Methods of Error Estimation for Delay Power Spectra in 21 cm Cosmology

DOI:
10.3847/1538-4365/ac0533

Draft version May 7, 2021
Typeset using LaTeX **twocolumn** style in AASTeX631

Methods of Error Estimation for Delay Power Spectra in 21 cm Cosmology

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

[1]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*
[2]*Department of Physics and McGill Space Institute, McGill University, Montreal, QC, Canada H3A 2T8*
[3]*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA*
[4]*Department of Astronomy, University of California, Berkeley, CA*
[5]*Cavendish Astrophysics, University of Cambridge, Cambridge, UK*
[6]*SKA-SA, Cape Town, South Africa*
[7]*School of Earth and Space Exploration, Arizona State University, Tempe, AZ*
[8]*Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa*
[9]*INAF-Istituto di Radioastronomia, via Gobetti 101, 40129 Bologna, Italy*
[10]*National Radio Astronomy Observatory, Charlottesville, VA*
[11]*School of Physics & Astronomy, Queen Mary University of London, London, UK*
[12]*Department of Physics, Brown University, Providence, RI*
[13]*National Radio Astronomy Observatory, Socorro, NM*
[14]*Department of Physics and Astronomy, University of California, Los Angeles, CA*
[15]*School of Physics, University of Melbourne, Parkville, VIC 3010, Australia*
[16]*Department of Physics, University of Washington, Seattle, WA*
[17]*eScience Institute, University of Washington, Seattle, WA*
[18]*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA*
[19]*School of Chemistry and Physics, University of KwaZulu-Natal, Westville Campus, Durban, South Africa*
[20]*Scuola Normale Superiore, 56126 Pisa, PI, Italy*
[21]*Center for Astrophysics, Harvard & Smithsonian, 60 Garden St., Cambridge, MA*
[22]*American Astronomical Society, 1667 K Street NW, Suite 800, Washington, DC 20006*

ABSTRACT

Precise measurements of the 21 cm power spectrum are crucial for understanding the physical processes of hydrogen reionization. Currently, this probe is being pursued by low-frequency radio interferometer arrays. As these experiments come closer to making a first detection of the signal, error estimation will play an increasingly important role in setting robust measurements. Using the delay power spectrum approach, we have produced a critical examination of different ways that one can estimate error bars on the power spectrum. We do this through a synthesis of analytic work, simulations of toy models, and tests on small amounts of real data. We find that, although computed independently,

Corresponding author: Jianrong Tan
jianrong@sas.upenn.edu

the different error bar methodologies are in good agreement with each other in the noise-dominated regime of the power spectrum. For our preferred methodology, the predicted probability distribution function is consistent with the empirical noise power distributions from both simulated and real data. This diagnosis is mainly in support of the forthcoming HERA upper limit, and also is expected to be more generally applicable.

1. INTRODUCTION

The Epoch of Reionization (EoR)—when neutral hydrogen in the intergalactic medium (IGM) was ionized by photons from early galaxies and active galactic nuclei—remains one of the most exciting frontiers in modern astrophysics and cosmology. Precise measurements of this era will significantly enhance our understanding on the origin of very first stars, the process of galaxy formation and the thermal history of the IGM (Barkana & Loeb 2001; Dayal & Ferrara 2018). Some measurements, such as those of the optical depth of Cosmic Microwave Background (CMB) photons (Planck Collaboration et al. 2020), the Gunn-Peterson trough in distant quasar spectra (Becker et al. 2001; Fan et al. 2006; Bolton et al. 2011; Becker et al. 2015), quasar damping wings (Davies et al. 2018), and the decrease in the number density and the clustering trends of Ly-α emitters at high redshifts (Stark et al. 2010; Ouchi et al. 2010; Bosman et al. 2018), have already established the basic parameters of the EoR. Collectively, they suggest that reionization is a process which probably began at $z \gg 10$ and ended around $z \approx 6$. However, the aforementioned probes paint an indirect and incomplete picture of the EoR. For example, CMB measurements are integral constraints over redshift, making the extraction of detailed information technically difficult (often involving subtle kinetic Sunyaev-Zel'dovich effect or polarization measurements); Lyα photons suffer from severely saturated absorption that makes it difficult for them to probe earlier times than the end of reionization; and low-mass galaxies (i.e., those thought to be responsible for supplying a large fraction of ionizing photons) are too faint to be directly detected. A complementary probe capable of making direct observations of the EoR is therefore desirable.

A strong candidate for a direct probe of reionization is the 21 cm line. Arising from the "spin flip" transition in the hyperfine structure of atomic hydrogen, the 21 cm line is a promising way to directly trace the evolution of HI regimes on different spatial scales and to eventually provide a comprehensive three-dimensional picture throughout the history of reionization (Furlanetto et al. 2006; Morales & Wyithe 2010; Pritchard & Loeb 2012; Liu & Shaw 2020). Current experimental efforts are focused on slightly more modest—but still ambitious—observables. One example is the global 21 cm signal, which is a single spectrum of 21 cm absorption or emission averaged over the entire angular area of the sky (Bowman et al. 2008; Singh et al. 2018). Recently, the Experiment to Detect the Global Epoch of reionization Step team (EDGES) reported a tentative detection of a 21 cm absorption signature at $z \sim 17$ (Bowman et al. 2018a), although this result remains controversial (Hills et al. 2018; Bowman et al. 2018b; Bradley et al. 2019; Singh & Subrahmanyan 2019; Sims & Pober 2020). Global signal measurements are complemented by experimental efforts to map spatial fluctuations in the 21 cm brightness temperature field. Most such efforts currently focus on a measurement of the power spectrum, i.e., the variance in Fourier space. Power spectrum measurements have the potential to significantly improve constraints on cosmological and astrophysical parameters of reionization models, and to potentially even discover new fundamental physics (e.g., McQuinn et al. 2006; Pober et al. 2014; Greig & Mesinger 2015; Pober et al. 2015; Kern et al. 2017; Greig & Mesinger 2017; Hassan et al. 2017; Park et al. 2019; Ghara et al. 2020). Typically, these measurements are pursued by low-frequency radio interferometer arrays, such as the Murchison Widefield Array[1] (MWA; Tingay et al. 2013; Bowman et al. 2013), the Low Frequency Array[2] (LOFAR; van Haarlem et al. 2013), the Donald C. Backer Precision Array for Probing the Epoch of Reionization[3] (PAPER; Parsons et al. 2010), the Hydrogen Epoch of Reionization Array[4] (HERA; DeBoer et al. 2017), and the Square Kilometre Array[5] (SKA; Mellema et al. 2013; Koopmans et al. 2015). Although no experiment has yet to claim a detection of the 21 cm power spectrum at redshifts relevant to the EoR, steady progress has been made in recent years in the form of increasingly stringent and robust upper limits(Dillon et al. 2014, 2015; Beardsley et al. 2016; Patil et al. 2017; Barry et al. 2019; Kolopanis et al. 2019; Li et al. 2019; Mertens et al. 2020; Trott et al. 2020).

[1] http://www.mwatelescope.org

[2] http://www.lofar.org

[3] http://eor.berkeley.edu

[4] https://reionization.org

[5] https://www.skatelescope.org

In this paper, we tackle the crucial problem of error estimation in the context of 21 cm power spectrum measurements. While an extensive literature on power spectrum error estimation exists for CMB measurements and galaxy surveys, there are several challenges that are unique to 21 cm cosmology. Chief amongst these is the fact that any measured signals will be strongly contaminated by the foregrounds, which are generally 4 to 5 orders of magnitude stronger in temperature (de Oliveira-Costa et al. 2008; Jelić et al. 2008; Bernardi et al. 2009). To overcome this obstacle, some collaborations pursue a strategy of foreground subtraction, where models of foreground emission are subtracted from the data (e.g., Harker et al. 2009; Bernardi et al. 2011; Cho et al. 2012; Chapman et al. 2012; Shaw et al. 2015). Different approaches to foreground subtraction make different assumptions (see Liu & Shaw 2020 for examples), but all face the same problem of attempting to subtract a large contaminant from a large raw signal to reveal a small cosmological signature. With empirical constraints on the low-frequency radio sky being relatively scarce and generally imprecise, the chances of mis-subtraction are high. Errors in such a subtraction process as well as the effects of subtraction residuals must therefore be propagated through to a final power spectrum estimate.

In this paper, however, we do not tackle the problem of error propagation in the context of foreground subtraction; instead, we consider error estimation in the context of foreground avoidance, where one aims to make cosmological measurements exclusively in Fourier modes where foregrounds are expected to be subdominant. Key to this is the notion of the foreground wedge, a regime in Fourier space beyond which spectrally smooth foregrounds cannot extend if observed using an ideal interferometer (Datta et al. 2010; Parsons et al. 2012b; Vedantham et al. 2012; Morales et al. 2012; Trott et al. 2012; Thyagarajan et al. 2013; Hazelton et al. 2013; Liu et al. 2014a). The limitation of foregrounds to the wedge is a theoretically robust notion (Liu & Shaw 2020), and in principle one can make foreground-free measurements simply by avoiding the regime. In practice, observations are never made using perfect interferometers, and instrumental systematics such as having non-identical antenna elements, cable reflections, and cross couplings (e.g., Kern et al. 2019, 2020a) complicate one's foreground mitigation efforts. These complications can result in the appearance of contaminants outside of the foreground wedge, and in this paper we define and tackle the problem of error estimation in two regimes: a noise-dominated regime and a signal-dominated regime (whether these signals could be foregrounds, systematics, or any other coherent signals).

Through a combination of analytic work, simulations of toy models, and tests on small amounts of real data, we critically examine different ways in which one can place error bars on 21 cm delay power spectra. Our goal is to produce a "buyer's guide" that enumerates the advantages and disadvantages of various error estimation methods. Understanding these strengths and weaknesses are crucial for setting upper limits, diagnosing systematics, interpreting the results of null tests, and for the design and optimization of future telescopes (Morales 2005; McQuinn et al. 2006; Parsons et al. 2012a). Although we will focus primarily on the delay power spectrum-style analysis (Parsons et al. 2012b) in support of recent HERA upper limits (HERA Collaboration 2021), we expect many of our results to be more generally applicable.

This paper is organized as follows: in Section 2, we review the basics of power spectrum estimation using the delay spectrum technique, establishing our notation. In Section 3 we propose several methods for estimating errors in 21 cm delay power spectra. These approaches are then compared and contrasted using simulations and real data in Section 4. We then discuss the strengths and weaknesses of each error estimation method in Section 5 before summarizing our conclusions in Section 6. For readers' convenience, we provide dictionaries for a number of quantities defined in this paper in Tables 1 and 2.

2. POWER SPECTRUM ESTIMATION VIA THE DELAY SPECTRUM

In this section we review the delay spectrum approach to 21 cm power spectrum estimation (Parsons et al. 2012b) using the the language of the quadratic estimator (QE) formalism (Liu & Tegmark 2011) that we adopt in this paper.

The delay spectrum technique enables power spectra to be estimated using just a single baseline of a radio interferometer, with fluctuations in the 21 cm signal probed primarily in the line-of-sight direction via spectral information. The starting point is the visibility $V(\boldsymbol{b},\nu)$ measured by an interferometer's baseline $\boldsymbol{b}$ at frequency ν. Under the flat-sky limit, it is given by

$$V(\boldsymbol{b},\nu) = \int I(\boldsymbol{\theta},\nu)A(\boldsymbol{\theta},\nu)\exp\left(-i2\pi\frac{\nu}{c}\boldsymbol{b}\cdot\boldsymbol{\theta}\right)\mathrm{d}^2\theta\,, \quad (1)$$

where c is the speed of light, $\boldsymbol{\theta}$ is the angular sky position, $I(\boldsymbol{\theta},\nu)$ is the source intensity function, and $A(\boldsymbol{\theta},\nu)$ is the primary beam function. If we express $I(\boldsymbol{\theta},\nu)$ in terms of its Fourier transform $\tilde{I}(\boldsymbol{u},\eta)$, i.e.,

$$I(\boldsymbol{\theta},\nu) = \int \tilde{I}(\boldsymbol{u},\eta)e^{i2\pi(\boldsymbol{u}\cdot\boldsymbol{\theta}+\eta\nu)}\mathrm{d}^2u\mathrm{d}\eta, \quad (2)$$

then our visibility equation becomes

$$V(\boldsymbol{b},\nu) = \int \tilde{I}(\boldsymbol{u},\eta) A(\boldsymbol{\theta},\nu) e^{i2\pi(\boldsymbol{u}\cdot\boldsymbol{\theta}+\eta\nu-\boldsymbol{b}_\lambda\cdot\boldsymbol{\theta})} \mathrm{d}^2u\mathrm{d}\eta\mathrm{d}^2\theta$$
$$= \int \tilde{I}(\boldsymbol{u},\eta)\tilde{A}(\boldsymbol{b}_\lambda - \boldsymbol{u},\nu) e^{i2\pi\eta\nu} \mathrm{d}^2u\mathrm{d}\eta, \quad (3)$$

where we have defined $\boldsymbol{b}_\lambda \equiv \frac{\nu}{c}\boldsymbol{b}$ as the normalized baseline vector for baseline $\boldsymbol{b}$ in units of wavelength. In the angular directions, we see that a visibility has a response to $\boldsymbol{u}$ modes centred around $\boldsymbol{b}_\lambda$. If the primary beam A is fairly broad, $\tilde{A}$ will be highly compact and the majority of the integral will be sourced from $\boldsymbol{u} \approx \boldsymbol{b}_\lambda$. We will use this fact later. From this, one sees that a visibility $V(\boldsymbol{b},\nu)$ is a linear function of $\tilde{I}(\boldsymbol{u},\eta)$. This quantity is directly related to the cylindrical power spectrum $P(\boldsymbol{u},\eta)$, which decomposes power into Fourier wavenumbers perpendicular to the line of sight ($\boldsymbol{u}$) and parallel to the line of sight (η), and is formally defined as

$$\langle \tilde{I}^*(\boldsymbol{u},\eta)\tilde{I}(\boldsymbol{u}',\eta')\rangle \equiv \delta^{\mathrm{D}}(\boldsymbol{u}-\boldsymbol{u}')\delta^{\mathrm{D}}(\eta-\eta')P(\boldsymbol{u},\eta). \quad (4)$$

Such a power spectrum can be recast into more conventional cosmological coordinates via the relations[6]

$$\boldsymbol{k}_\perp = \frac{2\pi\boldsymbol{u}}{D_\mathrm{c}}\,; \quad k_\parallel = \frac{2\pi\nu_{21}H_0E(z)}{c(1+z)^2}\eta, \quad (5)$$

where D_c is the line-of-sight comoving distance, ν_{21} is the rest frequency of the 21 cm line, H_0 is the Hubble parameter today, and $E(z) \equiv \sqrt{\Omega_\Lambda + \Omega_m(1+z)^3}$, with Ω_Λ and Ω_m as the normalized dark energy and matter density, respectively.

Since the power spectrum is a quadratic function of the Fourier representation of the sky, we expect that one should be able to estimate the power spectrum by forming some quadratic function of visibilities. However, directly squaring some functions of the visibilities will incur a noise bias because noise that is symmetrically distributed about zero will have a positive contribution that does not average down with cumulative samples. Fortunately, the noise bias can be avoided by cross-multiplying nominally identical measurements rather than by squaring a single measurement. For instance, one might choose to form quadratic combinations of data from adjacent time samples of a single baseline's time stream, or perhaps to cross-multiply the time streams from two redundant baselines that satisfy $\boldsymbol{b}_1 = \boldsymbol{b}_2 = \boldsymbol{b}$ for some $\boldsymbol{b}$. In this paper, we will consider power spectrum measurements that are formed from cross-multiplications in *both* time and different copies of an identical baseline. Utilizing both types of cross-multiplications has the advantage of avoiding skewness in the probability distributions of the measured power spectra, simplifying the interpretation of our results. This is discussed in Appendix A. In this section, however, we will—for simplicity—suppress explicit reference to the data time stream and use notation that explicitly refers to cross-correlating different baselines. Given a pair of redundant baselines $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, we stack their measuring visibilities at multiple frequencies ν_1, ν_2... at single time instants into two data vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, such that

$$\boldsymbol{x}_1 = \begin{pmatrix} V(\boldsymbol{b}_1,\nu_1) \\ V(\boldsymbol{b}_1,\nu_2) \\ \vdots \end{pmatrix}; \quad \boldsymbol{x}_2 = \begin{pmatrix} V(\boldsymbol{b}_2,\nu_1) \\ V(\boldsymbol{b}_2,\nu_2) \\ \vdots \end{pmatrix}. \quad (6)$$

To make an explicit connection between visibilities and power spectra, we must examine the statistical properties of these data vectors. For quadratic statistics the key quantity is the covariance matrix $\boldsymbol{C}^{12} \equiv \langle \boldsymbol{x}_1\boldsymbol{x}_2^\dagger\rangle$, which can be written as

$$\boldsymbol{C}^{12}_{ij} \equiv \langle V(\boldsymbol{b_1},\nu_i)V^*(\boldsymbol{b_2},\nu_j)\rangle$$
$$= \int P(\boldsymbol{u},\eta)\tilde{A}(\boldsymbol{b}_{\lambda 1i}-\boldsymbol{u},\nu_i)\tilde{A}^*(\boldsymbol{b}_{\lambda 2j}-\boldsymbol{u},\nu_j)$$
$$\times e^{i2\pi\eta(\nu_i-\nu_j)}\mathrm{d}^2u\mathrm{d}\eta$$
$$\approx \int P(\overline{\boldsymbol{b}}_\lambda,\eta)e^{i2\pi\eta(\nu_i-\nu_j)}\mathrm{d}\eta$$
$$\times \int \tilde{A}^*(\boldsymbol{b}_{\lambda 1i}-\boldsymbol{u},\nu_i)\tilde{A}(\boldsymbol{b}_{\lambda 2j}-\boldsymbol{u},\nu_j)\mathrm{d}^2u\,, \quad (7)$$

where $\boldsymbol{b}_{\lambda 1i}$ and $\boldsymbol{b}_{\lambda 2j}$ are the normalized baseline vectors for baseline $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ evaluated at frequencies ν_i and ν_j, respectively, and $\overline{\boldsymbol{b}}_\lambda$ is the mean of the two. In deriving Equation (7), we first substituted Equation (3) for the expressions of visibilities in the angle bracket, and then factored the evaluated cylindrical power spectrum out of the integral over $\boldsymbol{u}$. Next we replace the continuous integral on power spectra with discrete sums over a series of piecewise constant bandpowers $P(\overline{\boldsymbol{b}}_\lambda,\eta_\alpha)$, such that

$$\boldsymbol{C}^{12}_{ij} \approx \sum_\alpha P(\overline{\boldsymbol{b}}_\lambda,\eta_\alpha)\int_{\eta_\alpha} e^{i2\pi\eta_\alpha(\nu_j-\nu_i)}\mathrm{d}\eta$$
$$\times \int \tilde{A}(\boldsymbol{b}_{\lambda 1i}-\boldsymbol{u},\nu_i)\tilde{A}^*(\boldsymbol{b}_{\lambda 2j}-\boldsymbol{u},\nu_j)\mathrm{d}^2u$$
$$\approx \sum_\alpha P(\overline{\boldsymbol{b}}_\lambda,\eta_\alpha)e^{i2\pi\eta_\alpha(\nu_i-\nu_j)}\Delta\eta$$
$$\times \int e^{-i2\pi(\boldsymbol{b}_{\lambda 1i}-\boldsymbol{b}_{\lambda 2j})\cdot\boldsymbol{\theta}}A(\theta,\nu_i)A^*(\theta,\nu_j)\mathrm{d}^2\theta$$
$$\equiv \sum_\alpha P(\overline{\boldsymbol{b}}_\lambda,\eta_\alpha)\boldsymbol{Q}^{12,\alpha}_{ij}, \quad (8)$$

[6] In addition to mapping the arguments of P, there is also an additional multiplicative constant; see Liu et al. (2014a) for explicit expressions.

Henceforth, we will adopt the notation $P_\alpha \equiv P(\bar{\boldsymbol{b}}_\lambda, \eta_\alpha)$ to mean the value of the cylindrical power spectrum $P(\boldsymbol{u}, \eta)$ evaluated at $\boldsymbol{u} = \bar{\boldsymbol{b}}_\lambda$ and $\eta = \eta_\alpha$. The index α discretely runs over a series of bins in η, and as long as these bins are narrow compared to the scales over which the power spectrum changes, a piecewise constant treatment is appropriate.

Equation (8) shows the cross-baseline covariance matrix of visibilities encodes information about the power spectrum bandpowers via a family of response matrices $\boldsymbol{Q}^{12,\alpha}$ (with a different matrix for every value of the bandpower index α). Since the covariance is an ensemble-averaged quadratic function of the data, one might venture that estimators for the bandpowers can be constructed by forming quadratic combinations of the data, i.e.,

$$\hat{P}_\alpha = \boldsymbol{x}_1^\dagger \boldsymbol{E}^{12,\alpha} \boldsymbol{x}_2 \,, \tag{9}$$

where $\boldsymbol{E}^{12,\alpha}$ is a matrix that can be chosen (within certain limitations) by the data analyst. Taking the ensemble average on both sides and inserting Equation (8) then yields

$$\langle \hat{P}_\alpha \rangle = \sum_\beta \mathrm{tr}\left(\boldsymbol{E}^{12,\alpha}\boldsymbol{Q}^{21,\beta}\right) P_\beta \equiv \sum_\beta W_{\alpha\beta} P_\beta, \tag{10}$$

where $\boldsymbol{W}$ is the window function matrix. To ensure that our estimated bandpowers are correctly normalized, we require that each row of $\boldsymbol{W}$ sum to unity.

In the HERA power spectrum pipeline, we pick a family of $\boldsymbol{E}^{12}$ matrices of the form

$$\boldsymbol{E}^{12,\alpha} \equiv M_\alpha \boldsymbol{R}_1 \boldsymbol{Q}^{\mathrm{DFT},\alpha} \boldsymbol{R}_2, \tag{11}$$

where the matrix $\boldsymbol{Q}^{\mathrm{DFT},\alpha}_{ij} \equiv e^{i2\pi\eta_\alpha(\nu_i - \nu_j)}$ is responsible for taking the Fourier transform of the two copies of the data vectors in the quadratic estimator. The matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ are weighting matrices that act on visibilities from $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, respectively. In this paper, we use $\boldsymbol{R} = \boldsymbol{T}\boldsymbol{Y}$, where both $\boldsymbol{T}$ and $\boldsymbol{Y}$ are diagonal matrices. The former is used to impose a Blackman-Harris tapering function on the spectral data, and the latter propagates data flags. With a quadratic estimator of this form, the normalization scalar, M_α, should take the form

$$M_\alpha = \frac{1}{\sum_\beta \mathrm{tr}(\boldsymbol{R}_1 \boldsymbol{Q}^{\mathrm{DFT},\alpha} \boldsymbol{R}_2 \boldsymbol{Q}^{12,\beta})} \tag{12}$$

which ensures that the rows of $\mathbf{W}$ sum to unity, and therefore that the bandpowers are properly normalized. In our case, we do use this normalization, but we approximate the $\boldsymbol{Q}^{12,\beta}$ term in the denominator. Rather than evaluating the full integral in Equation (8), we make the approximation that $\mathbf{b}_{\lambda 1i} \approx \mathbf{b}_{\lambda 2i}$. In fact, this is the motivation for the use of $\boldsymbol{Q}^{\mathrm{DFT},\alpha}$ in Equation (11) rather than $\boldsymbol{Q}^{12}$; notice that if $\mathbf{b}_{\lambda 1i} = \mathbf{b}_{\lambda 2i}$, then $\boldsymbol{Q}^{12} \propto \boldsymbol{Q}^{\mathrm{DFT}}$. Over large bandwidths, this will fail for long baselines, since $\boldsymbol{b}_\lambda \equiv \nu \boldsymbol{b}/c$.

