the power spectrum. Current efforts include those of the Murchison Widefield Array (MWA; Tingay et al. 2013; Dillon et al. 2014; Beardsley et al. 2016; Ewall-Wice et al. 2016; Li et al. 2019; Barry et al. 2019b; Trott et al. 2020), the Low Frequency Array (LOFAR; van Haarlem et al. 2013; Patil et al. 2017; Gehlot et al. 2018; Mertens et al. 2020), the Long Wavelength Array (LWA; Eastwood et al. 2019), and the Hydrogen Epoch of Reionization Array (HERA; DeBoer et al. 2017). Prior work also includes the Giant Meter Wave Radio Telescope (GMRT; Paciga et al. 2013) and the Donald C. Backer Precision Array for Probing the Epoch of Reionization (PAPER; Kolopanis et al. 2019).

A persistent problem has been that the complexity of the measurement, combined with the novelty of analysis techniques, has created situations in which significant biases are created in the final power spectrum in ways that are not initially obvious. Due to these complications, there have been a number of limits that have required significant revision. These include the GMRT (Paciga et al. 2011 as amended by Paciga et al. 2013) and PAPER (Ali et al. 2015 as amended by Ali et al. 2018 and Cheng et al. 2018). Liu & Shaw (2020) provide a good overview of many of these issues.

In response, an increased effort has been made to explore the effects of choices in 21 cm analysis pipelines via simulation. These studies have attempted to isolate specific effects, for example, sky-based calibration errors (Barry et al. 2016; Ewall-Wice et al. 2017; Mouri Sardarabadi & Koopmans 2019), redundant calibration errors (Byrne et al. 2019; Orosz et al. 2019), instrumental coupling systematics (Kern et al. 2019), power spectrum estimation (Cheng et al. 2018), foreground modeling and subtraction (Chapman et al. 2013; Mertens et al. 2018; Kern & Liu 2021), interferometric imaging (Offringa et al. 2019b), the effect of radio frequency interference (RFI) (Wilensky et al. 2020), and data in-painting (Offringa et al. 2019a; Trott et al. 2020).

Increasing effort has also gone into connecting these isolated studies into more complete end-to-end simulations of the pipelines. For example, the MWA team has two parallel pipelines (Jacobs et al. 2016; Trott et al. 2016; Barry et al. 2019a). The reliability of the pipeline in recovering a mock power spectrum (but without including the effects of calibration) was tested in Beardsley et al. (2016) and more explicitly in Barry et al. (2019b) (Figure 8). The LOFAR limits published in Mertens et al. (2020) have had the method simulated in Mertens et al. (2018) and the effect of calibration considered in Mouri Sardarabadi & Koopmans (2019) and Mevius et al. (2022).

This paper details the current status of an end-to-end simulation effort for the HERA pipeline, as a companion paper to The HERA Collaboration et al. (2021, hereafter HC21) and specifically addresses the instrument configuration and systematic effects of HERA Phase I. Importantly, as will be expanded upon later, these validation tests aim to verify the accuracy of the HERA Phase I pipeline under the intrinsic assumptions of the pipeline itself. Furthermore, they provide a reproducible framework with which to evaluate future HERA analysis pipelines and data releases. These tests are in principle sufficient to avoid the algorithmic errors leading to revisions such as those in Cheng et al. (2018).

The outline of the paper is as follows: Section 2 briefly describes the HERA instrument and the software pipeline we are attempting to validate. Section 3 explains the underlying philosophy of software development and organization of the validation effort, while Section 4 outlines the simulation methods used for each individual portion of the pipeline and the results of isolated tests of those portions. Section 5 then shows the results for the end-to-end pipeline simulation and a comparison with an independent method of estimating the power spectrum. We conclude with a discussion of lessons learned and next steps in Section 6.

## 2. The HERA Instrument and Software Pipeline

### *2.1. The HERA Instrument*

HERA (DeBoer et al. 2017) is a dedicated instrument to measure the power spectrum of spatial fluctuations in the strength of the hyperfine signal of neutral hydrogen during the EoR and Cosmic Dawn. The final instrument, currently under construction at the SKA South Africa site, will comprise a core of three-hundred and twenty 14 m parabolic dishes in a fractured hexagonal-close-pack configuration (Dillon & Parsons 2016) with 30 outrigger antennas. It will operate over the frequency range 50–250 MHz ($27 < z < 5$).

Here we are concerned with the state of the instrument consistent with the HC21 data,[25] which comprises 39 active antennas operating from 100–200 MHz, using the feed type described in Fagnoni et al. (2021b) in the configuration shown

[25] This data set is referred to within the collaboration as H1C Internal Data Release (IDR) 2.2.

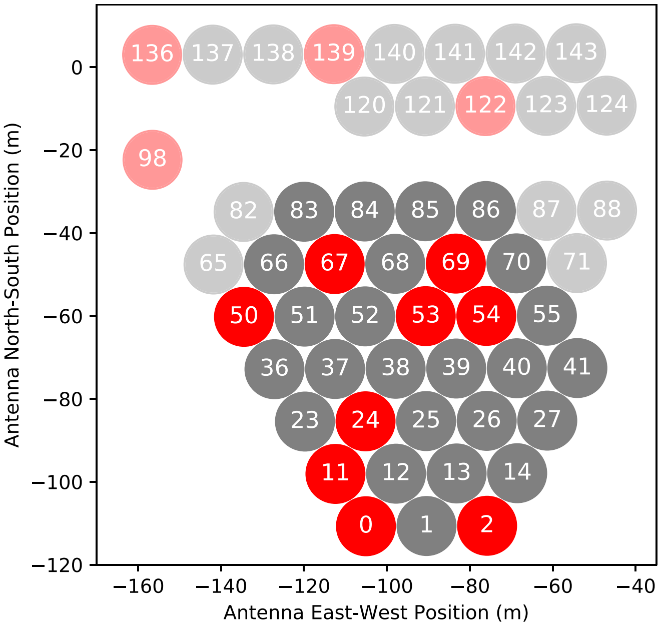

**Figure 1.** The array layout and antenna numbering scheme for the HERA Phase I data (see Figure 2 of HC21). The subset of antennas used in end-to-end validation is shown by dark colors; the additional antennas present in the real array are transparent in this figure. Flagged antennas, shown in red, match those flagged in the real data (Dillon et al. 2020).

in Figure 1. In particular, the systematic effects considered are specific to that instrument. The Phase II instrument under construction has an entirely different feed (Fagnoni et al. 2021a), signal path, and correlator, and thus will have very different behavior; the simulation and validation of the analysis pipeline for that instrument are the subjects of future work.

### 2.2. *Brief Overview of the HERA Analysis Pipeline*

The HERA data reduction pipeline takes raw data output from the correlator and delivers calibrated visibilities with bad antennas, RFI, and other anomalies flagged or removed, and produces delay-type (Parsons et al. 2012b; Morales et al. 2019) power spectra, with accompanying error bars and null tests. A full accounting is given in HC21, and the key steps are shown in Figure 2 (in the "Analysis Pipeline" and "PSpec Pipeline" columns). Steps in the pipeline that were not included in the validation are indicated with faded colors. We briefly outline the steps here, indicating the reason certain steps were not included and where further information can be found.

Raw data are delivered from the correlator (not shown in Figure 2, but equivalent to the output of the systematic simulation). A first round of quality checking attempts to identify antennas that are not performing correctly ("Antenna Metrics"), as the inclusion of these in the following step ("Redundant Calibration") adversely affects the results. We did not attempt to produce simulations of various kinds of antenna defect (hardware failure, incorrect wiring), but simply excluded a subset of possible antennas to simulate the effects of flagging the antennas. "Redundant calibration" uses the constraint of the repeated array configuration to solve for internal degrees of freedom by forcing identical baselines to have an identical response. The particular implementation is described in greater detail in Dillon et al. (2020). "Absolute Calibration" solves for the remaining degenerate parameters undetermined by redundant calibration. This is done by comparing the redundantly calibrated data to a set of model visibilities that have been absolutely calibrated. Those model visibilities had their absolute flux and phase determined by CASA, using a model of GLEAM sources, as described in Kern et al. (2020a). The model visibilities are then subsequently smoothed to not contain structure beyond the baseline's horizon delay (with an additional 50 ns buffer) or 150 ns, whichever is larger. Using a calibration based on an incomplete sky model is known to produce biases in the estimated power spectra (Barry et al. 2016). This is mitigated for HERA because the sky-based calibration is only used to determine the degenerate parameters (fewer degrees of freedom) and also because of the subsequent gain smoothing. Because we do not know the level to which the CASA model was actually incomplete, we do not simulate the effect of that error in this analysis, in keeping with our philosophy of simulating data that respects the pipeline assumptions. We discuss this as part of future work in Section 6.

At this point the data are flagged ("RFI Flagging") based on a number of metrics, including the output of the calibration steps, with the goal of removing particularly nonredundant behavior and RFI. An entirely separate study is required to understand the efficacy of this algorithm in removing RFI, which we defer to later work. While unflagged RFI is a concern for the power spectrum (e.g., Wilensky et al. 2020), we show in HC21 that there is no strong evidence for unflagged RFI at the current noise levels, and so here we concern ourselves primarily with the effect of gaps in the data resulting from RFI flagging. Thus, there is not a simulated RFI injection and removal step; we simply copy the flagging pattern from the real data and use it to create gaps in the simulated data.

Following this, the final gains are smoothed in frequency and time to avoid imparting spurious structure ("Gain Smoothing"), and a final calibrated data set is produced for each night. The data are then averaged over nights of observation ("LST Binning"), with data taken at the same LST for a given baseline averaged together. Any remaining gaps in the averaged baseline visibilities due to flagged data are filled ("Delay Inpainting") (see Section 4.4; Parsons & Backer 2009; Kern et al. 2020b). At this point two systematic effects are corrected ("Systematics Removal"): the presence of internal reflections, causing an "echo" of the signal at different time delays, effectively changing the gain solutions; and a cross-coupling between antennas creating an additive signal. These effects are described in Kern et al. (2019, 2020b). Baselines are then averaged in time ("Coherent Time Averaging") to produce the data set used in "Power Spectrum Estimation" (the pipeline used in HC21 uses `hera_pspec`, which we validate against an alternative power spectrum pipeline—`simpleDS`—in this study). Additional averaging occurs over different baseline types and the "cylindrical" average from $(k_{\parallel}, k_{\perp})$ to $k$ to produce 1D power spectra with associated errors. Further details about the pipeline can be found in HC21.

## 3. Methods

### 3.1. *Overview of Validation Effort*

The desire to ensure that the HERA analysis pipeline does not produce biased results motivated the creation of a separate "Validation" group within HERA (see Appendix A), which seeks to provide an ongoing framework for testing the pipeline

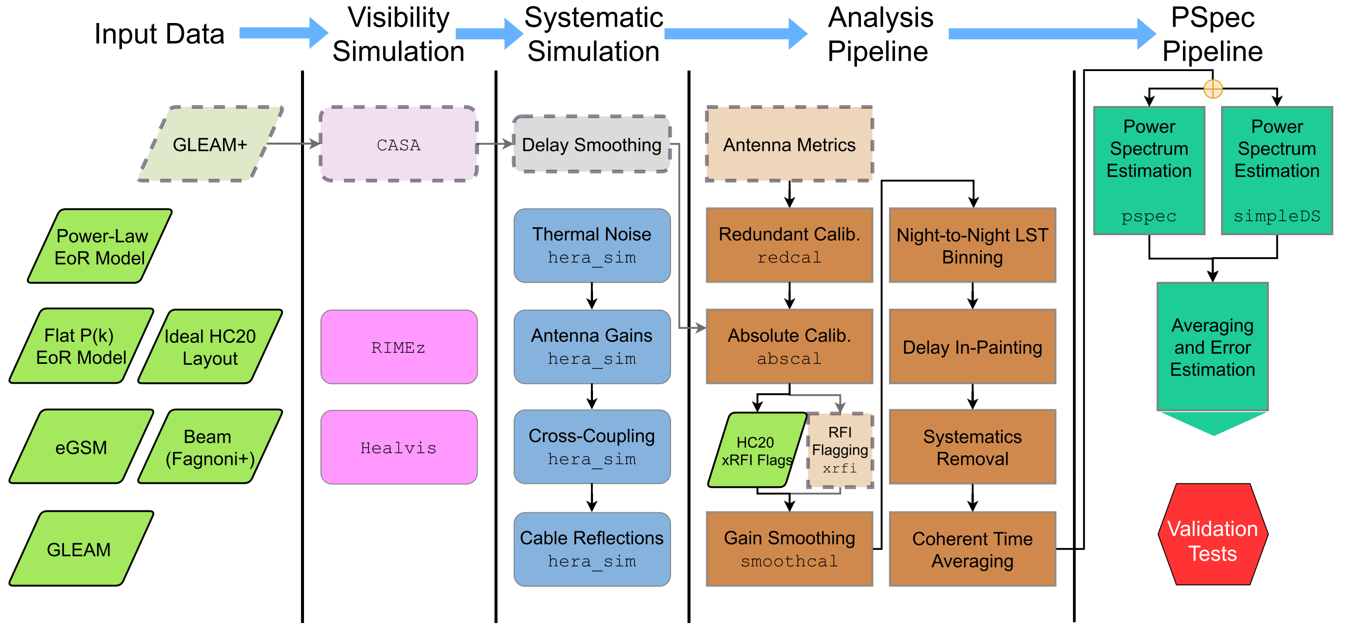

**Figure 2.** A schematic representation of the entire validation pipeline. See Figure 3 for the way in which individual components were tested. Simulation and analysis flow from left to right, and top to bottom where applicable, as indicated by the black arrows. Broadly, the output of the "Input Data" step is sky and beam models in image space, which are then input to the "Visibility Simulation" step, which produces "true" visibilities (Equation (16) as used in Equation (19)), spanning 24 hr of LST and the full frequency coverage of the instrument. At the "Systematic Simulation" step, these visibilities are transformed into the actual time sampling of the HERA observations over multiple days, downselected to the observed LST range, and have multiplicative and additive systematic and instrument effects applied. The input to the "Analysis Pipeline" step is then equivalent in format to the raw observed data, and the processing proceeds as if the simulated data were real, including using the same configuration parameters (e.g., smoothing scales and in-painting tolerances). At the "PSpec Pipeline" stage, further averaging and data selection occur before the estimation of the power spectrum in selected frequency (redshift) bands and LST ranges. All steps from the actual pipeline are shown, but not all were tested in this analysis. Steps not tested are indicated in a lighter shade and with dashed borders. In particular, in the "Input Data" and "Visibility Simulation" sections, we did not construct the absolute calibration (ABSCAL) model in the same way as Kern et al. (2020a) (as indicated by the "GLEAM+" and "CASA" boxes and gray arrows); rather in our pipeline, we use the exact simulated foreground model with delay smoothing applied. In the "Analysis Pipeline" section, we did not generate the kinds of antenna errors that would result in flagging from the "Antenna Metrics," and we did not simulate and extract RFI in the validation pipeline, but rather used preexisting real flags from the data to test the effect of flagging, as indicated by the "HC20 xRFI Flags" box in place of "RFI Flagging `xrfi`".

**Table 1**
Table of Repositories Tested in this Validation Effort

| NAME | URL (https://github.com/) | DESCRIPTION |
|---|---|---|
| PIPELINE | | |
| `hera_cal` | hera-team/hera_cal | Redundant and sky-based calibration routines. |
| `hera_pspec` | hera-team/hera_pspec | Robust foreground-avoidance power spectrum (and covariance) estimation. |
| SIMULATION | | |
| `RIMEz` | upenneor/rimez | Fast and accurate visibility calculation implementing multiple methods for different source and beam function definitions. |
| `spin1_beam_model` | upenneor/spin1_beam_model | Harmonic space decomposition of the HERA primary beam. |
| `healvis` | rasg-affiliates/healvis | Fast visibility simulation based on HEALPIX discretization. |
| `pyuvsim` | RadioAstronomySoftwareGroup/pyuvsim | Accurate visibility simulation of point sources with very limited approximations. |
| `gcfg` | zacharymartinot/redshifted_gaussian_fields | Consistently simulate cosmological Gaussian fields over the full sky |
| `hera_sim` | hera-team/hera_sim | Add HERA-specific instrumental systematics to visibilities. |

via realistic simulations. The scope of this paper is somewhat narrower. The HERA analysis and power spectrum pipeline described above is clearly a large and complex system. It is implemented, in part, by the public software repositories in Table 1, which comprise at least four complete, original, Python packages, with upwards of 40,000 standard lines of code between them. While each of the packages is written to a high collaboration standard (see Appendix A.1 for details), the interplay between the subcomponents is much more difficult to test. What we wish to do here is verify that the software pipeline used in HC21 performs as expected in the case where the simulated data match the underlying assumptions of the analysis. Importantly, we do not explore the effects of violating certain key assumptions of the pipeline, including perfect redundancy of antenna elements (Dillon et al. 2020), or systematic effects that differ substantially from Kern et al. (2019, 2020b). That is, we assume that the physical effects for which the analysis pipeline was designed to remove are the only (nonnegligible) effects in the data and that the modeling of these effects in the pipeline is comprehensive in principle. (In some cases, it was not possible to completely simulate what was done in HC21, and we have noted this.) Our key metric for this validation is the recovery of a known power spectrum, without significant bias in the recovered signal, at the level of error bars that is consistent with the known level of thermal noise and its coupling to the signal, following the error analysis in Tan et al. (2021).

The approach used here tests subcomponents of the analysis with multiple simulations but does not attempt a statistical suite of simulations of the full end-to-end result. This is partially the result of practical limitations of computation (many aspects of the simulation pipeline would need to be sped up), but also because we expect (and show) that in the absence of systematic effects that do not deviate from our assumptions, the output power spectrum is reproduced within the errors. A more thorough investigation of ensemble effects is appropriate as the limits continue to come down and in the exploration of the violation of pipeline assumptions.

At the current sensitivity of HERA, we do not expect to make high-significance detections of the EoR power spectrum. Consequently, our criterion for “how good is good enough?” in assessing the results of our simulations is that any systematic effect in the analysis is smaller than the expected error bars on the EoR spectrum, or the systematic errors in its calibration. In practice, that means we consider effects to be “small” if they cause a change in the power spectrum of less than 10%. This bound will clearly need to be tightened as we begin to move toward detections. It is worth commenting that errors may appear at different points in the analysis and may affect the calibration gains, the individual visibilities, or the power spectrum itself. Our metric is the power spectrum, and we note that errors in gains $g$ typically propagate to the power spectrum as $g^4$ and errors in visibilities $V$ as $V^2$. Thus, errors that affect the gains or visibilities must meet correspondingly smaller fractional error requirements so that the final effect on the power spectrum stays within the desired bound.

### *3.2. Schematic Overview of the Validation Effort*

We designed the validation effort to be incremental, building complexity in successive steps, and finally resulting in a simulation including a large fraction of the physical effects that the HERA pipeline attempts to address. We divided the various components required for a full simulation into steps and tested each. The various simulation components are outlined in Figure 2, and the testing steps in Figure 3. Section 4 describes these in more detail.

