The Astrophysical Journal Letters, 905:L6 (9pp), 2020 December 10 https://doi.org/10.3847/2041-8213/abca92

Model Dependence of Bayesian Gravitational-wave Background Statistics for Pulsar Timing Arrays

Jeffrey S. Hazboun[1], Joseph Simon[2,3], Xavier Siemens[4], and Joseph D. Romano[5]

[1] Physical Sciences Division, University of Washington Bothell, 18115 Campus Way NE, Bothell, WA 98011, USA; hazboun@uw.edu
[2] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA
[3] Department of Astrophysical and Planetary Sciences, University of Colorado, Boulder, CO 80309, USA
[4] Department of Physics, Oregon State University, Corvallis, OR 97331, USA
[5] Department of Physics and Astronomy, Texas Tech University, Lubbock, TX 79409-1051, USA

Received 2020 September 10; revised 2020 November 12; accepted 2020 November 15; published 2020 December 9

Abstract

Pulsar timing array (PTA) searches for a gravitational-wave background (GWB) typically include time-correlated "red" noise models intrinsic to each pulsar. Using a simple simulated PTA data set with an injected GWB signal we show that the details of the red noise models used, including the choice of amplitude priors and even which pulsars *have* red noise, have a striking impact on the GWB statistics, including both upper limits and estimates of the GWB amplitude. We find that the standard use of uniform priors on the red noise amplitude leads to 95% upper limits, as calculated from one-sided Bayesian credible intervals, that are less than the injected GWB amplitude 50% of the time. In addition, amplitude estimates of the GWB are systematically *lower* than the injected value by 10%–40%, depending on which models and priors are chosen for the intrinsic red noise. We tally the effects of model and prior choice and demonstrate how a "dropout" model, which allows flexible use of red noise models in a Bayesian approach, can improve GWB estimates throughout.

Unified Astronomy Thesaurus concepts: Gravitational waves (678); Gravitational wave astronomy (675); Gravitational wave detectors (676); Bayesian statistics (1900); Prior distribution (1927); Posterior distribution (1926); Astrostatistics strategies (1885); Pulsar timing method (1305); Red noise (1956); Time series analysis (1916); Millisecond pulsars (1062)

1. Introduction

Pulsar timing arrays (PTAs) are sensitive to gravitational waves (GWs) in the nanohertz frequency band (Sazhin 1978; Detweiler 1979; Foster & Backer 1990). The most promising GW sources in that band are supermassive binary black holes (SMBBHs) that are formed via mergers of massive galaxies (Rosado et al. 2015). Orbiting SMBBHs can produce a stochastic gravitational-wave background (GWB), individual periodic signals, and transient GW bursts (Burke-Spolaor et al. 2018). The GWB from SMBBHs manifests in pulsar timing data as a stochastic signal that is correlated both temporally and spatially between pulsars (Hellings & Downs 1983; Phinney 2001; Jaffe & Backer 2003). The spatial correlations are defined by the overlap reduction function known as the Hellings–Downs curve (Hellings & Downs 1983), and the temporal correlations follow a steep power-law spectrum (Phinney 2001; Jaffe & Backer 2003)—i.e., the GWB is a spatially correlated "red" noise process in pulsar timing data. The spatial correlations are a direct consequence of general relativity, originating from the quadrupolar nature of GWs.

However, there are also a number of potential non-GW sources of temporal correlations (red noise processes) in pulsar timing data. Intrinsic spin noise exists in many canonical pulsars (Cordes & Downs 1985) and at a lower level in millisecond pulsars (Cordes 2013). Clock errors and solar system ephemeris errors can manifest as spatially correlated sources of red noise (Champion et al. 2010; Tiburzi et al. 2016). Finally, unmodeled trends in radio-dependent propagation delays, e.g., dispersion and scattering due to the interstellar medium, will also produce red noise in pulsar timing data (Cordes & Shannon 2010). Since the stochastic GWB is modeled as a red noise process with a steep spectral index, unmodeled or mismodeled noise in pulsars can have an impact on GWB detection statistics and parameter estimation (Hazboun et al. 2020). It is therefore imperative to accurately model the noise in individual pulsars so that it does not contaminate the GWB signal.

As PTA data sets have matured and reached astrophysically interesting levels of sensitivity, 95% upper limits (ULs) on the amplitude of the GWB, $A_{\rm GWB}^{95\%}$, have been the flagship statistic quoted by PTA collaborations (Shannon et al. 2015; Arzoumanian et al. 2016, 2018a; Lentati et al. 2016). These ULs have been used to constrain model parameters for SMBBH populations (Arzoumanian et al. 2016; Simon & Burke-Spolaor 2016), such as the M–$M_{\rm bulge}$ relationship between SMBHs to their host galaxies, and to calculate the evidence for various SMBBH population models (Shannon et al. 2015; Sesana et al. 2018). The standard $A_{\rm GWB}^{95\%}$ quoted is the one-sided credible interval of a Bayesian analysis and is therefore subject to the choice of the signal+noise models, including the choice of prior probability distributions for the parameters associated with these models.

1.1. Statement of the Problem

In this Letter, we systematically investigate how the choice of different signal+noise models, including the choice of priors, affects the estimates and ULs of the GWB returned by PTA analyses. We shall see that of utmost importance is the choice of noise model for the individual pulsars,[6] as steep red

[6] As another pertinent example of the interplay of noise models and GWs, the development of the ECORR noise parameter (Arzoumanian et al. 2014) was in response to spurious single-source GW detections at frequencies higher than 1 $\rm yr^{-1}$ caused by noise correlated across frequencies on intraday timescales.

noise in one or several pulsars can masquerade as red noise in the GWB, potentially leading to a bias in the conditional median estimate of the amplitude of the background, $A_{\rm GWB}$, or its UL, $A_{\rm GWB}^{95\%}$. Not surprisingly, a marginally significant GWB can be absorbed by red noise models for individual pulsars across the PTA, reducing the estimated amplitude of the GWB. It is also possible for models that do not accurately model the pulsar noise to lead to spurious increases in the estimated GWB amplitude. As shown in Hazboun et al. (2020), the significance of the detection of a GWB signal is strongly affected by which red noise models are chosen for each pulsar.

Over the years the above considerations regarding red noise led the PTA community to adopt a conservative approach that includes a red noise model for *every* pulsar in the array (Yardley et al. 2011; Demorest et al. 2013; Lentati et al. 2015; Arzoumanian et al. 2016, 2018a; Lentati et al. 2016). An obvious alternative to this approach is to not model intrinsic pulsar red noise at all, which would effectively attribute any observed red noise to the GWB. This approach is arguably *more* conservative than the standard approach, since it results in the largest ULs on the GWB amplitude. A third choice is to allow a Bayesian analysis to *choose* which pulsars should include intrinsic red noise models—i.e., letting the data decide whether a red noise model for a particular pulsar is preferred to a white-noise-only model in a search for the GWB. There are a number of possibilities for implementing this more flexible pulsar noise model, including transdimensional models (Ellis & Cornish 2016), hierarchical modeling (Gelman & Hill 2007), and product space methods (Carlin & Chib 1995; Godsill 2001; Hee et al. 2015; Taylor et al. 2020). But for the analyses that we perform in this Letter, we adopt a *dropout* method (Aggarwal et al. 2019; discussed in more detail in Section 2.1) to decide which pulsars should be assigned red noise. As we shall see below, a pulsar red noise model that utilizes the dropout method, and allows the data to decide whether or not to include intrinsic red noise in each pulsar, gives results that most robustly and accurately return the injected $A_{\rm GWB}$.

