The NANOGrav 15 yr dataset: Posterior predictive checks for gravitational-wave detection with pulsar timing arrays

Gabriella Agazie, Akash Anumarlapudi, Anne M. Archibald, Zaven Arzoumanian, Jeremy George Baier, Paul T. Baker, Bence Bécsy, Laura Blecha, Adam Brazier, Paul R. Brook, Sarah Burke-Spolaor, J. Andrew Casey-Clyde, Maria Charisi, Shami Chatterjee, Katerina Chatziioannou, Tyler Cohen, James M. Cordes, Neil J. Cornish, Fronefield Crawford, H. Thankful Cromartie, Kathryn Crowter, Megan E. DeCesar, Paul B. Demorest, Heling Deng, Lankeswar Dey, Timothy Dolch, Elizabeth C. Ferrara, William Fiore, Emmanuel Fonseca, Gabriel E. Freedman, Emiko C. Gardiner, Nate Garver-Daniels, Peter A. Gentile, Kyle A. Gersbach, Joseph Glaser, Deborah C. Good, Kayhan Gültekin, Jeffrey S. Hazboun, Ross J. Jennings, Aaron D. Johnson, Megan L. Jones, Andrew R. Kaiser, David L. Kaplan, Luke Zoltan Kelley, Matthew Kerr, Joey S. Key, Nima Laal, Michael T. Lam, William G. Lamb, Bjorn Larsen, T. Joseph, W. Lazio, Natalia Lewandowska, Tingting Liu, Duncan R. Lorimer, Jing Luo, Ryan S. Lynch, Chung-Pei Ma, Dustin R. Madison, Alexander McEwen, James W. McKee, Maura A. McLaughlin, Natasha McMann, Bradley W. Meyers, Patrick M. Meyers, Chiara M. F. Mingarelli, Andrea Mitridate, Cherry Ng, David J. Nice, Stella Koch Ocker, Ken D. Olum, Timothy T. Pennucci, Benetge B. P. Perera, Nihan S. Pol, Henri A. Radovan, Scott M. Ransom, Paul S. Ray, Joseph D. Romano, Jessie C. Runnoe, Alexander Saffer, Shashwat C. Sardesai, Ann Schmiedekamp, Carl Schmiedekamp, Kai Schmitz, Brent J. Shapiro-Albert, Xavier Siemens, Joseph Simon, Magdalena S. Siwek, Sophia V. Sosa Fiscella, Ingrid H. Stairs, Daniel R. Stinebring, Kevin Stovall, Abhimanyu Susobhanan, Joseph K. Swiggum, Stephen R. Taylor, Jacob E. Turner, Caner Unal, Michele Vallisneri, Sarah J. Vigeland, Haley M. Wahl, Caitlin A. Witt, David Wright, and Olivia Young

(NANOGrav Collaboration)

(Received 2 August 2024; accepted 22 November 2024; published 28 February 2025)

Pulsar timing array experiments have reported evidence for a stochastic background of nanohertz gravitational waves consistent with the signal expected from a population of supermassive black hole binaries. Their analyses assume power-law spectra for intrinsic pulsar noise and for the background, as well as a Hellings-Downs cross-correlation pattern among the gravitational-wave-induced residuals across pulsars. These assumptions may not be realized in actuality. We test them in the NANOGrav 15 yr dataset using Bayesian posterior predictive checks. After fitting our fiducial model to real data, we generate a population of simulated dataset replications. We use the replications to assess whether the optimal statistic significance, interpulsar correlations, and spectral coefficients are extreme. We recover Hellings-Downs correlations in simulated datasets at significance levels consistent with the correlations measured in the NANOGrav 15 yr dataset. A similar test on spectral coefficients shows that their values in real data are not extreme compared to their distributions across replications. We also evaluate the evidence for the stochastic background using posterior predictive versions of the frequentist optimal statistic and of Bayesian model comparison and find comparable significance (3.2σ and 3σ respectively) to what was previously reported for the standard statistics. We conclude with novel visualizations of the reconstructed gravitational waveforms that enter the residuals for each pulsar. Our analysis strengthens confidence in the identification and characterization of the gravitational-wave background.

DOI: 10.1103/PhysRevD.111.042011

I. INTRODUCTION

In June 2023, four separate publications based on the observations of five pulsar timing array (PTA) collaborations reported strong evidence for a nanohertz gravitational-wave (GW) background [1–4], spurring interest in the implications of its spectral properties and spatial correlations for astrophysics and fundamental physics [5–8]. If the signal originates from a population of supermassive black hole binaries (SMBHBs), its spectrum is expected to approximate a power law [9,10], but deviations can be caused by a large number of potential effects. For example, at low frequencies, interactions between the binaries and the surrounding gas may result in a spectral turnover; at high frequencies, the

finite number of binaries emitting in each frequency bin may result in bin-to-bin fluctuations [5,7,11]. If the signal originates from new physics, the spectrum can point to the mechanism of its generation, and a large number of models are currently consistent with the data [6,12].

Spatial correlations between pulse times of arrival (TOAs) for different pulsars were found to be consistent with the Hellings-Downs function, the correlation pattern induced by an isotropic GW background [13–15]. Deviations could be caused by anisotropy in the background, by a signal from a loud individual SMBHB. Measuring anisotropy would constrain black hole population properties [16], while detecting an individual SMBHB would offer a prime target for multimessenger follow-up. However, dedicated searches for anisotropy and individual sources have so far produced null results [17–19]. Systematic errors could also induce correlations between pulsars, e.g., monopolar correlations due to clock errors or dipolar correlations induced by errors in the Solar System ephemeris [20,21]. There is slight evidence for monopolar correlations presented in the NANOGrav 15 yr dataset [1].

Simulations can address the expected level of anisotropy from a population of SMBHBs [16,22] and its detectability using standard PTA models, which assume an isotropic GW background with Gaussian statistics and a stationary power-law spectrum. Indeed, Refs. [23,24] found that the GW signal from a realistic SMBHB population would still be detected using standard models. Thus, current PTA observations [1–4,25] do not preclude the presence of astrophysically interesting deviations from power-law spectrum or isotropy.

In this paper, we ask whether the power-law and Hellings-Downs assumptions are supported by observed data, independent of any specific alternative physical model. Our starting point is a fiducial Bayesian analysis of NANOGrav's 15 yr dataset [26] under the standard power-law, Hellings-Downs model. We test these assumptions by way of "posterior predictive model checks" [27] as proposed in the context of PTA data in Refs. [28,29]. These checks consist of creating populations of *replicated* datasets from real-data parameter posteriors and using these replications to evaluate whether real data is "typical" (i.e., not a statistical outlier) according to a variety of detection, spectral, and correlation statistics. Similar types of checks are becoming increasingly common in the realm of binary black hole population analyses as well [30–38].

Specifically, following Ref. [28] we reevaluate the significance of Hellings-Downs correlations and search for alternative spatial correlations using a new detection statistic that marginalizes p-values over noise-parameter posteriors. Following Ref. [29] we test the power-law assumption by comparing intrinsic-noise and GW power-spectrum posteriors as computed for real and replicated data, and we perform a similar test for the binned angular correlations between pulsars. We also carry out leave-one-out cross validation to identify possible mismodeling in individual pulsars and to compute the pseudo-Bayes factor (a cross-validation metric of model comparison) between the standard Hellings-Downs model and a null model in which common excess power has no interpulsar correlations.

The rest of this paper is organized as follows. In Sec. II we describe our data and data model, and we introduce two sets of data replications that we will use for model checking. In Sec. III we test Hellings-Downs correlations using "Bayesian p-values" [28] for the optimal statistic [39,40]; these p-values are marginalized over GW and intrinsic-noise posteriors, and therefore account fairly for the risk of false positives when the null distribution is uncertain. We find evidence for Hellings-Downs correlations at the 3.2σ level. We also evaluate the evidence for additional background components with monopolar or dipolar correlations and find none.

In Sec. IV we compare real-data and replicated-data posteriors to search for deviations from a power-law spectrum and from Hellings-Downs correlations. We find no evidence that any individual frequency bin deviates from the power-law model for either intrinsic pulsar noise or the GW background, consistent with Ref. [41]. We also find no evidence that any of the binned interpulsar correlations deviate from the Hellings-Downs curve.

In Sec. V we examine the predictive power of the standard PTA model as fit to the NANOGrav 15 yr data. We perform a leave-one-out analysis where we fit Hellings-Downs and uncorrelated models to $N_p - 1$ pulsars and use the models to predict the N_p^{th} pulsar's data. The resulting pseudo-Bayes factors favor Hellings-Downs correlations at the 3σ level. Using simulations, we show that the distribution of the factors across pulsars is consistent with what would be expected for a power-law, Hellings-Downs-correlated GW background with parameters from our fiducial analysis.

Last, in Sec. VI we present the gravitational waveforms that can be reconstructed for each pulsar from our fiducial posteriors. These reconstructions are akin to the waveform reconstructions for stellar-binary coalescences based on LIGO data [42,43], with the distinction being that in this case we show the estimated realization of a broadband, spatially correlated stochastic signal, as opposed to the gravitational waveform produced by a single binary system. Pulsar J1909-3744 offers the best view so far of the GW background reported in [1–4,25]. In Sec. VII we offer concluding remarks.

II. DATA, MODEL, AND DATA REPLICATIONS

In this section we introduce the NANOGrav 15 yr dataset, the modeling that is performed on each pulsar,

and the full PTA models used to search for a GW background. We then discuss data replications based on our typical PTA models, which we use in subsequent sections to compare to the 15 yr dataset for the purposes of model checking and model comparison.

A. Data

We use the NANOGrav 15 yr dataset, which contains 67 pulsars that have been timed for more than 3 yr, with 16.03 yr of data between the first and the last time of arrival in the dataset [26]. We use the DMX dispersion measure noise model [44] and white noise parameters included in the NANOGrav 15 yr data release [26]. For each pulsar, a best-fit timing model is constructed that accounts for deterministic effects like Roemer delay, proper motion, parallax, binary orbits, etc., which is then subtracted from the TOAs to produce a set of timing residuals for each pulsar, $\delta\mathbf{t}$. Stochastic processes like achromatic intrinsic spin wandering and GW background-induced delays are included in this initial fit as a single "total red noise" contribution, as the first pass analysis is done on a pulsar-by-pulsar basis and so we cannot separate intrinsic pulsar noise from the GW background.

B. PTA model

In this subsection, we discuss the full PTA model that is used to search for a GW background. Readers familiar with this already can skip to Sec. III, although later we will make frequent reference to equations introduced in this section. For a more in-depth presentation of the PTA analyses, see Refs. [45–47].

The starting point for the analysis are the timing residuals, $\delta\mathbf{t}$. We characterize stochastic processes like intrinsic pulsar noise and the GW background in the frequency domain using a Fourier matrix $\mathbf{F}$ and associated amplitudes $\mathbf{a}$ [48]. The stochastic processes are covariant with elements of the timing model (specifically the frequency, spin-down, and dispersion measure variations), and so we also introduce deviations from the best-fitting timing model parameters, $\boldsymbol{\epsilon}$. We assume these deviations are small, such that changes in $\delta\mathbf{t}$ are linear in changes in $\boldsymbol{\epsilon}$ with a design matrix $\mathbf{M}$ made up of derivatives of $\delta\mathbf{t}$ with respect to the timing model parameters. Putting these effects together, we have a model for the residuals

$$\mathbf{r} = \delta\mathbf{t} - \mathbf{T}\mathbf{b}, \tag{1}$$

where we have consolidated the frequency domain representation and timing model corrections,

$$\mathbf{T} = [\,\mathbf{M} \quad \mathbf{F}\,], \tag{2}$$

$$\mathbf{b} = \begin{bmatrix} \boldsymbol{\epsilon} \\ \mathbf{a} \end{bmatrix}. \tag{3}$$

If radio frequency interference is effectively excised and standard pulse profiles are accurate, the resulting noise is dominated by radiometer noise and "pulse profile jitter" which is traditionally assumed to be frequency independent and Gaussian. This leads to a Gaussian likelihood for the timing residuals

$$\ln p(\delta\mathbf{t}|\mathbf{b}) = -\frac{1}{2}[\mathbf{r}^T\mathbf{N}^{-1}\mathbf{r} + \ln\det(2\pi\mathbf{N})], \tag{4}$$

where the covariance matrix $\mathbf{N}$ describes the measurement noise of the individual observations and is block diagonal. TOA at different radio frequencies from the same individual observation are correlated with one another due to pulse profile jitter [49], but TOAs from different observations are uncorrelated.

We assume that the GW background and the intrinsic pulsar noise are stationary, and so they can be characterized by the power spectrum of the GW background, correlations between pulsars, and the power spectrum of the intrinsic pulsar noise in each pulsar. The assumption of stationarity for the GW background should hold if the dominant contribution to the background is an ensemble of SMBHBs emitting at roughly constant frequencies. The assumption that intrinsic pulsar noise is stationary is one of expedience that should be tested. Tests on the European Pulsar Timing Array second data release show no signs of nonstationarity [50].