A row in Figure 3 indicates which elements of the simulation were included in the step. For each of these steps (except `2.0`), the primary metric of success involves the estimated power spectrum. As we progress through the steps, generally more elements are included, i.e., these are integration tests where we cumulatively test the interaction between components. This has the potential to uncover undetected errors concerning the interaction of individual components, but also the potential to hide errors that are negligible in the final power spectrum metric.

Each specific `major.minor` step tests a unique pathway through Figure 2, combining different inputs, simulator, and systematics with relevant pipeline steps. Figure 3 encodes each of these pathways in matrix form. Note that the steps were not meant to test every possible combination of components, but rather to coherently build toward the ultimate test, which essentially combines all of them. The reason for this incremental complexity building was pragmatic—defining a full end-to-end simulation containing all known physical effects is a large task, and it is not necessarily clear from the outset what form an effect will have on the final result. Furthermore, in the case where the test fails, it is difficult to disentangle effects and determine which component (or combination thereof) caused the failure. Increasing complexity gradually builds confidence in the individual analysis components before combining them.

## 4. Simulation Components

We now walk sequentially through the various components of the simulation, as shown in the first three columns of Figure 2, and describe how they are constructed, including the methods and codes (Table 1), as well as how they were tested within the rubric of Figure 3. With any simulation, there are physical and instrument effects that are ignored, either due to ignorance (i.e., unknown systematic effects) or practical limitations in including them in the simulation. We have included in Figure 2 a set of effects that encapsulate the most complete description of the sky and instrument that we are able to construct at this time. There are known additional effects that are the subject of future work, which we discuss further in Section 6.

### *4.1. Instrument and Foreground Models*

The first step in simulating instrumental output is to make models of the desired signal (“mock EoR”) and astrophysical foregrounds (point sources and diffuse emission), as well as to simulate the antenna response pattern and interferometer layout.

#### *4.1.1. Mock EoR*

To be able to test the unbiased recovery of an EoR power spectrum, it was highly desirable to produce a mock-EoR sky with a known, analytic power spectrum $P(k)$. It was deemed more important that the power spectrum be analytically known and that the simulated EoR be full sky, covering both $4\pi$ steradians and the full observation bandwidth, than that it be derived from a physical simulation. While this means that the mock EoR we inject will not have the most realistic power spectrum nor will it have a non-Gaussian component, these are second-order effects for ascertaining whether there is bias in the recovery of the power spectrum.

We chose to use for our mock-EoR signal a realization of a Gaussian random brightness temperature field $T_e(\vec{r}, z)$ (expressed in the emitted frame) with a power spectrum with a shape

$$P_e(k, z) = A_0 k^{-2}, \tag{1}$$

which approximates power spectra obtained from cosmological simulations. The observed field is given by

$$T(\hat{s}, \nu) = \frac{\nu}{\nu_e} T_e(\mathbf{r} = r_\nu \hat{s}, z = z_\nu), \tag{2}$$

where $\nu$ is the observed frequency, $\nu_e$ is the rest frequency, $z_\nu = \frac{\nu_e}{\nu} - 1$ is the redshift of the source point in the direction $\hat{s}$ on the sky, and

$$r_\nu = c \int_0^{z_\nu} \frac{dz}{H(z)} \tag{3}$$

is the comoving distance in terms of the Hubble function $H(z)$. Note that the cosmological parameters are implicit in the

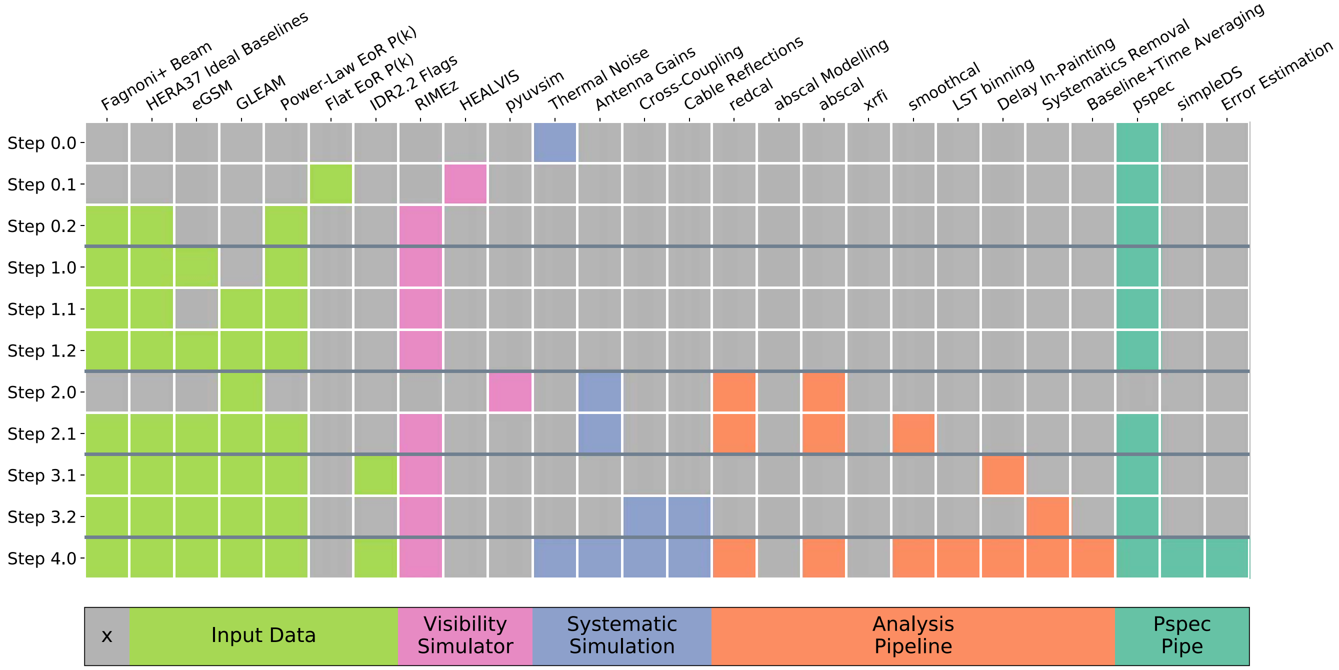

**Figure 3.** Components included in each validation step. Note the color scheme used here for the various components matches that of Figure 2, which describes the corresponding steps in the analysis pipeline. Thick horizontal lines represent boundaries between major steps. Varying colors represent different subcategories of the pipeline. Where applicable, components are simulated/applied from left to right. See Figure 2 for a more detailed flow diagram of the simulation and analysis components and products. “Input Data” steps are described in Sections 4.1 and 4.4, “Visibility Simulator” steps in Section 4.2, “Systematic Simulations” in Section 4.3, “Analysis Pipeline” in HC20, and “Pspec Pipe” in HC20 and Kolopanis et al. (2019) and Tan et al. (2021).

comoving distance $r_\nu$; we have used the same parameters as in subsequent power spectrum estimation.

We can create realizations of this temperature field by expanding in spherical harmonic modes

$$T(\hat{\boldsymbol{s}}, \nu) = \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{m} a_{\ell m}(\nu) Y_{\ell m}^{*}(\hat{\boldsymbol{s}}) \tag{4}$$

and generating $a_{\ell m}$ that satisfy

$$\langle a_{\ell m}(\nu) a_{\ell' m'}^{*}(\nu')\rangle = C_\ell(\nu, \nu')\delta_{\ell\ell'}\delta_{mm'}. \tag{5}$$

The cross-frequency angular power spectrum $C_\ell(\nu, \nu')$ is related to the original power spectrum $P_e(k)$ by

$$C_\ell(\nu, \nu') = \frac{2}{\pi}\frac{\nu\nu'}{\nu_e^2}\int_0^\infty dkk^2 j_\ell(r_\nu k) j_\ell(r_\nu' k) P_e(k). \tag{6}$$

For our chosen form of the power spectrum, Equation (1), this takes the simple form

$$C_\ell(\nu, \nu') = \frac{\nu\nu'}{\nu_e^2}\begin{cases} \dfrac{A_0}{2\ell+1}\dfrac{{r_\nu'}^{\ell}}{r_\nu^{\ell+1}} & \text{if } r_\nu' \leqslant r_\nu, \\ \dfrac{A_0}{2\ell+1}\dfrac{r_\nu^{\ell}}{{r_\nu'}^{\ell+1}} & \text{if } r_\nu' > r_\nu. \end{cases} \tag{7}$$

Our mock-EoR signal is then a realization of this cross-frequency spectrum. The maximum $\ell$necessary is determined by the effective angular band limit imposed by the natural spatial filtering of the simulated interferometric array and sufficient error control on the visibility calculation.

The final field on the sky is expressed as a specific intensity (in units of Jy str$^{-1}$) using the conversion

$$I_\nu(\hat{\boldsymbol{s}}, \nu) = \kappa(\nu) T(\hat{\boldsymbol{s}}, \nu). \tag{8}$$

The conversion factor is given by

$$\kappa(\nu) = 2k_B\frac{\nu^2}{c^2} \times 10^{26}\left[\frac{\text{Jy str}^{-1}}{\text{K}}\right] \tag{9}$$

$$= \frac{2k_B}{A(\nu)\Omega(\nu)} \times 10^{26}, \tag{10}$$

where $k_B$ is Boltzmann’s constant in SI units and $A(\nu)$ and $\Omega(\nu)$ are the effective area and solid angle of the beam, respectively.

To verify `hera_pspec`’s normalization conventions and cosmological conversions in going from visibilities to power spectra, we tested the recovery of $P(k) \propto k^{-2}$ in the absence of any foreground emission, noise, or instrumental corruption beyond the beam (designated Step 0.2 in Figure 3). The results are shown in Figure 4. While the agreement is generally good, the results highlight an important aspect of the power spectrum measurement. The estimated power spectrum $\hat{P}(k)$ is related to the true spectrum via a window function $W(k, k')$ via

$$\langle\hat{P}(k)\rangle = \int_0^\infty W(k, k')P(k')dk'. \tag{11}$$

In general, $W(k, k')$ is complicated and cannot be made equal to the ideal $\delta(k - k')$. A discussion of the window function is included in Appendix B, and a general expression is given in Equation (B5). A full calculation of the window function would naturally include effects such as the curvature of the sky (e.g., Liu et al. 2016) and the bandwidth and resolution of the frequency sampling of the data. The window function computed by `hera_pspec` (and given in HC21, Equation (19)) does not fully implement Equation (11) and consequently suffers from small biases at low and high $k$. In Figure 4, we show the size of these biases. In the range where we are most sensitive ($0.2 \leqslant k \leqslant 0.5\, h\,\text{Mpc}^{-1}$), these biases are intrinsically

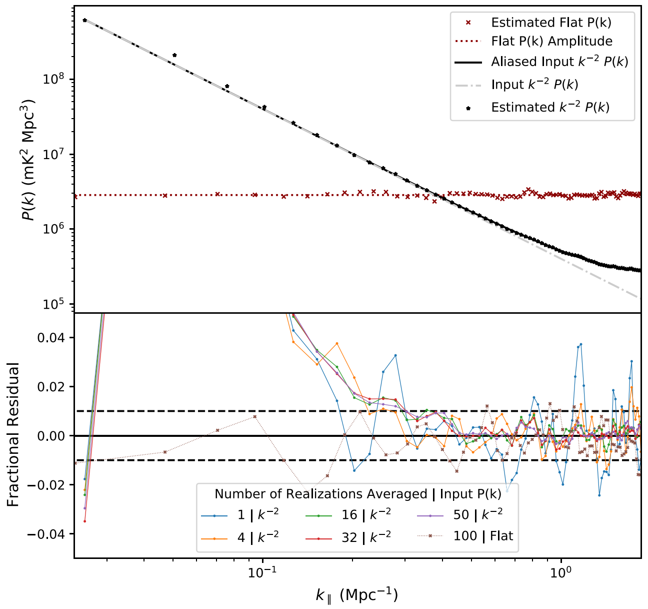

**Figure 4.** A baseline test of recovering the power spectrum for different power spectrum shapes. Top: an analytic $P(k) \propto k^{-2}$ (orange) was converted to its corresponding $C_\ell(\nu, \nu')$; harmonic realizations of this correlation were input to RIMEz to generate mock visibilities, and the delay spectrum was estimated using hera_pspec. The results of a single realization are shown (blue), along with the calculated deviation due to the approximate window function of Equation (B13). (The amplitude here is arbitrary and not related to the level used in the end-to-end test (Section 5).) Bottom: the fractional deviation between the input power spectrum and the recovered one, after correction by Equation (B13), as a function of the number of realizations averaged. Random fluctuations about the mean analytic form are expected due to cosmic variance; those fluctuations average down as shown. For $k \geqslant 0.2$, systematic deviations are $< 1\%$. The systematic deviations at low $k$ are due to not properly calculating the window function (Equation (B5)).

less than 1%. With a simple approximation of the aliasing effect (Equation (B12)), the bias for $k > 0.5$ can be reduced to a similar level. The bias for $0.03 < k < 0.2$ due to the window function is more severe. Part of the discrepancy is simply due to nearly uniform width windows in $k$ when integrated against a power law, but in general, the window functions become more complicated in this regime, leading to biases of both signs. We note that, in the present work, foregrounds will dominate for $k < 0.2$ and consequently this low-$k$ bias on the power spectrum is not detectable. The proper inclusion of the window function in hera_pspec to allow accurate estimation over all $k$ is left to future work.

### *4.1.2. Foreground Models*

The simulated foreground emission is constructed from two components, one of point-like sources and one of spatially smooth diffuse emission. The point-source component is composed of all sources in the GLEAM catalog (Hurley-Walker et al. 2017) for which a spectral model is provided with the catalog, or a number $N_{\rm src}$ of approximately 240,000 sources, each with a power-law emission spectrum. The GLEAM catalog lacks the brightest sources, so these were added separately as point sources according to the values in Table 2 of Hurley-Walker et al. (2017). Fornax A was added based on the model of McKinley et al. (2015). Explicitly, for the $N_{\rm src}$ sources each specified by their flux, spectral index, and position $(F_i, \alpha, \hat{\mathbf{s}}_i)$, this component is described by

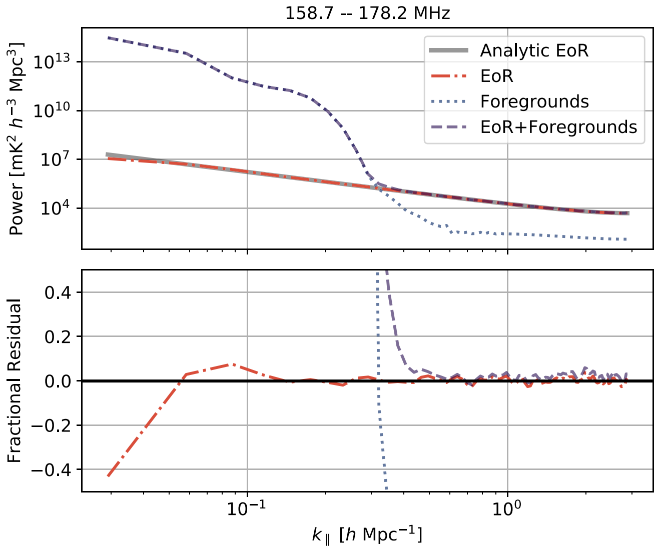

**Figure 5.** Noiseless power spectra estimated with hera_pspec, showing power-law EoR, foregrounds (point-source plus diffuse), and their sum. Also plotted is the analytic input $P_{\rm EoR}(k)$ (corrected for aliasing, see Figure 4.). The lower panel shows the residuals with respect to the analytic input. Note that this figure illustrates both that the simulated foregrounds have the requisite high dynamic range (i.e., spurious spectral structure induced by the simulator is negligible) and that the power spectrum estimator correctly handles the linear sum of EoR and foregrounds in visibility space. Note that the amplitude of $P_{\rm EoR}(k)$ is lower here than in the end-to-end test (see Section 5).

$$I_p(\nu, \hat{\mathbf{s}}) = \sum_{i=1}^{N_{\rm src}} F_i \delta \left(\frac{\nu}{\nu_0}\right)^{\alpha_i} \delta(1 - \hat{\mathbf{s}} \cdot \hat{\mathbf{s}}_i). \qquad (12)$$

The GLEAM catalog has significant gaps in regions nominally covered by the HERA observation, notably at R.A. ∼ 7 hr as the Galactic anticenter transits. Rather than inject artificial sources, we excluded from these simulations observing times where the GLEAM catalog was significantly incomplete in the primary beam. The diffuse emission component was simulated based on an improved version of the GSM (de Oliveira-Costa et al. 2008; Zheng et al. 2017; Kim et al. 2021). In this version of the eGSM, the spatial templates are smooth on 3° scales, and interpolation from the frequencies of the model maps to the desired frequency was done using a Gaussian process regression to ensure spectral smoothness.

The key requirements for the foreground model were that it should be representative of real foregrounds with respect to spectral smoothness and strength. The tests in Step 1 were primarily designed to check hera_pspec's ability to reproduce known input EoR power spectra in the presence of foregrounds for $k$'s outside the foreground-dominated modes, thereby demonstrating that there are no dynamic range limitations in either the visibility simulation or the power spectrum estimation. Figure 5 summarizes the results of this test for the combined GLEAM and eGSM sky model.

### *4.1.3. Antenna Beams*

The HERA antenna vector far-field beams $A^p_{j,\delta}(\hat{s}, \nu)$ are simulated from a detailed electrical and mechanical model using CST Microwave Studio (Fagnoni et al. 2021b). Each polarized feed $p$ of an antenna $j$ responds to incident radiation from infinity in the direction $(\alpha, \delta)$ with a complex vector antenna pattern

$$\boldsymbol{A}_j^p(\hat{s}, \nu) = A_{j,\delta}^p(\hat{s}, \nu)\hat{e}_\delta + A_{j,\alpha}^p(\hat{s}, \nu)\hat{e}_\alpha, \tag{13}$$

where $(\hat{\delta}, \hat{\alpha})$ define an orthogonal coordinate system on the sphere, here taken to be the standard R.A./decl. system. This antenna pattern is proportional to the simulated far-field beam patterns, by the reciprocity theorem. The antenna patterns are assembled into a Jones matrix per antenna as

$$\mathcal{J}_j = \begin{bmatrix} A_{j,\delta}^p(\hat{s}, \nu) & A_{j,\alpha}^p(\hat{s}, \nu) \\ A_{j,\delta}^q(\hat{s}, \nu) & A_{j,\alpha}^q(\hat{s}, \nu) \end{bmatrix}. \tag{14}$$

Note that because the antenna pattern is a vector field, the appropriate representation of it in harmonic space requires spin-1 spherical harmonics. We implemented this with custom code (see Table 1). In addition to producing a spatially smooth representation of the beam, independent of a particular pixelization, the interpolation of the spin-1 harmonic coefficients in frequency preserves the smooth frequency evolution of the beam.

