Calibrating approximate Bayesian credible intervals of gravitational-wave parameters

Ruiting Mao,[1] Jeong Eun Lee,[1] Ollie Burke,[2] Alvin J. K. Chua,[3,4,*] Matthew C. Edwards,[1] and Renate Meyer[1]

[1]*Department of Statistics, University of Auckland, Auckland 1010, New Zealand*
[2]*Laboratoire des 2 Infinis—Toulouse (L2IT-IN2P3), Université de Toulouse, CNRS, UPS, F-31062 Toulouse Cedex 9, France*
[3]*Department of Physics, National University of Singapore, Singapore 117551*
[4]*Department of Mathematics, National University of Singapore, Singapore 119076*

(Received 9 October 2023; accepted 27 February 2024; published 1 April 2024)

Approximations are commonly employed in realistic applications of scientific Bayesian inference, often due to convenience if not necessity. In the field of gravitational-wave (GW) data analysis, fast-to-evaluate but approximate waveform models of astrophysical GW signals are sometimes used in lieu of more accurate models to infer properties of a true GW signal buried within detector noise. In addition, a Fisher-information-based normal approximation to the posterior distribution can also be used to conduct inference in bulk, without the need for extensive numerical calculations such as Markov chain Monte Carlo simulations. Such approximations can generally lead to an inaccurate posterior distribution with poor statistical coverage of the true posterior. In this paper, we present a novel calibration procedure that calibrates the credible sets for a family of approximate posterior distributions, to ensure coverage of the true posterior at a level specified by the analyst. Tools such as autoencoders and artificial neural networks are used within our calibration model to compress the data (for efficiency) and to perform tasks such as logistic regression. As a proof of principle, we demonstrate our formalism on the GW signal from a high-mass binary black hole merger, a promising source for the near-future space-based GW observatory LISA.

DOI: 10.1103/PhysRevD.109.083002

I. INTRODUCTION

The ground-breaking observation of a gravitational-wave (GW) signal spectacularly opened the field of gravitational-wave astronomy on the 14th of September, 2015. The ground-based gravitational-wave observatories, the Laser Interferometer Gravitational-Wave Observatory (LIGO) in Livingston and Hanford, observed a GW signal emanating from the coalescence of two stellar-mass black holes (BHs) [1]. This feat was achieved through the use of sophisticated statistical signal processing algorithms *and* accurate waveform templates used to filter the data stream [2–4]. In a traditional (ground-based) matched-filtering search, template banks are used to detect the presence of a signal buried within the instrumental noise [5–8]. Once a candidate signal in the data stream is established, stochastic sampling algorithms, such as Markov chain Monte Carlo (MCMC), are used to estimate the parameter set that best describes the corresponding astrophysical source [9–11]. To do this both efficiently and accurately, we require astrophysical waveform models to be both cheap to generate and sufficiently detailed to describe the fully relativistic waveform in the data stream [12].

The current state-of-the-art waveform modeling technique for binary black holes with mass ratios $\lesssim 20$ is numerical relativity (NR), where the fully general Einstein equations are numerically solved for the space-time metric perturbations [13–16]. These NR simulations are the most accurate method to date we have to generate comparable mass BHs, but can take months to generate a small number of orbits [17,18]. For data analysis, which relies on generating hundreds of thousands of waveforms with multiple cycles, this is computationally infeasible. In order to circumvent this, various waveform approximations have been developed that rely on a hybrid between the post-Newtonian (PN) formalism and NR [19–24]. This has the major advantage that these waveform models are faster to generate, but with the cost that they are not truly faithful to the GW signal hidden within the data stream. Modeling errors can result in an overall reduced sensitivity in the detection of actual signals. Using nonfaithful waveforms can also impact inference: we may potentially recover biased parameter estimates, and/or claim incorrectly how precise we can measure the parameters in question [12,25,26]. Additionally, it can be shown that biases arising from approximate waveform models scale inversely in the limit of high signal-to-noise ratio (SNR) [26]. Further, waveform inaccuracies could result in systematic errors in

*alvincjk@nus.edu.sg

parameter estimates and even dominate them, especially with multiple overlapping signals [27–30].

In our work, we will restrict our attention to a particular class of sources expected to be observed by the space-based gravitational-wave observatory: the Laser Interferometer Space Antenna (LISA) [31,32]. In contrast to ground-based detectors, which are limited by seismic noise at lower frequencies, the LISA instrument will achieve optimum sensitivity in the mHz GW frequency band, providing the means to probe the rich structure of high mass binaries. One of the most promising sources of mHz gravitational radiation will be the collisions of comparable massive black hole binaries (MBHBs) with masses of 10^5–$10^7 M_\odot$ up to redshifts of $z \lesssim 20$. Observation of these MBHBs at specific redshifts gives one a means to probe various formation channels for MBHBs [33–36]. Unlike sources observed by ground-based detectors, MBHBs will be extremely loud, with SNRs up to ∼1000, offering strong constraints on the parameters that govern the signal [37,38]. Owing to the strength of these signals and the precision in which we can measure their parameters, they will provide powerful tests of general relativity (GR) [39–41].

With the immense strength of these signals, we will require exceptionally accurate waveform models to ensure that recovered parameters are not biased and that uncertainties are correctly quantified. For this reason, calibration techniques that provide a means to "correct" inference resulting from an approximate waveform model to an exact waveform model (such as a NR simulation) may be viewed as essential. An early example of this was presented in [12,25,26], which showed how to estimate the bias (and thus the accurate maximum *a posteriori* estimate) in the linear-signal regime. More recently, Gaussian process regression has been developed as a viable method for interpolating and marginalizing over model error in the GW likelihood, thus calibrating the likelihood itself before it is used for posterior estimation [42–45]. Similar studies can also be found with ground-based detectors, like LIGO and Virgo, as well as preparations for Cosmic Explorer and Einstein Telescope. Early Fisher studies, with an approximation to the full likelihood, may not be applicable for low SNR events. Full analyses into the full parameter space were needed when correcting the model uncertainties. Signal-specific calibration with marginalization in gravitational-wave inference was implemented [46]. The application of the Bayesian method was also proposed to marginalize the ignorance of (unknown) higher PN order terms and give general directive calibrations [47].

When the generative model lacks computational efficiency for executing the MCMC simulation, resorting to posterior density approximation becomes an appealing option, and the MCMC simulation can be avoided. This is often seen in Bayesian inference, and some well-known approximations include the mean field approximation in the variational Bayes [48] and Laplace approximation [49,50]. In this paper, the posterior density is normally approximated by imposing the Fisher matrix approach to the likelihood and the uniform prior. The Fisher matrix for the GW likelihood can be expressed in terms of waveform derivatives of the approximate model (which may need to be computed numerically, but using far fewer evaluations of the model than posterior sampling). It is widely used throughout GW astronomy to cheaply forecast precision statements on parameters [51] and to predict biases on parameter estimates through the use of nonfaithful waveform models [26]. As the Fisher matrix lends itself to calculations in bulk, it is often also used to approximate and hence study a family of posteriors over a space of GW signals [52–55]. Thus there is plenty of motivation for methods that can calibrate the family of approximate normal posteriors obtained via the Fisher matrix.

In this paper, we introduce a novel calibration technique that approaches the problem from another angle, using the formalism presented by [56]. In the Bayesian framework, the uncertainty about unknowns is probabilistically represented by the posterior distribution. When an approximate waveform generation model is used or a likelihood/posterior is approximated, the resulting inference is not exact. *Operational coverage* of a credible set of a parameter vector based on an approximate posterior measures how much of the exact posterior probability mass lies in this set, and it can be interpreted as an error estimate for the approximate posterior. A practical operational coverage estimator allows us to estimate an error of posterior approximation for any observed waveform generated from a prior distribution [56,57]. To perform the calibration formalism, it is necessary to generate a large number of posteriors using an approximate waveform model. This would clearly become computationally prohibitive with expensive posterior simulation methods. Owing to the usage of normal approximations via a Fisher-based formalism to the posteriors, expensive posterior simulation is not required, and the practical operational coverage estimator is acquired with a more budget-friendly computing cost.

Here, we present a practical estimator for gravitational-wave data analysis and demonstrate how to compute the calibrated credible set of an approximate posterior that corresponds to the desired exact posterior probability mass. Systematic studies usually focus on correcting a *single* posterior, generated via an approximate waveform model at a specific point in parameter space [28]. Instead, we propose a method that, after a training scheme (on the prior space of samples), near-instantaneous calibrated posterior estimates can be generated over the entire signal space. In other words, we devise a scheme that can calibrate a *family* of posteriors, rather than a single one.

This paper is organized as follows. In Sec. II we set notations and introduce the data analysis concepts that will be used throughout this work. In Sec. III, we summarize the

work of [56], outlining a framework that can be used to calibrate the statistical coverage of approximate posterior distributions to the exact posterior distribution as if parameter estimation was performed using a more faithful waveform model. In Sec. IV we demonstrate this calibration procedure using a simple toy example and finally show its general applicability on an MBHB source within the LISA framework in Sec. V. Our conclusions and scope for future work are presented in Sec. VI.

II. GRAVITATIONAL-WAVE DATA ANALYSIS

A. Noise modeling and likelihood

The typical time-domain data stream observed by the LISA instrument will be a combination of time delay interferometry (TDI) variables $X = \{A, E, T\}$, representing the response of the LISA instrument to the plus and cross polarizations of the incoming GW source in the transverse-traceless gauge [58,59]:

$$d_o^{(X)}(t) = h_{\rm e}^{(X)}(t;\boldsymbol{\theta}_0) + n^{(X)}(t), \qquad X = \{A, E, T\}. \tag{1}$$

Here d_o is the observed data stream, $\boldsymbol{\theta}_0$ are the true parameters of the true gravitational wave $h_{\rm e}^{(X)}$, and $n^{(X)}(t)$ are noise fluctuations arising from perturbations to the LISA instrument from unresolvable GW sources and non-GW instrumental perturbations. In our work, we will perform inference on a *single* waveform within the data streams $d^{(X)}$ and ignore potential multiple signals within the data stream, such as would be considered in the global fit [60]. We make the assumption that the noise $n^{(X)}$ in each channel is a weakly stationary Gaussian stochastic process with zero mean, colored by the power spectral density (PSD) of their respected TDI channel. A consequence of this is that the noise $n^{(X)}$ is uncorrelated in the frequency domain, resulting in a purely diagonal noise covariance matrix Σ [61–63]:

$$\Sigma(f, f') = \langle \hat{n}^{(X)}(f)(\hat{n}^{(X)}(f'))^\star \rangle \tag{2}$$

$$= \frac{1}{2}\delta(f - f')S_n^{(X)}(f'). \tag{3}$$

for $f \in (0, \infty)$. Here $\langle\cdot\rangle$ denotes an average ensemble over many noise realisations, δ is the Dirac delta function and $S_n^{(X)}$ is the PSD of the noise process within a channel X. Hatted quantities refer to the Fourier transform with convention,

$$\hat{h}(f) = \int_0^\infty {\rm d}t\, h(t) \exp(-2\pi {\rm i} f t). \tag{4}$$

Assuming that the arm lengths of the LISA interferometer are both equal and constant, it can be shown that the noise across channels X is independent and thus uncorrelated [58,59]. From Eq. (3), Whittle showed that the likelihood in the frequency domain takes the form [64]

$$p(n) = -\frac{1}{2}\sum_X (n|n)^{(X)} \tag{5}$$

with inner product [11,25]

$$(a|b)^{(X)} = 4{\rm Re}\int_0^\infty {\rm d}f \frac{\hat{a}^{(X)}(f)(\hat{b}^{(X)}(f'))^\star}{S_n^{(X)}(f')}. \tag{6}$$

Substituting equation (1) into (5), we obtain the usual likelihood used throughout gravitational-wave astronomy [10,11,63],

$$p(d|\boldsymbol{\theta}) = -\frac{1}{2}\sum_X (d - h_{\rm m}|d - h_{\rm m})^{(X)}, \tag{7}$$

where $h_{\rm m}$ are our approximate model templates, favorably quick to generate and used when inferring parameters $\boldsymbol{\theta}$.

The SNR, ρ, is a quantity used to determine the power of the signal when compared to noise. Within the framework of matched filtering [65,66], the optimal matched filtering SNR takes the form [11]

$$\rho_X^2 = (h_m|h_m)^{(X)} = 4\int_0^\infty {\rm d}f \frac{|\hat{h}(f)|^2}{S_n^{(X)}(f)}, \tag{8}$$

with total (squared) SNR across $X = \{A, E, T\}$ given by summing Eq. (8) in quadrature $\rho^2 = \sum_X (h_{\rm m}|h_{\rm m})^{(X)}$.

We now describe how we generate detector noise given that the noise is both stationary and we know, *a priori*, the PSD of the noise process $n^{(X)}$.

The frequency domain equation (2) can be discretized in the continuum limit to give the covariance of the noise between two frequency bins f_i and f_j:

$$\hat{\Sigma}_{ij} = \mathbb{E}_d[\hat{n}^{(X)}(f_i)(\hat{n}^{(X)}(f_j))^\star] \tag{9}$$

$$= S_n^{(X)}(f_i)\delta_{ij}/2\Delta f. \tag{10}$$

Here $f_i \in [0, \Delta f, \ldots, (\frac{N}{2})\Delta f]$ is an individual frequency bin, $\Delta f = 1/N\Delta t = 1/T_{\rm obs}$ the spacing between frequency bins, N the length of the time series, and Δt the sampling interval.

Equation (10) highlights that, for stationary Gaussian noise, the frequency bins for $i \neq j$ are uncorrelated. Focusing on the diagonal elements of (10), it is possible to show that the real and imaginary components of the noise follows a Gaussian distribution:

$${\rm Re}(\hat{n}^{(X)}(f_i) = N\left(0, \frac{S_n^{(X)}(f_i)}{4\Delta f}\right) \tag{11a}$$

$$\mathrm{Im}(\hat{n}^{(X)}(f_i) = N\left(0, \frac{S_n^{(X)}(f_i)}{4\Delta f}\right). \tag{11b}$$

To simulate noise, we thus draw components of the noise from Eqs. (11a) and (11b) given PSDs $S_n^{(X)}$ for each of the channels $X = \{A, E, T\}$. An exact signal is then generated, added to this specific noise realization, and this constructs the dataset $d^{(X)}(t)$ in (1).