Information about the power-law amplitude and spectral index for the intrinsic pulsar noise and the GW background is encoded in the covariance matrix of the sine and cosine amplitudes $\mathbf{a}$ across pulsars. We introduce a set of hyperparameters $\boldsymbol{\Lambda}$ to characterize these power laws. We place a Gaussian prior on $\mathbf{b}$,

$$\ln p(\mathbf{b}|\boldsymbol{\Lambda}) = -\frac{1}{2}[\mathbf{b}^T\mathbf{B}^{-1}\mathbf{b} + \ln\det(2\pi\mathbf{B})], \tag{5}$$

$$\text{where } \mathbf{B} = \begin{bmatrix} \infty & 0 \\ 0 & \boldsymbol{\varphi}(\boldsymbol{\Lambda}) \end{bmatrix}. \tag{6}$$

We use an improper uniform prior on $\boldsymbol{\epsilon}$ so that its posterior is determined by the likelihood. This prior is now broadcast across $\mathbf{b}$ parameters for each pulsar. The covariance matrix of the $\mathbf{a}$ coefficients is given by $\boldsymbol{\varphi}(\boldsymbol{\Lambda})$, which contains blocks corresponding to correlations of the Fourier modes between pulsars. Diagonal blocks encode information about the power spectrum of the total red noise for a given pulsar, including the intrinsic pulsar noise $\boldsymbol{\eta}_a(\boldsymbol{\Lambda})$ (where the a subscript labels the pulsar) and the GW background spectrum $\boldsymbol{\rho}(\boldsymbol{\Lambda})$. Off-diagonal blocks between pulsars a and b contain (scaled) contributions from the GW background. Putting all of this together, the covariance matrix for $\mathbf{a}$ is

$$\varphi(\boldsymbol{\Lambda})_{(ai,bj)} = \Gamma_{ab}\rho_i^2(\boldsymbol{\Lambda})\delta_{ij} + \eta_{ai}^2(\boldsymbol{\Lambda})\delta_{ij}\delta_{ab}, \tag{7}$$

where i and j label frequencies and Γ_{ab} corresponds to the correlations between pulsars. Different angular correlation patterns correspond to different models. In this paper we consider four models. The first states that Γ_{ab} follows the Hellings-Downs curve (HD model) that is expected from an isotropic GW background,

$$\Gamma_{ab} = \frac{1}{2}\delta_{ab} + \frac{1}{2} - \frac{\zeta_{ab}}{4} + \frac{3}{2}\zeta_{ab}\ln\zeta_{ab}, \tag{8}$$

$$\zeta_{ab} = \frac{1-\cos\theta_{ab}}{2}, \tag{9}$$

where θ_{ab} is the angle between pulsars a and b on the sky. The second is that $\Gamma_{ab} = \delta_{ab}$, which we call the common uncorrelated red noise (CURN) model. We will also consider a MONO model that is characterized by monopolar correlations, $\Gamma_{ab} = 1$, and a model with dipolar correlations (DIP) with $\Gamma_{ab} = \cos\theta_{ab}$. Theoretical models indicate that $\rho_i(\mathbf{\Lambda})$ will roughly take the form of a power law, and past empirical studies suggest that $\eta_{ai}(\mathbf{\Lambda})$ often follows a power law as well,

$$\eta^2_{ai}(\mathbf{\Lambda}) = \frac{A^2_{\rm rn,a}}{12\pi^2}\left(\frac{f_i}{f_{\rm yr}}\right)^{-\gamma_{\rm rn,a}}\frac{f_{\rm yr}^{-3}}{T}, \tag{10}$$

$$\rho^2_i(\mathbf{\Lambda}) = \frac{A^2_{\rm gw}}{12\pi^2}\left(\frac{f_i}{f_{\rm yr}}\right)^{-\gamma_{\rm gw}}\frac{f_{\rm yr}^{-3}}{T}, \tag{11}$$

where $A_{\rm gw}$ is the amplitude of the GW background at $f_{\rm yr} = (1\ \text{yr})^{-1}$, $\gamma_{\rm gw}$ is the negative spectral index, $A_{\rm rn,a}$ is the amplitude of intrinsic pulsar noise for pulsar a, and $\gamma_{\rm rn,a}$ its associated spectral index. The frequency is given by $f_i = i/T$, and T is the time between the first and last TOAs in the dataset. For intrinsic pulsar noise we use 30 frequencies, $i \in [1, 30]$ and for the GW background we use $i \in [1, 14]$. These numbers were chosen based on individual-pulsar fitting (for the intrinsic pulsar noise) and a dedicated CURN analysis that allows for the common spectrum to "flatten" at high frequencies, where it then becomes indistinguishable from white noise.

The power-law models for the GW background and intrinsic pulsar noise spectra have amplitudes and spectral indices associated with them: $A_{\rm gw}$, $\gamma_{\rm gw}$ for the GW background and $A_{\rm rn,a}$, $\gamma_{\rm rn,a}$ for each of the N_p pulsars in the array. We collectively denote these parameters as $\mathbf{\Lambda}$. To reduce the total number of parameters we need to infer, we typically marginalize over the model parameters $\mathbf{b}$, leaving a posterior on the hyperparameters

$$p(\mathbf{\Lambda}|\delta\mathbf{t}) = \int d\mathbf{b}\, p(\delta\mathbf{t}|\mathbf{b})p(\mathbf{b}|\mathbf{\Lambda})p(\mathbf{\Lambda}), \tag{12}$$

$$= \frac{p(\mathbf{\Lambda})}{\sqrt{\det(2\pi\mathbf{C})}}\exp\left(-\frac{1}{2}\delta\mathbf{t}^T\mathbf{C}^{-1}\delta\mathbf{t}\right). \tag{13}$$

The covariance matrix is now $\mathbf{C} = (\mathbf{N} + \mathbf{T}\mathbf{B}\mathbf{T}^T)$, and we introduced a prior on the hyperparameters $p(\mathbf{\Lambda})$. We also note that $p(\mathbf{b}|\delta\mathbf{t}, \mathbf{\Lambda}) \propto p(\delta\mathbf{t}|\mathbf{b})p(\mathbf{b}|\mathbf{\Lambda}) = \mathcal{N}(\hat{\mathbf{b}}, \mathbf{\Sigma})$, which is normal with mean and covariance given by

$$\hat{\mathbf{b}} = \mathbf{\Sigma}\mathbf{T}^T\mathbf{N}^{-1}\delta\mathbf{t}, \tag{14}$$

$$\mathbf{\Sigma} = (\mathbf{T}^T\mathbf{N}^{-1}\mathbf{T} + \mathbf{B}^{-1})^{-1}. \tag{15}$$

We estimate the marginalized posterior on $\mathbf{\Lambda}$ using stochastic sampling methods [51] because of the large dimension of $\mathbf{\Lambda}$ ($2N_p + 2$ in the case described above). This yields N_s samples $\{\mathbf{\Lambda}^s\}_{s=1}^{N_s}$ approximately drawn from the posterior,

$$\mathbf{\Lambda}^s \sim p(\mathbf{\Lambda}|\delta\mathbf{t}). \tag{16}$$

C. Data replications

Below we use $\delta\mathbf{t}$ to refer to generic timing residuals, $\delta\mathbf{t}^{15\ \rm yr}$ to refer to residuals from the 15 yr dataset, and $\delta\mathbf{t}^{\rm rep}$ to refer to data replications. We use two models to create sets of data replications to compare to the collected data. Each method proceeds along similar lines:

(1) Choose $\mathbf{\Lambda}$ by drawing randomly from $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr})$.
(2) Draw $\mathbf{b} \sim p(\mathbf{b}|\mathbf{\Lambda})$. The choice of $p(\mathbf{b}|\mathbf{\Lambda})$ depends upon the set of replications we are performing. We specify details below when we discuss individual replication sets. This method nominally calls for us to draw from the improper prior on $\boldsymbol{\epsilon}$, yielding unusable timing residuals. Therefore, we do not simulate timing model variations, and fix $\boldsymbol{\epsilon} \approx 0$.
(3) Draw $\delta\mathbf{t}^{\rm rep} \sim \mathcal{N}(\mathbf{T}\mathbf{b}, \mathbf{N})$ where $\mathbf{b}$ comes from the previous step.

The data replications use different models at each stage. We outline the different data replication sets, their purpose, and what models they use to carry out the procedure described above.

(i) CURNPosteriorDraws: We create simulated datasets based on the CURN model which we index with s. We draw $\mathbf{\Lambda}^s \sim p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr}, \text{CURN})$, and $\mathbf{b}^s \sim \mathcal{N}(0, \mathbf{B}(\mathbf{\Lambda}^s)|\text{CURN})$, Eqs. (5) and (6). The conditioning on CURN implies no correlations between pulsars, $\Gamma_{ab} = \delta_{ab}$. We do not simulate timing model variations, i.e., $\boldsymbol{\epsilon} = 0$. These sets of data replications are compared to the data and recovered model parameters.
(ii) HDPosteriorDraws: We create simulated datasets based on the HD model which we index with s. We draw $\mathbf{\Lambda}^s \sim p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr}, HD)$, and $\mathbf{b}^s \sim \mathcal{N}(0, \mathbf{B}(\mathbf{\Lambda}^s)|HD)$. The conditioning on HD implies we include Hellings-Downs correlations between pulsars during simulation. We do not simulate timing model variations, i.e., $\boldsymbol{\epsilon} = 0$. These sets of data replications are

compared to the real data and recovered model parameters to assess how consistent the data are with the HD model.

III. POSTERIOR PREDICTIVE NULL HYPOTHESIS TESTING

Given $\mathbf{\Lambda}^s \sim p(\mathbf{\Lambda}|\delta\mathbf{t})$, we perform "posterior predictive checks" by checking whether specific desired properties of the model are consistent in the data. To do this, we construct a test statistic $T(\delta\mathbf{t}, \mathbf{\Lambda})$ that is sensitive to the property we are interested in, and we compare that test statistic calculated in the 15 yr data to the same statistic calculated over data replications. Using this method we check (1) whether the 15 yr data are consistent with the lack of correlations assumed by the CURN model; (2) whether the 15 yr data have correlations that are consistent with the HD curve; and (3) whether the 15 yr data show evidence for alternative spatial correlations, e.g., monopolar or dipolar, inconsistent with both the HD model and the CURN model.

A. CURN model tests

The CURN model is characterized by a lack of spatial correlations between pulsars, $\Gamma_{ab} = \delta_{ab}$. To reject this model, we use the optimal statistic signal-to-noise ratio (SNR) as our test statistic [39,40,52,53]. The SNR, for a given choice of noise parameters, is distributed according to a generalized χ^2 distribution [54]; it is large when Hellings-Downs correlations are present and centered around zero when no spatial correlations are present.

The optimal statistic depends upon the total red noise in each pulsar (intrinsic pulsar noise and GW background), which we do not know *a priori*. Therefore, current analyses average the SNR over the posterior distribution on the noise parameters,

$$\begin{aligned}\overline{\mathrm{SNR}} &= \int \mathrm{d}\mathbf{\Lambda}\, p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \mathrm{yr}})\mathrm{SNR}(\delta\mathbf{t}^{15\ \mathrm{yr}};\mathbf{\Lambda}) \\ &\approx \frac{1}{N_s}\sum_{s=1}^{N_s}\mathrm{SNR}(\delta\mathbf{t}^{15\ \mathrm{yr}};\mathbf{\Lambda}^s),\end{aligned} \tag{17}$$

where in the second line we perform a Monte Carlo integral using a finite set of N_s posteriors samples [53]. $\overline{\mathrm{SNR}}$ is then used as the test statistic for null hypothesis testing. The motivation for using this "noise marginalized optimal statistic" is that we are marginalizing over uncertainty in the noise and signal parameters when we calculate statistical significance. One uses "phase shifts" [55], "sky scrambles" [56], or data replications to estimate the null distribution of $\overline{\mathrm{SNR}}$ and calculate a p-value for the measured $\overline{\mathrm{SNR}}$. In Ref. [1], $\overline{\mathrm{SNR}} \approx 5$, which falls at the $3.5-4\sigma$ level in the null distributions, indicating that the CURN model does not fully describe the data.

Here, we use a more conservative statistic that gives more weight to low-SNR outliers and lower weight to high-SNR outliers than the noise marginalized optimal statistic [28]. The cumulative distribution function is not linear in the SNR, and so we first calculate the p-value of the SNR calculated on each $\mathbf{\Lambda}^s$, and then average those p-values together. The resulting p-value will be less significant than the one calculated on $\overline{\mathrm{SNR}}$. Conceptually, this can be thought of as averaging over the risk of rejecting the null hypothesis by placing more weight on the most conservative noise realizations. By contrast, calculating a p-value on $\overline{\mathrm{SNR}}$ weighs high-SNR (and therefore less conservative noise realizations) equally to low-SNR noise realizations.

We compare the value of the optimal statistic on the observed data to its value on data replications from the posterior predictive distribution. We calculate a p-value on each $\mathrm{SNR}(\delta\mathbf{t}^{15\ \mathrm{yr}};\mathbf{\Lambda}^s)$ and average those p-values. This final, averaged p-value is referred to as a posterior predictive p-value or a Bayesian p-value because it is marginalized over the posterior predictive distribution for the data [27,57,58]. We can do this generically, when we do not know the distribution of the test statistic, by calculating

$$\begin{aligned}p_B = \int\int &\Theta[\mathrm{SNR}(\delta\mathbf{t}^{\mathrm{rep}},\mathbf{\Lambda}) - \mathrm{SNR}(\delta\mathbf{t}^{15\ \mathrm{yr}},\mathbf{\Lambda})] \\ &\times p(\delta\mathbf{t}^{\mathrm{rep}}|\mathbf{\Lambda},\mathrm{CURN})\,p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \mathrm{yr}})\mathrm{d}(\delta\mathbf{t}^{\mathrm{rep}})\mathrm{d}\mathbf{\Lambda}.\end{aligned} \tag{18}$$

Here Θ is the Heaviside function, and $p(\delta\mathbf{t}^{\mathrm{rep}}|\mathbf{\Lambda},\mathrm{CURN})$ could be one of the data replications described in Sec. II C, or it could be a set of "bootstrapped" data replications like sky scrambles or phase shifts.[1] If the analytic distribution for the SNR for a given choice of $\mathbf{\Lambda}^s$ is known, then we do not need to actually perform data replications, and $\int\Theta[\cdot]p(\delta\mathbf{t}^{\mathrm{rep}}|\mathbf{\Lambda},\mathrm{CURN})\mathrm{d}\delta\mathbf{t}^{\mathrm{rep}}$ is the inverse cumulative distribution function for the SNR. The Bayesian p-value reduces to

$$\begin{aligned}p_B(\delta\mathbf{t}^{15\ \mathrm{yr}}) &= \int P[\mathrm{SNR}(\delta\mathbf{t}^{\mathrm{rep}};\mathbf{\Lambda}) > \mathrm{SNR}(\delta\mathbf{t}^{15\ \mathrm{yr}};\mathbf{\Lambda})] \\ &\quad\times p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \mathrm{yr}})\mathrm{d}\mathbf{\Lambda} \\ &\approx \frac{1}{N_s}\sum_{s=1}^{N_s} P[\mathrm{SNR}(\delta\mathbf{t}^{\mathrm{rep},s};\mathbf{\Lambda}^s) > \mathrm{SNR}(\delta\mathbf{t}^{15\ \mathrm{yr}};\mathbf{\Lambda}^s)],\end{aligned} \tag{19}$$

[1]We use $\delta\mathbf{t}^{\mathrm{rep}}$ to also denote sky scrambled and phase shifted datasets, in addition to actual simulated datasets. This is somewhat poor notation. When performing sky scrambles and phase shifts the timing residuals themselves are the same as $\delta\mathbf{t}^{15\ \mathrm{yr}}$ and it is the sky position (sky scrambles) or $\mathbf{F}$ (phase shifts) that changes. Nevertheless, we use $\delta\mathbf{t}^{\mathrm{rep}}$ to refer generically to any datasets or schemes used to construct the null distribution, for simplicity.

