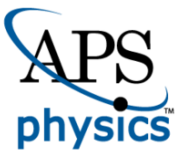


Reliability of parameter estimates in the first observing run of Advanced LIGO

Suman Kulkarni and Collin D. Capano

Phys. Rev. D **103**, 104002 — Published 3 May 2021

DOI: 10.1103/PhysRevD.103.104002

On the reliability of parameter estimates in the first observing run of Advanced LIGO

Suman Kulkarni*
Indian Institute of Science Education and Research, Homi Bhabha road, Pashan, Pune 411008, India

Collin D. Capano†
Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), D-30167 Hannover, Germany
Leibniz Universität Hannover, D-30167 Hannover, Germany

(Dated: March 22, 2021)

Accurate parameter estimation is key to maximizing the scientific impact of gravitational-wave astronomy. Parameters of a binary merger are typically estimated using Bayesian inference. It is necessary to make several assumptions when doing so, one of which is that the detectors output stationary Gaussian noise. We test the validity of these assumptions by performing percentile-percentile tests in both simulated Gaussian noise and real detector data in the first observing run of Advanced LIGO (O1). We add simulated signals to 512s of data centered on each of the three events detected in O1 — GW150914, GW151012, and GW151226 — and check that the recovered credible intervals match statistical expectations. We find that we are able to recover unbiased parameter estimates in the real detector data, indicating that the assumption of Gaussian noise does not adversely effect parameter estimates. However, we also find that both the parallel-tempered sampler `emcee_pt` and the nested sampler `dynesty` struggle to produced unbiased parameter estimates for GW151226-like signals, even in simulated Gaussian noise. The `emcee_pt` sampler does produce unbiased estimates for GW150914-like signals. This highlights the importance of performing percentile-percentile tests in different targeted areas of parameter space.

I. INTRODUCTION

To date, the LIGO [1] and Virgo [2] observatories have detected over 50 gravitational waves from binary black hole and binary neutron star mergers [3–7]. These detections have opened a new window on to the Universe, offering insights that would not be possible from electromagnetic observations alone. This is because gravitational waves carry information about the source binaries' parameters. By careful observation of a gravitational wave's frequency and amplitude evolution, one can estimate a binary's masses, spins, location, and other properties. From these measurements it is possible to infer the distribution of black holes in the universe [8], measure the equation of state of dense nuclear matter [9–11], constrain cosmological parameters [12–14], and (by allowing modifications to model waveforms) test general relativity (GR) in the strong-field regime [15–18]. Accurate parameter estimation is therefore critical to the success of gravitational-wave astronomy.

Bayesian inference is the standard method by which physical parameters are extracted from gravitational waves. Given some observed data and a gravitational-wave model, a posterior probability distribution is obtained on the parameters describing the binary. This posterior is then marginalized to obtain credible intervals on specific parameters. Stochastic samplers are needed to fully map out the posterior distribution due to the high dimensionality and complicated topology of the parameter space. Several assumptions are made in this process, such as: the template gravitational-wave model is an accurate representation of the signal; the digital output from the detectors is accurately calibrated to the strain induced by the passing gravitational wave; the stochastic sampler produces an accurate representation of the posterior.

One of the most critical assumptions made when estimating parameters is that, in the absence of a signal, the detectors output wide-sense stationary Gaussian noise. This leads to a canonical likelihood function that can be evaluated numerically, providing an algorithm for estimating posterior distributions. However, the LIGO and Virgo detectors are known to produce a number of non-Gaussian noise transients ("glitches") [19–23]. **Indeed, search pipelines must employ a number of bespoke statistics in order to overcome these glitches [24–27]. Non-Gaussian transients are generally less of a concern for parameter estimation. Since gravitational waves from binary mergers have finite duration within the LIGO/Virgo sensitive frequency band, estimating the parameters of a binary typically involves between O(10s) and O(100s) of data. At current sensitivity, both the signal rate and glitch rate is low enough that it is rare for a non-Gaussian transient to occur during the time of interest for a parameter estimation analysis. Even so, it has happened.**

A large transient occurred in the Livingston detector $\sim 1\,\mathrm{s}$ before the merger of GW170817 [22]. This nearly caused the signal to be missed by low-latency search pipelines [28]. For the followup parameter estimation

* suman.kulkarni@students.iiserpune.ac.in
† collin.capano@aei.mpg.de

analyses, it was necessary to remove the transient from the data by fitting a glitch model to the transient [22, 29]. Without this removal procedure the inferred parameters of the binary would have been biased [30]. This, in turn, would have yielded misleading answers to key questions, such as what the equation of state of dense nuclear matter is.

The transient that occurred during GW170817 was loud enough that it could be identified and removed. However, its presence raises the question, what if there is less-obvious non-Gaussian noise present in the data containing a binary merger? Such noise, being unmodelled, could adversely affect the reported credible intervals. This would be particularly problematic for tests of GR using gravitational waves.

Several previous studies have tested the validity of Bayesian inference tools used for gravitational-wave parameter estimation [31–36]. The standard test is to perform a percentile-percentile test. This involves generating a number of simulated signals, which are drawn from some prior distribution. These are added to some data; a Bayesian inference analysis is then performed on each signal separately, using the same prior. This yields credible intervals on each of the signals' parameters. If all of the assumptions made in the analysis are valid, then $X\%$ of the signals should fall within the X-percentile credible interval for any given parameter.

The primary focus of previous studies employing percentile-percentile tests was to validate software tools used for gravitational-wave Bayesian inference [31, 32, 34–36]. This was accomplished by adding simulated signals to simulated Gaussian noise, colored by power spectral densities (PSDs) representative of the LIGO and Virgo detectors. A violation of a percentile-percentile test under these conditions would indicate that some aspect of the software — e.g., the stochastic sampler used — was not properly recovering the posterior distribution. **A percentile-percentile test was also used in Ref. [33] to study the accuracy of parameter estimates of signals that were anticipated to be detected in Advanced LIGO given non-Gaussian noise. As Advanced LIGO had not begun yet, that study used Initial LIGO data that was recolored to resemble the expected power spectral density of Advanced LIGO.**

In this paper we, for the first time, perform percentile-percentile tests in real **Advanced LIGO** data. Our aim is to test the assumption that the detectors' output can be modelled as a gravitational-wave plus stationary Gaussian noise in data surrounding identified gravitational-wave events. We do this by performing identical percentile-percentile tests in real and simulated Gaussian noise. Should the test be violated in the former and not the latter, it would indicate that the detector data is not sufficiently Gaussian during the observation time. We apply this test to the three gravitational-wave events that were detected during the first observing run (O1) of Advanced LIGO: GW150914 [37], GW151012 [3, 4, 38], and GW151226 [39].

Our paper is structured as follows: in Section II we review Bayesian inference and its application to gravitational waves. In Section III we detail the methods used in this study. The results of our study are discussed in Section IV. Finally, in Section V, we discuss the implications of our results and prospects for future studies.