As will be discussed in Sec. II B, assuming that the likelihood is consistent with the noise model, the detector noise will encode a statistical fluctuation forcing a deviation between the recovered and true parameters. In reality, the recovered parameters will *not* be centred on the true parameters due to the presence of two features: waveform modeling errors and noise. The next section describes how one can compute the bias on parameters due to waveform modeling errors and statistical fluctuations due to the inclusion of noise.

B. Fisher matrix

In our work we will use a Fisher matrix formalism to generate approximate distributions on parameters given an observed data stream $p(\boldsymbol{\theta}|d_o)$ [11,51]. In the high SNR limit, it is expected that the distribution of parameter sets is a multivariate Gaussian defined by a mean vector and covariance matrix. Here we will show how one can approximate both the mean vector and covariance matrix using a Fisher matrix approach, rather than using costly MCMC simulations. Here we generalize the results from [12,25,26] to account for the three LISA channels A, E, and T.

We denote the best fit parameter $\theta^i_{\rm bf}$ as the *maximum likelihood estimate* that maximizes Eq. (7):

$$\sum_{X=\{A,E,T\}} (\partial_i h_m^{(X)}(t;\boldsymbol{\theta}_{\rm bf})|d^{(X)} - h_m^{(X)}(t;\boldsymbol{\theta}_{\rm bf})) = 0. \tag{12}$$

Here $\partial_i \coloneqq \partial/\partial\theta^i$ denotes a partial derivative with respect to θ^i. Now consider a small perturbation around the true parameters $\boldsymbol{\theta}_{\rm bf} = \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}$ for $\Delta\boldsymbol{\theta} = (\boldsymbol{\theta}_{\rm bf} - \boldsymbol{\theta}_0)$. By applying the linear signal approximation, an expansion in $\Delta\boldsymbol{\theta} \ll 1$ to first order in our model templates gives

$$h_m^{(X)}(\boldsymbol{\theta}_0) \approx h_m^{(X)}(\boldsymbol{\theta}_{\rm bf}) - \partial_i h_m^{(X)}(\boldsymbol{\theta}_{\rm bf})\Delta\theta^i. \tag{13}$$

From this point onward, we will drop the (fixed) time coordinate t for notational convenience. Equation (13) can then be used in the expression $d^{(X)} - h_m^{(X)}$ to find

$$\begin{aligned} &d^{(X)} - h_m^{(X)}(\boldsymbol{\theta}_{\rm bf}) \\ &= n^{(X)} + h_e^{(X)}(\boldsymbol{\theta}_0) - h_m^{(X)}(\boldsymbol{\theta}_{\rm bf}) \\ &= n^{(X)} + h_e^{(X)}(\boldsymbol{\theta}_0) - h_m^{(X)}(\boldsymbol{\theta}_0) + h_m^{(X)}(\boldsymbol{\theta}_0) - h_m^{(X)}(\boldsymbol{\theta}_{\rm bf}) \\ &\approx n^{(X)} + \delta h^{(X)}(\boldsymbol{\theta}_0) - \partial_i h_m^{(X)}(\boldsymbol{\theta}_{\rm bf})\Delta\theta^i + (\Delta\theta^i)^2 \end{aligned} \tag{14}$$

for $\delta h^{(X)}(\boldsymbol{\theta}_0) = h_e^{(X)}(\boldsymbol{\theta}_0) - h_m^{(X)}(\boldsymbol{\theta}_0)$ denoting residuals between the true waveform $h_e^{(X)}$ and the approximate waveform $h_m^{(X)}$. When there are no mismodeling errors present, the term $\delta h^{(X)} = 0$. Substituting Eq. (14) into (12), one obtains at first order in $\Delta\theta^i$

$$\sum_{X=\{A,E,T\}} [\Delta\theta^j(\partial_j h_m(\boldsymbol{\theta}_{\rm bf})|\partial_i h_m(\boldsymbol{\theta}_{\rm bf}))^{(X)} - (\partial_j h_m(\boldsymbol{\theta}_{\rm bf})|n)^{(X)} - (\partial_j h_m(\boldsymbol{\theta}_{\rm bf})|\delta h(\boldsymbol{\theta}_0))^{(X)}] = 0. \tag{15}$$

Defining the matrix

$$(\Gamma_{AET})_{ij} = \sum_{X=\{A,E,T\}} (\partial_i h_m(\boldsymbol{\theta}_{\rm bf})|\partial_j h_m(\boldsymbol{\theta}_{\rm bf}))^{(X)} \tag{16}$$

it is then possible to invert the matrix-vector equation (15) to calculate $\Delta\theta^i$:

$$\Delta\theta^i = (\Gamma^{-1}_{AET})^{ij}\left[\sum_X (\partial_j h_m(\boldsymbol{\theta}_{\rm bf})|n)^{(X)} + (\partial_j h_m(\boldsymbol{\theta}_{\rm bf})|\delta h(\boldsymbol{\theta}_0))^{(X)}\right]. \tag{17}$$

In the presence of noise fluctuations $n^{(X)}$ and waveform modeling errors $h_e \neq h_m$, there are two sources of discrepancy between the recovered parameters $\theta^i_{\rm bf}$ and true parameters θ^i_0 described by Eq. (17). Each of these terms represent a *statistical error*, determined by the presence of noise and the second a *systematic error*, a consequence of nonfaithful model templates h_m:

$$\Delta\theta^i_n = (\Gamma^{-1}_{AET})^{ij}\sum_X (\partial_j h_{\rm m}(\boldsymbol{\theta}_{\rm bf})|n)^{(X)}, \tag{18}$$

$$\Delta\theta^i_{\rm sys} = (\Gamma^{-1}_{AET})^{ij}\sum_X (\partial_j h_{\rm m}(\boldsymbol{\theta}_{\rm bf})|\delta h(\boldsymbol{\theta}_0))^{(X)}. \tag{19}$$

The first term (18) enforces a statistical fluctuation to the recovered parameters. Since the noise has mean zero, the statistic $\theta^i_{\rm n,bf} = \theta^i_0 + \widehat{(\Delta\theta^i_n)}_{AET}$ is an unbiased estimator of the true parameters such that $\mathbb{E}[\theta^i_{\rm n,bf}] = \theta^i_0$. The quantity (19) governs the bias in the recovered parameters due to using inaccurate waveform models where $\delta h^{(X)}(\boldsymbol{\theta}_0) \neq 0$. The overall bias is thus given by the second term in Eq. (19):

$$\mathbb{E}[\theta^i_{\rm bf}] = \theta^i_0 + (\Gamma^{-1}_{AET})^{ij}\sum_X (\partial_j h_{\rm m}(\boldsymbol{\theta}_{\rm bf})|\delta h(\boldsymbol{\theta}_0)^{(X)}). \quad (20)$$

Note that since $h \sim \rho$ and $S_n(f) \sim \rho^0$, we have that $\Gamma^{-1} \sim \rho^{-2}$ giving the scaling relationship for the statistical uncertainty $(\widehat{\Delta\theta^i_n})_{AET} \sim O(1/\rho)$. Similarly, for the systematic error, the scaling relationship is given by $(\widehat{\Delta\theta^i_{\rm sys}})_{AET} \sim O(\rho^0)$ and is thus independent of the SNR of the source. Therefore, if the signal-to-noise ratio of the underlying signal is large, then the (relative) magnitude of the systematic error will be larger when compared to the statistical fluctuation. Further details can be found in [12,25,26,28].

In the limit of small waveform modeling errors $|\delta h^{(X)}|^2 \ll \rho^0$ and high SNR, the Fisher matrix yields an approximation to the predicted covariance matrix of the posterior distribution

$$\mathbb{E}[(\Delta\theta^i_{\rm bf})(\Delta\theta^j_{\rm bf})] \approx (\Gamma^{-1}_{AET})^{ij}. \quad (21)$$

with rooted diagonal elements an approximation to how well one can constrain the parameters of the system. The Fisher matrix is widely used within gravitational-wave astronomy to cheaply compute precision measurements on parameters of interest. Precision measurements are given by the rooted diagonals of the inverse of the Fisher matrix

$$\Delta\theta^i_{\rm stat} = \sqrt{(\Gamma^{-1})^{ii}} \text{no sum}. \quad (22)$$

where $\Delta\theta^i_{\rm stat}$ is the statistical error, the 1σ deviation (through expectation) in the recovered parameters due to noise fluctuations. For systematic studies in Cutler-Valisneri framework [26], the ratio between (19) and (22) is computed. If the bias on parameters exceeds the statistical uncertainty, then the proposed waveform model is not suitable for parameter estimation.

In our work, we will not focus on error induced due to approximate waveform models. Instead, we will focus on the notion of *coverage* given by an approximate posterior distribution. The "coverage" of a posterior density describes the probability that the true parameters are contained within the posteriors credible set. The Cutler Valisneri formalism can be discussed in terms of coverage of a posterior: If the (assumed normal) approximate posterior has a 68% coverage, then the waveform model will be deemed suitable for parameter estimation.[1] The primary focus of our work is to *calibrate* the coverage of an approximate posterior to a much higher level, specified by the user. The final (calibrated) credible region will then contain the true parameters with higher level of probability the original credible region. In other words, the coverage of the new calibrated approximate posterior will be larger, indicating greater certainty that we have recovered the correct parameters to a reasonable certainty. To discuss this in more detail, it is necessary to introduce the Bayesian tools we will use to perform parameter estimation.

The Fisher matrix formalism leading to Eqs. (18), (19), and (21) can be used to compute precision measurements on parameters and potential sources of bias on the recovered parameters $\boldsymbol{\theta}_{\rm bf}$. However, given an observed data stream d_o, we wish to make inference on parameters $\boldsymbol{\theta}$ that govern the structure of the underlying dataset (and thus the GW signal).

C. Bayesian theory

The standard procedure used within GW astronomy to estimate parameters of a signal $h^{(X)}$ given observation of a set of data streams $d^{(X)}_o$ is Bayesian inference. At the heart of Bayesian theory lies Bayes' theorem:

$$p(\boldsymbol{\theta}|d_o) = \frac{p(d_o|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(d_o)} \quad (23)$$

$$\propto p(d_o|\boldsymbol{\theta})p(\boldsymbol{\theta}), \quad (24)$$

where $p(\boldsymbol{\theta}|d_o)$ is the posterior density of unknown parameters $\boldsymbol{\theta}$ given the observation of a data stream d_o, $p(d_o|\boldsymbol{\theta})$ the likelihood function, and $p(\boldsymbol{\theta})$ the prior distribution, reflecting our beliefs on parameters $\boldsymbol{\theta}$ before observing the data. The marginal likelihood $p(d_o) = \int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(d_0|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ is a constant over the parameter space and is unnecessary for our work.

Stochastic sampling algorithms, such as MCMC, are used to obtain random samples $\boldsymbol{\theta}$ from the posterior density $p(\boldsymbol{\theta}|d_o)$ by constructing a Markov chain whose steady-state distribution is the posterior distribution of interest. The posterior distribution is then summarized using Monte Carlo integration to compute moments such as the posterior mean $\mathbb{E}_{p(\boldsymbol{\theta}|d_o)}[\boldsymbol{\theta}]$ or quantifying levels of precision on how well we can constrain parameters. In this work, we use the MCMC ensemble sampler `emcee` [67] to obtain samples from $p(\boldsymbol{\theta}|d_o)$.

To obtain the *exact* posterior density, one would use the likelihood function (7) with model templates $h_{\rm m}$ precisely equal to the true waveform within the data stream $h_{\rm e}$. This would yield an unbiased result in the recovered parameters, a consequence of generating an *exact* posterior density $p(\boldsymbol{\theta}|d_o)$. However, in the context of gravitational-wave astronomy, this is unfeasible. The two-body problem in general relativity has no exact solution, and the most numerically accurate are NR waveforms [16] that are computationally prohibitive for MCMC algorithms. Instead, we must make do with approximate models that

[1] In other words, for an approximate one-dimensional Gaussian distribution, if the true parameters are contained within the 68% level credible interval $[\hat{\mu}-\hat{\sigma}, \hat{\mu}+\hat{\sigma}]$, the model is deemed suitable for parameter estimation. Here hatted quantities are estimates of the posterior means and standard deviations.

are both *fast* to generate and *faithful* with their true counterpart. Hence, in the statistical inference procedure of parameter estimation in gravitational-wave astronomy, we in fact sample from an approximate posterior density $\tilde{p}(\boldsymbol{\theta}|d_o)$:

$$\tilde{p}(\boldsymbol{\theta}|d_o) \propto \tilde{p}(d_o|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{25}$$

with $h_m \neq h_e$ in the approximate likelihood function (7). As discussed in Sec. II B, this would result in a biased set of parameters.

Generating samples from an approximate posterior distribution is very common in Bayesian inference [48,68–70]. When an approximate likelihood $\tilde{p}(d_o|\boldsymbol{\theta})$ is used, the resulting posterior inference will be distorted[2] with respect to the exact posterior distribution $p(\boldsymbol{\theta}|d_o)$. Within the statistical literature, various estimators have been proposed to measure this distortion [71–73]. Most of the existing methods are based on [74,75]. Their original motivation was to check for correct sampling from the posterior distribution based on the following equality for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$:

$$\mathbb{E}_{d_o,\boldsymbol{\theta}'}[p(\boldsymbol{\theta}|d_o)] = p(\boldsymbol{\theta}), \tag{26}$$

i.e., the integral of the exact posterior density with respect to the generative model $p(d_o|\boldsymbol{\theta}')p(\boldsymbol{\theta}')$ is equal to the prior. Then they constructed statistical tests to check the validity of this relationship when replacing $p(\boldsymbol{\theta}|d_o)$ by an approximate posterior density $\tilde{p}(\boldsymbol{\theta}|d_o)$. However, these types of tests can falsely accept the hypothesis of correct sampling even if the approximate likelihood is far from the exact likelihood, see, e.g., [56,72]. Moreover, they do not quantify the distortion for a particular observation.