The approximation that we have just made is equivalent to the delay spectrum approximation (Parsons et al. 2012b; Liu et al. 2014a). To see this, we can write our estimator in the continuous limit. Our current form for $\boldsymbol{E}^{12,\alpha}$ is separable into the product of two matrices that each involve only one of the two baselines. In particular, if $\gamma(\nu)$ is the functional form of the Blackman-Harris taper, then we have $\boldsymbol{E}^{12,\alpha}_{ij} = \gamma_1(\nu_i) e^{i2\pi\eta_\alpha(\nu_i-\nu_j)} \gamma_2(\nu_j)$, and its action on each baseline's visibilities in Equation (9) is to compute the quantity

$$\sum_i V(\boldsymbol{b}, \nu_i) \gamma(\nu_i) e^{-2\pi\eta\nu_i} \Delta\nu, \tag{13}$$

which is just a discrete approximation to

$$\tilde{V}(\boldsymbol{b}, \eta) = \int V(\boldsymbol{b}, \nu) \gamma(\nu) e^{-i2\pi\eta\nu} \mathrm{d}\nu \,. \tag{14}$$

Note Equation (14) is an equivalent expression of the delay transform in Parsons et al. (2012b). Therefore

$$\begin{aligned} \hat{P}_\alpha &= \boldsymbol{x}_1^\dagger \boldsymbol{E}^{12,\alpha} \boldsymbol{x}_2 \\ &\propto \sum_{ij} V^*(\boldsymbol{b}_1, \nu_i)\gamma_1(\nu_i) V(\boldsymbol{b}_2, \nu_j)\gamma_2(\nu_j) e^{i2\pi\eta_\alpha(\nu_i-\nu_j)} \\ &= \tilde{V}^*(\boldsymbol{b}_1, \eta_\alpha)\tilde{V}(\boldsymbol{b}_2, \eta_\alpha) \,. \end{aligned} \tag{15}$$

Equation (15) just indicates that the quadratic estimator is proportional to the product of delay-transformed visibilities. This is an estimator that is based on Fourier transforming the visibility spectra from individual baselines, rather than combining information from different baselines. In principle, only the latter can probe truly rectilinear Fourier modes on the sky, since $\boldsymbol{k}_\perp \propto \boldsymbol{b}_\lambda$ (which is a frequency-dependent quantity), and thus to probe the same $\boldsymbol{k}_\perp$ at multiple frequencies—which is needed to perform the Fourier transform along the line-of-sight direction—one needs multiple baselines. The delay spectrum approach uses the fact that $\boldsymbol{b}_\lambda$ evolves only slowly with frequency for short baselines to form an approximate power spectrum estimator. We make this approximation throughout this paper, as this is the choice that has been made for the next iteration of power spectrum upper limits from HERA observations. In recognition of this, we will henceforth use τ to index our line-of-sight Fourier modes (as is customary for delay spectra) instead of η (which is generally used to denote true rectilinear line-of-sight wavenumbers) (Morales et al. 2012, 2019).

In the language of the delay spectrum, the foreground wedge becomes particularly simple to describe: smooth

Quantity	Definition/Meaning	First Appearance
$\boldsymbol{b}$; $\boldsymbol{b}_p$	Baseline vector; Vector of the pth index baseline	Equation (1)
$\boldsymbol{\theta}$	Angular sky position	Equation (1)
$\nu; \nu_i$	Frequency; Frequency of the ith index channel	Equation (1)
$\boldsymbol{b}_\lambda$; $\boldsymbol{b}_{\lambda pi}$	Normalized baseline vector in units of wavelength; Normalized vector for baseline $\boldsymbol{b}_p$ at frequency ν_i	Equation (3)
$\boldsymbol{u}$	Fourier dual to $\boldsymbol{\theta}$	Equation (2)
η; η_α	Fourier dual to ν; the αth index η mode	Equation (2)
τ; τ_α	Delay, i.e., Fourier dual to ν on a single baseline; the αth index delay mode	Equation (16)
$A(\boldsymbol{\theta}, \nu)$	Primary beam function at position θ and frequency ν	Equation (1)
$\tilde{A}(\boldsymbol{u}, \nu)$	Spatial Fourier Transform Dual of primary beam function	Equation (3)
$\gamma(\nu)$	Spectral tapering function at frequency ν	Equation (14)
$N_{\rm time}; N_{\rm blp}$	Number of time instants; Number of baseline-pairs	Equation (18)
$N_{\rm boot}$	Number of bootstrapping sample sets	Equation (24)
$I(\boldsymbol{\theta}, \nu)$	Sky source intensity function at position θ and frequency ν	Equation (1)
$\tilde{I}(\boldsymbol{u}, \eta)$	Fourier transform of I at angular wavenumber $\boldsymbol{u}$ and line-of-sight wavenumber η	Equation (2)
$V(\boldsymbol{b}, \nu)$	Visibility measured by baseline $\boldsymbol{b}$ at frequency ν	Equation (1)
$P(\boldsymbol{u}, \eta)$	Cylindrical power spectrum at angular wavenumber $\boldsymbol{u}$ and line-of-sight wavenumber η	Equation (4)
P_α	The αth bandpower $P_\alpha \equiv P(\overline{\boldsymbol{b}}_\lambda, \eta_\alpha)$	Equation (8)
$\hat{P}_\alpha$	The estimator for the αth bandpower P_α	Equation (9)
M_α	The normalization scalar of the estimator for the αth bandpower	Equation (11)
$\tilde{V}(\boldsymbol{b}_p, \tau_\alpha), \tilde{x}_p(\tau_\alpha)$	Delay spectra of baseline $\boldsymbol{b}_p$ at delay mode τ_α	Equation (15)
$\tilde{V}_{\rm signal}(\boldsymbol{b}_p, \tau_\alpha), \tilde{s}_p(\tau_\alpha)$	The signal component of $\tilde{V}$ of baseline $\boldsymbol{b}_p$ at delay mode τ_α	Equation (16)
$\tilde{V}_{\rm noise}(\boldsymbol{b}_p, \tau_\alpha), \tilde{n}_p(\tau_\alpha)$	The noise component of $\tilde{V}$ of baseline $\boldsymbol{b}_p$ at delay mode τ_α	Equation (16)
$P_{\tilde{x}_1\tilde{x}_2}$	Power spectra formed from visbilities $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$	Equation (30)

Table 1. Dictionary of highlighted scalars and functions.

Quantity	Definition/Meaning	Size	First Appearance
$\boldsymbol{x}_p$	Stacked visibilities at multiple frequencies of baseline $\boldsymbol{b}_p$	$N_{\rm freq}$	Equation (6)
$\boldsymbol{C}^{pq}$	Covariance matrices $\boldsymbol{C}^{pq} \equiv \langle \boldsymbol{x}_p \boldsymbol{x}_q^\dagger \rangle$	$N_{\rm freq} \times N_{\rm freq}$	Equation (7)
$\boldsymbol{Q}^{pq,\alpha}$	Response of covariance $\boldsymbol{C}^{pq}$ to the αth bandpower	$N_{\rm freq} \times N_{\rm freq}$	Equation (8)
$\boldsymbol{E}^{pq,\alpha}$	Matrix for quadratic estimator of bandpower P_α, i.e., $\hat{P}_\alpha = \boldsymbol{x}_p^\dagger \boldsymbol{E}^{pq,\alpha} \boldsymbol{x}_q$	$N_{\rm freq} \times N_{\rm freq}$	Equation (9)
$\boldsymbol{W}$	Window function matrix	$N_{\rm delay} \times N_{\rm delay}$	Equation (10)
$\boldsymbol{R}_p$	Weighting matrix acting on $\boldsymbol{x}_p$	$N_{\rm freq} \times N_{\rm freq}$	Equation (11)
$\boldsymbol{Q}^{\rm DFT,\alpha}$	Matrix taking Fourier Transform in the estimator	$N_{\rm freq} \times N_{\rm freq}$	Equation (11)
$\boldsymbol{U}^{pq}$	two-point correlation matrices $\boldsymbol{U}^{pq} \equiv \langle \boldsymbol{x}_p \boldsymbol{x}_q^T \rangle$	$N_{\rm freq} \times N_{\rm freq}$	Equation (33)
$\boldsymbol{G}^{pq}$	two-point correlation matrices $\boldsymbol{G}^{pq} \equiv \langle \boldsymbol{x}_p^* \boldsymbol{x}_q^\dagger \rangle$	$N_{\rm freq} \times N_{\rm freq}$	Equation (33)

Table 2. Dictionary of highlighted vectors and matrices.

spectrum foregrounds simply contaminate all modes below a particular delay, the value of which depends on the baseline length (Parsons et al. 2012b; Liu et al. 2014a; Liu & Shaw 2020). Suppose we decompose the delay transformed visibility into the signal component $\tilde{V}_{\rm signal}$ (mainly foregrounds, and we are neglecting the much weaker EoR signal here) and the noise component $\tilde{V}_{\rm noise}$, such that

$$\begin{aligned}\tilde{V}(\boldsymbol{b}_1,\tau_\alpha) &\equiv \tilde{x}_1(\tau_\alpha)\\ &\equiv \tilde{V}_{\rm signal}(\boldsymbol{b}_1,\tau_\alpha)+\tilde{V}_{\rm noise}(\boldsymbol{b}_1,\tau_\alpha)\\ &\equiv \tilde{s}_1(\tau_\alpha)+\tilde{n}_1(\tau_\alpha). \end{aligned} \tag{16}$$

Since we are working on redundant baselines, we will henceforth drop the subscript on $\tilde{s}$, as the two baselines used in Equation (15) should measure identical signals. Mathematically, then, the statement that the smooth spectrum foregrounds contaminate only low delay modes is given by

$$\hat{P}_\alpha \approx \begin{cases} \tilde{s}^*\tilde{s}+\tilde{s}^*\tilde{n}_2+\tilde{n}_1^*\tilde{s} & \text{if } |\tau_\alpha|<\tau_0 \\ \tilde{n}_1^*\tilde{n}_2 & \text{otherwise,} \end{cases} \tag{17}$$

where τ_α is the delay corresponding to the αth bandpower, and τ_0 is some critical delay value that separates parts of the power spectrum that are foreground-dominated from those that are not. In general, τ_0 will depend on the properties of one's instrument as well as the extent to which the assumption of smooth foregrounds is good. At delays less than τ_0, we have assumed that the foreground signal is so large that the noise-noise cross term can be neglected.

Throughout the rest of this paper, we will appeal to Equation (17) for intuition when contemplating the behaviour of our power spectrum estimates at different delays. For now, we note two of its important properties. First, while the power spectrum of a signal $\tilde{s}^*\tilde{s}$ will be always real valued, the overall estimator $\hat{P}_\alpha$ is complex. It is possible to write down symmetrized estimators that give real power spectra. However, since the imaginary part is sourced by noise, it is a useful diagnostic quantity to examine. Second, even though the noise-noise terms may be negligible in the signal dominated regimes, there will still be a considerable uncertainty here that enters via the signal-noise cross terms.

Until now, we have focused on power spectra estimated from visibilities measured at single time instants. Given data from multiple times, we can average the power spectra estimated from individual measurements together. For a drift scan telescope, this averaging of power spectra from different time samples is tantamount to invoking statistical isotropy to justify the spherical averaging of power spectra over different wavevector $\mathbf{k}$ directions. In addition to averaging in time, if we have multiple pairs of baselines within the same redundant group of baselines, we may average over the power spectrum estimates from multiple baseline pairs. The simplest way to do this is to perform an unweighted average:

$$\overline{\hat{P}}_\alpha = \frac{1}{N_{\rm time}N_{\rm blp}}\sum_{\rm time,blp}\hat{P}_\alpha({\rm time},{\rm blp})\,, \tag{18}$$

where $N_{\rm time}$ is the number of time integrations, $N_{\rm blp}$ is the number of baseline pairs, $\hat{P}_\alpha({\rm time},{\rm blp})$ is the power spectrum estimate (given by previous equations in this section) at a time instant and a baseline pair ("blp"), and $\overline{\hat{P}}_\alpha$ is the average of estimates. The type of averaging performed here may be termed an "incoherent average", to distinguish it from a "coherent average", where one averages over visibilities (or converts them into a single image) before squaring them in power spectrum estimation. The latter provides greater sensitivity—if calibration errors and other systematic effects can be brought under control (Morales et al. 2019). The former retains the ability to inspect the contributions from particular baseline pairs and time until right before the final result, making some systematics easier to diagnose. However, note that by employing a suitable fringe-rate filtering of the time-stream data, it is in principle possible to recover the lost sensitivity from a "square-then-add" approach (Parsons et al. 2016). In this paper, we will focus on the error statistics of the incoherent average approach, as this is what is currently used in the HERA pipeline (HERA Collaboration 2021).

Before we move into the discussion on error estimation methods in the next section, it is worth noting that Equation (18) is not the optimal way to obtain average power spectra with the least variance. Generally, given a set of estimates $\hat{\boldsymbol{P}}_\alpha$ for bandpower P_α with measurement errors $\boldsymbol{\sigma}$, such that

$$\hat{\boldsymbol{P}}_\alpha = \boldsymbol{D}P_\alpha+\boldsymbol{\epsilon}\,, \tag{19}$$

an linear estimator of P_α is written as

$$\overline{\hat{P}}_\alpha = \boldsymbol{K}\hat{\boldsymbol{P}}_\alpha\,. \tag{20}$$

Here $\boldsymbol{D}$ is a column vector of 1s. We need to select $\boldsymbol{K}$ such that $\boldsymbol{KD}=\boldsymbol{I}$ in order to achieve an unbiased constraint that satisfies $\langle\overline{\hat{P}}_\alpha\rangle = P_\alpha$. For an arbitrary matrix $\boldsymbol{K}$, the error bar $\Sigma_\alpha \equiv \langle|\overline{\hat{P}}_\alpha - P_\alpha|^2\rangle = \boldsymbol{K}\boldsymbol{\epsilon}\boldsymbol{K}^t$, where the error covariance matrix $\boldsymbol{\epsilon}\equiv\langle\boldsymbol{\sigma}\boldsymbol{\sigma}^t\rangle$. The superscript "$t$" used here and along in this paper refers to the matrix transposition. Note that Equation (18) is just a special case where $\boldsymbol{K}=[\boldsymbol{D}^t\boldsymbol{D}]^{-1}\boldsymbol{D}^t$. When Σ_α is minimized (optimal), $\overline{\hat{P}}_\alpha$ and the corresponding Σ_α should take the form of (Tegmark 1997; Dillon et al. 2014)

$$\widehat{\overline{P}}_\alpha = [\boldsymbol{D}^t \boldsymbol{\epsilon}^{-1} \boldsymbol{D}]^{-1} \boldsymbol{D}^t \boldsymbol{\epsilon}^{-1} \hat{\boldsymbol{P}}_\alpha \tag{21}$$

$$\Sigma_\alpha = [\boldsymbol{D}^t \boldsymbol{\epsilon}^{-1} \boldsymbol{D}]^{-1}\,, \tag{22}$$

which amounts to an inverse covariance weighting of the data in averaging it down. Equation (21) brings us the ability to propagate the full covariance information over samples to obtain an least-variance average result. The diagonal elements of $\boldsymbol{\epsilon}$ are easily interpreted as the variance in each individual measurement, while the off-diagonal elements, reflected by the coherency between time samples and baseline-pair samples, are far more complicated. If estimating the covariance matrix $\boldsymbol{\epsilon}$ of the pre-averaged data is difficult, one may opt to weight the data using some other matrix $\boldsymbol{\Gamma}$ instead of $\boldsymbol{\epsilon}$ in Equation (21). In this case, the final variance Σ_α ends up being

$$\Sigma_\alpha = [\boldsymbol{D}^t \boldsymbol{\Gamma}^{-1} \boldsymbol{D}]^{-1} \boldsymbol{D}^t \boldsymbol{\Gamma}^{-1} \boldsymbol{\epsilon} \boldsymbol{\Gamma}^{-t} \boldsymbol{D} [\boldsymbol{D}^t \boldsymbol{\Gamma}^{-t} \boldsymbol{D}]^{-1}\,. \tag{23}$$

In principle, one could model the off-diagonal elements of $\boldsymbol{\epsilon}$. This is particularly important in the cosmic-variance dominated regime where the sky signal—which is what sources a cosmic variance error—is slowly drifting through HERA's field of view over the course of the day, thus inducing strong correlations between different time samples. In this paper we do not consider the modelling of off-diagonal covariances in $\boldsymbol{\epsilon}$ (or between different α values in $\widehat{\overline{P}}_\alpha$). We assume diagonal covariance matrices and set $\Gamma = \boldsymbol{I}$, i.e., we use Equation (18) when computing the "incoherently-averaged" power spectra, and here we are acknowledging other possibilities only for completeness.

3. ERROR ESTIMATION METHODOLOGY

Placing robust error bars on power spectra is crucial to our data analysis, whether it is for setting upper limits, diagnosing experimental systematics, or eventually declaring a detection of the cosmological 21 cm signal. Generally, contributions to the error bars of observed power spectra come from three sources: the EoR signal, noise, and foregrounds (Thyagarajan et al. 2013; Trott 2014; Dillon et al. 2014, 2015; Lanman & Pober 2019). Of course, this is all complicated by the response of one's instrument, and ultimately, one's ability to place reliable error bars rests on one's ability to understand the behaviour of each data source in the context of the instrument.

The intrinsic variance of the EoR signal, also known as "cosmic variance", is the ensemble covariance on all possible realizations of the 21-cm temperature field. If the field is Gaussian, then its cosmic variance is proportional to the square of the power spectrum amplitude over the number of independent modes. Lanman & Pober (2019), for example, estimate the cosmic variance could go as high as $\sim 35\%$ of the EoR signal for HERA-like fields of view with eight hours of local sidereal time (LST) observations using only the shortest (14.6-m) baselines of HERA. This uncertainty due to cosmic variance is brought down to a few percent level for the spherically averaged power spectrum when using all types of baselines. Importantly, as reionization evolves, the 21-cm temperature field is expected to become highly non-Gaussian, and the excess contribution from the non-Gaussian component could lift the cosmic variance in Gaussian part staggeringly, which is significant and should be considered for future high-sensitivity measurements (Mondal et al. 2016, 2017; Shaw et al. 2019). In this paper, however, we assume that at our current levels of precision the cosmic variance is subdominant to noise and foregrounds.

For instrumental noise, we assume that the noise in the visibility from each baseline is independent and Gaussian-distributed. This is what one might expect based on the statistics of correlator outputs in a radio interferometer, but is also an assumption that we will see borne out in our empirical data in Section 4. With these well-understood statistical properties, the noise-dominated delays (recall Equation 17) are relatively easy to model, at least in principle.

The low-delay, foreground-dominated regimes are trickier to model. One key problem is that the statistics of foregrounds are not well-understood, particularly at the low frequencies relevant to us. There are different approaches that one can take to this roadblock. The first is where one attempts to make a measurement of the cosmological 21 cm signal only, by proactively subtracting (or simultaneously fitting) a foreground model. To properly set error bars on such a power spectrum, it is necessary to propagate uncertainties (accounting for the possibility of mis-subtractions) in the foreground model to the final errors (or in the case of a simultaneous fitting, to allow the errors on the cosmological signal to be appropriately inflated as one marginalizes over foreground uncertainties). While conceptually straightforward, these steps are difficult to implement in practice without a deep understanding of foreground statistics.

Instead, in this paper we treat foregrounds as additive systematics on the total sky emission. Crucially, this means we only require empirical knowledge of the foregrounds themselves, and not their full probability distribution. We simply quantify the error bars on a measurement of total sky emission due to instrumental

Name	Description	Definition
$\sigma_{\rm bs}$	Error bar of the average power spectra by bootstrapping over the collection of samples	Equation (24)
$P_{\rm diff}$	Power spectra from differenced visibility used as a form of error bar	Equation (26)
$P_{\rm N}$	Analytic noise power spectrum	Equation (27)
$P_{\rm SN}$	Error bar based on $P_{\rm N}$ but including the extra signal-noise cross term	Equation (30)
$\sigma_{\rm QE\text{-}N}$	Error bar from the output covariance in QE formalism including only noise-noise term	Equation (37)
$\sigma_{\rm QE\text{-}SN}$	Error bar from the output covariance in QE formalism including noise-noise term and signal-noise term	Equation (38)
$\tilde{P}_{\rm SN}$	Same as $P_{\rm SN}$ but with an adjustment for noise double-counting	Equation (31)
$\tilde{\sigma}_{\rm QE\text{-}SN}$	Same as $\sigma_{\rm QE\text{-}SN}$ but with an adjustment for noise double-counting	Equation (39)

Table 3. Dictionary of error bars.

noise, rather than what the error bars on the cosmological signal due to foreground uncertainties and noise. Some understanding of foregrounds is still needed for setting our errors because of the signal-noise cross terms in Equation (17). Implicit in this approach is a strategy of foreground avoidance in the hunt for a cosmological signal detection, where it is hoped that the separation between foreground-dominated and foreground-negligible regimes in Equation (17) is a clean one. It is important to note, however, that we seek to compute error bars that transition smoothly between the regimes and are valid even if the conceptual separation is not a clean one in practice.[7]

In addition to foregrounds, one can treat instrumental systematics in the same way. In other words, interpreting systematics as additive "signals", the signal-noise cross term in the variance of power spectra is sourced by not just foregrounds, but also other systematics such as cable reflections and cross couplings (Kern et al. 2019, 2020a). We can apply some models to remove systematics from the signal, but the residuals due to mis-subtraction will still increase the total uncertainties via the signal-noise cross term. Note, however, that in this paper we do not develop a comprehensive model to account for all systematics, which is particularly difficult when unknown modeling errors are present in complicated effects (e.g. direction-dependent gains). We will instead argue that a procedure of using the measured visibility itself to model the foregrounds and systematics allows us to set robust upper bounds, provided certain safeguards are in place to avoid biases. We will leave more exquisite *a priori* characterizations of foregrounds and systematics in the signal-noise cross terms for the future.

Finally, one might worry that the averaging of power spectra from multiple measurements together like Equation (18) might complicate the statistics. Appendix B shows an example of this. There, we show that when averaging over redundant baseline-pairs, the variance of average power spectra in the foreground-dominated regime goes down roughly with $N_{\rm blp}^{-1/2}$ and not $N_{\rm blp}^{-1}$ because some baselines will appear in multiple baseline *pairs*. In other words, in foreground-dominated (or systematics-dominated) regimes, one cannot assume that baseline pairs average together in an independent fashion. This has consequences for certain methods of error bar computation, such as the bootstrapping approach discussed in the next subsection, which will tend to underestimate error bars in these regimes. To avoid this, one might just use *pairs* in which each baseline only appears once in all baseline *pairs*, or to compute a correction factor on the final results. In contrast to the foreground/signal-dominated regime, in the noise-dominated regime one obtains correct final error bars by assuming that the baseline-pair samples are independent (even if they are not for the aforementioned reasons). In this paper, to avoid averaging power spectra over correlated samples, we will concentrate on the averaging of power spectra of a single baseline-pair over multiple time samples.

We will have a more extensive discussion of the meaning of our error bars in Section 5. For concreteness, however, we will now propose several different methods for generating error bars based on the HERA power spectrum pipeline before performing quantitative com-

[7] We stress that our analysis does not cease to apply at a certain delay—it is simply the case that at high delays, there is less of a pressing need to construct detailed models for foreground subtraction, which to some extent mitigates the need to consider the complicated statistical properties of this subtraction. It is likely that our formalism can be generalized to encompass some foreground subtraction, but detailed work beyond the scope of this paper would be necessary. As an example, suppose one were to use information at $\tau = 0$ and an instrument model to subtract off leakage from other low (but non-zero) delay modes. In such a scenario, one would need to account for the fact that the noise contributions between different delay modes are now coupled. This can in principle be accommodated with appropriate covariance matrix modeling, but we leave this to future work.

parisons in Section 4. For the convenience of our readers, we provide Table 3 as a quick preview.

