

Physical approach to the marginalization of LIGO calibration uncertainties

Salvatore Vitale^{1,2,*} Carl-Johan Haster^{3,2,†} Ling Sun^{4,5} Ben Farr⁶ Evan Goetz^{4,7}
 Jeff Kissel⁸ and Craig Cahillane⁴

¹*LIGO Laboratory, Massachusetts Institute of Technology, 185 Albany Street Cambridge, Massachusetts 02139, USA*

²*Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA*

³*LIGO Laboratory, Massachusetts Institute of Technology, 185 Albany Street, Cambridge, Massachusetts 02139, USA*


⁴*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*

⁵*OzGrav-ANU, Centre for Gravitational Astrophysics, College of Science, The Australian National University, Australian Capital Territory 2601, Australia*

⁶*Department of Physics, University of Oregon, Eugene, Oregon 97403, USA*

⁷*University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada*

⁸*LIGO Hanford Observatory, Richland, Washington 99352, USA*

 (Received 23 September 2020; accepted 16 February 2021; published 15 March 2021)

The data from ground-based gravitational-wave detectors such as Advanced LIGO and Virgo must be calibrated to convert the digital output of photodetectors into a relative displacement of the test masses in the detectors, producing the quantity of interest for inference of astrophysical gravitational-wave sources. Both statistical uncertainties and systematic errors are associated with the calibration process, which would in turn affect the analysis of detected sources, if not accounted for. Currently, source characterization algorithms either entirely neglect the possibility of calibration uncertainties or account for them in a way that does not use knowledge of the calibration process itself. We present *physiCal*, a new approach to account for calibration errors during the source characterization step, which directly uses all the information available about the instrument calibration process. Rather than modeling the overall detector's response function, we consider the individual components that contribute to the response. We implement this method and apply it to the compact binaries detected by LIGO and Virgo during the second observation run, as well as to simulated binary neutron stars for which the sky position and distance are known exactly. We find that the *physiCal* model performs as well as the method currently used within the LIGO-Virgo Collaboration, but additionally it enables improving the measurement of specific components of the instrument control through astrophysical calibration.

DOI: [10.1103/PhysRevD.103.063016](https://doi.org/10.1103/PhysRevD.103.063016)

I. INTRODUCTION

The advanced gravitational-wave (GW) detectors LIGO [1,2] and Virgo [3] have concluded their third observation run as of March 2020, reporting the detection of 56 candidate GW sources [4], most of which, if confirmed, are binary black holes (BBHs). Owing to planned increases in sensitivity for LIGO and Virgo, and the addition of the Japanese detector KAGRA [5] to the global network, the detection rate will be even higher in the next few years [6]. Having access to a large number of GW sources will allow for unprecedented measurements of the mass and spin distribution of compact objects, as well as their formation channels [7]. The potential of detecting many binary

neutron star mergers (BNSs) together with electromagnetic (EM) counterparts opens the way to precise measurements of the Hubble constant [8–12]. Some of the detected sources will have a high signal-to-noise ratio (SNR), which would enable precise tests of general relativity and of the nature of individual objects.

For gravitational-wave astrophysics to fulfill its potential, one must control all of the (known) sources of systematics. In this work we focus on instrumental calibration uncertainties. The complex function that relates the voltage measured at the output of LIGO and Virgo photodetectors to the strain needed for astrophysical inference is the response function, $R(f)$. In the Fourier domain, the relation between these quantities is simply

$$d(f) \equiv \frac{\Delta L}{L} = R(f)v(f), \quad (1)$$

*salvo@mit.edu
 †haster@mit.edu

where $v(f)$ is the photodetector readout, $d(f)$ is the gravitational-wave strain, ΔL is the differential arm (DARM) displacement of the mirrors, and L is the nominal length of the interferometer arms [13]. The calibration process includes collecting a set of measurements performed on the detectors to inform a reference model of the response function, $R^{(\text{model})}$ [14]. This includes tracking the slow time dependence of the detector response with respect to that model [15]; using that model to create a near-real-time data stream as an estimate of $d(f)$ at any time [16]; and characterizing the systematic error and statistical uncertainty in the model, or equivalently in the data stream used for astrophysical analysis [17]. The fundamental reference fiducials for the calibration process are independent laser systems, called *photon calibrators* (Pcal), to calibrate LIGO and Virgo by applying a known radiation pressure directly into the test masses [14,18,19]. Errors, bias, or uncertainty in any part of this calibration process to develop the estimated strain (including that in the Pcal systems) directly affect the strain, and hence if unaccounted for, bias the estimation of the source parameters. Reference [20] has shown how the parameters that would suffer the largest biases are those mostly related to the amplitude of GW signals. For compact binaries coalescences (CBCs), those would be luminosity distance (D_L), orbital inclination (i), and sky position. In turn, those parameters are related to some of the key science goals mentioned above: identification of an EM counterpart and cosmology.

Statistical uncertainties and systematic errors in the measurement of the response function result in both amplitude and phase offsets, so that the model response function at a specific time and frequency is related to the true response function by

$$R^{(\text{true})}(f, t) = (1 + \delta A(f, t))e^{i\delta\phi(f, t)}R^{(\text{model})}(f, t), \quad (2)$$

where δA is the *relative* amplitude error and $\delta\phi$ the phase error. In turn, this affects the GW data strain as

$$d^{(\text{true})} = d^{(\text{model})}(1 + \delta A(f, t))e^{i\delta\phi(f, t)}, \quad (3)$$

where $d^{(\text{model})}$ is the calibrated data strain produced using the model response function.

Here we are explicitly reporting a time dependence to stress that the behavior of GW detectors, and hence their transfer functions, varies over timescales of minutes [15]. Therefore, while it is generally a good approximation to treat the response function as a constant in time (not in frequency) when analyzing a single CBC event, since its duration will usually be shorter than 2 minutes (for a BNS detected by advanced detectors), one should not assume that the response function is the same throughout an observing run. In fact, the response function of the LIGO and Virgo detectors is characterized continuously

in a few small frequency bins throughout the run, and across all frequencies weekly, as a precaution against unexpected changes [17,21].

Currently, the results presented by the LIGO-Virgo Collaboration (LVC) obtained with the LALInference [22] or BILBY [23,24] source characterization algorithms marginalize over calibration errors with a spline interpolant informed by the frequency-dependent 68% credible interval contours of the systematic error and uncertainty in each response function [25] (henceforth *splineCal* method). While that approach has the advantage of accounting for calibration uncertainties, it also has some limitations. First, it introduces a significant number of nuisance parameters that must be marginalized over numerically: roughly 20 parameters per interferometer. Second, the frequencies at which the spline points are anchored do not use any information about characteristic physical correlation lengths in the instrument (they are simply chosen uniformly in log space). Third, the spline marginalization method treats the uncertainties in the phase and amplitude of the response function as independent and uncorrelated. Fourth, should any constraints be placed on the response function through a so-called astrophysical calibration (see below) it would be hard or impossible to relate those constraints to specific components of the detector.