Reference [56] showed how to quantify the distortion of an approximate posterior credible interval conditional on the observed data by estimating its operational coverage as described in Secs. III A and III B. We present a practical operational coverage estimator for gravitational-wave problems in Sec. III B and how to calibrate approximate credible set in Sec. III C.

III. GENERAL CALIBRATION METHODOLOGY

A. Ideal operational coverage estimation

Let d_o represent the observed data. When we refer to "coverage," we are describing the posterior probability that a credible set, determined by a specific prior and likelihood function forming the posterior, contains the true parameters we intend to estimate.

Let $\tilde{C}_{d_o}$ and C_{d_o} be the level α posterior credible sets calculated using $\tilde{p}(\boldsymbol{\theta}|d_o) \propto \tilde{p}(d_o|\boldsymbol{\theta})p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|d_o) \propto p(d_o|\boldsymbol{\theta})p(\boldsymbol{\theta})$, respectively, i.e.,

$$\alpha = P(\boldsymbol{\theta} \in C_{d_o}) = \int \mathbb{1}_{C_{d_o}}(\boldsymbol{\theta})p(\boldsymbol{\theta}|d_o)\mathrm{d}\boldsymbol{\theta}$$
$$\alpha = \tilde{P}(\boldsymbol{\theta} \in \tilde{C}_{d_o}) = \int \mathbb{1}_{\tilde{C}_{d_o}}(\boldsymbol{\theta})\tilde{p}(\boldsymbol{\theta}|d_o)\mathrm{d}\boldsymbol{\theta}, \tag{27}$$

where $\mathbb{1}$ denotes the indicator function, i.e., $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 otherwise. A coverage of α is only guaranteed if the data are distributed according to the assumed generative model. This means that if the data d_o are actually generated from the specified likelihood $p(d_o|\boldsymbol{\theta})$, then C_{d_o} achieves the nominal level α.

However, since the likelihood $\tilde{p}(d_o|\boldsymbol{\theta})$ of the approximate posterior does not correspond to the generative model, its level α credible set $\tilde{C}_{d_o}$ does not achieve the nominal level α but only an *operational* coverage probability

$$b(d_o) = P(\boldsymbol{\theta} \in \tilde{C}_{d_o}) = \int \mathbb{1}_{\tilde{C}_{d_o}}(\boldsymbol{\theta})p(\boldsymbol{\theta}|d_o)\mathrm{d}\boldsymbol{\theta} \tag{28}$$

that is not generally equal to the nominal coverage, α.

If $b(d_o) \gg \alpha$ or $b(d_o) \ll \alpha$, this would indicate a poor approximation. Thus $|\alpha - b(d_o)|$ measures the distortion or discrepancy in coverage of the credible set at the observed data d_o. If an approximation is not good enough for a user, there are two possible approaches to fix this: using a different approximation, such as using a more accurate but more costly waveform model, or correcting the posterior itself [76].

In practice, we often generate samples from the approximate posterior $\tilde{p}(\boldsymbol{\theta}|d_o)$ using MCMC and estimate $\tilde{C}_{d_o}$. If we denote this estimator of $\tilde{C}_{d_o}$ by $\hat{C}_{d_o}$, we get the *realized operational coverage* probability

$$b_r(d_o) = P(\boldsymbol{\theta} \in \hat{C}_{d_o}). \tag{29}$$

The realized operational coverage probability in Eq. (29) can be estimated using the standard Monte Carlo method, i.e., sampling from the exact posterior $p(\boldsymbol{\theta}|d_o)$ and taking the proportion of the samples that are inside credible intervals $\hat{C}_{d_o}$. This estimate takes the Monte Carlo error of estimating the credible set into account. However, this procedure will not be practical because it needs samples from the exact posterior $p(\boldsymbol{\theta}|d_o)$, which may be expensive and impractical to sample from. An example here would be generating an exact posterior density $p(\boldsymbol{\theta}|d)$ using the most accurate, but computationally prohibitive numerical relativity waveforms for massive black holes (MBHs). Instead, using techniques from regression, we show that it is possible to provide operational coverage estimators without sampling from the exact posterior distribution $p(\boldsymbol{\theta}|d)$ in the next section.

B. Practical operational coverage estimation

Operational coverage estimators that do not require simulation from the exact posterior have been suggested [56,57].

[2]By distortion we refer to the nonzero statistical distance between two distributions, say p_1 and p_2. For example, such a distortion (and thus statistical distance) could be measured by the Kullback-Leibler divergence.

These are based on logistic regression (binary classification) and (annealed) importance sampling. In this paper, we use the logistic regression as the operational coverage.

The setup is the following. For $j = 1, \ldots, J$, we sample $\boldsymbol{\theta}_{(j)}$ from the prior, $\boldsymbol{\theta}_{(j)} \sim p(\boldsymbol{\theta})$ and generate data $d_{(j)} = h_{\mathrm{e}}(\boldsymbol{\theta}_{(j)}) + n$. For each $d_{(j)}$, we estimate a credible set $\hat{C}_{d_{(j)}}$ of $\tilde{p}(\boldsymbol{\theta}_{(j)}|d_{(j)})$ and, $c_j = \mathbb{1}(\boldsymbol{\theta}_{(j)} \in \hat{C}_{d_{(j)}})$ which is regarded as a Bernoulli trial with success probability $b_r(d_{(j)})$ associated with the data themselves, i.e.,

$$c_j \sim \mathrm{Bernoulli}(b_r(d_{(j)})).$$

If we can fit a logistic regression to c_j with $d_{(j)}$ as predictors, one can use the model to predict $b_r(d_o)$ with the observed data d_o. We denote this prediction as $\bar{b}_r(d_o)$. In literature, a semiparametric regression [56] and a Bayesian additive regression tree [57] were used.

Theoretical properties of traditional parametric and nonparametric regression often assume that the number of samples is larger than the dimension of predictor, i.e., $|d| \ll J$. As is often the case within gravitational-wave data analysis, this assumption is violated when the length of $d_{(j)}$ is larger than the training dataset size J. In this paper, we use an artificial neural network (ANN) with the sigmoid activation function to fit the binary classification for practicality. To enhance the efficiency of the ANN [77], we utilize an autoencoder to project $d_{(j)}$ into a lower-dimensional subspace while capturing the main features of the data [78]. The autoencoder comprises two neural networks: an encoder network that maps the input data into a lower-dimensional latent space, and a decoder network that recreates the input from the encoded representation.

For realistic applications of data analysis, the choice of the nominal value α is chosen by the user. The proposed estimators by [56,57] are conditioned on a nominal value and restrict the operational coverage estimate to a particular nominal coverage. In an attempt to tackle this issue, the nominal level α is also taken as an input to fit the classification. The training set is $\{d_{(j)}, c_j, \alpha_j\}_{j=1}^{J}$ where $\alpha_j \sim p(\alpha)$, $d_{(j)} = h_{\mathrm{e}}(\boldsymbol{\theta}_{(j)}) + n$ and $c_j = \mathbb{1}_{\hat{C}_{d_{(j)},\alpha_j}}(\boldsymbol{\theta}_{(j)})$. Here, $\hat{C}_{d_{(j)},\alpha_j}$ is an estimated credible set of $\tilde{p}(\boldsymbol{\theta}|d_{(j)})$ with a nominal level, α_j, i.e., $\alpha_j = \int \mathbb{1}_{\hat{C}_{d_{(j)},\alpha_j}}(\boldsymbol{\theta})\tilde{p}(\boldsymbol{\theta}|d_{(j)})d\boldsymbol{\theta}$.

Figure 1 shows the procedure for constructing the operational coverage estimator. It consists of two components, dimensional reduction and classification. First, we train the encoder function in the autoencoder to reduce the dimension of $\{d_{(j)}\}_{j=1}^{J}$. Then the compressed data and nominal values are fed to an ANN as an input feature. To obtain the target output $c_1, \ldots, c_J$, the classifier is trained. The operational coverage, which is the success probability of the fitted classifier, can be predicted with d_o at a desired nominal level of α, and it is denoted by $\bar{b}_r(d_o, \alpha)$.

We should point out that data stream $d_{(j)}$, as an input of the proposed classifier, can be either an actual signal $(h_e + n)$ or discrete Fourier transform of signal $(\hat{h}_e + \hat{n})$. Alternatively, an adequate summary of signals can be used, and it is denoted by $S(d_{(j)})$. If a summary is used, the

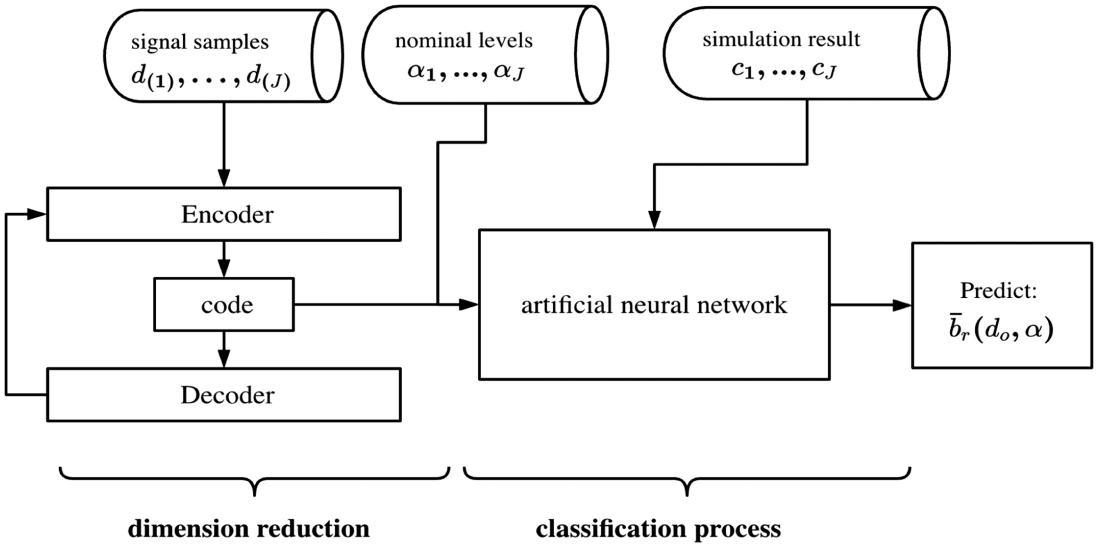

FIG. 1. The flow chart describing the overall procedure of our operational coverage estimator. For $j = 1, \ldots, J$, generate $\boldsymbol{\theta}_{(j)} \sim p(\boldsymbol{\theta})$, and data streams generated using an exact model with noise $d_{(j)} = h_e^{(X)}(t; \boldsymbol{\theta}_{(j)}) + n^{(X)}(t)$ are passed into an encoder where dimensional reduction is applied. After this process, nominal levels are simulated $\alpha_j \sim p(\alpha)$ and $\hat{C}_{d_{(j)},\alpha_j}$ of an approximate posterior distribution using the Fisher based parameter estimation scheme returns c_j, where $c_j = 1$ if $\boldsymbol{\theta}_{(j)} \in \hat{C}_{d_{(j)},\alpha_j}$ and 0 otherwise. The classifier is subsequently trained, enabling us to predict $\bar{b}_r(d_o, \alpha)$ at a specified nominal level α.

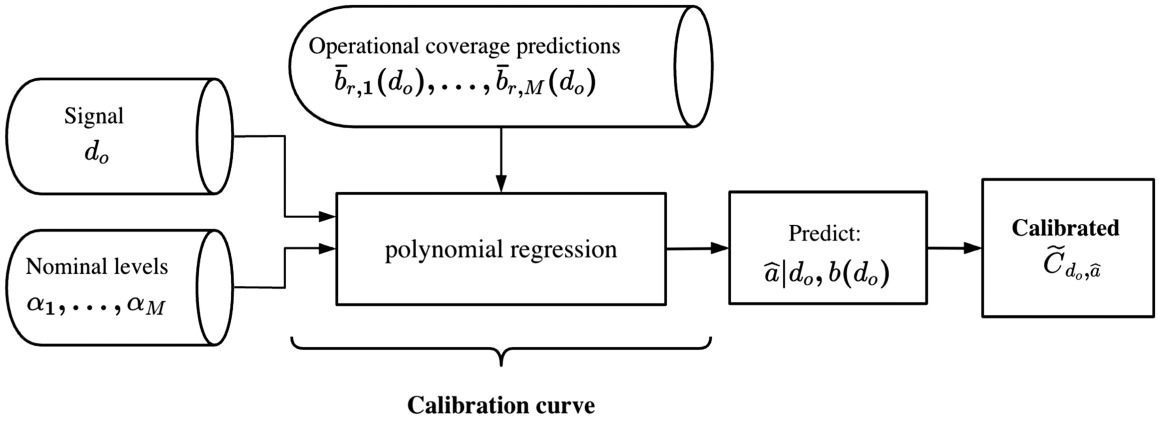

FIG. 2. The flow chart describes the overall procedure of the calibration curve. After the procedure given in Fig. 1 is completed, for a nominal level sample $\alpha_j \sim p(\alpha)$ the operational coverage is predicted $\bar{b}_{r,j}(d_o)$ using the estimator, $j = 1, \ldots, M$. Taking $\bar{b}_{r,j}(d_o)$ as a predictor and α_j as a response, polynomial regression is fitted, and then the calibrated nominal level $\hat{\alpha}$ is predicted at the desired operational coverage $b(d_o)$.

training set is $\{S(d_{(j)}), c_j, \alpha_j\}_{j=1}^{J}$ and, the prediction is made with a nominal level of α and the summary of the observed data, $\bar{b}_r(S(d_o))$. Our choice of the summary for numerical simulation studies is included in Secs. IV and V.