II. PARAMETER ESTIMATION USING BAYESIAN INFERENCE

Consider a network of N detectors labelled with indices i = 1,..., N. The data collected at the i^{th} detector is a time series comprising of a signal under some waveform model H along with the detector noise:

$$d_i(t) = n_i(t) + s_i(t).$$

Here, $n_i(t)$ is the noise observed at the i^{th} detector and $s_i(t)$ is the gravitational waveform obtained under the model used. We denote the collection of data at all detectors by $\mathcal{D}(t)$ and the set of parameters for the waveform model by $\mathbf{v}$.

Applying the Bayes' Theorem, the posterior probability density function is

$$\underbrace{P(\mathbf{v}|\mathcal{D}(t),H)}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}(t)|\mathbf{v},H)}^{\text{Likelihood}}\overbrace{P(\mathbf{v}|H)}^{\text{Prior}}}{\underbrace{P(\mathcal{D}(t)|H)}_{\text{Evidence}}}. \quad (1)$$

The prior indicates our knowledge of the parameters in a given model before analysing the data. **Assuming circular orbits,** models describing binary black hole mergers involve 15 parameters: the component masses $m_{1,2}$, the magnitude and orientation of the component spins, the luminosity distance d_L, the right ascension α, the declination δ, the polarization ψ, the binary inclination angle ι, the coalescence time t_c, and the phase at the time of coalescence ϕ.

We use similar priors as in Ref. [6]. For BBH mergers it is common to use uniform priors on the component source masses, choosing the bounds such that all regions with non-zero posterior support are within the boundaries. For the magnitudes of each component spin vector, we use uniform priors on $a_{1,2} \in [0.0, 0.99)$. For the other two components of each spin vector, we use a uniform solid angle prior, which assumes a uniform distribution for the azimuthal angle $\theta_{1,2}^{azimuthal} \in [0, 2\pi]$ and a sine-angle distribution for the polar angle $\theta_{1,2}^{polar}$. We make use of a uniform sky location prior, which assumes a uniform distribution on $\alpha \in [0, 2\pi)$ and a cosine-angle distribution for δ. The polarization angle ψ is uniform $\in [0, 2\pi)$ and the inclination ι uses a sine-angle prior. Uniform priors are used for the coalescence time ± 0.1 s around the time of the merger. A uniform prior on $[0, 2\pi)$ is used for ϕ and analytically marginalized over, as discussed later in this section.

For the luminosity distance, a uniform prior on comoving volume was used in Ref. [6] (which was converted to luminosity distance by assuming a standard ΛCDM cosmology [40]), with bounds chosen to enclose the posterior for each event. However, in this work, we use a prior uniform in the $\log_{10}$ of the comoving volume. We do this so as to sample higher SNRs; see Section III B for details.

We assume that each detector outputs independent, wide-sense stationary Gaussian noise in the absence of a signal. Under this assumption the likelihood function is

$$P(\mathcal{D}(t)|\mathbf{v}, H) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\langle \tilde{d}_i(f) - \tilde{s}_i(f,\mathbf{v}),\, \tilde{d}_i(f) - \tilde{s}_i(f,\mathbf{v})\rangle\right] \quad (2)$$

The inner product $\langle \tilde{a}_i(f),\, \tilde{b}_i(f)\rangle$ is defined as

$$\langle \tilde{a}_i(f),\, \tilde{b}_i(f)\rangle = 4\Re\int_0^\infty \frac{\tilde{a}_i^*(f)\tilde{b}_i(f)}{S_n^{(i)}(f)}df$$

where $S_n^{(i)}(f)$ is the PSD of the noise in the i^{th} detector.

We use the median-mean variation of Welch's method as described in Ref. [41] to estimate the PSD in each detector. As in Ref. [6], ±256 s of data centered on each simulated signal is broken up into overlapping 8 s segments for this purpose. The O1 LIGO detectors' PSD **grows very rapidly** at frequencies below ∼ 20 Hz. Consequently, we use a lower-frequency cutoff of 20 Hz when generating template waveforms $\tilde{s}(f)$ and evaluating inner products.

As was done in Ref. [6], we use the IMRPhenomPv2 waveform model [42, 43] to generate template waveforms when evaluating the likelihood. We also use this model to generate our simulated signals. IMRPhenomPv2 models the inspiral, merger, and ringdown of the dominant gravitational-wave mode emitted by circular, precessing binary black holes. Due to some simplifying assumptions made in the model, the waveform's dependence on the coalescence phase ϕ can be written as

$$\tilde{s}_i(\mathbf{v}, f, \phi) = \tilde{s}_i^0(\mathbf{v}, f, \phi = 0)e^{i\phi}. \quad (3)$$

It is possible to analytically marginalize over the phase for signals of this form. Substituting Eq. (3) into the likelihood Eq. (2) and marginalizing the posterior over ϕ using a uniform prior yields

$$\log P(\mathbf{v}|\mathcal{D}) \propto \log P(\mathbf{v}) + \log I_0\left[\left|\sum_i O(s_i^0, d_i)\right|\right] - \frac{1}{2}\sum_i\left[\langle s_i^0,\, s_i^0\rangle + \langle d_i,\, d_i\rangle\right], \quad (4)$$

where

$$O(s_i^0, d_i) \equiv 4\int_0^\infty \frac{\tilde{s}_i^*(f; \mathbf{v}, \phi = 0)\tilde{d}_i(f)}{S_n^{(i)}(f)}\mathrm{d}f,$$

and I_0 is the modified Bessel function of the first kind. We use this form of the likelihood function in our analysis, obviating the need to sample over phase. This substantially reduces computational cost, as the phase is a difficult parameter for stochastic samplers to measure.

III. METHODS

A. Obtaining Samples

We use the PyCBC Inference software library to perform our analysis [34]. PyCBC Inference provides a collection of tools for performing Bayesian inference on gravitational waves, as well as doing percentile-percentile tests on simulated signals. It has support for multiple stochastic samplers. In this analysis, we use the parallel-tempered, ensemble Markov-chain Monte Carlo (MCMC) sampler `emcee_pt` [44, 45], with 200 walkers and 20 temperatures.

To ensure that the samples we obtain for the posterior are independent of their initial position of the chains, we use the `max_posterior` and `n_acl` burn-in tests. The `max_posterior` test requires that all chains sample at least one point with a prior-weighted log likelihood within $N_D/2$ of the maximum over all chains, where N_D is the number of dimensions. The `n_acl` test ensures that the length of each chain is greater than 10 times the autocorrelation length, as calculated using the second half of the chain. If so, samples in the second half of the chain are retained. The sampler is considered burned-in at the first iteration **that passes both tests**. Post burn-in samples are thinned by their autocorrelation length so that we only use independent samples to estimate the posterior probability density function. We run the sampler until we obtain 1400 – 2000 independent samples.

B. Credible intervals and the percentile-percentile test

Once the posterior is estimated, we obtain a credible interval on each parameter by marginalizing over all of the other parameters. The $X\%$ credible interval gives the region of parameter space that contains $X\%$ of the marginalized posterior probability of that parameter. In other words, we expect the true value of the parameter to be within the $X\%$ credible interval with $X\%$ probability. Hence, the credible intervals provide a useful way to test whether a parameter estimation analysis is biased.