3.1. *Bootstrap*

Bootstrapping is a natural method for computing the error bars on the final averaged power spectrum with only minimal *a priori* modeling assumptions. Within the 21 cm cosmology literature, it has previously been used to set error bars on power spectrum upper limits (Parsons et al. 2014; Ali et al. 2015; although see Cheng et al. 2018 for caveats on these limits). Bootstrapping is a process that goes hand in hand with the averaging step described in Equation (18). Rather than performing a single average, we repeatedly form a new set of pre-averaged data by resampling the original set with replacement (i.e., allowing repeated entries). A new estimate of the final average, $\overline{\hat{P}}^{(k)}$, can be produced from the kth draw. The scatter in the realizations of the final averaged power spectrum is then quoted as an error bar $\sigma_{\rm bs}$, such that

$$\sigma_{\rm bs}^2 = \frac{1}{N_{\rm boot}} \sum_k \left[\overline{\hat{P}}^{(k)} - \frac{1}{N_{\rm boot}} \sum_l \overline{\hat{P}}^{(l)} \right]^2 , \quad (24)$$

where $N_{\rm boot}$ is the number of bootstrapping sample sets. In essence, one is using the data itself as an empirical estimate of the distribution from which the data is drawn (Efron & Tibshirani 1994; Press et al. 2007).

If the input data samples are independent and identically distributed, bootstrapping will give the same error bars as the true ones from ensemble average. However, this assumption is likely to be violated with our data. Consider the two axes that we have at our disposal. One possibility is to bootstrap over different time samples. Over short timescales, different time integrations have relatively uncorrelated noise realizations. However, as our drift scan telescope moves across different local sidereal time (LST) values, the sky brightness seen by the telescope changes, leading to slow changes in the noise level for a sky-noise dominated telescope. An alternative to bootstrapping over time is to bootstrap over different copies of an identical ("redundant") baseline group. Here, the downside is that it remains an open question as to how truly redundant current interferometric arrays are (Dillon et al. 2020), and precisely what the consequences of non-redundancy are (Choudhuri et al. 2021).

With correlated data samples, bootstrapping tends to underestimate the true error bars on a final averaged power spectrum (Cheng et al. 2018). On the other hand, non-stationary effects such as non-redundancy can inflate bootstrap errors rather than revealing the fact that the data in fact come from multiple distributions. In later sections, we will compute error bars that come from bootstrapping over different LSTs, but will interpret these results with caution given the caveats we have just outlined. Of course, these caveats by no means diminish the value of bootstrap errors as yet another consistency check, particularly when one is diagnosing systematic effects (e.g., Kolopanis et al. 2019).

3.2. *Direct Noise Estimation By Visibility Differencing*

The foreground and EoR signal varies relatively slowly in time (or frequency), such that after differencing the integrated visibility between very close LSTs (or frequencies), the normalized residual,

$$V_{\rm diff} = \frac{V(\boldsymbol{b}, \nu, t_1) - V(\boldsymbol{b}, \nu, t_2)}{\sqrt{2}}$$

or

$$V_{\rm diff} = \frac{V(\boldsymbol{b}, \nu_1, t) - V(\boldsymbol{b}, \nu_2, t)}{\sqrt{2}} , \quad (25)$$

is almost noise-like. We can propagate such $V_{\rm diff}$ through power spectrum estimation pipelines to generate a "noise-like" power spectrum $P_{\rm diff}$, such that

$$P_{\rm diff} \propto \tilde{V}_{\rm diff}^* \tilde{V}_{\rm diff} , \quad (26)$$

where appropriate proportionality/normalization constants allow $P_{\rm diff}$ to have the same units as—and therefore be directly comparable to—power spectra. This quantity can be viewed as a random variable that represents random realizations of the noise in the system, which can be used to at least roughly estimate error bars in noise-dominated regimes (see Appendix C for more details). It can be computed from either time-differenced or frequency-differenced visibilities. However, by differencing neighbouring points in frequency, we are in fact applying a high-pass filter in the delay space, which means that power is suppressed at low delay modes. This is illustrated in Figure 1, and for this reason that the time-differencing method is preferred for empirical noise uncertainty estimation. However, it is important to note that many correlators do not dump data to disk fast enough for this to be feasible, as the sky changes non-negligibly on the timescale of a few seconds. The maximum time length of a single integration before reaching a decorrelation threshold depends on the baseline length, thus ones need particular simulations for their instruments to determine the suitable time scale (Wijnholds et al. 2018). For the upgraded HERA correlator, it will be able to produce time-differenced visibilities on the milli-second timescale for accurate, empirical noise estimates.

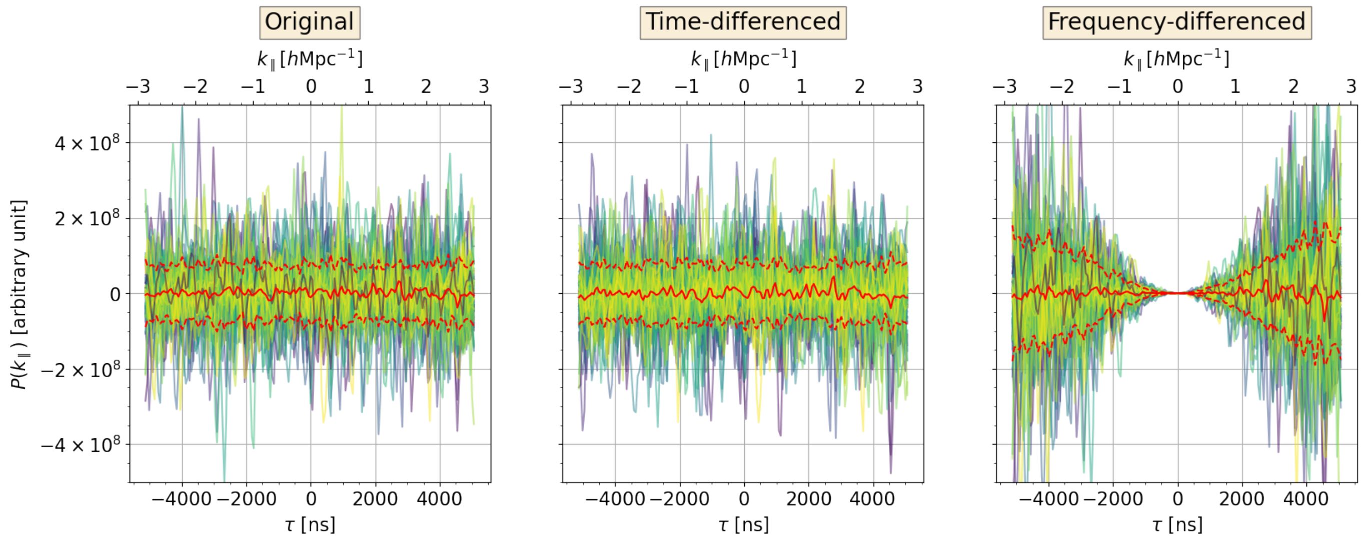

Figure 1. Here we generate ~ 60 realizations of time streams of white-Gaussian-noise visibilities, and compute the time-differenced visibilities and frequency-differenced visibilities respectively. *Left*: Power spectra from the original visibilities. *Center*: Power spectra from time-differenced visibilities. *Right*: Power spectra from frequency-differenced visibilities. In each panel, we plot the power spectra from every realization, along with the mean (solid red) and the standard deviation (dashed red) of power spectra over all realizations. We see power spectra from frequency-differenced visibilities are highly suppressed at low delays.

3.3. *Power Spectrum Method*

With appropriate approximations (see Liu & Shaw 2020 for details), it is possible to write down an analytic expression for the noise power spectrum given a system temperature, $T_{\rm sys}$ in units of Kelvin:

$$P_{\rm N} = \frac{X^2 Y \Omega_{\rm eff} T^2_{\rm sys}}{t_{\rm int} N_{\rm coherent} \sqrt{2 N_{\rm incoherent}}}\,, \tag{27}$$

where $X \equiv D_{\rm c}$ and $Y \equiv \frac{c(1+z)^2}{\nu_{21} H_0 E(z)}$ are conversion factors from sky angles and frequencies to cosmological coordinates, $\Omega_{\rm eff}$ is the effective beam area, $t_{\rm int}$ is the integration time, $N_{\rm coherent}$ is the number of samples averaged at the level of visibility while $N_{\rm incoherent}$ is the numbers of samples averaged at the level of power spectrum (Zaldarriaga et al. 2004; Pober et al. 2013; Cheng et al. 2018; Kern et al. 2020a). This is an estimate of the root-mean-square (RMS) of a power spectrum measurement in the limit that it is purely thermal noise dominated. The system temperature, $T_{\rm sys} = T_{\rm sky} + T_{\rm rcvr}$, is the sum of the sky and receiver temperature and describes the total noise content of the visibilities formed between cross-correlating data from different antennas (Thompson et al. 2017).

There are many ways in which the key quantity $T_{\rm sys}$ can be estimated. For example, we can take advantage of the differenced visibilities discussed in the previous subsection. These differences can then be converted into an estimate of $T_{\rm sys}$ via the relation

$$V_{\rm RMS}(\{p,q\}) = \frac{2k_b \nu^2 \Omega_p}{c^2} \frac{T_{\rm sys,\{p,q\}}}{\sqrt{B\Delta t}}\,, \tag{28}$$

where k_b is the Boltzmann constant, Ω_p is the integrated beam area, B is the bandwidth, and Δt is the integration time at a single time sample. The "RMS" subscript signifies taking the root-mean-square of the differenced visibilities and p and q are indices denoting two different antennas that form a baseline $\{p,q\}$. This serves to emphasize the fact that we can have a distinct system temperature for every baseline.

Another way to estimate $T_{\rm sys}$—which we use in this paper—is to use auto-correlation visibilities, i.e., visibilities formed by correlating a single antenna's data with itself. The system temperature on a non-auto correlation baseline $\{p,q\}$ is then related to the geometric mean of the auto-correlation visibilities of the two constituent antennas as (Jacobs et al. 2015)

$$\sqrt{V(\{p,p\})V(\{q,q\})} = \frac{2k_b \nu^2 \Omega_p}{c^2} T_{\rm sys,\{p,q\}}\,. \tag{29}$$

In Figure 2 we plot the system temperatures predicted using both methods for some HERA data. The lower scatter with the second method is why we recommend its usage.

The noise power spectrum $P_{\rm N}$ correctly describes the error bars *assuming that our instrument measures nothing but noise.* This may be a suitable approximation for noise-dominated delays. More generally, however,

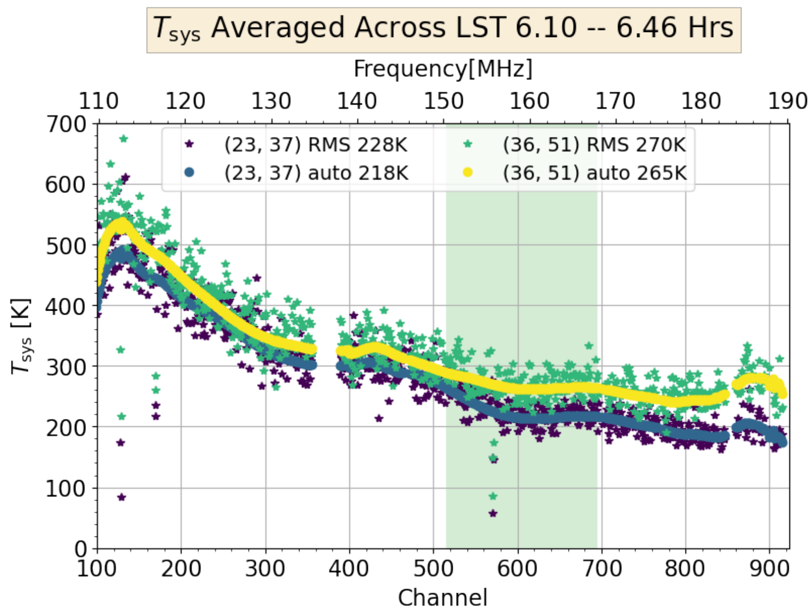

Figure 2. Comparison of two ways to estimate the system temperature based on HERA data. The system temperatures of cross-correlation visibilities on two 14.6 m baselines [indexed by HERA antenna numbers (23, 37) and (36, 51)] are averaged across the LST range of 6.10 to 6.46 hours. The green regime, from frequency channel number #515 to 695, show the HERA data band used for analysis in this paper. The label "autos" and "RMS" indicate the method (either from products of auto-visibilities or the RMS of differenced visibilities) by which the curves of system temperatures are calculated. And the values of temperatures shown in labels are the average values over the band specified by the green regime. We see the results from two methods are consistent to 5%, though the curves from auto-correlations are far less scattered.

when a signal (be it foregrounds or systematics) exists, the cross terms of Equation (17) provide an additional contribution to the noise scatter/error bars.[8] This term exists regardless of whether one's foreground mitigation strategy is based on subtraction or avoidance. In the former case, the foreground residuals after subtracting a model from data enter into the final expression; in the latter case, the whole foreground contribution is propagated as a systematic signal in the data. We show how to take this into account in Appendix D, where we define $P_{\rm SN}$ as

$$P^2_{\rm SN} \equiv \sqrt{2}\mathrm{Re}(P_{\tilde{x}_1\tilde{x}_2})P_{\rm N} + P^2_{\rm N} \tag{30}$$

which serves as a characterization of the error bars on the total sky emission, consistent with the form derived in Kolopanis et al. (2019). Here, $\mathrm{Re}(P_{\tilde{x}_1\tilde{x}_2})$, the real part of power spectra formed from $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, serves as a stand-in for a signal-only power spectrum $P_{\rm S}$ assuming that the signal dominates the noise (whether this "signal" takes the form of foregrounds, systematics, or the cosmological signal).

[8] We stress that this scatter/error is still due to instrumental noise and not the variance of the signal term. Even for a perfectly constant and known signal, the presence of the cross term alters the uncertainty, essentially having the signal term act as a multiplicative amplifier for noise fluctuations.

Using real data helps us approximate the true $P_{\rm S}$ when we do not possess good *a priori* models. However, by using real data our estimate of the first term of Equation (30) can in principle be negative because $\tilde{x}_1$ and $\tilde{x}_2$ contain different noise realizations. This can cause problems, since the signal-only power spectrum is expected to be non-negative. We thus enforce a hard prior on this term and set negative values of $\mathrm{Re}(P_{\tilde{x}_1\tilde{x}_2})$ to zero. In this way $P^2_{\rm SN}$ is always positive and the error bar $P_{\rm SN}$ is at worst a conservative estimate. When we average power spectra with error bars, this conservatism leads to a substantial bias between $P_{\rm SN}$ and $P_{\rm N}$ in our final error estimates in the noise-dominated regime. This is due to $\mathrm{Re}(P_{\tilde{x}_1\tilde{x}_2})$ in the first term of Equation (30) is empirical—and therefore contains noise—which effectively yields a double-counting of the noise-noise term in the variance. This double-counting does not result in an average bias if one does not enforce our prior, since in a noise-dominated regime $\mathrm{Re}(P_{\tilde{x}_1\tilde{x}_2})$ has zero mean. Our prior ensures that $P_{\rm SN} > P_{\rm N}$. Despite this, we will show that Equation (30) is a reasonable approximation over broad swaths of the power spectrum. Moreover, if we understand the statistics of noise fluctuations, one can simply predict—and correct for—the double-counting bias in $P_{\rm SN}$. In the noise-dominated regime, $P_{\rm N}$ characterizes the scatter in $\mathrm{Re}(P_{\tilde{x}_1\tilde{x}_2})$. Thus one can estimate the expectation value of the extra noise contribution from the first term of Equation (30) by computing

$$\begin{aligned}&\sqrt{2}\langle \mathrm{Re}(P_{\tilde{x}_1\tilde{x}_2})\rangle P_\mathrm{N}\\ =&\sqrt{2}\left[\frac{1}{\sqrt{2\pi}P_\mathrm{N}}\int_0^\infty y\exp\left(-y^2/2P_\mathrm{N}^2\right)\mathrm{d}y\right]P_\mathrm{N}\\ =&P_\mathrm{N}^2/\sqrt{\pi}\,. \end{aligned} \tag{31}$$

The integral runs over only positive values since we are imposing a non-negative prior. Note that here where we have neglected any complicated window function effects in inserting the measured power spectrum, essentially assuming that all power is locally sourced at the delay where it is measured. In principle, these effects can be taken into account in a more general derivation within the quadratic estimator formalism, but we leave this for future work.

We see from Equation (31) that the excess of P_SN above P_N in the noise-dominated regime is proportional to P_N; thus, we can just subtract it from the initially computed P_SN. We then define a modified "P_SN" free from the double-counting noise bias as[9]

$$\tilde{P}_\mathrm{SN} \equiv P_\mathrm{SN} - \left(\sqrt{1/\sqrt{\pi}+1}-1\right)P_\mathrm{N}\,. \tag{32}$$

The reduction of double-counting noise bias in this way also holds where signal dominates over noise. Since P_N, P_SN, and $\tilde{P}_\mathrm{SN}$ are all either power spectra or constructed from products of power spectra, we name this methodology of error estimation the "Power Spectrum Method".

3.4. *Covariance Method*

The quadratic estimator formalism leads to a natural way to write down an analytic form of error bars by propagating the input covariance matrices on visibilities into the output covariance matrices on bandpowers, which we name "Covariance Method" (see Appendix E for more details). Provided three set of matrices below containing the full frequency-frequency two-point correlation information of complex visibilities

$$\begin{aligned}\boldsymbol{C}^{12}_{ij} \equiv&\langle \boldsymbol{x}_{1,i}\boldsymbol{x}^*_{2,j}\rangle\,,\\ \boldsymbol{U}^{12}_{ij} \equiv&\langle \boldsymbol{x}_{1,i}\boldsymbol{x}_{2,j}\rangle\,,\\ \boldsymbol{G}^{12}_{ij} \equiv&\langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}^*_{2,j}\rangle\,,\end{aligned} \tag{33}$$

the variance in the real part of $\hat{P}_\alpha$ is

$$\begin{aligned}&\mathrm{var}\left[\mathrm{Re}(\hat{P}_\alpha)\right]\\ &=\frac{1}{4}\Big\{\mathrm{tr}\Big[(\boldsymbol{E}^{12,\alpha}\boldsymbol{U}^{22}\boldsymbol{E}^{21,\alpha*}\boldsymbol{G}^{11}+\boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{21}\boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{21})\\ &+2\times(\boldsymbol{E}^{12,\alpha}\boldsymbol{U}^{21}\boldsymbol{E}^{12,\alpha*}\boldsymbol{G}^{21}+\boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{22}\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{11})\\ &+(\boldsymbol{E}^{21,\alpha}\boldsymbol{U}^{11}\boldsymbol{E}^{12,\alpha*}\boldsymbol{G}^{22}+\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{12}\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{12})\Big]\Big\}\,,\end{aligned} \tag{34}$$

and the variance in the imaginary part of $\hat{P}_\alpha$ is

$$\begin{aligned}&\mathrm{var}\left[\mathrm{Im}(\hat{P}_\alpha)\right]\\ &=\frac{-1}{4}\Big\{\mathrm{tr}\Big[(\boldsymbol{E}^{12,\alpha}\boldsymbol{U}^{22}\boldsymbol{E}^{21,\alpha*}\boldsymbol{G}^{11}+\boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{21}\boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{21})\\ &-2\times(\boldsymbol{E}^{12,\alpha}\boldsymbol{U}^{21}\boldsymbol{E}^{12,\alpha*}\boldsymbol{G}^{21}+\boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{22}\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{11})\\ &+(\boldsymbol{E}^{21,\alpha}\boldsymbol{U}^{11}\boldsymbol{E}^{12,\alpha*}\boldsymbol{G}^{22}+\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{12}\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{12})\Big]\Big\}\,,\end{aligned} \tag{35}$$

To get the final error bar on power spectra, we should accurately model input covariance matrices on visibilities and propagate them into output covariance matrix on bandpowers. Generally, we assume that the input covariance matrices can be decomposed as $\boldsymbol{C} \equiv \boldsymbol{C}_\mathrm{signal} + \boldsymbol{C}_\mathrm{noise}$.

Assuming the distributions of the real and imaginary parts of noise in visibilities are independently and identically distributed (IID) at the same frequency and are uncorrelated between different frequency channels, our expressions simplify considerably. With these assumptions, $\boldsymbol{C}^{11}_\mathrm{noise}$ and $\boldsymbol{C}^{22}_\mathrm{noise}$ are diagonal and $\boldsymbol{C}^{12}_\mathrm{noise}$, $\boldsymbol{U}^{11}_\mathrm{noise}$, $\boldsymbol{U}^{22}_\mathrm{noise}$, $\boldsymbol{U}^{12}_\mathrm{noise}$, $\boldsymbol{G}^{11}_\mathrm{noise}$, $\boldsymbol{G}^{22}_\mathrm{noise}$ and $\boldsymbol{G}^{12}_\mathrm{noise}$ are all zero. Analogous to Equation (29), one can estimate the diagonal terms of $\boldsymbol{C}^{11}_\mathrm{noise}$ and $\boldsymbol{C}^{22}_\mathrm{noise}$ using the amplitudes of auto-correlation visibilities. For a baseline $\{p,q\}$ composed by two antennas p and q, its $\boldsymbol{C}_\mathrm{noise}$ is

$$\begin{aligned}\boldsymbol{C}^{\{p,q\},\{p,q\}}_{\mathrm{noise},ii}(t) \equiv&\ \langle V_\mathrm{noise}(\{p,q\},\nu_i,t)V^*_\mathrm{noise}(\{p,q\},\nu_i,t)\rangle\\ \approx&\ \left|\frac{V(\{p,p\},\nu_i,t)V(\{q,q\},\nu_i,t)}{N_\mathrm{nights}B\Delta t}\right|\,,\end{aligned} \tag{36}$$

where $B\Delta t$ is the product of the channel bandwidth and the integration time, and N_nights is the total number of nights of data analyzed from a drift scan telescope.

[9] Here we derived the correction factor $\sqrt{1/\sqrt{\pi}+1}-1 \approx 0.251$ assuming $\mathrm{Re}(P_{\tilde{x}_1\tilde{x}_2})$ follows Gaussian distribution. This is appropriate assuming that enough power spectra formed from data at different times have been incoherently averaged together for the Central Limit Theorem to apply (we will examine this point further in Section 4.1). For a single snapshot in time, the measured power spectrum follows a Laplacian distribution (again, see Section 4.1) and the correction factor becomes $\sqrt{3/2}-1 \approx 0.225$. Since the difference is small and in practice we operate in the Gaussianized regime anyway we use $\sqrt{1/\sqrt{\pi}+1}-1$ in our definition.

Inserting only $\boldsymbol{C}_{\text{noise}}$ for $\boldsymbol{C}$ in Equations (34) and (35), we have another estimate on the noise power variance as

$$\begin{aligned}\text{var}\left[\text{Re}(\hat{P}_\alpha)\right] &= \text{var}\left[\text{Im}(\hat{P}_\alpha)\right] \\ &= \frac{1}{2}\left\{\text{tr}\left[\boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{22}_{\text{noise}}\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{11}_{\text{noise}}\right]\right\} \\ &= \sigma^2_{\text{QE-N}}\,. \end{aligned} \tag{37}$$

By taking the trace on the products of matrices, we have in fact taken a weighted average of covariance information over frequencies. The quantity $\sigma_{\text{QE-N}}$ should be equal to P_{N} from the previous subsection, provided that in computing T_{sys} using Equation (27) we average over frequencies to obtain an effective T_{sys} in the same way. In this way, we see that the analytic noise power spectrum essentially reduces to a special case of Equation (37).