We use k-fold cross validation to predict success probabilities $\bar{b}_r(d_o)$ and class at d_o, but the uncertainty of the operational coverage estimate is not attainable (i.e., only a point estimate is available with this approach). For a simple logistic regression, the Delta method [79] and the bootstrap method [80] are therefore applied to measure the prediction errors in Sec. IV. We emphasize here that there does not exist any unbiased and universal estimator of the variance of k-fold cross validation that is valid under all distributions [81].

C. Credible interval calibration via operational coverage estimation

There are no formal guidelines on how to interpret operational coverage. One may consider correcting an approximate posterior distribution. Within the approximate Bayesian computation framework, adjustment procedures for bias and frequentist coverage of Bayesian credible sets were suggested when the credible set does not have the correct nominal coverage probability in the frequentist sense [82]. With a focus on coverage in the Bayesian sense, a distortion map estimator using machine learning methods was proposed by [76], and is limited to one-dimensional problems.

The application of an operational coverage estimator enables us to determine the approximate credible set level that yields the desired posterior coverage level. Since an exact inverse map of the trained ANN is not always feasible or necessary, the simplest way is to read off from a plot of operational coverage estimates against nominal values at observed signal d_o.

For an observed signal d_o, the calibration curve is constructed with nominal levels and operational coverage predictions, which the nominal level is parametrized by operational coverage, and the procedure is summarized in Fig. 2. The training data $\{\alpha_j, \bar{b}_{r,j}(d_o)\}_{j=1}^{M}$ are constructed by finding the operational coverage at each of the M nominal levels in $[\alpha_{\min}, \alpha_{\max}]$ from the estimator (Fig. 1). Taking a nominal level (α_j) as a response and operational coverage $(\bar{b}_{r,j}(d_o))$ as a predictor, the K-degree polynomial regression is fitted. The value for K is chosen by minimizing the residual mean squared error, i.e.,

$$\alpha_j = c_0 + \sum_{k=1}^{K} c_k \bar{b}_{r,j}(d_o)^k, \qquad j = 1, \ldots, M.$$

Given the desired operational coverage $b(d_o)$, the *calibrated nominal coverage* level, $\hat{\alpha}$, is estimated from the calibration curve for d_o and, with the calibrated approximate credible set $\hat{C}_{d_o,\hat{\alpha}}$, $P(\theta \in \hat{C}_{d_o,\hat{\alpha}}) = b(d_o)$.

We demonstrate how to estimate an operational coverage and calibrate a credible set using a simple toy example in Sec. IV and a massive black hole problem in Sec. V.

IV. APPLICATION: SIMPLE TOY MODEL

In this section, we present a simple toy model to illustrate our calibration procedure discussed in Sec. III.

A. Setup and Fisher matrix validation

In this section, we will consider a data stream of the form

$$d(t) = h_e(t; \boldsymbol{\theta}) + n(t) \tag{30}$$

with an exact template of the form

$$h_e(t; a_0, f_0, \dot{f}_0) = a_0 \sin\left(2\pi t\left[f_0 + \frac{1}{2}\dot{f}_0 t\right]\right). \quad (31)$$

Here, we have the true values for parameters $\boldsymbol{\theta}_0 = \{a_0 = 5\times 10^{-21}, f_0 = 10^{-3}\ \text{Hz}, \dot{f}_0 = 10^{-8}\ \text{Hz/s}\}$.

We use model templates

$$h_m(t; a, f, \dot{f}, \epsilon) = a \sin\left(2\pi t\left[f + \frac{1}{2}\dot{f} t\right](1-\epsilon)\right). \quad (32)$$

Here $\epsilon \ll 1$ is used as a tuneable parameter allowing deviations from the exact model (31) $h_{\rm e}(t;\boldsymbol{\theta}, \epsilon = 0)$ given by an approximate model $h_{\rm m}(t;\boldsymbol{\theta}, \epsilon \neq 0)$. For simplicity, we do not take into account the response function through TDI as outlined in Sec. II. Hence, when calculating the inner product (6), which is used in likelihood (5), Fisher matrix (16) and SNR (8) calculations, we only consider a single data stream and use the approximate LISA-like PSD defined by Eq. (1) of [83].

Using Eqs. (11a) and (11b), we generate stationary Gaussian noise in the frequency domain to construct the toy model data stream (30). In this example, we will set $\epsilon = 10^{-6}$, allowing for a discrepancy between the exact model h_e and approximate model h_m. For $\epsilon = 0$ and over a time of observation of 30 hours sampled with cadence $\Delta t \sim 200$ seconds, we observe an optimal matched filtering SNR $\rho \sim 188$. The length of the data stream is $N = 2^{16}$.

Our calibration technique requires multiple parameter estimation simulations using an approximate model to then estimate the operational coverage. As discussed in Sec. I, this is extremely computationally intensive so we approximate samples from the approximate posterior density using a Fisher matrix approach instead.

To validate our Fisher matrix approach, we inject an exact model with $\epsilon = 0$ and recover with an approximate model with $\epsilon = 10^{-6}$ using the `emcee` algorithm. We generate 31,000 samples from the approximate posterior under $h_m(t;\boldsymbol{\theta}), \epsilon = 10^{-6})$, and then discard 6,000 samples as burn in. In parallel, we compute the Fisher matrix (16) and then sample from a multivariate Gaussian,

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0 + \boldsymbol{\theta}_{\rm bias}, \Gamma^{-1}(h_m)),$$
$$\theta^i_{\rm bias} = [\Gamma^{-1}(h_m)]^{ij}(\partial_j h_{\rm m} | \delta h + n), \quad (33)$$

with $\theta^i \in \boldsymbol{\theta}_{\rm bias}$. Here Eq. (33) is evaluated at the true parameters $\boldsymbol{\theta}_0$ and $\delta h = h_e(t;\boldsymbol{\theta}_0) - h_m(t;\boldsymbol{\theta}_0, \epsilon = 10^{-6})$. We plot the approximate posterior densities in Fig. 11. The blue curve is generated via MCMC, and the green curve generated via the Fisher matrix computation. This simulation has shown that the Fisher matrix can be used as a suitable approximation to the posterior density. Having verified the Fisher matrix is a suitable approximation, we then use it to generate approximate posteriors in bulk in order to apply the calibration procedure discussed in Sec. III. This is the focus of the next section.

B. Calibration procedure

The calibration technique described in Sec. III requires a training set, built from the prior space of samples and the resultant generation of a family of approximate posteriors. We first focus our attention on a single-parameter study, then generalize to the three-parameter study at the end of this section.

For a single-parameter study, $\dot{f}$ is unknown, and the true values for a, f are used, i.e., $\boldsymbol{\theta} = \dot{f}$. The uniform prior is assigned for $\dot{f}$,

$$\dot{f} \sim \text{U}[\dot{f}_0 - 10^{-13}, \dot{f}_0 + 10^{-13}]\ \text{Hz/s}. \quad (34)$$

The training data $\{S(d_{(j)}), c_j, \alpha_j\}_{j=1}^{5000}$ are generated to obtain a practical operational coverage estimator. For $j = 1, \ldots, 5000$, $\boldsymbol{\theta}_{(j)}$ is generated from the prior (34) and $\alpha_j \sim U[0.78, 0.97]$. For each of prior sample $\boldsymbol{\theta}_{(j)}$, the data stream $d_{(j)} = \hat{h}_{e,(j)}(f;\boldsymbol{\theta}_{(j)}) + \hat{n}_{(j)}$ is generated by adding a noise $\hat{n}_j$ through (10) to the exact reference signal $\hat{h}_{e,(j)}$. An approximate posterior density is in the form of a multivariate Gaussian (33) using the inverse Fisher matrix (21) and expectation for the bias (17) at $\boldsymbol{\theta}_{(j)}$. An output c_j is obtained from the α_j credible set of the approximate posterior density. Instead of $d_{(j)}$, the real part of discrete Fourier transform of signal is used to find the practical operational coverage estimator, i.e, $S(d_{(j)}) = \text{Re}(d_{(j)})$. We tried $|d_{(j)}|^2$ and $|d_{(j)}|$ and did not gain any improvement in the results.

We use autoencoders to reduce the dimension of $S(d_{(j)})$ in order to apply the calibration procedure. The autoencoder is trained with an Adam optimizer [84] and a learning rate, 10^{-4}, is chosen by minimizing the mean squared error. The size of $S(d_{(j)})$ is then reduced from 2^{16} to 2^3, which is far less than the size of the overall training dataset. From our preliminary study, reducing the size of the dataset any less than 2^3 gave a significantly poor fit. For classification, an ANN with one fully connected layer is trained using the cross entropy loss function. The calibration curve is obtained from the feed-forward ANN inversion. We tried multiple layers for the ANN but found no real gain from using more than one layer for this toy example.

The left panel of Fig. 3 presents the performance of the operational coverage estimator (with summary of the procedure given in Fig. 1) using 30 approximate signals with noise and on average absolute errors are relatively small. The right panel of Fig. 3 compares the operational coverage estimates using the *practical estimator* $\bar{b}_r(S(d_o))$ (Sec. III B) and the *realized operational coverage* $b_r(d_o)$ (29) for the test signal. At a given nominal level α, the practical operational coverage and the realized operational coverage agree to excellent precision. Absolute errors tend to be less than 0.06 in general and, for the test data, the

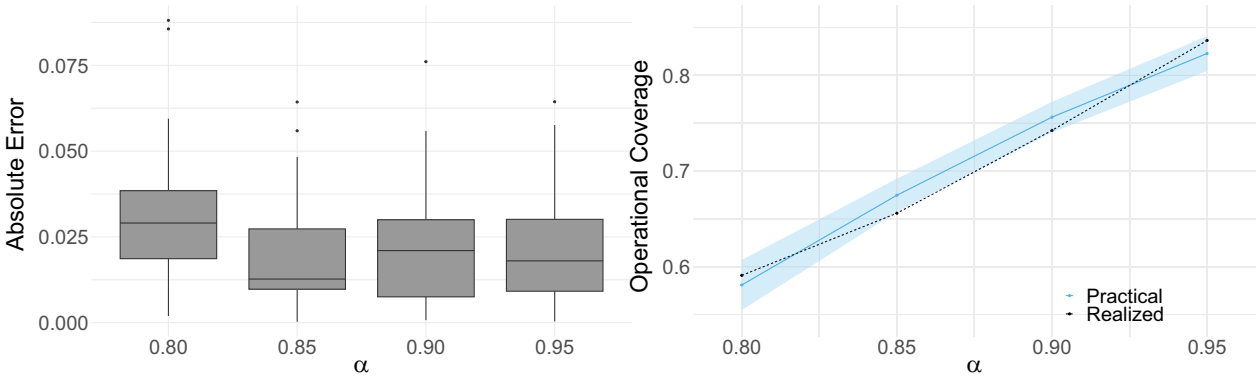

FIG. 3. Estimation of $\dot{f}$-only. Absolute error from 30 replicates against α (left) and operational coverage estimates $\bar{b}_r(S(d_o))$ with 2-SE error and *realized operational coverage* $b_r(d_o)$ (right) for the test waveform d_o.

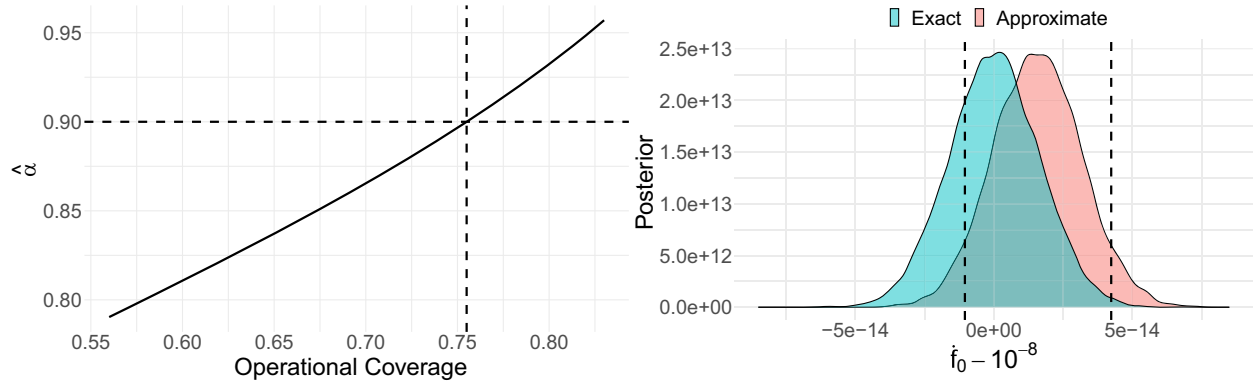

FIG. 4. Left panel: calibration curve. Right panel: exact and approximate posterior densities of $\dot{f}_0$. The dashed lines on the left figure represent the operational coverage for the 90% credible interval of approximate posterior ($\tilde{C}_{d_o,0.9}$). This is calculated by using the calibration curve to map the calibrated nominated level of $\hat{\alpha} = 90\%$ back to an operational coverage level, which is ≈76% for the test waveform.

coverage of posterior density over the 0.9 approximate credible set is 0.76.

Exact and approximate inferences on the true signal (31) are compared in Fig. 4. It is observed that the exact and approximate posterior densities do not overlap completely and there is some deviation between them. The calibration curve shows how the calibrated level changes with the desired operational coverage and the calibrated level $\hat{\alpha}$ is higher than $b(d_o)$, i.e., $\bar{b}_r(S(d_o))$ is smaller than α (right panel of Fig. 3). The main conclusion of this study is that we have calibrated an approximate credible set (arising from an approximate posterior) to achieve 0.76 coverage of the posterior.