In this paper we propose a new approach to account for uncertainties in the response function, which builds upon recent progress in measurement and modeling of the response function, and does not suffer from the same limitations of the spline approach. We implement the new method, called “physical calibration” (henceforth, *physiCal*) in the LALInference software and apply it to the CBCs detected by LIGO and Virgo in their second observing run, as well as on simulated binary neutron star sources.

The rest of this paper is organized as follows: in Sec. II A we summarize the measurements and algorithms used to calibrate the LIGO instruments; in Sec. II B we present the implementation of the *physiCal* method; in Sec. III A and Sec. III B we report results from the analysis of LIGO-Virgo sources and simulated signals, respectively; finally in Sec. IV we summarize the main conclusions.

II. METHOD

A. Calibration physical model

While a full description of systematic error and uncertainty in the calibration of the LIGO detectors is beyond the scope of this paper, we will review the main points and refer the interested reader to Ref. [17] for more details.

In the frequency domain, the complex-valued detector response can be written as

$$R(f) = \frac{1}{C(f)} + A(f)D(f). \quad (4)$$

The sensing function C converts the suppressed DARM residual displacement¹ to digitized photodetector output signals. The actuation function A converts the requested digital control signal to the force applied to the test masses, producing a control displacement meant to suppress the DARM displacement. The total A function consists of three actuation stages, the upper intermediate (U), penultimate (P), and test mass (T) stages in the quadruple suspension [26]. The function D represents a set of digital, feedback control filters, which can be assumed as perfectly known. The DARM strain, and thus the calibrated data in Eq. (1), are reconstructed using the modeled sensing and actuation functions, $C^{(\text{model})}$ and $A^{(\text{model})}$, in the detector calibration pipeline. Here $A^{(\text{model})}$ denotes the model of the total A function, in which each stage A_a is modeled independently ($a = U, P, T$). The time-dependent, frequency-dependent systematic errors on our model of the response function are written as

$$\eta_R = \frac{R^{(\text{true})}}{R^{(\text{model})}}, \quad (5)$$

where $R^{(\text{true})}$ is the true detector response and $R^{(\text{model})} = 1/C^{(\text{model})} + A^{(\text{model})}D$ is the modeled response [17]. The relative amplitude error and phase error in Eq. (3) can thus be written as

$$\begin{aligned} \delta A &= |\eta_R| - 1, \\ \delta \phi &= \angle \eta_R. \end{aligned} \quad (6)$$

where $\angle z$ indicates the phase of the complex number z . Throughout the observing run, η_R and its associated uncertainty is evaluated at a 1 hour cadence.

The models $C^{(\text{model})}$ and $A^{(\text{model})}$ contain many parameters representing the entire DARM control loop, from the basic properties of signal processing electronics to complex actuator dynamics. Most parameters can be measured independently to high precision and do not dominantly contribute to the systematic error and/or uncertainty in $R^{(\text{model})}$. However, a set of physical parameters related to specific properties of the instrument must be determined from interferometric measurements taken while the detectors are in the most sensitive, nominal operational state [17]. These parameters, discussed as follows, highly depend on the loosely controlled alignment and thermal state of the detector and may vary slowly over time. Hence they are difficult to measure and are likely to introduce systematic errors in the calibration model. For the sensing function, we write the physical parameter vector as

$$\lambda^C = [H_C, f_{cc}, f_s, Q, \delta\tau_C], \quad (7)$$

where H_C is the overall gain of the sensing function, f_{cc} is the differential coupled-cavity pole frequency, f_s and Q are, respectively, the pole frequency and quality factor of an optical springlike response of any detuning between the coupled Fabry-Pérot arm cavities and signal recycling cavity [27], and $\delta\tau_C$ is the residual time delay in C . For the a th stage of the actuation function ($a = U, P, T$), the physical parameter vector is

$$\lambda_a^A = [H_a, \delta\tau_a], \quad (8)$$

where H_a is the overall gain for the a th stage actuator and $\delta\tau_a$ is the residual time delay in that stage. Some parameters in C and A vary slowly over time, on a timescale of minutes to days, due to various physical mechanisms [28]. The overall gain variation of H_C is tracked by a real-valued scalar factor $\kappa_C(t)$. Parameters f_{cc} , f_s , and Q in the sensing function are also time varying. The variation of H_a ($a = U, P, T$) is tracked by scalar factors $\kappa_U(t)$, $\kappa_P(t)$, and $\kappa_T(t)$ for each corresponding actuation stage. A full description of C and A , as well as all the time-independent and time-dependent factors therein is given in Ref. [17].

While $R^{(\text{model})}$ does an excellent job at reproducing $R^{(\text{true})}$, the residual systematic error η_R and its uncertainty need to be quantified through the frequency-dependent, time-independent residuals $\eta_C = C^{(\text{true})}/C^{(\text{model})}$ and $\eta_{A_a} = A_a^{(\text{true})}/A_a^{(\text{model})}$, where $C^{(\text{true})}$ and $A^{(\text{true})}$ are the true sensing and actuation functions inferred from large collections of interferometric measurements, and the subscript a indexes the actuation stages ($a = U, P, T$). This set of residuals is computed via Gaussian process regression (GPR) [29,30], by taking into account potential model-agnostic but physically motivated frequency-dependent correlations. The posterior results from the GPR indicate the residual errors and uncertainties in the sensing and actuation models. In a perfect calibration model, η_C , η_{A_a} , and hence η_R are at unity in magnitude and zero in phase.

At any given time t , measurements of the various physical quantities that we have just described and that affect the response function (which we will collectively refer to as *physiCal* parameters from now on) are used to assess the complex-valued, frequency-dependent systematic error in the detector response and its associated uncertainty. Using interferometric measurements, we apply Markov chain Monte Carlo (MCMC) methods to obtain the posterior probability density functions (PDFs) of λ^C and λ_a^A ($a = U, P, T$). The *physiCal* parameters are estimated jointly within each vector, i.e., the posterior probability density can be written as

$$p(\lambda^C, \lambda_U^A, \lambda_P^A, \lambda_T^A) = p(\lambda^C) \prod_{a \in \{U, P, T\}} p(\lambda_a^A). \quad (9)$$

¹That is, the residual differential displacement of the mirrors after the control signal has been applied; see, e.g., Fig. 3 of Ref. [17].

The maximum *a posteriori* values of λ^C and λ_a^A are used to form the model functions $C^{(\text{model})}$ and $A^{(\text{model})}$, and thus $R^{(\text{model})}$. Since λ^C and λ_a^A are time varying, the time-dependent corrections are taken into consideration when constructing $R^{(\text{model})}$ for any given analysis time. We use 10^4 fair draws from the posterior PDFs of λ^C and λ_a^A to create a distribution of draws from R as described below. These R samples, once divided by $R^{(\text{model})}$, yield a posterior distribution for $\eta_R(f; t)$. The generic i th sample for the inferred response function posterior reads [17]

$$R_i(f; t) = \eta_{\text{Pcal}_i} \left[\frac{1}{\eta_{C_i}(f) C(\lambda_i^C; f; t)} + \eta_{A_i}(f) A(\lambda_i^A; f; t) D(f) \right]. \quad (10)$$

The samples for the sensing and actuation functions $C(\lambda_i^C; f; t)$ and $A(\lambda_i^A; f; t)$ are derived from the MCMC posterior distributions of λ^C and λ_a^A ($a = U, P, T$). The samples $\eta_{C_i}(f)$ and $\eta_{A_i}(f)$ are, respectively, drawn from the GPR posterior distributions. Here in Eq. (10), we do not explicitly split out the three stages in A , and use i to denote the sample in total A for simplicity. In practice, the samples in each stage of A are drawn independently. The 1σ uncertainties of the time-dependent factors applied in C and A at time t are taken into account. The real-valued scale factor η_{Pcal} accounts for the systematic error and uncertainty of the photon calibrator, common to all interferometric measurements in a detector.