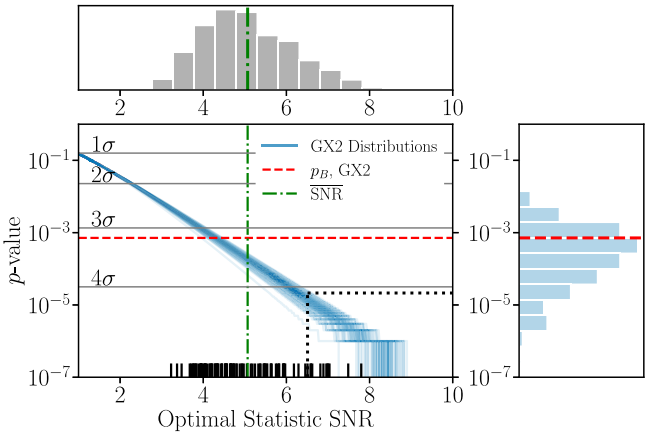

FIG. 1. Null hypothesis testing results using the GX2 distribution. The inverse CDF curve is shown in blue for 100 draws from a CURN posterior distribution. The optimal statistic SNR for each of those draws is indicated by the black lines at the bottom, and histogrammed in black in the top panel. We show p_B, averaged over p-values calculated using the GX2 distribution, with the red dashed line. The histogram of the p-value calculated for each draw is shown in the blue histogram in the right panel. The green dot-dashed line indicates $\overline{\text{SNR}}$. The gray horizontal lines correspond to different Gaussian equivalent σ levels.

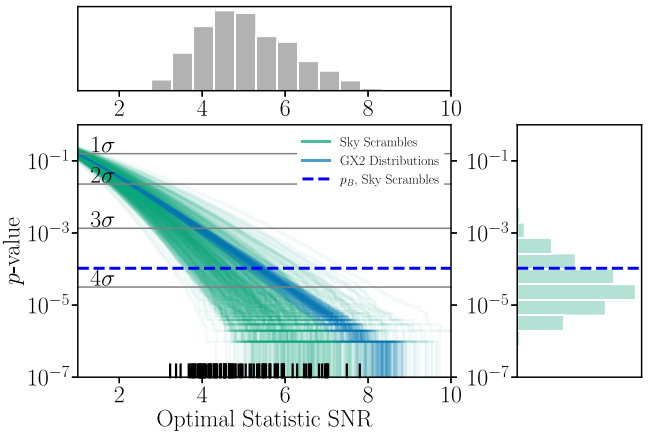

FIG. 2. Same as Fig. 1, but for sky scrambled distributions. Each green line corresponds to an inverse CDF for a given $\mathbf{\Lambda}^s$. The green histogram in the right panel corresponds to the p-value of the $\text{SNR}(\delta\mathbf{t}^{15\text{ yr}};\mathbf{\Lambda}^s)$ for each $\mathbf{\Lambda}^s$, and the blue dashed line in the middle and right panels corresponds to the average of those p-values, which is p_B for sky scrambles. In the center panel, we have included the blue GX2 inverse CDFs from Fig. 1 for reference. The sky scrambles result in a more significant p-value and p_B.

where in the second line we evaluate the integral numerically using draws from $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\text{ yr}})$. The superscript "rep" indicates that the inverse cumulative distribution function on the measured $\text{SNR}(\delta\mathbf{t}^{15\text{ yr}};\mathbf{\Lambda})$ is calculated over (theoretical or actual) data replications or sky scrambles.

The probability distribution function for the optimal statistic SNR for fixed $\mathbf{\Lambda}^s$, under the noise model, is a generalized χ^2 distribution [54] which we will refer to as GX2 moving forward. In Fig. 1 we show $\text{SNR}(\delta\mathbf{t}^{15\text{ yr}};\mathbf{\Lambda}^s)$ for 100 draws along the bottom, and each blue curve is $P[\text{SNR}^{\text{rep,s}}(\delta\mathbf{t}^{\text{rep,s}};\mathbf{\Lambda}^s) > \text{SNR}(\delta\mathbf{t}^{15\text{ yr}};\mathbf{\Lambda}^s)]$, calculated using the GX2 distribution. The dashed red line gives $p_B = 7\times10^{-4}$, which corresponds to 3.2σ significance in favor of rejecting the CURN model. By contrast, the significance of the SNR maximum-likelihood draw from $p(\mathbf{\Lambda}|\delta\mathbf{t})$ calculated using GX2 was $\approx 2\times 10^{-4}$ or 3.5σ.

In using the GX2 distribution, we assumed that the noise model is correct. Instead, we can construct replications of our dataset that break correlations due to GWs, but preserve any mismodeling in the noise that might cause large SNR values in favor of correlations. The two main methods for doing this are sky scrambles [56] and phase shifts [55]. For each $\mathbf{\Lambda}^s$, we perform 400,000 sky scrambles, where we artificially move the location of the pulsars to different positions on the sky drawn uniformly on the two-sphere and calculate a "new" Hellings-Downs curve using these new positions. Using these sky scrambles, we build a null distribution and calculate significance. We repeat this for 100 draws of $\mathbf{\Lambda}^s$ and average the inverse cumulative distribution functions (CDFs) as in Eq. (19). Under this procedure, we find $p_B = 1\times10^{-4}$, which corresponds to an equivalent Gaussian significance of 3.7σ. Using sky scrambles, Ref. [1] found $p = 5\times10^{-5}$ using the traditional procedure of building a null distribution for $\overline{\text{SNR}}$. The results are shown in Fig. 2, where the inverse CDFs for sky scrambles (green) in general fall off faster than for the GX2. However, there are some outliers resulting in p_B being larger than the p-value calculated on $\overline{\text{SNR}}$ using scrambles.

It is unclear why the inverse CDF for sky scrambles generally falls off faster than for the GX2; this is an open area of investigation [59]. In previous work, methods of generating a background distribution from sky scrambling or phase shifting use a "match statistic," in an attempt to use scrambles or shifts that are quasi-independent of one another and the Hellings-Downs curve [60,61]. Recently, in Ref. [62], the authors suggested only sky scrambles that produce correlation curves that are independent of one another should be used, where independence is achieved by insisting the match statistic disappear. In Ref. [1], the condition is that the match threshold between any one sky scramble and all others is $\lesssim 0.2$. Here we do not use a match statistic, as our goal is to estimate the probability that the pulsars would be arranged on the sky in such a way that noise fluctuations would produce Hellings-Downs correlations. To test this, we must draw the positions uniformly on the sky. How to produce reliable null distributions for datasets that preserve potentially unmodeled noise is still subject to exploration.

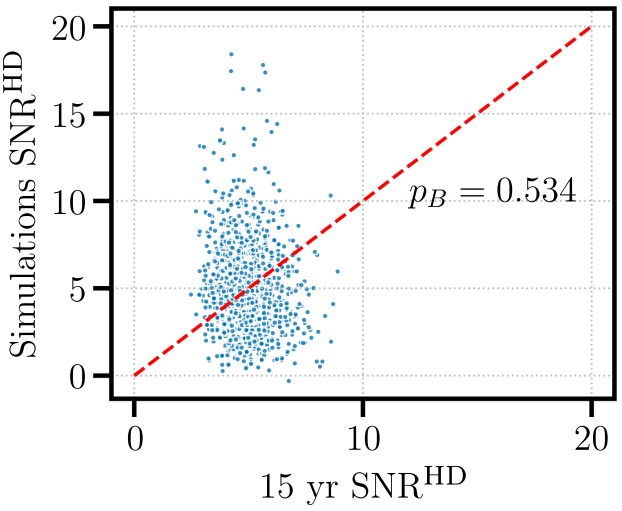

FIG. 3. We show a comparison of $\mathrm{SNR}(\delta\mathbf{t}^{15\,\mathrm{yr}}, \mathbf{\Lambda}^s)$ (x-axis) with $\mathrm{SNR}(\delta\mathbf{t}^{\mathrm{rep}}, \mathbf{\Lambda}^s)$ (y-axis). Each point corresponds to a draw from the posterior $\mathbf{\Lambda}^s$. The fact that roughly half the points fall above the line $y = x$, and $p_B = 0.53$ indicates that the measured SNR for Hellings-Downs correlations on the 15 yr data is consistent the 1000 HDPosteriorDraws replications.

B. Consistency with the HD model

In the previous subsection, we reject the null hypothesis of the CURN model at the 3.2σ level. In this section, we use the same scheme to test whether the data are consistent with Hellings-Downs correlations. Given that we only have an analytic *null* distribution for the optimal statistic, we use a Monte Carlo integral for Eq. (18) to evaluate p_B in the presence of a potential signal,

$$p_B \approx \frac{1}{N}\sum_{s=1}^{N} \Theta[\mathrm{SNR}(\delta\mathbf{t}^{\mathrm{rep}}, \mathbf{\Lambda}^s) - \mathrm{SNR}(\delta\mathbf{t}^{15\,\mathrm{yr}}, \mathbf{\Lambda}^s)], \tag{20}$$

where we average over $N = 1000$ HDPosteriorDraws replications. We show a scatter plot in Fig. 3, where the x-axis shows $\mathrm{SNR}(\delta\mathbf{t}, \mathbf{\Lambda}^s)$ and the y-axis shows $\mathrm{SNR}(\delta\mathbf{t}^{\mathrm{rep,s}}, \mathbf{\Lambda}^s)$, and p_B corresponds to the fraction of points above the line $y = x$. In this case, we find $p_B = 0.534$, indicating that the 15 yr NANOGrav data are consistent with data replications that assume Hellings-Downs correlations.

C. Additional spatial correlations

The model for a GW background assumes spatial correlations that follow the HD curve, but other spatial correlations could arise either from statistical fluctuations or due to mismodeling. Monopolar correlations could arise due to an error in the clock at each site, corresponding to a correlated offset that is common to all pulsars. Dipolar correlations could arise due to an error in the effective location and motion of the Solar System barycenter [20]. We use the multiple component optimal statistic [52], which estimates the amplitude of monopolar, dipolar, and Hellings-Downs correlations simultaneously, to test whether our estimate of these correlations in the 15 yr dataset is consistent with data replications from a pure HD model. The results and methods here follow Appendix H of Ref. [1]. The analysis is nearly identical, but with more simulations used to calculate p_B. The results and conclusion are the same as that analysis, but we include it here both for completeness and because it is strongly related to the rest of the new tests we have performed in this section.

The multiple component optimal statistic *simultaneously* produces the SNR for all three spatial correlations, $\mathrm{SNR}^{\mathrm{MONO}}_{\mathrm{MC}}$, $\mathrm{SNR}^{\mathrm{DIP}}_{\mathrm{MC}}$, and $\mathrm{SNR}^{\mathrm{HD}}_{\mathrm{MC}}$ where the superscript corresponds to the spatial correlation and the subscript indicates that we are using the multiple component optimal statistic. We again use 1000 HDPosteriorDraws data replications described in Sec. II and calculate Eq. (20) substituting $\mathrm{SNR}^{\mathrm{MONO}}_{\mathrm{MC}}$ for SNR to produce p_B^{MONO}. We produce p_B^{DIP} and p_B^{HD} for dipole and Hellings-Downs correlations defined analogously.

We show similar visualizations to the previous section in Fig. 4. We find $p_B^{\mathrm{HD}} = 0.64$, which again indicates that the HD SNR calculated on the 15 yr data is consistent

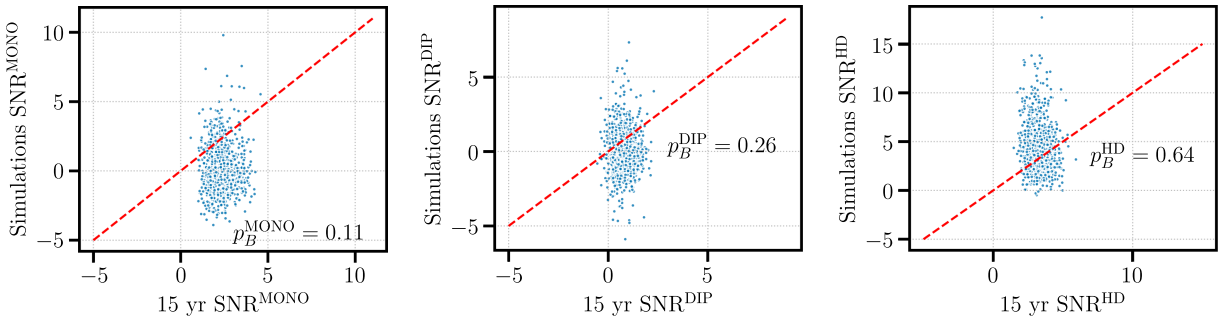

FIG. 4. We compare $\mathrm{SNR}^{\mathrm{MONO}}(\delta\mathbf{t}^{15\,\mathrm{yr}}, \mathbf{\Lambda}^s)$ (x-axis) with $\mathrm{SNR}^{\mathrm{MONO}}(\delta\mathbf{t}^{\mathrm{rep}}, \mathbf{\Lambda}^s)$ (y-axis) in the left panel (DIP and HD in center and right panels respectively), using the HDPosteriorDraws data replications. There is broad consistency between $\mathrm{SNR}^{\mathrm{HD}}$ in replications and 15 yr dataset. For the DIP and MONO models we find that the recovered SNR is consistent with data replications that include only Hellings-Downs correlations.

with what we expect from the pure HD model. Likewise, we find $p_B^{\rm DIP} = 0.26$, consistent with no dipolar correlations. Finally, we find $p_B^{\rm MONO} = 0.11$, which is largely consistent with no monopolar correlations.

IV. TESTING SPECTRUM AND CORRELATION MODELS

In this section, we assess the power-law assumption for the GW background and the intrinsic pulsar noise. We recap how to estimate the posterior distribution on the Fourier coefficients for the red noise, $\mathbf{a}$ at each frequency and for each pulsar for both the intrinsic pulsar noise and the GW background. Using the posterior on $\mathbf{a}$, we construct the posterior distribution on the intrinsic pulsar noise and GW background power spectrum in each pulsar, which can now deviate from a power-law but are subject to a power-law prior distribution. We then test for deviations in the intrinsic pulsar noise spectrum for each pulsar and in the total GW background spectrum.