Of course, the fully covariant treatment here also implicitly includes the signal-noise cross terms discussed in previous sections. Including both $\boldsymbol{C}_{\text{signal}}$ and $\boldsymbol{C}_{\text{noise}}$ in $\boldsymbol{C}$ gives

$$\begin{aligned}\text{var}\left[\text{Re}(\hat{P}_\alpha)\right] &= \text{var}\left[\text{Im}(\hat{P}_\alpha)\right] \\ &= \frac{1}{2}\left\{\text{tr}\left[\boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{22}_{\text{noise}}\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{11}_{\text{noise}}\right.\right. \\ &\quad + \boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{22}_{\text{signal}}\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{11}_{\text{noise}} \\ &\quad \left.\left. + \boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{22}_{\text{noise}}\boldsymbol{E}^{21,\alpha}\boldsymbol{C}^{11}_{\text{signal}}\right]\right\} \\ &= \sigma^2_{\text{QE-SN}}\,. \end{aligned} \tag{38}$$

Since we have assumed only $\boldsymbol{C}^{11}_{\text{noise}}$ and $\boldsymbol{C}^{22}_{\text{noise}}$ are nonzero, the extra signal-noise cross terms entering into the expression are just their couplings with the signal counterparts. For that last contribution, we estimate $\boldsymbol{C}_{\text{signal}}$ as

$$\boldsymbol{C}^{11}_{\text{signal},ij} = \boldsymbol{C}^{22}_{\text{signal},ij} = \frac{1}{2}\left[\boldsymbol{x}_{1,i}\boldsymbol{x}^*_{2,j} + \boldsymbol{x}_{2,i}\boldsymbol{x}^*_{1,j}\right]\,. \tag{39}$$

Note that this way of modelling $\boldsymbol{C}_{\text{signal}}$ is Hermitian and noise-bias free when taking the ensemble average, but not positive definite. With a similar argument to P_{SN} in subsection 3.3, we enact a hard non-negative prior on $\boldsymbol{C}_{\text{signal}}$, where rows and columns containing negative diagonal elements are set to zero. This procedure can be shown to give signal-noise cross terms in Equation (38) that are always non-negative. However, this means that $\sigma_{\text{QE-SN}}$ suffers from the same double-counting noise bias with P_{SN}, and analogously we may construct a modified "$\sigma_{\text{QE-SN}}$" which is also free from the bias as

$$\tilde{\sigma}_{\text{QE-SN}} = \sigma_{\text{QE-SN}} - \left(\sqrt{1/\sqrt{\pi}+1} - 1\right)\sigma_{\text{QE-N}}\,. \tag{40}$$

Generally speaking, the power spectrum method of the previous subsection is a special case of the covariance method of this subsection. For example, if we estimate P_{N} in a way that carefully accounts for the frequency dependence of T_{sys}, we should find that when we insert it into the expression for P_{SN} that $P_{\text{SN}} = \sigma_{\text{QE-SN}}$. The covariance method has the advantage of providing off-diagonal covariances between different bandpowers in addition to variances.

3.5. *Summary*

The methods of error bar estimation introduced in this section can be categorized into two groups:

- $\sigma_{\text{bs}}, P_{\text{SN}}, \sigma_{\text{QE-SN}}$: these estimate error bars on the total emission, including both contributions from signal-noise cross terms and noise-noise terms.
- P_{diff}, $P_{\text{N}}, \sigma_{\text{QE-N}}$: these estimate the error bar in the limit of noise-dominated (or noise level), only including contributions from the noise-noise terms.

Before we jump into a quantitative discussion using the HERA power spectrum pipeline to compute these error bars in the next section, it is important to stress that there are other methods of error estimation that we do not cover in this paper. For example, LOFAR has used the Stokes V parameter as an estimator of noise level (Patil et al. 2017; Gehlot et al. 2019; Mertens et al. 2020) since the astrophysical sky is expected to exhibit only extremely weak circular polarization. However, reliably estimating Stokes V power requires more accurate polarization calibration solutions than that are currently available for HERA (Kohn et al. 2019). Since one of our goals is to test our error estimation methods on HERA data, we will omit discussion of Stokes V techniques in this paper.

4. TESTS

In this section, we quantitatively examine the error estimation methods introduced in Section 3. We apply them to 21 cm delay power spectra estimated from both simulated data and HERA Phase I data. We directly compare the relative amplitudes of the error bars predicted by each method, delay mode by delay mode. We also study how the error bars respond to systematics and foregrounds in different regimes of delay space.

4.1. *Simulations from a Toy Model*

We start with simulations from a toy model. This allows us to generate a large number of realizations, with which we can numerically test the validity of our error bars in the ensemble-averaged limit. Our simulated visibilities include only the foregrounds and noise. For the

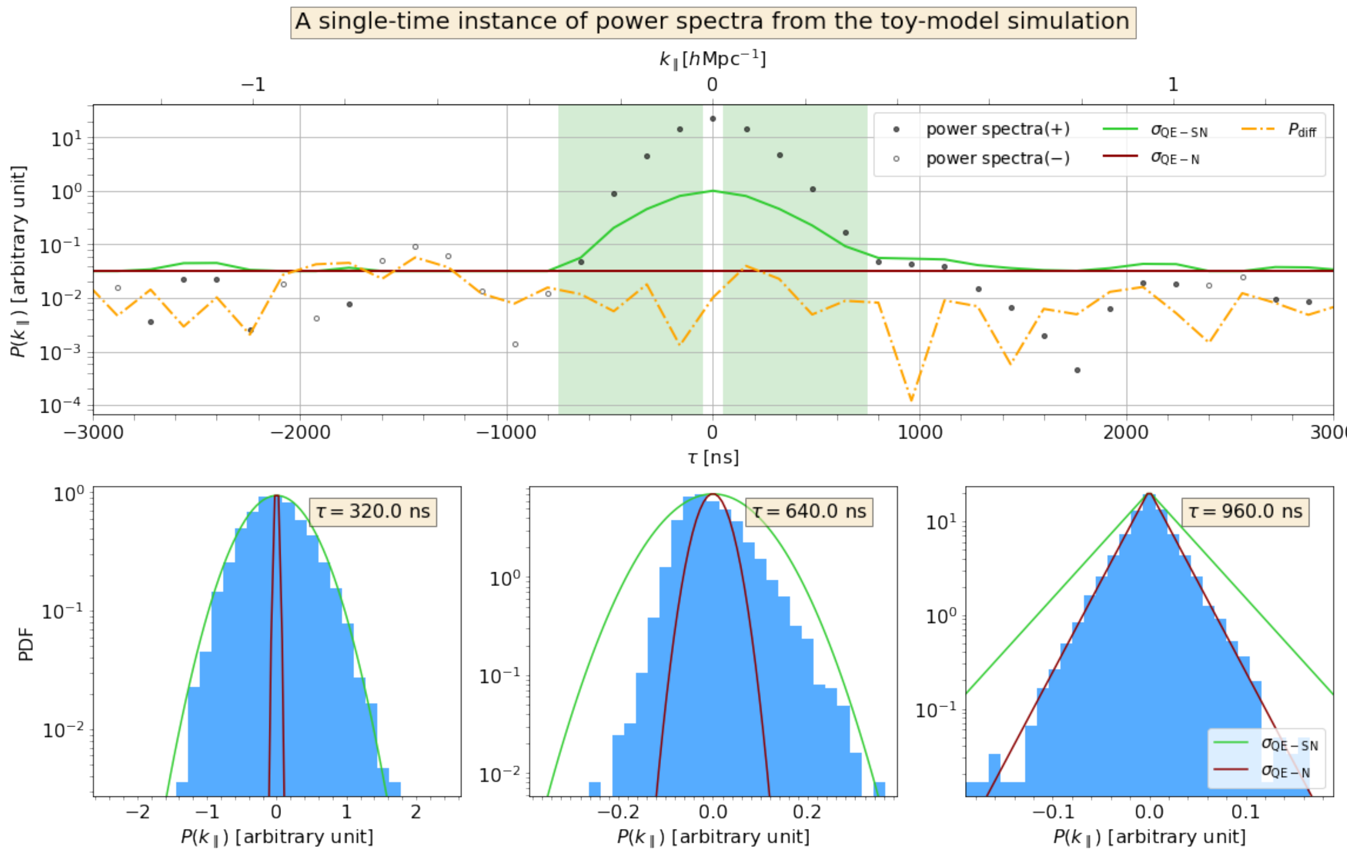

Figure 3. Error bars on single-baseline-pair power spectra at one timestamp from simulations described in Section 4.1. *Top*: We plot power spectra together with error bar types $P_{\mathrm{diff}}, \sigma_{\text{QE-SN}}$ and $\sigma_{\text{QE-N}}$. The green shaded regime ranges from ± 50 ns to ± 750 ns, where the foreground power is dominant over the noise power. *Bottom*: We plot histograms of bandpowers from ~ 10000 realizations at $\tau = 320.0$ (strongly foreground-dominated regime), 640.0 (transition regime), 960.0 (noise-dominated regime) ns respectively, along with probability distribution function (PDF) curves predicted using the $\sigma_{\text{QE-SN}}$ and $\sigma_{\text{QE-N}}$ values at the same delay. At $\tau = 320.0, 640.0$ ns, the PDF takes a Gaussian form. At $\tau = 960.0$ ns, the PDF takes the form of a Laplacian. The $P(k_\parallel)$ values used in the histograms have been subtracted from the mean value of all realizations. We can see error bars are roughly comparable to each other in amplitudes in the noise-dominated regime. At $\tau = 320.0$, the envelope of the histogram matches exactly with the PDF using $\sigma_{\text{QE-SN}}$. At $\tau = 960.0$, the envelope of the histogram matches the PDF using $\sigma_{\text{QE-N}}$, while we see the PDF using $\sigma_{\text{QE-SN}}$ is broader. Therefore, using $\sigma_{\text{QE-SN}}$ will lead to a more conservative estimate of errors in this delay regime.

foreground portion of the visibilities we draw a random visibility from a frequency-frequency covariance matrix of the form $\boldsymbol{C}_{ij} = A\exp\left[-(\nu_i - \nu_j)^2/l^2\right]$, where A and l characterize the amplitude and correlation length of the foreground signal, respectively. The adopted covariance model creates smoothly varying functions in frequency space, which is roughly in accordance with the relatively flat spectral structure of real foregrounds. Here we simulate visibilities on two redundant baselines for 20 consecutive timestamps. We set $A = 25$ and $l = 5$MHz, and the foreground visibilities are kept the same on each baseline and over all timestamps. The noise components of the visibilities on each baseline at each timestamp are independently drawn from the same white Gaussian distribution $\mathcal{N}(0, \sigma^2 = 1)$. We produce ~ 10000 realizations of such visibilities and then use `hera_pspec` code[10] to estimate the delay power spectra and to compute the error bars discussed previously.

In Figure 3, we plot power spectra together with a few of the error bar types computed from one timestamp of data from the simulations. We compute P_{diff} by differencing visibilities between the one timestamp and the next. We use Equation (37) and (38) to calculate error bars of the "covariance method", while we evaluate $\boldsymbol{C}_{\mathrm{noise}}$ using the exact covariance matrix from which noise visibilities are drawn, since we did not simulate visibilities on auto-correlation baselines. In the top panel of Figure 3, the green shaded regime (which

[10] https://github.com/HERA-Team/hera_pspec

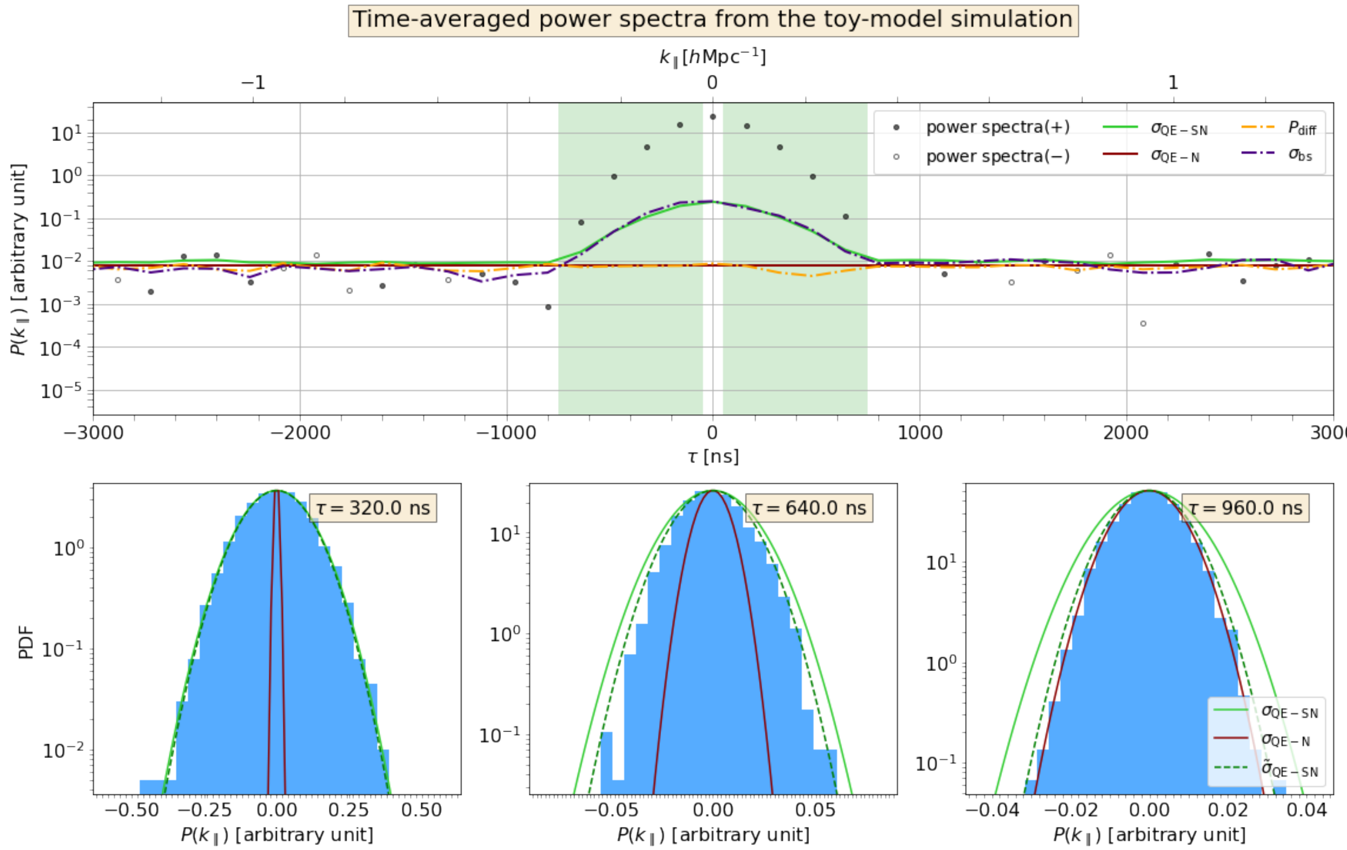

Figure 4. Error bars on time-averaged power spectra over 20 timestamps from simulations in Section 4.1. The figure follows similar conventions to Figure 3, except *Top*: σ_{bs} is added; *Bottom*: All PDFs take the forms of Gaussian and the ones specified by $\tilde{\sigma}_{\text{QE-SN}}$ are appended. We observe good agreement between σ_{bs} and $\sigma_{\text{QE-SN}}$ in the foreground-dominated regime, and the consistency of all types of labeled error bars in the noise-dominated regime. After the incoherent average, we see histograms at all delays become Gaussian. Additionally, $\tilde{\sigma}_{\text{QE-SN}}$ is clearly different from $\sigma_{\text{QE-SN}}$ where the signal is less dominant. Especially at $\tau = 960.0$ ns, the PDF using $\tilde{\sigma}_{\text{QE-SN}}$ is closer to the exact noise-dominated version using $\sigma_{\text{QE-N}}$.

ranges from $\pm 50\,$ns to $\pm 750\,$ns) is where the foreground power is dominant over the noise power. We see that P_{diff} and $\sigma_{\text{QE-N}}$ are insensitive to the foreground power in this regime, and when moving to higher delays, the noise levels characterized by P_{diff}, $\sigma_{\text{QE-N}}$, and $\sigma_{\text{QE-SN}}$ are very close to one another. Compared to the other two, P_{diff} shows much more scatter from delay to delay since it is a more empirical estimation of noise based on examining what amounts to noise *realizations*. Notice also that as expected by construction, the $\sigma_{\text{QE-SN}}$ curve always lies above $\sigma_{\text{QE-N}}$, due to the fact we enforce a zero clipping on the signal-noise cross term.

In the bottom panel of Figure 3, we plot histograms of power spectra at three delays ($\tau = 320.0, 640.0$ and $960.0\,$ns) by accumulating data points from ~ 10000 realizations. The results here are therefore representative of ensemble-averaged expectations. At each delay, we also plot theoretical predictions for the probability distribution functions (PDFs). Precisely what form these PDFs take will depend on the delay. In the low-delay regime, Equation (17) shows the variation comes from single powers of visibility noise, which we assume is Gaussian. (Recall that we are not modelling the signal as a random field, in the sense that it does not participate in our ensemble average.) The result is a Gaussian PDF. At high delays Equation (17) shows that the power spectrum is the cross-multiplication of two independent realization of noise. The resulting PDF is a Laplacian. Both of these distributions take one free parameter (the standard deviation of power) and we show predictions where this standard deviation is specified by $\sigma_{\text{QE-SN}}$ and $\sigma_{\text{QE-N}}$. At $\tau = 320.0$ and $640.0\,$ns, we plot Gaussian reference PDFs. At $\tau = 960.0\,$ns, we plot a Laplacian reference PDF. We see at $\tau = 320.0\,$ns, where foreground power is overwhelmingly dominant, the shape of the histogram is indeed Gaussian-like, and its envelope matches the PDF curves using $\sigma_{\text{QE-SN}}$. At $\tau = 960.0$ where noise is dominant, the shape of the histogram is indeed Laplacian-like, and its envelope matches the PDF curves using $\sigma_{\text{QE-N}}$ (since $\sigma_{\text{QE-N}}$ does not suffer from the conservatism of $\sigma_{\text{QE-SN}}$ discussed in Section 3.3). With $\tau = 640.0\,\text{ns}$ we have a transition case between the two extremes. The distribution of power spectra will be skewed since neither the signal nor the noise dominates in this occasion (for a mathematical proof of the skewness see Appendix F). The histogram does not match the PDF predicted by either standard deviation, but note from the widths of the PDFs that an error bar given by $\sigma_{\text{QE-SN}}$ is a conservative error, as we designed it to be.

In Figure 4, we present the same types of error bars plus a bootstrapped one on power spectra which were formed by incoherently averaging over 20 timestamps. We see in the green regime that σ_{bs} agrees with $\sigma_{\text{QE-SN}}$. All the different kinds of error bars agree well with each other in the noise dominated regime, and with the extra time averaging step (compared to Figure 3) P_{diff} exhibits less scatter. Again, we plot histograms of the averaged power spectra from Monte-Carlo simulations against Gaussian PDF curves at $\tau = 320.0, 640.0$ and $960.0\,\text{ns}$. One feature to note from the histogram is that each distribution has become nearly Gaussian. This is simply due to the Central Limit Theorem as power spectra are averaged together incoherently. In addition to $\sigma_{\text{QE-SN}}$ and $\sigma_{\text{QE-N}}$, we also plot the PDFs using $\tilde{\sigma}_{\text{QE-SN}}$ which eliminates the double-counting bias in $\sigma_{\text{QE-SN}}$. It is as expected that the PDF using $\tilde{\sigma}_{\text{QE-SN}}$ is more close to the one using $\sigma_{\text{QE-N}}$ at the noise-dominated delay mode.

4.2. *Application to HERA Phase I Data*

The HERA Phase I data used for analysis in this paper consists of 18 observing nights taken in the Karoo Desert, South Africa from December 10th to 28th, 2017. The HERA array consisted of ~ 40 functional antennas during observations, which were taken across a 100 to 200 MHz band comprised of 1024 channels and dual polarization "X" and "Y" feeds. [See Table 1 of Kern et al. (2020b) for more details on the array and correlator specifications during the observations.] The data used in this work were first preprocessed with the HERA analysis pipeline (internally called H1C IDR2.2[11]). This includes automated metric evaluation and data flagging for faulty antennas and radio frequency interference (RFI). In addition, the data are redundantly calibrated (Dillon et al. 2020), absolutely calibrated (Kern et al. 2020b), binned and averaged across observing nights, in-painted over RFI gaps in frequency and then treated for known instrumental systematics (Kern et al. 2020a).

[11] http://reionization.org/manual_uploads/HERA069_IDR2.2_Memo_v3.html

We pick a slice of HERA Phase I visibilities taken from a 14.6-m redundant baseline group during an LST range of 5.75 to 6.10 hours. The visibilities in each timestamp are integrated over ~ 10 seconds. We select visibilities falling within a 150.3 to 167.8 MHz band to compute power spectra. We use pseudo-Stokes I visibilities V_{pI}, which are constructed by combining the visibilities from a cross correlation of two X feeds ("XX") and a cross-correlation two Y feeds ("YY") as follows:

$$V_{\text{pI}} = \frac{1}{2}\left(V_{\text{XX}} + V_{\text{YY}}\right). \tag{41}$$

In forming the delay power spectra we cross correlate visibilities from different baselines (e.g., $\boldsymbol{b}_1$-$\boldsymbol{b}_2$, $\boldsymbol{b}_1$-$\boldsymbol{b}_3$, $\boldsymbol{b}_2$-$\boldsymbol{b}_3$, etc.) and between odd and even timestamps (e.g., t_1-t_2, t_3-t_4, t_5-t_6, etc.) to form delay power spectra. In this way, we obtain power spectra on 253 baseline-pairs at 30 timestamps.

We show the power spectra from one baseline-pair at one timestamp in Figure 5, together with error bar types $P_{\text{diff}}, \sigma_{\text{QE-SN}}, \sigma_{\text{QE-N}}, P_{\text{SN}}$, and P_{N}. The P_{diff} errors are computed from time-differenced visibilities, e.g., for power spectra at the cross timestamp $t_1 - t_2$ we form $V_{\text{diff}} \propto V(t_2) - V(t_1)$ and then we cross multiply V_{diff} from two different baselines to obtain the corresponding P_{diff} for that baseline pair. We calculate $\sigma_{\text{QE-SN}}$ and $\sigma_{\text{QE-N}}$ using Equations (38) and (37) with $\mathbf{C}_{\text{signal}}$ and $\mathbf{C}_{\text{noise}}$ specified by Equation (39) and (36). Equations (30) and (27) give the expressions for P_{SN} and P_{N}. See `hera_pspec` for detailed implementation.

In the top panel of Figure 5, we see all error bars agree well with each other in the noise-dominated regime (the red curve for P_{N} is almost exactly underneath the brown curve for $\sigma_{\text{QE-N}}$, making the former difficult to see; the same is true for the teal curve for P_{SN} versus the bright green curve for $\sigma_{\text{QE-SN}}$). The green shaded regime ranging from $\pm 20\,\text{ns}$ to $\pm 200\,\text{ns}$ is where foregrounds are expected to dominate. Here we see that P_{diff} also responds to the foreground power, similar to P_{SN} and $\sigma_{\text{QE-SN}}$. This tells us that the time-differenced visibilities contain non-negligible foreground residuals, which is not surprising since the sky is expected to evolve non-negligibly over the ~ 10 seconds of difference between our time samples.