The median frequency-dependent value of the 10^4 samples from the distribution of $\eta_R(f; t)$ represents our best estimate for the systematic difference between $R^{(\text{true})}$ and $R^{(\text{model})}$ at time t , and thus the systematic error in the calibrated data $d(f; t)$. Meanwhile, the 16th and 84th percentiles represent the bounds of systematic error and 1σ statistical uncertainty in the modeled detector response, and thus $d(f; t)$.

For each of the LIGO detectors, we perform the above procedure and store to file the 10^4 posterior samples from the posteriors of the *physiCal* parameters, together with the resulting posterior samples for the frequency-dependent response function, Eq. (10). The Virgo detector does not have as sophisticated an infrastructure, but the detector response can be modeled in the same way [31]. The next section describes how these are used in the source characterization algorithm.

B. Implementation in source characterization code

Given a stretch of interferometric data d containing a CBC signal, one wants to estimate the posterior distribution of the unknown source parameters θ (masses, spins, sky position, etc. See, e.g., Ref. [22]). Bayes theorem allows us to write the posterior probability density as

$$p(\theta|d) = \frac{p(d|\theta)\pi(\theta)}{p(d)}, \quad (11)$$

where $\pi(\theta)$ is the prior distribution of the CBC parameters (in what follows we will use the standard priors used by the LVC [32]) and $p(d)$ is the evidence of the data, which will not play a role in parameter estimation [22]. The remaining term is the likelihood of the data given θ . If one assumes that the interferometric noise is stationary and Gaussian, then the likelihood in the Fourier domain reads

$$p(d|\theta) \propto e^{-\langle d|d \rangle - \langle d|\theta \rangle - \langle \theta|\theta \rangle}, \quad (12)$$

where we have defined the noise-weighted inner product

$$\langle a|b \rangle \equiv 2 \int df \frac{ab^* + a^*b}{S(f)} \quad (13)$$

and $h(f, \theta)$ is the gravitational-wave template calculated at θ . The likelihood weights the difference between data and GW template (i.e., the data residuals) by the noise power spectral density (PSD) $S(f)$ [33,34], i.e., the noise auto-correlation. These expressions are written for a generic interferometer, and since noise is expected to be uncorrelated between detectors, it is extended to a network by taking the product of likelihoods calculated for each interferometer [22].

If one wants to explicitly account for statistical uncertainties and systematic errors in the response function, the likelihood in Eq. (12) needs to be modified by correcting the data, Eq. (3), or—which is equivalent [25]—by modifying the GW template $h(f, \theta)$:

$$h(f, \theta) \rightarrow h(f, \theta) (1 + \delta A(\lambda^A, \lambda^C, f)) e^{i\delta\phi(\lambda^A, \lambda^C, f)}. \quad (14)$$

As mentioned in Sec. II A, the calibration pipelines produce draws from the posterior distribution of the response function errors, which can be used to obtain frequency-dependent medians and standard deviations for amplitude and phase errors, δA and $\delta\phi$. Current LVC results are produced by only using these medians and 1-sigma uncertainties to inform the position and width of the Gaussian priors of the calibration spline points [32,35].

Instead, we wish to augment LALInference so that it can directly use *individual* draws from $R(f)$, i.e., for δA and $\delta\phi$ as defined by Eq. (6). In addition to the interferometer-dependent amplitude errors, we include the possibility of a *common* offset in the amplitude of the response functions of both LIGO detectors introduced by the calibration of LIGO's Pcal lasers against the same reference from the National Institute of Standards [18]. We will use the variable η_{NIST} to indicate this common offset.

We will thus work with the following template for the likelihood of LIGO's data:

$$h^I(f, \boldsymbol{\theta}) \rightarrow \eta_{\text{NIST}} h^I(f, \boldsymbol{\theta}) [1 + \delta A^I(\boldsymbol{\lambda}^A, \boldsymbol{\lambda}^C, f)] \times e^{i\delta\phi^I(\boldsymbol{\lambda}^A, \boldsymbol{\lambda}^C, f)}, \quad (15)$$

where an index $I = H$ (Hanford) or L (Livingston) is used to label quantities which are instrument dependent.

To run a source characterization analysis on a CBC event detected at some time t we thus proceed in two steps. First, we build the distribution of frequency dependent systematic error, η_R , described in Sec. II A for each of the LIGO detectors at time t . As described above, this produces a file with 10^4 samples from the posteriors of the *physiCal* parameters and their corresponding response function, which, given $R^{(\text{true}),I}$ at time t , can be recast into posteriors for δA^I and $\delta\phi^I$ following Eqs. (5) and (6). We then deploy a modified version of `LALInference` to generate posterior PDFs for both the CBC parameters $\boldsymbol{\theta}$ and the *physiCal* parameters.

More specifically we modify the likelihood function, priors, and the sampler of `LALInference` so that it can use the files containing η_R^I and $R^{I,(\text{model})}$ directly. For each of the LIGO detectors:

- (i) We load to memory the corresponding *physiCal* file. We label each of the samples produced by the calibration pipeline with an integer from 1 to 10^4 .
- (ii) We introduce a new sampling parameter, an integer between 1 and 10^4 , and assign it a uniform prior. We call it the *physiCal* ID of this interferometer.

The common η_{NIST} parameter is assigned a uniform prior in the range $[-0.9914, 1.0086]$ consistent with the uncertainties on the calibration of the LIGO photon calibrators at the time of our analysis. With these changes implemented the parameter estimation algorithm proceeds as usual: at each iteration of the MCMC chain (or update of a nested sampling live point [22,36]), we update $\boldsymbol{\theta}$, the calibration *physiCal* IDs, η_{NIST} , calculate the modified waveform templates for each interferometer, and hence the corresponding likelihood. Our updates allow the user to use a different calibration marginalization scheme (*splineCal*, *physiCal*, no marginalization) for each detector independently when running a network analysis. For the runs described in the remainder of this paper we only use the *physiCal* method for the LIGO detectors and the spline method for Virgo. In total our scheme introduces a single new parameter for each instrument for which the *physiCal* method is used, plus η_{NIST} . This should be compared with the ~ 20 new parameters used for each instrument if the spline method is used.