We perform a Bayesian analysis on a set of simulated events. For each parameter, we plot the fraction of signals in which the true parameter value lies in a credible interval as a function of credible intervals. This is called a percentile-percentile plot. If the recovery is unbiased, we expect the plot for each parameter to follow the $y = x$ line, with some fluctuations due to noise. To quantify the deviations from the expected $y = x$ line, we perform

a Kolmogorov-Smirnov (KS) test. This gives us a two-tailed p-value for each parameter.

The p-value gives the probability of obtaining a percentile-percentile curve at least as extreme as the observed curve under the null hypothesis that the analysis provides an unbiased estimate of the parameter. A very small p-value (lower than a level of significance α) points to an outcome which is very unlikely under the null hypothesis, allowing us to reject the null hypothesis with a maximum Type I error of α. We apply this test to each parameter, yielding 13 p-values for each event. If the analysis is unbiased, we in turn expect these p-values to be uniformly distributed between 0 and 1. We therefore perform another KS-test on the set of p-values to obtain a single "p-value of p-values", by which we can evaluate the probability that the analysis is unbiased. This method was used in Ref. [34] to verify that `emcee_pt` provides an unbiased estimate of gravitational-wave parameters in Gaussian noise, using a prior similar to that used for GW150914.

For each event (GW150914, GW151012 and GW151226), we generate 100 simulated signals ("injections") with parameters drawn from the same mass and spin priors used for that event in Ref. [6]. The coalescence times of the injections are drawn from a prior uniform in an interval of $\pm 256\,\mathrm{s}$ centered on the reported coalescence time of each event. We choose this time window because it was the amount of time used to estimate the PSD for each event in the original analysis [46]. As was done in the original analyses, we use a uniform prior on the coalescence time $t_c \in t_0 + [-0.1, 0.1)\,\mathrm{s}$, where t_0 is GPS time of the simulated signal. Since the prior is centered on signals' injected time, we do not include t_c in the percentile-percentile tests.

The original analysis used a prior uniform in comoving volume to encapsulate the posterior [6]. Instead, we use a prior uniform in $\log_{10}$ of the comoving volume. This is done to avoid having too many injections with low SNR. We choose the bounds such that the majority of the injections have SNRs representative of their actual event while ensuring that the 90% credible interval of the reported distances are contained in the bounds. The distribution of the SNRs of the resulting injections is shown in Fig. 1. The prior bounds used on the source masses and the comoving volume (expressed in terms of the luminosity distance) are uniform over the intervals listed in Table I. For the remaining parameters, we use the priors discussed in Section II.

The amount of time analyzed for each event needs to be large enough to encapsulate the longest-duration waveform allowed by the prior. The in-band waveform duration of binary mergers is approximately inversely proportional to the chirp mass of the binary. For GW150914 and GW151012, we use an analysis duration of 8 seconds around each simulated signal, which is sufficiently long given our low frequency cutoff of 20 Hz. GW151226, being lower mass, requires a longer analysis time. To limit the amount of time that needed to be analyzed, a constraint on the detector-frame chirp mass $\mathcal{M}^{\mathrm{det}} = (m_1^{\mathrm{det}} m_2^{\mathrm{det}})^{3/5}/(m_1^{\mathrm{det}} + m_2^{\mathrm{det}})^{1/5}$ was applied to the GW151226 analysis in both Ref. [6] and in the original publication by the LIGO and Virgo collaborations [39]. We include the same constraint — $\mathcal{M}^{det} \in [8.7, 10.7]\,\mathrm{M}_\odot$ — as used in Ref. [6] here [46]. So as not to bias the percentile-percentile test, this constraint is applied both when analyzing each simulated signal, and when drawing the parameters for the signals. With this constraint in this place, we need only to analyze 12 seconds of data around each signal in the GW151226 percentile-percentile test.

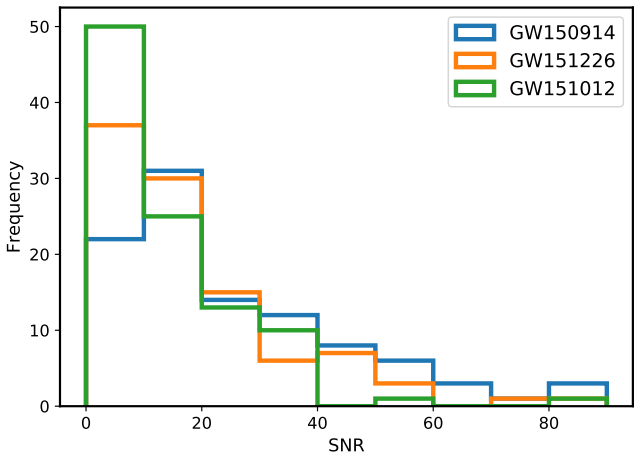


FIG. 1. Distribution of the "optimal" SNRs of the injections used in our analysis. Here, SNR is defined as $\sqrt{\sum_i \langle \tilde{s}_i, \tilde{s}_i \rangle}$, where the sum is over the number of detectors and s_i is the simulated waveform. The bounds on the comoving volume prior are chosen so that majority of the injections have an SNR close to that of the actual event.

We set up two types of runs for each event — one in stationary Gaussian noise and one in real detector noise. In the former, we generate stationary Gaussian noise colored by the PSDs representative of the sensitivity of the LIGO detectors at the time of detection of the respective event. We then add the same simulated signals to the two noise runs and perform a parameter estimation analysis on each signal. We obtain credible intervals on all the parameters and construct the percentile-percentile plot. We also find the p-values of p-values as described in Section III B. **We checked that the presence of the original events in the real detector noise runs do not significantly affect the p-values and their interpretation.**

When analyzing each simulated signal, we re-estimate the PSDs using a window of $\pm 256\,\mathrm{s}$ centered on that signal, as described above. Consequently, our test makes use of up to $\pm 512\,\mathrm{s}$ of real data centered on the original events. Data is downloaded from the Gravitational-wave Open Science Center (GWOSC) [47].

Parameter	GW150914	GW151226	GW151012
m_1 ($M_\odot$)	[10, 80]	[7, 50]	[10, 80]
m_2 ($M_\odot$)	[10, 80]	[3, 15]	[10, 80]
V_C converted to d_L (Mpc)	[50,700]	[150, 700]	[300, 1500]

TABLE I. Source mass and comoving volume priors used in our analysis. The comoving volume is expressed in terms of the luminosity distance.