A. Method

In this subsection, we summarize the methods outlined in Ref. [29]. Given $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr})$, we calculate $\mathbf{a}$ in two ways. In one method, $\mathbf{a}$ are conditioned on $\delta\mathbf{t}^{15\ \rm yr}$ and $\mathbf{\Lambda}$, which we refer to as the "inferred" coefficients (subscript "inf") because they are drawn from the inferred posterior on $\mathbf{a}$ using information from both the power-law spectrum prior and the real data. In the other method, $\mathbf{a}$ are conditioned only on $\mathbf{\Lambda}$, which we refer to as "predicted" coefficients (subscript "pre") because these are the coefficients predicted by the power-law spectrum prior. In both cases, we marginalize over $\mathbf{\Lambda}$ and $\boldsymbol{\epsilon}$. We illustrate the workflow in Fig. 5. The posteriors on the inferred and predicted parameters are formally given by

$$p_{\rm inf}(\mathbf{a}|\delta\mathbf{t}^{15\ \rm yr}) = \int d\mathbf{\Lambda} d\boldsymbol{\epsilon}\, p(\mathbf{a},\boldsymbol{\epsilon}|\mathbf{\Lambda},\delta\mathbf{t}^{15\ \rm yr},HD) \times p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr},HD), \tag{21}$$

$$p_{\rm pre}(\mathbf{a}|\delta\mathbf{t}^{15\ \rm yr}) = \int d\mathbf{\Lambda} d\boldsymbol{\epsilon}\, p(\mathbf{a},\boldsymbol{\epsilon}|\mathbf{\Lambda},HD)\, p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr},HD) = \int d\mathbf{\Lambda}\, p(\mathbf{a}|\mathbf{\Lambda},HD)\, p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr},HD). \tag{22}$$

The first term in the integrand differs between the predicted (no dependence on $\delta\mathbf{t}^{15\ \rm yr}$) and inferred posteriors (which are conditioned on $\delta\mathbf{t}^{15\ \rm yr}$). We discuss specifics of how these terms are evaluated below. The second term in the integrand is the posterior on $\mathbf{\Lambda}$, e.g., power-law amplitudes and spectral indices.

In both Eqs. (21) and (22), we evaluate the posterior with a Monte Carlo integral. We first draw a sample (labeled with "s" superscript) $\mathbf{\Lambda}^s \sim p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr})$. We then draw from the first term in the integrand. For the inferred coefficients we draw from $p(\mathbf{a},\boldsymbol{\epsilon}|\mathbf{\Lambda}^s,\delta\mathbf{t}^{15\ \rm yr})$, which is a Gaussian with mean and covariance that depend on both the prior and the data and are given by Eqs. (14) and (15). For the predicted coefficients, we draw from just the prior distribution (i.e., no dependence on data), $p(\mathbf{a},\boldsymbol{\epsilon}|\mathbf{\Lambda}^s)$, which is a zero-mean Gaussian given by Eqs. (5) and (6). More details on this scheme are discussed in Ref. [29].

For each $\mathbf{\Lambda}^s$ we split $\mathbf{a}$ into $\mathbf{a}^s_{{\rm rn},a}$ for intrinsic pulsar noise and $\mathbf{a}^s_{{\rm gw},a}$ for GWs for each pulsar a. We then reconstruct the power spectrum for the intrinsic pulsar noise in each pulsar and the GW power observed by each pulsar. Each pulsar sees a different realization of the GW background, but $\mathbf{a}^s_{{\rm gw},a}$ are drawn from the same distribution for all pulsars a. By contrast, the intrinsic pulsar noise is a different power spectrum for each pulsar, and therefore

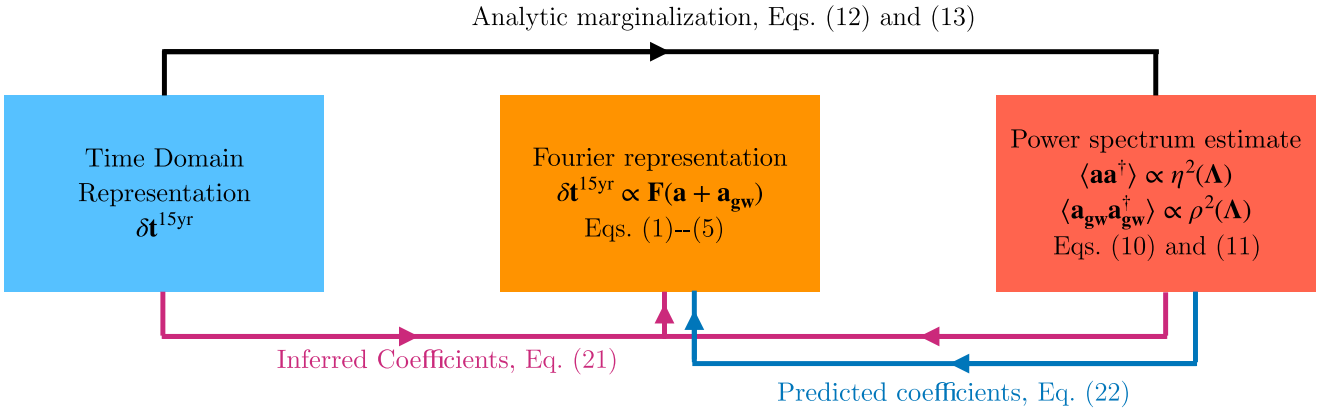

FIG. 5. Workflow for the analysis of Sec. IV based on predicted and inferred Fourier coefficients for the GW and individual-pulsar red noise spectrum. The black line corresponds to estimating $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \rm yr})$ directly after analytically marginalizing over the Fourier coefficients and directly estimating the amplitude and spectral index for the power-law GW background and intrinsic pulsar noises. The blue line corresponds to generating predicted coefficients using the power spectrum to simulate Fourier coefficients. The maroon line indicates the inferred coefficients, which use the power-law power spectrum as a prior, combined with $\delta\mathbf{t}$, to further constrain $\mathbf{a}$.

$\mathbf{a}^s_{\mathrm{rn},a}$ are drawn from a different distribution for each individual pulsar.

For the inferred coefficients, by conditioning on the data, the power spectrum will deviate from a power law if the true data-generating process differs from a power law. For the predicted spectrum, we obtain different realizations of a power spectrum that are consistent with a power law. In Sec. IV B, we discuss results for the inferred and predicted intrinsic pulsar noise and GW spectra for each pulsar and compare them to an "excess noise" analysis done in Ref. [41]. At each frequency, we use a modified version of the optimal statistic[2] to combine individual-pulsar coefficients to estimate the total GW power across the PTA in that frequency bin [63]. We present the results in Sec. IV C 1.

Finally, for each $\mathbf{\Lambda}^s$, we produce pulsar pairwise correlations and compare them to the expected Hellings-Downs curve. To do this, we draw coefficients from $p_{\mathrm{inf}}(\mathbf{a}|\delta\mathbf{t}^{15\ \mathrm{yr}}, HD)$ and $p_{\mathrm{pred}}(\mathbf{a}|\delta\mathbf{t}^{15\ \mathrm{yr}}, HD)$ and use the optimal statistic to construct pairwise correlations. In the case of the predicted parameters, this will give us an expected spread on correlation vs angular separation for a given model. For the inferred parameters, by conditioning on the data, the correlations can deviate from the model. In Sec. IV C 2, we look at reconstructions of the pairwise correlations as a function of angular separation and search for deviations from the Hellings-Downs curve.

B. Power spectra of individual pulsars

Power spectra for each pulsar in the NANOGrav array are explored in Ref. [41], it is therefore worth contrasting the two results. First, Ref. [41] simultaneously estimated the *total* red noise ($\rho_i^2 + \eta_{ai}^2$) in each frequency bin i, performing a separate analysis for each individual pulsar a. A Savage-Dickey Bayes factor was calculated to estimate the significance of the total red noise at each frequency for each pulsar. Next, both the common red noise and intrinsic pulsar noise were fixed to the maximum likelihood values estimated from a CURN analysis that assumes these spectra follow a power law. Once fixing these parameters in their model, excess noise in each frequency bin for each pulsar was searched for. No evidence for excess noise was found, the power-law model for intrinsic pulsar noise and the common red noise processes are therefore sufficient.

In this work, we instead separate intrinsic pulsar noise and GW background contributions when drawing parameters from the inferred and predicted distributions, and we produce a posterior distribution on both contributions in each frequency bin for each pulsar. This way we are testing both the intrinsic pulsar noise and GW background power-law assumptions at the same time, while constructing full posteriors on the intrinsic pulsar noise and the GW background. This is in contrast to the search for excess noise on top of a power-law common red noise and intrinsic pulsar noise. Another difference is that in this work, individual frequency bin estimates are subject to a prior that follows a power law, while Ref. [41] used a log-uniform prior on the power in each frequency bin.

In Fig. 6, we show results for two pulsars with strong intrinsic pulsar noise, B1937 + 21 and J1012 + 5307 (top two panels), and J1909-3744 which has no measurable intrinsic pulsar noise, but a contribution attributed to the GW background. The green boxes correspond to the estimates of the total red noise power from [41], which was discussed above. The blue and pink boxes correspond to the inferred intrinsic pulsar noise and GW background contributions respectively, orange boxes correspond to the predicted intrinsic pulsar noise, and yellow boxes correspond to the predicted GW background in the bottom panel. Similar plots for each pulsar are included in Supplemental Material [64].

In the top two panels, the intrinsic pulsar noise (inferred in blue boxes, predicted in orange boxes) typically agrees with the total red noise (green boxes) from Ref. [41]—indicating that the total red noise is dominated by intrinsic pulsar noise. The GW background is significantly below the intrinsic pulsar noise and the total red noise. Additionally, the orange distributions, corresponding to predicted intrinsic pulsar noise, agree with the inferred intrinsic pulsar noise. We quantify this agreement below. In the bottom panel, the total red noise agrees with the GW background, while the intrinsic pulsar noise is significantly lower. This is consistent with the GW background contributing significantly to $\delta\mathbf{t}^{15\ \mathrm{yr}}$ in this pulsar, with no intrinsic pulsar noise contribution. In a few cases the free-spectrum total red noise (green boxes) deviates further from the power law than the inferred intrinsic pulsar noise (blue boxes). This is due to the power-law prior used for the inferred intrinsic pulsar noise, which will tend to move those parameters closer to the power law.

In a few situations, e.g., the first and fifth through seventh bins for J1909-3744, the estimated total red noise (green boxes) appears to be lower than the predicted (yellow boxes) and inferred spectra (pink boxes) for the GW background. The low total noise are consistent with the predicted distributions, which follow a χ^2 with 2 degrees of freedom, and therefore have large support at those low values (despite what the box and whiskers show). The inferred distributions broadly agree with the predicted distributions, but both do look inflated compared to the total red noise. However, simply combining the GW background and intrinsic pulsar noise spectra in each bin may not reproduce the green boxes for a few reasons. First, interference between these two contributions may cancel or amplify the estimated total red noise above or below what we would expect from naively adding their contributions incoherently. Second, a log-uniform prior on the power in

[2]See Sec. IV B 2 in Ref. [29] for a discussion.

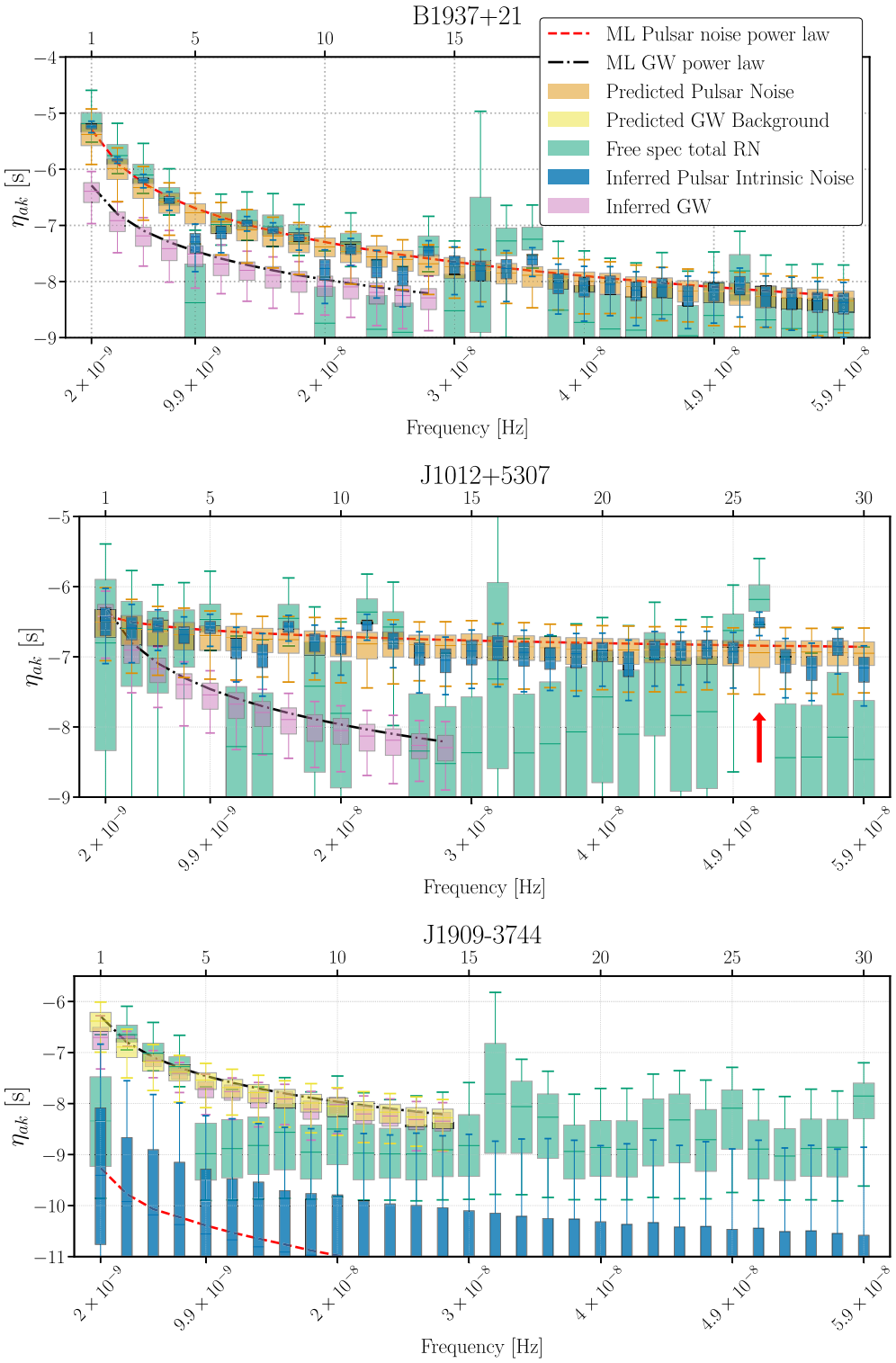

FIG. 6. We show the total red noise (green, $\rho_i^2 + \eta_{ai}^2$) [41], the inferred and predicted intrinsic pulsar noise (blue, $\eta_{ai,\mathrm{inf}}^2$; orange, $\eta_{ai,\mathrm{pred}}^2$), and inferred and predicted GW background (pink, $\rho_{ai,\mathrm{inf}}^2$; yellow, $\rho_{ai,\mathrm{pred}}^2$) for B1937 + 21 (top), J1012 + 5307 (middle), J1909-3744 (bottom). The boxes indicate the 50% credible interval, while the whiskers show the 5th and 95th percentiles. The red dashed line shows the maximum likelihood (ML) intrinsic pulsar noise power law, and the black dashed line shows the same for the GW background. To reduce clutter, we only show $\eta_{ai,\mathrm{pred}}^2$ in the top two panels and $\rho_{ai,\mathrm{pred}}^2$ in the bottom panel. The top two pulsars exhibit strong intrinsic pulsar noise, that is larger than the estimated background, because the green and blue boxes are larger than the pink ones. In the bottom panel, total red noise (green) is dominated by the GW background (pink, yellow), while the intrinsic pulsar noise (blue) is not detectable. In some frequency bins, there appears to be a lack of red noise, but this is consistent with what is expected from a power-law model. There is little evidence for excess noise, the strongest evidence being the 26th bin for J1012 + 5307 (marked with a red arrow), which has a Bayesian p-value of 0.03, which we discuss in the text.