### 4.2. Visibility Simulation

We had several requirements for the simulation of visibilities from HERA. One was that our visibility simulator be able to produce visibilities based on full-sky models of the instrument beam (including treatment of the full Jones matrix, Equation (13)) and sky (including both diffuse and point-source emission). Another was that it correctly handle drift-scan visibilities and be able to compute visibilities at the cadence of HERA time sampling over the bulk of a sidereal day, and at the full frequency resolution and bandwidth of HERA. The resulting visibilities should do a reasonable job of reproducing the observed HERA visibilities, though we do not demand sufficient fidelity that we would be able to calibrate HERA data to the simulated visibilities. Crucially, however, when considering the representation of the visibilities in their Fourier dual spaces (delay for frequency, fringe rate for time), the simulator should not produce numerical artifacts that adversely affect the dynamic range between bright foregrounds and regions in the EoR window. Specifically, with the assumption of spectrally "smooth" (i.e., compact in delay space) input models of the sky and beam, the simulator should not generate numerical errors that scatter foreground power to high delays. Finally, it should be able to do these calculations in a reasonable time.

Our primary simulation engine is `RIMEz`, an internally developed program that meets these requirements. We describe the unique features of `RIMEz` in the next section (Section 4.2.1). Because any simulator will need to make approximations to allow computation in a reasonable time and will leave out some instrument effects, we describe in Section 4.2.2 independent checks of the simulator with reference calculations and make a qualitative comparison against HERA data.

#### 4.2.1. Simulator Method

We take the fundamental visibility measurement equation for all four correlation products from a single baseline of a a polarized interferometer to be

$$\mathcal{V}_{jk}(\nu, t) = \begin{bmatrix} V_{jk}^{pp} & V_{jk}^{pq} \\ V_{jk}^{qp} & V_{jk}^{qq} \end{bmatrix} \tag{15}$$

$$= \int \mathcal{J}_j \mathcal{C} \mathcal{J}_k^\dagger \exp(-2\pi i \nu \boldsymbol{b} \cdot \hat{s}/c) d\hat{s}, \tag{16}$$

where the integration is taken over the full sphere. The coherency matrix is given by

$$\mathcal{C} = \begin{bmatrix} I(\hat{s}, \nu) + Q(\hat{s}, \nu) & U(\hat{s}, \nu) - iV(\hat{s}, \nu) \\ U(\hat{s}, \nu) + iV(\hat{s}, \nu) & I(\hat{s}, \nu) - Q(\hat{s}, \nu) \end{bmatrix}, \tag{17}$$

where $I$, $Q$, $U$, and $V$ are the images of the Stokes sky, expressed relative to the same coordinate system as Equation (13). While `RIMEz` is capable of fully polarized simulations, in this work we assume $Q = U = V = 0$.

In order to actually compute Equation (16), and in particular to address the wide field of view of HERA and nontrivial contribution from diffuse emission over the full sky, it is necessary to treat the integration over the sphere carefully. `RIMEz` evaluates the integral by summation over a harmonic representation of the beam, fringe, and sky terms, rather than evaluating these terms on a pixelization of the sphere, similar to the formalism in Shaw et al. (2014). The `RIMEz` implementation is based on the SSHT code for computing spherical harmonic transforms (McEwen & Wiaux 2011). Computing the visibility integral in harmonic space (for appropriate values of the maximum $\ell$) naturally handles the spherical quadrature correctly for continuous functions on the sphere, like diffuse emission and the beam. Point sources are also included by computing their harmonic space representation; summation of the coefficients for all point sources in the simulation allows compressing an entire catalog into a single set of harmonic coefficients. The relative orientations of the sky and the beam are handled via the *m*-mode formalism for transit telescopes (e.g., Shaw et al. 2014; Eastwood et al. 2019) to calculate visibilities as a function of time. Self-consistent autocorrelations are also produced by including the $\boldsymbol{b} = 0$ term in Equation (16).

#### 4.2.2. Simulator Validation

Because of the many choices inherent in implementing Equation (16) as a numerical calculation, we independently tested `RIMEz` with the goal of ensuring any systematics introduced by the simulator are below the dynamic range inherent to the `hera_pspec` power spectrum estimation in the presence of foregrounds. Undesired chromatic modulation of foregrounds $10^4$ times brighter than the background is the most challenging, but not the only, aspect to consider. At this dynamic range, approximations and errors usually ignored in radio interferometry become important; calculation of phases, pixelization, sky geometry, and simple coding or math errors can all be significant. We checked these issues first by comparison of the `RIMEz` calculations against analytic calculations of the visibility phase and amplitude of simple, unpolarized diffuse, and point-source terms in Equation (16). These tests revealed small numerical errors at the $10^{-10}$ level, consistent with the expected precision of the internal Fourier transforms. An additional test compared against `pyuvsim` reference simulations including only point sources,[26] revealing

[26] See the repository in Table 1 for a description of the reference simulations.

differences in amplitude and phase that were primarily due to small differences in the calculation of current epoch source positions from a source catalog.[27] While source positions are important for, e.g., sky-based calibration, they are not relevant for the validation process because the coordinate system difference is a small rotation of the sky, and thus does not affect the visibility smoothness nor the resulting power spectra.[28] More comprehensive validation tests of the visibility simulators will be required in the future and are currently ongoing.

To check for spectral and time smoothness, as part of Steps 0 and 1, RIMEz was used to generate visibilities for mock EoR and foregrounds and hera_pspec to generate the corresponding power spectra. For both the EoR only (Figure 4) and for high-dynamic-range simulations of EoR in the presence of foregrounds (Figure 5), we were able to recover the input EoR, showing that the simulator was not adding additional spectral structure, as measured by the power spectrum. We also inspected the delay and fringe-rate transforms of the data for anomalous structure (Figure 8) and compared simulated visibilities against real data for qualitative agreement (Figure 9).

For the test in Step 0.1, we also used healvis,[29] which takes a nearly orthogonal approach to RIMEz, computing Equation (16) by pixelizing the beam and sky using HEALPIX (Górski et al. 2005) and performing a simple Riemann sum (Lanman & Pober 2019; Lanman et al. 2020; A. Lanman et al. 2021, in preparation). The interested reader should refer to those papers for a discussion of the accuracy of the healvis simulator.

### 4.3. Noise and Instrument Systematic Simulation

#### 4.3.1. Thermal Noise

Thermal noise is generated from a Gaussian distribution whose variance is determined on a per-time, per-frequency basis according to the amplitude of the simulated noise-free autocorrelation[30] with an added receiver temperature, $T_{\rm rx}$, which is the same for each antenna, and constant with frequency. The standard deviation of the noise in a given (time, frequency) sample of the autocorrelation is calculated via the radiometer equation

$$\sigma(\nu, t) = \kappa(\nu)\Omega(\nu)\frac{[T_{\rm auto}(\nu, t) + T_{\rm rx}]}{\sqrt{\Delta\nu\Delta t}} \quad {\rm Jy}, \tag{18}$$

where $\Delta\nu$ is the channel width, $\Delta t$ is the integration time, and $\kappa(\nu)\Omega(\nu)$ converts to Jy units (see Equation (8)).

Let $V_{apbq,t}$ be the visibility measured by the cross-correlation of feed $p$ on antenna $a$ with feed $q$ on antenna $b$ in time ($t$, $t+\Delta t$) (we drop the implicit dependence of $\nu$ for notational clarity). We assume the thermal noise is uncorrelated between baselines and polarizations, so we may write the noisy visibilities as

$$V_{apbq,t}^{\rm noisy} = V_{apbq,t}^{\rm true} + V_{apbq,t}^{\rm noise}, \quad a \neq b, \tag{19}$$

where the real and imaginary parts of $V_{apbq}^{\rm noise}$ are drawn from $\mathcal{N}(0, \sigma^2)$, with the variance given by Equation (18). Henceforth, we use "true" to denote the simulated visibility from Equation (16), including all sources of astrophysical emission, but excluding noise and instrumental effects except the primary beam. The autocorrelations just have the receiver temperature added (with a signal-to-noise ratio of ~1000 in the autocorrelations, this is a very good approximation); this ensures they remain real and positive definite. The ability to reconstruct the correct power spectrum level given pure input noise was tested in Step 0.0.[31]

#### 4.3.2. Gains

Each antenna feed is assumed to have a direction-independent gain, representing the effects of the electronics and cables. We assumed each antenna was represented by a diagonal Jones matrix (ignoring possible cross-coupling between the feeds). The average bandpass of each feed is described by a degree 6 polynomial fit to measurements of the gain derived from HERA data.[32] each feed receives a unique bandpass by perturbing this average bandpass via convolution with a complex white-noise realization and subsequent application of a phase factor with a randomly generated delay and phase offset. That is, if $g_0(\nu)$ is the bandpass polynomial evaluated at the simulation frequencies $\nu$, then the antenna-based bandpass gains are given by

$$g_{ad}(\nu) = [g_0(\nu)^* K_{ad}(\nu)]\exp(i2\pi\nu\tau_{ad} + i\phi_{ad}), \tag{20}$$

where $*$ indicates convolution in frequency, $K_{ap} \sim \mathcal{N}(0, 1)$ is a complex white-noise convolution kernel, and $\tau_{ap}$ and $\phi_{ap}$ are the randomly selected delay and phase offset, respectively, for antenna $a$ on day $d$ (note that the same random bandpass gain was used for each feed on an antenna). Note that the bandpass gains are randomized per day rather than per time. This formulation ensures that the gains all vary between antennas but retain the same overall average shape. The hera_sim package was used to generate these gains. Using these gains, we determine "uncalibrated" visibilities per frequency, baseline, polarization, and time:

$$V_{apbq,t}^{\rm uncal} = g_{a,d\ni t}\, g_{b,d\ni t}^{*}\, V_{apbq,t}^{\rm noisy}. \tag{21}$$

Step 2 tests demonstrated that redundant and absolute calibration return the known input gains to machine precision, in the absence of noise.[33] We note that because we assume the gains are band limited in delay space, we are not testing the

[27] These small position errors have since been corrected in the RIMEz source code.

[28] The complete results of this simulation comparison can be found at https://nbviewer.jupyter.org/github/HERA-Team/hera-validation/blob/test-neg1.1.0/test-series/-1/test-neg1.1.0.ipynb.

[29] At the time the simulations were done, pyuvsim could not simulate diffuse emission. For the purposes of long-term support, this functionality of healvis has been incorporated into pyuvsim. The standalone package is deprecated and not recommended for new projects.

[30] The antennas are assumed to all have identical beam patterns in this work, so each antenna shares the same autocorrelation visibility prior to the application of systematic effects.

[31] Test notebook available at https://github.com/HERA-Team/hera-validation/blob/master/test-series/0/test-0.0.0_noise_pspec.ipynb.

[32] The HERA bandpass was measured by differencing in time and frequency the cross-correlation visibilities of 19 antennas to estimate thermal noise. This noise was matched to a theoretical foreground and receiver temperature model to generate per-antenna, per-frequency gains. The resulting bandpasses for each antenna are fit by a single shared polynomial multiplied by an independent scalar amplitude per antenna.

[33] Test notebooks available at https://github.com/HERA-Team/hera-validation/blob/master/test-series/2/test-2.0.0.ipynb and https://github.com/HERA-Team/hera-validation/blob/master/test-series/2/test-2.1.0.ipynb.

effect of real instrument gains that have structure beyond the smoothing scale of our `smoothcal` step.

#### 4.3.3. Cross-coupling and Cable Reflection Systematics

Cable reflections are captured as a per-antenna gain-like term described as

$$\tilde{g}_a = \prod_j^M (1 + A_{a,j} \exp(i 2\pi\nu\tau_{a,j} + i\phi_{a,j})), \tag{22}$$

where each reflection is characterized by an amplitude ($A$), delay ($\tau$), and phase offset ($\phi$). The overall effect of reflections is the product of the $M$ different reflections (per antenna) present in the analog signal chain. Note that the reflection gains are generated per antenna and are unchanging with respect to feed and time.

Reflection gains result in the visibilities

$$V^{\rm refl}_{apbq,t} = \tilde{g}_a \tilde{g}_b^* V^{\rm uncal}_{apbq,t}. \tag{23}$$

The cross-coupling systematic present in the H1C data (Kern et al. 2020b), $V^{cc}_{ab}$, is described as

$$V^{\rm cc}_{apbq,t} = V^{\rm refl}_{apap,t} \left( \sum_j^N A^{d,j}_{apbq} \exp(i2\pi\nu\tau^{d,j}_{apbq} + i\phi^{d,j}_{apbq}) \right)_{d\ni t}, \tag{24}$$

where each of the $N$ couplings between the autocorrelation $V_{aa}$ and the cross-correlation $V_{ab}$ is characterized as a per-baseline reflection term. For simplicity, we only apply the cross-coupling systematic to the cross-correlations—this may be thought of as a leading-order approximation to the cross-coupling seen in the data, as cross-coupling shows up as a much smaller effect in the autocorrelations than in the cross-correlations (Kern et al. 2019). Each of the cross-coupling parameters $A$, $\tau$ and $\phi$ are drawn randomly (see Section 5.1 for details) per antenna $a$, feed $p$ and day $d$ (similar to bandpass gains). Note that whether or not the cross-coupling term is also subject to reflection depends on the exact physical origin and placement in the signal chain of the cross-couplings and reflections. In this model, the cross-coupling does not reflect. Such a term would be second order in the (already small) coefficients, so we do not expect it to be significant relative to the current noise level. The final "corrupted data" that are the input to the analysis pipeline are

$$\tilde{V}^{\rm corrupt}_{apbq,t} = V^{\rm refl}_{apbq,t} + \begin{cases} V^{\rm cc}_{apbq,t} & a \neq b \\ 0 & a = b \end{cases}. \tag{25}$$

It is worth noting that we do not remove the cross-coupling systematics by fitting an equation of the form of Equation (24), but rather by using the method in Kern et al. (2020b).

### 4.4. Flagging

The various steps in the analysis pipeline follow very closely the description in Section (2.2). We discuss in detail here the only real departure from the actual data analysis pipeline, which was our treatment of flagged data.

We chose to use the flagging patterns from the real data to test the question of how cutting data affects the power spectrum. Recall that we did not simulate RFI or other effects that would normally be caught by the data quality portions of the pipeline and produce gaps in time and frequency. The delayed in-painting process (Parsons & Backer 2009; Kern et al. 2020b) allows us to "fill in the gaps" in the frequency domain by estimating the values of the underlying Fourier modes in the delay domain. Because the H1C pipeline does not attempt to remove foregrounds, the bright foregrounds contribute extremely large side lobes in delay-space if the frequency axis is Fourier-transformed with step-function-like flagging gaps. Filling in the gaps with an informed estimate of their true value allows us to apply Fourier techniques to (de) flagged data and significantly reduce these side lobes. However, a concern is that errors in the process will propagate power from inside the wedge into the EoR window. There is an additional concern that using in-painted estimates (which have no EoR signal in them) will bias the resulting power spectrum. We show that this effect is negligible with current flagging via the end-to-end analysis. Figure 6 shows the results of the Step 3.1 investigation of this process.[34]

We consider a variety of flagging patterns in time and frequency taken from the data for a variety of different baseline lengths and orientations. We consider how accurately (in the absence of noise) the input power spectrum (foregrounds + EoR) can be reconstructed after in-painting the gappy data. We then cast that in terms of an effective dynamic range (relative to the foreground amplitude at $\tau = 0$ ns) for the reconstruction. Generically, we find that large gaps in frequency near the center of a spectral window limit the dynamic range severely, but that for data with flagging more like the spectral windows chosen in HC21, the dynamic range exceeds $10^9$ in the power spectrum. In the ideal case, the middle row of Figure 6 would show low fractional error in the recovered power spectrum at all $\tau$ (or $k_{\|}$). For the choice of in-painting parameters used in Step 3.1, we chose a 10% deviation of the recovered power spectrum from the flag-free power spectrum as a fiducial marker to estimate the dynamic range, but recovery better than this would, of course, be preferred. In the bottom row of Figure 6, each line crosses 1 at $\tau = \tau_{10\%}$, where the recovered power spectrum deviates by 10% in the absolute fractional error. Better performance thus appears as a larger amplitude at $\tau = 0$ ns. Both the absolute fractional error and dynamic range are plotted in Figure 6 because they represent different aspects of our ability to recover the true power spectrum. Small values of the absolute fractional error are required to accurately recover the power spectrum. Large dynamic range values are required to suppress foreground contamination at higher delays.

In interpreting Figure 6, it is important to note that the in-painting parameters are tunable. For the work in HC21, the in-painting parameters were tuned for recovering power spectra at a dynamic range set by the expected thermal noise of the observations. In the case of the noiseless simulated data used in Step 3.1, the in-painting parameters were tuned further, at the cost of lower computational performance, to demonstrate the ability of the in-painting process to obtain a larger dynamic range. Thus, it is difficult to translate this noise-free test directly into an impact on the measured power spectrum in the presence of noise for the full HERA analysis, and so our end-to-end simulations offer the fullest justification that the additional power added by this process at $k$'s in the EoR window averages down. The reason we performed this test with noise-free

[34] Full test notebook available at https://github.com/HERA-Team/hera-validation/blob/master/test-series/3/test-3.1.0.ipynb.