III. RESULTS

A. Analysis of LIGO-Virgo's sources

In this section we apply the *physiCal* method to all of the CBCs found by the LVC during their second observing run (O2), using the corresponding public data

release [32,35,37].² LIGO-only data are available for GW170104, GW170608, and GW170823, whereas LIGO-Virgo data are available for GW170729, GW170809, GW170814, GW170817, and GW170818.

The Bayesian priors on the CBC parameters are chosen to match those used by the LVC, whereas the priors on the *physiCal* parameters have been described in the previous section. We use the `IMRPhenomPv2` waveform approximant [38–40] for all BBH analyses, with the reduced order quadrature (ROQ) likelihood implementation [41], while we use `IMRPhenom_NRTidal` [38–40,42,43] for the binary neutron star merger GW170817.³

For all sources, we find that the posterior distributions of the astrophysical parameters $\boldsymbol{\theta}$ obtained with the *physiCal* method are virtually indistinguishable from those reported by the LVC using the *splineCal* method. To quantify the level of similarity, we compute the Jensen-Shannon (JS) divergence [44], a general symmetrized extension of the Kullback-Leibler divergence [45], between the two sets of one-dimensional probability distributions. The JS divergence is defined between 0 bits of information difference (i.e., the distributions are statistically identical) and 1 bit (no statistical overlap). The maximum JS divergence for the astrophysical parameters inferred from the O2 LVC observations is calculated to be 0.012 bits, with the vast majority of divergences more than an order of magnitude smaller than this (cf. Table IV of [46] where a similar conclusion is reached). Hence, we conclude that the *physiCal* and *splineCal* methods recover posterior distributions that are similar enough that no astrophysical statement would depend on the method used. For example, in Fig. 1 we show the marginal posterior distribution of the luminosity distance of GW170729, the most distant of the sources in the GWTC-1 catalog, obtained with *physiCal* and with the spline method.

This can be explained by noticing that for both the spline and the *physiCal* methods no constraints can be placed on any of the calibration parameters, and the respective priors are recovered. Since the priors are informed by the same underlying calibration model, the two approaches yield consistent results. The CBC sources LIGO and Virgo detected in O2 [32] had network SNRs in the range $\sim [10, 33]$. This suggests that even higher SNRs and/or some auxiliary information about the sources is needed to constrain the *physiCal* parameters (see Sec. III B and Sec. IV of [46]). The authors of Ref. [47] analyzed the BNS GW170817 with a different approach, and similarly

²We cannot reanalyze the sources detected in the first observing run, since the distribution of systematic error and uncertainty, η_R , was not recorded.

³The ROQ likelihood in `LALInference` is distinct from the likelihood that is used for most waveform families. Our implementation of the *physiCal* method works for both the ROQ and the “classic” likelihood.

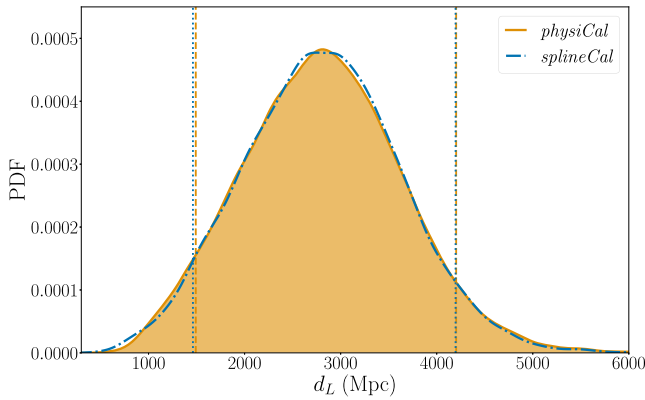


FIG. 1. Marginal posterior density function for the luminosity distance of GW170729 inferred using the LVC’s spline marginalization of the calibration uncertainty (*splineCal* [25,35]) and the *physiCal* method described in this work. The vertical lines denote the 90% credible interval for each analysis.

found that nothing can be learned about the response function.

In Fig. 2 we show a comparison of the posteriors for the response function’s errors when analyzing GW170814.