C. Test on simulated glitches

As a proof of principle, we perform a percentile-percentile test on simulated non-Gaussian noise, which we create by adding glitches to Gaussian noise. We use the same realization of Gaussian noise and the same injections as used for the GW150914 analysis. Glitches are created by using the BayesWave [29] reconstruction of the glitch that occurred in the Livingston detector during GW170817 [22, 30, 48]. We add the transient at random times to our simulated Hanford and Livingston data, with an average rate of one glitch per 16 seconds in each detector. Each glitch is given a random phase offset that is drawn uniformly in $[0, 2\pi)$, and we randomly scale the amplitude of each transient by using a uniform prior in $[0, 1]$ for the scale factor. The glitch times, phase, and amplitude are uncorrelated between the two detectors.

The percentile-percentile plot for the simulated glitch data is shown in Fig. 2. We obtain a p-value of p-values for this analysis of 0.001. Compared to the results using Gaussian noise — see Fig. 3 and the first column of Table II — the non-Gaussian noise is clearly failing the percentile-percentile test, as expected.

IV. RESULTS

The results are summarized in Tables II and III for the Gaussian noise and real noise runs, respectively. Percentile-percentile plots for GW150914 are shown in Fig. 3 and for GW151226 in Fig. 4. The percentile-percentile plot for GW151012 (not shown) is qualitatively similar to GW150914.

For GW150914 and GW151012, we find a p-value of p-values of 0.86 and 0.98 in Gaussian noise, respectively. We therefore find no reason to reject the null hypothesis that the sampler provides an unbiased estimate of the parameters. In real noise, we obtain a p-value of p-values of 0.25 and 0.94 for the two events, respectively. Although the p-value of p-values is lower for GW150914 in real noise, it is not small enough to reject the null hypothesis that the data containing GW150914 is stationary and Gaussian.

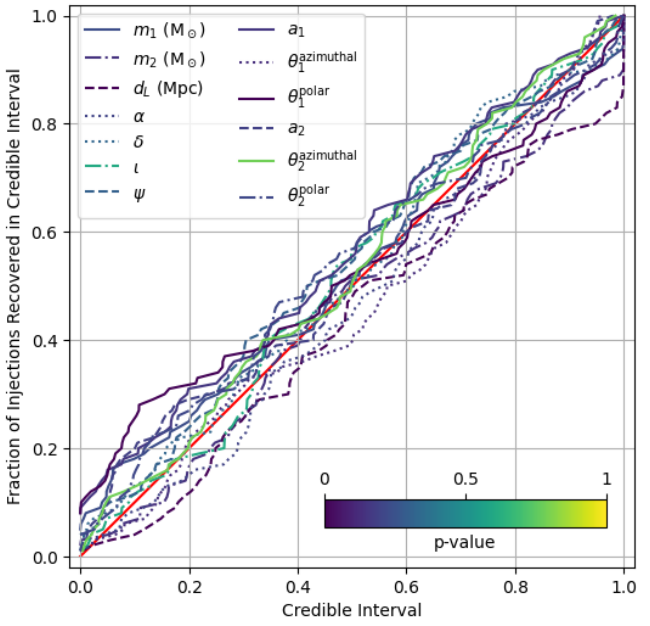


FIG. 2. Percentile-percentile plot for simulated non-Gaussian noise. Shown are the fraction of simulated signals with parameter values recovered in a credible interval as a function of credible intervals for each parameter. If the parameters were recovered exactly, the plot would follow the line $y{=}x$ as indicated by the diagonal line. Simulated glitches were added to the same Gaussian noise as used in Fig 3, and the same simulated signals were used to perform the test. We perform a KS-test to obtain a p-value for each parameter, which is indicated by the color bar. We then perform a KS-test on the set of p-values to obtain a p-value of p-values of 0.001, indicating that the noise fails the percentile-percentile test, as expected.

The results from GW151226 are less clear, however. The p-value of p-values is less than 0.1 in both Gaussian and real noise. This indicates that the `emcee_pt` sampler is not fully converging on the posterior under this prior. Given the low p-values from `emcee_pt`, we decided to try the `dynesty` nested sampler [49] with the GW151226 prior in Gaussian noise.[1] Several previous studies have shown that `dynesty` can produce unbiased parameter estimates of gravitational wave sources [35, 36, 50]. However, in this case, we find that `dynesty` yields even worse results than `emcee_pt`, with a p-value of p-values of 0.001. We did not try to analyze real noise with `dynesty` as a result.

The low-frequency noise in the Livingston detector had larger amplitude during GW151226 than it did during GW150914. In order to test whether the small p-values for GW151226 are due to the different PSD at

[1] Due to the way `dynesty` samples the parameter space it was necessary to remove the constraint on chirp mass when testing it. This meant increasing the analyzed time to 30 seconds, and generating a new set of simulated signals.

those times, we shifted the coalescence times of the injections used in the GW150914 runs to those used in the GW151226 runs. We then perform the percentile-percentile test in Gaussian noise using the same noise realizations as was used to analyze the GW151226-like injections, and with the `emcee_pt` sampler. The resulting percentile-percentile plot is shown in Fig. 5; p-values are reported in the last column of Table II. Doing this, we obtained a p-value of p-values of 0.92. The low p-values for the GW151226-like signals is therefore not due to the changing PSD.

Since the GW150914 injections shifted to the GW151226 times passed the percentile-percentile test in Gaussian noise, we repeat the analysis in the real noise around GW151226, again using the `emcee_pt` sampler. The results are reported in Fig. 5 and the last column of Table III. We obtain a p-value of p-values of 0.31 in this case. We therefore cannot rule out the null hypothesis that the real noise around GW151226 is stationary and Gaussian.

V. SUMMARY AND OUTLOOK

We have performed percentile-percentile tests in both simulated and real noise using the same priors that were used to analyze GW150914, GW151012, and GW151226. Simulated signals were added to a 512 s block of time centered on each event. Comparing percentile-percentile test results from real data to simulated noise, we find no reason to reject the null hypothesis that the detector data was sufficiently stationary and Gaussian at these times to produce unbiased parameter estimates for these events.

We find that both the `emcee_pt` sampler and the `dynesty` nested sampler struggle to produce unbiased parameter estimates for signals similar to that of GW151226. Previous studies have shown that both of these samplers pass percentile-percentile tests for GW150914-like signals [34–36]. In addition, Ref. [50] recently showed that the `dynesty` sampler passes percentile-percentile tests for binary neutron star and neutron-star–black-hole binaries. This suggests that the difficulty with GW151226-like signals is not due to their low mass, or even large mass ratio (Ref. [50] allowed mass ratios up to 10:1). One major difference from our work is that Ref. [50] only considered aligned-spin signals. That reduced the number of parameters involved in their analysis as compared to ours. In addition, spin precession adds more structure to waveforms, leading to a more complicated likelihood topology. This is particularly true of lower-mass and larger mass-ratio signals, which is targeted by the GW151226 prior. However, while our results are suggestive, determining if precession is the primary difficulty for these samplers will require more study.