We now consider the full model with three parameters, i.e., $\boldsymbol{\theta} = \{a_0, f_0, \dot{f}_0\}$. Tight priors are assigned on these three parameters:

$$a \sim \mathrm{U}[a_0 - 10^{-22}, a_0 + 10^{-22}]$$
$$f \sim \mathrm{U}[f_0 - 10^{-7}, f_0 + 10^{-7}]\ \mathrm{Hz}$$
$$\dot{f} \sim \mathrm{U}[\dot{f}_0 - 10^{-13}, \dot{f}_0 + 10^{-13}]\ \mathrm{Hz/s}.$$

We increase the size of the training data to 10,000, $\{S(d_{(j)}), c_j, \alpha_j\}_{j=1}^{10000}$ where $S(d_{(j)}) = \mathrm{Re}(d_{(j)})$, and it is generated similarly. For approximate posterior density, the multivariate Gaussian form of density approximation for logged parameter values was imposed. We used three fully connected layers with one dropout layer in both the encoder and decoder to reduce the dimension of $S(d_{(j)})$ from 2^{16} to 2^6. A one-layer ANN classifier with the l_1 penalty on weights of neurons is fitted.

The performance of the operational coverage estimator using 30 approximate signals with noise is shown in the left panel of Fig. 5, and absolute errors are less than ≈0.075 in general. The right panel of Fig. 5 compares $\bar{b}_r(S(d_o))$ and $b_r(d_o)$ for the test signal. The practical and realized operational coverage agrees relatively well at a given α. Exact and approximate posterior densities for the test data d_o are compared in Fig. 6. The calibration curve for the test waveform d_o is shown in Fig. 7, and the calibrated nominal level $\hat{\alpha}$ is higher than the desired operational coverage value, i.e., $b_r(d_o)$ is smaller than α. For d_o, the coverage of

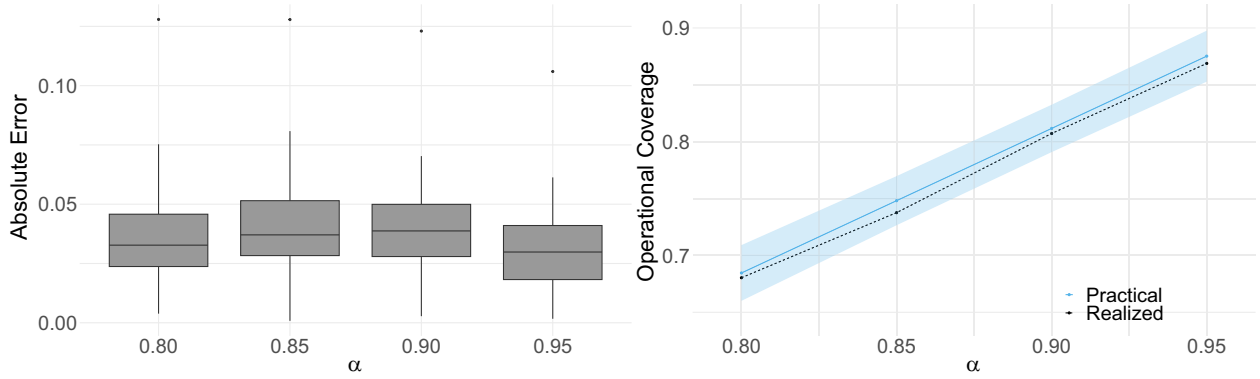

FIG. 5. Estimation of the full model, i.e., $a_0, f_0, \dot{f}_0$. Absolute error from 30 replicates against α (left) and operational coverage estimates $\bar{b}_r(S(d_o))$ with 2-SE error and *realized operational coverage* $b_r(d_o)$ (right) for the test waveform d_o.

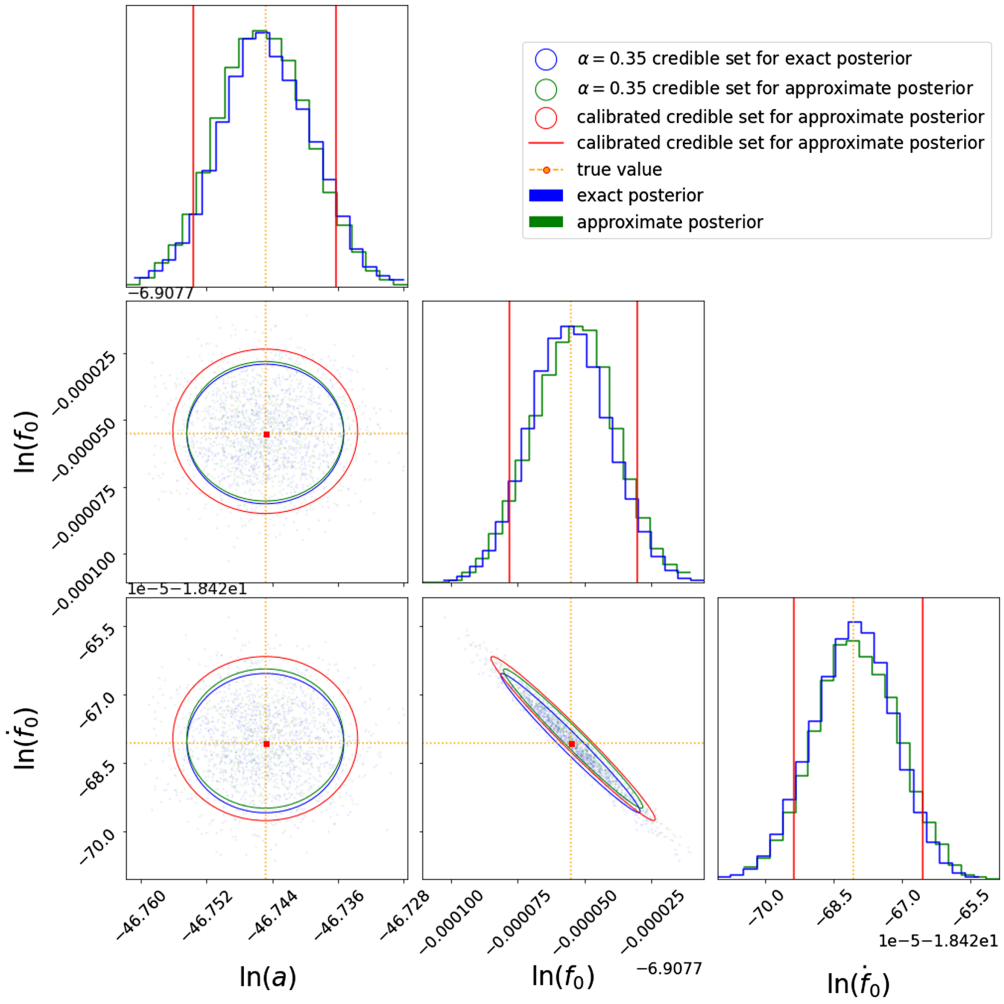

FIG. 6. Corner plot summary of samples from the exact and approximate posterior densities in the toy model. The panels on the diagonal show the exact (blue) and approximate (green) empirical marginal posterior densities with the true value (yellow dashed line) and calibrated approximate credible interval (red line). The off-diagonal panels shows the exact (blue contour) and approximate (green contour) credible sets with a nominal level of $\alpha = 0.8$ and, calibrated approximate credible set (red contour) with the desired operational coverage of 0.8 based on exact (blue points) and approximate (green points) posterior sample. The true values $\boldsymbol{\theta}_0$ are marked by red points.

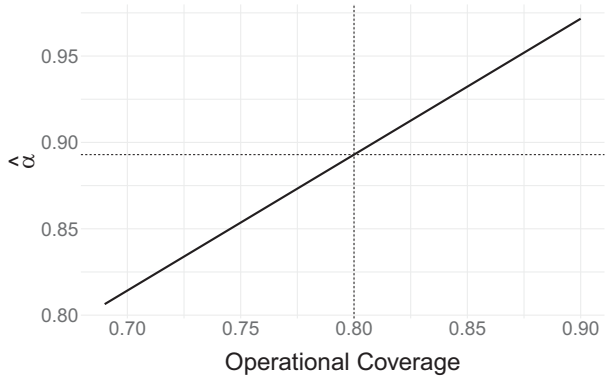

FIG. 7. Calibration curve for d_o. The dashed lines represent the calibrated nominal level $\hat{\alpha} = 0.89$ for the target operational coverage $b(d_o) = 0.8$.

posterior density over the approximate credible set with a nominal level of 0.89 is 0.8.

Having demonstrated our calibration procedure through an illustrative example, we will now apply it to a more realistic gravitational-wave scenario.

V. APPLICATION: MASSIVE BLACK HOLES

In this section, we apply the calibration procedure discussed in Sec. III and exemplified in Sec. IV on a realistic massive black hole binary signal. Using `lisabeta` developed in [85,86], we generate complete inspiral-merger-ringdown frequency domain spin-aligned massive black holes in the solar system barycenter frame with the LISA response applied.

A. Setup

In the Bayesian inference, we incorporate higher modes $\mathcal{H}_e = \{(2,2),(2,1),(3,3),(3,2),(4,4),(4,3)\}$ for the true MBH waveform denoted by $h_e(t;\boldsymbol{\theta})$. Approximate waveforms denoted by $h_m(t;\boldsymbol{\theta})$ are generated by removing a single harmonic $\mathcal{H}_m = \mathcal{H}_e \backslash \{(4,3)\}$, giving a deviation from the exact model and the approximate model. For our exact signal with modes $\mathcal{H}_e$, we define the true parameters: the total mass $M = m_1 + m_2 = 3 \times 10^7 M_\odot$; mass ratio $q = m_1/m_2 = 2$; the two effective spin parameters of the two component masses $\chi_1 = 0.5$ and $\chi_2 = 0.5$; the time of coalescence $t_c = 10^5$ seconds; luminosity distance 6.67 Gpc; initial phase at coalescence $\phi_c = 1.1$; sky position $(\beta = 0.3, \lambda = 0.8)$ in ecliptic coordinates; and polarization angle $\psi = 1.7$. In our example, we will focus on a subset of parameters $\boldsymbol{\theta} = \{M, q, \chi_1, \chi_2, t_c\}$, to demonstrate the calibration procedure as a proof of principle.

We choose true parameters given by $\boldsymbol{\theta}_0 = \{M_0 = 10^7 M_\odot, q_0 = 2, \chi_{1,0} = 0.5, \chi_{2,0} = 0.5, t_{c,0} = 10^5 \text{ seconds}\}$. The observation time of our signals will be $\sim$1 day, sampled with cadence $\Delta t = 200$ seconds. Owing to the large total mass, the frequencies emitted by the massive black hole signal are low, even at the larger harmonics. This allows us to analyze very short data segments with large sampling intervals thus reducing computational costs. The length of our datasets are $N = 2^{12}$. We found that the optimal matched filtering SNR using (8) over both the A and E channels are given by $\rho_A \sim 2022.02$ and $\rho_E \sim 1702.57$ giving a total SNR over both A and E as $\rho_{AE} \sim 2643.35$.

Before applying the calibration procedure, we will show that our Fisher matrix calculations are not subject to numerical instabilities. We inject a signal with true parameters $\boldsymbol{\theta}_0$ defined earlier into a two noiseless data streams corresponding to TDI channels $X = \{A, E\}$. Including the T channel is unnecessary for our purposes since the contribution of SNR is low with respect to the A and E channels. For inference, we use the likelihood defined in (7) with model template h_m with an incomplete set of modes $\mathcal{H}_m = \mathcal{H}_e/\{(4,3)\}$. Starting close to the true parameters, we use the `emcee` sampling algorithm to generate samples from the approximate posterior density $\tilde{p}(\boldsymbol{\theta}|d)$. To compute the Fisher matrix, we use the numerical procedure of finite differences to compute derivatives of the MBH waveform with respect to parameters. After computing the matrix (16), we apply a log transformation to reduce the condition number of the matrix prior to computing the inverse. From Eq. (20), we then sample from a multivariate Gaussian,

$$\begin{aligned} &\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0 + \boldsymbol{\theta}_{\text{bias}}, \Gamma_{AE}^{-1}(h_m)), \\ &\theta^i_{\text{bias}} = [\Gamma_{AE}(h_m)^{-1}]^{ij} \sum_{X=\{A,E\}} (\partial_j h^{(X)}_{\text{m}} | \delta h^{(X)} + n^{(X)}), \end{aligned} \tag{35}$$

for θ^i_{bias} a component of $\boldsymbol{\theta}$. We remind the reader that each of the quantities in (35) is evaluated at the true parameters $\boldsymbol{\theta}$. We then plot the histogram of samples alongside the approximate posterior density in Fig. 12. What we learn here is that the Fisher matrix is a suitable approximation to the posterior density and can be used to approximate posterior distributions generated using an approximate waveform model.