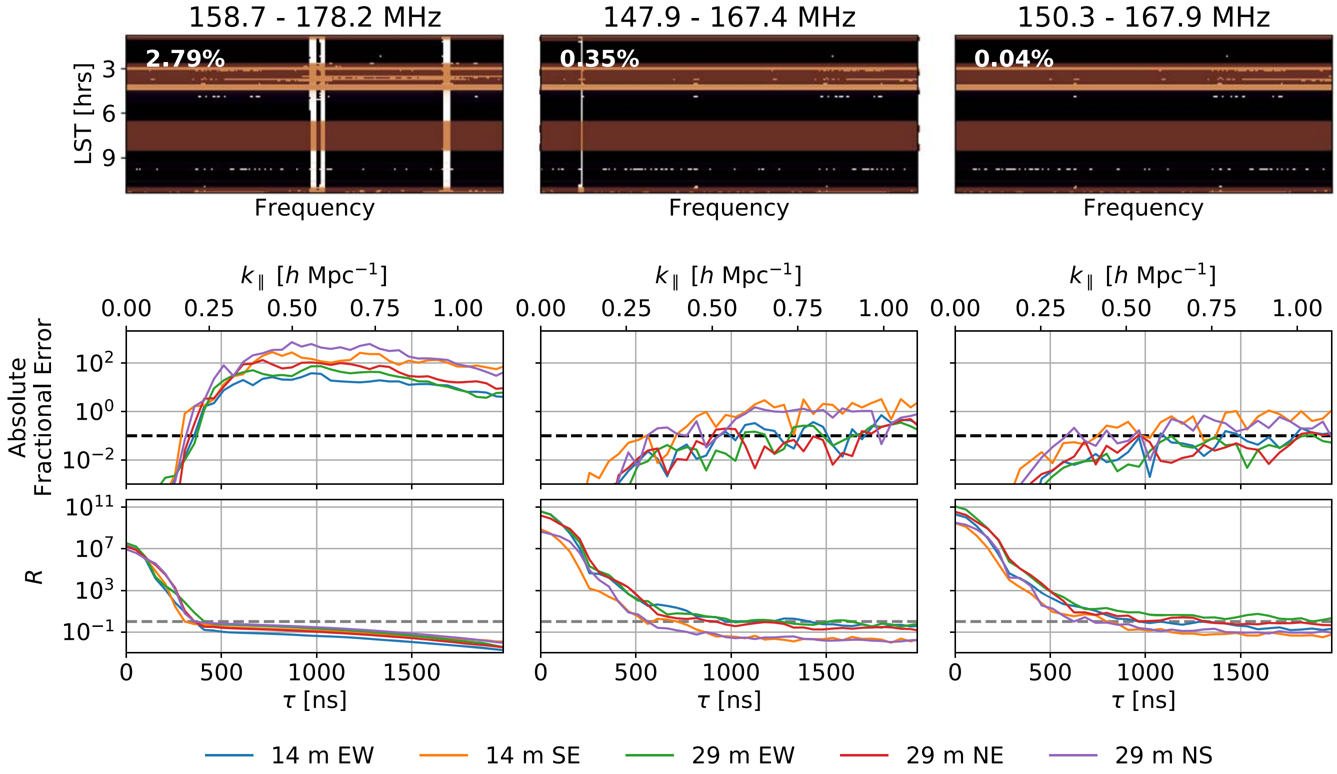

**Figure 6.** In-painting demonstration from Step 3.1. Columns correspond to different spectral windows indexed by channel indices. From left to right, the frequency channel indices corresponding to each spectral window are (600, 800), (490, 690), and (515, 695). Top row: frequency vs. LST flagging waterfalls. White pixels represent flags in data, and red strips are integrations unused due to either flags from the in-painting process itself, flag occupancy cuts, or manual LST cuts. The percentages in the top-left corner of each waterfall represent the number of remaining flags in nonfully flagged integrations, i.e., flags within integrations not highlighted in red. Middle row: magnitude of fractional error of recovered (in-painted) $\hat{P}(\tau)$ relative to known input as a function of delay for various baseline types (colors). The dashed black line marks 10% and the delay at which the fractional error meets or exceeds 10% is denoted as $\tau_{10\%}$. Bottom row: a metric for the dynamic range in the recovered power spectrum, defined as $R = \hat{P}(\tau)/\hat{P}(\tau_{10\%})$. Good dynamic range in recovery corresponds to higher values at $\tau = 0$ ns and greater values of $\tau_{10\%}$ (i.e., crossing the gray reference line farther to the right). Note the generally increasing dynamic range and increasing $\tau_{10\%}$ with fewer flags in the center of the band and intrinsic variability in the recovery for different baseline types.

visibilities was to isolate the accuracy of in-painting; finding the composite effects of in-painting errors interacting with thermal noise effectively requires the level of integration of an end-to-end test.

## 5. End-to-end Simulations

Here we summarize the end-to-end test, which combines the independent steps discussed previously into a single run of our calibration and power spectrum pipeline. We start by discussing the production of the simulation and systematics, and then present the results of the calibration pipeline and the power spectrum pipeline. Lastly, we present additional tests that look carefully at the amount of possible signal loss induced by the analysis pipeline.

### 5.1. The Simulated Data Set

The end-to-end simulation is not an exact replica of the data set in HC21. The simulated data set contains fewer days, a narrower LST range, fewer antennas, and shorter baselines (see Figure 1). The LST restriction is in part due to the limitations of the GLEAM sky model, which lacks sources for R.A. > 7 in HERA's observing patch. We show the differences between the simulated data set and HC21 in Table 2. Nevertheless, the simulated data set is sufficiently complete to capture most of the features of the real data and to reach a comparable depth after all averaging steps were completed.

The corrupted data were created to match the data from the H1C observing season with the computing and software resources currently available. There were three major steps in creating the corrupted data: first, we combined simulated observations of foreground emission and a reionization signal to form the true visibilities, $V^{\rm true}$; next, we modified the set of true visibilities to match the H1C array and observing parameters (modulo nonredundancy);[35] finally, the modified simulations were used to generate 10 days' worth of visibilities that were corrupted with the systematic effects outlined in Sections 4.3–4.3.3.

The base simulation consists of three components: point-source foreground emission, diffuse foreground emission, and visibilities appropriate for a power spectrum of $P_{\rm eor}(k) = 200(k/0.2\ h\ {\rm Mpc}^{-1})^{-2}$. This power spectrum amplitude was chosen to produce a strong detection at $k \sim 0.2\,h\,{\rm Mpc}^{-1}$, but dominated in turn by foregrounds, systematics, and thermal noise at other scales. The modified true simulations were corrupted as described in Section 4.3.3. Here we specify the values used in the data corruption for the end-to-end test. Recall that $V_{ab}^{\rm uncal}$ has two parameters per antenna: the delay, $\tau_a$, and the phase offset, $\phi_a$. We drew delays randomly from a uniform distribution spanning (−20, +20) ns,

[35] This was necessary due to the expensive original simulations having been defined for a slightly different set of antennas within the HERA array. We selected a maximal overlapping subset of the simulated baselines, interpolated to match the intrinsic H1C observation times, for use in the rest of the analysis.

**Table 2**
Comparison of the Parameters of the Real Data Set and the End-to-end Simulation

| | Simulation | Data |
|---|---|---|
| Number of days | 10 | 18 |
| LST range (hr) | 1.5–7 | 1–10 |
| Number of antennas | 33 (8 flagged) | 52 (13 flagged) |
| Total number of baselines | 300 | 741 |
| Maximum baseline length | 84 m | 118 m |

chosen via manually tuning to matched observed features in H1C data, and the phase offsets from a uniform distribution on $[0, 2\pi)$. The bandpass gains used for this work vary only between antennas, nights, and as a function of frequency; we did not add any LST variability to the gains.

Recall that the effect of a single cable reflection is characterized by an amplitude, delay, and phase offset (Equation (22)). As implemented, the reflections were split into two categories: single cable reflections and a reflection "shoulder" meant to model a series of subreflections. The former consisted of two relatively high-amplitude reflections with random per-antenna delays of $(200 + 10\epsilon_{200})$ ns and $(1200 + 30\epsilon_{1200})$ ns, with relative gain amplitudes of $3 \times 10^{-3}$ and $8 \times 10^{-4}$, respectively, and $\epsilon$ a standard normal variable. These delays and relative amplitudes were chosen to match the observed systematics in the H1C system (Kern et al. 2020b). The subreflection shoulder consisted of 20 individual reflections uniformly located between a delay range of 200–1000 ns, with amplitudes following a power law in delay from $10^{-3}$ to $10^{-4}$. Each reflection's delay is perturbed by a Gaussian offset with a scale of 30 ns, and their amplitudes were perturbed randomly by 1% of their assigned value. All reflection terms have their phases drawn from a uniform distribution from $[0, 2\pi)$ and are not varied across frequency, time, and observing night. An example of the reflection gains in delay space is shown in Figure 7.

The cross-coupling model is simulated in a similar manner to the reflection shoulder, meant to match systematics observed in the H1C system. We generated an independent set of reflection terms (amplitude, delay, phase) for each baseline, per night. We characterize the cross-coupling with 10 reflections (each determined by an amplitude, delay, and phase). The 10 delays were spaced linearly between 900 and 1300 ns. Amplitudes are regular in log-space, $A = 10^{-(\tau-100)/200}$ (going from $10^{-4}$ at $\tau = 900$ ns to $10^{-6}$ at $\tau = 1300$ ns), but offset by a random normal variable with scale 0.01% of the amplitude. The phase was drawn from a uniform distribution from $[0, 2\pi)$. Cross-coupling is simulated as described by Equation (24), where the systematic is the product of the autocorrelation visibility with each of the reflection coupling terms. The phase of the reflections is similarly drawn randomly for each reflection and antenna from a uniform distribution from $[0, 2\pi)$.

Figure 8 shows the various components described above in fringe-rate/delay space, for a 51 m WNW-oriented baseline. The top panel shows the EoR component, demonstrating that it primarily populates positive fringe rates (as it is a statistically isotropic and sky-locked signal, and the baseline has a negative EW component). The foreground component has a positive fringe-rate component as well but also demonstrates significant power at fringe rates near zero, peaking at a delay corresponding to the length of the baseline. The "pitchfork effect" due to the monopole described by Thyagarajan et al. (2015) also shows excess power at the delay corresponding to the baseline length; the observed effect here is clearly related to this when resolved in both fringe rate and delay. Importantly, the fringe rates occupied are dependent on the EW component of the baseline, which is why cross-coupling cannot be removed from baselines with a small EW component without incurring signal loss. The cable reflection plot shows the convolution of the reflection terms with the foreground signal, showing that it acts to smear the foreground signal horizontally across the delay. Lastly, the cross-coupling signal is fairly independent of the other terms, and occupies near-zero fringe-rate modes.

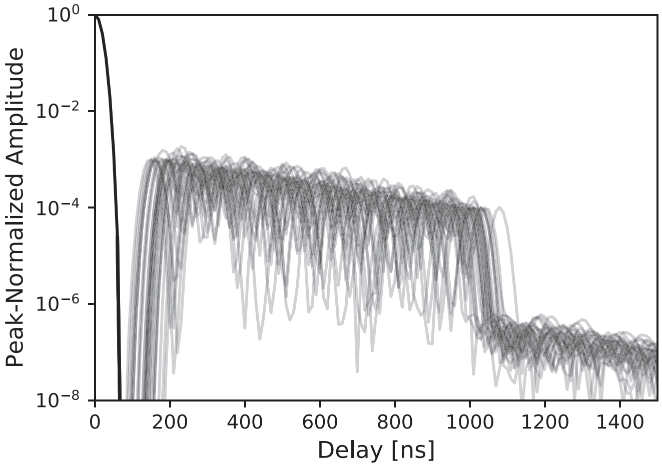

**Figure 7.** Peak-normalized delay spectra of simulated reflections. Each line is a different antenna. Note the relatively narrow spread in the reflection at 200 ns contrasted with the relatively large spread in the reflection at 1200 ns. Low-level features beyond 1200 ns are second-order reflection terms.

In Figure 9, we show a comparison of the simulated data to real data, highlighting the simulation's ability to capture the general features observed in the data.

### 5.2. Calibration Results

Each night of the corrupted simulated data are first passed through the H1C calibration pipeline, which employs a direction-independent calibration of the XX and YY polarizations for each 10.7 s time integration. The performance of redundant calibration on the H1C system is described in Dillon et al. (2020) and can be summarized by the final reduced $\chi^2$ of the gain solutions. Figure 10 shows the reduced $\chi^2$ (blue) of the estimated gains from the simulated data set compared to an idealized, pure-noise distribution (dashed gray), showing a slight positive bias indicative of excess variance due to the presence of baseline-dependent cross-coupling systematics that break the redundancy condition.

When we compare the estimated gains to the true gains used to corrupt the data, we find that the gain phases are recovered to good precision, while the recovered gain amplitudes are biased slightly high (Figure 11). Gains biased slightly high will result in the calibrated data being biased slightly low. This bias comes from a time-varying signal-to-noise ratio (S/N) of the visibilities and our choice of absolute calibration technique (see Figure 12). Calibrating noisy data with a low S/N can lead to biased estimates of the gain amplitude when using a logarithm to linearize the antenna-based calibration equation (Boonstra & van der Veen 2003), as we employ here. Comparing the gain

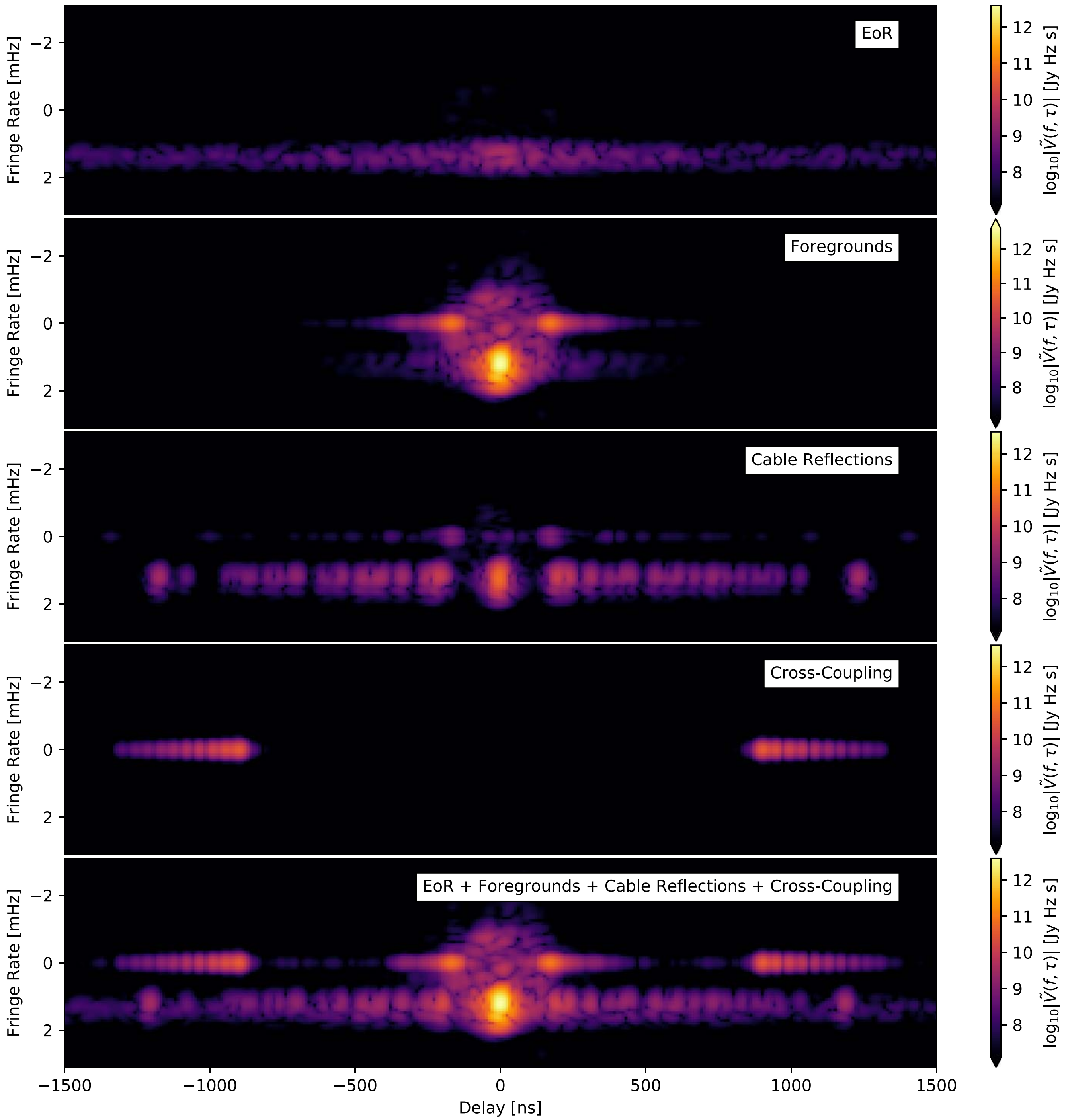

**Figure 8.** Delineation of different simulation components for a 51 m WNW-oriented baseline (44 m westward east–west projection, 25 m northward north–south projection), Fourier transformed along the frequency and time and frequency axes ($x$- and $y$-axes, respectively). The fringe-rate/delay basis is useful for highlighting unique physical characteristics of each component, though note that there is overlap between some components in this basis (e.g., the foregrounds and cable reflections).

solutions to the known simulated gains we find solutions biased high by roughly 4% (left panel of Figure 11).

This absolute flux-scale bias can also be well quantified by imaging the true model visibilities and the postcalibrated data and comparing the fluxes of bright point sources near beam center. This reveals an amplitude biased low by ∼8% and varying slowly with frequency.[36]

To correct this, we multiply all estimated power spectra and their $1\sigma$ error bars by the measured bias of 1.11 for the low band and 1.15 for the midband (Table 3).

Other sources of uncertainty from absolute calibration can also impact the overall error budget. These effects include (i) the overall uncertainty on the flux scale at ∼10% (Hurley-Walker et al. 2017) and (ii) the change in the gain amplitude due to ambient temperature drift during a nightly observation, which is not corrected for in the H1C pipeline and is estimated to be roughly 5% (Kern et al. 2020a). While these sources of

[36] In the real H1C data set, this bias is closer to 4%, which is due to the larger range in R.A. included in the real H1C data compared to the validation data set.

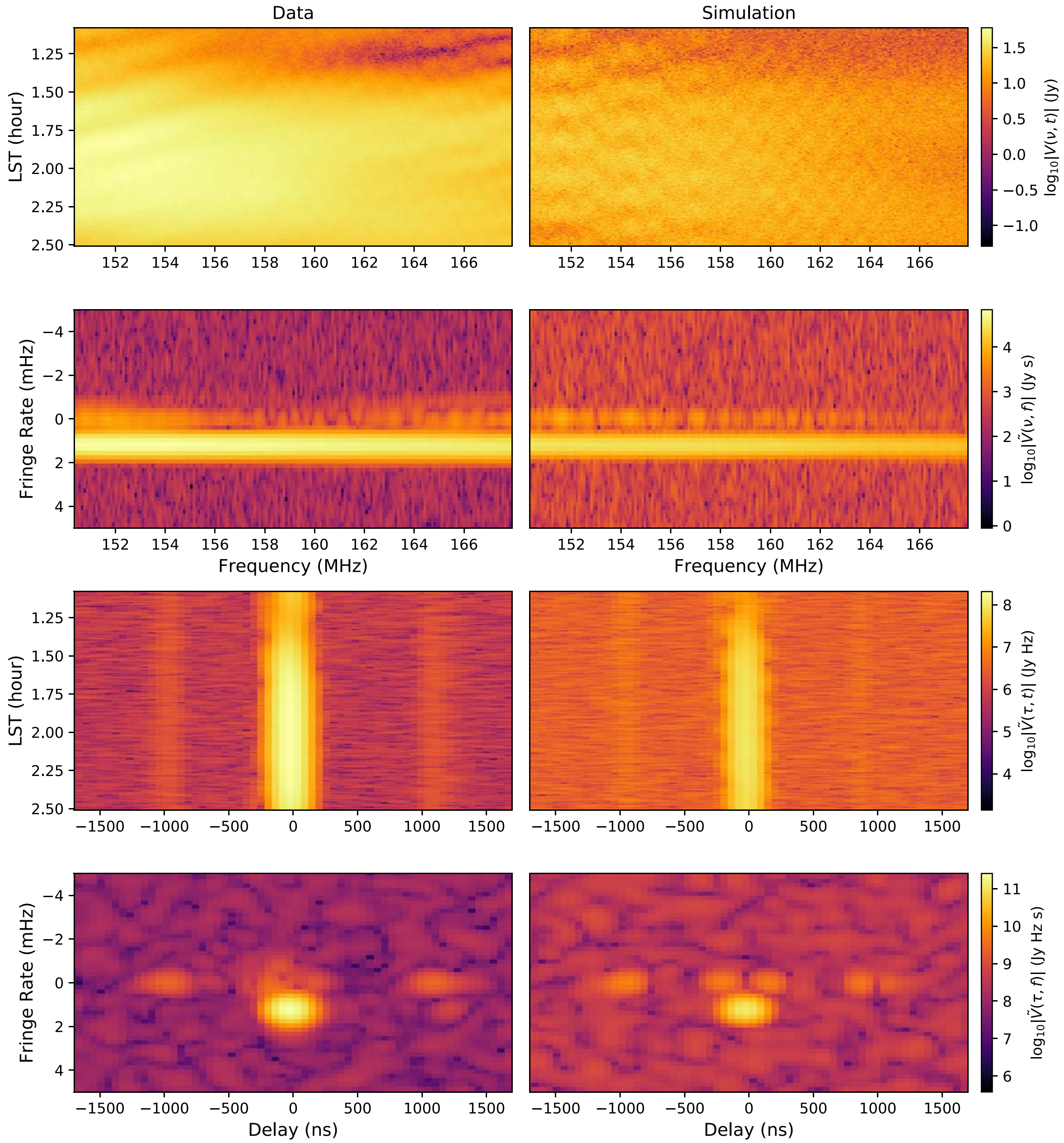

**Figure 9.** Comparison of simulated data (right) with observation data (left) for the same LST range, spectral window, and baseline. Each set of plots shows the four possible choices of Fourier-transformed versions of the data. The observation data has been redundantly and absolutely calibrated, but not treated for cable reflections or the cross-coupling systematic. The observation data have also been LST binned, so it has a substantially lower noise level than the simulation. While the simulation and observation data look strikingly similar, there is a clear difference in the qualities of the high-delay systematic: the simulated version is much more symmetric in delay, and it appears to be somewhat brighter than it is in the observed data.

uncertainty are added in quadrature with the error bars in HC21, we do not do so here as these effects have not been included in simulation.