Regardless of the cause, the poor percentile-percentile test results we obtain for GW151226-like signals with both `emcee_pt` and `dynesty` highlights the need for better stochastic samplers. Even if a sampler is shown to produce unbiased parameter estimates for some region of parameter space, as both these samplers have, it does not mean that the sampler will do so for all parts of parameter space. For this reason, the gravitational-wave community should strive to continually perform these tests as new waveform models and more sensitive detectors become available. The primary hurdle to performing percentile-percentile tests is the computational cost involved. However, new methods for fast likelihood estimation [51], and the ease with which newer inference toolkits [34, 36] can parallelize over many cores make regular percentile-percentile tests more feasible.

We emphasize that our results in real data do not *prove* that the detectors' noise is stationary and Gaussian, only that we have no reason to doubt that they are. It is possible that a non-Gaussian noise component exists in the data during the inspiral and merger of one these events that is simply missed by our analysis, or is not detected due to the statistical nature of our test. Even so, our results give confidence that assuming stationary Gaussian noise has not lead to biased parameter estimates in O1. We plan to extend this test to other detected events in the future.

ACKNOWLEDGEMENTS

We thank Sumit Kumar, Alexander Nitz, and Badri Krishnan for useful suggestions and insights. This work was made possible by the summer intern research program at the AEI Hannover. SK would like to acknowledge the funding provided by DAAD-WISE. We are grateful to the computing team at the AEI Hannover for maintaining the Atlas computer cluster, which was used to carry out all analyses. This research has made use of data obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org), a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. LIGO Laboratory and Advanced LIGO are funded by the United States National Science Foundation (NSF) as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain.

[1] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Class. Quantum Grav. **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].

[2] F. Acernese *et al.* (VIRGO), Advanced Virgo: a second-generation interferometric gravitational wave detector, Class. Quantum Grav. **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].

[3] B. P. Abbott *et al.* (LIGO Scientific, Virgo), 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), arXiv:1811.12907 [astro-ph.HE].

[4] A. H. Nitz, C. Capano, A. B. Nielsen, S. Reyes, R. White, D. A. Brown, and B. Krishnan, 1-OGC: The first open gravitational-wave catalog of binary mergers from analysis of public Advanced LIGO data, Astrophys. J. **872**, 195 (2019), arXiv:1811.01921 [gr-qc].

[5] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, New Binary Black Hole Mergers in the Second Observing Run of Advanced LIGO and Advanced Virgo, (2019), arXiv:1904.07214 [astro-ph.HE].

[6] A. H. Nitz, T. Dent, G. S. Davies, S. Kumar, C. D. Capano, I. Harry, S. Mozzon, L. Nuttall, A. Lundgren, and M. Tpai, 2-OGC: Open Gravitational-wave Catalog of binary mergers from analysis of public Advanced LIGO and Virgo data, Astrophys. J. **891**, 123 (2019), arXiv:1910.05331 [astro-ph.HE].

[7] R. Abbott *et al.* (LIGO Scientific, Virgo), GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run, (2020), arXiv:2010.14527 [gr-qc].

[8] R. Abbott *et al.* (LIGO Scientific, Virgo), Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog, (2020), arXiv:2010.14533 [astro-ph.HE].

[9] S. De, D. Finstad, J. M. Lattimer, D. A. Brown, E. Berger, and C. M. Biwer, Tidal Deformabilities and Radii of Neutron Stars from the Observation of GW170817, Phys. Rev. Lett. **121**, 091102 (2018), [Erratum: Phys.Rev.Lett. 121, 259902 (2018)], arXiv:1804.08583 [astro-ph.HE].

[10] B. Abbott *et al.* (LIGO Scientific, Virgo), GW170817: Measurements of neutron star radii and equation of state, Phys. Rev. Lett. **121**, 161101 (2018), arXiv:1805.11581 [gr-qc].

[11] C. D. Capano, I. Tews, S. M. Brown, B. Margalit, S. De, S. Kumar, D. A. Brown, B. Krishnan, and S. Reddy, Stringent constraints on neutron-star radii from multimessenger observations and nuclear theory, Nature Astron. **4**, 625 (2020), arXiv:1908.10352 [astro-ph.HE].

[12] B. F. Schutz, Determining the Hubble constant from gravitational wave observations, Nature (London) **323**, 310 (1986).

[13] B. Abbott *et al.* (LIGO Scientific, Virgo, 1M2H, Dark Energy Camera GW-E, DES, DLT40, Las Cumbres Observatory, VINROUGE, MASTER), A gravitational-wave standard siren measurement of the Hubble constant, Nature **551**, 85 (2017), arXiv:1710.05835 [astro-ph.CO].

[14] B. Abbott *et al.* (LIGO Scientific, Virgo), A gravitational-wave measurement of the Hubble constant following the second observing run of Advanced LIGO and Virgo, (2019), arXiv:1908.06060 [astro-ph.CO].

[15] R. Abbott *et al.* (LIGO Scientific, Virgo), Tests of General Relativity with Binary Black Holes from the second LIGO-Virgo Gravitational-Wave Transient Catalog, (2020), arXiv:2010.14529 [gr-qc].

[16] B. Abbott *et al.* (LIGO Scientific, Virgo), Tests of General Relativity with the Binary Black Hole Signals from the LIGO-Virgo Catalog GWTC-1, Phys. Rev. D **100**, 104036 (2019), arXiv:1903.04467 [gr-qc].

[17] B. Abbott *et al.* (LIGO Scientific, Virgo), Tests of General Relativity with GW170817, Phys. Rev. Lett. **123**, 011102 (2019), arXiv:1811.00364 [gr-qc].

[18] B. Abbott *et al.* (LIGO Scientific, Virgo), Tests of general relativity with GW150914, Phys. Rev. Lett. **116**, 221101 (2016), [Erratum: Phys.Rev.Lett. 121, 129902 (2018)], arXiv:1602.03841 [gr-qc].

[19] L. K. Nuttall *et al.*, Improving the Data Quality of Advanced LIGO Based on Early Engineering Run Results, Class. Quantum Grav. **32**, 245005 (2015), arXiv:1508.07316 [gr-qc].

[20] B. Abbott *et al.* (LIGO Scientific, Virgo), Characterization of transient noise in Advanced LIGO relevant to gravitational wave signal GW150914, Class. Quant. Grav. **33**, 134001 (2016), arXiv:1602.03844 [gr-qc].

[21] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), Effects of data quality vetoes on a search for compact binary coalescences in Advanced LIGOs first observing run, Class. Quant. Grav. **35**, 065010 (2018), arXiv:1710.02185 [gr-qc].

[22] B. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, Phys. Rev. Lett. **119**, 161101 (2017), arXiv:1710.05832 [gr-qc].

[23] M. Cabero *et al.*, Blip glitches in Advanced LIGO data, Class. Quant. Grav. **36**, 155010 (2019), arXiv:1901.05093 [physics.ins-det].

[24] B. Allen, A chi**2 time-frequency discriminator for gravitational wave detection, Phys. Rev. D **71**, 062001 (2005), arXiv:gr-qc/0405045 [gr-qc].

[25] A. H. Nitz, T. Dent, T. Dal Canton, S. Fairhurst, and D. A. Brown, Detecting binary compact-object mergers with gravitational waves: Understanding and Improving the sensitivity of the PyCBC search, Astrophys. J. **849**, 118 (2017), arXiv:1705.01513 [gr-qc].