In Section 3, we argued that the "covariance method" and the "power spectrum method" should be equivalent to each other. In the middle and bottom panels of Figure 5, we compute the relative difference in magnitudes between error bars, setting $\sigma_{\text{QE-SN}}$ and $\sigma_{\text{QE-N}}$ as the benchmarks respectively. We see that P_{SN} differs from $\sigma_{\text{QE-SN}}$ and P_{N} from $\sigma_{\text{QE-N}}$ by less than 1%, so they are essentially equivalent in our pipeline. On the other hand, P_{diff} can differ from $\sigma_{\text{QE-N}}$ at more than the 10%-

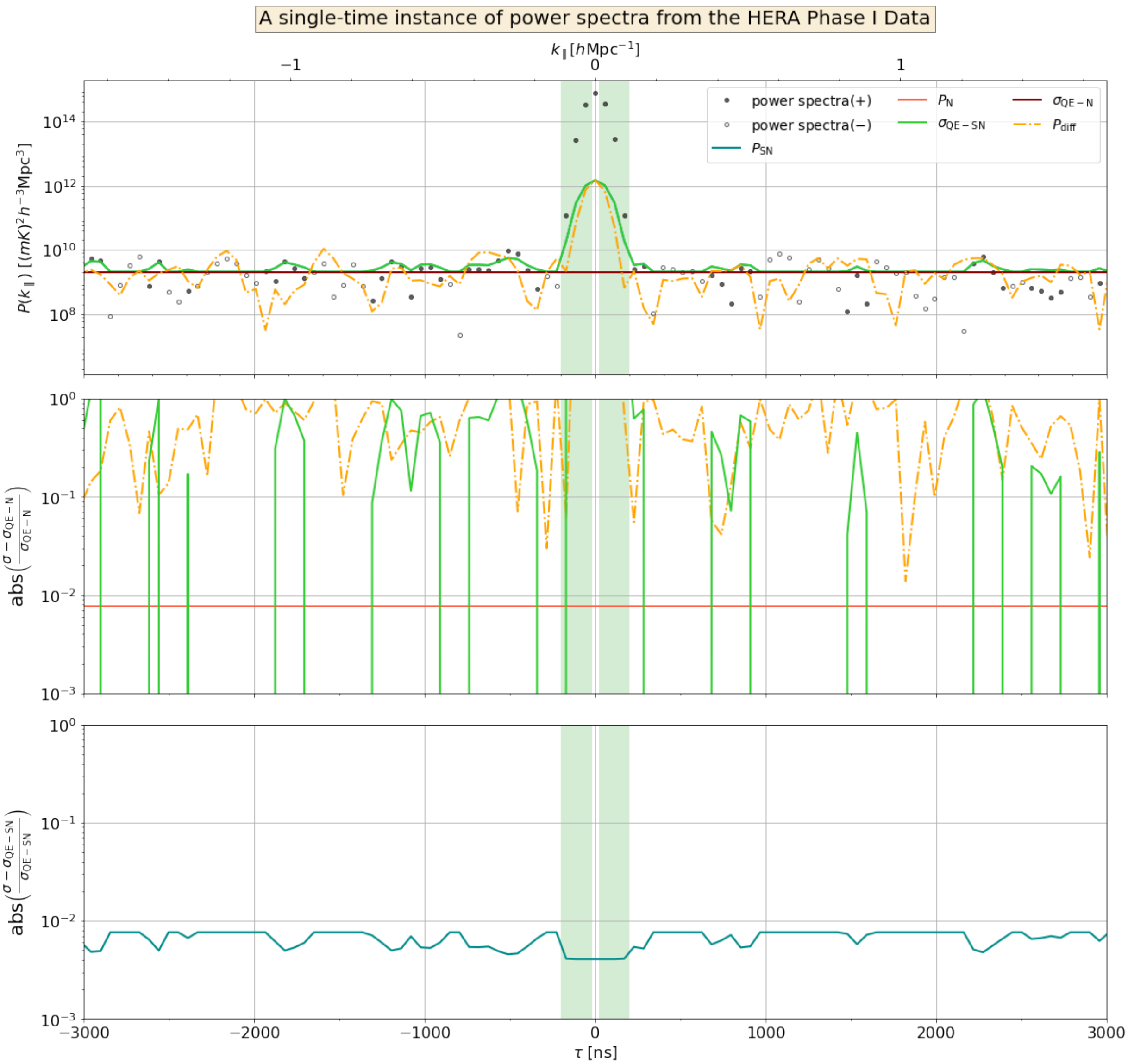

Figure 5. Error bars on single-baseline-pair power spectra at one timestamp from HERA Phase I data. The visibilities are selected from a band spanning 150.3 to 167.8 MHz. *Top*: Power spectra with error bars. The green shaded regime ranging from ±20 ns to ±200 ns is expected to be foreground dominated. *Middle*: Absolute relative difference between selected error bars with $\sigma_{\text{QE-N}}$. *Bottom*: Absolute relative difference between selected error bars with $\sigma_{\text{QE-SN}}$. We see numerically that P_{SN} differs from $\sigma_{\text{QE-SN}}$ by less than 1% and that the same is true for P_{N} and $\sigma_{\text{QE-N}}$.

level due to the fact that it is highly scattered. Note that $\sigma_{\text{QE-SN}}$ and P_{SN} are also scattered at some delays, whereas they are equal to $\sigma_{\text{QE-N}}$ and P_{N} at other delays. This is due to our imposition of a non-negative prior on the signal-noise cross term.

In Figure 6, we show the power spectra with error bars on the same baseline-pair as Figure 5, but with the further step of incoherently averaging over 30 time samples. We still see that all error bars (with bootstrap errors σ_{bs} added) agree well in the noise-dominated regime. At low delays, σ_{bs} peaks at an even higher value than $\sigma_{\text{QE-SN}}$. This is because the sky is not unchanged over different timestamps, so the bootstrapped error bars over time samples are inflated. After incoherently averaging, we still see P_{SN} differing from $\sigma_{\text{QE-SN}}$ and P_{N} differing from $\sigma_{\text{QE-N}}$ by less than 1%. On the other hand, P_{diff} and σ_{bs} differ from $\sigma_{\text{QE-N}}$ at roughly the 10% level in the noise-dominated regime. We also see that in the limit of noise domination, $\sigma_{\text{QE-SN}}$ has a relative bias over $\sigma_{\text{QE-SN}}$ by about 30%. Therefore, using $\sigma_{\text{QE-SN}}$ or P_{SN} leads to a conservative estimate of one's errors, as we expected. For comparing, we also plot results of $\tilde{\sigma}_{\text{QE-SN}}$, which eliminates the double-counting noise bias in $\sigma_{\text{QE-SN}}$. The relative difference between $\tilde{\sigma}_{\text{QE-SN}}$ and $\sigma_{\text{QE-N}}$ is re-

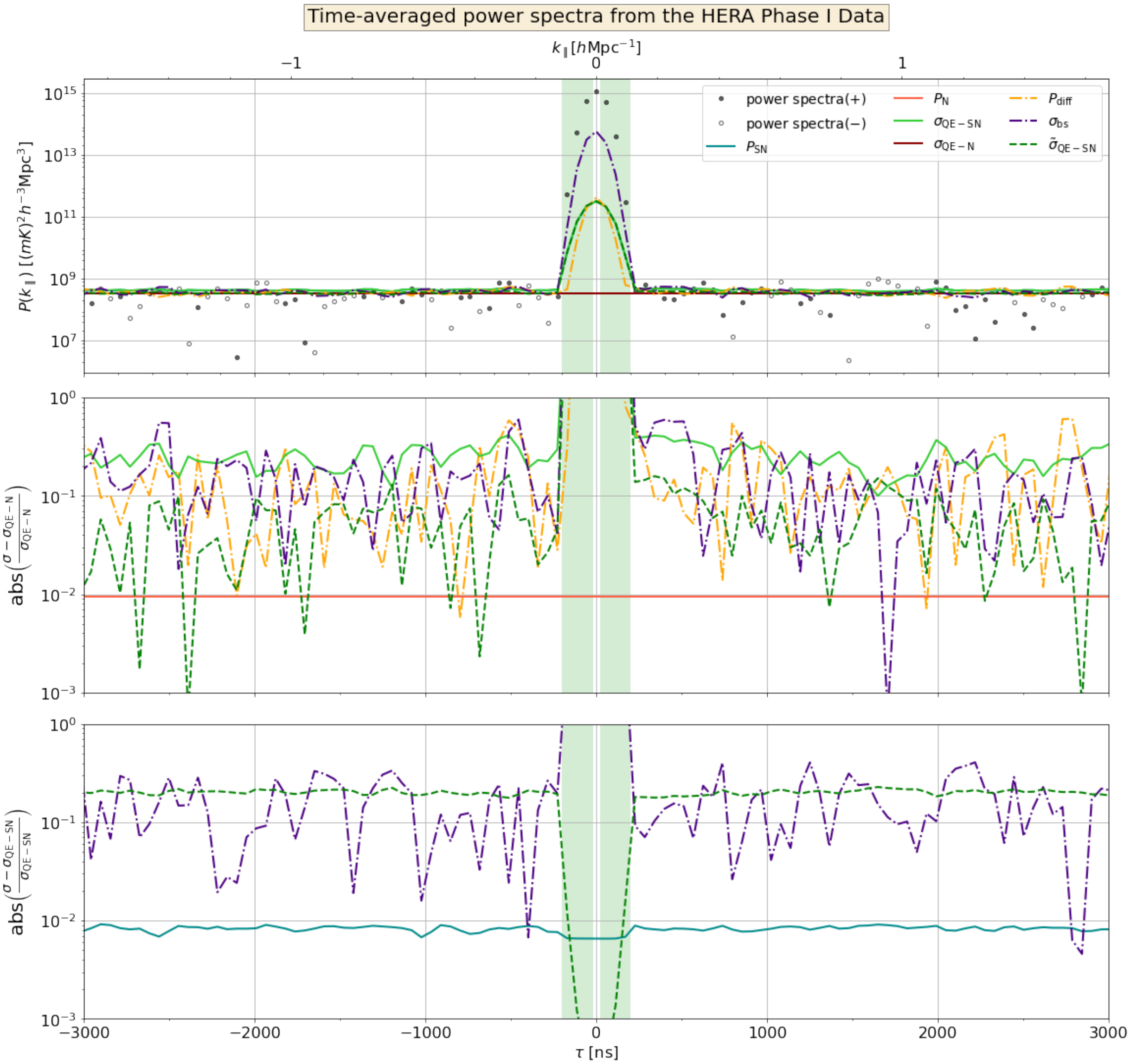

Figure 6. Error bars on single-baseline-pair power spectra incoherently averaged over 30 time samples from the same slice of HERA Phase I data as Figure 5. Our plotting conventions also follow those of Figure 5 for other conventions. We add results from $\tilde{\sigma}_{\text{QE-SN}}$ in each panel. In the center panel we see the relative difference between $\tilde{\sigma}_{\text{QE-SN}}$ and $\sigma_{\text{QE-N}}$ drops remarkably from $\sim 30\%$ to a few percent compared to the $\sigma_{\text{QE-SN}}$, demonstrating the effectiveness of our noise-double-counting bias removal. On the other hand, in the bottom panel we see that going from $\sigma_{\text{QE-SN}}$ to $\tilde{\sigma}_{\text{QE-SN}}$ results in significant changes only at the noise-dominated delays, and thus there one can always elect to use $\tilde{\sigma}_{\text{QE-SN}}$ even in foreground-dominated regimes.

duced to a few percents in the noise-dominated regime. While $\tilde{\sigma}_{\text{QE-SN}}$ is not significantly modified from $\sigma_{\text{QE-SN}}$ in the foreground-dominated regime. Thus if we want a compromise on reflecting the properties of the signal-noise cross term while not introducing noise bias, $\tilde{\sigma}_{\text{QE-SN}}$ might be our choice.

What we have established so far is the *relative* agreement (or lack thereof) between different types of error bars in different regimes. However, we have not yet established the *absolute* validity of these error bars on real data (i.e., we have not ruled out the possibility that they are all incorrect in the same way). For simulated power spectra we were able to compare the Monte-Carlo histograms with the PDF curves predicted from the error bars. The good match between the two gave us confidence in applying our error estimation methods. Might we perform similar analyses for power spectra from real data? Unfortunately, in real observations we only have one realization of the sky so that we cannot reach ensemble average limit by accumulating data points from a large number of realizations. Also, unlike simulated data with understood statistics, real data will contain

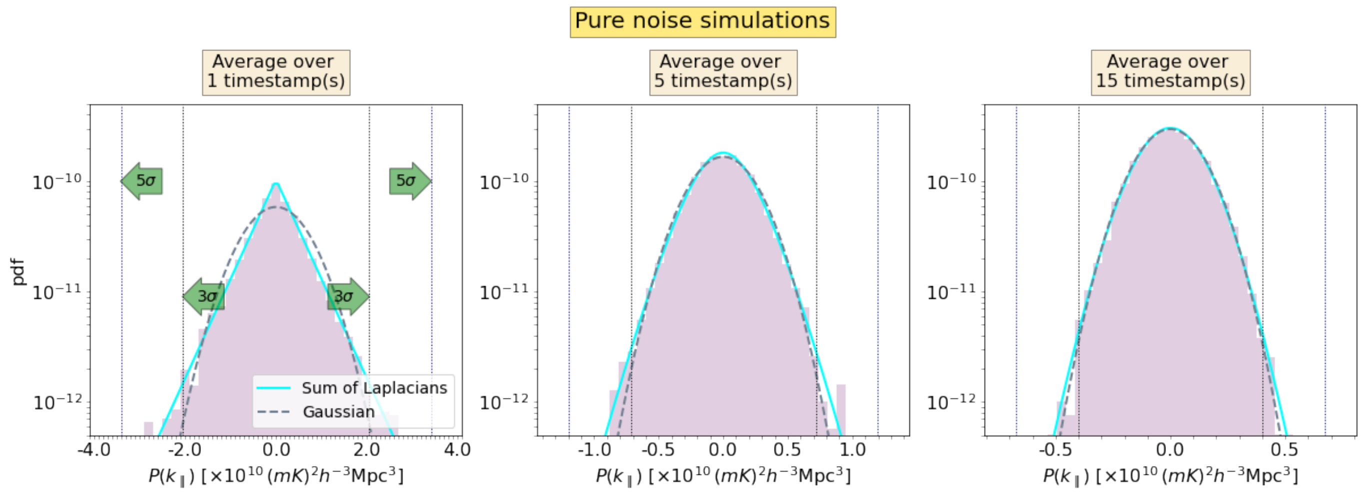

Figure 7. We plot the histograms of incoherently averaged power spectra over certain timestamps from pure noise simulations. The histogram in each column contains ~ 10000 data points. We compute $\sigma_{\text{QE-N}}$ and refer to Equation (G33) to evaluate the "Sum of Laplacians" PDF. Data points have been subtracted from the mean over all realizations. We also plot the equivalent Gaussian PDF with the same variance as the "Sum of Laplacians" PDF. The green arrows point to the dotted vertical lines representing "3σ" and "5σ", where σ is the square root of the variance of the predicted PDF. We see the envelopes of the histograms match the PDFs predicted using (G33) very well. As a check, the fractions of outliers beyond 3σ in each histogram are $(1.27\%, 0.57\%, 0.25\%)$, while the corresponding values from the predicted PDFs are $(1.34\%, 0.58\%, 0.22\%)$—a very close agreement. And with more time samples to be incoherently averaged, the shape of the histogram becomes increasingly Gaussian, which is a consequence of the central-limit theorem. As expected, we also see the distribution get narrower with more samples averaged together.

systematics that make their statistics more complicated and difficult to understand (although this may change as the field of 21 cm cosmology continues to mature).

For now, we may partially achieve our goal by checking the distributions of noise-like modes in our power spectra of real data. The noise-like modes refer to power spectra at higher delays where noise power is thought to be dominant and systematics are negligible. As we discussed in Section 3, we expect the noise visibilities to be Gaussian-distributed. This makes it possible to analytically compute the resultant statistics of power spectra. In Appendix G, we derive the mathematical form of the PDF of incoherently averaged noise-dominated power spectra. The final result, Equation (G33), shows that the correct PDF is a weighted sum of a series of Laplacian distributions. As a numeric test of the derivation, we produce Monte-Carlo histograms of incoherently averaged power spectra from pure Gaussian noise visibilities with an increasing number of averaged samples in Figure 7. We generate ~ 10000 realizations of power spectra with multiple time samples, and evaluate the power spectra at a single timestamp, as well as what it would be if incoherently averaged over 5 or 15 timestamps. For realizations at each time sample, we can calculate the error bar $\sigma_{\text{QE-N}}$ of the power spectra and substitute them into Equation (G33). It is clear that the predicted PDF matches the envelope of the histograms and that the shape of the histograms of averaged power spectra become increasingly Gaussian when averaging is over more timestamps. This is again a result of the Central Limit Theorem.

Confronting our results with real data, we use the power spectra from the same HERA Phase I data set as Figures 5 and 6 to generate the histograms. To accumulate sufficient data points for a histogram, we view all noise-like modes in power spectra over different redundant baseline-pairs as independent realizations. And we carry out the incoherent average over the time axis. Because the noise level at different baseline pairs may differ, all power spectra are first normalized by being divided over their corresponding $\sigma_{\text{QE-N}}$ and then subtracted from the mean of all data points. After the normalization, we have a uniform error bar $\sigma_{\text{QE-N}}$ for all data points at each time sample. We then make histograms and compare their envelopes with the PDF of "Sum of Laplacians" predicted using Equation (G33).

Before we jump to the results, we first take a look at the data set which includes RFI gap inpainting but without the removal of systematics. For histograms drawn in Figure 8, we evaluate the distributions of power spectra at delays larger than 2000 ns and at delays between 500 and 1500 ns, respectively. In the former case, we see the shape of histograms are perfectly consistent with the predicted PDF, and the distributions become more Gaussian and narrower with increasing number of averaged samples, similar to what we saw in Fig-

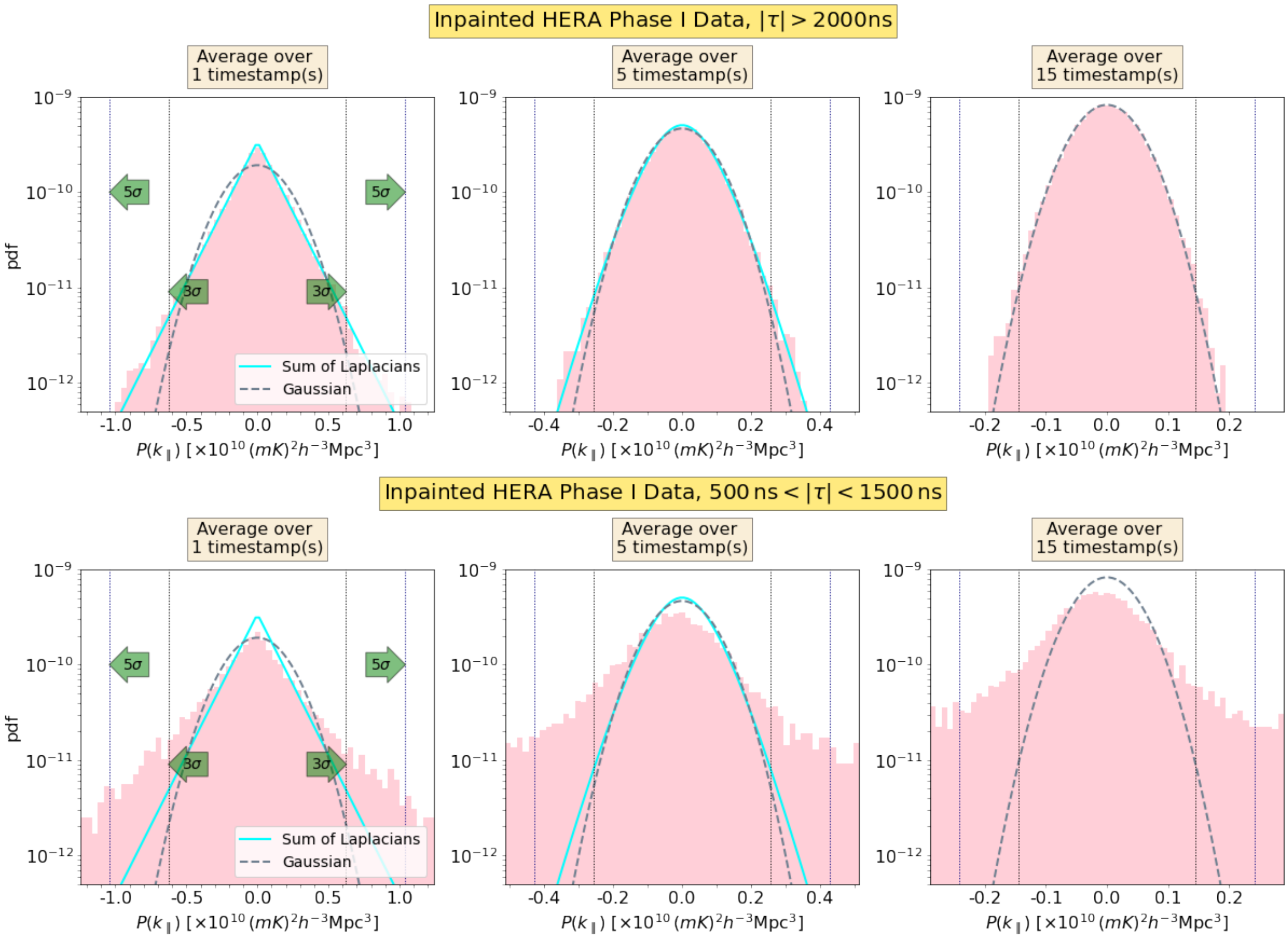

Figure 8. Histograms of power spectra at noise-like modes from the same HERA Phase I data used in Figure 5 and 6, including RFI gap inpainting, but without the removal of systematics. The data points are accumulated from power spectra at the same delays from different redundant baseline-pairs. Because their noise levels may differ, they are first normalized by dividing out their corresponding $\sigma_{\text{QE-N}}$ and then having the mean of all data points subtracted off. In this way we have a uniform $\sigma_{\text{QE-N}}$ for all points, and we use Equation (G33) to compute the "Sum of Laplacians" PDF. Refer to Figure 7 for other plotting conventions. *Top*: histograms from power spectra at all delays larger than 2000 ns, where there are ~ 27000 points in each column. *Bottom*: histograms from power spectra at delays between 500 and 1500 ns, where there are ~ 9000 points in each column. As a check, in the top panel, the fractions of outliers beyond 3σ in each histogram are $(1.49\%, 0.65\%, 0.40\%)$, which are close to the corresponding values from the predicted PDFs $(1.36\%, 0.57\%, 0.24\%)$. In the bottom panel, the fractions of outliers beyond 3σ in each histogram are $(7.95\%, 10.70\%, 11.46\%)$, which greatly exceed corresponding values from the predicted PDFs $(1.36\%, 0.57\%, 0.24\%)$.

ure 7. While in the latter case, we observe the histograms are flattened and much wider compared to the predicted PDF and there exist evidently hefty wings on either ends. Numerically, the fractions of outliers beyond 3σ in each histogram are $(7.95\%, 10.70\%, 11.46\%)$, which greatly exceed corresponding values from predicted PDFs $(1.36\%, 0.57\%, 0.24\%)$. This is a remarkable proof that significant systematics exist at lower delays in inpainted only data, as we expect.