1.2. Outline of the Letter

To compare the effects of different pulsar noise models on the statistics of the GWB, we analyze 400 realizations of simulated PTA data consisting of a GWB signal injected into white timing noise (WN). Details of the simulations, signal +noise models, and data analysis methods used are discussed in detail in Section 2. We allow for different priors for both the amplitude of the GWB and the red noise of the individual pulsars, as well as whether or not a red noise model for a particular pulsar should be included, i.e., the dropout method. The results are described in Section 3 for both GWB parameter estimation (Section 3.1) and 95% UL calculations (Section 3.2). It turns out that the choice of the individual red noise models has a surprisingly strong effect, especially in the case of UL analyses. We consider more realistic simulations in Section 4, where we inject red timing noise for a handful of pulsars, and show that the dropout method can also handle this more realistic scenario without any problems. Finally, in Section 5, we reanalyze the NANOGrav 11 yr data set using the dropout method, obtaining a revised 95% UL, $A_{\rm GWB}^{95\%} = 3.0 \times 10^{-15}$. This is more than twice as large as the value reported in Arzoumanian et al. (2018a).

2. Simulated Data and Signal+Noise Models

To investigate the effect of different models and priors for the intrinsic red noise in pulsars, we performed a number of simulations and analyses. The software `libstempo` (Vallisneri 2020) is used to simulate 400 realizations of a GWB with amplitude $A_{\rm GWB} = 1.4 \times 10^{-15}$ in WN at a level of 1 μs for a simple PTA data set based on the pulsars in the IPTA's second mock data challenge (Hazboun et al. 2018). The WN is simulated using the time of arrival (TOA) errors, identical in amplitude for all pulsars. (Pulse TOAs are the fiducial data used in PTA analyses.) The amplitude of the WN is treated as a known quantity for all of the analyses. We then compare the results of the different models and priors to the results of a model that incorporates only what we know to be in the simulated data: a red GWB signal plus WN.

The construction of the likelihood and analysis methods match exactly those used in recent PTA data analysis work such as Arzoumanian et al. (2018b) and developed over the past decade in the literature (Ellis et al. 2013; Taylor et al. 2013, 2017; Arzoumanian et al. 2014, 2016; van Haasteren & Vallisneri 2014). Therefore, we do not describe the analysis methods in detail here except to note a few features connected to the signal+noise models relevant to this work.

The GWB is modeled as a Gaussian process (Williams & Rasmussen 2006) in the Fourier domain (van Haasteren & Vallisneri 2014; Lentati et al. 2016) with a power spectral density given by a power law. The main results quoted from searches for the GWB assume a *fixed* spectral index $\gamma = 13/3$ for the induced timing residuals:

$$P_g(f) = \frac{A_{\rm GWB}^2}{12\pi^2}\left(\frac{f}{f_{\rm yr}}\right)^{-\gamma} {f_{\rm yr}}^{-3}, \tag{1}$$

where $f_{\rm yr} \equiv 1\,{\rm yr}^{-1}$ is a reference frequency. The choice $\gamma = 13/3$ corresponds to a spectral index of $-2/3$ for the characteristic strain of the GWB, appropriate for inspiraling binaries (Phinney 2001; Jaffe & Backer 2003). Typical analyses, in addition to the GWB, include a separate red noise model for each pulsar, parameterized in the same way as the GWB,

$$P_{\rm RN}(f) = \frac{A_{\rm RN}^2}{12\pi^2}\left(\frac{f}{f_{\rm yr}}\right)^{-\gamma_{\rm RN}} {f_{\rm yr}}^{-3}, \tag{2}$$

where *both* the spectral index $\gamma_{\rm RN}$ and red noise amplitude $A_{\rm RN}$ are allowed to vary for each pulsar. This adds $2N_{\rm pulsars}$ parameters to the search. The prior on the spectral index is taken to be uniform from 0 to 7. This covers the range from white noise ($\gamma = 0$) to the the steepest power spectral density for which the quadratic spin down removes dependence on any lower cutoff frequency in that power spectral density (Blandford et al. 1984; van Haasteren & Levin 2013). In principle, the prior on the spectral index could also be chosen differently (see, e.g., Callister et al. 2017), but we do not investigate the effects of those choices here. However, we do consider the effect of different priors on the *amplitudes* of both the GWB and pulsar red noise, $A_{\rm GWB}$ and $A_{\rm RN}$. We use either *uniform* or *log-uniform* (uniform in log space) probability distributions for these individual amplitudes, defined over the range of values 10^{-18} to 10^{-14}. Table 1 lists the various signal

Table 1
Different Signal+noise Models and Prior Probability Distributions Used in Our Analyses

Model	Signal Prior	Noise Priors
$\mathrm{GWB_{CRN}}$-only	$\pi(A_{\mathrm{GWB}}) = \mathrm{logunif}(10^{-18}, 10^{-12})$	⋯
$\mathrm{GWB_{CRN}}$-only	$\pi(A_{\mathrm{GWB}}) = \mathrm{unif}(10^{-18}, 10^{-12})$	⋯
$\mathrm{GWB_{CRN}+RN}$	$\pi(A_{\mathrm{GWB}}) = \mathrm{logunif}(10^{-18}, 10^{-12})$	$\pi(A_{\mathrm{RN}}) = \mathrm{logunif}(10^{-20}, 10^{-11})$ $\pi(\gamma_{\mathrm{RN}}) = \mathrm{unif}(0, 7)$
$\mathrm{GWB_{CRN}+RN}$	$\pi(A_{\mathrm{GWB}}) = \mathrm{unif}(10^{-18}, 10^{-12})$	$\pi(A_{\mathrm{RN}}) = \mathrm{unif}(10^{-20}, 10^{-11})$ $\pi(\gamma_{\mathrm{RN}}) = \mathrm{unif}(0, 7)$
$\mathrm{GWB_{CRN}+RN_{DO}}$	$\pi(A_{\mathrm{GWB}}) = \mathrm{logunif}(10^{-18}, 10^{-12})$	$\pi(A_{\mathrm{RN}}) = \mathrm{logunif}(10^{-20}, 10^{-11})$ $\pi(\gamma_{\mathrm{RN}}) = \mathrm{unif}(0, 7)$
$\mathrm{GWB_{CRN}+RN_{DO}}$	$\pi(A_{\mathrm{GWB}}) = \mathrm{unif}(10^{-18}, 10^{-12})$	$\pi(A_{\mathrm{RN}}) = \mathrm{unif}(10^{-20}, 10^{-11})$ $\pi(\gamma_{\mathrm{RN}}) = \mathrm{unif}(0, 7)$
$\mathrm{GWB_{HD}+RN}$	$\pi(A_{\mathrm{GWB}}) = \mathrm{logunif}(10^{-18}, 10^{-12})$	$\pi(A_{\mathrm{RN}}) = \mathrm{logunif}(10^{-20}, 10^{-11})$ $\pi(\gamma_{\mathrm{RN}}) = \mathrm{unif}(0, 7)$

Note. RN stands for the pulsar red noise model, DO stands for the red noise dropout model, CRN stands for a signal model without spatial correlations, and HD stands for a signal model that includes spatial Hellings–Downs correlations. For the pulsar red noise models, there are different amplitude and spectral index parameters for *each* pulsar, for a total of $2N_{\mathrm{pulsars}}$ parameters.