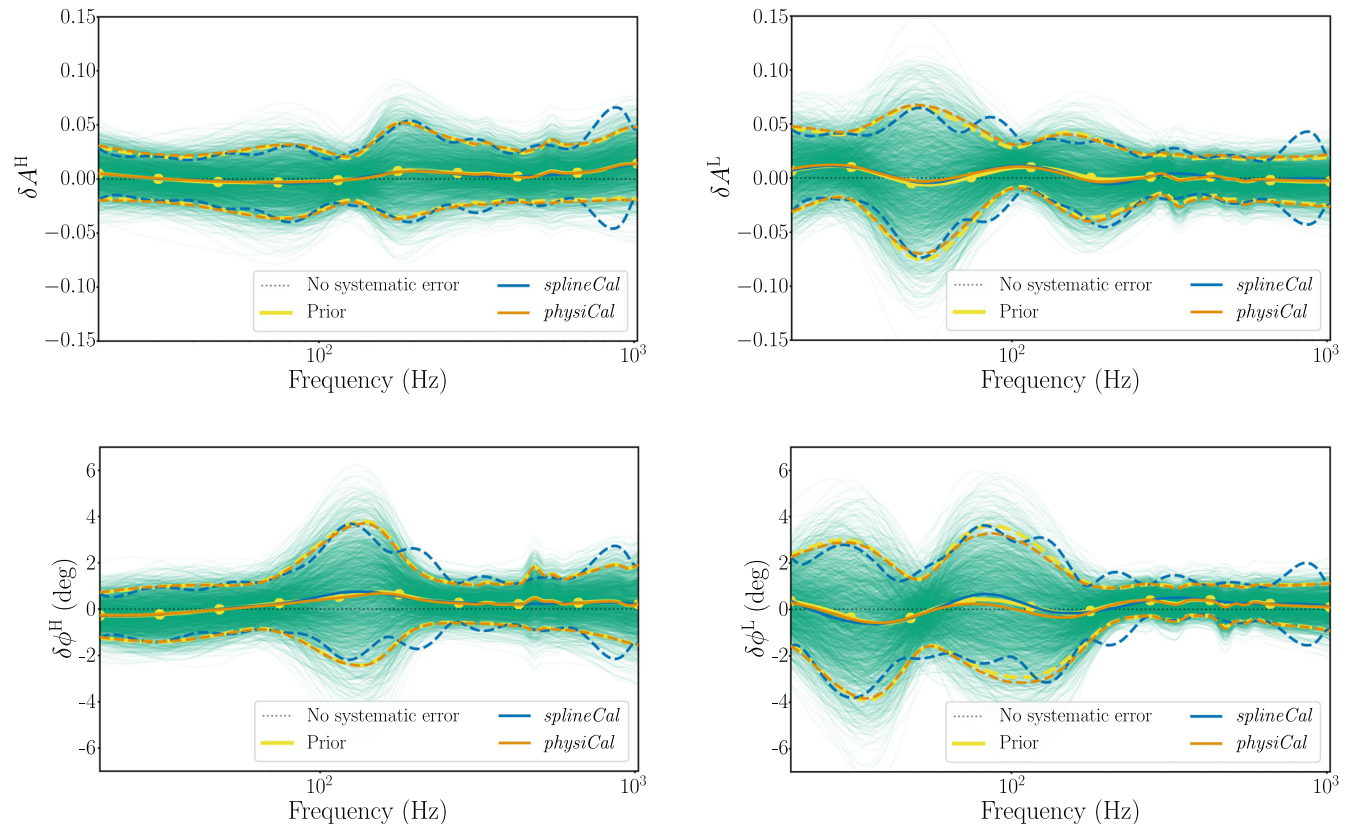


FIG. 2. PDFs for the amplitude (top) and phase (bottom) of the response function’s errors at the time of GW170814 for the LIGO Hanford (left) and LIGO Livingston (right) detectors. The gray dotted line indicates the ideal case with no systematic error. All PDFs are represented by their median (solid line) and 90% credible interval (dashed lines). The prior distributions are shown in yellow. The yellow dots indicate the frequencies where the *splineCal* variables are defined. The *splineCal* posteriors are shown in blue and the *physiCal* posteriors are shown in orange. For the *physiCal* method, we also show 2048 individual draws from the posteriors (green semitransparent curves).

Amplitude and phase errors are reported—for the two LIGO detectors individually—in the top and bottom rows, respectively. The blue lines refer to the spline method and the orange lines to the *physiCal* method. In both cases, the solid lines are the medians and the dashed lines mark the 90% credible intervals. For the *physiCal* method, we also show 2048 individual draws from the posteriors (semi-transparent green curves).

B. Simulated events

The results we obtained for the O2 sources show that with the “typical” CBC source of medium-low SNR for which most or all of the astrophysical parameters are unknown, no information can be gained about the *physiCal* parameters, and we just recover the priors. This can likely be attributed to the fact that calibration errors mostly affect the amplitude of the response function, and hence the signal [20]. On the other hand, analysis of CBC signals cannot usually constrain amplitude parameters (mainly distance and orbital inclination) as precisely as parameters that affect the phase evolution of the system (e.g., masses and spins) [32]. Furthermore, the *physiCal* parameters have Gaussian priors, which are

TABLE I. The true values of some selected parameters for the two BNS sources described in Sec. III B. An index H , L or V is used to refer to the LIGO Hanford, LIGO Livingston or Virgo detector respectively.

ID	$m_1[M_\odot]$	$m_2[M_\odot]$	D_L [Mpc]	t	ι [rad]	SNR_H	SNR_L	SNR_V
1	1.98	1.78	58.8	1167560557.32	0.22	21.4	21.9	n/a
2	1.99	1.69	74.4	1187057243.40	0.71	13.5	25.9	3.3

stronger than the priors of the CBC amplitude parameters (isotropic of the orbital inclination, uniform in comoving volume for the distance) [48–50]. This implies that even if the response function were off, it might be “easier” for a Bayesian algorithm to compensate for it by biasing distance and inclination, which might come at a smaller prior expense. We plan to thoroughly explore the topic of biases in a future publication.

On the other hand, if extra astrophysical information is obtained that better constrains CBC parameters that are correlated with calibration parameters (e.g., sky location, distance), better constraints on the *physiCal* parameters could be possible. While the idea of “astrophysical calibration,” i.e., of learning something about the detector using particularly loud or otherwise exceptional events, is not new [47,51], we stress that the best one can do using the spline approach is to verify that something is wrong with the overall response function. With the *physiCal* method instead, we can hope to say something about specific parts of the sensing and actuation systems, as described in Sec. II A above and Refs. [16,17,21].

To test this we add 200 simulated BNSs into *real* LIGO-Virgo interferometric data from O2 [37] (we only consider BNSs and not BBHs because we will want to assume the source extrinsic parameters can be constrained with EM data; see below). The signals’ merger times are randomly chosen to be in the 3600 seconds preceding or following the eight CBC sources detected in LIGO-Virgo’s second observing run.⁴ Rather than producing the full distribution of η_R for each simulated event, we reuse the distributions at the time of the eight O2 sources. For each of the simulated signals we thus use the output of the calibration pipeline as calculated for the nearest of the O2 sources. This implies that the largest possible time interval between the time a simulated signal is added into the data and the assigned O2 event time for which its η_R was produced is one hour. The simulated events are assigned random sky positions and orbital orientation, and are placed uniformly in comoving volume. This implies that the resulting SNRs are representative of realistic detections in the second and third observing runs (i.e., with network SNRs in the approximate range [10, 40] and with most sources having SNR near the minimum). For these analyses we use the IMRPhenomPv2 waveforms to simulate the signals that are added both into the data and for parameter estimation. The neutron stars are

⁴If the simulated signal precedes the O2 detection, we leave enough time between them to avoid overlaps.

assigned randomly oriented spins with (dimensionless) magnitude uniform in the range [0, 0.2] and component masses uniform in the range [1.8–2.4] M_\odot .⁵ We do not include tidal effects either in simulating signals or in the subsequent source characterization analysis.

To mimic a situation where a successful electromagnetic counterpart has been found, which yields the source’s 3D position, we run the source characterization algorithm by assuming that the sky position *and* the luminosity distance of the sources are perfectly known. This neglects potential uncertainties introduced by the cosmology used to convert the source redshift into a luminosity distance; however, here we are interested in a somewhat optimistic scenario to show what this method can theoretically do. If, as it is more realistic, the distance to the source is only known within an uncertain range, the overall amplitude parameter η_{NIST} would not be constrained. While it is possible to obtain some constraints about the source’s orbital inclination by folding in external information about the source [52,53], that inference would not be very precise and would depend on detailed modeling of the EM emission. Therefore, instead of assuming the inclination angle is perfectly known, we restrict its prior to a $\pm 20^\circ$ interval symmetric around the true value excluding unphysical values (i.e., $\iota < 0$ rads and $\iota > \pi$). Having fixed luminosity distance and sky position to their true values, the inclination angle is thus the only CBC parameter that significantly affects the amplitude of the signals in our analysis.⁶ It is worth stressing that even for LIGO-only analyses, η_{NIST} is not perfectly degenerate with the (cosine of the) inclination angle, since the latter affects the two GW polarizations each in a different way [33], while the former is an overall amplitude offset. This would be different if the luminosity distance were also a free parameter, since in that case η_{NIST} and distance would be perfectly degenerate in a LIGO-only analysis, and only the combination η_{NIST}/D_L would be measurable.

We will not report extensively on these simulations because for the overwhelming majority of them, owing to the low SNRs, nothing is learned about the *physiCal* parameters. Instead, we will just focus on two high-SNR signals, one in LIGO-Virgo data, and the other in LIGO data. The true values of some of their parameters are

⁵This range of mass was not chosen to be representative of a realistic mass distribution, but rather to optimize the runtime of LALInference with the ROQ likelihood.

⁶Intrinsic parameters also affect the GW amplitude. However, they are usually measured from the GW phase well enough that they can be thought as known when considering the signal’s amplitude.