We produce histograms for the systematics-removed data, as we used for Figures 5 and 6, in Figure 9. At delays larger than 2000 ns, we still see a good match between the Monte-Carlo histograms with the predicted PDFs. While at delays between 500 and 1500 ns, we see the deviations between histograms and PDFs are highly suppressed, compared to Figure 8. This is not surprising since we have exerted systematics removal. Though there is still a little excess above PDFs in histograms on far ends, this does not substantially affect the error bars that one might quote on a power spectrum measurement (which serve as a summary statistic for the main bulk of the PDF rather than its wings). However, such deviations are worth keeping an eye on, especially when performing rigorous jackknife or null tests in an

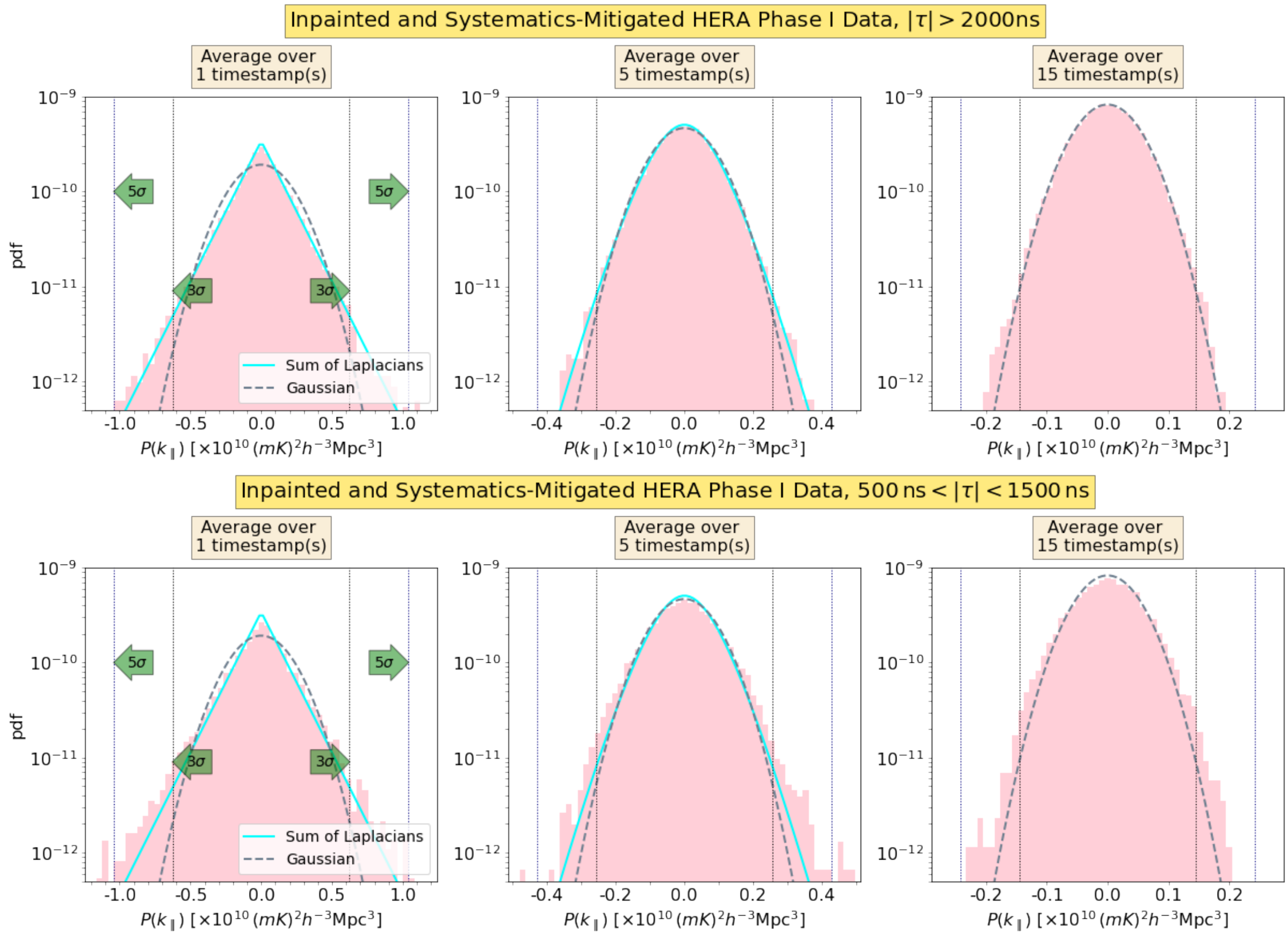

Figure 9. Histograms of power spectra at noise-like modes from inpainted and systematics-mitigated HERA Phase I data. The power spectra used here come from exactly the same data set as Figure 5 and 6. As a check, in the top panel the fractions of outliers beyond 3σ in each histogram are $(1.48\%, 0.63\%, 0.39\%)$, which are close to the corresponding values from the predicted PDFs $(1.36\%, 0.57\%, 0.24\%)$. And in the bottom panel, the fractions of outliers beyond 3σ in each histogram are $(2.19\%, 1.32\%, 0.80\%)$, which slightly exceed the corresponding values from the predicted PDFs $(1.36\%, 0.57\%, 0.24\%)$, but at a much lower level than the disagreement seen in Figure 8.

attempt to understand the systematics in one's instrument. As noted above, the excessive wings of the histograms in the bottom panel of Figure 8 can serve as a diagnostic tool for systematics that lead to deviations from Gaussian noise-like visibilities. They may also be used to investigate the related question of how instrumental systematics (e.g., Kern et al. 2019, 2020a) might affect the validity of one's error bars. There are of course limitations of this analysis, but we show the systematics does not effectively change the noise properties of power spectra at high delays. Readers should interpret Figure 8 and 9 as a quality check of HERA Phase I data, which shows the power spectra at high delays ($>$ 2000 ns) and at middle delays (500-1500 ns) after systematics mitigation are close to the predicted behaviors of Gaussian noise visibilities. Thus $\sigma_{\text{QE-N}}$ (along with other equivalent methods) validates itself a successful tool to characterize the noise statistics in real data. However, we will still quote $\tilde{\sigma}_{\text{QE-SN}}$ as a more robust error bar on reporting EoR upper limits at those delays. One should be aware that not all systematics can be cleanly corrected for, which mean that in principle the statistics can be much more complicated than the simple Gaussian distribution shown here. Along this theme, we urge readers to always perform consistency checks on the data, including but not limited to the ones we have performed here.

5. DISCUSSION

In previous sections, we have examined a number of different methods for assigning error bars to a HERA power spectrum. Here, we perform a comparison of the different types of error bars, highlighting the advantages and disadvantages of each.

We first consider the error bars using the "covariance method" ($\sigma_{\text{QE-N}}$ and $\sigma_{\text{QE-SN}}$) to those computed using the "power spectrum method" (P_{N} and P_{SN}).

- The "covariance method" error bars analytically take the covariance of the input visibilities and propagate them through to the output covariance of the bandpowers, via general formulae given by Equations (34) and (35). There are two weaknesses to this approach. First, the output errors will only be as good as the modeling of the input covariances. This modeling is particularly difficult for foregrounds and systematics, which can have statistical properties that are not entirely understood. In this paper, we adopt a strategy where we view systematics as non-random, and empirically estimate them from the real data. The other weakness of our "covariance method" is that our derivations rely on Gaussianity (Indeed, it would be strange for this method to only require an input *covariance*—a two-point function—if it were capable of capturing the effects of non-Gaussianity). This assumption will also be violated by foregrounds and systematics as well the cosmological signal (which is an effect that was modeled in Mondal et al. 2016, 2017; Shaw et al. 2019).

  Sidestepping these modeling restrictions on the "covariance method" are the noise-dominated bandpowers at high delays. In this regime, we use an input covariance matrix that is C_{noise} that is diagonal, with the diagonal elements set by the auto-correlation visibilities as Equation (36). The resulting error bars we call $\sigma_{\text{QE-N}}$ (see Table 3 for a reminder of our notation). These error bars are confirmed by tests on simulations and real data in Figure 7 and Figure 9, which verify that the error bars do properly account for the spread seen in an ensemble of Monte Carlo simulations. Further bolstering our confidence in using the "covariance method" are their agreement with other error metrics at our disposal. Figures 5 and 6 show that in the noise-dominated regime, the error bars using the "covariance method" are in excellent agreement with the bootstrap errors σ_{bs}, error bars using the 'power spectrum method', and the power spectrum of differenced data P_{diff}.

- The agreement between these different error estimation methods raises the question of why one might favour the "covariance method" over others. Consider first a comparison between $\sigma_{\text{QE-N}}$ and P_{N} from the "power spectrum method". These two methods are in fact quite similar, because P_{N} is also an analytically propagated measurement of error, as one can see for instance in the derivation of Zaldarriaga et al. (2004). The difference is one of generality, whether in the inputs, the intermediate steps, and the outputs. On the input side, P_{N} assumes uncorrelated noise between visibilities whose amplitude is governed by the radiometer equation; $\sigma_{\text{QE-N}}$ can accept an arbitrary input covariance (even though in our tests we take it to be diagonal). During the actual propagation of errors, the derivation of P_{N} assumes that fluctuations in $uv\nu$ space are uncorrelated; $\sigma_{\text{QE-N}}$ makes no such approximations. Finally, on the output side, the "power spectrum method" returns a single error bar; the 'covariance method' provides a full bandpower covariance matrix.

Of course, in reality not all delay modes are noise-dominated, and reliable error bars need to be placed even in signal-dominated regimes (whether this signal comes in the form of instrument systematics, foregrounds, or—ultimately—the cosmological signal). It is difficult to place rigorous error bars on bandpowers in these regimes: unless one has a physical model for all the systematics involved (with knowledge of their probability distributions), it is an ill-defined problem to ask how errors propagate. Unfortunately, the presence of unexplained (or at least not fully explained) systematics is the current state of affairs in 21 cm cosmology, and truly rigorous error bars will need to wait for future work on the modeling of systematics.

Even with well-defined (if not perfectly characterized) systematics, the meaning of one's error bars is subtle. For instance, foregrounds such as a continuum of unresolved point sources can be appropriately treated as a random field. Given this, one's approach might be to say that the unresolved point sources contribute some effective power spectrum to the measurement. With such a formalism, there is a fundamental limit to how well these foregrounds can be characterized, since they come with their own form of cosmic variance. In other words, if one is trying to place constraints on foregrounds, one must account for the fact that the particular realization of foregrounds that we see may not be representative of foregrounds in general. This sort of error is difficult to compute in general, as the squared nature of the power spectrum means that the non-Gaussian—and therefore non-trivial—four-point function of the foregrounds needs to be known.

A goal of characterizing the general statistical properties of all possible foregrounds, however, may be unnecessarily ambitious. In particular, for a cosmological measurement one is not particularly concerned with the behaviour of a "typical" foreground; one is primarily concerned with how our particular realization of foregrounds affect our observations. As a concrete example, if our Galaxy's synchrotron emission happens to be anomalously bright compared to a typical galaxy's synchrotron emission, it is our own brighter foregrounds that we need to deal with! With such a mindset, it is more appropriate to consider all foregrounds as non-random components of our data. By this, we do not mean that the foregrounds need to be spatially or spectrally constant; rather, we mean that in hypothetical random draws for taking ensemble averages, the cosmological signal and the instrumental noise change with each new realization, but the foregrounds remain the same. If the foregrounds are not formally random, our error bars are the result of instrumental noise (and in principle cosmic variance of the cosmological signal, although this contribution is small for current upper limits).

It is important to stress, however, that even if our error bars are due to the randomness of instrumental noise, the resulting error bars are *not* simply what one obtains from imagining a noise-only measurement and propagating the noise fluctuations through to a power spectrum. This is because the power spectrum is a squared statistic. Thus, in the squaring of a measurement that contains both noise and a (non-random) signal, there are signal-noise cross-terms to contend with. These terms are zero in expectation, but do not have non-zero *variance*. This means that knowledge of the signal (whether from systematics or foregrounds) is needed to correctly account for instrumental noise errors in non-noise-dominated regimes.

- In short, even if we lower our ambitions and forgo incorporating knowledge about signal *statistics* into our error calculations, understanding the signal itself is necessary for computing noise-sourced error bars. This requirement is where noise-only computations like $P_{\rm N}$ and $\sigma_{\rm QE\text{-}N}$ fall short.

- This shortcoming is remedied by generalized versions of $P_{\rm N}$ and $\sigma_{\rm QE\text{-}N}$, which we dub $P_{\rm SN}$ and $\sigma_{\rm QE\text{-}SN}$. These are given by Equations (30) and (38). The key idea is that in signal dominated regimes, the measured data itself can be a good approximation to the signal. Thus, we may reinsert the data in an appropriate way to capture signal terms in our general expressions. Figures 3 and 4 show that these error bars work well in both signal-dominated and noise-dominated regimes.

- Although we treat foregrounds and systematics as a single signal term that is directly estimated from measured data in this paper, we note that for future high-sensitivity detections, more elaborate modeling of both are needed. Of course, there is also the possibility of unknown systematic effects, which our formalism does not account for.

- However Moreover, two cautionary warnings are in order when applying Equations (30) and (38). The first is that because the measured data are now part of the error bars themselves, it can be dangerous to use these error bars to inform data weightings for downstream averages in one's pipeline (e.g., in further incoherent time averaging of power spectra or in incoherent averaging of power spectra from different baselines). If the data weightings are coupled to the data themselves, our so-called quadratic estimators are no longer quadratic. As shown in Cheng et al. (2018), a blind application of the usual methods for normalizing quadratic estimators leads to power spectrum estimates that are biased low ("signal loss"). For this reason, while $P_{\rm SN}$ and $\sigma_{\rm QE\text{-}SN}$ are fine ways to compute error bars, we recommend that any error-motivated data weightings be based on $P_{\rm N}$ and $\sigma_{\rm QE\text{-}N}$ instead.

- The second warning is that there almost certainly exist regimes that are neither signal- nor noise-dominated, where signal and noise are comparable in magnitude. Here, it becomes necessary to contend with the fact that a noisy measurement of the signal can be unphysically negative. Said differently, if our estimate of the signal itself contains noise, we are in effect double counting the noise in our error computations. One approach is to enact a hard prior on the positivity of the signal. This is what was done in all computations of $P_{\rm SN}$ and $\sigma_{\rm QE\text{-}SN}$ in this paper. However, Figures 3 and 4 show that this has the effect of inflating the error bars. Given that this is a conservative bias on the errors, this may or may not be appropriate depending on one's application.

- A slightly more accurate approach is to assume that instrumental noise is Gaussian distributed and to quantitatively predict and correct the noise bias in the errors. Implementing this correction gives $\tilde{P}_{\rm SN}$ and $\tilde{\sigma}_{\rm QE\text{-}SN}$, which are given by Equations (32) and (40) respectively. Figures 3 and 4 show that this corrects the bias and gives error bars that are no longer overly conservative. However, because this correction is designed to give

Error Bar Type	Pros	Cons
Bootstrap ($\sigma_{\rm bs}$)	Easy to implement with minimal *a priori* assumptions; can be useful as a reference statistics in diagnosis of systematics	Not strictly applicable in the presence of non-independent and non-statistically stationary data samples
Power spectra from differenced visibilities ($P_{\rm diff}$)	Data product close to raw data	Provides noise *realizations* rather than direct error bars, resulting in considerable scatter
Power spectrum method ($P_{\rm N}$ and $P_{\rm SN}$)	Accurately captures variances/error bars in noise-dominated regimes (both $P_{\rm N}$ and $P_{\rm SN}$) and signal-dominated regimes ($P_{\rm SN}$)	Does not contain covariance information between different bandpowers; $P_{\rm SN}$ requires non-negativity prior on the signal, which slightly inflates errors; downstream data weightings using $P_{\rm SN}$ at risk of signal loss
Covariance method ($\sigma_{\rm QE\text{-}N}$ and $\sigma_{\rm QE\text{-}SN}$)	Same accuracy as $P_{\rm N}$ and $P_{\rm SN}$ for variance information and additionally provides full covariance information	Derivation assumes data is Gaussian; $\sigma_{\rm QE\text{-}SN}$ requires non-negativity prior on the signal, which slightly inflates errors; downstream data weightings using $\sigma_{\rm QE\text{-}SN}$ at risk of signal loss
Modified covariance method ($\tilde{\sigma}_{\rm QE\text{-}SN}$) and modified power spectrum method $\tilde{P}_{\rm SN}$	Eliminates conservative double counting of noise in noisy estimates of the signal	Occasional error predictions that are slightly smaller than instrumental noise expectations from $\sigma_{\rm QE\text{-}N}$ and $P_{\rm N}$

Table 4. A summary of the advantages and disadvantages of different error estimation methods in 21 cm power spectrum estimation.

unbiased errors *in expectation*, it will occasionally give error bars that are slightly smaller than the error predicted by noise-only estimators such as $P_{\rm N}$. In practice, however, we find that this is a reasonably rare occurence.

With the aforementioned difficulties with error estimation in the presence of poorly characterized signals, one may be tempted to make use of more empirically based error estimates. These estimates also come with their strengths and weaknesses:

- As discussed in Section 3.2, $P_{\rm diff}$ from frequency-differenced data suffers from a bias at low delays. Figure 1 shows that even at reasonably high delays ~ 1500 ns, the bias can be significant. Thus, while $P_{\rm diff}$ from frequency-differenced data is a useful asymptotic check at high delays, it is not a robust estimator of our errors. Implementing $P_{\rm diff}$ using time-differenced data does not have the delay-dependent bias, as one can also see in Figure 1. However, care must be taken to ensure that the time differencing is small enough to suppress any sky signal that is coherent between adjacent time samples (Dillon et al. 2015). In addition, with a differencing scheme one is ultimately constructing noise *realizations*, not noise statistics. The resulting error bars thus show considerable scatter. In that sense, the analytically propagated error bars vary in a more physically plausible—smoother—way with time and frequency.

- The problem of a noisy error bar estimate persists with $\sigma_{\rm bs}$. However, bootstrapping has several appealing features that makes it a crucial check on the analytically propagated error bars. First, no assumptions are made regarding Gaussianity of the input data. Thus, the fact that our $\sigma_{\rm bs}$ agree with our analytically propagated errors—which assumed the input noise in the visibilities—is an essential validation of our assumptions. In a similar way, $\sigma_{\rm bs}$ may potentially capture increased variance due to systematics since it is a measure of uncertainties of total sky emission. However, the bootstrap method is known to suffer from some important limitations. For example, as noted in Appendix B, if systematics are correlated between samples, the bootstrap method tends to underestimate errors. Also, bootstrapped error bars will be inflated from non-stationary effects such as sky brightness changes and non-redundancies between nominally identical baselines. Precisely how these non-stationary effects should be folded into one's error estimation is reserved for future work, but the correct approach will certainly be more sophisticated than a simple inflation of errors. That said, this increase in bootstrap errors due to nonstationarity can serve as a useful diagnostic for further examination of unexpected systematics.

In Table 4 we summarize the discussion in this section with an succinct listing of the pros and cons of each error estimation method.

6. CONCLUSIONS

In this paper, we have systematically studied a variety of error bar methodologies in 21 cm power spectrum estimation. We have synthesized some of the common techniques in the literature, outlining their relative strengths and weaknesses in quantifying noise levels and in accounting for residual systematics. Specifically, we considered a variety of types of error estimators, including

- Power spectrum methods. This includes the standard $P_{\rm N}$ estimator for the noise power spectrum found in the literature (Zaldarriaga et al. 2004; Parsons et al. 2012a; Pober et al. 2013; Cheng et al. 2018; Kern et al. 2020b) and the $P_{\rm SN}$ estimator that involves cross products with signal power spectrum $P_{\rm S}$, as detailed in Kolopanis et al. (2019). Here we set $P_{\rm S}$ to be the real values of experimentally observed power spectrum, which is a good approximation when the signal dominates the noise. Our implementation of $P_{\rm SN}$ leads to a double-counting bias compared to $P_{\rm N}$ which is considerable in noise-dominated regimes, and we show how a modified form $\tilde{P}_{\rm SN}$ can eliminate this bias.

- Covariance methods. This consists of propagating a data covariance matrix between frequencies per timestamp and per baseline-pair through the quadratic estimator (QE) formalism to the bandpower covariance matrix (Liu & Tegmark 2011; Dillon et al. 2014; Liu et al. 2014a,b), including error metrics described here: $\sigma_{\rm QE\text{-}N}$ for noise-dominated spectra and $\sigma_{\rm QE\text{-}SN}$ that include signal-noise terms. These have identical *variance* predictions as $P_{\rm N}$ and $P_{\rm SN}$ by construction but also provide bandpower covariance information.

- Other methods. Other methods studied in this work includes the bootstrapping method that can lead to misreported errors when not handled carefully (Cheng et al. 2018), as well as the method of using differenced visibilities as noise realizations propagated through a power spectrum estimator. We show that differencing in frequency is ill-advised for this approach. Differencing in time avoids some problems, but either differencing scheme generates error estimates that are rather scattered. However, we stress the importance of these more empirically based methods are useful cross-checks (e.g., in the manner performed in this paper) that can also be helpful diagnostics for systematics (e.g., Kolopanis et al. 2019).

Using simulations and real HERA Phase I data, we show that these methods are generally in agreement with each other, demonstrating their robustness and their applicability to future delay power spectrum measurements from HERA. Importantly, we show that for bandpowers that are not completely dominated by noise, one needs to go beyond the standard thermal noise estimates and account for signal-noise cross terms in order to fully describe the uncertainty on the band power. In a series of Appendices, we also examine sources of skewness in probability distributions of measured power spectrum bandpowers (Appendices A and F), derive exact expressions for the probability distributions of incoherently summed delay power spectra (Appendix G), and examine whether common baselines in the cross multiplication of multiple baseline *pairs* affects assumptions about error independence (Appendix B). The insights gained in this paper regarding error estimation are applicable in 21 cm cosmology beyond HERA. They provide a foundation upon which to develop rigorous error estimation methods which will prove to be key in unlocking the potential of the 21 cm line as a powerful probe of our high redshift universe.

This material is based upon work supported by the National Science Foundation under Grant Nos. 1636646 and 1836019 and institutional support from the HERA collaboration partners. This research is funded in part by the Gordon and Betty Moore Foundation. HERA is hosted by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. Parts of this research were supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. G. Bernardi acknowledges funding from the INAF PRIN-SKA 2017 project 1.05.01.88.04 (FORECaST), support from the Ministero degli Affari Esteri della Cooperazione Internazionale - Direzione Generale per la Promozione del Sistema Paese Progetto di Grande Rilevanza ZA18GR02 and the National Research Foundation of South Africa (Grant Number 113121) as part of the ISARP RADIOSKY2020 Joint Research Scheme, from the Royal Society and the Newton Fund under grant NA150184 and from the National Research Foundation of South Africa (grant No. 103424). P. Bull acknowledges funding for part of this research from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 948764), and from STFC Grant ST/T000341/1. J.S. Dillon gratefully acknowledges the support of the NSF AAPF award #1701536. N. Kern acknowledges support from the MIT Pappalardo fellowship. A. Liu acknowledges support from the New Frontiers in Research Fund Exploration grant program, the Canadian Institute for Advanced Research (CIFAR) Azrieli Global Scholars program, a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and a Discovery Launch Supplement, the Sloan Research Fellowship, and the William Dawson Scholarship at McGill.

APPENDIX

A. SKEWNESS IN POWER SPECTRA ESTIMATED FROM MULTIPLE IDENTICAL BASELINES

In this Appendix, we consider a source of skewness in probability distributions of delay spectra. In particular, we consider the noise properties of power spectra formed from a set of identical ("redundant") baselines. We show that even if each baseline is measuring Gaussian random noise with mean zero, the resulting power spectra will exhibit some skewness. We emphasize, however, that this skewness vanishes if one additionally splits the data into two distinct set of time samples (e.g., even and odd time stamps) and estimates power spectra that are not only cross-baselines but also cross-times.

As a concrete example, suppose that on the ith copy of a particular baseline we measure $\tilde{x}_i \equiv c_i + id_i$ after taking the delay transform, where c_i and d_i are independently Gaussian distributed random variables with variance $\sigma^2/2$. This represents the behavior of $\tilde{x}_i$ at noise-dominated delays. If only two identical baselines were available, cross multiplying them to obtain a power spectrum would yield

$$\tilde{x}_1\tilde{x}_2^* = (c_1 + id_1)(c_2 - id_2) = (c_1c_2 + d_1d_2) + i(d_1c_2 - c_1d_2). \tag{A1}$$

Consider the real part. Since c_1 and c_2 are independent random variables, c_1c_2 is a symmetric distribution about zero (and in fact is given by K_0, the zeroth modified Bessel function of the second kind). The same reasoning holds for the d_1d_2 term. Since $\{c_i\}$ and $\{d_i\}$ are independent, it follows that c_1c_2 and d_1d_2 are also independent. The result is that the probability distribution for $c_1c_2 + d_1d_2$ is given by the convolution of the distributions for the individual terms. With the two contributing distributions both symmetric about zero, their convolution inherits this property, and is in fact given by the Laplacian distribution discussed in Section 4.1.