B. Calibration procedure

As a proof of principle, we will apply our calibration procedure on a five-dimensional space. We will choose the five parameter set $\boldsymbol{\theta} = \{M, q, \chi_1, \chi_2, t_c\}$. For this study, tight uniform priors are set for the five parameters as follows:

$$\begin{aligned} M &\sim \text{U}[M_0 - 5 \times 10^4, M_0 + 5 \times 10^4] \\ q &\sim \text{U}[q_0 - 2.5 \times 10^{-3}, q_0 + 2.5 \times 10^{-3}] \\ \chi_1 &\sim \text{U}[\chi_{1,0} - 5 \times 10^{-4}, \chi_{1,0} + 5 \times 10^{-4}] \\ \chi_2 &\sim \text{U}[\chi_{2,0} - 5 \times 10^{-4}, \chi_{2,0} + 5 \times 10^{-4}] \\ t_c &\sim \text{U}[t_{c,0} - 0.25, t_{c,0} + 0.25]. \end{aligned}$$

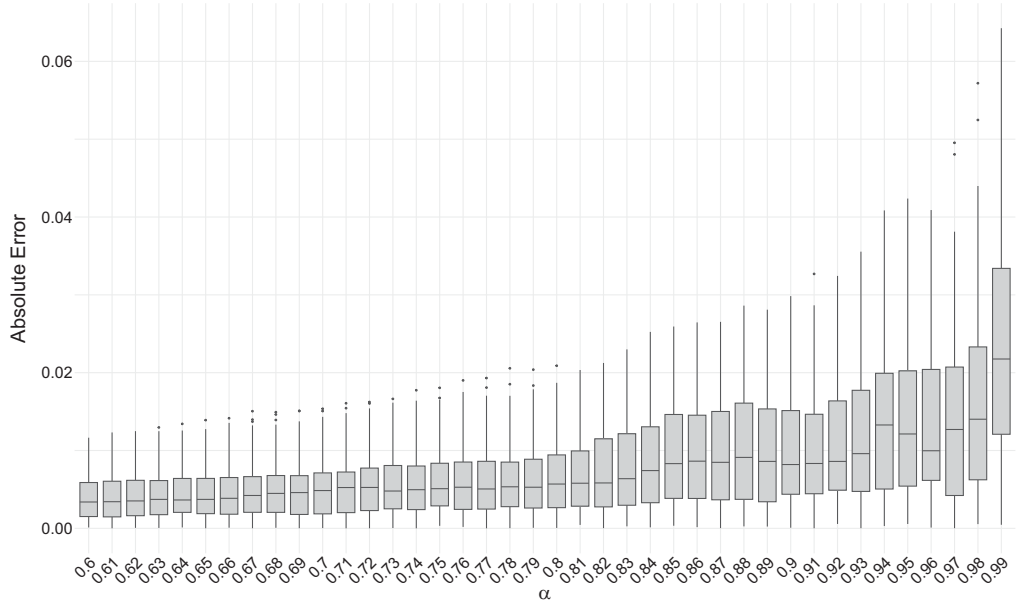

FIG. 8. Absolute error between operational coverage estimate and realized operational coverage against α from 30 replicates. The black dots are outlier absolute errors at a specific nominal level α.

Following a similar procedure outlined in Sec. IV B, the data stream is $d = \hat{h}_e(\boldsymbol{\theta}) + \hat{n}$, and the test waveform is $d_o = \hat{h}_e(\boldsymbol{\theta}_0)$. The training data $\{S(d_{(j)}), c_j, \alpha_j\}_{j=1}^{10^5}$ where $S(d_{(j)}) = \mathrm{Re}(d_{(j)})$ is generated to get a practical operational coverage estimator. We also considered $|d_{(j)}|^2$ and did not gain any significant improvements in the results and, the corresponding result is not included in the paper. We reduce the input feature size of $N = 2^{12}$ to $N = 2^8$ using a three-layer fully connected encoder and decoder network. For classification, an ANN using three fully connected hidden layers and an output layer with a sigmoid activation function are used. For the *realized operational coverage* estimation b_r, we used 26,880 exact posterior samples, using the complete inspiral-merger-ringdown waveform with full harmonic structure $\mathcal{H}_e$, with 32 parallel chains with a thinning factor of 5.

In general, the operational coverage estimator $\bar{b}_r$ exhibits generally small absolute errors in Fig. 8. Although the absolute errors tend to increase with α, relative errors are likely to be less variable. The absolute error is small as 0.013 and large as 0.065, i.e., the smallest and largest relative absolute errors are $0.013/0.6 = 0.0217$ and $0.065/0.99 = 0.0657$ respectively. For the test waveform d_o, the estimate $\bar{b}_r(S(d_o))$ from the practical operational estimator is compared to the realistic operational coverage $b_r(d_o)$ in the left plot of Fig. 9. The calibration curve in Fig. 9 was modeled by a polynomial regression with the degree 7. We observe a very small operational coverage in

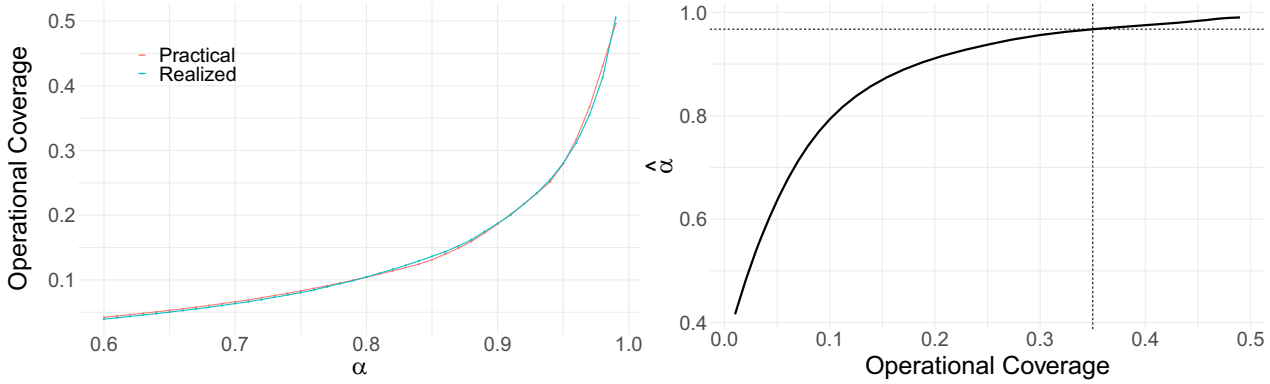

FIG. 9. Left panel: operational coverage estimates $\bar{b}_r(S(d_o))$ (red) and realized operational coverage $b_r(d_o)$ (blue) for the test data d_o against α. Right panel: calibration curve for d_o [Calibrated nominal levels $\hat{\alpha}$ against the target operational coverage $b(d_o)$]. The dashed lines represent the calibrated nominal level $\hat{\alpha} = 0.97$ for the target operational coverage $b(d_o) = 0.35$.

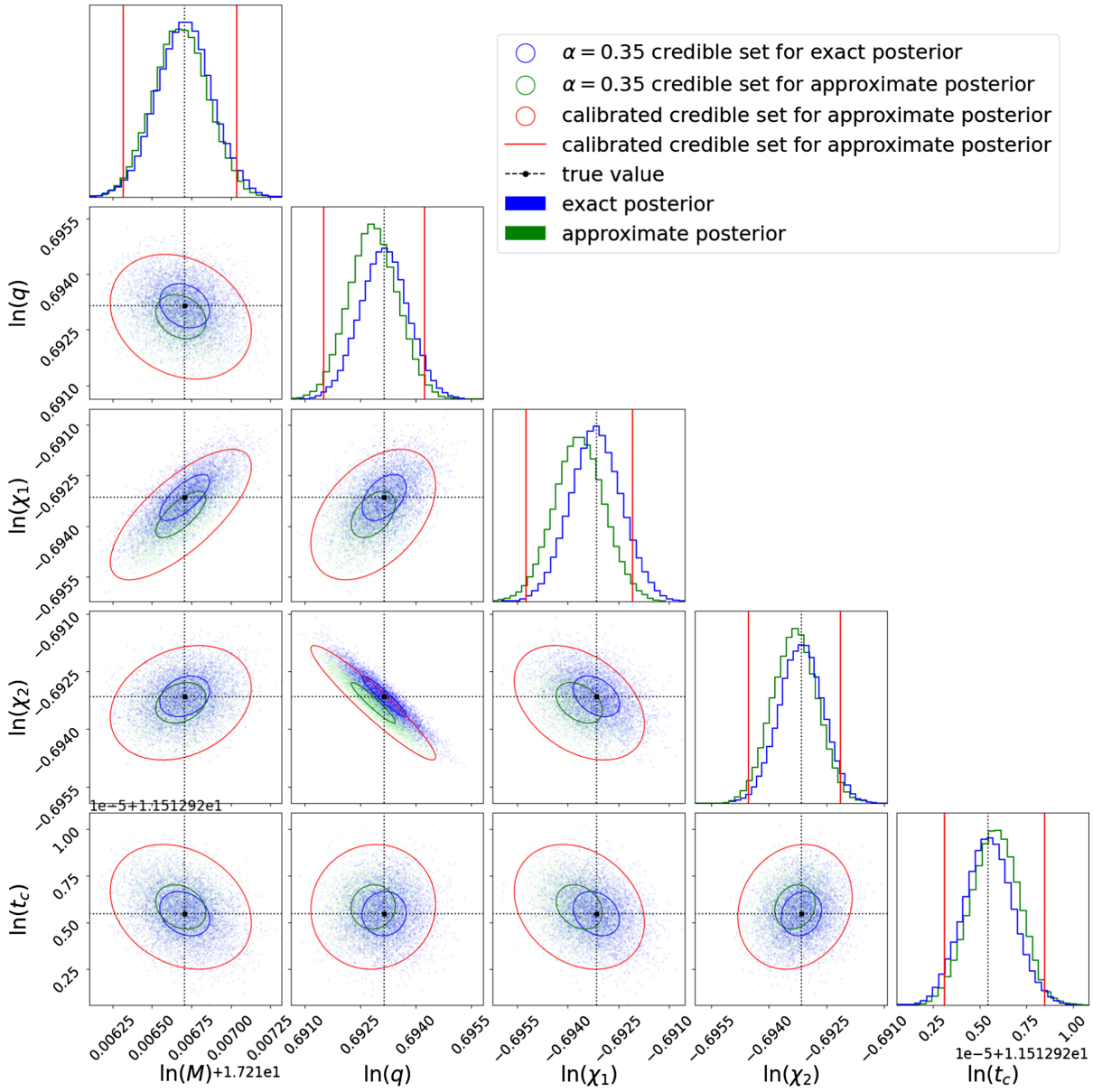

FIG. 10. Corner plot summary of samples from the exact and approximate posterior densities for the massive black hole binary signal d_o. The panels on the diagonal show the exact (blue) and approximate (green) empirical marginal posterior densities with the true value (black dashed line) and calibrated approximate credible interval (red line). The off-diagonal panels show the exact (blue contour) and approximate (green contour) credible set with a nominal level of $\alpha = 0.35$ and, the calibrated approximate credible set (red contour) with the desired operational coverage of 0.35 based on exact (blue points) and approximate (green points) posterior samples. The true values $\boldsymbol{\theta}_0$ are marked by black points.

comparison to the nominal level ($\bar{b}_r(S(d_o)) \ll \alpha$), and this is due to the small overlap between the exact and approximate posterior distributions in Fig. 10, in particular q and χ_2. For example, the posterior coverage of the 0.967 credible set of an approximate posterior distribution is 0.35, and this is also confirmed from the calibration curve that the calibrated level is 0.97 for the target operational coverage of 0.35. As result, the calibrated credible set of an approximate posterior (red contour) is larger in order to achieve the target 0.35 operational coverage, which is about 1.8σ.

VI. DISCUSSION AND CONCLUSIONS

In this paper, we have presented a Bayesian calibration procedure for the operational coverage of an approximate family of GW posteriors and applied the method to both a toy model and a more realistic analysis of MBHB signals for LISA. Specifically, the posteriors that are calibrated in our work are Fisher-information-based approximations (i.e., normal approximations) to the posteriors that are obtained by using an approximate waveform model in the GW likelihood. Such approximations are frequently used in

GW astronomy to perform bulk calculations in exploratory data analysis studies; our proposed method then allows users to rapidly obtain corrected credible sets that correspond to some desired coverage level ("correct" relative to a more accurate waveform model).

At present, our proposed method is novel in the GW literature and has no comparable counterpart since other methods for inference calibration (i) focus on correcting the posterior itself in terms of coverage and (ii) deal only with a single posterior at a time. For example, one would have to evaluate the accurate model in bulk in order to use the method proposed by Cutler and Vallisneri [26] for estimating inference biases on the fly. Even if a regression model is fit directly to the biases over the waveform model space, that would still require learning a parameter vector rather than a single coverage number, which would require more complex computation and stringent accuracy requirements on the fit.

In addition to our proposed usage of the method in bulk studies over a space of GW signals, the calibration procedure could potentially also be used on actual data containing a signal. With a model that is pretrained on more accurate waveforms and realistic detector noise, one could rapidly compute calibrated credible sets for the approximate waveforms used in template banks for ground-based observing. This could be useful for applications such as the rapid sky localization of sources for low-latency electromagnetic follow-up. The future third-generation ground-based GW detectors, such as the Einstein Telescope and Cosmic Explorer, will have enhanced sensitivity to gravitational wave signals in the Hz frequency band $f \in (5, 2000)$ Hz. These instruments can exploit the proposed operational coverage discussed in this paper in order to quantify the systematic biases and thus investigate the impact of inaccurate waveforms on tests of GR in an efficient way [87]. The estimated operational coverage can be regarded as a criterion to set requirements for the sensitivity of the detectors to yield unbiased parameter inference for the target system. The forward-thinking calibrated result gives reasonable and meaningful information for future GW research.

The MBH example we presented was restricted to five parameters of the full waveform model. However, the computational complexity of the calibration procedure—in particular, the operational coverage estimator—will typically require a significant increase in computing resources as additional parameters are considered. Increasing the computational efficiency of the operational coverage estimator for high-dimensional problems, and thus improving the scalability of the calibration method to the dimensionalities of both the parameter space and the data representation, is an avenue for future research.

We should point out that the approximate posterior distribution is not too different to the true one. At least, the overlap of the credible intervals exists, which is not always the case. As the divergence between posteriors becomes large, the nominal credible level to achieve any operational coverage goes to 1, which is trivial and manifests the inability to train the model. Therefore, it is encouraged to estimate the maximum operational coverage when obtaining the calibration curve; if it is small, for example, less than 0.5σ, more accurate waveforms should be considered.

The PYTHON code for the one-dimensional toy example in Sec. IV is provided at [88].