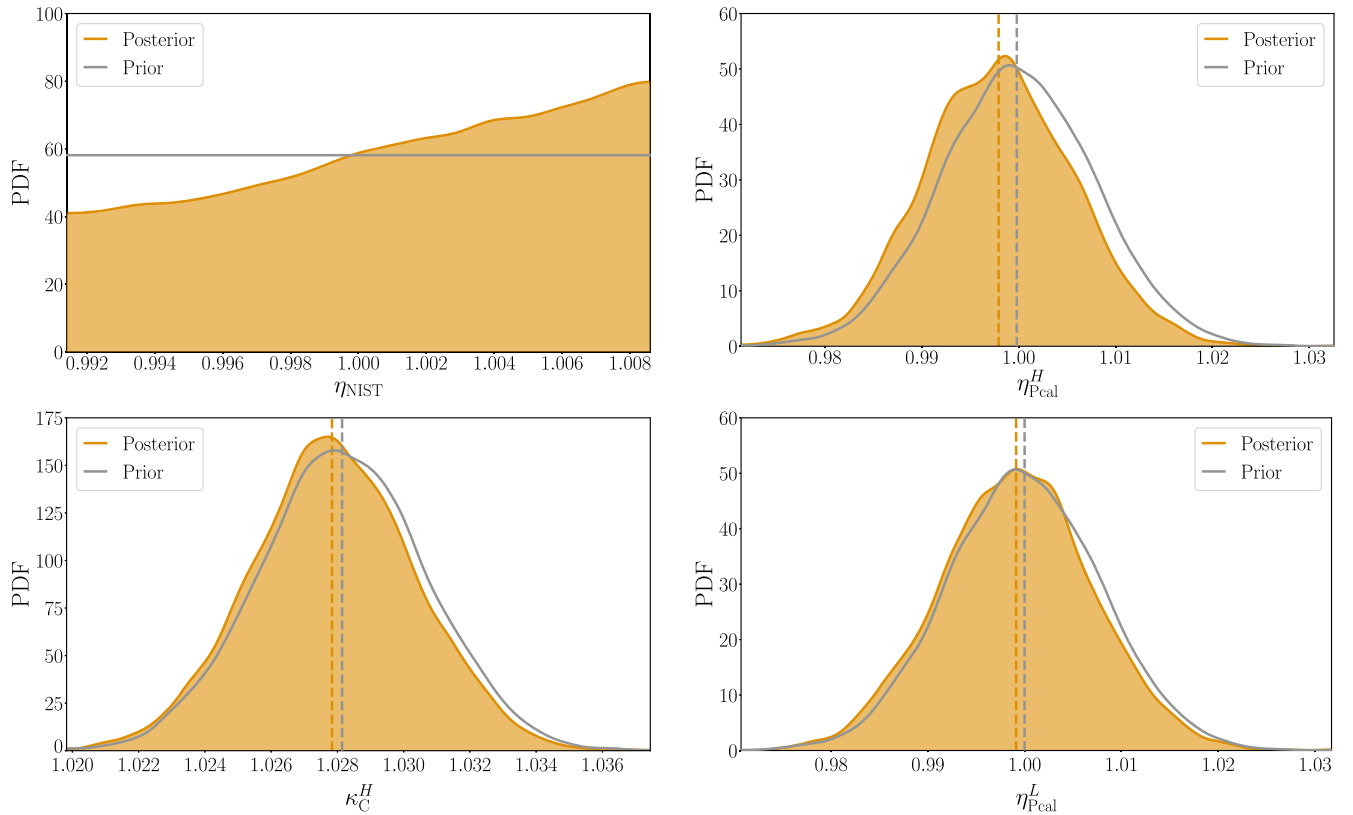


FIG. 3. Posterior distributions for the *physiCal* parameters for which information is gained relative to the priors, for the BNS #1 (see Table I). The respective priors are shown as solid gray lines. The median of each PDF is shown as a dashed vertical line.

reported in Table I, together with the ID we will use to refer to each.

The BNS #1 is added into LIGO-only data, since Virgo was not operating at the time. While for most of the *physiCal* parameters the prior is returned, a handful of posterior distributions are informative and are shown in Fig. 3, together with their priors. We see that the posterior of η_{NIST} , while still broad and with support in the whole prior range, does have some support for values larger than one. Meanwhile, the posterior for η_{Pcal}^H , which controls the overall amplitude of the response function in LIGO Hanford, is clearly different from its Gaussian prior and prefers slightly smaller values. For η_{Pcal}^L , the corresponding parameter for LIGO Livingston, the effect is not as significant. The other parameter that shows a slight departure from its priors is κ_C^H , a time-dependent parameter related to the sensing function of LIGO Hanford [17]. We again use the JS divergence [44] to quantify the statistical similarity between the prior and posterior distributions. For the *physiCal* parameters shown in Fig. 3, the JS divergences are 0.11 bits (η_{Pcal}^H), 0.09 bits (η_{NIST}), 0.05 bits (κ_C^H), and 0.05 bits (η_{Pcal}^L), respectively. In all these cases, we see that the offsets are much smaller than the statistical uncertainties. The posteriors of all other *physiCal* parameters are either even more similar to or undistinguishable from their priors.

When considering the BNS #2 we find instead that all of the *physiCal* parameters return exactly the prior, except η_{Pcal}^H (shown in Fig. 4), for which the JS divergence is 0.06 bits. Thus, despite a comparable network SNR and the presence of Virgo, less information is gained about the *physiCal* parameters for the BNS #2 than for the BNS #1. This suggests that the SNR is not the only figure of merit to predict if and what can be learned with astrophysical calibration. Instead, this might be suggestive of the fact

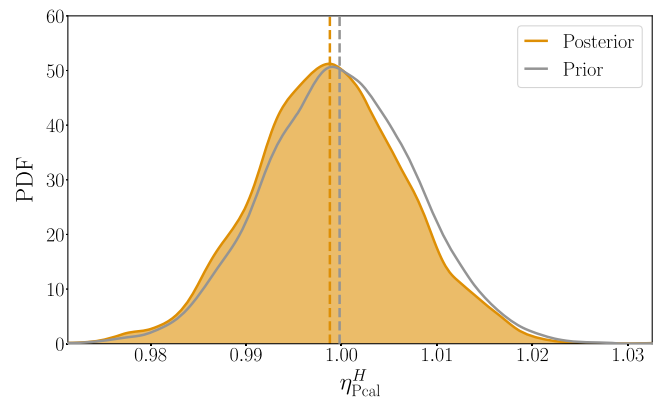


FIG. 4. Posterior distribution of η_{Pcal}^H for the BNS #2 (see Table I). The prior is shown as a solid gray line. The median of each PDF is shown as a dashed vertical line.