+noise models and prior probability distributions used in our analyses. Note that in most cases the GW signal does not include Hellings–Downs spatial correlations as these considerably decrease the computational efficiency. The alternative signal model searches for a common red noise process (Arzoumanian et al. 2018a) are subscripted in Table 1 with CRN.

All searches use the software ENTERPRISE (Ellis et al. 2019) and enterprise_extensions for modeling the PTA data likelihood, the GWB, and the various signal+noise models. We used the Parallel Tempering Markov Chain Monte Carlo (MCMC) sampler PTMCMCSampler (Ellis & van Haasteren 2017) for sampling the likelihood.

2.1. Pulsar Red Noise Dropout Method

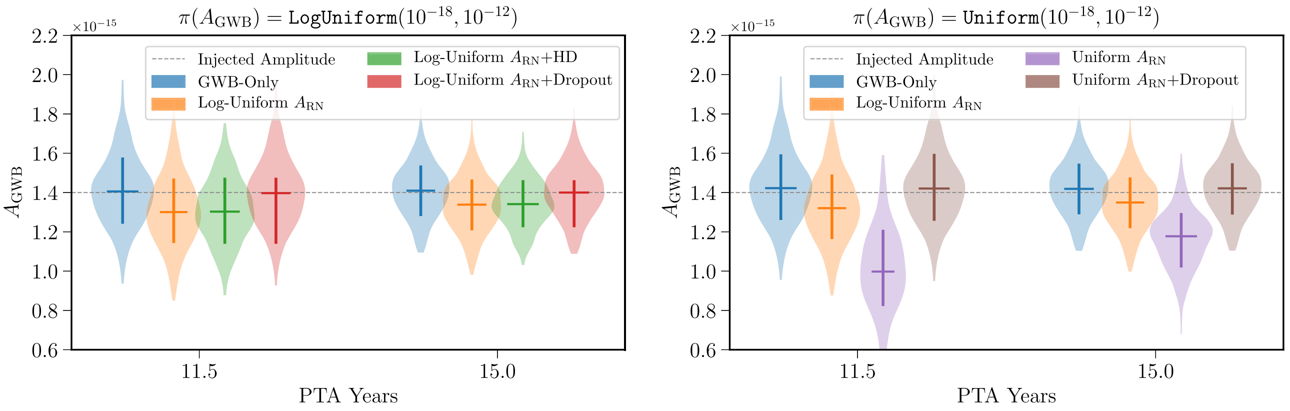
In addition to the two models where we only search for the GWB (GWB-only in Table 1) or where we model intrinsic red noise for *every* pulsar (GWB+RN in Table 1), we consider a more flexible per-pulsar noise model that uses the data to determine whether or not an individual pulsar should be modeled as having red noise ($\mathrm{GWB+RN_{DO}}$ in Table 1). This model is implemented using the so-called *dropout* method (Arzoumanian et al. 2020) on the red noise model, which uses a discrete parameter to switch the red noise model for a particular pulsar on or off during the MCMC sampling of a Bayesian analysis. This extremely flexible tool has been used for investigating the support of deterministic and stochastic signals in particular pulsars (Aggarwal et al. 2019, 2020). The $\mathrm{GWB+RN_{DO}}$ analyses therefore include a red noise model for each pulsar with an amplitude and spectral index along with a dropout parameter. The dropout parameter is sampled from a uniform distribution over the unit interval. If the dropout parameter samples above a certain threshold, then the red noise model acts as usual. If the threshold is not met, then the red noise model is turned off completely. The threshold defines the prior odds ratio for a red noise model to be turned on.

Throughout this work we use a threshold of $10/11$ for the red noise model to be turned on. This means that given no support for red noise, the red noise model will be turned on only $1/11$ of the time. This threshold effectively set an odds ratio of 10:1 as a hurdle to overcome in order to use a red noise model for a particular pulsar. Although one can consider using different threshold values for the dropout model, we do not investigate here how these different values affect the GWB statistics.

2.2. Sensitivity to Choice of Priors

As discussed in Section 1.1, the statistics derived from Bayesian analyses depend on the choice of signal+noise model, including the choice of prior probability distributions for the parameters associated with those models (Kass & Wasserman 1996). Given sufficiently informative data the prior choices will matter little; however, PTA data sets are not yet at this point (Arzoumanian et al. 2018a). As the data sets continue to increase in duration and sensitivity increases, it is critical to understand any potential pitfalls and limitations of our analysis.

Specifically, the simulations and analyses chosen for this work allow us to compare the performance of different models, and their fidelity in returning injected GWB parameters, when applied to a relatively simple data set. If the signal+noise model matches that used in the simulations, we should obtain, on average, the expected coverage for our credible intervals. For signal+noise models that do not match the simulations, over- or undercoverage is possible. Thus, the analyses that we perform here can be thought of as a "sensitivity analysis" (Efron 2015), which checks the robustness of our statistical inference results to the choice of models and priors. However, these simulations do not completely test the "coverage" of different signal+noise models. Coverage is the fidelity of Bayesian credible intervals (or likewise frequentist confidence intervals) over many iterations of an analysis (Heinrich et al. 2004), which allows us to answer the question "does the injected value of a parameter fall within an $X\%$ interval in $X\%$ of simulations?" In order to formally check the coverage of a Bayesian pipeline, one would need to sample from the prior on A_{GWB}, as well as look at different realizations, something that would be too prohibitive to do across all of the models


$\pi(A_{\rm GWB}) = \texttt{LogUniform}(10^{-18}, 10^{-12})$

$\pi(A_{\rm GWB}) = \texttt{Uniform}(10^{-18}, 10^{-12})$

Figure 1. Comparison of the distributions of median values of $A_{\rm GWB}$ for various signal+noise models and prior distributions, obtained from analyzing 400 realizations of the GWB+WN simulations. The left and right panels correspond to log-uniform and uniform prior distributions for $A_{\rm GWB}$. The two groups of four violin plots per panel correspond to two different observation times (11.5 and 15 yr). The horizontal bars show the median of the recovered median values of $A_{\rm GWB}$, while the vertical bars show the central 68% credible interval around that median value. The different color violin plots correspond to the different signal+noise models and prior distributions listed in Table 1.

considered here. As we shall see below, choosing the wrong model—in this case, whether or not some or all pulsars have red noise—can skew the final statistics.

3. Results of Analyses on Simulated Data

The following two subsections describe how the choice of different signal+noise models and prior probability distributions affects both GWB parameter estimation and UL calculations.