ACKNOWLEDGMENTS

J. E. L., M. C. E., and R. M. acknowledge support by the Marsden Grant No. MFP-UOA2131 from New Zealand Government funding, administered by the Royal Society Te Aparangi. A. J. K. C. acknowledges previous support from the NASA LISA Preparatory Science Grant No. 20-LPS20-0005. O. B. acknowledges support from the French space agency CNES in the framework of LISA. He also thanks Sylvain Marsat for giving permission to use `lisa-beta`. All computations are performed on a virtual machine with 32 GB RAM, 16 VCPUs, and an Ubuntu Linux operating system. The autoencoder and ANN were implemented using PYTHON package Tensorflow and SCIKIT-LEARN. We thank the Center for eResearch (CeR) at the University of Auckland for providing access to and assistance with the Nectar Research Cloud.

R. M.: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing—original draft. J. E. L.: Conceptualization, data curation, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing—original draft. O. B.: Conceptualization, data curation, formal analysis, investigation, software: Fisher matrix, MCMC & `lisa-beta`, methodology, resources, software, supervision, validation, visualization, writing—original draft. A. J. K. C.: Conceptualization, formal analysis, methodology, project administration, supervision, writing—original draft, writing—review & editing. M. C. E.: Conceptualization, methodology, project administration, software, visualization, writing—original draft, writing—review & editing. R. M.: Conceptualization, methodology, funding acquisition, project administration, supervision, writing—original draft, writing—review & editing.

APPENDIX: FISHER MATRIX COMPUTATIONS

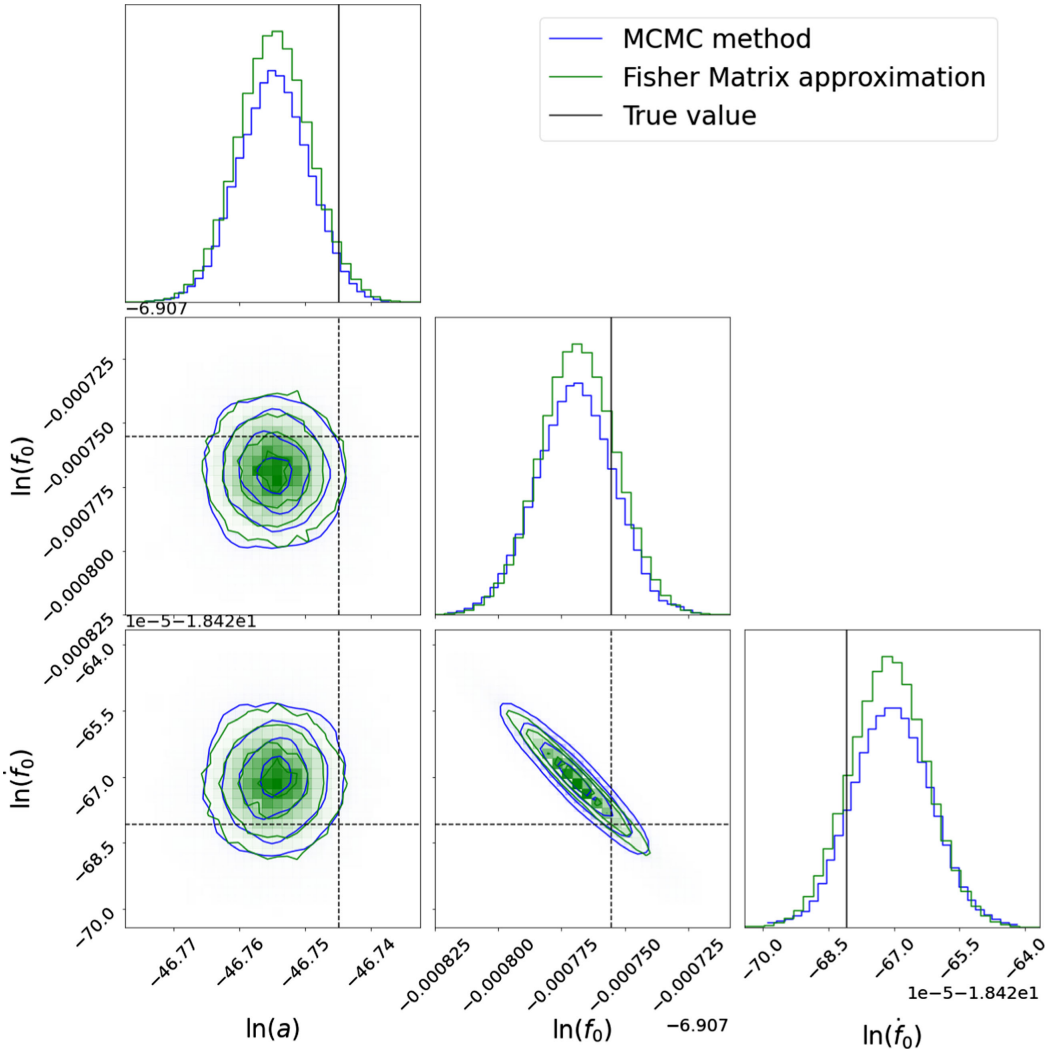

FIG. 11. The blue curve represents a posterior distribution on parameters $\boldsymbol{\theta}$ when inferring parameters of an injected exact signal $h_e(t;\boldsymbol{\theta}_0)$ into Gaussian noise with an approximate model template $h_m(t;\boldsymbol{\theta}, \epsilon = 10^{-6})$. The green curve represents an approximation to the posterior, computed via the Fisher matrix formalism. We highlight here that the computation of the Fisher matrix accurately describes the posterior, implying we are not subject to numerical instabilities when calculating derivatives/inverses. The black line indicates the value of the true parameters in the study.

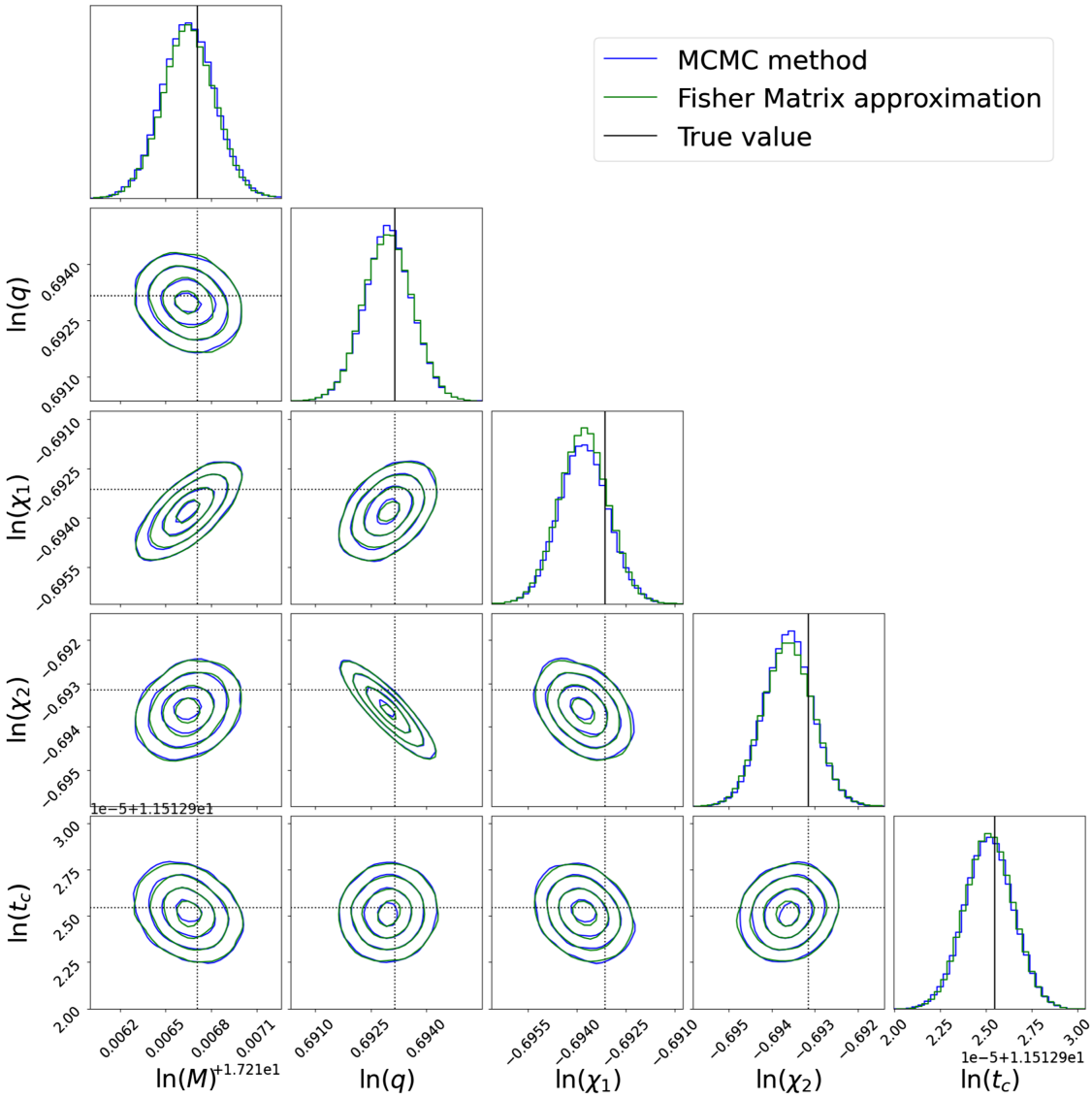

FIG. 12. Comparison of posterior distribution between the MCMC method (blue) and the Fisher Matrix approximation (green) at the massive black hole binary signal d_o, with black vertical/horizontal lines denoting the true parameters. We highlight here an excellent agreement between our MCMC simulation and the approximate Fisher matrix approach. We see that both the fluctuations to recovered parameters induced through noise via Eq. (18) and precision measurements on parameters encoded by (16) are well described by the Fisher matrix formalism.

[1] Benjamin P. Abbott, Richard Abbott, T. D. Abbott, M. R. Abernathy, Fausto Acernese, Kendall Ackley, Carl Adams, Thomas Adams, Paolo Addesso, R. X. Adhikari *et al.*, Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116**, 061102 (2016).

[2] John Veitch, Vivien Raymond, Benjamin Farr, Will Farr, Philip Graff, Salvatore Vitale, Ben Aylott, Kent Blackburn, Nelson Christensen, Michael Coughlin *et al.*, Parameter estimation for compact binaries with ground-based gravitational-wave observations using the lalinference software library, Phys. Rev. D **91**, 042003 (2015).

[3] Benjamin P. Abbott, Rich Abbott, Thomas D. Abbott, Sheelu Abraham, Fausto Acernese, Kendall Ackley, Carl Adams, Vaishali B. Adya, Christoph Affeldt, Michalis Agathos *et al.*,

A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals, Classical Quantum Gravity **37**, 055002 (2020).
[4] Benjamin P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C Adams, T Adams, P Addesso, R. X. Adhikari *et al.*, GW150914: First results from the search for binary black hole coalescence with Advanced LIGO, Phys. Rev. D **93**, 122003 (2016).
[5] Prayush Kumar, Ilana MacDonald, Duncan A. Brown, Harald P. Pfeiffer, Kipp Cannon, Michael Boyle, Lawrence E. Kidder, Abdul H. Mroué, Mark A. Scheel, Béla Szilágyi *et al.*, Template banks for binary black hole searches with numerical relativity waveforms, Phys. Rev. D **89**, 042002 (2014).
[6] Samantha A. Usman, Alexander H. Nitz, Ian W. Harry, Christopher M. Biwer, Duncan A. Brown, Miriam Cabero, Collin D. Capano, Tito Dal Canton, Thomas Dent, Stephen Fairhurst *et al.*, The pycbc search for gravitational waves from compact binary coalescence, Classical Quantum Gravity **33**, 215004 (2016).
[7] P. Ajith, N. Fotopoulos, S. Privitera, A. Neunzert, N. Mazumder, and A. J. Weinstein, Effectual template bank for the detection of gravitational waves from inspiralling compact binaries with generic spins, Phys. Rev. D **89**, 084041 (2014).
[8] T. Adams, D. Buskulic, V. Germain, G. M. Guidi, F Marion, Matteo Montani, B. Mours, Francesco Piergiovanni, and Gang Wang, Low-latency analysis pipeline for compact binary coalescences in the advanced gravitational wave detector era, Classical Quantum Gravity **33**, 175012 (2016).
[9] Nelson Christensen and Renate Meyer, Parameter estimation with gravitational waves, Rev. Mod. Phys. **94**, 025001 (2022).
[10] Renate Meyer, Matthew C. Edwards, Patricio Maturana-Russel, and Nelson Christensen, Computational techniques for parameter estimation of gravitational wave signals, Wiley Interdiscip. Rev. **14**, e1532 (2022).
[11] Lee S. Finn, Detection, measurement, and gravitational radiation, Phys. Rev. D **46**, 5236 (1992).
[12] Mark Miller, Accuracy requirements for the calculation of gravitational waveforms from coalescing compact binaries in numerical relativity, Phys. Rev. D **71**, 104016 (2005).
[13] Luis Lehner and Frans Pretorius, Numerical relativity and astrophysics, Annu. Rev. Astron. Astrophys. **52**, 661 (2014).
[14] Frans Pretorius, Numerical relativity using a generalized harmonic decomposition, Classical Quantum Gravity **22**, 425 (2005).
[15] Luis Lehner, Numerical relativity: A review, Classical Quantum Gravity **18**, R25 (2001).
[16] Michael Boyle, Daniel Hemberger, Dante A. B. Iozzo, Geoffrey Lovelace, Serguei Ossokine, Harald P. Pfeiffer, Mark A. Scheel, Leo C. Stein, Charles J. Woodford, Aaron B. Zimmerman *et al.*, The SXS collaboration catalog of binary black hole simulations, Classical Quantum Gravity **36**, 195006 (2019).
[17] Carlos O. Lousto and Yosef Zlochower, Orbital evolution of extreme-mass-ratio black-hole binaries with numerical relativity, Phys. Rev. Lett. **106**, 041101 (2011).
[18] Jonathan Blackman, Scott E. Field, Chad R. Galley, Béla Szilágyi, Mark A Scheel, Manuel Tiglio, and Daniel A. Hemberger, Fast and accurate prediction of numerical relativity waveforms from binary black hole coalescences using surrogate models, Phys. Rev. Lett. **115**, 121102 (2015).
[19] Alessandra Buonanno and Thibault Damour, Effective one-body approach to general relativistic two-body dynamics, Phys. Rev. D **59**, 084006 (1999).
[20] Alessandro Nagar and Piero Rettegno, Efficient effective one body time-domain gravitational waveforms, Phys. Rev. D **99**, 021501 (2019).
[21] Roberto Cotesta, Sylvain Marsat, and Michael Pürrer, Frequency-domain reduced-order model of aligned-spin effective-one-body waveforms with higher-order modes, Phys. Rev. D **101**, 124040 (2020).
[22] Sebastian Khan, Sascha Husa, Mark Hannam, Frank Ohme, Michael Pürrer, Xisco Jiménez Forteza, and Alejandro Bohé, Frequency-domain gravitational waves from non-precessing black-hole binaries. II. A phenomenological model for the advanced detector era, Phys. Rev. D **93**, 044007 (2016).
[23] Sascha Husa, Sebastian Khan, Mark Hannam, Michael Pürrer, Frank Ohme, Xisco Jiménez Forteza, and Alejandro Bohé, Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal, Phys. Rev. D **93**, 044006 (2016).
[24] Geraint Pratten, Sascha Husa, Cecilio Garcí a-Quirós, Marta Colleoni, Antoni Ramos-Buades, Héctor Estellés, and Rafel Jaume, Setting the cornerstone for a family of models for gravitational waves from compact binaries: The dominant harmonic for nonprecessing quasicircular black holes, Phys. Rev. D **102**, 064001 (2020).
[25] Eanna E. Flanagan and Scott A. Hughes, Measuring gravitational waves from binary black hole coalescences: 2. The Waves' information and its extraction, with and without templates, Phys. Rev. D **57**, 4566 (1998).
[26] Curt Cutler and Michele Vallisneri, LISA detections of massive black hole inspirals: Parameter extraction errors due to inaccurate template waveforms, Phys. Rev. D **76**, 104018 (2007).
[27] Qian Hu and John Veitch, Accumulating errors in tests of general relativity with gravitational waves: Overlapping signals and inaccurate waveforms, Astrophys. J. **945**, 103 (2023).
[28] Andrea Antonelli, Ollie Burke, and Jonathan R. Gair, Noisy neighbours: Inference biases from overlapping gravitational-wave signals, Mon. Not. R. Astron. Soc. **507**, 5069 (2021).
[29] Elia Pizzati, Surabhi Sachdev, Anuradha Gupta, and B. S. Sathyaprakash, Toward inference of overlapping gravitational-wave signals, Phys. Rev. D **105**, 104016 (2022).
[30] Qian Hu and John Veitch, Accumulating errors in tests of general relativity with gravitational waves: Overlapping signals and inaccurate waveforms, Astrophys. J. **945**, 103 (2023).
[31] Pau Amaro-Seoane *et al.*, Laser interferometer space antenna (ESA Publications Division c/o Estec, 2017).