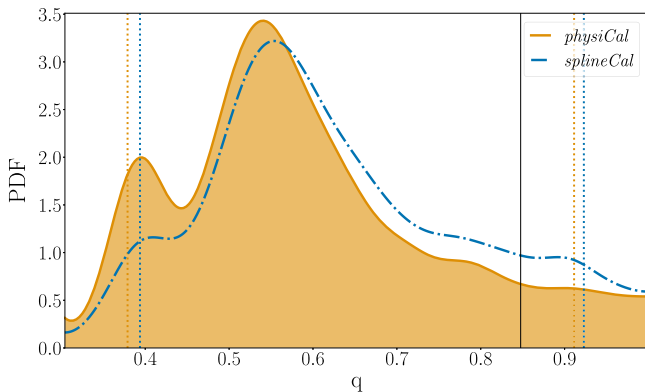


FIG. 5. Posterior distribution of q for the BNS #2 (see Table I) obtained with the *physiCal* (solid orange curve) and *splineCal* (dot-dashed blue curve) methods. Vertical dotted lines denote the 90% credible interval, whereas the solid vertical line indicates the true value.

that the *model* for the response function was adequate for the BNS #2, whereas it was not for the BNS #1. Theoretically, it is possible that even if the model for the response function is correct, we could beat the statistical uncertainties on the *physiCal* parameters, i.e., obtain posteriors which are centered at the same positions as their priors, but are narrower. We speculate that a similar measurement would require even higher SNRs, and we will explore that possibility in a future publication.

Next, we check if the posteriors for the CBC parameters of BNS #1 and BNS #2 are consistent with what would be obtained using the *splineCal* method (which was summarized in Sec. I). As for the O2 sources, we find a good consistency between the two methods: the highest value of the JS entropy for BNS#1 is 0.008 bits (for the arrival time at the geocenter), whereas for BNS#2 the highest value is 0.012 bits (asymmetric mass ratio and arrival time at the geocenter). Figure 5 shows a comparison of the asymmetric mass ratio posterior for the BNS#2: the value of the JS entropy is driven by the different support that the two methods find for the secondary peak at $q \sim 0.39$. The presence of secondary peaks in some parameters is not unusual, even for loud events, when analyzing real data (e.g., the tidal deformability of GW170817 [54]).

IV. CONCLUSIONS AND OUTLOOK

In this paper we have proposed a different and more physical approach to marginalizing over possible systematic error associated with the calibration of ground-based gravitational-wave detectors, called *physiCal*. We account for departures from the nominal value of the instruments' response functions using directly the output of the calibration pipeline of LIGO's instruments (the method can be extended to other detectors, even though we have not done it for this study). This method improves the existing

approach, which relies on a spline-based phenomenological model of calibration errors, hence discarding some of the available information about the detectors and their response functions.

We have augmented the LALInference source characterization algorithm with the *physiCal* method, and used it to analyze the eight CBC signals in the public data from the second observing run of the LVC. We find that the posteriors for the CBC parameters obtained with *physiCal* are extremely similar to those produced by the LVC with the existing spline method. This is not surprising since, at the expected SNRs of detections given the current detectors' sensitivities, the data are not informative enough to constrain the parameters of either calibration model better than the well-informed priors that are the result of the extensive efforts to calibrate the advanced LIGO and Virgo detectors. We then looked at the possibility of astrophysical calibration, i.e., the idea that a high SNR CBC observation, with perfectly known extrinsic parameters derived from an accompanying electromagnetic characterization, can be used to learn something about systematic error in each detector's calibration. We created a set of simulated BNS signals and added them to real public data from the LVC's second observing run. For all analyses, we assumed that the sources' sky positions and luminosity distances are perfectly known, whereas the orbital inclination angles are known to within 20° , mimicking a very successful EM campaign which provides information about position and orientation of the binaries. We find that for most of the simulations nothing can be learned about the *physiCal* parameters, and the posteriors are very similar to their priors. Only for the loudest BNSs we considered, with network SNRs around 30, were the posteriors for some of the *physiCal* parameters clearly, though not dramatically, different from their priors. Furthermore, we found that the SNR is not the only relevant parameter to forecast how informative any given source will be, and we showed that two BNSs with virtually the same SNRs can yield quite different posteriors for the *physiCal* parameters. Ultimately, both a high SNR and an imperfect model for the response function at the time of the simulated event are necessary for the data to be informative, as shown in Fig. 1 of [46]. In the representative system we chose, the parameters that were most different from their priors were the overall amplitude and two of the parameters associated with the sensing function in the LIGO Hanford detector. We argue that this is one of the main advantages of the *physiCal* method over the spline-based method: astrophysical calibration can potentially yield information about *specific* components involved with the calibration process, rather than about the response function as a whole. While we observed some departure from the modeled response function for some of the loudest BNSs we considered, the uncertainty in the *physiCal* parameters was not narrower than the prior uncertainty established by the calibration pipeline. That is, some of the posteriors shifted relative to their priors, but maintained the same shape. It is

possible that with even louder signals one could decrease the prior statistical uncertainty in the *physiCal* parameters. A large scale study will be necessary to explore the parameter space more systematically to fully understand which sources would yield the best astrophysical calibration, and which of the *physiCal* parameters are more likely to be constrained. Another possible avenue to improve our understanding of the response function is combining multiple detections. In fact, even though for most of the weaker sources very little is learned about the instrument, one can potentially combine all detected signals and build joint posteriors for the subset of the *physiCal* parameters that do not depend on time, and are thus expected to have the same value throughout a science run. Both of these prospects will be explored in a future publication.

ACKNOWLEDGMENTS

We thank Reed Essick, Paul Lasky, Ethan Payne, Colm Talbot, and Eric Thrane for useful discussion and for sharing an early version of their manuscript. We also thank the journal referees for their useful comments. S. V., C. J. H., L. S., and J. K. acknowledge support of the National Science Foundation and the LIGO Laboratory. L. S. also acknowledges the support of the Australian Research Council Centre of Excellence for Gravitational Wave Discovery (OzGrav), Project No. CE170100004. E. G. acknowledges the support of the Natural Sciences and

Engineering Research Council (NSERC) of Canada. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the United States National Science Foundation and operates under cooperative agreement PHY-1764464. Advanced LIGO was built under Grant No. PHY-0823459. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. This research has made use of data, software, and/or web tools obtained from the Gravitational Wave Open Science Center [55], a service of LIGO Laboratory, the LIGO Scientific Collaboration, and the Virgo Collaboration. This analysis was made possible by the LALSuite [56], NumPy [57,58], SciPy [59], and Matplotlib [60] software packages. The authors thank all of the essential workers who put their health at risk during the COVID-19 pandemic, without whom we would not have been able to complete this work. This is LIGO Document No. DCC-P2000293.

Note added.—After this work had begun, an independent group started exploring the possibility of using importance sampling to marginalize over physical calibration parameters [46]. As we have indicated in this paper, the two methods yield consistent results.