3.1. Effect of Signal+Noise Models on GWB Parameter Estimation

We begin by showing the effects of different signal+noise models and priors on estimates of the amplitude $A_{\rm GWB}$ of the GWB. The first analysis that we performed, which serves as the base model for all comparisons, uses the GWB-only signal +noise model—i.e., we only search for the GWB and use TOAs weighted by their WN errors. Not surprisingly, given that we are analyzing the data using the same model that we used to produce the simulations, we recover posterior distributions for the median value of $A_{\rm GWB}$ that agree well with the injected value, $A_{\rm GWB}^{\rm inj} = 1.4 \times 10^{-15}$ (see the blue violin plots in the two panels of Figure 1). The two panels correspond to two different prior probability distributions for $A_{\rm GWB}$, either log-uniform or uniform over the range 10^{-18} to 10^{-14}. Analyses using log-uniform priors are usually referred to as "detection runs" in the PTA literature as they are especially effective for obtaining a Savage–Dickey approximation (Dickey 1971) to the Bayes factor for weak signals. Uniform priors on the amplitude of the GWB are usually used in "upper limit" runs, in order to provide more conservative ULs on $A_{\rm GWB}$.

The other violin plots in Figure 1 show the recovered distributions of the median value of $A_{\rm GWB}$ at two different observation times (11.5 and 15 yr) for the other signal+noise models and priors listed in Table 1. From these recovered distributions we are able to draw several conclusions:

1. As the signal-to-noise ratio increases the choice of prior on the signal amplitude (in this case $A_{\rm GWB}$) has little effect on the median value of $A_{\rm GWB}$, which is expected in a Bayesian analysis. This can be seen by comparing the GWB-only and log-uniform $A_{\rm RN}$ results (blue and orange violin plots, respectively) in the two panels of Figure 1. The choice of prior on $A_{\rm GWB}$ does not considerably change the distribution of median values.
2. The choice of prior on $A_{\rm RN}$ has a dramatic effect on the recovery of the median value of $A_{\rm GWB}$. This can be seen by comparing the orange and purple violin plots in the right panel of Figure 1. These correspond to log-uniform and uniform priors on $A_{\rm RN}$, respectively.
3. With log-uniform priors on both $A_{\rm GWB}$ and $A_{\rm RN}$, recovered median values of $A_{\rm GWB}$ are systematically lower than the injected value 1.4×10^{-15} (orange violin plots in the left panel of Figure 1). This bias remains even if HD correlations are included in the signal model (green violin plots in the left panel of Figure 1).
4. The dropout model mitigates the effects of the $A_{\rm RN}$ prior on the posterior. It does that by including an intrinsic RN model only when it is really needed, returning results consistent with the injected amplitude of the GWB (red and brown violin plots, respectively, in the two panels of Figure 1). The difference between the brown and purple distributions shows that even uniform priors on *both* $A_{\rm GWB}$ and $A_{\rm RN}$ return fairly accurate median values for $A_{\rm GWB}$ when the dropout method is used.

While the differences in the third point above seem fairly small (i.e., the orange and green/second and third violin plots in each set of the left panel of Figure 1 are shifted lower by about 7%), these shifts can have fairly drastic results when considering the interpretation of a single posterior from real data. Instead of distributions of the median, one can tally the quantile position of the injected value for each of the "detection runs" (which use log-uniform priors for $A_{\rm GWB}$). In the models where red noise is assumed for all pulsars, the injected value falls higher than a given credible interval 3–7 times more often than it falls lower than the same credible interval. In other words, one is 3–7 times more likely to *underestimate* $A_{\rm GWB}$ than overestimate it. However, when the red noise dropout method is used, the discrepancy between the injected value falling higher or lower than the credible interval is reduced considerably, giving only a 20% difference, corresponding to a factor of 1.2.

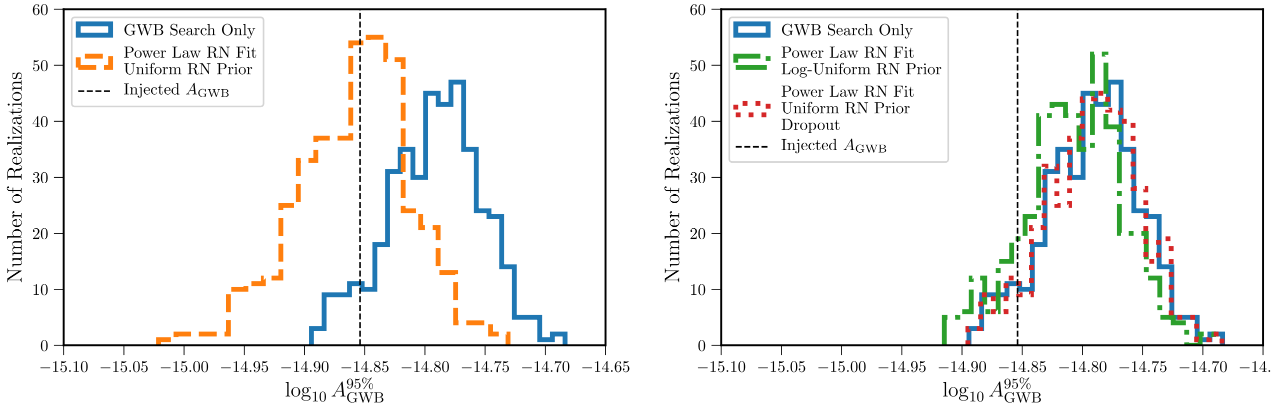


Figure 2. Comparison of the distributions of the 95% ULs for various signal+noise models and prior distributions, obtained from analyzing 400 realizations of the GWB+WN simulations. The left panel shows results for both the simple GWB-only signal+noise model (blue solid) and the standard PTA UL analysis (orange dashed), which additionally includes red noise models for each pulsar with a uniform prior on the amplitude $A_{\rm RN}$. The right panel also shows results for a signal+noise model that uses log-uniform priors for $A_{\rm RN}$ (green dotted–dashed), and a dropout model for uniform red noise priors (red dotted).

3.2. Effect of Signal+Noise Models on GWB Upper Limits

We also determine the effects of the different signal+noise models and prior distributions on the 95% UL, $A^{95\%}_{\rm GWB}$. For all of the UL analyses, we use the standard convention of using a uniform prior on $A_{\rm GWB}$ (Arzoumanian et al. 2016, 2018a; Lentati et al. 2016). For the GWB-only signal+noise model, we recover results for $A^{95\%}_{\rm GWB}$ consistent with expectations (see the solid blue curve in either the left or right panel of Figure 2). As discussed in Section 3.1, this is because the GWB-only signal+noise model agrees with that used to produce the simulated data. The distribution of UL values calculated for the 400 realizations of the simulated data has values that are greater than the injected value of the background roughly 93% of the time (within error of the expected value of 95%).

We then perform analyses using the standard model for PTA GWB searches that includes an intrinsic red noise model for each pulsar. Recall that these models introduce $2N_{\rm pulsars}$ additional parameters (an amplitude, $A_{\rm RN}$, and spectral index, $\gamma_{\rm RN}$, for each pulsar). To the best of our knowledge most Bayesian PTA ULs to date have used uniform priors on $A_{\rm RN}$ and $A_{\rm GWB}$. From the dashed orange curve shown in the left panel of Figure 2, we see that the distribution of 95% ULs is shifted to significantly lower values, with basically *even odds*, i.e., 50%, that the calculated UL is above or below the injected value.