[26] S. A. Usman *et al.*, The PyCBC search for gravitational waves from compact binary coalescence, Class. Quant. Grav. **33**, 215004 (2016), arXiv:1508.02357 [gr-qc].

[27] S. Sachdev, S. Caudill, H. Fong, R. K. L. Lo, C. Messick, D. Mukherjee, R. Magee, L. Tsukada, K. Blackburn, P. Brady, P. Brockill, K. Cannon, S. J. Chamberlin, D. Chatterjee, J. D. E. Creighton, P. Godwin, A. Gupta, C. Hanna, S. Kapadia, R. N. Lang, T. G. F. Li, D. Meacher, A. Pace, S. Privitera, L. Sadeghian, L. Wade, M. Wade, A. Weinstein, and S. L. Xiao, The gstlal search analysis methods for compact binary mergers in advanced ligo's second and advanced virgo's first observing runs, (2019), arXiv:1901.08580 [gr-qc].

[28] B. P. Abbott *et al.* (GROND, SALT Group, OzGrav, DFN, INTEGRAL, Virgo, Insight-Hxmt, MAXI Team, Fermi-LAT, J-GEM, RATIR, IceCube, CAAS-

TRO, LWA, ePESSTO, GRAWITA, RIMAS, SKA South Africa/MeerKAT, H.E.S.S., 1M2H Team, IKI-GW Follow-up, Fermi GBM, Pi of Sky, DWF (Deeper Wider Faster Program), Dark Energy Survey, MASTER, AstroSat Cadmium Zinc Telluride Imager Team, Swift, Pierre Auger, ASKAP, VINROUGE, JAGWAR, Chandra Team at McGill University, TTU-NRAO, GROWTH, AGILE Team, MWA, ATCA, AST3, TOROS, Pan-STARRS, NuSTAR, ATLAS Telescopes, BOOTES, CaltechNRAO, LIGO Scientific, High Time Resolution Universe Survey, Nordic Optical Telescope, Las Cumbres Observatory Group, TZAC Consortium, LOFAR, IPN, DLT40, Texas Tech University, HAWC, ANTARES, KU, Dark Energy Camera GW-EM, CALET, Euro VLBI Team, ALMA), Multi-messenger Observations of a Binary Neutron Star Merger, Astrophys. J. **848**, L12 (2017), arXiv:1710.05833 [astro-ph.HE].

[29] N. J. Cornish and T. B. Littenberg, BayesWave: Bayesian Inference for Gravitational Wave Bursts and Instrument Glitches, Class. Quant. Grav. **32**, 135012 (2015), arXiv:1410.3835 [gr-qc].

[30] C. Pankow *et al.*, Mitigation of the instrumental noise transient in gravitational-wave data surrounding GW170817, Phys. Rev. D **98**, 084016 (2018), arXiv:1808.03619 [gr-qc].

[31] T. Sidery *et al.*, Reconstructing the sky location of gravitational-wave detected compact binary systems: methodology for testing and comparison, Phys. Rev. D **89**, 084060 (2014), arXiv:1312.6013 [astro-ph.IM].

[32] J. Veitch, V. Raymond, B. Farr, W. Farr, P. Graff, S. Vitale, B. Aylott, K. Blackburn, N. Christensen, M. Coughlin, W. Del Pozzo, F. Feroz, J. Gair, C.-J. Haster, V. Kalogera, T. Littenberg, I. Mandel, R. O'Shaughnessy, M. Pitkin, C. Rodriguez, C. Röver, T. Sidery, R. Smith, M. Van Der Sluys, A. Vecchio, W. Vousden, and L. Wade, Robust parameter estimation for compact binaries with ground-based gravitational-wave observations using the lalinference software library, Phys. Rev. D **91**, 042003 (2015), arXiv:1409.7215 [gr-qc].

[33] C. P. L. Berry, I. Mandel, H. Middleton, L. P. Singer, A. L. Urban, A. Vecchio, S. Vitale, K. Cannon, B. Farr, W. M. Farr, P. B. Graff, C. Hanna, C.-J. Haster, S. Mohapatra, C. Pankow, L. R. Price, T. Sidery, and J. Veitch, Parameter estimation for binary neutron-star coalescences with realistic noise during the advanced ligo era, The Astrophysical Journal **804**, 114 (2015), arXiv:1411.6934 [astro-ph.HE].

[34] C. M. Biwer, C. D. Capano, S. De, M. Cabero, D. A. Brown, A. H. Nitz, and V. Raymond, PyCBC Inference: A Python-based parameter estimation toolkit for compact binary coalescence signals, Publ. Astron. Soc. Pac. **131**, 024503 (2019), arXiv:1807.10312 [astro-ph.IM].

[35] R. J. Smith, G. Ashton, A. Vajpeyi, and C. Talbot, Massively parallel Bayesian inference for transient gravitational-wave astronomy, Mon. Not. Roy. Astron. Soc. **498**, 4492 (2020), arXiv:1909.11873 [gr-qc].

[36] I. Romero-Shaw *et al.*, Bayesian inference for compact binary coalescences with BILBY: Validation and application to the first LIGO–Virgo gravitational-wave transient catalogue 10.1093/mnras/staa2850 (2020), arXiv:2006.00714 [astro-ph.IM].

[37] B. P. Abbott *et al.* (Virgo, LIGO Scientific), Observation of Gravitational Waves from a Binary Black Hole Merger, Phys. Rev. Lett. **116**, 061102 (2016), arXiv:1602.03837 [gr-qc].

[38] B. P. Abbott *et al.* (Virgo, LIGO Scientific), Binary Black Hole Mergers in the first Advanced LIGO Observing Run, Phys. Rev. X **6**, 041015 (2016), [erratum: Phys. Rev.X8,no.3,039903(2018)], arXiv:1606.04856 [gr-qc].

[39] B. P. Abbott *et al.* (Virgo, LIGO Scientific), GW151226: Observation of Gravitational Waves from a 22-Solar-Mass Binary Black Hole Coalescence, Phys. Rev. Lett. **116**, 241103 (2016), arXiv:1606.04855 [gr-qc].

[40] P. A. R. Ade *et al.* (Planck Collabortion), Planck 2015 results. XIII. Cosmological parameters, (2015), arXiv:1502.01589 [astro-ph.CO].

[41] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An Algorithm for detection of gravitational waves from inspiraling compact binaries, Phys. Rev. D **85**, 122006 (2012), arXiv:0509116 [gr-qc].

[42] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Simple Model of Complete Precessing Black-Hole-Binary Gravitational Waveforms, Phys. Rev. Lett. **113**, 151101 (2014), arXiv:1308.3271 [gr-qc].

[43] P. Schmidt, F. Ohme, and M. Hannam, Towards models of gravitational waveforms from generic binaries II: Modelling precession effects with a single effective precession parameter, Phys. Rev. D **91**, 024043 (2015), arXiv:1408.1810 [gr-qc].