The model visibilities used in absolute calibration are not constructed in an identical manner to HC21, which used a CASA-based pipeline to calibrate a few independent fields, then stitched them together to construct a set of model visibilities for all LSTs while finally low-pass filtering the visibilities across frequency to reduce noise and fill in gaps due to RFI. Instead, here we simply take the sum of the foreground, EoR, and noise visibilities as our representative model visibilities, sidestepping the question of calibration uncertainties due to an incomplete sky model for the time being. However, as demonstrated in Kern et al. (2020a) and Dillon et al. (2020), the gain-smoothing procedure applied to the post redcal + abscal gains is meant to filter off any fast time and frequency structure in the gains that might be generated by such issues.

Each night is calibrated independently and then binned onto a uniform grid in LST and coherently averaged together (known as LST binning). We show via two methods in Figure 13 that to within a few percent, the noise in our LST-binned visibilities matches our expectations. One way to estimate the noise in LST-binned visibilities is simply to measure the variance of all visibilities that are to be binned together after rephasing to a common phase center. In our case, because our LST-binned data set has a cadence of 21.4 s (twice as long as the nightly simulations with a 10.7 s cadence), we

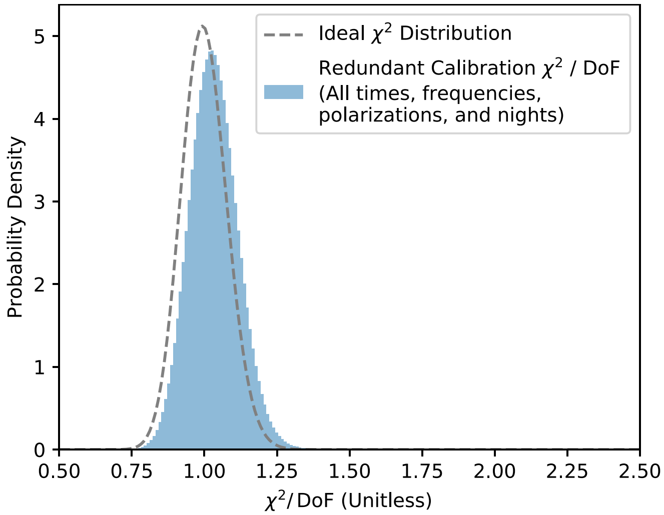

**Figure 10.** The success of redundant-baseline calibration can be assessed by examining the difference between raw visibilities and the visibility model by assuming redundancy and antenna-based gains. This is quantified by $\chi^2$ per degree of freedom, which was defined in (Dillon et al. 2020) and can be compared to a theoretical expectation (in this case with DoF = 164). Here we can see that the simulated distribution of $\chi^2$ nearly matches the expected distribution; we expect a mean value of $\chi^2$/DoF of 1, we observe 1.03. This is substantially better the observed distribution, which peaks around 1.3–1.4 (Dillon et al. 2020). The key difference is that while both the validation simulation and real data contain baseline-dependent cross-talk systematics (an additive effect that breaks the assumption of redundancy), the simulation does not contain any antenna position errors or antenna-to-antenna variation of the primary beam, which likely accounts for most of the observed deviation from 1.

simply compute the variance of all (up to) 20 visibilities in a single LST bin from the 10 nights simulated. Because the LST binner is also estimating the mean visibility, we use the unbiased estimator of the variance (i.e., we use Bessel's correction). The left panel of Figure 13 shows that this matches closely the noise we expect on each input visibility, as inferred from the calibrated autocorrelations using Equation (18). Another way of estimating the noise in the LST-binned visibilities is to take the frequency-interleaved difference—in this case, $V_{ij}(\nu) - \frac{1}{2}V_{ij}(\nu - \Delta\nu) - \frac{1}{2}V_{ij}(\nu + \Delta\nu)$. This gives an estimate of the noise variance at frequency $\nu$, though discontinuities in $N_{\text{samples}}$ complicate it slightly. Regardless, the right panel of Figure 13 shows that the observed noise again matches the expectation for how the noise in the visibilities should integrate down quite well, accounting for the number of samples. In both panels we drop any time or frequencies with $N_{\text{samples}} < 10$ before averaging in order to account for RFI gaps.

After binning, the averaged visibilities are passed through systematics treatment (Kern et al. 2019, 2020b). This involves modeling the smooth foregrounds and in-painting the model in the remaining RFI flags in the data (Section 4.4). Next, the autocorrelations are used to model antenna-based reflections in the signal chain. A total of 28 signal chain reflection terms are iteratively solved for, chosen by visual inspection of the residuals, and the algorithm is only provided the rough location in delay space where we expect reflections to appear (150–1500 ns). After calibrating out the reflection terms, we apply the Kern et al. (2019) procedure for modeling and subtracting off the slowly time-varying cross-coupling systematics. Note that this cannot be done reliably for baselines with a projected east–west length less than 14 m without substantial signal loss (Kern et al. 2019), so we flag all baselines that do not meet this requirement.

Next, the visibilities are coherently averaged in time with a 214 s averaging window, having first phased the different time integrations to a common pointing center. Lastly, the instrumental XX and YY visibility polarizations are summed to construct a pseudo-Stokes *I* visibility as $V_I = (V_{\text{XX}} + V_{\text{YY}})/2$. Recall that many of these analysis steps are tested individually (Figure 3), but here we present the effects of these steps on the fully integrated power spectrum.

### 5.3. Power Spectrum Recovery

Power spectra are formed in the same manner as described in HC21. To summarize, we form delay spectra (Parsons et al. 2012b) in two spectral windows, which we refer to as Band 1 and Band 2, spanning 117–132 MHz and 150–168 MHz, respectively. Note that because we apply a Blackman–Harris apodization function across each spectral window, their equivalent bandwidth is half of the full bandwidth. Power spectra are formed by cross-multiplying every pair of nonidentical baselines within a redundant set. We then calculate two sets of error bars: a theoretical noise rms given the measured system temperature ($P_N$) and a semiempirical error bar that accounts for the signal-and-noise cross-terms in the power spectrum, $\tilde{P}_{\text{SN}}$ (see Table 3 of Tan et al. 2021). All incoherent averaging (i.e., averaging after forming the power spectra) is weighted inversely by $P_N^2$, but the final quoted error bars come from $\tilde{P}_{\text{SN}}$. The justification for using these error bars is outlined in the discussion (Section 5) of Tan et al. (2021). HC21 outline three broad LST ranges (or "fields") that are used for forming an averaged power spectrum. With the slightly smaller LST coverage studied in this work, we look at two similar fields, spanning 1.5–2.8 hr LST and 4.4–6.4 hr LST. Power spectra are formed from three data sets: the full data without systematics treatment, the full data with systematics treatment, and just the EoR component of the data.

The first check on our power spectra is to ensure that our noise estimates agree with the data at different stages of integration. This has recently been studied and validated for HERA simulations and real data (Kern et al. 2020b; Tan et al. 2021), but we repeat the exercise here for completeness. Figure 14 demonstrates the impact of incoherent averaging within a single redundant set. We show successive averages of redundant baselines with an increasing number of baselines in each average. We also plot the propagated thermal noise uncertainty $P_N$ (dashed), which agrees well with the power spectra outside of the foreground-dominated region for $k > 0.2\,h\,\text{Mpc}^{-1}$. Note that the final average (magenta) shows an increase in power at low $k$, which is foreshadowing a low-level detection of the EoR signal in the simulated data.

After averaging all baseline pairs within a redundant group, we average the remaining time bins within each LST range, leading to a cylindrically averaged $P(k_{\parallel}, k_{\perp})$ power spectrum per field per spectral window. Figure 15 shows this for Band 1 of the first LST range. The left panel shows the full data set, the middle panel shows the data after systematic treatment, and the right panel shows the EoR-only data set. In all panels, the gray dashed line shows the extent of the foreground wedge from the baseline horizon. We expect some amount of leakage beyond this line simply due to the side lobes of the apodization function applied before taking the visibility Fourier transform. What is apparent from Figure 15 are the systematics at

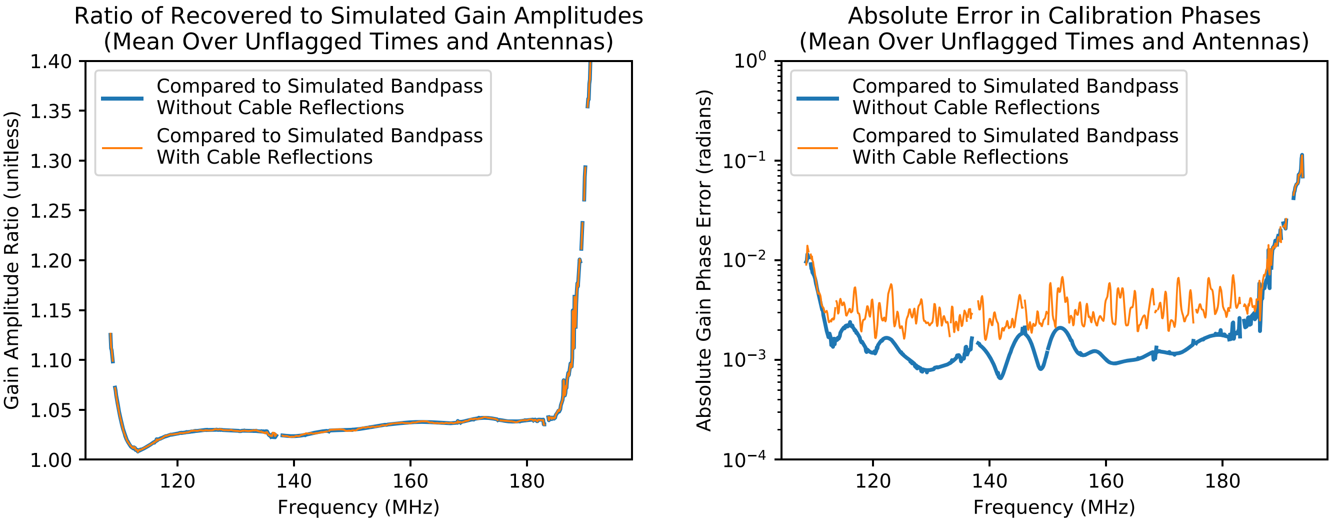

**Figure 11.** After the calibration pipeline, the inferred gain solutions are quite close to the true simulation gains, with amplitude errors at the few-percent level and phase errors at the few-milliradian level. Calibration errors are likely due to a range of factors including thermal noise, cross-talk systematics, and smoothing of features in the inferred calibration solution at the 100 ns delay scale. The latter explains both the spectral structure in the phase errors, which is largely contained to higher harmonics and the increased errors at the band edges, as the smoothing was performed with a Tukey window ($\alpha = 0.3$), which downweights discrepancies at high and low frequencies. The cable reflections, which dominate the true gains at high delay, are intentionally smoothed out of the calibration solutions and corrected only after LST binning. The smoothing out of real spectral structure from cable reflections produces the dominant phase calibration error at most frequencies but is subdominant to the few-percent-level amplitude bias seen in the left panel, which is due to a small bias in absolute calibration (see Figure 12). While this level of gain error and bias should be factored into a final power spectrum and errors, it is unlikely to produce substantial signal loss from decoherence.

$k_\parallel \sim 0.5\, h\,\mathrm{Mpc}^{-1}$ that are effectively suppressed by the modeling and removal step. This systematics treatment also suppresses power just beyond the foreground wedge. As discussed in Kern et al. (2020b), this is a result of the fact that foregrounds entering from the horizon (and thus lying near the wedge border in $k_\parallel$, $k_\perp$ space) are slowly time-variable and are therefore partially filtered off with the cross-coupling filter. So although Figure 15 becomes less wedge like with the application of the cross-coupling filter, this is due to filtering that impacts the edge of the wedge. The wedge will manifest on longer baselines (e.g., Kern et al. 2020b).

Lastly, we group the cylindrical power spectra in bins of constant $|k|$ and spherically average them to get our final 1D power spectra. Figure 16 shows these results for the first LST cut (top panels) and the second LST cut (bottom panels), for both Band 1 (left panels) and Band 2 (right panels). We plot the data before systematic treatment (blue points), after systematic treatment (orange) points with $2\sigma$ error bars, as well as the EoR-only data set (gray). Open circles denote negative band powers, which are plotted as positive for visual clarity. The subpanels show the data after systematic treatment divided by the EoR-only data set, with the $2\sigma$ error bars overlaid. Recall that the amplitude of the EoR signal was chosen specifically to allow for detection of the signal at low $k$, with its significance decreasing at higher $k$. The salient points we draw from Figure 16 are as follows: (1) the EoR signal is recovered to within the error bars across all $k$ modes,[37] and (2) the systematics at $k \sim 0.45\, h\,\mathrm{Mpc}^{-1}$ are suppressed down to the measured EoR amplitude. Importantly, the recovered power spectra (orange) match the EoR signal at low $k$ where we detect the signal at ∼10 times the noise floor and at high $k$ the power spectra are consistent with noise.

Using the data products at hand, we also perform additional tests targeted at particularly sensitive components of our analysis pipeline. One analysis step not quantified in the unbiased recovery seen in Figure 16 is the amount of loss induced by coherent time averaging (i.e., LST averaging or fringe-rate filtering). Over the course of a drift-scan observation, one can coherently average different time integrations that are closely spaced together relative to the overall beam-crossing time after rephasing them to a common pointing center. Using Monte Carlo simulations of an ensemble set of mock, $P(k) \propto k^0$ EoR observations, HC21 claim that they can coherently average their visibilities over a 528 s window and only induce ∼1% signal loss in the measured EoR power. We use the data products in this work (which, recall, use a $P(k) \propto k^{-2}$ EoR model) to confirm that this specification is met. Figure 17 shows this test, comparing the ratio of the EoR-only power spectra having first averaged the visibilities over a 528 s window over the power spectra with a 43 s averaging window. We show that, as expected, this induces a ∼1% loss in power that is constant across Fourier $k$ modes.

Another somewhat sensitive step in our analysis chain is the filtering of cross-coupling systematics, which is performed by applying a high-pass filter across the time axis. Recall that such a filter is designed to reject the slowly variable systematics while retaining the vast majority of the EoR sky signal (Parsons et al. 2016; Kern et al. 2019, 2020b). One complication to this is the impact of the time edges when working with finitely sampled data. Near the time edges, the properties of the sharp Fourier filter are degraded, and in our case we observe slightly more loss than the original specification (Kern et al. 2019). We can mitigate this effect by flagging the time bins near the edges after filtering. Figure 18 shows a demonstration of this on a data set that contains only the EoR signal and a cross-coupling systematic. We remove the systematic in the same way, but now in averaging the power spectra we flag all time bins within

[37] For the second LST cut there is an outlier at $k \sim 0.85\, h\,\mathrm{Mpc}^{-1}$ for both Band 1 and Band 2. It is odd that these outliers occur at the same Fourier mode, although an outlier or two is not entirely unexpected as the error bars plotted are $\pm 2\sigma$. At the very least, it is a high outlier, so concerns about signal loss are not an issue. More work is needed to understand if this is a statistical or systematic outlier.

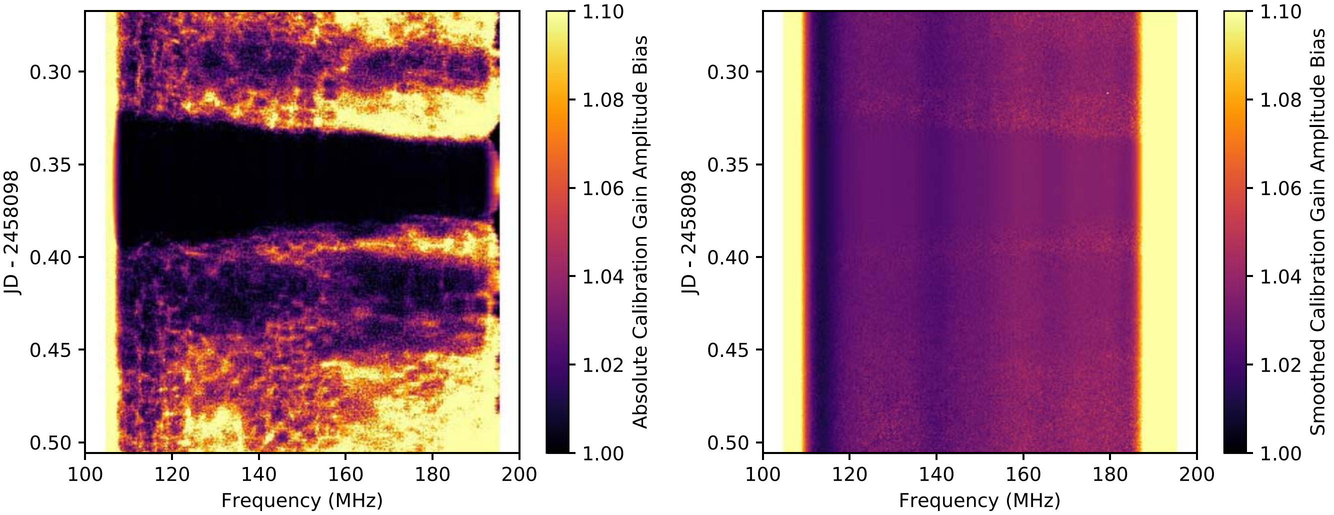

**Figure 12.** Here we show the bias in our amplitude calibration of a single simulated night (JD 2458098), averaged over all unflagged antennas. Because we are calibrating the overall amplitude of a noisy data set with a lower-noise set of "model" visibilities using a logarithmic linearization of the calibration equation, the time and frequency bins with low S/N return gain amplitudes that are biased high (Boonstra & van der Veen 2003). The opposite is observed at the transit of Fornax A around JD 2458098.36, which produces very high-S/N visibilities and suppresses the bias, leading to near-perfect gain amplitude recovery. The right panel shows the gain bias after frequency and time smoothing of the gain. We see that the bias is now effectively time independent but has a slight dependence on frequency, which is accounted for when forming power spectra over different parts of the band.