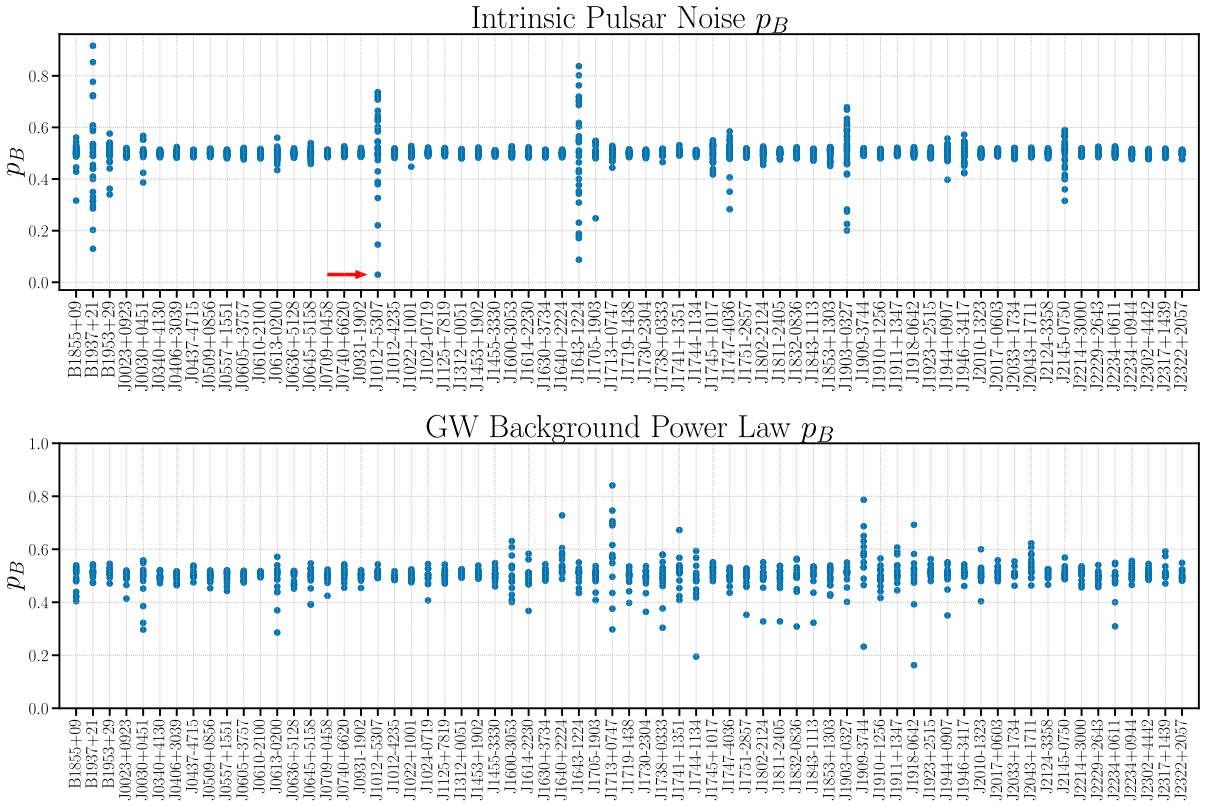

FIG. 7. For each pulsar on the horizontal axis we show p_B for all 30 (14) frequency bins in the top (bottom) panel. For most pulsars p_B is near 0.5, indicating that the inferred and predicted power spectra agree with one another, and so a power law is an appropriate model for the intrinsic pulsar noise. For several pulsars, there is a broader spread in p_B, including two of the pulsars that we show in Fig. 6.

each frequency will likely reduce the upper limit on the estimated power in that bin when no red noise detection is made when compared to the upper limit we would set using a prior informed by the power-law model.

We then quantify excess intrinsic pulsar noise in individual frequency bins for each pulsar. For each pulsar a and frequency bin i we calculate

$$p_B = \frac{1}{N}\sum_{s=1}^{N} \Theta(\eta^2_{\mathrm{inf},ai} - \eta^2_{\mathrm{pred},ai}), \tag{23}$$

where $\eta^2_{\mathrm{inf},ai}$ are drawn from $p_{\mathrm{inf}}(\mathbf{a}|\delta\mathbf{t}^{15\ \mathrm{yr}})$ and correspond to the inferred power spectrum, while $\eta^2_{\mathrm{pred},ai}$ are drawn from $p_{\mathrm{pred}}(\mathbf{a}|\delta\mathbf{t}^{15\ \mathrm{yr}})$ and correspond to the predicted power spectrum due to a power law. Each η^2 carries with it an implicit s index, which we have suppressed. We also use the HDPosteriorDraws simulations to calculate

$$p_B^{\mathrm{sim}} = \frac{1}{N}\sum_{s=1}^{N} \Theta[\eta^2_{\mathrm{inf},ai} - (\eta^2_{\mathrm{inf},ai})^{\mathrm{rep}}], \tag{24}$$

where the superscript "rep" indicates it is the inferred estimate on the power in that frequency calculated on the replicated data. Note that $\eta^2_{\mathrm{inf},ai}$ and $(\eta^2_{\mathrm{inf},ai})^{\mathrm{rep}}$ are calculated using the same $\mathbf{\Lambda}^s$. We find that these two methods produce nearly identical results, and so we report results for p_B instead of p_B^{sim}.

For intrinsic pulsar noise across all pulsars and frequencies, we find a minimum of $p_B = 0.03$, for J1012 + 5307, $f = 51$ nHz, which is the box that is visibly above the max likelihood curve in the middle panel of Fig. 6, marked with the red arrow. This is not a significant p-value, given that we are analyzing 67 pulsars and 30 frequency bins, and so we cannot conclude that this represents a deviation from a power law. This is the same conclusion as Ref. [41]. The minimum and maximum p_B for the GW background power spectrum across all pulsars are 0.16 and 0.84.

Deviations from the power-law model may not just take the form of excess noise at individual frequencies. For example, one could have a broken power law or excess noise across multiple frequencies that are not individually detectable. We do not develop a statistic to measure this here, as it requires a specific model to compare to the power-law model and there are a broad range of potential models. However, such an analysis should be done in the future.

We show a plot of p_B for intrinsic pulsar noise for each pulsar in the top panel of Fig. 7 and for the GW background

in each pulsar in the bottom panel. The horizontal axis corresponds to each pulsar, while the vertical axis corresponds to p_B; each point represents a p_B for each pulsar and each frequency bin. When the inferred and predicted power-spectrum estimates agree at a given frequency bin, we expect $p_B \approx 0.5$. We see that a few noisy pulsars show p_B values that stray away from 0.5, including the pulsars in the top two panels of Fig. 6. As stated before, no individual frequency bin shows an extreme value of p_B, e.g., $p_B < 0.01$ or $p_B > 0.99$, meaning we cannot state there are individual bins with excess noise. However, pulsars like B1937 + 21 and J1012 + 5307 do show quite a few frequency bins with p_B deviating from 0.5, meaning they may benefit from a more flexible model in the future. It is currently prohibitively computationally expensive to estimate intrinsic pulsar noise separately in each frequency bin for each pulsar when we estimate $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\text{ yr}})$. However, with future computational improvements, we may be able to do this for a limited number of pulsars, and this method provides a good starting point for choosing those pulsars.

C. Full-array results for the gravitational-wave background

1. Spectral shape

To get a full-PTA estimate of the GW background power in each frequency bin, we use a modified version of the optimal statistic [29,40,53] that estimates the Hellings-Downs-correlated GW power in each individual frequency bin. The details of this statistic are discussed in Sec. IV B 2 of Refs. [29,63].

In Fig. 8 we show results for the estimated power in each frequency bin for the inferred parameters (blue), the predicted parameters (orange), and a fully frequentist estimate that depends only on the data (green). The boxes correspond to the interquartile range of the GW power in each bin, estimated over draws from $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\text{ yr}})$, and the whiskers are the 5th and 95th percentiles. In showing these together, we compare predicted, inferred, and data-only results. There is no visible evidence for a deviation from a power law, which is indicated by the gray shaded region, which encompasses draws from $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\text{ yr}})$. The interquartile ranges for the predicted, inferred, and data-only power overlap in most frequency bins. In a few places, the data-only results appear to differ from the inferred and predicted results, e.g., 9th–11th bins, but the data are weakly informative and the inferred parameters are closer to the predicted parameters.

The per-frequency optimal statistic, used to combine $\mathbf{a}_{\text{gw}}$ across pulsars, allows for negative power in situations where the data are uninformative. Like the traditional optimal statistic, when no correlated signal is present, the distribution of the statistic is GX2 centered at zero. This is why the whiskers for several frequencies leave the bottom of the plot, and in two cases (bins 9 and 14) the data-only interquartile ranges are negative. This does not change our conclusions, as we find that our results at frequencies where we know we should see correlated GW power show such power (specifically the lowest five frequency bins). This is consistent with the HD free-spectrum results in Ref. [1].

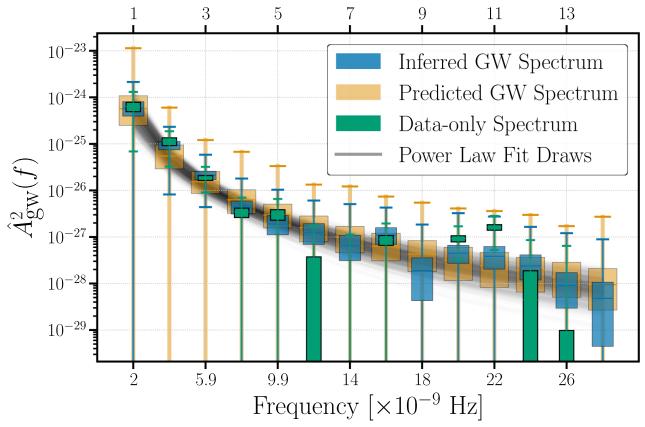

FIG. 8. We show reconstructed GW power in each frequency bin for inferred (blue) and predicted (orange) coefficients and a data-only (green) reconstruction. The boxes correspond to interquartile ranges and the whiskers are the 5th and 95th percentiles. Power-law draws from $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\text{ yr}})$ are shown in light black. The blue and orange distributions agree with one another at most frequencies. In a few places, the green boxes differ from the predicted or inferred distribution (e.g., 20 and 22 nHz), but the data are weakly informative, as the inferred and predicted distributions agree with one another in those cases. In a few frequencies the data-only distribution shows evidence for negative power, this is to be expected when the data are not informative (discussed further in the text).

We use the HDPosteriorDraws data replications to compare simulations from a power-law model to the results in Fig. 8. We find that the spectral results are consistent with a power-law model with Hellings-Downs correlations—the lowest and highest p_B comparing inferred parameters from simulations with inferred parameters on the 15 yr dataset are 0.30 and 0.72.

2. Spatial correlations

In Sec. III, we showed broad consistency between the data and the HD model, and we showed that there is no evidence for additional monopolar or dipolar correlations. Those tests compare plausible alternative analytic correlation models to the expected correlation model. In this section, we search for isolated deviations in the binned spatial correlations from the Hellings-Downs curve. We use the optimal statistic on the inferred and predicted coefficients, as well as directly on the data, and compare the inferred, predicted, and data-only binned reconstructions to search for potential deviations from the Hellings-Downs curve that are not just monopolar or dipolar.

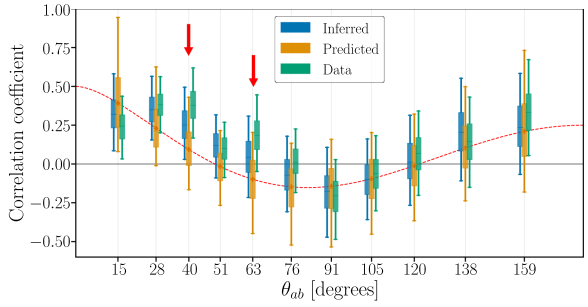

FIG. 9. We show reconstructed binned spatial correlations. The predicted (orange) show the expected spread around the Hellings-Downs curve that we might expect across many realizations. The inferred (blue) and data-only (green) recoveries broadly agree with the orange. There are two bins (third and fifth) that show some deviation between the green and orange bins. We find that these bins are not statistically significant when comparing the inferred and predicted distributions, and that changing the binning does not have a significant affect.

To construct the binned correlations, we perform an inverse-noise-weighted average over correlations for all pulsar pairs whose angular separation falls in a given angular-separation bin. An example for 11 bins of equal width is shown in Fig. 9. We compare the inferred (blue), predicted (orange), and data-only (green) estimates of the binned correlations. The spread comes from calculating the mean and variance of the correlation across pulsar pairs for a given posterior draw, sampling from a univariate Gaussian with that mean and variance, and then repeating over many draws from $p(\mathbf{\Lambda}|\delta\mathbf{t})$. The variance for the bin for a given draw includes covariance between pairs of pulsars due to the non-zero GW background [65]. The bars indicate the 5th and 95th percentiles of the resulting distribution.

We find that the data and inferred correlations are consistent with the predicted correlations. We also use the HDPosteriorDraws replications and find that none of the inferred or data-only binned correlations differ significantly from those calculated with the data replications. The two bins with the most extreme p_B[3] are the third and fifth bins, with $p_B = 0.87$, 0.88 respectively, indicating that, if we take Hellings-Downs correlations as our prior, we do not have evidence for deviations from the Hellings-Downs curve. This does not mean that we are fully consistent with Hellings-Downs correlations (as subtle changes in each bin could result in a different overall correlation pattern), but it does indicate that there are no obvious "spikes" in correlations on small angular scales. Changing the choice of binning does not change the qualitative conclusion.

[3]Because we are comparing two distributions, we consider both large and small p-values to be extreme.