[44] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, emcee: The MCMC Hammer, Publ. Astron. Soc. Pac. **125**, 306 (2013), arXiv:1202.3665 [astro-ph.IM].

[45] W. D. Vousden, W. M. Farr, and I. Mandel, Dynamic temperature selection for parallel tempering in Markov chain Monte Carlo simulations, Monthly Notices of the Royal Astronomical Society **455**, 1919 (2015), https://academic.oup.com/mnras/article-pdf/455/2/1919/18514064/stv2422.pdf.

[46] A. H. Nitz and C. e. a. Capano, 2-OGC Open Gravitational-wave Catalog, `www.github.com/gwastro/2-ogc` (2019).

[47] R. Abbott *et al.* (LIGO Scientific, Virgo), Open data from the first and second observing runs of Advanced LIGO and Advanced Virgo, (2019), arXiv:1912.11716 [gr-qc].

[48] t. . B. h. . `https://dcc.ligo.org/LIGO-T1700406-v3/public`. LIGO Scientific Collaboration, year = "2018",.

[49] J. S. Speagle, dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences, Monthly Notices of the Royal Astronomical Society **493**, 3132 (2020), https://academic.oup.com/mnras/article-pdf/493/3/3132/32890730/staa278.pdf.

[50] D. Finstad and D. A. Brown, Fast Parameter Estimation of Binary Mergers for Multimessenger Followup, (2020), arXiv:2009.13759 [astro-ph.IM].

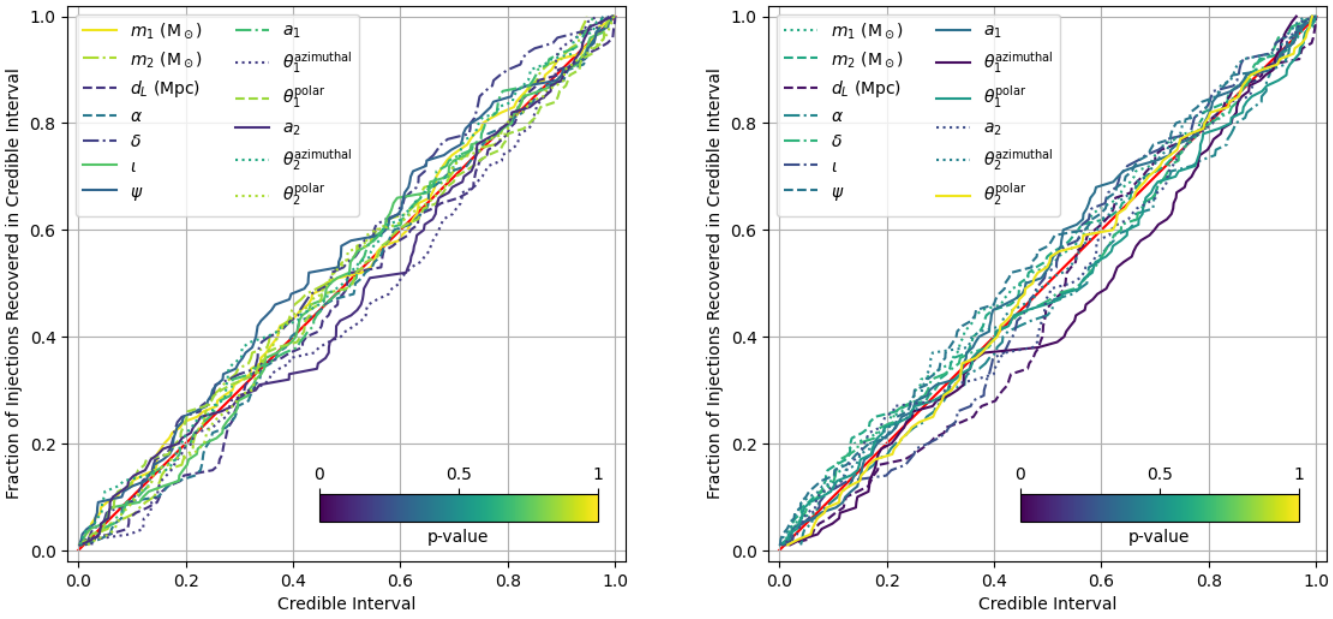
[51] B. Zackay, L. Dai, and T. Venumadhav, Relative Binning and Fast Likelihood Evaluation for Gravitational Wave Parameter Estimation, (2018), arXiv:1806.08792 [astro-ph.IM].

Parameter	GW150914 prior	GW151226 prior	GW151012 prior	GW150914 prior t_c shifted
m_1 ($M_\odot$)	0.975	0.314	0.814	0.137
m_2 ($M_\odot$)	0.876	0.850	0.483	0.990
d_L (Mpc)	0.154	0.377	0.032	0.248
α	0.405	0.076	0.701	0.103
δ	0.192	0.001	0.342	0.990
ι	0.731	0.076	0.122	0.567
ψ	0.326	0.238	0.268	0.296
a_1	0.693	0.518	0.419	0.567
$\theta_1^{\rm azimuthal}$	0.176	0.064	0.693	0.176
$\theta_1^{\rm polar}$	0.850	0.333	0.966	0.659
a_2	0.134	0.065	0.829	0.359
$\theta_2^{\rm azimuthal}$	0.609	0.487	0.021	0.775
$\theta_2^{\rm polar}$	0.873	0.184	0.526	0.775
p-values of p-values	0.855	0.019	0.978	0.915

TABLE II. P-values obtained for each parameter when recovered from Gaussian noise colored by the power-spectral densities representative of the detectors at the time of the respective events. The last column corresponds to results from the injections used in the GW150914-like run shifted to the coalescence times used in the GW151226-like injections. We perform a KS-test comparing the set of p-values obtained for each event to a uniform distribution and report the p-value of p-values in the last row.

Parameter	GW150914 prior	GW151226 prior	GW151012 prior	GW150914 prior t_c shifted
m_1 ($M_\odot$)	0.609	0.420	0.901	0.253
m_2 ($M_\odot$)	0.600	0.351	0.722	0.584
d_L (Mpc)	0.054	0.290	0.282	0.405
α	0.464	0.619	0.169	0.357
δ	0.651	0.668	0.487	0.659
ι	0.243	0.101	0.051	0.651
ψ	0.398	0.219	0.276	0.398
a_1	0.371	0.501	0.973	0.668
$\theta_1^{\rm azimuthal}$	0.035	0.434	0.494	0.089
$\theta_1^{\rm polar}$	0.542	0.356	0.676	0.449
a_2	0.233	0.078	0.802	0.405
$\theta_2^{\rm azimuthal}$	0.449	0.332	0.034	0.419
$\theta_2^{\rm polar}$	0.975	0.071	0.827	0.895
p-values of p-values	0.242	0.069	0.938	0.310