It is worth pointing out that there is nothing wrong with the ULs produced by this procedure, as long as one is explicit about the model being used in the Bayesian analysis. However, if the signal+noise model that we are using differs greatly from what gave rise to the data, then statistical inferences can be systematically (and significantly) inaccurate. The standard PTA UL analysis is based on the "conservative" assumption that all pulsars have substantial red noise. In our simulations this assumption leads to individual pulsar red noise models that are able to absorb a substantial amount of the common red process, i.e., the GWB, and thus produce an overall *smaller* $A^{95\%}_{\rm GWB}$. The effects of the "conservative" assumption can be somewhat mitigated by using instead a log-uniform prior on $A_{\rm RN}$. This choice of prior decreases the bias, as one can see from the dotted–dashed green distribution of $A^{95\%}_{\rm GWB}$ in the right panel of Figure 2, which has 87.25% of its ULs above the injected value. The log-uniform prior, however, is still part of a signal+noise model that assumes that *all* pulsars have at least some level of measurable red noise.

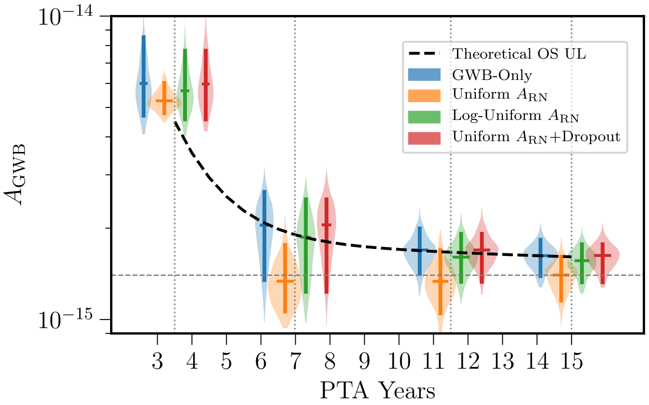


Figure 3. Comparison of the distributions of the 95% ULs for various signal+noise models and prior distributions, obtained from analyzing 400 realizations of the GWB+WN simulations. The dashed black line shows the theoretical evolution of the UL based on the frequentist optimal statistic; see Equation (A5). The violin plots show the distributions of Bayesian 95% ULs for four different signal+noise models and priors at four different time slices of the full data set. The vertical bar within a violin plot shows the central 90% credible interval. The horizontal dashed gray line shows the injected amplitude of 1.4 × 10^{-15}.

As we have already seen in Section 3.1, a better option is to use a red noise pulsar model in conjunction with the dropout method, which allows the data to decide whether a given pulsar should be modeled to include intrinsic red noise. The dropout analysis turns off the red noise models in almost all cases for this simple simulation of GWB+WN-only, returning us to results commensurate with the GWB-only search (see the dotted red distribution in the right panel of Figure 2, which has 93.25% of its ULs above the injected value). The dropout method does this by using the threshold value to effectively set a prior on the presence (or absence) of intrinsic red noise in our pulsars, allowing the data to inform the choice of noise model.

In Figure 3 we show the evolution of the UL as a function of time for the various analyses. The injected $A_{\rm GWB}$ is specifically chosen so that the GWB begins to have power greater than the WN at the lowest frequencies when there is 7 yr of data. Before this time the ULs are all still well above the injected value, but as soon as one enters the intermediate signal regime,[7] biases in the standard GWB UL analysis (orange violin plots) are clearly manifest.

In comparison, the distributions for the dropout analysis (red violin plots) match those for the GWB-only search (blue violin plots), which also match the expected 95% confidence-level UL (black dashed line) predicted by the frequentist optimal statistic (OS) developed in Ellis et al. (2013), Siemens et al. (2013), Chamberlin et al. (2015), and Vigeland et al. (2018). The scaling laws of Siemens et al. (2013) are used to construct an analytic expression for $A_{\rm GWB}^{95\%}$ as a function of time using Equation (A5). Thus, just as we saw in Section 3.1, the pulsar red noise dropout analysis performs better than the standard PTA analysis (uniform amplitude priors for both $A_{\rm GWB}$ and $A_{\rm RN}$) for 95% ULs as well.

4. Application to More Realistic Data Sets

As discussed earlier, the simple simulations studied in the previous sections only included WN and a red GWB signal. For these simulations, the dropout model successfully turns off the red noise models in all pulsars, as can be seen in the top panel of Figure 4. In this panel, the vertical height of the dots shows the fraction of time the red noise dropout model is turned on for a given realization at a given slice of the data set. A data set that is completely ambivalent about the presence of red noise would lie along the horizontal thin dashed line, i.e., the pulsars will have their red noise model turned on $1/11$ of the time, while the presence of most dots below the line shows that red noise is disfavored. Note the realization dependence for these parameters. A given white noise realization or GWB realization may be modeled better by the red noise model and result in a dropout parameter that is turned on a majority of the time. In these cases these parameters have smaller spectral indices, and hence are picking up whiter noise.

In order to assess the full abilities of the red noise dropout model, another set of simulations were run where a handful of pulsars were injected with red noise characteristic of that seen in real PTA data sets. Various amplitudes and spectral indices of red noise were injected and are detailed in the caption of Figure 4. The middle panel shows the analogous results to the top panel, but with the successful modeling of red noise in the pulsars where it has been injected. Depending on the spectral index and amplitude, it may take longer in the data set to resolve the red noise in the pulsar, as shown by the dependence of Equation (A2) on γ. However, for the full 15 yr of data, those pulsars with injected red noise have red noise dropout models turned on through most steps of the MCMC. The bottom left panel of Figure 4 shows the distribution of medians for $A_{\rm GWB}$ from both dropout analyses. As one can see, the red noise dropout model does just as well at estimating $A_{\rm GWB}$ whether there is red noise present in some of the pulsars or not. The *P*–*P* plot in the bottom right panel of Figure 4 shows the fraction of parameter recoveries for which the injected value appeared at or below a given percentile, e.g., in both cases the injected $A_{\rm GWB}$ appeared at or below the 55th percentile in about 55% of the realizations. There is little bias in these analyses across the 400 realizations examined.

5. Reassessing the NANOGrav 11 yr Analysis

Finally we turn to real PTA data and apply the red noise dropout model to the NANOGrav 11 yr data set (Arzoumanian et al. 2018b). In addition to adding a dropout parameter for each of the pulsars, we also use the BAYESEPHEM (Vallisneri et al. 2020) solar system ephemeris model so that our results are directly comparable to those of Arzoumanian et al. (2018a). This analysis includes the Hellings–Downs spatial correlations for the GWB. Looking at the dropout parameters in the top panel of Figure 5 we see that the analysis largely agrees with the noise analysis used in Arzoumanian et al. (2018b). The only pulsar deemed significant in Arzoumanian et al. (2018b), where the red noise model is turned off a majority of the time during the dropout analysis, is PSR J1909–3744. In a single pulsar noise analysis the red noise parameters in this pulsar are significant and similar to those expected for a GWB. In a standard full PTA GWB analysis, where all pulsars have red noise models, the red noise posteriors returned are very uninformative, i.e., the red noise in the pulsar is modeled as the common process. In this dropout analysis the results are similar, except that rather than return uninformative priors, the analysis now turns off the individual red noise model, in favor of the red noise power going into the common process.