**Table 3**
Systematic Loss in Analysis

| Analysis Step | Fractional Bias |
|---|---|
| Absolute Calibration | −11% (−15%) |
| Cross-coupling Filtering | −3% (−1%) |
| LST Time Averaging | −1% (−1%) |

**Note.** Percentage loss in power for Band 1 (Band 2), which is corrected for after forming the power spectrum and is constant for all $k$ modes. Redundant-baseline averaging is also explored in HC20 as a possible source of percent-level loss, but is not studied in this work.

some buffer of the time edge, showing that the amount of loss converges with a 30 minute buffer. In the end, this results in a residual 3% (1%) scale-independent loss in power after cross-coupling subtraction for Band 1 (Band 2).

All of the steps discussed in this work that were found to lead to loss are summarized in Table 3, most of which are on the order of a percent. The bias discussed in Section 5.2 is not technically signal loss in the traditional sense but is still a bias that results in an underreporting of the EoR signal, therefore we include it in this table. Note that HC21 also explore the impact of coherent baseline averaging within a redundant group, which can in principle lead to signal loss. We do not explore this currently as baseline nonredundancy is not within the intended scope of this work, but future work will incorporate this aspect into the validation pipeline presented here.

Lastly, another metric we can pin down with our validation simulations is the expected level of cosmic variance on the EoR signal after all of our coherent and incoherent averaging. Lanman & Pober (2019) quantify this in a similar manner using Monte Carlo simulations of a mock-EoR field and find that for a HERA-37 spherical power spectrum averaged over 8 hr LST the cosmic variance (1$\sigma$) peaks at around 2% of fractional power. Pushing our own EoR simulation through the analysis pipeline discussed in this work and averaging the power spectra over the second LST range (2 hr spanning 4.4–6.4 hr LST), we find a fractional (1$\sigma$) cosmic variance uncertainty of ∼5.5% for both Band 1 and Band 2 (Figure 19). As discussed in Tan et al. (2021), this is currently a subdominant contributor to the total error budget.

### 5.4. Blind Test with Parallel Pipeline

To double-check that our primary power spectrum analysis pipeline indeed induces minimal signal loss and makes a clear detection of the input 21 cm power spectrum, we performed a blind analysis of the mock data with an alternate power spectrum estimator.[38]

Parallel, or shadow, analysis is a powerful validation technique that has been adopted for several published results (Jacobs et al. 2016; Barry et al. 2019a; Trott et al. 2020). These analyses are expensive in both researcher and computer time and thus are often limited in the amount of data processed in parallel and may share common preprocessing steps. Nevertheless, they provide some measure of confidence that the reported result is not unique to a particular analysis. Errors made in the absence of such testing have commonly been associated with power spectrum estimation (Paciga et al. 2013; Cheng et al. 2018), so this is where we choose to focus our efforts. We perform a parallel power spectrum analysis of the calibrated and LST averaged simulation product using the SIMPLEDS pipeline (Kolopanis et al. 2019), verifying that it also reproduces the expected result.

Our shadow analysis followed the procedure described in Kolopanis et al. (2019). Power spectra were formed by cross-multiplying redundant baselines and errors estimated by calculating the expected sensitivity according to Pober et al. (2014), simulating noise using the autocorrelations as a measure of variance and bootstrapping across the many possible pairs of baselines. As this last step is computationally expensive, scaling with the amount of $uv$ space analyzed, we

[38] Use of the word blind here might not be preferred by blind people;however, the term is used pervasively in science to describe a technical procedure that is not well served by synonyms. We keep the term for now to avoid confusion as the practice is introduced to the field of 21 cm cosmology.

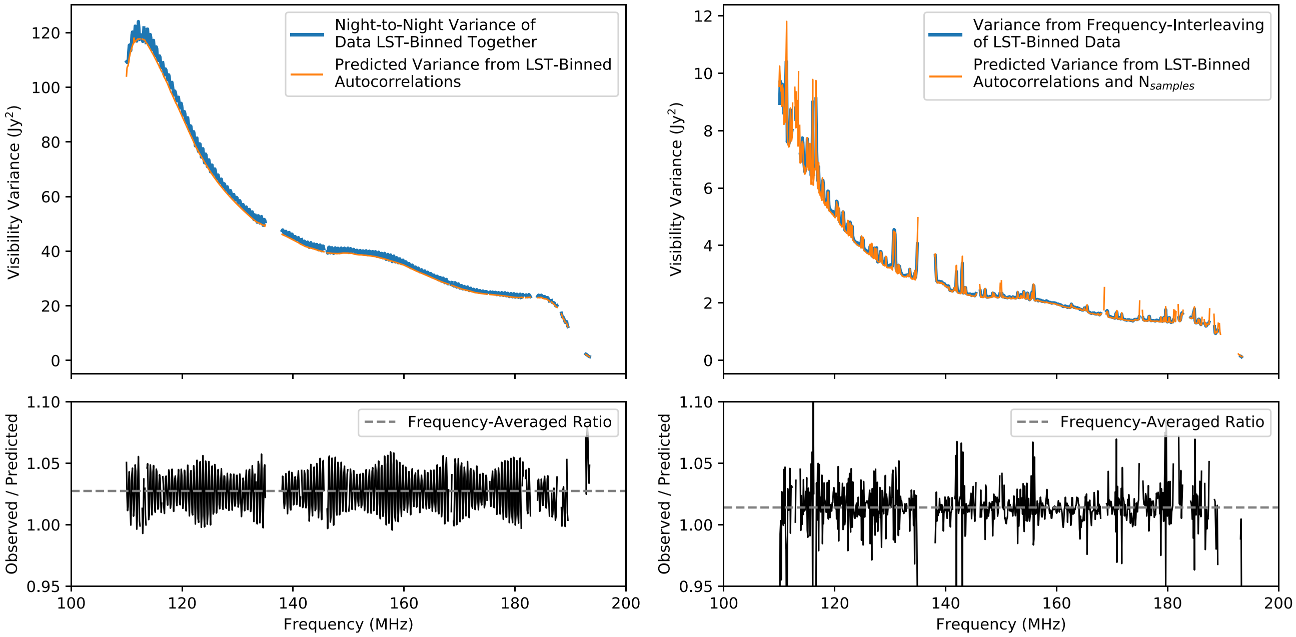

**Figure 13.** Confirmation that noise in LST-binned visibilities matches expectations. Left panel: night-to-night variance over 10 nights (averaged over unflagged baselines and times in the LST range of 6.464–6.817 hr) compared against noise predicted by LST-binned autocorrelations (see Equation (18)). Right panel: variance calculated from the same data using frequency differencing compared to normalized predicted noise from autocorrelations and $N_{\rm samples}$. In both panels, we drop any time or frequencies with $N_{\rm samples} < 10$ before averaging in order to account for RFI gaps. Both metrics indicate a close match with the predictions.

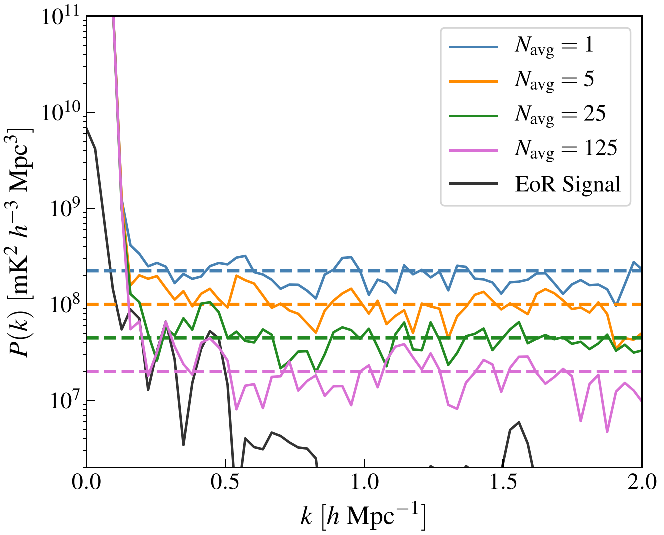

**Figure 14.** Power spectra from Band 1 after successive incoherent redundant-baseline averages. In each case, we plot the averaged power spectra (solid) and their corresponding $P_N$ given the amount of averaging (dashed), which marks the noise amplitude if the power spectra are noise dominated. In each case we see good agreement between the two for $k > 0.2\ h\ {\rm Mpc}^{-1}$, except for the final case where at low $k$ we begin to detect a signal; note the black line, which is the fully integrated spectrum for this baseline type.

limited our baseline selection to three vectors of length ∼28 m, of differing orientation. These are the shortest baselines included in the mainline analysis. This restriction to a narrow *uv* range is the largest divergence from the main power spectrum processing.

As the alternate pipeline was not developed in close proximity to the simulation, we were afforded an opportunity to trial blind testing. A blind test using realistic simulations provides an opportunity to test our judgment in identifying whether data points that are not noise dominated arise from foregrounds, systematics, or true 21 cm signal.

A small subgroup, disconnected from the main Validation team and blind to the preparation of the mock data, was set a challenge in which they were to distinguish between two data sets that were the same in every respect, except that one had 21 cm signal and the other did not. These simulation products were blinded by changing filenames and removing metadata and provided to the shadow-pipeline team after the “Coherent Time Averaging” step (see Figure 2).

Figure 20 summarizes the results. In the first analysis, no cosmological signal could be clearly identified in the data (see left panel of Figure 20). Residuals were strong enough to make all data sets look roughly the same. Having finalized and reported this blind result to the rest of the Validation team, a meeting was held in which the topics discussed were intentionally limited to a comparison of data selection between pipelines and some clarification of the meaning of certain metadata. Importantly, the form and amplitude of the 21 cm signal were kept hidden. Each change discussed during these conversations was recorded and tested one at a time. The final resulting power spectrum estimate, obtained as a result of these limited discussions, is shown in the right column of Figure 20. This figure shows clear improvement over the fully blinded analysis shown in the left column and indeed confirms that the alternate pipeline is able to accurately detect the input signal and differentiate that detection from data without the signal.

The largest improvement between the left and right panels of Figure 20 came from correctly interpreting sample-count metadata. The main analysis pipeline assigns flagged channels a sample count of zero but then in-paints some of these channels (as described in Section 4.4). These in-painted data points are meant to be used when computing the delay spectra but not to contribute toward the estimation of noise. The alternate pipeline was erroneously reflagging these channels,

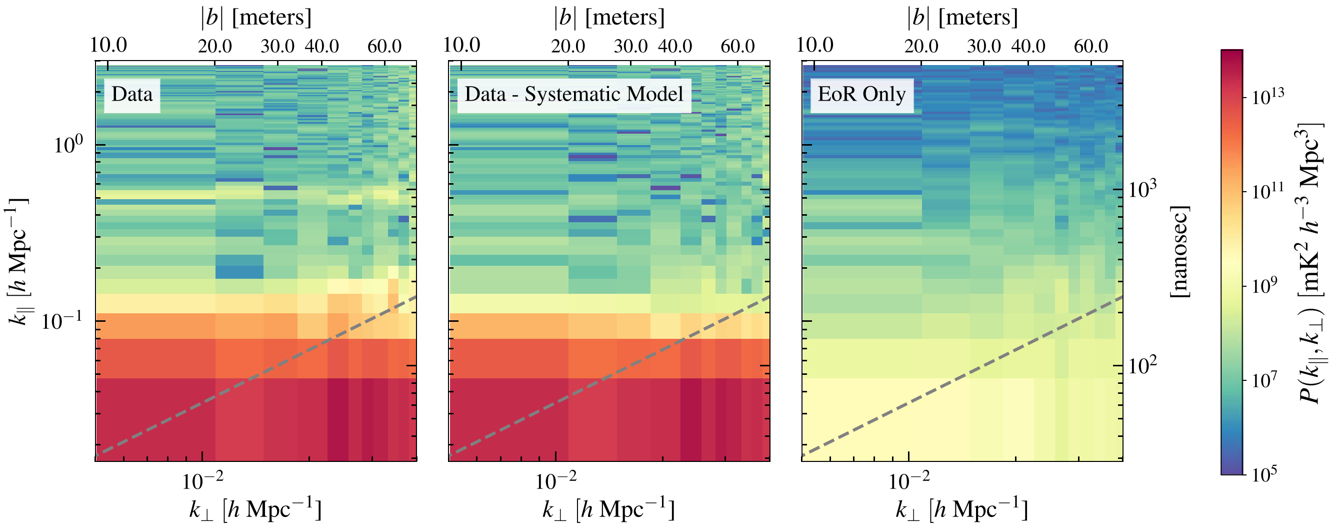

**Figure 15.** Two-dimensional delay spectra for the end-to-end test (step 4). The left panel shows the result having skipped the systematic subtraction step (and thus contains all of the extra instrumental systematics injected into the data), the center panel shows the full end-to-end run with systematic subtraction, and the right panel shows the EoR-only data set. The gray dashed line marks the foreground horizon (i.e., the wedge).

due to a misunderstanding of the intention of the sample count. With this misinterpretation, the flagged channels produced a large level of "ringing" when Fourier transformed, appearing as anomalous power in the estimated power spectrum.

Despite the aforementioned meeting to obtain clarification on the metadata, some differences between the pipelines were intentionally maintained. The single biggest difference between the pipelines was the selection of the LST range, which was smaller for the shadow pipeline (about an hour shorter). Furthermore, the shadow analysis averaged both fields together. Clearly, these differences do not significantly affect the conclusions of the test. Indeed, they further strengthen the case that the analysis is not highly sensitive to the precise choice of LST range.

This test increases confidence in the power spectrum portion of the analysis, reinforces the conclusion that loss within the calibration pipeline is minimal, and provides a guide for how blind comparison between simulation and data can be employed to assess the relative likelihood that an observed power level is due to a true background.

### *5.5. Accuracy of Error Bars*

The level of the injected EoR power spectrum is such that there are four regimes at the final noise level (integrating all times and baselines): foreground dominated for $k < 0.2$, EoR dominated for $0.2 < k < 0.4$, systematics dominated (before subtraction) for $0.4 < k < 0.55$, and noise dominated for $k > 0.6$. Consequently, we can assess the consistency of the recovered data points with the error bars in a manner similar to that of HC21 (Equation (26) and Table 5). The null hypothesis here is that the data points are consistent with zero, given the error bars reported. Performing this test, we find a significant detection ($p < 0.001$) in all bands and fields when including all $k > 0.2$, consistent with detections of the injected EoR. For $k > 1$, all bands and fields had a $p$-value consistent with the null hypothesis, except for Band 2, Field 2, which has two $2\sigma$ outliers at the highest $k$. The imaginary component of the power spectrum is consistent with the null hypothesis for all $k > 0.2$ and all bands and fields.

## 6. Discussion and Conclusions

### *6.1. Basic Conclusions*

In general, we have found that the HERA H1C software pipeline successfully reproduces known analytic input power spectra, under the assumptions it adopts; we did not find major issues with any of the pipeline steps we investigated here.

We performed power spectrum estimation with a full end-to-end mock data set including a wide range of realistic instrumental effects and foregrounds (see Section 5). In this test, mock visibility data were self-consistently generated from a known analytic power spectrum (see Section 4.1.1 and Section 4.2), obscured with realistic Galactic and extragalactic foreground models (see Section 4.1.2) and contaminated with almost all known instrumental effects relevant to the HERA instrument (see Section 4.3). A summary of the included components can be found in Figure 3 and the top panel of Table 4. These mock data, simulated to be broadly consistent with H1C observing parameters, were passed through the full H1C analysis and power spectrum estimation pipeline, with all analysis parameters consistent with those used for processing actual data (see HC21). As our primary result, we demonstrated that the pipeline produces power spectrum estimates that are consistent with the known analytic input to within thermal noise levels (at the $2\sigma$ level) for $k > 0.2\,h\,\mathrm{Mpc}^{-1}$ for both bands and fields considered (see Figure 16).

To test the pipeline in various regimes in which different components dominate, the analytic input spectrum was intentionally amplified to enable a strong "detection" at $k \sim 0.2\,h\,\mathrm{Mpc}^{-1}$—at the level of $\sim 25\sigma$—with foregrounds dominating on larger scales, thermal noise dominating at smaller scales, and systematics dominating (before subtraction) in between. The pipeline successfully detected this amplified input signal, after suppressing foregrounds with a dynamic range (foreground to noise ratio) of $\gtrsim 10^7$. Additionally, the noise-dominated power spectrum at high $k$ was found to be consistent with the predicted noise power. With the possible exception of a single $k$-bin in "Band 1, Field 2", systematics were mitigated to below the noise level of the simulation.