[32] John Baker, Jillian Bellovary, Peter L. Bender, Emanuele Berti, Robert Caldwell, Jordan Camp, John W. Conklin, Neil Cornish, Curt Cutler, Ryan DeRosa *et al.*, The laser interferometer space antenna: Unveiling the millihertz gravitational wave sky, arXiv:1907.06482.
[33] Emanuele Berti, Vitor Cardoso, and Clifford M. Will, Gravitational-wave spectroscopy of massive black holes with the space interferometer LISA, Phys. Rev. D **73**, 064030 (2006).
[34] Scott A. Hughes, Untangling the merger history of massive black holes with LISA, Mon. Not. R. Astron. Soc. **331**, 805 (2002).
[35] Alberto Sesana, Francesco Haardt, Piero Madau, and Marta Volonteri, The gravitational wave signal from massive black hole binaries and its contribution to the LISA data stream, Astrophys. J. **623**, 23 (2005).
[36] Alberto Vecchio, LISA observations of rapidly spinning massive black hole binary systems, Phys. Rev. D **70**, 042001 (2004).
[37] Neil J. Cornish and Edward K. Porter, The search for massive black hole binaries with LISA, Classical Quantum Gravity **24**, 5729 (2007).
[38] Edward K. Porter and Neil J. Cornish, Effect of higher harmonic corrections on the detection of massive black hole binaries with LISA, Phys. Rev. D **78**, 064005 (2008).
[39] Kent Yagi and Leo C. Stein, Black hole based tests of general relativity, Classical Quantum Gravity **33**, 054001 (2016).
[40] Enrico Barausse, Emanuele Berti, Thomas Hertog, Scott A. Hughes, Philippe Jetzer, Paolo Pani, Thomas P. Sotiriou, Nicola Tamanini, Helvi Witek, Kent Yagi *et al.*, Prospects for fundamental physics with LISA, Gen. Relativ. Gravit. **52**, 1 (2020).
[41] Jonathan R. Gair, Michele Vallisneri, Shane L. Larson, and John G. Baker, Testing general relativity with low-frequency, space-based gravitational-wave detectors, Living Rev. Relativity **16**, 1 (2013).
[42] Christopher J. Moore and Jonathan R. Gair, Novel method for incorporating model uncertainties into gravitational wave parameter estimates, Phys. Rev. Lett. **113**, 251101 (2014).
[43] Christopher J. Moore, Christopher P. L. Berry, Alvin J. K. Chua, and Jonathan R. Gair, Improving gravitational-wave parameter estimation using Gaussian process regression, Phys. Rev. D **93**, 064001 (2016).
[44] Alvin J. K. Chua, Natalia Korsakova, Christopher J. Moore, Jonathan R. Gair, and Stanislav Babak, Gaussian processes for the interpolation and marginalization of waveform error in extreme-mass-ratio-inspiral parameter estimation, Phys. Rev. D **101**, 044027 (2020).
[45] Miaoxin Liu, Xiao-Dong Li, and Alvin J. K. Chua, Improving the scalability of Gaussian-process error marginalization in gravitational-wave inference, Phys. Rev. D **108**, 103027 (2023).
[46] Jocelyn Read, Waveform uncertainty quantification and interpretation for gravitational-wave astronomy, Classical Quantum Gravity **40**, 135002 (2023).
[47] Caroline B. Owen, Carl-Johan Haster, Scott Perkins, Neil J. Cornish, and Nicolás Yunes, Waveform accuracy and systematic uncertainties in current gravitational wave observations, Phys. Rev. D **108**, 044018 (2023).
[48] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, An introduction to variational methods for graphical models, Mach. Learn. **37**, 183 (1999).
[49] D. J. C. Mackay, Choice of basis for laplace approximation, Mach. Learn. **33**, 77 (1998).
[50] A. M. Walker, On the asymptotic behaviour of posterior distributions, J. R. Stat. Soc. Ser. B **31**, 80 (1969).
[51] Michele Vallisneri, Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects, Phys. Rev. D **77**, 042001 (2008).
[52] Antoine Klein, Enrico Barausse, Alberto Sesana, Antoine Petiteau, Emanuele Berti, Stanislav Babak, Jonathan Gair, Sofiane Aoudia, Ian Hinder, Frank Ohme, and Barry Wardell, Science with the space-based interferometer eLISA: Supermassive black hole binaries, Phys. Rev. D **93**, 024003 (2016).
[53] Stanislav Babak, Jonathan Gair, Alberto Sesana, Enrico Barausse, Carlos F Sopuerta, Christopher P. L. Berry, Emanuele Berti, Pau Amaro-Seoane, Antoine Petiteau, and Antoine Klein, Science with the space-based interferometer LISA. V. Extreme mass-ratio inspirals, Phys. Rev. D **95**, 103012 (2017).
[54] Christopher P. L. Berry, Scott A. Hughes, Carlos F. Sopuerta, Alvin J. K. Chua, Anna Heffernan, Kelly Holley-Bockelmann, Deyan P. Mihaylov, M Coleman Miller, and Alberto Sesana, The unique potential of extreme mass-ratio inspirals for gravitational-wave astronomy, arXiv:1903.03686.
[55] Monica Colpi, Kelly Holley-Bockelmann, Tamara Bogdanovic, Priya Natarajan, Jillian Bellovary, Alberto Sesana, Michael Tremmel, Jeremy Schnittman, Julia Comerford, Enrico Barausse *et al.*, The gravitational wave view of massive black holes, arXiv:1903.06867.
[56] Jeong Eun Lee, Geoff K. Nicholls, and Robin J. Ryder, Calibration procedures for approximate Bayesian credible sets, Bayesian Anal. **14**, 1245 (2019).
[57] H. Xing, G. Nicholls, and J. Lee, Calibrated approximate Bayesian inference, in *Proceedings of the 36th International Conference on Machine Learning, PMLR* (PMLR, Long Beach, US, 2019), Vol. 97, pp. 6912–6920.
[58] Massimo Tinto and Sanjeev Dhurandhar, Time-delay interferometry, Living Rev. Relativity **24**, 1 (2021).
[59] Massimo Tinto, Sanjeev Dhurandhar, and Dishari Malakar, Second-generation time-delay interferometry, Phys. Rev. D **107**, 082001 (2023).
[60] Tyson B. Littenberg and Neil J. Cornish, Prototype global analysis of LISA data with multiple source types, Phys. Rev. D **107**, 063004 (2023).
[61] Norbert Wiener, Generalized harmonic analysis, Acta Math. **55**, 117 (1930).
[62] Alexander Khintchine, Korrelationstheorie der stationären stochastischen prozesse, Math. Ann. **109**, 604 (1934).
[63] Alexander Iain Burke and Ollie Burke, Extreme precision and extreme complexity: Source modelling and data analysis development for the laser interferometer space antenna, Edinburgh U., 2021.
[64] P. Whittle, Curve and periodogram smoothing, J. R. Stat. Soc. **19**, 38 (1957).
[65] Philip M. Woodward, *Probability and Information Theory, with Applications to Radar*, International Series

of Monographs on Electronics and Instrumentation Vol. 3 (Elsevier, New York, 2014).
[66] George Turin, An introduction to matched filters, IRE Trans. Inf. Theory **6**, 311 (1960).
[67] Daniel Foreman-Mackey, David W. Hogg, Dustin Lang, and Jonathan Goodman, emcee: The MCMC Hammer, Publ. Astron. Soc. Pac. **125**, 306 (2013).
[68] J. K. Pritchard, S. M. T, A. Perez-Lezaun, and M. W. Feldman, Population growth of human y chromosomes: A study of y chromosome microsatellites, Mol. Biol. Evol. **16**, 1791 (1999).
[69] J. Besag, Statistical analysis of non-lattice data, Statistician **24**, 179 (1975).
[70] S. N. Wood, Statistical inference for noisy nonlinear ecological dynamic systems, Nature (London) **466**, 1102 (2010).
[71] P. Fearnhead and D. Prangle, Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation, J. R. Stat. Soc. **74**, 419 (2012).
[72] D. Prangle, M. G. B. Blum, G. Popovic, and S. A. Sisson, Diagnostic tools for approximate Bayesian computation using the coverage property, Aust. N. Z. J. Stat. **56**, 309 (2014).
[73] Y. Yao, A. Vehtari, D. Simpson, and A. Gelman, Yes, but did it work?: Evaluating variational inference, In *In Proceedings of the 35th International Conference on Machine Learning* (PMLR, Stockholm, Sweden, 2018), Vol. 80, pp. 5581.
[74] J. Geweke, Getting it right: Joint distribution tests of posterior simulators, J. Am. Stat. Assoc. **99**, 799 (2004).
[75] S. R. Cook, A. Gelman, and D. B. Rubin, Validation of software for Bayesian models using posterior quantile, J. Comput. Graph. Stat. **15**, 675 (2006).
[76] H. Xing, G. Nicholls, and J. Lee, Distortion estimates for approximate Bayesian inference, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, edited by Jonas Peters and David Sontag, Proceedings of Machine Learning Research Vol. 124 (PMLR, 2020), pp. 1208–1217.
[77] G. E. Hintonand and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science **313**, 504 (2006).
[78] Quentin Fournier and Daniel Aloise, Empirical comparison between autoencoders and traditional dimensionality reduction methods, in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (IEEE, Sardinia, Italy, 2019), pp. 211–214.
[79] Jun Xu and J Scott Long, Using the delta method to construct confidence intervals for predicted probabilities rates, and discrete changes, Stata J. **5**, 537 (2005).
[80] Bradley Efron and Robert Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Stat. Sci. **1**, 54 (1986).
[81] Yoshua Bengio and Yves Grandvalet, No unbiased estimator of the variance of k-fold cross validation, J. Mach. Learn. Res. **5**, 1089 (2004).
[82] P. Menendez, Y. Fan, P. Garthwaite, and S. Sisson, Simultaneous adjustment of bias and coverage probabilities for confidence intervals, Computational Statistics and Data Analysis **70**, 35 (2014).
[83] Travis Robson, Neil J. Cornish, and Chang Liu, The construction and use of LISA sensitivity curves, Classical Quantum Gravity **36**, 105011 (2019).
[84] Diederik P. Kingma and Jimmy Ba, Adam: A method for stochastic optimization, 2014, https://api.semanticscholar.org/CorpusID:6628106.
[85] Sylvain Marsat, John G. Baker, and Tito Dal Canton, Exploring the Bayesian parameter estimation of binary black holes with LISA, Phys. Rev. D **103**, 083011 (2021).
[86] Sylvain Marsat and John G. Baker, Fourier-domain modulations and delays of gravitational-wave signals, arXiv:1806.10734.
[87] Michael Pürrer and Carl-Johan Haster, Gravitational waveform accuracy requirements for future ground-based detectors, Phys. Rev. Res. **2**, 023151 (2020).
[88] https://github.com/bpandamao/calibration_case_study.