The situation is different when we have more than two baselines. Taking all possible pairwise combinations (excluding the multiplication of a baseline with itself to eliminate noise bias), we obtain

$$\mathrm{Re}[\tilde{x}_1\tilde{x}_2^* + \tilde{x}_1\tilde{x}_3^* + \tilde{x}_2\tilde{x}_3^*] = (c_1c_2 + c_1c_3 + c_2c_3) + (d_1d_2 + d_1d_3 + d_2d_3), \tag{A2}$$

where we have grouped our result into two terms that can be considered separately because $\{c_i\}$ and $\{d_i\}$ are independent. Consider the first term. It has zero mean:

$$\langle c_1c_2 + c_1c_3 + c_2c_3\rangle = \langle c_1\rangle\langle c_2\rangle + \langle c_1\rangle\langle c_3\rangle + \langle c_2\rangle\langle c_3\rangle = 0 \tag{A3}$$

because the different $\{c_i\}$ are independent. However, the resulting distribution has a skewness to it, which can be seen by the fact that the third moment is non-zero:

$$\begin{aligned}\langle(c_1c_2 + c_1c_3 + c_2c_3)^3\rangle &= \langle c_2^3c_1^3 + c_3^3c_1^3 + 3c_2c_3^2c_1^3 + 3c_2^2c_3c_1^3 + 3c_2c_3^3c_1^2 + 6c_2^2c_3^2c_1^2 + 3c_2^3c_3c_1^2 + 3c_2^2c_3^3c_1 + 3c_2^3c_3^2c_1 + c_2^3c_3^3\rangle \\ &= 6\langle c_2^2c_3^2c_1^2\rangle = 6\langle c_2^2\rangle\langle c_3^2\rangle\langle c_1^2\rangle \neq 0\end{aligned} \tag{A4}$$

[Of course, in principle we should be taking the cube of Equation (A2) in its entirety, not just the first term. However, the independence of $\{c_i\}$ and $\{d_i\}$ means we reach the same conclusion.] The non-zero third moment shown here arises because the three terms that make up the sum are correlated as a triplet, even though each pair has no average covariance. For instance, the covariance between c_1c_2 and c_1c_3 is

$$\langle c_1c_2c_1c_3\rangle - \langle c_1c_2\rangle\langle c_1c_3\rangle = \langle c_1^2\rangle\langle c_2\rangle\langle c_3\rangle = 0. \tag{A5}$$

This implies that even though c_1c_2, c_1c_3, and c_2c_3 are not independent, for the purposes of computing the variance of the final result, one obtains the same result even if one pretends that these contributions are independent. This result is explored in more detail in the first half of Appendix B

To summarize, the different moments of the distribution provide different insights into power spectrum estimation with different baseline pair combinations. The mean of the distribution is zero, indicating that there is no bias (as one might expect for cross-correlation spectra). The variance turns out to be the same expression as if we had completely independent baseline pairs, so the noise averages down with the number of baseline pairs as one might naively have expected them to (without worrying about correlations). However, the skewness is non-zero. This complicates the interpretation of null tests that implicitly assume that the probability distributions of noise-dominated delays are symmetric.

Importantly, these considerations do not apply when we consider the imaginary part, which is given by

$$\mathrm{Im}[\tilde{x}_1\tilde{x}_2^* + \tilde{x}_1\tilde{x}_3^* + \tilde{x}_2\tilde{x}_3^*] = c_2d_1 + c_3d_1 - c_1d_2 + c_3d_2 - c_1d_3 - c_2d_3. \tag{A6}$$

This has a third moment given by $\langle(c_2d_1 + c_3d_1 - c_1d_2 + c_3d_2 - c_1d_3 - c_2d_3)^3\rangle$. To get terms that are non-zero under the expectation value, we require terms that contain *squares* of the random variables when we multiply out the polynomial. For example, the first term c_2d_1 must be multiplied onto c_2d_3, because there is no other c_2 term in the expression to pair to. This gives us $c_2^2d_1d_3$. However, we now need to multiply this onto d_1d_3, or we end up with a stray d_1 and a stray d_3. But none of the terms are the product of two $\{d_i\}$, so no matter what terms we pair this up with, it will average to zero. This logic applies to any of the terms, so the distribution of the imaginary part will not be skewed. Because of this, statistical tests involving the imaginary part of a power spectrum estimator can be more easily interpreted using symmetric distributions.

Our result here has implications for how one should avoid the noise bias in power spectrum measurements. Two commonly used methods for doing so are to cross-multiply either different identical baselines together or different time stamps together. Here we have shown that employing only one of these will incur a skewness. (While our discussion above focused on cross multiplying different baselines, the same conclusions hold if we consider cross multiplying more than two groups in time—after all, the indices in our mathematical expressions can simply be considered timestamp indices instead of baseline indices.) However, if we perform cross-multiplications across both time and baseline axes, the skewness vanishes. To see this, imagine that we split our data into odd and even time samples, labeled with superscripts "o" and "e" respectively. Equation (A2) then becomes

$$\mathrm{Re}[\tilde{x}_1^e\tilde{x}_2^{o*}+\tilde{x}_1^e\tilde{x}_3^{o*}+\tilde{x}_2^e\tilde{x}_3^{o*}]=(c_1^ec_2^o+c_1^ec_3^o+c_2^ec_3^o)+(d_1^ed_2^o+d_1^ed_3^o+d_2^ed_3^o), \tag{A7}$$

and cubing this expression as before to compute the third moment, one finds no non-zero terms after taking the ensemble average.

B. VARIANCE OF AVERAGED POWER SPECTRA FROM DEPENDENT BASELINE-PAIR SAMPLES

In this Appendix, we consider the effect of having common baselines between different baseline *pairs* used to form power spectra. Inside a redundant baseline group consisting of N_{bl} different baselines, then we can construct up to $N_{\mathrm{blp}}=\frac{1}{2}N_{\mathrm{bl}}(N_{\mathrm{bl}}-1)$ different baseline pairs and we can form a power spectrum using each pair. Consider the averaged power spectrum over these baseline pairs and the variance of this average. The form of the averaged power spectrum is

$$\overline{P}=\frac{\sum_{(p,q>p)}P_{pq}}{\frac{1}{2}N_{\mathrm{bl}}(N_{\mathrm{bl}}-1)}\,, \tag{B8}$$

where the sum is over all possible (p,q) pairs of baselines. The variance of the averaged power spectrum does not simply go down with N_{blp}^{-1} because the data being averaged together are not fully independent of each other. For example, P_{12} and P_{13} both carry information from baseline #1.

Let the signal be $\tilde{s}\equiv a+bi$, and $\tilde{n}_p\equiv c_p+d_pi$ and $\tilde{n}_q\equiv c_q+d_qi$ be the noise realizations in the pth and qth baselines. The signal $\tilde{s}$ is identical in each baseline, since we are assuming that we are combining data from identical ("redundant") baselines. The random variables c_p,d_p,c_q,d_q... are IID normal variables with variance σ^2. In the foreground-negligible regime, recall from Equation (17) that the average power spectrum is given by

$$\overline{P}=\frac{\sum_{(p,q>p)}n_p^*n_q}{\frac{1}{2}N_{\mathrm{bl}}(N_{\mathrm{bl}}-1)}=\frac{\sum_{(p,q>p)}c_pc_q+d_pd_q}{\frac{1}{2}N_{\mathrm{bl}}(N_{\mathrm{bl}}-1)}+i\frac{\sum_{(p,q>p)}c_pd_q-c_qd_p}{\frac{1}{2}N_{\mathrm{bl}}(N_{\mathrm{bl}}-1)}\,. \tag{B9}$$

We notice

$$\begin{aligned}\mathrm{Var}\left(\sum_{(p,q>p)}c_pc_q\right)=&\left\langle\sum_{(p,q>p)}c_pc_q\sum_{(r,t>r)}c_rc_t\right\rangle-\left[\left\langle\sum_{(p,q>p)}c_pc_q\right\rangle\right]^2=\left\langle\sum_{(p,q>p)}c_pc_q\sum_{(r,t>r)}c_rc_t\right\rangle\\=&\sigma^4\left[\sum_{(p,q>p,r,t>r)}(\delta_{pr}\delta_{qt}+\delta_{pt}\delta_{qr})\right]=\frac{N_{\mathrm{bl}}(N_{\mathrm{bl}}-1)}{2}\sigma^4\,,\end{aligned} \tag{B10}$$

which means that the variance in the real part of $\overline{P}$ is $\frac{4\sigma^4}{N_{\mathrm{bl}}(N_{\mathrm{bl}}-1)}$. For the imaginary part we compute

$$\begin{aligned}\mathrm{Var}\left(\sum_{(p,q>p)}c_pd_q-c_qd_p\right)=&\left\langle\sum_{(p,q>p)}\{c_pd_q-c_qd_p\}\sum_{(r,t>r)}\{c_rd_t-c_td_r\}\right\rangle-\left[\left\langle\sum_{(p,q>p)}\{c_pd_q-c_qd_p\}\right\rangle\right]^2\\=&\left\langle\sum_{(p,q>p)}\{c_pd_q-c_qd_p\}\sum_{(r,t>r)}\{c_rd_t-c_td_r\}\right\rangle=\sigma^4\left[\sum_{(p,q>p,r,t>r)}(2\delta_{pr}\delta_{qt}-2\delta_{pt}\delta_{qr})\right]\\=&N_{\mathrm{bl}}(N_{\mathrm{bl}}-1)\sigma^4\,,\end{aligned} \tag{B11}$$

so that the variance of the imaginary part of $\overline{P}$ is also $\frac{4\sigma^4}{N_{\rm bl}(N_{\rm bl}-1)}$. Since the number of baseline pairs is given by $N_{\rm bl}(N_{\rm bl}-1)/2$ and $2\sigma^4$ is the variance we would expect to get from a single baseline pair, we can see that $\overline{P}$ averages down in a manner that is identical to the scenario where the baseline pairs are independent.

In foreground-dominant regimes, the average power spectrum goes to

$$\overline{P} = \frac{\sum_{(p,q>p)} s^* s + s^* n_q + n_p^* s}{\frac{1}{2}N_{\rm bl}(N_{\rm bl}-1)} = \frac{\sum_{(p,q>p)} a^2 + b^2 + a(c_p + c_q) + b(d_p + d_q)}{\frac{1}{2}N_{\rm bl}(N_{\rm bl}-1)} + i\frac{\sum_{(p,q>p)} a(d_q - d_p) + b(c_p - c_q)}{\frac{1}{2}N_{\rm bl}(N_{\rm bl}-1)} \,. \tag{B12}$$

The variance in the real part is $\frac{4(a^2+b^2)\sigma^2}{N_{\rm bl}}$ and the variance in the imaginary part is $\frac{4(N_{\rm bl}+1)(a^2+b^2)\sigma^2}{3N_{\rm bl}(N_{\rm bl}-1)}$. They now go down roughly as $N_{\rm blp}^{-1/2}$ and are larger than the variance from independent samples by factors of $(N_{\rm bl}-1)$ and $(N_{\rm bl}+1)/3$ respectively.

C. TIME-DIFFERENCED VISIBILITIES AS NOISE ESTIMATORS

In this Appendix, we establish the validity of using time-differenced visibilities as a way to estimate noise error bars. The key idea is that if we form residuals of data vectors $x_p(\nu, t)$ by subtracting data from the pth baseline in adjacent time bins (t_1 and t_2) from each other, the result should be noise dominated. The same holds true for delay-transformed visibilities, where the residual can be written as $\tilde{n}_p(\tau, t_2) - \tilde{n}_p(\tau, t_1)$. Suppressing τ and demoting the time variable to a subscript for notational brevity, we write $\tilde{n}_{p,t} = c_{p,t} + d_{p,t} i$, where $c_p, d_p ...$ are IID normal variables with variance σ^2. The power spectra constructed from such residuals are

$$\begin{aligned} P_{\rm diff} &= \frac{(\tilde{n}_{1,t2} - \tilde{n}_{1,t_1})^*}{\sqrt{2}} \frac{(\tilde{n}_{2,t2} - \tilde{n}_{2,t_1})}{\sqrt{2}} \\ &= \left[\frac{(c_{1,t2} - c_{1,t1})}{\sqrt{2}} \frac{(c_{2,t2} - c_{2,t1})}{\sqrt{2}} + \frac{(d_{1,t2} - d_{1,t1})}{\sqrt{2}} \frac{(d_{2,t2} - d_{2,t1})}{\sqrt{2}}\right] \\ &\quad + \left[\frac{(c_{1,t2} - c_{1,t1})}{\sqrt{2}} \frac{(d_{2,t2} - d_{2,t1})}{\sqrt{2}} - \frac{(c_{2,t2} - c_{2,t1})}{\sqrt{2}} \frac{(d_{1,t2} - d_{1,t1})}{\sqrt{2}}\right] i \,. \end{aligned} \tag{C13}$$

From this, we see that

$$\left\langle [{\rm Re}(P_{\rm diff})]^2 \right\rangle = \left\langle \left[\frac{(c_{1,t2} - c_{1,t1})}{\sqrt{2}} \frac{(c_{2,t2} - c_{2,t1})}{\sqrt{2}} + \frac{(d_{1,t2} - d_{1,t1})}{\sqrt{2}} \frac{(d_{2,t2} - d_{2,t1})}{\sqrt{2}}\right]^2 \right\rangle = \langle c_1^2 \rangle \langle c_2^2 \rangle + \langle d_1^2 \rangle \langle d_2^2 \rangle = 2\sigma^4. \tag{C14}$$

This is again the variance expected for a noise-dominated power spectrum. Therefore, what we have shown is that $|{\rm Re}(P_{\rm diff})|$ can serve as an estimator that *in expectation* is equal to the correct noise errors for the measured power spectrum $P_{\tilde{x}_1\tilde{x}_2}$ in noise-dominated regimes. However, since this result only holds in expectation, we expect that in practice it will exhibit considerable scatter as an error estimate.

D. SIGNAL DEPENDENT ERROR BAR FROM POWER SPECTRUM METHOD

In this Appendix we derive an expression for the variance on the power spectrum in the presence of foregrounds or systematics (or any "signal"). A similar derivation is presented in Kolopanis et al. (2019). Given two delay spectra $\tilde{x}_1 = \tilde{s} + \tilde{n}_1$ and $\tilde{x}_2 = \tilde{s} + \tilde{n}_2$, the power spectra formed from $\tilde{x}_1^* \tilde{x}_2$ is

$$\begin{aligned} P_{\tilde{x}_1\tilde{x}_2} &= \tilde{s}^*\tilde{s} + \tilde{s}^*\tilde{n}_2 + \tilde{n}_1^*\tilde{s} + \tilde{n}_1^*\tilde{n}_2 \\ &= \left[a^2 + b^2 + a(c_1 + c_2) + b(d_1 + d_2) + c_1 c_2 + d_1 d_2\right] + \left[a(d_2 - d_1) + b(c_1 - c_2) + d_2 c_1 - d_1 c_2\right] i \,, \end{aligned} \tag{D15}$$

where we have written $\tilde{s} = a + bi$, $\tilde{n}_1 = c_1 + d_1 i$ and $\tilde{n}_2 = c_2 + d_2 i$.

Consistent with the rest of the paper, we assume that a and b are not random variables, so that $\langle s \rangle = s$. The true sky power spectrum is then given by $P_{\tilde{s}\tilde{s}} = a^2 + b^2$, and c_1, d_1, c_2 and d_2 in noise parts are IID random normal variables. We then have

$$\begin{aligned} {\rm Var}\left[{\rm Re}(P_{\tilde{x}_1\tilde{x}_2})\right] &= {\rm Var}\left[a^2 + b^2 + a(c_1 + c_2) + b(d_1 + d_2) + c_1 c_2 + d_1 d_2\right] \\ &= 2(a^2 + b^2)\langle c_1^2 \rangle + 2\langle c_1^2 \rangle^2 = \sqrt{2} P_{\tilde{s}\tilde{s}} P_{\rm N} + P_{\rm N}^2 = \sqrt{2}\langle {\rm Re}(P_{\tilde{x}_1\tilde{x}_2})\rangle P_{\rm N} + P_{\rm N}^2 = P_{\rm SN}^2 \,. \end{aligned} \tag{D16}$$

In the above we have used the relation ${\rm var}(c_1 c_2 + d_1 d_2) = 2\langle c_1^2 \rangle^2 = P_{\rm N}^2$, where $P_{\rm N}$ is the analytic noise power spectrum. We have also used $P_{\tilde{s}\tilde{s}} = \langle {\rm Re}(P_{\tilde{x}_1\tilde{x}_2})\rangle$. This shows that $P_{\rm SN}$ is a general form for error bars in the existence of foregrounds or systematics (or again, any "signal").

E. COVARIANCE METHOD

In this Appendix we provide more explicit derivations of the expressions quoted in Section 3.4 for the covariance method of error estimation.

E.1. *Variance*

If $\hat{P}_\alpha$ is a complex number representing a power spectrum estimate of the αth bandpower, its real part and imaginary part are given by $\frac{1}{2}(\hat{P}_\alpha + \hat{P}^*_\alpha)$ and $\hat{P}_\alpha$ is $\frac{1}{2i}(\hat{P}_\alpha - \hat{P}^*_\alpha)$ respectively. The variance in the real part of $\hat{P}_\alpha$ is

$$\frac{1}{4}\left\{(\langle \hat{P}_\alpha \hat{P}_\alpha\rangle - \langle \hat{P}_\alpha\rangle\langle \hat{P}_\alpha\rangle) + 2(\langle \hat{P}_\alpha \hat{P}^*_\alpha\rangle - \langle \hat{P}_\alpha\rangle\langle \hat{P}^*_\alpha\rangle) + (\langle \hat{P}^*_\alpha \hat{P}^*_\alpha\rangle - \langle \hat{P}^*_\alpha\rangle\langle \hat{P}^*_\alpha\rangle)\right\}, \tag{E17}$$

while the variance in the imaginary part of $\hat{P}_\alpha$ is

$$-\frac{1}{4}\left\{(\langle \hat{P}_\alpha \hat{P}_\alpha\rangle - \langle \hat{P}_\alpha\rangle\langle \hat{P}_\alpha\rangle) - 2(\langle \hat{P}_\alpha \hat{P}^*_\alpha\rangle - \langle \hat{P}_\alpha\rangle\langle \hat{P}^*_\alpha\rangle) + (\langle \hat{P}^*_\alpha \hat{P}^*_\alpha\rangle - \langle \hat{P}^*_\alpha\rangle\langle \hat{P}^*_\alpha\rangle)\right\}. \tag{E18}$$

Recall that $\hat{P}_\alpha$ is defined as $\hat{P}_\alpha = \boldsymbol{x}_1^\dagger \boldsymbol{E}^{12,\alpha}\boldsymbol{x}_2 = \sum_{ij}\boldsymbol{x}^*_{1,i}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{x}_{2,j}$. We define three set of matrices containing the whole two-point correlation information for the complex estimator $\boldsymbol{C}^{12}$, $\boldsymbol{U}^{12}$ and $\boldsymbol{G}^{12}$, such that

$$\boldsymbol{C}^{12}_{ij} \equiv \langle \boldsymbol{x}_{1,i}\boldsymbol{x}^*_{2,j}\rangle; \qquad \boldsymbol{U}^{12}_{ij} \equiv \langle \boldsymbol{x}_{1,i}\boldsymbol{x}_{2,j}\rangle; \qquad \boldsymbol{G}^{12}_{ij} \equiv \langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}^*_{2,j}\rangle, \tag{E19}$$

Equipped with these definitions, we can generate the following equations

$$\begin{aligned}
\langle \hat{P}_\alpha \hat{P}_\beta\rangle - \langle \hat{P}_\alpha\rangle\langle \hat{P}_\beta\rangle =& \sum_{ijkl}\langle \boldsymbol{x}^*_{1,i}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{x}_{2,j}\boldsymbol{x}^*_{1,k}\boldsymbol{E}^{12,\beta}_{kl}\boldsymbol{x}_{2,l}\rangle - \langle \boldsymbol{x}^*_{1,i}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{x}_{2,j}\rangle\langle \boldsymbol{x}^*_{1,k}\boldsymbol{E}^{12,\beta}_{kl}\boldsymbol{x}_{2,l}\rangle \\
=& \sum_{ijkl}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{E}^{12,\beta}_{kl}(\langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}_{2,j}\boldsymbol{x}^*_{1,k}\boldsymbol{x}_{2,l}\rangle - \langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}_{2,j}\rangle\langle \boldsymbol{x}^*_{1,k}\boldsymbol{x}_{2,l}\rangle) \\
=& \sum_{ijkl}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{E}^{12,\beta}_{kl}(\langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}^*_{1,k}\rangle\langle \boldsymbol{x}_{2,j}\boldsymbol{x}_{2,l}\rangle + \langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}_{2,l}\rangle\langle \boldsymbol{x}^*_{1,k}\boldsymbol{x}_{2,j}\rangle) \\
=& \sum_{ijkl}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{E}^{12,\beta}_{kl}(\boldsymbol{G}^{11}_{ik}\boldsymbol{U}^{22}_{jl} + \boldsymbol{C}^{21}_{li}\boldsymbol{C}^{21}_{jk}) \\
=& \sum_{ijkl}(\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{U}^{22}_{jl}\boldsymbol{E}^{21,\beta*}_{lk}\boldsymbol{G}^{11}_{ki} + \boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{C}^{21}_{jk}\boldsymbol{E}^{12,\beta}_{kl}\boldsymbol{C}^{21}_{li}) \\
=& \mathrm{tr}(\boldsymbol{E}^{12,\alpha}\boldsymbol{U}^{22}\boldsymbol{E}^{21,\beta*}\boldsymbol{G}^{11} + \boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{21}\boldsymbol{E}^{12,\beta}\boldsymbol{C}^{21}),
\end{aligned} \tag{E20}$$

$$\begin{aligned}
\langle \hat{P}_\alpha \hat{P}^*_\beta\rangle - \langle \hat{P}_\alpha\rangle\langle \hat{P}^*_\beta\rangle =& \sum_{ijkl}\langle \boldsymbol{x}^*_{1,i}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{x}_{2,j}\boldsymbol{x}_{1,k}\boldsymbol{E}^{12,\beta*}_{kl}\boldsymbol{x}^*_{2,l}\rangle - \langle \boldsymbol{x}^*_{1,i}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{x}_{2,j}\rangle\langle \boldsymbol{x}_{1,k}\boldsymbol{E}^{12,\beta*}_{kl}\boldsymbol{x}^*_{2,l}\rangle \\
=& \sum_{ijkl}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{E}^{12,\beta*}_{kl}(\langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}_{2,j}\boldsymbol{x}_{1,k}\boldsymbol{x}^*_{2,l}\rangle - \langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}_{2,j}\rangle\langle \boldsymbol{x}_{1,k}\boldsymbol{x}^*_{2,l}\rangle) \\
=& \sum_{ijkl}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{E}^{12,\beta*}_{kl}(\langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}^*_{2,l}\rangle\langle \boldsymbol{x}_{1,k}\boldsymbol{x}_{2,j}\rangle + \langle \boldsymbol{x}^*_{1,i}\boldsymbol{x}_{1,k}\rangle\langle \boldsymbol{x}_{2,j}\boldsymbol{x}^*_{2,l}\rangle) \\
=& \sum_{ijkl}\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{E}^{12,\beta*}_{kl}(\boldsymbol{G}^{12}_{il}\boldsymbol{U}^{12}_{kj} + \boldsymbol{C}^{11}_{ki}\boldsymbol{C}^{22}_{jl}) \\
=& \sum_{ijkl}(\boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{U}^{21}_{jk}\boldsymbol{E}^{12,\beta*}_{kl}\boldsymbol{G}^{21}_{li} + \boldsymbol{E}^{12,\alpha}_{ij}\boldsymbol{C}^{22}_{jl}\boldsymbol{E}^{21,\beta}_{lk}\boldsymbol{C}^{11}_{ki}) \\
=& \mathrm{tr}(\boldsymbol{E}^{12,\alpha}\boldsymbol{U}^{21}\boldsymbol{E}^{12,\beta*}\boldsymbol{G}^{21} + \boldsymbol{E}^{12,\alpha}\boldsymbol{C}^{22}\boldsymbol{E}^{21,\beta}\boldsymbol{C}^{11}),
\end{aligned} \tag{E21}$$

and

$$\begin{aligned}
\langle \hat{P}^*_\alpha \hat{P}^*_\beta \rangle - \langle \hat{P}^*_\alpha \rangle \langle \hat{P}^*_\beta \rangle =& \sum_{ijkl} \langle \boldsymbol{x}_{1,i} \boldsymbol{E}^{12,\alpha *}_{ij} \boldsymbol{x}^*_{2,j} \boldsymbol{x}_{1,k} \boldsymbol{E}^{12,\beta *}_{kl} \boldsymbol{x}^*_{2,l} \rangle - \langle \boldsymbol{x}_{1,i} \boldsymbol{E}^{12,\alpha *}_{ij} \boldsymbol{x}^*_{2,j} \rangle \langle \boldsymbol{x}_{1,k} \boldsymbol{E}^{12,\beta *}_{kl} \boldsymbol{x}^*_{2,l} \rangle \\
=& \sum_{ijkl} \boldsymbol{E}^{12,\alpha *}_{ij} \boldsymbol{E}^{12,\beta *}_{kl} (\langle \boldsymbol{x}_{1,i} \boldsymbol{x}^*_{2,j} \boldsymbol{x}_{1,k} \boldsymbol{x}^*_{2,l} \rangle - \langle \boldsymbol{x}_{1,i} \boldsymbol{x}^*_{2,j} \rangle \langle \boldsymbol{x}_{1,k} \boldsymbol{x}^*_{2,l} \rangle) \\
=& \sum_{ijkl} \boldsymbol{E}^{12,\alpha *}_{ij} \boldsymbol{E}^{12,\beta *}_{kl} (\langle \boldsymbol{x}_{1,i} \boldsymbol{x}_{1,k} \rangle \langle \boldsymbol{x}^*_{2,j} \boldsymbol{x}^*_{2,l} \rangle + \langle \boldsymbol{x}_{1,i} \boldsymbol{x}^*_{2,l} \rangle \langle \boldsymbol{x}^*_{2,j} \boldsymbol{x}_{1,k} \rangle) \\
=& \sum_{ijkl} \boldsymbol{E}^{12,\alpha *}_{ij} \boldsymbol{E}^{12,\beta *}_{kl} (\boldsymbol{G}^{22}_{jl} \boldsymbol{U}^{11}_{ik} + \boldsymbol{C}^{12}_{il} \boldsymbol{C}^{12}_{kj}) \\
=& \sum_{ijkl} (\boldsymbol{E}^{21,\alpha}_{ji} \boldsymbol{U}^{11}_{ik} \boldsymbol{E}^{12,\beta *}_{kl} \boldsymbol{G}^{22}_{lj} + \boldsymbol{E}^{21,\alpha}_{ji} \boldsymbol{C}^{12}_{il} \boldsymbol{E}^{21,\beta}_{lk} \boldsymbol{C}^{12}_{kj}) \\
=& \mathrm{tr}(\boldsymbol{E}^{21,\alpha} \boldsymbol{U}^{11} \boldsymbol{E}^{12,\beta *} \boldsymbol{G}^{22} + \boldsymbol{E}^{21,\alpha} \boldsymbol{C}^{12} \boldsymbol{E}^{21,\beta} \boldsymbol{C}^{12})\,,
\end{aligned} \tag{E22}$$

where $\boldsymbol{E}^{12,\alpha *}_{ij} = \boldsymbol{E}^{21,\alpha}_{ji}$. Setting $\alpha = \beta$ in these equations then allows us to evaluate Equations (E17) and (E18).