The posterior on $A_{\rm GWB}$ is compared to the standard PTA analysis of the Arzoumanian et al. (2018b) data set in Figure 5. Any evidence for a detection using the dropout analysis, though slightly better, is still marginal. However, as might be expected from the analyses in Section 3.1, the maximum a posteriori (MAP) value for $A_{\rm GWB}$ using the dropout analysis is $A_{\rm GWB} = 1.4 \times 10^{-15}$, which is larger than that from the standard analysis, $A_{\rm GWB} = 1.1 \times 10^{-15}$, and more similar to the 95% UL obtained in Arzoumanian et al. (2018a). We can reweight the samples in either of the analyses (Gelman et al. 2013) shown in Figure 5 to obtain new 95% ULs. Reweighting the samples from the standard PTA analysis is equivalent to the log-uniform $A_{\rm RN}$ analysis discussed in Section 3.2 and gives $A_{\rm GWB}^{95\%} = 2.1 \times 10^{-15}$. If instead we reweight the samples from the dropout analysis from Figure 5, we obtain results comparable to the dropout analysis from Section 3.2, which appears to be the most trustworthy model examined here for obtaining ULs. This gives a UL for the NANOGrav 11 yr data set of $A_{\rm GWB}^{95\%} = 3.0 \times 10^{-15}$, *which is more than twice as large* as the UL quoted in Arzoumanian et al. (2018a). This result is dependent on the threshold set for the dropout parameter, as discussed in Section 2.1, and should not be taken as a concrete astrophysical result, but rather an example of how the GWB statistics can shift when using this new model. For the most up-to-date GWB results see Arzoumanian et al. (2020).

6. Conclusions

Here we have shown explicitly how the choice of prior on $A_{\rm RN}$, and indeed whether pulsars have red noise models at all, can have unanticipated consequences on the statistics of $A_{\rm GWB}$. In a simulated data set with WN and a GWB, the effect of a standard GWB search with uniform priors for the pulsar intrinsic red noise amplitudes on $A_{\rm GWB}^{95\%}$ is drastic, returning a 95% UL *lower* than the injected value in about half of all realizations. As we have seen, putting a uniform prior on $A_{\rm RN}$ biases the noise model to steal power from the GWB model. The biases of other estimators, such as the conditional median, that occur for parameter estimation are smaller, but still show a

[7] Defined in Siemens et al. (2013) as beginning when the power in the lowest frequency bin of the GWB is greater than that of the WN.

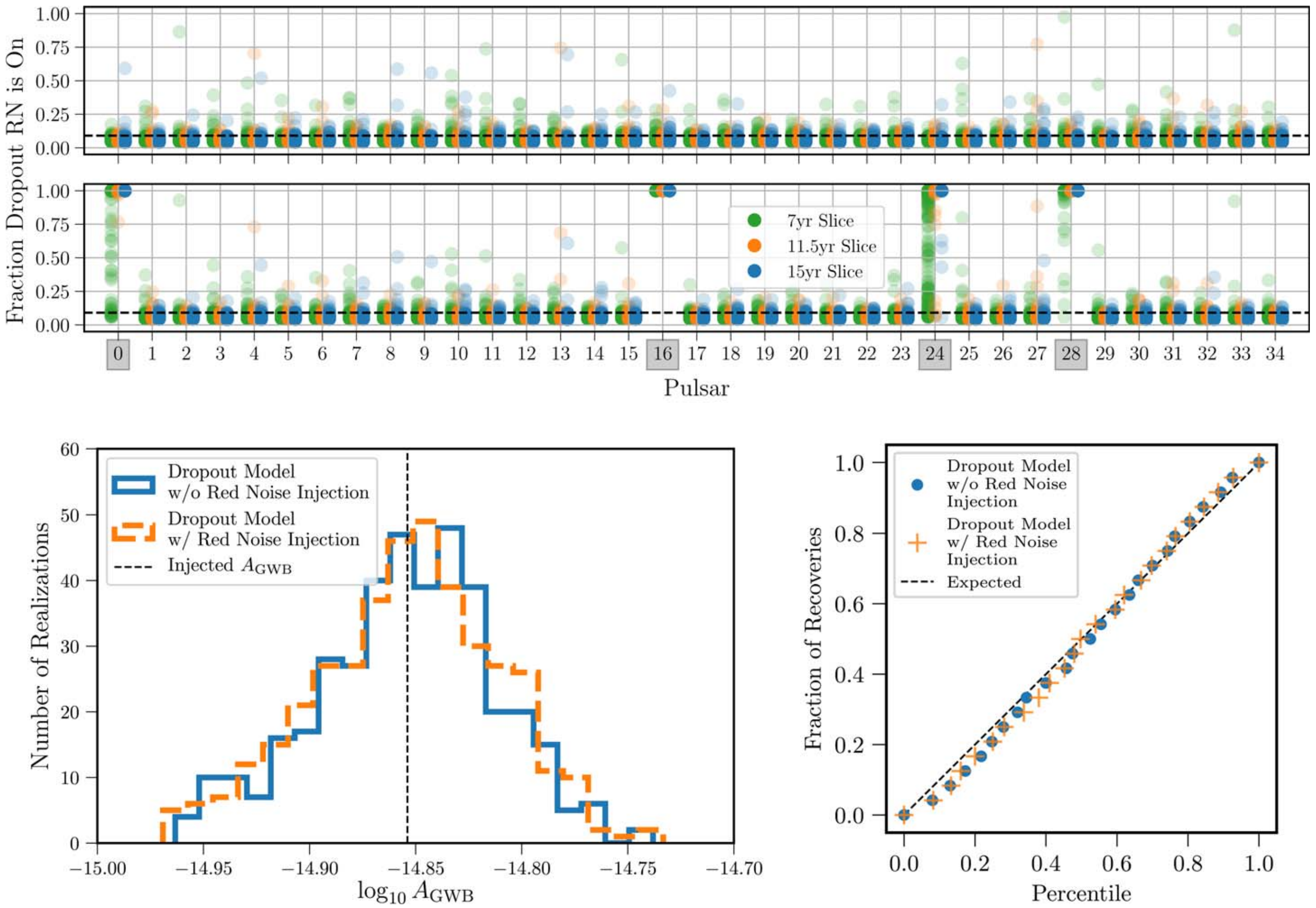


Figure 4. Summary of dropout analyses for simulated data set analyses. In the top two panels each dot represents the fraction of the samples from a particular realization/analysis when the red noise model is turned on; dots below the dashed black line indicate that a particular pulsar's red noise model is disfavored. The topmost panel shows the fraction turned on for the analysis of the GWB+WN-only simulations. The middle panel is for a simulation where additional red noise is injected into pulsars 0, 16, 24, and 28 (highlighted in gray) with $A_{\rm RN}$ equal to (10^{-15}, 3.4×10^{-13}, 8×10^{-15}, and 7×10^{-14}) and spectral indices $\gamma_{\rm RN}$ equal to (7, 2, 5, and 3) respectively. The bottom left panel shows the distribution of the median values for $A_{\rm GWB}$ for the dropout analysis applied to both the GWB+WN-only and GWB+WN+RN simulations. The bottom right panel shows a probability–probability (*P*–*P*) plot showing the cumulative distribution of injection percentiles for both dropout analyses.

consistent shift in the same direction, i.e., to smaller values of $A_{\rm GWB}$. We have implemented a simple solution to these problems—the so-called dropout method—which is a flexible red noise model for pulsars that allows the intrinsic red noise model in each pulsar to be turned off during the course of the Bayesian analysis if there is not sufficient evidence in the data to warrant its presence.