V. LEAVE-ONE-OUT ANALYSES

Although individual pulsars can exhibit unique chromatic noise features, profile changes, and red noise properties, similar noise models are fit to each pulsar in the array. In this section, we seek to identify whether certain pulsars are poorly fit by these models. We perform a leave-one-out analysis, where we calculate a posterior predictive likelihood for the timing residuals in one pulsar, given the data in all other pulsars [29]. This analysis is similar to the one in Ref. [1], with a few key differences. First, we use 14 frequency bins for the analysis, and we include the negative spectral index of the GW background, γ, in the initial fit. In Ref. [1], γ is fixed to 13/3. We also use a larger number of CURN simulations to evaluate the significance of the GW background and perform a new comparison between simulated and real data on each individual pulsar.

Both the CURN and the HD models can be used to predict features in one pulsar, given a model fit to the other pulsars in the array. The CURN model, for example, can only predict the variance of the common-process-induced timing residuals, while the HD model, which includes GW-induced correlations, makes a prediction for both the variance of the timing residuals and their waveform.[4]

We compare the predictive power of these two models by calculating a pseudo-Bayes factor (PBF), which is the ratio of the posterior predictive likelihood for the HD and the CURN models. We calculate this on both a pulsar-by-pulsar basis, to identify potential pulsars that are not well predicted by the models, and also across the full pulsar timing array to construct a new detection statistic.

We denote the "left-out" pulsar with subscript a and the rest of the dataset excluding that pulsar with a subscript $-a$. The posterior predictive likelihood is

$$p(\delta\mathbf{t}_a|\delta\mathbf{t}_{-a}) = \int \mathrm{d}\mathbf{\Lambda}\mathrm{d}\mathbf{a}\mathrm{d}\boldsymbol{\epsilon}\, p(\delta\mathbf{t}_a|\mathbf{a},\boldsymbol{\epsilon},\mathbf{\Lambda})p(\mathbf{a},\boldsymbol{\epsilon},\mathbf{\Lambda}|\delta\mathbf{t}_{-a}). \tag{25}$$

As in Ref. [29], we split up the parameters and hyperparameters into separate pieces based on whether they correspond to pulsar a or pulsars $-a$, and whether they describe GW or red noise coefficients, $\mathbf{\Lambda} = [\mathbf{\Lambda}_a, \mathbf{\Lambda}_{-a}, \mathbf{\Lambda}_{gw}]$, $\mathbf{a} = [\mathbf{a}_{\mathrm{gw},a}, \mathbf{a}_{\mathrm{gw},-a}, \mathbf{a}_a, \mathbf{a}_{-a}]$. We use this new notation and evaluate Eq. (25) for the HD and the CURN models to find, see Appendix A of [29],

[4]This prediction is limited by the strength of the Hellings-Downs correlations. We cannot predict the pulsar-term fluctuations, but Earth-term predictions are informative.

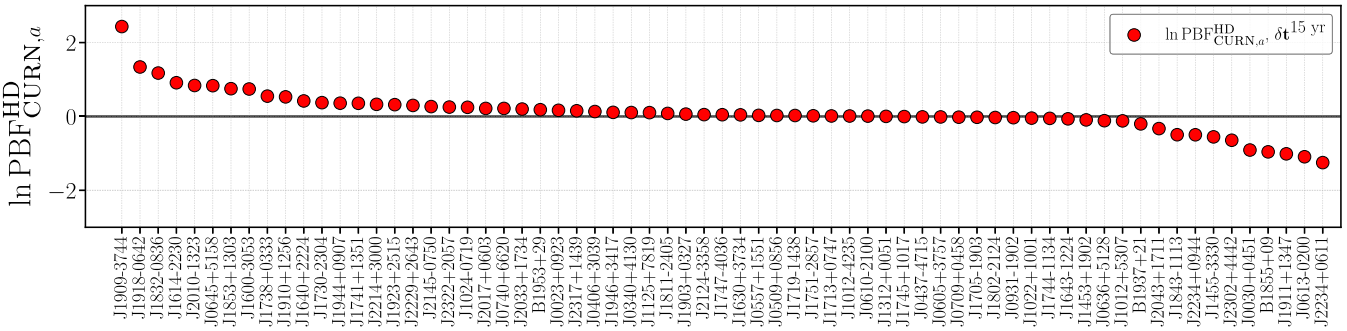

FIG. 10. We show $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ the 67 pulsars. The pulsar best predicted by the HD model is J1909-3744, whose red noise is predominantly due to a GW background. There are 43 pulsars with $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a} > 0$ and 25 with $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a} < 0$. As discussed in the text, we find the number and level of the pulsars with $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a} < 0$ to be consistent with simulations that have a GW background as strong as what we find in the data.

$$p_{\mathrm{HD}}(\delta \mathbf{t}_a|\delta \mathbf{t}_{-a}) \approx \frac{1}{N_s}\sum_{s=1}^{N_s}\int \mathrm{d}\boldsymbol{\Lambda_a}\mathrm{d}\mathbf{a}_{\mathrm{gw},a}\, p(\delta \mathbf{t}_a|\boldsymbol{\Lambda_a}, \mathbf{a}_{\mathrm{gw},a}) \times p(\mathbf{a}_{\mathrm{gw},a}|\boldsymbol{\Lambda}^s_{\boldsymbol{gw}}, \boldsymbol{\Lambda}^s_{\boldsymbol{-a}}, \delta \mathbf{t}_{-a}) p(\boldsymbol{\Lambda_a}), \tag{26}$$

$$p_{\mathrm{CURN}}(\delta \mathbf{t}_a|\delta \mathbf{t}_{-a}) \approx \frac{1}{N_s}\sum_{s=1}^{N_s}\int \mathrm{d}\boldsymbol{\Lambda_a}\, p(\delta \mathbf{t}_a|\boldsymbol{\Lambda_a}, \boldsymbol{\Lambda}^s_{\boldsymbol{gw}}) p(\boldsymbol{\Lambda_a}). \tag{27}$$

In both cases, we perform a Monte Carlo integral over the hyperparameter posterior

$$\boldsymbol{\Lambda}^s_{\boldsymbol{gw}}, \qquad \boldsymbol{\Lambda}^s_{\boldsymbol{-a}} \sim p(\boldsymbol{\Lambda}^s_{\boldsymbol{gw}}, \boldsymbol{\Lambda}^s_{\boldsymbol{-a}}|\delta \mathbf{t}_{-a}). \tag{28}$$

The main difference between Eqs. (26) and (27) is that for the HD model, the $-a$ pulsars can produce a prediction for $\mathbf{a}_{\mathrm{gw},a}$ due to the Hellings and Downs correlations, while the CURN model cannot.

The ratio of the posterior predictive likelihoods for the CURN and HD models is the PBF and it can be used to compare the two models. We first calculate the PBF pointwise across pulsars

$$\mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a} = \frac{p_{\mathrm{HD}}(\delta \mathbf{t}_a|\delta \mathbf{t}_{-a})}{p_{\mathrm{CURN}}(\delta \mathbf{t}_a|\delta \mathbf{t}_{-a})}, \tag{29}$$

and then the *total* PBF as a pointwise product

$$\mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN}} = \prod_a \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}. \tag{30}$$

A full discussion of the differences and similarities between a typical Bayes Factor and the PBF is given in Ref. [29], but we summarize a few key points here. Unlike the Bayes Factor, the PBF is not sensitive to parts of the parameter space that have no likelihood support. The PBF compares how well the models predict new data, while the Bayes factor is a summary statistic comparing how well two models fit existing data. Both statistics, however, are uncalibrated—meaning it is unclear how to interpret statistical significance as a function of the value of the statistic. In this section, similar to previous sections, we use data replications to assess the significance of the PBF.

Importantly, we can calculate the PBF on each pulsar individually and identify whether certain pulsars are "outliers" that are not well predicted by a given model. This is similar to the "dropout factor" analysis in [1,66]. In this work, we calculate a separate predictive likelihood for each model for each pulsar, while the dropout factor analysis samples an indicator variable that chooses whether to model a pulsar with the CURN or HD model. The interpretation of the results are similar to pointwise results.

Across the array, multiplying all of the "leave-out" PBFs we find $\mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN}} = 873$. This is on a similar scale to the 14 frequency Bayes factor comparing HD and CURN [1], but as with typical Bayes factors, there is no natural scale to use to "calibrate" this level of significance. In Ref. [1], several methods are used to generate a null distribution for detection statistics, including sky scrambles [56], phase shifts [55], and simulated datasets. In this work, we again resort to simulated datasets. Using 600 CURNPosteriorDraws simulations, we calculate $\mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN}}$ on each of the simulations. We find a Gaussian equivalent p-value of 3.0σ in favor of Hellings-Downs correlations on the 15 yr NANOGrav data.

We also use the HDPosteriorDraws draws to test whether this result is consistent with the HD model. We find that $\mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN}}$ falls in the 27th percentile of the HD simulations, again confirming that our results are inconsistent with the CURN model and are consistent with the HD model.

We show $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ for each pulsar in Fig. 10. There are more pulsars with $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a} > 0$ than the reverse, because the HD model is better at predicting new data than the CURN model. There are several pulsars with $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a} < 0$. We expect this in a few pulsars due to the specific realization of intrinsic pulsar noise and the pulsar term from the GW background, which we cannot predict. To understand whether the number of pulsars with

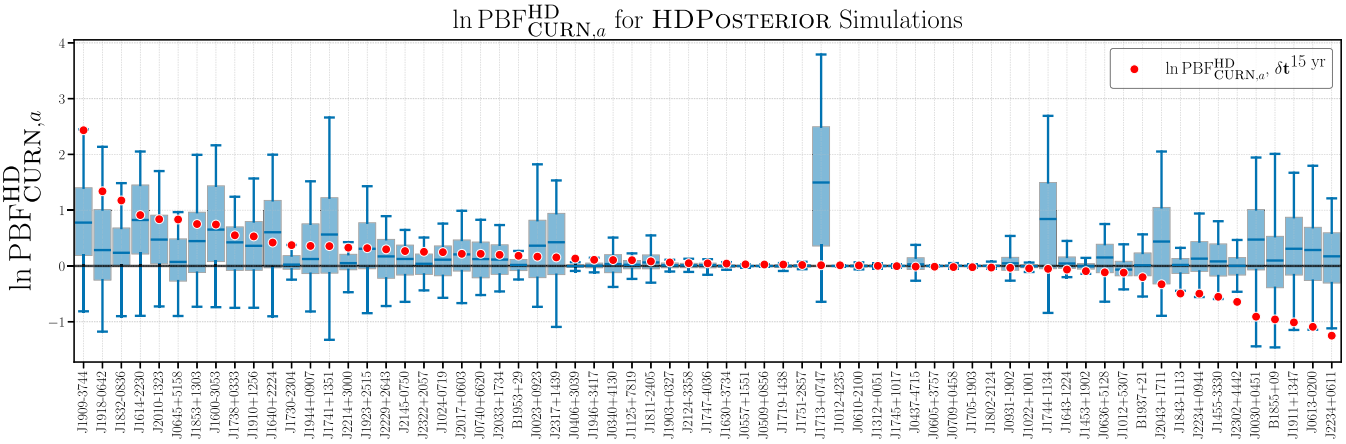

FIG. 11. We show $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ on $\delta\mathbf{t}^{15\ \mathrm{yr}}$ in red. The blue box and whisker plot show the interquartile range and 5th and 95th percentiles of the distribution of $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ for 200 of the HDPosteriorDraws simulations for each pulsar. For most pulsars the median falls above zero for these simulations, indicating that HD is the preferred model, as expected. Calculating the percentile of the red point within the blue distribution for each pulsar yields a set of percentiles that are consistent with a uniform distribution between 0 and 1, which means $\mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ on $\delta\mathbf{t}^{15\ \mathrm{yr}}$ is consistent with what we expect from a model with HD correlations and intrinsic pulsar noise consistent with $p(\mathbf{\Lambda}|\delta\mathbf{t}^{15\ \mathrm{yr}})$.

$\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a} < 0$ is expected, and whether the typical scale of those downward fluctuations is "representative" of what we would expect from a GW background, we perform simulations. We do 200 HDPosteriorDraws simulations and calculate $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ for each of those simulations to understand what the typical PBF is for the "best" and "worst" predicted pulsars if we have a GW background consistent with our posteriors.

We show the results of those simulations in Figs. 11 and 12. In Fig. 11, we plot $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ for each pulsar in red in the same order as Fig. 10. We show the distribution of $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ for each pulsar across 200 HDPosteriorDraws simulations in the blue box and whisker plots. The boxes and whiskers correspond to the 50% credible interval and the 5th and 95th percentiles respectively. The red points broadly agree with these distributions. The median simulated distribution for each pulsar falls above zero, corresponding to the HD model being preferred. Calculating the percentile of the red point ($\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ on $\delta\mathbf{t}^{15\ \mathrm{yr}}$) in each distribution yields a set of percentiles that are consistent with a uniform distribution between 0 and 1. This is what we expect if each pulsar is well predicted by all of the others. In general, the pulsars with broader distributions and larger (positive or negative) values of $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ correspond to the longest-timed and lowest-noise pulsars that have the greatest effect on the analysis.

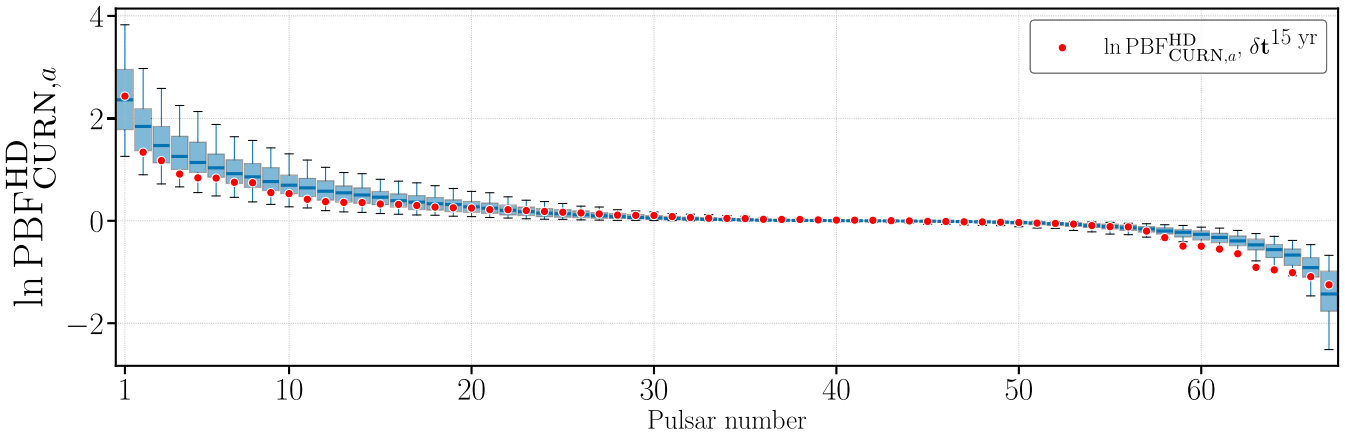

FIG. 12. We compare results on 200 HDPosteriorDraws simulations to the results on $\delta\mathbf{t}^{15\ \mathrm{yr}}$. The blue box and whisker plots correspond to the distribution of the pulsar with the ith highest value of $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ for each simulation. For example, to construct the far left box we find the maximum $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ for each simulation, and then build a distribution across simulations. For the far right box, we find the minimum $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ for each simulation and build a distribution, and so on. So the $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ going into each blue box could be for a different pulsar for each simulation.