This does not guarantee that there are no inaccuracies remaining, but we can be confident that any are unlikely to

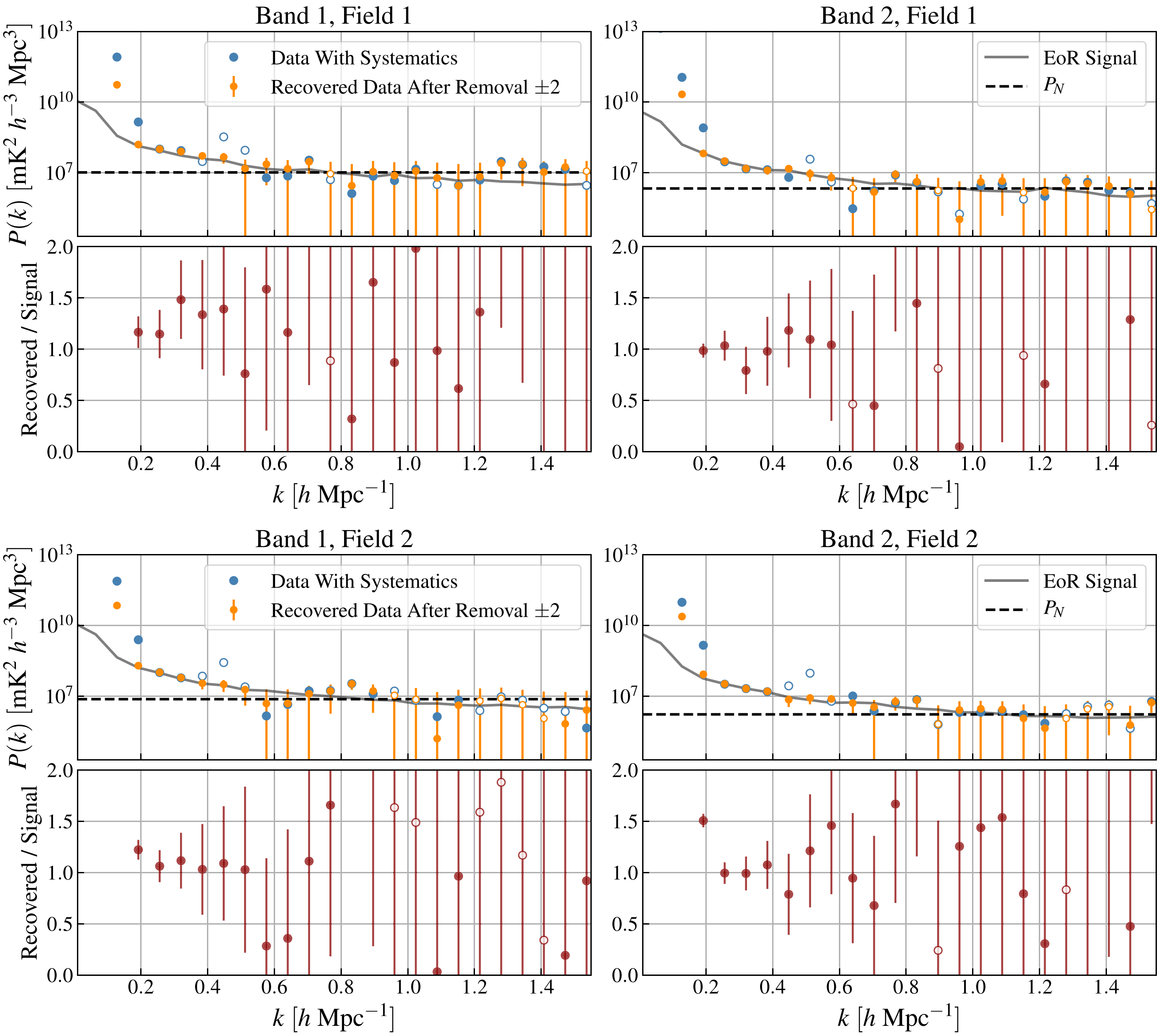

**Figure 16.** Recovered power spectra of the end-to-end test (step 4). We plot the data before systematic treatment (blue), after systematic treatment (orange) with its 2$\sigma$ error bars, as well as the intrinsic EoR signal (gray) and the noise floor (black dashed). The top panel shows the first LST cut (1.5–2.8 hr) and the bottom panel the second LST cut (4.4–6.4 hr). The subpanels plot the ratio of the recovered (orange) over the EoR (gray), showing unbiased recovery of the intrinsic EoR signal at all $k$ to within the estimated error bars.

have major effects on the HC21 results. Recall that the goal of this effort was to validate the *software and algorithms*, not the *data*. Thus, there may yet be subtle effects present in the real data, which we did not adequately represent in the simulation, or analysis choices that do not perform correctly when the assumptions of the pipeline are violated. However, substantial issues like those found in Cheng et al. (2018), which were the result of the algorithm, independent of the data, seem unlikely.

A number of small problems and unanticipated effects were discovered (e.g., those given in Table 3) and these have led either to improvements in the existing pipeline which eliminate them, or inclusion in a list of effects to continue investigating.

### 6.2. Scope of Future Work

As the HERA pipeline improves and changes, the validation effort will need to continue to include simulations that effectively test the new software and challenge the assumptions made. There are some obvious axes along which the validation effort will need to be extended or modified for future work; we briefly list some of them here.

In this work, we have compared portions of the analysis between two pipelines: the HERA standard one and `simpleDS`. This comparison was done nearly “blind,” i.e., the group analyzing using `simpleDS` did not know anything about the data sets that were prepared for it and analyzed the data as if they were from the real instrument. Both aspects of this cross-check should be kept in future validation efforts, namely the existence of a parallel pipeline and the independent, blind analysis of the simulated data sets. This turns up both differences due to the different algorithms but is also revealing of different implicit assumptions in the analyses.

A more complete simulation of RFI and the effectiveness of the flagging is clearly essential as the limits get deeper to

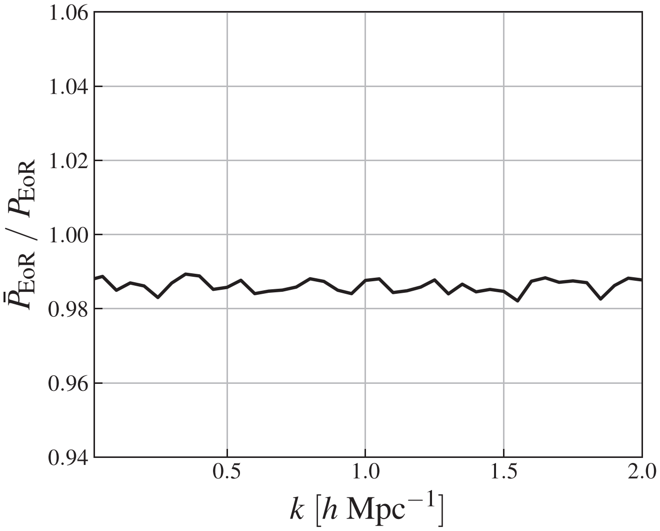

**Figure 17.** Signal-loss test for the LST averaging step in the H1C analysis pipeline. This measures the amount of signal loss induced by coherent averaging of the visibilities across the LST. The numerator of the ratio is the power spectrum of the EoR-only visibilities having been averaged over a 7 minute window, while the denominator is the same data product with a time averaging window of only 43 seconds. This step induces a ∼1% signal loss, which is deemed negligible compared to other limiting uncertainties. This result has been verified against different visibility simulators and different EoR models (Kern et al. 2019).

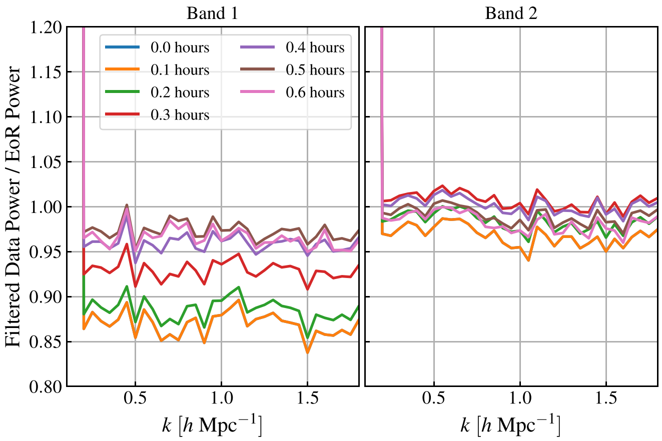

**Figure 18.** A measurement of the amount of loss induced by the cross-coupling high-pass time filter as a function of how much of the data near the edges of our time axis we flag. Because of edge effects, the cross-coupling filter leads to more loss for time bins near the bounds of our time axis. We show here that by flagging 30 minutes on either side, we minimize this loss, with a residual loss of 3% (1%) in power for Band 1 (Band 2).

ensure that there is no significant effect due to unflagged RFI. In part, this requires devising and implementing a suite of null tests on the real data, because the complexity of actual RFI will probably always exceed our ability to simulate it. Nevertheless, it should be possible to gain considerably more insight via a detailed simulation step into how well our current algorithms are doing and whether there are likely gaps in their effectiveness (The HERA Collaboration & Wilenksy 2019).

The simulations here have several aspects that are specific to the instrument configuration. An entirely new feed system is currently being commissioned (de Lera Acedo et al. 2020; Fagnoni et al. 2021a), which will necessitate a new set of investigations of systematic effects, new simulations of them, and tests of the methods proposed to correct or mitigate them.

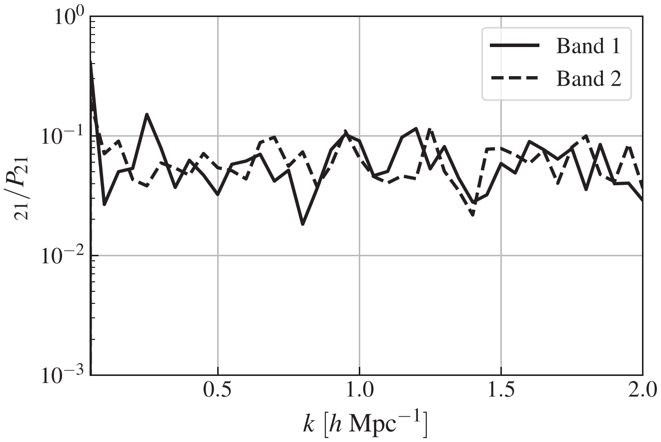

**Figure 19.** The fractional (1$\sigma$) cosmic variance uncertainty on the power spectrum for 2 hr of drift-scan observations with HERA-37 is on average ∼5.5%. This is computed by taking the rms of the EoR-only power spectra across the 2 hr time range and then dividing by the effective degrees of freedom set by the HERA beam-crossing time of 1 hr.

The continued improvement in our understanding of foregrounds, including both better point-source catalogs and better models of diffuse emission, particularly below 100 MHz, will be folded into future validation simulations. The foreground models should also be broadened to include Faraday-rotated polarized emission, to simulate the effects of polarized leakage that could be comparable to the level of EoR, especially for the delay spectrum approach, which does not segregate the polarized signal via an image-based analysis (Nunhokee et al. 2017; Asad et al. 2016, 2018). The effects of the ionosphere may also be important, particularly its interaction with the polarized foregrounds (Martinot et al. 2018).

Another complication not directly addressed by our analysis is the impact of an incomplete or incorrect sky and/or beam model on the HERA absolute calibration step. In addition to the flux-scale issues (quantified in Kern et al. 2020a and accounted for in HC21), this effect can introduce spurious spectral structure into the calibration solutions (Barry et al. 2016; Byrne et al. 2019), which can be mitigated by including only short baselines (Ewall-Wice et al. 2017; Orosz et al. 2019) in the calibration. Because the reference visibilities used in absolute calibration differ little from the true visibilities (they are filtered at high delay as explained in Section (2.2)), the impact of sky-model error cannot be quantified here. However, because calibration solutions are smoothed spectrally at delays larger than 100 ns, we avoid spectral structure from modeling error by simply not trying to calibrate any true spectral structure in the instrument response beyond 100 ns unless it can be modeled as the sum of reflections and inferred from the autocorrelations. While this smoothing was primarily motivated by the desire to mitigate the effect of cross-coupling on the gains (Kern et al. 2020a), it makes the spectral impact of modeling error largely irrelevant for this analysis. If additional degrees of freedom are admitted in the calibration solutions in the future (e.g., by increasing the delay threshold for smoothing), this question needs to be revisited.

Such considerations also point to the complex open question of how to simulate the effects of violations of the assumptions of the analysis, particularly with regard to incomplete knowledge of the primary beams of the antennas and failure of redundancy between nominally redundant baselines for various reasons. The ability to simulate a different primary beam for each antenna is

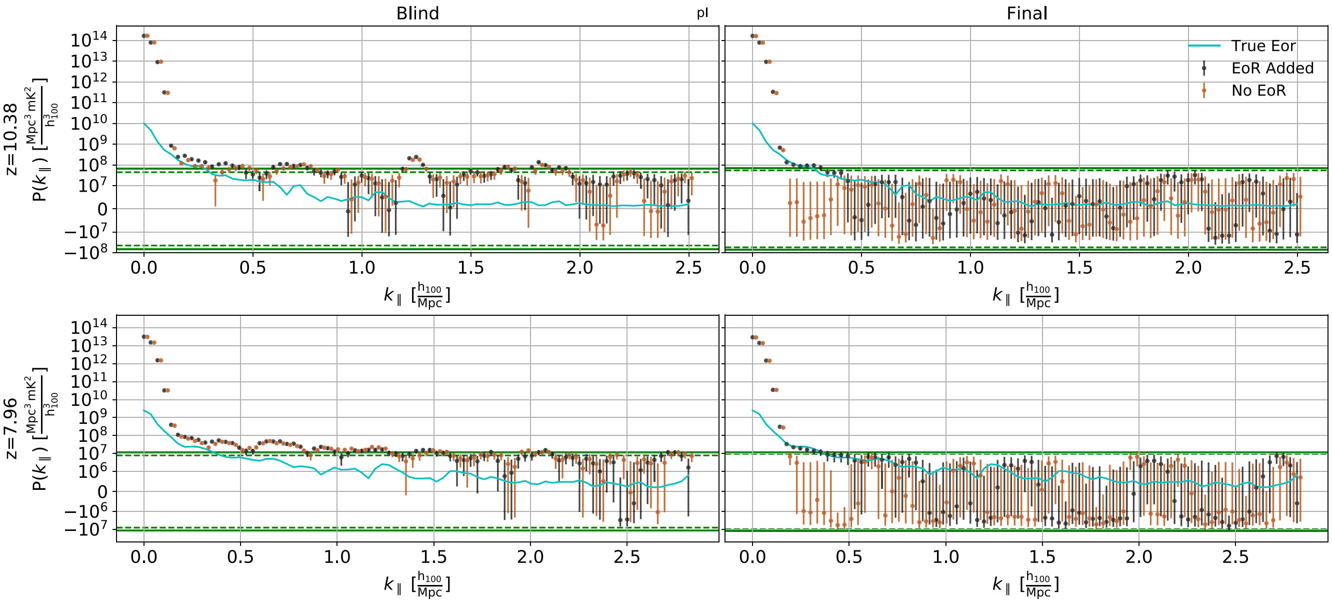

**Figure 20.** Two blinded simulations were processed using an independently developed shadow pipeline (simpleDS; Kolopanis et al. 2019). One contained a detectable 21 cm signal. The initial result by a completely blinded team (left) was dominated by a strong systematic common to both simulations. After a carefully limited discussion between the simulation team and the shadow group about weighting and data selection, but not unblinding the files, a strong distinction emerged (right), which the group interpreted as evidence for a (simulated) 21 cm signal. This was confirmed in unblinding by adding the expected 21 cm signal. Green lines show theoretically predicted noise power ($P_N$) at the $1\sigma$ and $2\sigma$ levels.

**Table 4**
Effects Included and Not Included in This Analysis

| |
|---|
| **SIMULATED EFFECTS** |
| Known bright point sources |
| Point-source foregrounds from the GLEAM catalog |
| Diffuse foregrounds on scales $>3°$ |
| Sky-derived thermal noise |
| Simple receiver noise |
| Realistic EM-simulated beam |
| Direction-independent gains |
| (per feed, time invariant) |
| Cross-coupling model |
| Per-antenna cable reflections |
| Realistic flagging patterns |
| **IGNORED EFFECTS (MAJOR)** |
| Misidentification of dysfunctional antennas |
| Misidentification of RFI |
| Generation of abscal model from images |
| Antenna nonredundancy (in primary beams) |
| Fully realistic antenna cross-coupling |
| Antenna position errors |
| **IGNORED EFFECTS (MINOR)** |
| Digital (correlator) artifacts |
| Confused point sources |
| Fully polarized direction-independent gains |
| Time variation in direction-independent gains |
| (due to, e.g., temperature variations) |
| Polarized and/or transient sources |
| Full suite of possible shapes for $P_{\mathrm{EoR}}(k)$ |
| Ionosphere |

included in our simulation packages (RIMEz and pyuvsim), but how to represent realistic variations is currently a topic of active research (e.g., Choudhuri et al. 2021). Another example of this is simulating time-variable gains that capture the actual instrument behavior.

Another consideration not addressed in this validation suite, but firmly in place for future tests, is the applicability of the pipeline to markedly different (but physically reasonable) shapes for $P_{\mathrm{eor}}(k)$. For instance, one might imagine that a "sharp" feature in $P_{\mathrm{eor}}(k)$ might cause difficulties for power spectrum estimation. The interaction of the window functions with the power spectrum also needs to be considered more carefully.

Finally, the end-to-end approach here only considered a simulation of a single data set of approximately the same size as the actual one. As the HERA data grow, this will be an increasingly difficult task, to say nothing of the need for exploring errors via multiple realizations of noise, systematic effects, and cosmological signal. In particular, our criteria for what constitutes a successful end-to-end test will need to be more rigorously tied to keeping systematic errors from the analysis to less than the random errors due to instrument noise (and its coupling to signal), combined with the expected cosmic variance (e.g., Lanman & Pober 2019). This will also have an effect on how we assess the errors on tests of portions of the pipeline (the "steps" in Figure 3). We will need to investigate further which aspects of the pipeline truly require a simulated data set comparable to the full one and which require multiple realizations to understand the statistical effects. This is particularly important with respect to systematic effects whose exact parameterization is difficult to quantify (e.g., primary beam nonredundancy).

This material is based upon work supported by the National Science Foundation under grant Nos. 1636646 and 1836019 and institutional support from the HERA collaboration partners. This research is funded in part by the Gordon and Betty Moore Foundation. HERA is hosted by the South African Radio Astronomy Observatory, which is a facility of the

National Research Foundation, an agency of the Department of Science and Innovation. Parts of this research were supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. J.E.A. acknowledges funding from NSF CAREER grant AST-1455151. G.B. acknowledges funding from the INAF PRIN-SKA 2017 project 1.05.01.88.04 (FORECaST), support from the Ministero degli Affari Esteri della Cooperazione Internazionale—Direzione Generale per la Promozione del Sistema Paese Progetto di Grande Rilevanza ZA18GR02 and the National Research Foundation of South Africa (grant No. 113121) as part of the ISARP RADIOSKY2020 Joint Research Scheme, from the Royal Society and the Newton Fund under grant NA150184 and from the National Research Foundation of South Africa (grant No. 103424). P.B. acknowledges funding for part of this research from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 948764) and from STFC grant ST/T000341/1. J.S.D. gratefully acknowledges the support of the NSF AAPF award #1701536. N.K. acknowledges support from the MIT Pappalardo fellowship. A. Liu acknowledges support from the New Frontiers in Research Fund Exploration grant program, the Canadian Institute for Advanced Research (CIFAR) Azrieli Global Scholars program, a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Discovery Launch Supplement, the Sloan Research Fellowship, and the William Dawson Scholarship at McGill. M.G.S. acknowledges support from the South African Square Kilometre Array Project and National Research Foundation (grant No. 84156).

## Contributions of the Authors

J.E.A. leads the HERA Validation group and was responsible for the overall direction and organization of the validation effort presented here. S.G.M. created and curates the `hera-validation` repository, directed validation of the visibility simulators (Step −1.1), and ran the foreground tests (Steps 1.1 and 1.2). R.P. created the end-to-end simulations and ran the Step 0.1 and 2.1 tests. Z.E.M. wrote `RIMEz` and `gcfg` and created the base visibility simulation components and ran the Step 0.2 test. J.B. tested the in-painting and systematic subtraction in Step 3.1. J.S.D. ran the calibration simulations/pipelines and their tests (Steps 2.0 and 4). M.K. and D.C. J. analyzed the blind test. D.C.J. wrote the results of the blind test and outside simulator validation. NSK ran the power spectrum pipeline on the end-to-end simulations and performed the signal loss tests (Section 5). L.W. ran the foreground tests (Step 1.0) and the `RIMEz` validation (Step −1.1). The HERA Builder's list is included alphabetically as authors because of the dependence of this work on the combined efforts of collaborators on the various software repositories as well as the necessity of using HERA data to build the models used in this paper.