In light of the offsets in parameter estimation uncovered by this work, it is worth revisiting constraints inferred on the SMBBH population from PTA data sets. The impact of the choice of prior on the intrinsic red noise amplitudes can be seen directly by comparing the astrophysical interpretation done in NANOGrav's 9 yr GWB constraint paper (Arzoumanian et al. 2016), which used uniform priors on red noise amplitude, with that done in its 11 yr GWB constraint paper (Arzoumanian et al. 2018a), which used log-uniform priors. Even though the 11 yr constraint on $A_{\rm GWB}$ is a smaller value, the astrophysical inference is less constraining. This is partially due to the differing models and analysis techniques. However, viewed through the lens of this work, the weakening of constraints, specifically on the M–$M_{\rm bulge}$ relationship, were certainly impacted by the choice of prior used for the red noise amplitudes. Beyond direct constraints using $A_{\rm GWB}$, previously reported $A_{\rm GWB}^{95\%}$ upper limits have been used in concert with electromagnetic observations to make statements about various SMBBH population models (Holgado et al. 2018; Sesana et al. 2018). Moving forward, astrophysical statements derived from past PTA constraints on $A_{\rm GWB}$ will need to be more cautious in assessing the strength of their inference.

The idea that credible intervals and parameter estimation are dependent on the choice of model and priors is a common refrain in Bayesian statistics. As a result, data analysts should strive to produce models and priors that robustly represent and quantify the underlying physical processes being investigated.

This work was supported by the NSF NANOGrav Physics Frontier Center (NSF PHY-1430284). The authors thank Paul Baker, Michael Lam, Timothy Pennucci, Stephen Taylor, Michele Vallisneri, and Sarah Vigeland for useful discussions and detailed comments on early drafts of this manuscript. 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. J.S. acknowledges support from the JPL RTD program.

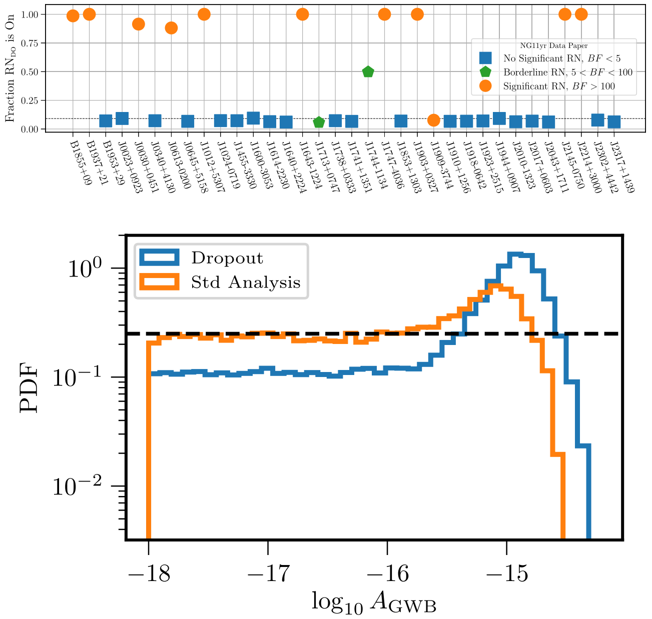


Figure 5. Summary of the dropout analysis for the NANOGrav 11 yr data. In the top panel, the various symbols show the fraction of samples where the red noise model is turned on for a given pulsar, color coded by the red noise Bayes factor. The bottom panel compares the GWB posteriors, including Hellings–Downs spatial correlations, for the NANOGrav 11 yr data with (blue) and without (orange) the red noise dropout analysis applied. The maximum a posteriori value of the GWB amplitude is $A_\mathrm{GWB} = 1.41 \times 10^{-15}$ when the dropout analysis is applied; it equals 1.1×10^{-15} without the dropout analysis.

Finally, we are grateful for computational resources provided by the Leonard E Parker Center for Gravitation, Cosmology and Astrophysics at the University of Wisconsin–Milwaukee (NSF PHY-1626190).

Facilities: Arecibo, GBT.

Software: astropy (Astropy Collaboration et al. 2018), libstempo (Vallisneri 2020), Tempo2 (Hobbs & Edwards 2012), la_forge, ENTERPRISE (Ellis et al. 2019), enterprise_extensions, PTMCMCSampler (Ellis & van Haasteren 2017).

Appendix
Time Evolution of a Frequentist GWB Upper Limit

Here we derive an expression for the expected value of the frequentist 95% confidence-level UL calculated from the optimal statistic (Ellis et al. 2013; Siemens et al. 2013; Chamberlin et al. 2015; Vigeland et al. 2018). Although this is a frequentist UL, it provides a good analytic approximation to the Bayesian 95% ULs calculated in this Letter.

The expected signal-to-noise ratio ρ derived from the optimal statistic can be written in the frequency domain as (Chamberlin et al. 2015)

$$\rho \equiv \sqrt{\langle \rho^2 \rangle} = \left(2T \sum_{IJ} \chi_{IJ}^2 \int_{f_L}^{f_H} df \frac{P_g^2(f)}{P_I(f) P_J(f)} \right)^{1/2}, \qquad \text{(A1)}$$

where the indices I and J label the individual pulsars, $P_I(f)$ is the total autocorrelated power spectral density for pulsar I, $P_g(f)$ is the power spectral density for the GWB, T is the time span of the data, and χ_{IJ} are the overlap reduction function coefficients, here assumed to be the quadrupolar spatial correlations induced by a GWB (Hellings & Downs 1983).

The expression for the signal-to-noise ratio can be simplified considerably for the main set of simulations considered in this work where all of the pulsars have the same level of white noise, cadence, and observing time span, and there is no red noise injected into the pulsars:

$$\rho = \left(2T \sum_{IJ} \chi_{IJ}^2 \int_{f_L}^{f_H} df \frac{b^2 f^{-2\gamma}}{(bf^{-\gamma} + 2\sigma^2 \Delta t)^2} \right)^{1/2}. \qquad \text{(A2)}$$

Here σ is the TOA error value for all the pulsars, Δt is the sampling period (the cadence), and b subsumes various constants,[8]

$$b \equiv \frac{A_\mathrm{GWB}^2}{12\pi^2} \left(\frac{1}{f_\mathrm{yr}} \right)^{-\gamma+3}. \qquad \text{(A3)}$$

One can relate the aforementioned scaling laws to a UL by using the complementary error function (Allen & Romano 1999; Hazboun et al. 2020)

$$A_\mathrm{UL}^2 = \hat{A}_\mathrm{GWB}^2 + \frac{\sqrt{2}\,\sigma_0}{\sqrt{T}} \operatorname{erfc}^{-1}[2(1-\mu)], \quad \rho \equiv \frac{A_\mathrm{GWB}^2}{\sigma_0/\sqrt{T}}, \qquad \text{(A4)}$$

where μ is the confidence level (e.g., $\mu = 0.95$ for a 95% confidence-level UL), and σ_0 is the effective noise level defined in terms of ρ, A_GWB, and T. The expectation value of Equation (A4) yields

$$\langle A_\mathrm{UL}^2 \rangle = A_\mathrm{GWB}^2 \left(1 + \frac{\sqrt{2}\, \operatorname{erfc}^{-1}[2(1-\mu)]}{\rho} \right). \qquad \text{(A5)}$$

For the simple GWB+WN simulations studied in the majority of this Letter, this UL is calculated analytically and plotted in Figure 3 as the dashed black line.