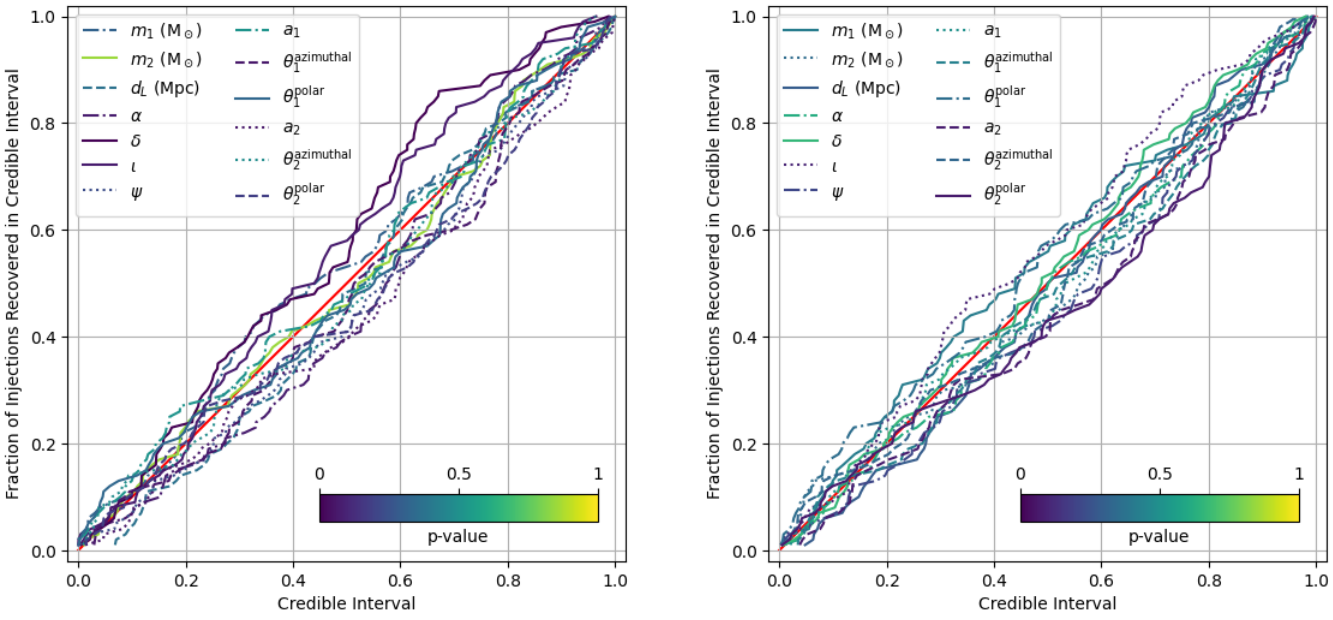
TABLE III. P-values obtained for each parameter when recovered from the actual detector noise at the time of the respective events. The last column corresponds to results from the injections used in the GW150914-like run shifted to the coalescence times used in the GW151226-like injections. We perform a KS-test comparing the set of p-values obtained for each event to a uniform distribution and report the p-value of p-values in the last row.



a) Gaussian Noise

b) Real Noise

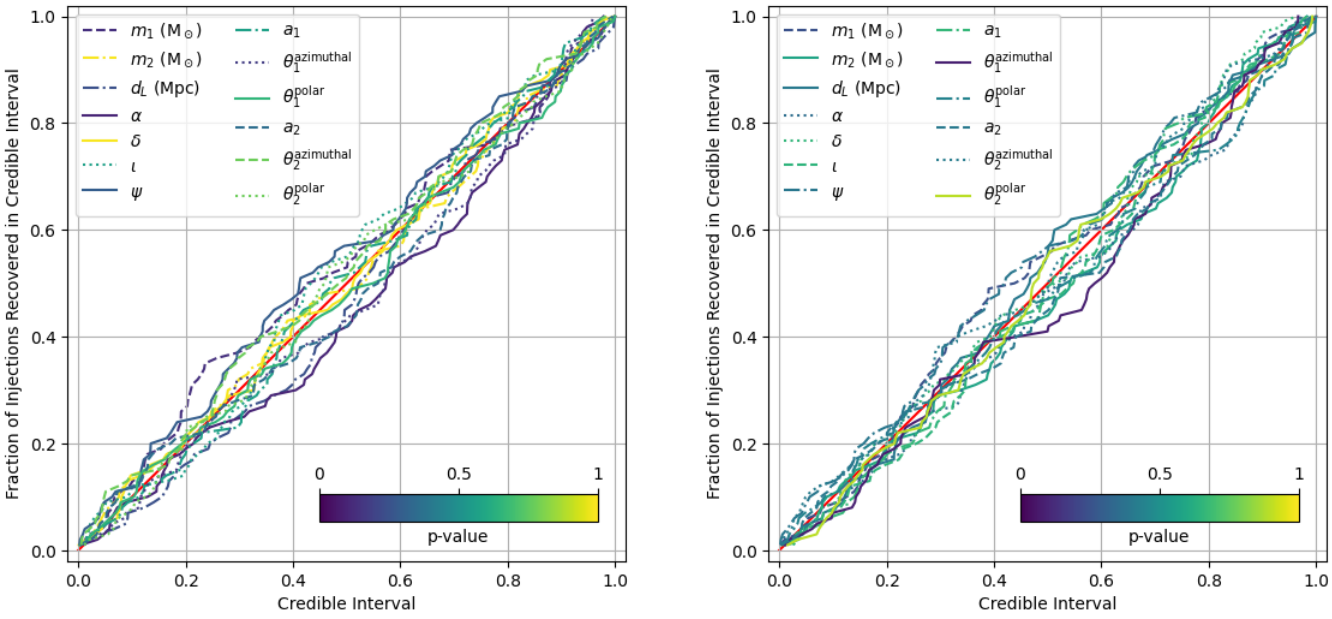
FIG. 3. **GW150914:** Plot of the fraction of simulated signals with parameter values recovered in a credible interval as a function of credible intervals for each parameter. If the parameters were recovered exactly, the plot would follow the line $y{=}x$ as indicated by the diagonal line. For each parameter, a KS test is performed between the recovered curve and the diagonal line to obtain a two-tailed p-value, **which is indicated by the color bar.**.



a) Gaussian Noise

b) Real Noise

FIG. 4. **GW151226:** Plot of the fraction of simulated signals with parameter values recovered in a credible interval as a function of credible intervals for each parameter. If the parameters were recovered exactly, the plot would follow the line $y{=}x$ as indicated by the diagonal line. For each parameter, a KS test is performed between the recovered curve and the diagonal line to obtain a two-tailed p-value, **which is indicated by the color bar.**



a) Gaussian Noise

b) Real Noise

FIG. 5. **GW150914 Shifted:** Plot of the fraction of simulated signals with parameter values recovered in a credible interval as a function of credible intervals for each parameter. If the parameters were recovered exactly, the plot would follow the line $y{=}x$ as indicated by the diagonal line. For each parameter, a KS test is performed between the recovered curve and the diagonal line to obtain a two-tailed p-value, **which is indicated by the color bar.**