## Appendix A
## The HERA Validation Subsystem

The HERA collaboration has placed a high emphasis on detailed validation by establishing a dedicated Validation team, formalized as an essential HERA subsystem. HERA "subsystems" are the major components of the HERA experiment and have a dedicated team associated with each. In addition to the "Validation" subsystem, others include "Power Spectrum Estimation," "Analysis," "Quality Metrics," and "Inclusion/Diversity."

The scope of this effort is clearly wide ranging: ultimately it is to verify that the reported power spectra from the HERA collaboration are free from defects, whether from code bugs, poor algorithmic choices, or inappropriate physical assumptions. At the same time, the goal is not to merely internally validate, but also to ensure that the pipeline is reproducible and understandable by the wider community, in order to build confidence in reported upper limits or detections.

### *A.1. Code Standards*

HERA has adopted a set of high-standard open-source software practices that encourage transparency, reproducibility, interoperability, and peer-verification. All systems-level HERA code is hosted open-source on a single GitHub organization.[39] A set of well-defined software standards is applicable across the organization, encouraging a certain degree of homogeneity between project-level packages. Among these standards are

1. Documentation: Python code is self-documented (i.e., includes "docstrings"[40] for all public modules, functions, classes, and methods), using a uniform docstring format (typically NUMPYDOC). Extra tutorials and examples are also encouraged.
2. Testing: all systems-level HERA packages are thoroughly unit-tested,[41] and kept at >95% code coverage.[42] Testing is performed continuously via an online Continuous Integration provider (e.g., Travis or Github Actions).
3. Formatting: all code is PEP8 compliant[43] (often enforced by the use of external tools such as BLACK[44] PRE-COMMIT[45]), making each package more homogeneous (important when there are many contributors to the repository) and easy to read. This is important for transparency both within and without the collaboration.
4. Review: each package uses the GitHub flow[46] as a software delivery workflow. In brief, in this workflow the "master"[47] branch is considered protected and is disabled for direct code changes on GitHub. This requires new code additions (and bug fixes) to be developed in a branch that is "not master" and a formal "pull request" (PR) to be created and accepted before merging back into the protected "master" branch. All repositories have an option enabled in which PRs must be first reviewed and

[39] https://github.com/hera-team—note that not all repositories found here are considered "systems-level".

[40] Docstrings are a Python construct for documenting code objects in place in the code and can be used to automatically create up-to-date online documentation.

[41] Unit tests are functions that assert specific conditions on the behavior of the basic units of the software (e.g., functions or class methods) and can be collected and run together in an automated fashion. This is in contrast to integration tests, which assert conditional behavior of combinations of the basic units.

[42] Code coverage represents the percentage of standard lines of code in the package that are run during the execution of the test suite.

[43] https://www.python.org/dev/peps/pep-0008/

[44] https://black.readthedocs.io

[45] https://pre-commit.com/

[46] https://guides.github.com/introduction/flow/

[47] In the near future, the `master` branch will be renamed to `main`, as is now widely endorsed.

accepted by a person other than the author before they can be merged. PRs must also satisfy a host of other status checks, such as passing Continuous Integration tests and satisfying coverage checks. Such reviews lessen the probability that subtle bugs enter the code (especially those that are only apparent when one has familiarity with a different part of the code), but also serve to increase the overall familiarity with the code base, as it evolves, of the wider collaboration.

### *A.2. The Validation Code Repository*

Following the lead of the wider collaboration, the HERA Validation team has established a public repository in which all pipeline validation tests are performed and archived.[48] We have defined a comprehensive set of tests of the pipeline, moving from simplistic analyses through to a full end-to-end simulation and analysis (see Section 3.2). Each of these tests is performed and documented in a Jupyter notebook,[49] developed and archived in our GitHub repository. Jupyter notebooks allow combining arbitrary documentation and code execution in order to generate a full analysis. We utilize this ability, adopting a certain template for each test that includes listing the full provenance of all data used, the exact package versions of all dependent software (down to the git hash), a summary description, a set of criteria to meet for the test to pass, and a list of suggestions for follow-up tests. These sections promote reproducibility and clarity.

Each test is recorded via a three-digit identifier: `major.minor.test`, in which the `major` digit identifies a broad class of physical effects being tested, the `minor` digit identifies variations on that class of physical effects, and the `test` digit represents an iteration in the testing procedure (e.g., a test may fail and require rerunning with a bugfix, or with a slight alteration in the assumptions). Each notebook contains a single test. Although all tests are version-controlled, we do *not* overwrite test notebooks when an updated test is performed. The failed or outdated tests are kept at the top level of the repository to make it easy to determine the history of the test.[50]

In keeping with the standards of the rest of the collaboration, validation tests are required to be reviewed and accepted by the rest of the group before being merged into the master branch.

Extra features of GitHub have also been used to aid in the organization of the Validation effort. In particular, newly proposed tests are created as GitHub issues, where they are discussed before accepting them into the test-suite canon. A set of custom tags has been specified, explicitly defining each simulation and analysis component the test would validate (see Figure 3).

This system has served well in this particular validation effort and will continue to be used to develop further validation tests for upcoming data releases.

## Appendix B
## Window Functions and Aliasing

This appendix seeks to explain the discrepancy between the analytic input and estimated power spectrum and present the definition of the "aliased" power spectrum in Figure 4 and why it is much closer to the estimated power spectrum.

In general the power spectrum estimates $\widehat{P}(k)$ produced with `hera_pspec` in this paper can be described by

$$\widehat{P}(k) = \sum_{\alpha}\sum_{\beta} E(k, \alpha, \beta) V(\alpha) V^{*}(\beta), \tag{B1}$$

where $V$ is the visibility function and $\alpha$, $\beta$ are indices over the set of points $\alpha = (t, \nu, \vec{\boldsymbol{b}})$ at which the visibility function is measured. Let the visibility $V$ be sourced by only the cosmological signal, as in the simulation that produces Figure 4. The covariance matrix of the data is then a linear functional of the power spectrum, which may be written

$$\langle V(\alpha) V^{*}(\beta)\rangle = \int_{0}^{\infty} \frac{\partial\langle V(\alpha) V^{*}(\beta)\rangle}{\partial P(k)} P(k) dk. \tag{B2}$$

The expectation value of the power spectrum estimate is thus

$$\langle\widehat{P}(k)\rangle = \sum_{\alpha\beta} E(k, \alpha, \beta)\langle V(\alpha) V^{*}(\beta)\rangle \tag{B3}$$

$$\langle = \int_{0}^{\infty} \sum_{\alpha\beta} E(k, \alpha, \beta) \frac{\partial\langle V(\alpha) V^{*}(\beta)\rangle}{\partial P(k')} P(k') dk'. \tag{B4}$$

We define the "window function"

$$\langle W(k, k') \equiv \sum_{\alpha\beta} E(k, \alpha, \beta) \frac{\partial\langle V(\alpha) V^{*}(\beta)\rangle}{\partial P(k')}, \tag{B5}$$

and hence an unbiased estimator for the power spectrum—i.e., an estimator such that

$$\langle\langle\widehat{P}(k)\rangle = P(k) \tag{B6}$$

–would be one such that the window function is

$$\langle\sum_{\alpha\beta} E(k, \alpha, \beta) \frac{\partial\langle V(\alpha) V^{*}(\beta)\rangle}{\partial P(k')} = \delta(k - k'). \tag{B7}$$

With a countable number of samples $\alpha$ and finite bandwidths of our measurements, it is not possible to achieve such a window function exactly—any real measurement will be "corrupted" by a window function $W(k, k')$ with a finite width. The best that can be accomplished is that the window function is localized so that the estimate of $\widehat{P}(k)$ has contributions from only $k'$ nearby to $k$. The delay spectrum estimator applied in Figure 4 has the form

$$\langle E(k, t, \nu_n, \vec{\boldsymbol{b}}_i, t', \nu_m, \vec{\boldsymbol{b}}_j) = \mathcal{N} w_n w_m \exp(i\mathcal{A}k(n - m))\delta_{t,t'}\delta_{ij} \tag{B8}$$

(i.e., $\mathcal{N}$ is the delay spectrum normalization scalar and $\mathcal{A}$ is the frequency band dependent delay conversion factor), which produces a localized window function when $w_n$ is a frequency taper like the Blackmann-Harris used in our power spectrum estimates.

It is this effect that causes the evident discrepancy between the input and estimated power spectrum in Figure 4. In the low-$k$ regime ($k \sim 0.1$) the window function has an approximately constant width (in linear units of $k$), except for the lowest several $k$ points. Each power spectrum estimate is an integral over the true power spectrum within that constant width, but as the estimated $k$ point decreases toward low $k$, the intrinsic power-law power spectrum increasingly varies over the width

[48] https://github.com/hera-team/hera-validation

[49] https://jupyter.org

[50] Jupyter notebooks are also not particularly well-suited for granular version control.

of the window function. This causes the estimate to be increasingly biased high with respect to the central value of the analytic input. However, at the lowest two or three $k$ points, the window functions become much less well behaved, i.e., more oscillatory and less localized, which produces the dip in the lowest $k$ point in Figure 4.

At high $k$ the effect of the window function is approximated by a classical aliasing calculation—aliasing in a DFT-based power spectrum estimate is described by a window function, and in this case the aliasing window function is a decent approximation of the true window function in our test. To see this, consider the window function induced by the classical aliasing in a DFT and the resulting power spectrum. If a Gaussian random function $f(r)$ has a power spectrum $P(k)$, i.e.,

$$\langle f(r) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} e^{2\pi i k r} \tilde{f}(k), \tag{B9}$$

$$\langle\langle \tilde{f}(k) \tilde{f}^{(} k')\rangle = 2\pi P(k) \delta(k - k'), \tag{B10}$$

and $f(r)$ is sampled at a rate of $2k_s$ with samples $f_n = f(r_n)$ and the power spectrum is estimated from the DFT estimate

$$\langle \tilde{f}(k_m) \propto \sum_{n=1}^{N} e^{2\pi i \frac{n}{N} m} f_n, \tag{B11}$$

then the measured power spectrum $\hat{P}(k_m) = |\tilde{f}(k_m)|^2$ has an expectation value $\langle \hat{P} \rangle \approx P_{\rm aliased}$ (for $N \gtrsim 100$) where, according to the well-known equation,

$$\langle P_{\rm aliased}(k) = P(k) + \sum_{n=1}^{\infty} P(2nk_s - k) + P(2nk_s + k). \tag{B12}$$

This can be expressed as the effect of a window function defined as

$$\langle \begin{array}{l} W(k, k') = \delta(k - k') + \sum_{n=1}^{\infty} \delta(2nk_s - k - k') \\ \quad +\delta(2nk_s + k - k') \end{array} \tag{B13}$$

and then for $k \in [0, k_s)$

$$\langle P_{\rm aliased}(k) = \int_{0}^{\infty} W(k, k') P(k') dk'. \tag{B14}$$

The function $P_{\rm aliased}$ is the "aliased" power spectrum in Figure 4.

## ORCID iDs

James E. Aguirre https://orcid.org/0000-0002-4810-666X
Steven G. Murray https://orcid.org/0000-0003-3059-3823
Joshua S. Dillon https://orcid.org/0000-0003-3336-9958
Daniel C. Jacobs https://orcid.org/0000-0002-0917-2269
Nicholas S. Kern https://orcid.org/0000-0002-8211-1892
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
Adrian Liu https://orcid.org/0000-0001-6876-0928
Adam P. Beardsley https://orcid.org/0000-0001-9428-8233
Gianni Bernardi https://orcid.org/0000-0002-0916-7443
Judd D. Bowman https://orcid.org/0000-0002-8475-2036
Richard F. Bradley https://orcid.org/0000-0003-1172-8331
Philip Bull https://orcid.org/0000-0001-5668-3101
Chris L. Carilli https://orcid.org/0000-0001-6647-3861
David R. DeBoer https://orcid.org/0000-0003-3197-2294
Aaron Ewall-Wice https://orcid.org/0000-0002-0086-7363
Steven R. Furlanetto https://orcid.org/0000-0002-0658-1243
Deepthi Gorthi https://orcid.org/0000-0002-0829-167X
Bradley Greig https://orcid.org/0000-0002-4085-2094
Bryna J. Hazelton https://orcid.org/0000-0001-7532-645X
Jacqueline N. Hewitt https://orcid.org/0000-0002-4117-570X
Joshua Kerrigan https://orcid.org/0000-0002-1876-272X
Saul A. Kohn https://orcid.org/0000-0001-6744-5328
Andrei Mesinger https://orcid.org/0000-0003-3374-1772
Miguel F. Morales https://orcid.org/0000-0001-7694-4030
Abraham R. Neben https://orcid.org/0000-0001-7776-7240
Nipanjana Patra https://orcid.org/0000-0002-9457-1941
Jonathan C. Pober https://orcid.org/0000-0002-3492-0433
Mario G. Santos https://orcid.org/0000-0003-3892-3073
Peter Sims https://orcid.org/0000-0002-2871-0413
Saurabh Singh https://orcid.org/0000-0001-7755-902X

## References

Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2015, ApJ, 809, 61
Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2018, ApJ, 863, 201
Asad, K. M. B., Koopmans, L. V. E., Jelić, V., et al. 2016, MNRAS, 462, 4482
Asad, K. M. B, Koopmans, L. V. E., Jelić, V., et al. 2018, MNRAS, 476, 3051
Barry, N., Beardsley, A. P., Byrne, R., et al. 2019a, PASA, 36, e026
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. 2019b, ApJ, 884, 1
Beardsley, A., Hazelton, B., Sullivan, I., et al. 2016, ApJ, 833, 102
Boonstra, A., & van der Veen, A. 2003, ITSP, 51, 25
Byrne, R., Morales, M. F., Hazelton, B., et al. 2019, ApJ, 875, 70
Chapman, E., Abdalla, F. B., Bobin, J., et al. 2013, MNRAS, 429, 165
Cheng, C., Parsons, A. R., Kolopanis, M., et al. 2018, ApJ, 868, 26
Choudhuri, S., Bull, P., & Garsden, H. 2021, MNRAS, 506, 2066
Datta, A., Bhatnagar, S., & Carilli, C. L. 2009, ApJ, 703, 1851
de Lera Acedo, E., Pienaar, H., & Fagnoni, N. 2020, arXiv:2003.10733
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
Dillon, J., Liu, A., Williams, C., et al. 2014, PhRvD, 89, 23002
Dillon, J. S., & Parsons, A. R. 2016, ApJ, 826, 181
Dillon, J. S., Lee, M., Ali, Z. S., et al. 2020, MNRAS, 499, 5840
Eastwood, M. W., Anderson, M. M., Monroe, R. M., et al. 2019, AJ, 158, 84
Ewall-Wice, A., Dillon, J. S., Liu, A., & Hewitt, J. 2017, MNRAS, 470, 1849
Ewall-Wice, A., Dillon, J., Hewitt, J., et al. 2016, MNRAS, 460, 4320
Fagnoni, N., de Lera Acedo, E., DeBoer, D. R., et al. 2021b, MNRAS, 500, 1232
Fagnoni, N., de Lera Acedo, E., Drought, N., et al. 2021a, ITAP, 69, 8143
Gehlot, B. K., Koopmans, L. V. E., de Bruyn, A. G., et al. 2018, MNRAS, 478, 1484
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Hurley-Walker, N., Callingham, J. R., Hancock, P. J., et al. 2017, MNRAS, 464, 1146
Jacobs, D. C., Hazelton, B. J., Trott, C. M., et al. 2016, ApJ, 825, 114
Kern, N., & Liu, A. 2021, MNRAS, 501, 1463
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, D., Liu, A., & Switzer, E. R. 2021, AAS Meeting, 231, 153.09
Kolopanis, M., Jacobs, D. C., Cheng, C., et al. 2019, ApJ, 883, 133
Lanman, A. E., & Pober, J. C. 2019, MNRAS, 487, 5840
Lanman, A. E., Pober, J. C., Kern, N. S., et al. 2020, MNRAS, 494, 3712
Li, W., Pober, J. C., Barry, N., et al. 2019, ApJ, 887, 141
Liu, A., & Shaw, J. R. 2020, PASP, 132, 062001
Liu, A., Zhang, Y., & Parsons, A. R. 2016, ApJ, 833, 242
Martinot, Z. E., Aguirre, J. E., Kohn, S. A., & Washington, I. Q. 2018, ApJ, 869, 79
McEwen, J. D., & Wiaux, Y. 2011, ITSP, 59, 5876
McKinley, B., Yang, R., López-Caniego, M., et al. 2015, MNRAS, 446, 3478
Mertens, F. G., Ghosh, A., & Koopmans, L. V. E. 2018, MNRAS, 478, 3640

Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, MNRAS, 493, 1662

Mevius, M., Mertens, F., Koopmans, L. V. E., et al. 2022, MNRAS, 509, 3693

Morales, M. F., Beardsley, A., Pober, J., et al. 2019, MNRAS, 483, 2207

Mouri Sardarabadi, A., & Koopmans, L. V. E. 2019, MNRAS, 483, 5480

Nunhokee, C. D., Bernardi, G., Kohn, S. A., et al. 2017, ApJ, 848, 47

Offringa, A. R., Mertens, F., & Koopmans, L. V. E. 2019a, MNRAS, 484, 2866

Offringa, A. R., Mertens, F., van der Tol, S., et al. 2019b, A&A, 631, A12

Orosz, N., Dillon, J. S., Ewall-Wice, A., Parsons, A. R., & Thyagarajan, N. 2019, MNRAS, 487, 537

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

Parsons, A., Pober, J., McQuinn, M., Jacobs, D., & Aguirre, J. 2012a, ApJ, 753, 81

Parsons, A. R., & Backer, D. C. 2009, AJ, 138, 219

Parsons, A. R., Liu, A., Ali, Z. S., & Cheng, C. 2016, ApJ, 820, 51

Parsons, A. R., Pober, J. C., Aguirre, J. E., et al. 2012b, ApJ, 756, 165

Patil, A. H., Yatawatta, S., Koopmans, L. V. E., et al. 2017, ApJ, 838, 65

Pober, J. C., Liu, A., Dillon, J. S., et al. 2014, ApJ, 782, 66

Shaw, J. R., Sigurdson, K., Pen, U.-L., Stebbins, A., & Sitwell, M. 2014, ApJ, 781, 57

Tan, J., Liu, A., Kern, N. S., et al. 2021, ApJS, 255, 26

The HERA Collaboration, Abdurashidova, Z., Aguirre, J. E., et al. 2021, MNRAS, arXiv:2108.02263

The HERA Collaboration, & Wilenksy, M. 2019, HERA Memo #82. 2019, HERA Memo Series, http://reionization.org/science/memos/

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

Trott, C. M., Jordan, C. H., Midgley, S., et al. 2020, MNRAS, 493, 4711

Trott, C. M., Pindor, B., Procopio, P., et al. 2016, ApJ, 818, 139

van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2

Wilensky, M. J., Barry, N., Morales, M. F., Hazelton, B. J., & Byrne, R. 2020, MNRAS, 498, 265

Zheng, H., Tegmark, M., Dillon, J. S., et al. 2017, MNRAS, 464, 3486