ORCID iDs

Jeffrey S. Hazboun https://orcid.org/0000-0003-2742-3321
Joseph Simon https://orcid.org/0000-0003-1407-6607
Xavier Siemens https://orcid.org/0000-0002-7778-2990
Joseph D. Romano https://orcid.org/0000-0003-4915-3246

References

Aggarwal, K., Arzoumanian, Z., Baker, P. T., et al. 2019, ApJ, 880, 116
Aggarwal, K., Arzoumanian, Z., Baker, P. T., et al. 2020, ApJ, 889, 38
Allen, B., & Romano, J. D. 1999, PhRvD, 59, 102001
Arzoumanian, Z., Baker, P. T., Blumer, H., et al. 2020, arXiv:2009.04496
Arzoumanian, Z., Baker, P. T., Brazier, A., et al. 2018a, ApJ, 859, 47
Arzoumanian, Z., Brazier, A., Burke-Spolaor, S., et al. 2014, ApJ, 794, 141
Arzoumanian, Z., Brazier, A., Burke-Spolaor, S., et al. 2016, ApJ, 821, 13
Arzoumanian, Z., Brazier, A., Burke-Spolaor, S., et al. 2018b, ApJS, 235, 37
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Blandford, R., Narayan, R., & Romani, R. W. 1984, JApA, 5, 369
Burke-Spolaor, S., Taylor, S. R., Charisi, M., et al. 2018, arXiv:1811.08826
Callister, T., Biscoveanu, A. S., Christensen, N., et al. 2017, PhRvX, 7, 041058
Carlin, B. P., & Chib, S. 1995, J. R. Stat. Soc. Ser. B, 57, 473
Chamberlin, S. J., Creighton, J. D. E., Siemens, X., et al. 2015, PhRvD, 91, 044048
Champion, D. J., Hobbs, G. B., Manchester, R. N., et al. 2010, ApJL, 720, L201
Cordes, J. M. 2013, CQGra, 30, 224002

[8] Notice the difference in minus sign of the exponent corrected from Hazboun et al. (2020).

Cordes, J. M., & Downs, G. S. 1985, ApJS, 59, 343
Cordes, J. M., & Shannon, R. M. 2010, arXiv:1010.3785
Demorest, P. B., Ferdman, R. D., Gonzalez, M. E., et al. 2013, ApJ, 762, 94
Detweiler, S. 1979, ApJ, 234, 1100
Dickey, J. M. 1971, Ann. Math. Stat., 42, 204
Efron, B. 2015, J. R. Stat. Soc. Ser. B, 77, 617
Ellis, J., & van Haasteren, R. 2017, jellis18/PTMCMCSampler: Official Release, v. 1.0.0, Zenodo, doi:10.5281/zenodo.1037579
Ellis, J. A., & Cornish, N. J. 2016, PhRvD, 93, 084048
Ellis, J. A., Siemens, X., & van Haasteren, R. 2013, ApJ, 769, 63
Ellis, J. A., Vallisneri, M., Taylor, S. R., & Baker, P. T. 2019, ENTERPRISE: Enhanced Numerical Toolbox Enabling a Robust PulsaR Inference SuitE, Astrophysics Source Code Library, ascl:1912.015
Foster, R. S., & Backer, D. C. 1990, ApJ, 361, 300
Gelman, A., Carlin, J., Stern, H., et al. 2013, Bayesian Data Analysis, Third Edition, Chapman & Hall/CRC Texts in Statistical Science (London: Taylor and Francis), https://books.google.com/books?id=ZXL6AQAAQBAJ
Gelman, A., & Hill, J. 2007, Data Analysis Using Regression and Multilevel/hierarchical Models, Vol. Analytical Methods for Social Research, xxii (New York: Cambridge Univ. Press), 625
Godsill, S. J. 2001, J. Comput. Graphical Stat., 10, 230
Hazboun, J. S., Mingarelli, C. M. F., & Lee, K. 2018, arXiv:1810.10527
Hazboun, J. S., Simon, J., Taylor, S. R., et al. 2020, ApJ, 890, 108
Hee, S., Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015, MNRAS, 455, 2461
Heinrich, J., Blocker, C., Conway, J., et al. 2004, arXiv:physics/0409129
Hellings, R. W., & Downs, G. S. 1983, ApJL, 265, L39
Hobbs, G., & Edwards, R. 2012, Tempo2: Pulsar Timing Package, Astrophysics Source Code Library, ascl:1210.015
Holgado, A. M., Sesana, A., Sandrinelli, A., et al. 2018, MNRAS, 481, L74
Jaffe, A. H., & Backer, D. C. 2003, ApJ, 583, 616
Kass, R. E., & Wasserman, L. 1996, J. Am. Stat. Assoc., 91, 1343
Lentati, L., Taylor, S. R., Mingarelli, C. M. F., et al. 2015, MNRAS, 453, 2576
Lentati, L., Shannon, R. M., Coles, W. A., et al. 2016, MNRAS, 458, 2161
Phinney, E. S. 2001, arXiv:astro-ph/0108028
Rosado, P. A., Sesana, A., & Gair, J. 2015, MNRAS, 451, 2417
Sazhin, M. V. 1978, SvA, 22, 36
Sesana, A., Haiman, Z., Kocsis, B., & Kelley, L. Z. 2018, ApJ, 856, 42
Shannon, R. M., Ravi, V., Lentati, L. T., et al. 2015, Sci, 349, 1522
Siemens, X., Ellis, J., Jenet, F., & Romano, J. D. 2013, CQGra, 30, 224015
Simon, J., & Burke-Spolaor, S. 2016, ApJ, 826, 11
Taylor, S. R., Gair, J. R., & Lentati, L. 2013, PhRvD, 87, 044035
Taylor, S. R., Lentati, L., Babak, S., et al. 2017, PhRvD, 95, 042002
Taylor, S. R., van Haasteren, R., & Sesana, A. 2020, arXiv:2006.04810
Tiburzi, C., Hobbs, G., Kerr, M., et al. 2016, MNRAS, 455, 4339
Vallisneri, M. 2020, libstempo: Python wrapper for Tempo2, Astrophysics Source Code Library, ascl:2002.017
Vallisneri, M., Taylor, S. R., Simon, J., et al. 2020, ApJ, 893, 112
van Haasteren, R., & Levin, Y. 2013, MNRAS, 428, 1147
van Haasteren, R., & Vallisneri, M. 2014, PhRvD, 90, 104012
Vigeland, S. J., Islo, K., Taylor, S. R., & Ellis, J. A. 2018, PhRvD, 98, 044003
Williams, C. K., & Rasmussen, C. E. 2006, Gaussian Processes for Machine Learning, 2 (Cambridge, MA: The MIT Press), 4
Yardley, D. R. B., Coles, W. A., Hobbs, G. B., et al. 2011, MNRAS, 414, 1777