In Fig. 12 we present results from the same simulations, but we look at the distribution of the order statistics of $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$. That is, the blue box and whisker to the furthest left correspond to the distribution of the maximum $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ across all pulsars for each simulation, so for each simulation we find the maximum $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$, and across simulations we build a distribution for that maximum. The second from left corresponds to the second largest $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$ in each simulation, the furthest to the right corresponds to the minimum value, and so on. We see that our results are consistent with the simulations from the HD model, and that in general we expect more pulsars to be better predicted by the HD model. Crucially, simulations always result in a few pulsars that are better predicted by the CURN model, i.e., negative $\ln \mathrm{PBF}^{\mathrm{HD}}_{\mathrm{CURN},a}$. Therefore, negative $\ln \mathrm{PBF}$ values are not immediately cause for concern as long as they are consistent with what we expect from simulations, which is the case here.

VI. GW BACKGROUND WAVEFORMS

The HD model is preferred to the CURN model using the optimal statistic, Bayes factors, and PBFs. In this section, we show reconstructions of the HD model compared to $\delta\mathbf{t}^{15\ \mathrm{yr}}$. We also highlight covariances between different parts of the model to better understand the relationship between the GW background, the intrinsic pulsar noise, and the timing model for each pulsar. The figures presented in this section are meant to be representative and interpreted qualitatively to illustrate the contribution of different models and the covariances between those models; similar to the waveform reconstructions shown in Refs. [42,67,68], for example. Similar figures have been shown before for noise models, e.g., [67–69], but not for the GW background model.

We first draw $\mathbf{b}^s \sim p(\mathbf{b}|\mathbf{\Lambda}^s, \delta\mathbf{t})$ using Eqs. (14) and (15), and then construct a model fit to the data by $\delta\mathbf{t}^s = \mathbf{T}\mathbf{b}^s$ for each pulsar. As in the previous section, we separate the Gaussian process coefficients for intrinsic pulsar noise $\mathbf{a}^s$, GW background $\mathbf{a}^s_{\mathrm{gw}}$, and timing model corrections $\boldsymbol{\epsilon}^s$, which we use to inspect contributions from each part of the model independently.

We show waveform reconstructions for pulsar J1909-3744 in Fig. 13. In each panel we show $\delta\mathbf{t}^{15\ \mathrm{yr}}$ (averaged over day-long timescales to reduce the number of points) and the contribution of one piece of our model. For this pulsar, the total red noise is primarily due to GWs, indicated by the lack of intrinsic pulsar noise in the top right panel and the fact that the GW background in the top left panel broadly follows $\delta\mathbf{t}^{15\ \mathrm{yr}}$ plotted in blue. In the bottom left, we show the timing model in cyan. The spin-down and spin frequency of the pulsar are covariant with the lowest frequencies of the GW background. This results in the broad uncertainties on the individual contributions from these models, but the narrow uncertainty on the

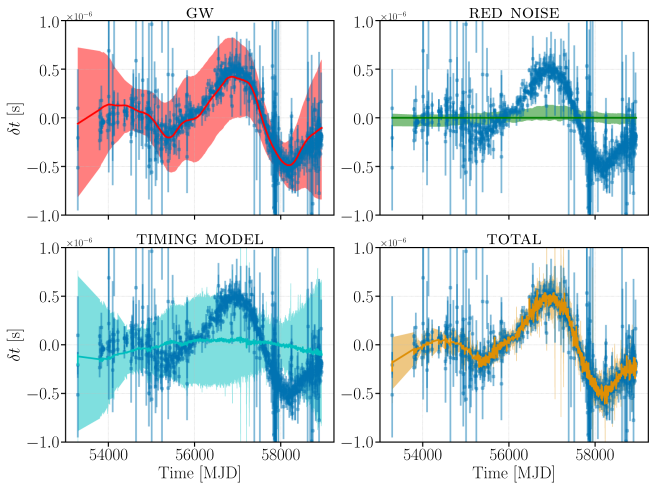

FIG. 13. Waveform reconstruction for J1909-3744 (shaded regions) and timing residuals (blue dots). The solid line corresponds to the median reconstruction, and the shaded regions correspond to the 90% credible interval. The blue points correspond to epoch-averaged residuals. In the top left is the contribution from the GW background, top right shows intrinsic pulsar noise, bottom left shows the timing model, and bottom right shows the combined total model. There is little evidence for intrinsic pulsar noise for this pulsar, and we can see that the frequency and spin-down components of the timing model (which give linear and quadratic offsets) are covariant with the lowest (and strongest) frequencies in the GW background. Regardless, the total model (bottom right) closely follows the data.

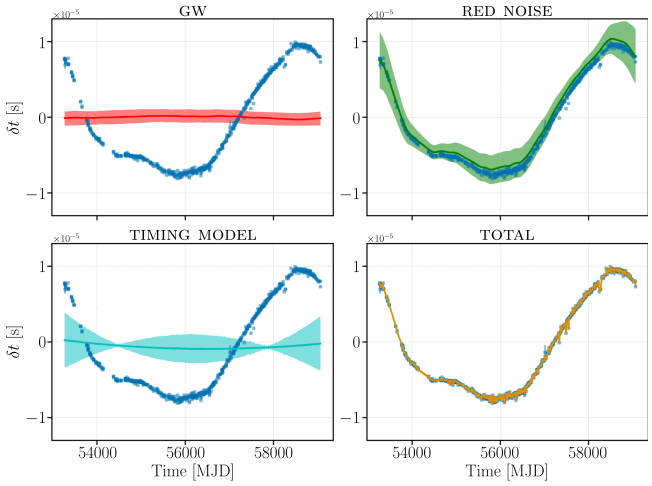

FIG. 14. Waveform reconstruction for B1937 + 21 (shaded regions) and timing residuals (blue dots). In the top left is the contribution from the GW background, top right shows intrinsic pulsar noise, bottom left shows the timing model, and bottom right shows the combined total model. There is strong intrinsic pulsar noise in this pulsar, and in this case frequency and spin-down components of the timing model (which give linear and quadratic offsets) are covariant with the lowest (and strongest) frequencies in the intrinsic pulsar noise, while the GW background is significantly smaller than the noise. Again the total model closely tracks the data (bottom right).

combined contributions of all of the models in the bottom right (orange), which tracks the timing residuals closely.

We show a similar waveform reconstruction for pulsar B1937 + 21 in Fig. 14. The intrinsic pulsar noise dominates the pulsar's total red noise, as expected based on Fig. 6. Waveform reconstructions for each pulsar are included as Supplemental Material [64]. We find that, in all cases, models represent reasonable fits to the data, which is expected based on the residual plots made with similar (single-pulsar) models in Ref. [26]. These figures are meant to illustrate the different contributions of each part of the model to the overall fit we make to each pulsar.

VII. CONCLUSIONS

The standard probabilistic model used to establish evidence for a GW background in Ref. [1] makes two assumptions motivated by theoretical expectations but also by computational convenience: that the background follows a power-law spectrum and that its interpulsar correlations conform to the Hellings-Downs pattern. Deviations from these assumptions are expected from SMBHB astronomy and astrophysics, and in certain fundamental-physics scenarios, although it is unclear whether the deviations would be measurable in current datasets.

In this paper, we examine the NANOGrav 15 yr dataset [26] within the framework of Bayesian predictive model checking [28,29], with the goal of testing the assumptions without comparison to alternative, more complex models. The *modus operandi* of Refs. [28,29] is that of using the fiducial model to simulate a population of replicated datasets from the real-data parameter posteriors and then comparing the values of multiple statistics of interest in real data and across the replications.

The optimal statistic [40,53] was used in Ref. [1] to establish the presence of interpulsar timing-residual correlations. Within the replication framework, we can account fully for the dependence of the optimal statistic on the uncertain noise parameters [28], building a Bayesian p-value that falsifies the no-correlation hypothesis at the 3.2σ level for the NANOGrav dataset. That is, we find that data replications obtained from a spatially uncorrelated model can rarely reproduce the value of the optimal statistic seen for real data. The Bayesian p-value is averaged over the noise-parameter posterior, accounting fairly for the overall risk of false rejection. If instead we build our replications from the Hellings-Downs model, we find a p-value ~ 0.5, as expected if that model is correct. We also find no anomalies when we use optimal statistic variants built to be sensitive to monopolar or dipolar correlations.

Moving on from the frequentist flavor of this optimal statistic analysis to Bayesian model comparison, we evaluate the relative predictive performance of the Hellings-Downs and spatially uncorrelated models by way of the leave-one-out cross-validation pseudo-Bayes factor [29]. We find that the Hellings-Downs model is favored at the 3σ level. That is, we find that data replications obtained from a spatially uncorrelated model can rarely reproduce the pseudo-Bayes factor seen for real data. We also verify that the binned correlation coefficients estimated from real data are consistent with the distribution expected under the Hellings-Downs hypothesis. Altogether, we find that the 15 yr NANOGrav dataset is consistent with the hypothesis of Hellings-Downs correlations, with no evidence for alternative correlation patterns.

We test the assumption that the GW background has a power-law spectrum by comparing the real-data posteriors of the spectral coefficients (i.e., the root-mean-square Fourier amplitudes at each frequency) with their distribution across replicated datasets. Although some spikes are evident in the spectral plots, we find that they are not statistically significant—they are not unlikely in the replicated population. As a by-product of this analysis, Fourier-amplitude posteriors provide a probabilistic reconstruction of the putative GW signal, as seen most strikingly in Fig. 13 for pulsar J1909-3744.

This paper details an extensive but certainly not exhaustive reanalysis of the NANOGrav 15 yr dataset. Our overall finding is that the data are consistent with a simple power-law GW background with isotropic Hellings-Downs correlations. Future more expansive and sensitive datasets will require more sophisticated data models; the framework introduced in Refs. [28,29] and exemplified here can tell us when we have reached that threshold.

ACKNOWLEDGMENTS

The authors thank Rutger van Haasteren and two anonymous referees for their constructive comments on the manuscript. The NANOGrav Collaboration receives support from National Science Foundation (NSF) Physics Frontiers Center Awards No. 1430284 and No. 2020265, the Gordon and Betty Moore Foundation, NSF AccelNet Award No. 2114721, an NSERC Discovery Grant, and CIFAR. The Arecibo Observatory is a facility of the NSF operated under cooperative agreement (AST-1744119) by the University of Central Florida (UCF) in alliance with Universidad Ana G. Méndez (UAGM) and Yang Enterprises (YEI), Inc. The Green Bank Observatory is a facility of the NSF operated under cooperative agreement by Associated Universities, Inc. The National Radio Astronomy Observatory is a facility of the NSF operated under cooperative agreement by Associated Universities, Inc. Part of this research was performed at the Jet Propulsion Laboratory, under contract with the National Aeronautics and Space Administration. Copyright 2024. L. B. acknowledges support from the National Science Foundation under Award No. AST-1909933 and from the Research Corporation for Science Advancement under Cottrell Scholar Award No. 27553. P. R. B. is supported by the Science and Technology Facilities Council, Grant No. ST/W000946/1. S. B. gratefully acknowledges the

support of a Sloan Fellowship and the support of NSF under Award No. 1815664. M. C. and S. R. T. acknowledge support from NSF AST-2007993. M. C. and N. S. P. were supported by the Vanderbilt Initiative in Data Intensive Astrophysics (VIDA) Fellowship. K. Ch., A. D. J., and M. V. acknowledge support from the Caltech and Jet Propulsion Laboratory President's and Director's Research and Development Fund. K. Ch. and A. D. J. acknowledge support from the Sloan Foundation. Support for this work was provided by the NSF through the Grote Reber Fellowship Program administered by Associated Universities, Inc./National Radio Astronomy Observatory. Pulsar research at UBC is supported by an NSERC Discovery Grant and by CIFAR. K. Cr. is supported by a UBC Four Year Fellowship (6456). M. E. D. acknowledges support from the Naval Research Laboratory by NASA under Contract No. S-15633Y. T. D. and M. T. L. are supported by an NSF Astronomy and Astrophysics Grant (AAG) Award No. 2009468. E. C. F. is supported by NASA under Award No. 80GSFC21M0002. G. E. F., S. C. S., and S. J. V. are supported by NSF Award No. PHY-2011772. K. A. G. and S. R. T. acknowledge support from an NSF CAREER Award No. 2146016. The work of N. L., X. S., and D. W. is partly supported by the George and Hannah Bolinger Memorial Fund in the College of Science at Oregon State University. N. La. acknowledges the support from Larry W. Martin and Joyce B. O'Neill Endowed Fellowship in the College of Science at Oregon State University. Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). D. R. L. and M. A. M. are supported by NSF No. 1458952. M. A. M. is supported by NSF No. 2009425. C. M. F. M. was supported in part by the National Science Foundation under Grants No. NSF PHY-1748958 and No. AST-2106552. A. Mi. is supported by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy—EXC 2121 Quantum Universe—390833306. The Dunlap Institute is funded by an endowment established by the David Dunlap family and the University of Toronto. K. D. O. was supported in part by NSF Grant No. 2207267. T. T. P. acknowledges support from the Extragalactic Astrophysics Research Group at Eötvös Loránd University, funded by the Eötvös Loránd Research Network (ELKH), which was used during the development of this research. H. A. R. is supported by NSF Partnerships for Research and Education in Physics (PREP) Award No. 2216793. S. M. R. and I. H. S. received support from a CIFAR Fellowship. Portions of this work performed at NRL. were supported by ONR 6.1 basic research funding. J. D. R. also acknowledges support from start-up funds from Texas Tech University. J. S. is supported by an NSF Astronomy and Astrophysics Postdoctoral Fellowship under Award No. AST-2202388 and acknowledges previous support by the NSF under Award 1847938. C. U. acknowledges support from a BGU Kreitman Fellowship and a Council for Higher Education and Israel Academy of Sciences and Humanities Excellence Fellowship. C. A. W. acknowledges support from Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA), the Adler Planetarium, and the Brinson Foundation through a CIERA-Adler Postdoctoral Fellowship. O. Y. is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2139292.