-
- [1] G. M. Harry (LIGO Scientific Collaboration), *Classical Quantum Gravity* **27**, 084006 (2010).
 - [2] J. Aasi *et al.* (LIGO Scientific Collaboration), *Classical Quantum Gravity* **32**, 074001 (2015).
 - [3] F. Acernese *et al.* (Virgo Collaboration), *Classical Quantum Gravity* **32**, 024001 (2015).
 - [4] LIGO Scientific and Virgo Collaborations, LIGO/Virgo O3 Public Alerts (2020).
 - [5] T. Akutsu *et al.* (KAGRA Collaboration), *Nat. Astron.* **3**, 35 (2019).
 - [6] B. Abbott *et al.* (KAGRA, LIGO Scientific, and VIRGO Collaborations), *Living Rev. Relativity* **21**, 3 (2018).
 - [7] B. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Astrophys. J. Lett.* **882**, L24 (2019).
 - [8] B. F. Schutz, *Nature (London)* **323**, 310 (1986).
 - [9] B. Abbott *et al.* (LIGO Scientific, Virgo, Fermi-GBM, and INTEGRAL Collaborations), *Astrophys. J. Lett.* **848**, L13 (2017).
 - [10] B. Abbott *et al.* (LIGO Scientific, Virgo, 1M2H, Dark Energy Camera GW-E, DES, DLT40, Las Cumbres Observatory, and VINROUGE, MASTER Collaborations), *Nature (London)* **85**, 425 (2017).
 - [11] H.-Y. Chen, M. Fishbach, and D. E. Holz, *Nature (London)* **562**, 545 (2018).
 - [12] B. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), arXiv:1908.06060.
 - [13] B. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Phys. Rev. Lett.* **116**, 131103 (2016).
 - [14] B. Abbott *et al.* (LIGO Scientific Collaboration), *Phys. Rev. D* **95**, 062003 (2017).
 - [15] D. Tuyenbayev *et al.*, *Classical Quantum Gravity* **34**, 015002 (2017).
 - [16] A. Viets *et al.*, *Classical Quantum Gravity* **35**, 095015 (2018).
 - [17] L. Sun *et al.*, *Classical Quantum Gravity* **37**, 225008 (2020).
 - [18] S. Karki *et al.*, *Rev. Sci. Instrum.* **87**, 114503 (2016).
 - [19] D. Bhattacharjee, Y. Leconte, S. Karki, J. Betzwieser, V. Bossilkov, S. Kandhasamy, E. Payne, and R. L. Savage, *Classical Quantum Gravity* **38**, 015009 (2021).
 - [20] S. Vitale, W. Del Pozzo, T. G. Li, C. Van Den Broeck, I. Mandel, B. Aylott, and J. Veitch, *Phys. Rev. D* **85**, 064034 (2012).
 - [21] C. Cahillane *et al.* (LIGO Scientific Collaboration), *Phys. Rev. D* **96**, 102001 (2017).
 - [22] J. Veitch *et al.*, *Phys. Rev. D* **91**, 042003 (2015).
 - [23] G. Ashton *et al.*, *Astrophys. J. Suppl. Ser.* **241**, 27 (2019).
 - [24] I. M. Romero-Shaw *et al.*, *Mon. Not. R. Astron. Soc.* **499**, 3295 (2020).

- [25] W. Farr, B. Farr, and T. Littenberg, Modelling Calibration Errors In CBC Waveforms, 2014, <https://dcc.ligo.org/LIGO-T1400682/public>.
- [26] S. M. Aston *et al.*, *Classical Quantum Gravity* **29**, 235004 (2012).
- [27] E. D. Hall, C. Cahillane, K. Izumi, R. J. E. Smith, and R. X. Adhikari, *Classical Quantum Gravity* **36**, 205006 (2019).
- [28] E. Goetz, S. Kandhasamy, J. Kissel, A. Viets, and S. Anand, Update to tracking temporal variations in DARM loop model parameters: Individual actuation stage tracking, cancelled lines, and SRC detuning, Technical Report No. LIGO-T1700106, LIGO Laboratory, 2019.
- [29] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning (MIT Press, Cambridge, MA, USA, 2006), p. 248.
- [30] F. Pedregosa *et al.*, *J. Machine Learning Res.* **12**, 2825 (2011), <https://www.jmlr.org/papers/v12/pedregosa11a.html>.
- [31] D. Estevez, B. Mours, L. Rolland, and D. Verkindt, Online $h(t)$ reconstruction for Virgo O3 data: start of O3, Technical Report No. VIR-0652B-19, Virgo, 2019.
- [32] B. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Phys. Rev. X* **9**, 031040 (2019).
- [33] B. Sathyaprakash and B. Schutz, *Living Rev. Relativity* **12**, 2 (2009).
- [34] K. Chatziioannou, C.-J. Haster, T. B. Littenberg, W. M. Farr, S. Ghonge, M. Millhouse, J. A. Clark, and N. Cornish, *Phys. Rev. D* **100**, 104004 (2019).
- [35] LIGO Scientific and Virgo Collaborations, Parameter Estimation samples, Power spectral densities and calibration uncertainty envelope release for GWTC-1, 2019.
- [36] J. Skilling, *Bayesian Anal.* **1**, 833 (2006).
- [37] LIGO Scientific and Virgo Collaborations, The O2 Data Release (2019), <https://www.gw-openscience.org/O2/>.
- [38] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, *Phys. Rev. Lett.* **113**, 151101 (2014).
- [39] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044006 (2016).
- [40] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044007 (2016).
- [41] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, *Phys. Rev. D* **94**, 044031 (2016).
- [42] T. Dietrich, S. Bernuzzi, and W. Tichy, *Phys. Rev. D* **96**, 121501 (2017).
- [43] T. Dietrich *et al.*, *Phys. Rev. D* **99**, 024029 (2019).
- [44] J. Lin, *IEEE Trans. Inf. Theory* **37**, 145 (1991).
- [45] S. Kullback and R. A. Leibler, *Ann. Math. Stat.* **22**, 79 (1951).
- [46] E. Payne, C. Talbot, P. D. Lasky, E. Thrane, and J. S. Kissel, *Phys. Rev. D* **102**, 122004 (2020).
- [47] R. Essick and D. E. Holz, *Classical Quantum Gravity* **36**, 125002 (2019).
- [48] B. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Phys. Rev. Lett.* **116**, 241102 (2016).
- [49] H.-Y. Chen, S. Vitale, and R. Narayan, *Phys. Rev. X* **9**, 031028 (2019).
- [50] S. A. Usman, J. C. Mills, and S. Fairhurst, *Astrophys. J.* **877**, 82 (2019).
- [51] M. Pitkin, C. Messenger, and L. Wright, *Phys. Rev. D* **93**, 062002 (2016).
- [52] I. Mandel, *Astrophys. J. Lett.* **853**, L12 (2018).
- [53] D. Finstad, S. De, D. A. Brown, E. Berger, and C. M. Biwer, *Astrophys. J. Lett.* **860**, L2 (2018).
- [54] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Phys. Rev. X* **9**, 011001 (2019).
- [55] <https://www.gw-openscience.org>.
- [56] LIGO Scientific Collaboration, LIGO Algorithm Library—LALSuite, free software (GPL), 2019.
- [57] T. Oliphant, *NumPy: A guide to NumPy* (Trelgol Publishing, USA, 2006).
- [58] C. R. Harris *et al.*, *Nature (London)* **585**, 357 (2020).
- [59] P. Virtanen *et al.*, *Nat. Methods* **17**, 261 (2020).
- [60] J. D. Hunter, *Comput. Sci. Eng.* **9**, 90 (2007).