E.2. *Covariance*

The covariance between the real part of $\hat{P}_\alpha$ and the real part of $\hat{P}_\beta$ is

$$\frac{1}{4}\left\{ (\langle \hat{P}_\alpha \hat{P}_\beta \rangle - \langle \hat{P}_\alpha \rangle \langle \hat{P}_\beta \rangle) + (\langle \hat{P}_\alpha \hat{P}^*_\beta \rangle - \langle \hat{P}_\alpha \rangle \langle \hat{P}^*_\beta \rangle) + (\langle \hat{P}^*_\alpha \hat{P}_\beta \rangle - \langle \hat{P}^*_\alpha \rangle \langle \hat{P}_\beta \rangle) + (\langle \hat{P}^*_\alpha \hat{P}^*_\beta \rangle - \langle \hat{P}^*_\alpha \rangle \langle \hat{P}^*_\beta \rangle) \right\}, \tag{E23}$$

and the covariance between the imaginary part of $\hat{P}_\alpha$ and the imaginary part of $\hat{P}_\beta$ is

$$\frac{1}{4}\left\{ (\langle \hat{P}_\alpha \hat{P}_\beta \rangle - \langle \hat{P}_\alpha \rangle \langle \hat{P}_\beta \rangle) - (\langle \hat{P}_\alpha \hat{P}^*_\beta \rangle - \langle \hat{P}_\alpha \rangle \langle \hat{P}^*_\beta \rangle) - (\langle \hat{P}^*_\alpha \hat{P}_\beta \rangle - \langle \hat{P}^*_\alpha \rangle \langle \hat{P}_\beta \rangle) + (\langle \hat{P}^*_\alpha \hat{P}^*_\beta \rangle - \langle \hat{P}^*_\alpha \rangle \langle \hat{P}^*_\beta \rangle) \right\}. \tag{E24}$$

These can be evaluted in the same way as the variances above.

F. SKEWNESS IN DISTRIBUTIONS OF POWER SPECTRA AT INTERMEDIATE DELAYS

In this Appendix, we consider the probability distribution functions of power spectra where neither signals (e.g., foregrounds) or noise are dominant and both must be considered. Using the same notation as Appendix D, the power spectra formed from $\tilde{x}_1 = \tilde{s} + \tilde{n}_1$ and $\tilde{x}_2 = \tilde{s} + \tilde{n}_2$ is

$$\begin{aligned}
P_{\tilde{x}_1 \tilde{x}_2} &= \tilde{s}^* \tilde{s} + \tilde{s}^* \tilde{n}_2 + \tilde{n}^*_1 \tilde{s} + \tilde{n}^*_1 \tilde{n}_2 \\
&= \left[a^2 + b^2 + a(c_1 + c_2) + b(d_1 + d_2) + c_1 c_2 + d_1 d_2\right] + \left[a(d_2 - d_1) + b(c_1 - c_2) + d_2 c_1 - d_1 c_2\right] i\,.
\end{aligned} \tag{F25}$$

Note that a and b are constants and c_1, d_1, c_2 and d_2 are IID randomly normal variables. For the real part of $P_{\tilde{x}_1 \tilde{x}_2}$, we have

$$\langle \mathrm{Re}(P_{\tilde{x}_1 \tilde{x}_2}) \rangle = a^2 + b^2\,. \tag{F26}$$

After subtracting from the mean, its third moment is

$$\left\langle \left[\mathrm{Re}(P_{\tilde{x}_1 \tilde{x}_2}) - (a^2 + b^2)\right]^3 \right\rangle = \left\langle \left[a(c_1 + c_2) + b(d_1 + d_2) + c_1 c_2 + d_1 d_2\right]^3 \right\rangle = 6\langle a^2 c_1^2 c_2^2 + b^2 d_1^2 d_2^2 \rangle > 0\,. \tag{F27}$$

This non-vanishing third moment implies that the probability distribution of power spectra is skewed. This skewness disappears for either signal-dominated or noise-dominated cases. These results are evident in the histograms shown in Figure 3.

G. PROBABILITY DISTRIBUTION FOR AN INCOHERENT SUM OF DELAY TRANSFORM-ESTIMATED POWER SPECTRA

In this Appendix, we derive the probability distribution for noise in a power spectrum that has been formed by the incoherent (i.e., after squaring) averaging of power spectra from individual time integrations. The resulting probability distribution is used in Figures 7, 8, and 9 to validate our error bar methodology.

For a noise-dominated delay power spectrum estimate, the power spectrum value u measured at one instant in time is distributed as a double exponential:

$$p(x) = \frac{1}{\sigma\sqrt{2}} \exp\left(-\frac{\sqrt{2}|u|}{\sigma}\right), \tag{G28}$$

where it is assumed that the power spectra are estimated by cross-correlation—thus eliminating noise bias—and where σ is the standard deviation on the resulting power spectrum.

Now suppose we average together a number of these power spectra. Let the power spectrum value at the ith time step be given by u_i. The average value is then

$$z \equiv \sum_i w_i u_i, \tag{G29}$$

where $\{w_i\}$ are a set of weights. Note that the error on each x_i may be different, so we define

$$p_i(u_i) = \frac{1}{\sigma_i\sqrt{2}} \exp\left(-\frac{\sqrt{2}|u_i|}{\sigma_i}\right). \tag{G30}$$

We now write down the probability distribution $p_+(z)$ for z. First we define $y_i \equiv w_i u_i$, such that

$$p_i(y_i) = \frac{1}{w_i\sigma_i\sqrt{2}} \exp\left(-\frac{\sqrt{2}|y_i|}{w_i\sigma_i}\right). \tag{G31}$$

With this notation, $z = \sum_i y_i$, and we can write down z by using the fact that the probability distribution of a sum of two random variables is the convolution of their individual distributions. By the convolution theorem, this is equivalent to multiplying the Fourier transforms of the individual probability distributions $\widetilde{p}_i(k)$, and thus

$$p_+(z) = \int \frac{dk}{2\pi} e^{ikz} \prod_i \widetilde{p}_i(k) = \int \frac{dk}{2\pi} e^{ikz} \prod_i \frac{1}{1 + w_i^2\sigma_i^2 k^2/2}, \tag{G32}$$

where we have used the fact that in our case, $\widetilde{p}_i(k) = (1 + w_i^2\sigma_i^2 k^2/2)^{-1}$.This integral can be evaluated by contour integration, giving

$$p_+(z) = \sum_j \frac{e^{-|z|\sqrt{2}/w_j\sigma_j}}{w_j\sigma_j\sqrt{2}} \prod_{i\neq j} \frac{1}{1 - w_i^2\sigma_i^2/w_j^2\sigma_j^2}. \tag{G33}$$

This is a weighted sum of double exponential distributions, and the curves in Figures 7, 8, and 9 labeled "Sum of Laplacians" are plots of this formula.

In closing, we note one peculiarity about this derivation—our contour integration assumed that none of the $w_i\sigma_i$ values were exactly equal. In principle, this is a reasonable assumption, since for a drift scan telescope that is sky noise dominated the noise power is continually changing from one time integration to the next. In practice, however, if this change is happening slowly, two adjacent time integrations may have similar enough noise properties to make Equation (G33) numerically problematic. If this is indeed the regime that one is in, it is advisable to instead use an approximate expression by letting $\sqrt{2}/w_i\sigma_i \equiv \kappa + \varepsilon_i$ and then Taylor expanding to leading order in ε_i.

REFERENCES

Ali, Z. S., Parsons, A. R., Zheng, H., et al. 2015, ApJ, 809, 61, doi: 10.1088/0004-637X/809/1/61

Barkana, R., & Loeb, A. 2001, PhR, 349, 125, doi: 10.1016/S0370-1573(01)00019-9

Barry, N., Wilensky, M., Trott, C. M., et al. 2019, ApJ, 884, 1, doi: 10.3847/1538-4357/ab40a8

Beardsley, A. P., Hazelton, B. J., Sullivan, I. S., et al. 2016, ApJ, 833, 102, doi: 10.3847/1538-4357/833/1/102

Becker, G. D., Bolton, J. S., Madau, P., et al. 2015, MNRAS, 447, 3402, doi: 10.1093/mnras/stu2646

Becker, R. H., Fan, X., White, R. L., et al. 2001, AJ, 122, 2850, doi: 10.1086/324231

Bernardi, G., Mitchell, D. A., Ord, S. M., et al. 2011, MNRAS, 413, 411, doi: 10.1111/j.1365-2966.2010.18145.x

Bernardi, G., de Bruyn, A. G., Brentjens, M. A., et al. 2009, A&A, 500, 965, doi: 10.1051/0004-6361/200911627

Bolton, J. S., Haehnelt, M. G., Warren, S. J., et al. 2011, MNRAS, 416, L70, doi: 10.1111/j.1745-3933.2011.01100.x

Bosman, S. E. I., Fan, X., Jiang, L., et al. 2018, MNRAS, 479, 1055, doi: 10.1093/mnras/sty1344

Bowman, J. D., Rogers, A. E. E., & Hewitt, J. N. 2008, ApJ, 676, 1, doi: 10.1086/528675

Bowman, J. D., Rogers, A. E. E., Monsalve, R. A., Mozdzen, T. J., & Mahesh, N. 2018a, Nature, 555, 67, doi: 10.1038/nature25792

—. 2018b, Nature, 564, E35, doi: 10.1038/s41586-018-0797-4

Bowman, J. D., Cairns, I., Kaplan, D. L., et al. 2013, PASA, 30, e031, doi: 10.1017/pas.2013.009

Bradley, R. F., Tauscher, K., Rapetti, D., & Burns, J. O. 2019, ApJ, 874, 153, doi: 10.3847/1538-4357/ab0d8b

Chapman, E., Abdalla, F. B., Harker, G., et al. 2012, MNRAS, 423, 2518, doi: 10.1111/j.1365-2966.2012.21065.x

Cheng, C., Parsons, A. R., Kolopanis, M., et al. 2018, ApJ, 868, 26, doi: 10.3847/1538-4357/aae833

Cho, J., Lazarian, A., & Timbie, P. T. 2012, ApJ, 749, 164, doi: 10.1088/0004-637X/749/2/164

Choudhuri, S., Bull, P., & Garsden, H. 2021, arXiv e-prints, arXiv:2101.02684. https://arxiv.org/abs/2101.02684

Datta, A., Bowman, J. D., & Carilli, C. L. 2010, ApJ, 724, 526, doi: 10.1088/0004-637X/724/1/526

Davies, F. B., Hennawi, J. F., Bañados, E., et al. 2018, ApJ, 864, 142, doi: 10.3847/1538-4357/aad6dc

Dayal, P., & Ferrara, A. 2018, PhR, 780, 1, doi: 10.1016/j.physrep.2018.10.002

de Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., et al. 2008, MNRAS, 388, 247, doi: 10.1111/j.1365-2966.2008.13376.x

DeBoer, D. R., Parsons, A. R., Aguirre, J. E., et al. 2017, PASP, 129, 045001, doi: 10.1088/1538-3873/129/974/045001

Dillon, J. S., Liu, A., Williams, C. L., et al. 2014, PhRvD, 89, 023002, doi: 10.1103/PhysRevD.89.023002

Dillon, J. S., Neben, A. R., Hewitt, J. N., et al. 2015, PhRvD, 91, 123011, doi: 10.1103/PhysRevD.91.123011

Dillon, J. S., Lee, M., Ali, Z. S., et al. 2020, MNRAS, doi: 10.1093/mnras/staa3001

Efron, B., & Tibshirani, R. J. 1994, An introduction to the bootstrap (CRC press)

Fan, X., Carilli, C. L., & Keating, B. 2006, ARA&A, 44, 415, doi: 10.1146/annurev.astro.44.051905.092514

Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, PhR, 433, 181, doi: 10.1016/j.physrep.2006.08.002

Gehlot, B. K., Mertens, F. G., Koopmans, L. V. E., et al. 2019, MNRAS, 488, 4271, doi: 10.1093/mnras/stz1937

Ghara, R., Giri, S. K., Mellema, G., et al. 2020, MNRAS, 493, 4728, doi: 10.1093/mnras/staa487

Greig, B., & Mesinger, A. 2015, MNRAS, 449, 4246, doi: 10.1093/mnras/stv571

—. 2017, MNRAS, 472, 2651, doi: 10.1093/mnras/stx2118

Harker, G., Zaroubi, S., Bernardi, G., et al. 2009, MNRAS, 397, 1138, doi: 10.1111/j.1365-2966.2009.15081.x

Hassan, S., Davé, R., Finlator, K., & Santos, M. G. 2017, MNRAS, 468, 122, doi: 10.1093/mnras/stx420

Hazelton, B. J., Morales, M. F., & Sullivan, I. S. 2013, ApJ, 770, 156, doi: 10.1088/0004-637X/770/2/156

HERA Collaboration. 2021

Hills, R., Kulkarni, G., Meerburg, P. D., & Puchwein, E. 2018, Nature, 564, E32, doi: 10.1038/s41586-018-0796-5

Jacobs, D. C., Pober, J. C., Parsons, A. R., et al. 2015, ApJ, 801, 51, doi: 10.1088/0004-637X/801/1/51

Jelić, V., Zaroubi, S., Labropoulos, P., et al. 2008, MNRAS, 389, 1319, doi: 10.1111/j.1365-2966.2008.13634.x

Kern, N. S., Liu, A., Parsons, A. R., Mesinger, A., & Greig, B. 2017, ApJ, 848, 23, doi: 10.3847/1538-4357/aa8bb4

Kern, N. S., Parsons, A. R., Dillon, J. S., et al. 2019, ApJ, 884, 105, doi: 10.3847/1538-4357/ab3e73

—. 2020a, ApJ, 888, 70, doi: 10.3847/1538-4357/ab5e8a

Kern, N. S., Dillon, J. S., Parsons, A. R., et al. 2020b, ApJ, 890, 122, doi: 10.3847/1538-4357/ab67bc

Kohn, S. A., Aguirre, J. E., La Plante, P., et al. 2019, ApJ, 882, 58, doi: 10.3847/1538-4357/ab2f72

Kolopanis, M., Jacobs, D. C., Cheng, C., et al. 2019, ApJ, 883, 133, doi: 10.3847/1538-4357/ab3e3a

Koopmans, L., Pritchard, J., Mellema, G., et al. 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14), 1. https://arxiv.org/abs/1505.07568

Lanman, A. E., & Pober, J. C. 2019, MNRAS, 487, 5840, doi: 10.1093/mnras/stz1639

Li, W., Pober, J. C., Barry, N., et al. 2019, ApJ, 887, 141, doi: 10.3847/1538-4357/ab55e4

Liu, A., Parsons, A. R., & Trott, C. M. 2014a, PhRvD, 90, 023018, doi: 10.1103/PhysRevD.90.023018

—. 2014b, PhRvD, 90, 023019, doi: 10.1103/PhysRevD.90.023019

Liu, A., & Shaw, J. R. 2020, PASP, 132, 062001, doi: 10.1088/1538-3873/ab5bfd

Liu, A., & Tegmark, M. 2011, PhRvD, 83, 103006, doi: 10.1103/PhysRevD.83.103006

McQuinn, M., Zahn, O., Zaldarriaga, M., Hernquist, L., & Furlanetto, S. R. 2006, ApJ, 653, 815, doi: 10.1086/505167

Mellema, G., Koopmans, L. V. E., Abdalla, F. A., et al. 2013, Experimental Astronomy, 36, 235, doi: 10.1007/s10686-013-9334-5

Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, MNRAS, 493, 1662, doi: 10.1093/mnras/staa327

Mondal, R., Bharadwaj, S., & Majumdar, S. 2016, MNRAS, 456, 1936, doi: 10.1093/mnras/stv2772

—. 2017, MNRAS, 464, 2992, doi: 10.1093/mnras/stw2599

Morales, M. F. 2005, ApJ, 619, 678, doi: 10.1086/426730

Morales, M. F., Beardsley, A., Pober, J., et al. 2019, MNRAS, 483, 2207, doi: 10.1093/mnras/sty2844

Morales, M. F., Hazelton, B., Sullivan, I., & Beardsley, A. 2012, ApJ, 752, 137, doi: 10.1088/0004-637X/752/2/137

Morales, M. F., & Wyithe, J. S. B. 2010, ARA&A, 48, 127, doi: 10.1146/annurev-astro-081309-130936

Ouchi, M., Shimasaku, K., Furusawa, H., et al. 2010, ApJ, 723, 869, doi: 10.1088/0004-637X/723/1/869

Park, J., Mesinger, A., Greig, B., & Gillet, N. 2019, MNRAS, 484, 933, doi: 10.1093/mnras/stz032

Parsons, A., Pober, J., McQuinn, M., Jacobs, D., & Aguirre, J. 2012a, ApJ, 753, 81, doi: 10.1088/0004-637X/753/1/81

Parsons, A. R., Liu, A., Ali, Z. S., & Cheng, C. 2016, ApJ, 820, 51, doi: 10.3847/0004-637X/820/1/51

Parsons, A. R., Pober, J. C., Aguirre, J. E., et al. 2012b, ApJ, 756, 165, doi: 10.1088/0004-637X/756/2/165

Parsons, A. R., Backer, D. C., Foster, G. S., et al. 2010, AJ, 139, 1468, doi: 10.1088/0004-6256/139/4/1468

Parsons, A. R., Liu, A., Aguirre, J. E., et al. 2014, ApJ, 788, 106, doi: 10.1088/0004-637X/788/2/106

Patil, A. H., Yatawatta, S., Koopmans, L. V. E., et al. 2017, ApJ, 838, 65, doi: 10.3847/1538-4357/aa63e7

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A6, doi: 10.1051/0004-6361/201833910

Pober, J. C., Parsons, A. R., DeBoer, D. R., et al. 2013, AJ, 145, 65, doi: 10.1088/0004-6256/145/3/65

Pober, J. C., Liu, A., Dillon, J. S., et al. 2014, ApJ, 782, 66, doi: 10.1088/0004-637X/782/2/66

Pober, J. C., Ali, Z. S., Parsons, A. R., et al. 2015, ApJ, 809, 62, doi: 10.1088/0004-637X/809/1/62

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2007, Numerical recipes 3rd edition: The art of scientific computing (Cambridge university press)

Pritchard, J. R., & Loeb, A. 2012, Reports on Progress in Physics, 75, 086901, doi: 10.1088/0034-4885/75/8/086901

Shaw, A. K., Bharadwaj, S., & Mondal, R. 2019, MNRAS, 487, 4951, doi: 10.1093/mnras/stz1561

Shaw, J. R., Sigurdson, K., Sitwell, M., Stebbins, A., & Pen, U.-L. 2015, PhRvD, 91, 083514, doi: 10.1103/PhysRevD.91.083514

Sims, P. H., & Pober, J. C. 2020, MNRAS, 492, 22, doi: 10.1093/mnras/stz3388

Singh, S., & Subrahmanyan, R. 2019, ApJ, 880, 26, doi: 10.3847/1538-4357/ab2879

Singh, S., Subrahmanyan, R., Shankar, N. U., et al. 2018, Experimental Astronomy, 45, 269, doi: 10.1007/s10686-018-9584-3

Stark, D. P., Ellis, R. S., Chiu, K., Ouchi, M., & Bunker, A. 2010, MNRAS, 408, 1628, doi: 10.1111/j.1365-2966.2010.17227.x

Tegmark, M. 1997, ApJL, 480, L87, doi: 10.1086/310631

Thompson, A. R., Moran, J. M., & Swenson, Jr., G. W. 2017, Interferometry and Synthesis in Radio Astronomy, 3rd Edition (Springer International Publishing), doi: 10.1007/978-3-319-44431-4

Thyagarajan, N., Udaya Shankar, N., Subrahmanyan, R., et al. 2013, ApJ, 776, 6, doi: 10.1088/0004-637X/776/1/6

Tingay, S. J., Goeke, R., Bowman, J. D., et al. 2013, PASA, 30, e007, doi: 10.1017/pasa.2012.007

Trott, C. M. 2014, PASA, 31, e026, doi: 10.1017/pasa.2014.23

Trott, C. M., Wayth, R. B., & Tingay, S. J. 2012, ApJ, 757, 101, doi: 10.1088/0004-637X/757/1/101

Trott, C. M., Jordan, C. H., Midgley, S., et al. 2020, MNRAS, 493, 4711, doi: 10.1093/mnras/staa414

van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2, doi: 10.1051/0004-6361/201220873

Vedantham, H., Udaya Shankar, N., & Subrahmanyan, R. 2012, ApJ, 745, 176, doi: 10.1088/0004-637X/745/2/176

Wijnholds, S. J., Willis, A. G., & Salvini, S. 2018, MNRAS, 476, 2029, doi: 10.1093/mnras/sty360

Zaldarriaga, M., Furlanetto, S. R., & Hernquist, L. 2004, ApJ, 608, 622, doi: 10.1086/386327