This paper uses a decade's worth of pulsar timing observations and is the product of the work of many people. P. M. M. helped conceive the project, wrote and developed code to perform the analysis, created all figures, and wrote and edited the text. M. V. helped conceive the project, developed code, performed parts of the analysis, ran preliminary analyses, and helped write and edit the text. K. Ch. helped conceive the project, guided direction of the analysis, and wrote and edited the text. B. L., S. V., T. D., and D. R. S. gave constructive comments that improved the manuscript, as did members of the NANOGrav Detection Working Group. G. A., A. A, A. M. A., Z. A., P. T. B., P. R. B., H. T. C., K. C., M. E. D., P. B. D., T. D., E. C. F., W. F., E. F., G. E. F., N. G. D., D. C. G., P. A. G., J. G., R. J. J., M. L. J., D. L. K., M. K., M. T. L., D. R. L., J. L., R. S. L., A. M., M. A. M., N. M., B. W. M., C. N., D. J. N., T. T. N., B. B. P. P., N. S. P., H. A. R., S. M. R., P. S. R., A. S., C. S., B. J. S. A., I. H. S., K. S., A. S., J. K. S., and H. M. W. developed timing models and ran observations for the NANOGrav 15 yr dataset.

[1] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Evidence for a gravitational-wave background, Astrophys. J. Lett. **951**, L8 (2023).

[2] D. J. Reardon *et al.*, Search for an isotropic gravitational-wave background with the parkes pulsar timing array, Astrophys. J. Lett. **951**, L6 (2023).

[3] H. Xu *et al.*, Searching for the nano-hertz stochastic gravitational wave background with the Chinese pulsar timing array data release I, Res. Astron. Astrophys. **23**, 075024 (2023).

[4] J. Antoniadis *et al.* (EPTA Collaboration), The second data release from the European pulsar timing array III. Search for

gravitational wave signals, Astron. Astrophys. **678**, A50 (2023).
[5] J. Antoniadis *et al.* (EPTA and InPTA Collaborations), The second data release from the European pulsar timing array—IV. Implications for massive black holes, dark matter, and the early Universe, Astron. Astrophys. **685**, A94 (2024).
[6] A. Afzal *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Search for signals from new physics, Astrophys. J. Lett. **951**, L11 (2023).
[7] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Constraints on supermassive black hole binaries from the gravitational-wave background, Astrophys. J. Lett. **952**, L37 (2023).
[8] G. Agazie *et al.*, The NANOGrav 15 yr data set: Looking for signs of discreteness in the gravitational-wave background, arXiv:2404.07020.
[9] E. S. Phinney, A practical theorem on gravitational wave backgrounds, arXiv:astro-ph/0108028.
[10] A. Sesana, Systematic investigation of the expected gravitational wave signal from supermassive black hole binaries in the pulsar timing band, Mon. Not. R. Astron. Soc. **433**, L1 (2013).
[11] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Search for transverse polarization modes in the gravitational-wave background, Astrophys. J. Lett. **964**, L14 (2024).
[12] C. Smarra *et al.* (European Pulsar Timing Array Collaboration), Second data release from the European pulsar timing array: Challenging the ultralight dark matter paradigm, Phys. Rev. Lett. **131**, 171001 (2023).
[13] M. V. Sazhin, Opportunities for detecting ultralong gravitational waves, Sov. Astron. **22**, 36 (1978), https://ui.adsabs.harvard.edu/abs/1978SvA....22...36S/abstract.
[14] S. Detweiler, Pulsar timing measurements and the search for gravitational waves, Astrophys. J. **234**, 1100 (1979).
[15] R. W. Hellings and G. S. Downs, Upper limits on the isotropic gravitational radiation background from pulsar timing analysis, Astrophys. J. Lett. **265**, L39 (1983).
[16] E. C. Gardiner, L. Z. Kelley, A.-M. Lemke, and A. Mitridate, Beyond the background: Gravitational-wave anisotropy and continuous waves from supermassive black hole binaries, Astrophys. J. **965**, 164 (2024).
[17] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Bayesian limits on gravitational waves from individual supermassive black hole binaries, Astrophys. J. Lett. **951**, L50 (2023).
[18] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Search for anisotropy in the gravitational-wave background, Astrophys. J. Lett. **956**, L3 (2023).
[19] J. Antoniadis *et al.* (EPTA Collaboration), The second data release from the European pulsar timing array V. Search for continuous gravitational wave signals, Astron. Astrophys. **690**, A118 (2024).
[20] C. Tiburzi, G. Hobbs, M. Kerr, W. A. Coles, S. Dai, R. N. Manchester, A. Possenti, R. M. Shannon, and X. P. You, A study of spatial correlations in pulsar timing array data, Mon. Not. R. Astron. Soc. **455**, 4339 (2016).
[21] M. Vallisneri, S. R. Taylor, J. Simon, W. M. Folkner, R. S. Park, C. Cutler, J. A. Ellis, T. J. W. Lazio, S. J. Vigeland, K. Aggarwal *et al.*, Modeling the uncertainties of solar system ephemerides for robust gravitational-wave searches with pulsar-timing arrays, Astrophys. J. **893**, 112 (2020).
[22] B. Bécsy, N. J. Cornish, and L. Z. Kelley, Exploring realistic nanohertz gravitational-wave backgrounds, Astrophys. J. **941**, 119 (2022).
[23] B. Bécsy *et al.*, How to detect an astrophysical nanohertz gravitational wave background, Astrophys. J. **959**, 9 (2023).
[24] S. Valtolina, G. Shaifullah, A. Samajdar, and A. Sesana, Testing strengths, limitations, and biases of current pulsar timing arrays' detection analyses on realistic data, Astron. Astrophys. **683**, A201 (2024).
[25] G. Agazie *et al.* (International Pulsar Timing Array Collaboration), Comparing recent pulsar timing array results on the nanohertz stochastic gravitational-wave background, Astrophys. J. **966**, 105 (2024).
[26] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Observations and timing of 68 millisecond pulsars, Astrophys. J. Lett. **951**, L9 (2023).
[27] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis* (CRC Press, Boca Raton, USA, 2013).
[28] M. Vallisneri, P. M. Meyers, K. Chatziioannou, and A. J. K. Chua, Posterior predictive checking for gravitational-wave detection with pulsar timing arrays. I. The optimal statistic, Phys. Rev. D **108**, 123007 (2023).
[29] P. M. Meyers, K. Chatziioannou, M. Vallisneri, and A. J. K. Chua, Posterior predictive checking for gravitational-wave detection with pulsar timing arrays. II. Posterior predictive distributions and pseudo-Bayes factors, Phys. Rev. D **108**, 123008 (2023).
[30] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Binary black hole population properties inferred from the first and second observing runs of Advanced LIGO and Advanced Virgo, Astrophys. J. Lett. **882**, L24 (2019).
[31] M. Fishbach, W. M. Farr, and D. E. Holz, The most massive binary black hole detections and the identification of population outliers, Astrophys. J. Lett. **891**, L31 (2020).
[32] M. Fishbach and D. E. Holz, Minding the gap: GW190521 as a straddling binary, Astrophys. J. Lett. **904**, L26 (2020).
[33] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Population properties of compact objects from the second LIGO-Virgo gravitational-wave transient catalog, Astrophys. J. Lett. **913**, L7 (2021).
[34] R. Abbott *et al.* (KAGRA, VIRGO, and LIGO Scientific Collaborations), Population of merging compact binaries inferred using gravitational waves through GWTC-3, Phys. Rev. X **13**, 011048 (2023).
[35] R. Essick, A. Farah, S. Galaudage, C. Talbot, M. Fishbach, E. Thrane, and D. E. Holz, Probing extremal gravitational-wave events with coarse-grained likelihoods, Astrophys. J. **926**, 34 (2022).
[36] T. A. Callister, S. J. Miller, K. Chatziioannou, and W. M. Farr, No evidence that the majority of black holes in binaries have zero spin, Astrophys. J. Lett. **937**, L13 (2022).
[37] E. Payne and E. Thrane, Model exploration in gravitational-wave astronomy with the maximum population likelihood, Phys. Rev. Res. **5**, 023013 (2023).
[38] S. J. Miller, Z. Ko, T. Callister, and K. Chatziioannou, Gravitational waves carry information beyond effective spin

parameters but it is hard to extract, Phys. Rev. D **109**, 104036 (2024).
[39] M. Anholm, S. Ballmer, J. D. E. Creighton, L. R. Price, and X. Siemens, Optimal strategies for gravitational wave stochastic background searches in pulsar timing data, Phys. Rev. D **79**, 084030 (2009).
[40] S. J. Chamberlin, J. D. E. Creighton, X. Siemens, P. Demorest, J. Ellis, L. R. Price, and J. D. Romano, Time-domain implementation of the optimal cross-correlation statistic for stochastic gravitational-wave background searches in pulsar timing data, Phys. Rev. D **91**, 044048 (2015).
[41] G. Agazie *et al.* (NANOGrav Collaboration), The NANOGrav 15 yr data set: Detector characterization and noise budget, Astrophys. J. Lett. **951**, L10 (2023).
[42] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116**, 061102 (2016).
[43] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GWTC-1: A gravitational-wave transient catalog of compact binary mergers observed by LIGO and Virgo during the first and second observing runs, Phys. Rev. X **9**, 031040 (2019).
[44] M. L. Jones *et al.*, The NANOGrav nine-year data set: Measurement and analysis of variations in dispersion measures, Astrophys. J. **841**, 125 (2017).
[45] S. R. Taylor, The nanohertz gravitational wave astronomer, arXiv:2105.13270.
[46] R. van Haasteren and Y. Levin, Understanding and analysing time-correlated stochastic signals in pulsar timing, Mon. Not. R. Astron. Soc. **428**, 1147 (2013).
[47] Z. Arzoumanian *et al.* (NANOGrav Collaboration), The NANOGrav nine-year data set: Limits on the isotropic stochastic gravitational wave background, Astrophys. J. **821**, 13 (2016).
[48] L. Lentati, P. Alexander, M. P. Hobson, S. Taylor, J. Gair, S. T. Balan, and R. van Haasteren, Hyper-efficient model-independent Bayesian method for the analysis of pulsar timing data, Phys. Rev. D **87**, 104021 (2013).
[49] Z. Arzoumanian *et al.* (NANOGrav Collaboration), The NANOGrav nine-year data set: Observations, arrival time measurements, and analysis of 37 millisecond pulsars, Astrophys. J. **813**, 65 (2015).
[50] M. Falxa *et al.*, Modeling nonstationary noise in pulsar timing array data analysis, Phys. Rev. D **109**, 123010 (2024).
[51] A. D. Johnson *et al.* (NANOGrav Collaboration), NANOGrav 15-year gravitational-wave background methods, Phys. Rev. D **109**, 103012 (2024).
[52] S. C. Sardesai, S. J. Vigeland, K. A. Gersbach, and S. R. Taylor, Generalized optimal statistic for characterizing multiple correlated signals in pulsar timing arrays, Phys. Rev. D **108**, 124081 (2023).
[53] S. J. Vigeland, K. Islo, S. R. Taylor, and J. A. Ellis, Noise-marginalized optimal statistic: A robust hybrid frequentist-Bayesian statistic for the stochastic gravitational-wave background in pulsar timing arrays, Phys. Rev. D **98**, 044003 (2018).
[54] J. S. Hazboun, P. M. Meyers, J. D. Romano, X. Siemens, and A. M. Archibald, Analytic distribution of the optimal cross-correlation statistic for stochastic gravitational-wave-background searches using pulsar timing arrays, Phys. Rev. D **108**, 104050 (2023).
[55] S. R. Taylor, L. Lentati, S. Babak, P. Brem, J. R. Gair, A. Sesana, and A. Vecchio, All correlations must die: Assessing the significance of a stochastic gravitational-wave background in pulsar-timing arrays, Phys. Rev. D **95**, 042002 (2017).
[56] N. J. Cornish and L. Sampson, Towards robust gravitational wave detection with pulsar timing arrays, Phys. Rev. D **93**, 104047 (2016).
[57] A. Gelman, X.-L. Meng, and H. Stern, Posterior predictive assessment of model fitness via realized discrepancies, Stat. Sin. **6**, 733 (1996), https://www.jstor.org/stable/24306036.
[58] A. Gelman, Two simple examples for understanding posterior p-values whose distributions are far from uniform, Electron. J. Stat. **7**, 2595 (2013).
[59] V. Di Marco, A. Zic, R. M. Shannon, and E. Thrane, Systematic errors in searches for nanohertz gravitational waves, Mon. Not. R. Astron. Soc. **532**, 4026 (2024).
[60] L. Sampson, N. J. Cornish, and S. T. McWilliams, Constraining the solution to the last parsec problem with pulsar timing, Phys. Rev. D **91**, 084055 (2015).
[61] S. R. Taylor, J. Simon, and L. Sampson, Constraints on the dynamical environments of supermassive black-hole binaries using pulsar-timing arrays, Phys. Rev. Lett. **118**, 181102 (2017).
[62] V. Di Marco, A. Zic, M. T. Miles, D. J. Reardon, E. Thrane, and R. M. Shannon, Toward robust detections of nanohertz gravitational waves, Astrophys. J. **956**, 14 (2023).
[63] K. A. Gersbach, S. R. Taylor, P. M. Meyers, and J. D. Romano, Spatial and spectral characterization of the gravitational-wave background with the PTA optimal statistic, arXiv:2406.11954.
[64] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevD.111.042011 for total noise, predicted and inferred intrinsic pulsar noise, and predicted and inferred GW background contributions for each pulsar. These are analogous to Fig. 6. We also show waveform reconstructions for each individual pulsar. These are analogous to Figs. 13 and 14.
[65] B. Allen and J. D. Romano, Hellings and downs correlation of an arbitrary set of pulsars, Phys. Rev. D **108**, 043026 (2023).
[66] Z. Arzoumanian *et al.* (NANOGrav Collaboration), The NANOGrav 12.5 yr data set: Search for an isotropic stochastic gravitational-wave background, Astrophys. J. Lett. **905**, L34 (2020).
[67] L. Lentati *et al.*, From spin noise to systematics: Stochastic processes in the first international pulsar timing array data release, Mon. Not. R. Astron. Soc. **458**, 2161 (2016).
[68] B. Goncharov *et al.*, Identifying and mitigating noise sources in precision pulsar timing data sets, Mon. Not. R. Astron. Soc. **502**, 478 (2021).
[69] B. Larsen *et al.*, The NANOGrav 15 yr data set: Chromatic Gaussian process noise models for six pulsars, Astrophys. J. **972**, 49 (2024).