# Gravitational-wave astronomy with a physical calibration model

Ethan Payne,[1,2,*] Colm Talbot,[3,1,2] Paul D. Lasky,[1,2] Eric Thrane,[1,2] and Jeffrey S. Kissel[4]

[1]*School of Physics and Astronomy, Monash University, Clayton VIC 3800, Australia*
[2]*OzGrav: The ARC Centre of Excellence for Gravitational Wave Discovery, Clayton VIC 3800, Australia*
[3]*LIGO, California Institute of Technology, Pasadena, California 91125, USA*
[4]*LIGO Hanford Observatory, Richland, Washington 99352, USA*

(Received 22 September 2020; accepted 10 November 2020; published 18 December 2020)

We carry out astrophysical inference for compact binary merger events in LIGO-Virgo's first gravitational-wave transient catalog (GWTC-1) using a physically motivated calibration model. We demonstrate that importance sampling can be used to reduce the cost of what would otherwise be a computationally challenging analysis for signal-to-noise ratios of current gravitational-wave detections. We show that including the physical estimate for the calibration error distribution has negligible impact on the inference of parameters for the events in GWTC-1. Studying a simulated signal with matched filter signal-to-noise ratio SNR $= 200$, we project that a calibration error estimate typical of GWTC-1 is likely to be negligible for the current generation of gravitational-wave detectors. We argue that other sources of systematic error—from waveforms, prior distributions, and noise modeling—are likely to be more important. Finally, using the events in GWTC-1 as standard sirens, we infer an astrophysically informed improvement on the estimate of the calibration error in the LIGO interferometers.

DOI: 10.1103/PhysRevD.102.122004

## I. INTRODUCTION

The burgeoning field of gravitational-wave astronomy [1–3] is advancing our understanding in multiple fields of astrophysics, including cosmology [4], galactic and stellar evolution [5], and strong-field gravity [6]. As the sensitivity of observatories improve, and gravitational waves are observed with increasingly large signal-to-noise ratios (SNRs) [7–9], an understanding of systematic effects will become ever more important. Sources of systematic biases include errors associated with gravitational waveforms [10], imperfect prior distributions, incorrect estimates for the noise power-spectral density [11–13], and errors associated with the calibration of the detectors [14]. Here, we focus on errors associated with calibration.

Calibration is defined as the process of converting the detector's primary control system error signal due to differences in the lengths of the interferometer's arms to an estimate of strain on the detector [15–18]. Imperfect knowledge of the interferometer's control system and the response to differential arm length changes leads to systematic error in the amplitude and phase of the calibration. This error is estimated by conducting a vast suite of measurements of the control system, and propagating the results of those measurements into a physically informed model. The resulting error estimation is represented by a frequency-dependent probability distribution. In order to avoid bias, the estimated probability distribution of calibration errors must be taken into account when inferring the astrophysical parameters of gravitational-wave signals [14]. Unfortunately, marginalizing over calibration error distributions can dramatically increase the number of parameters used in astrophysical inference: from the 15 required to describe a binary black hole to >50 [19]. This increase in parameter space can lead to a significant increase in computational cost and convergence issues, which has somewhat limited efforts to carry out astrophysical inference that include an estimate of calibration errors up to this point.

In this work, we demonstrate a computationally efficient implementation of the original physical calibration model [16,18, see Sec. II] for astrophysical inference. Following [20], we first evaluate the posterior distribution of astrophysical parameters without any estimated calibration error distribution, and then employ importance sampling to reweight approximate results to include this contribution. Importance sampling [21,22] is the technique of constructing weights for individual samples which determine each sample's contribution to the inferred probability distribution. Having verified the analysis procedure, we carry out a study of gravitational-wave signals from the first LIGO-Virgo Gravitational-Wave Transient Catalog (GWTC-1) [3] using estimates of the calibration error at the time of those events. Combining data from multiple events, we infer an astrophysically informed calibration error estimate [23], showing that it is possible to learn about the Advanced

*ethan.payne@ligo.org

 

LIGO interferometers [24] using gravitational waves as standard sirens.

With the assumption that the estimated systematic error present in GWTC-1 will be typical in the future, we demonstrate the effect of the calibration error estimate on a simulation of an SNR = 200 binary black hole merger—approximately the loudest event that will be observed by the second-generation interferometer network during its operational lifetime [7,25]. In this regime, naive application of our importance sampling algorithm becomes inefficient. However, we find that the calibration error estimate still has only a marginal effect on the posterior distributions of the intrinsic astrophysical parameters. The localization parameters are the most affected by the inclusion of the calibration error distribution; the sky map credible region approximately doubles in size for a SNR = 200 event. We conclude that the impact of calibration error estimates will likely be small compared to other previously mentioned sources of systematic error in astrophysical parameter estimation.

The remainder of this paper is organized as follows. In Sec. II, we summarize the physically motivated calibration model, its parameters, and the collective error estimate from [16,18]. In Sec. III, we describe our methodology for efficiently marginalizing over the probability distribution of calibration errors with importance sampling. In Sec. IV, we demonstrate our implementation using simulated data while in Sec. V, we analyze data from GWTC-1. We end with concluding remarks in Sec. VI.

## II. CALIBRATION MODEL

In this section, we summarize the physical calibration model for the LIGO detectors described in [16,18]. Though the Virgo detector [26] is similar to the LIGO detectors, the systematic error probability distribution used in this study is informed by the 68% confidence interval bounds on the systematic error described in Ref. [17].

### A. The physical model

The LIGO detectors are dual-recycled, kilometer-scale Fabry-Pérot Michelson interferometers, most sensitive between 10 and 2000 Hz [24]. The passage of a gravitational-wave signal induces differential displacement between the two arms of the interferometer. This differential arm displacement is measured at the output of the detector, where interfering laser light reflected from the resonant arm cavities is incident on a set of photodiodes [27]. The photodiode signals are summed and digitized to form a signal which includes both detector noise and gravitational waves. The digitized signal also serves as the residual error signal of the feedback control system for the changes in differential arm length. This, among other control loops in the detector, ensures that external noise sources do not force the interferometer cavities off-resonance. This is achieved through differentially actuating on the arm cavity mirrors, or test masses, and their suspension systems [28–30]. Below ∼100 Hz, the control systems actuation forces suppress the interferometer's response to differential arm displacement. Above this frequency, the response is free of control system influence and depends on both the interferometric response to differential arm displacement as well as the signal processing electronics of the photodiodes. Thus, in order to reconstruct the measured strain from the digitized photodiode signal over the entire sensitive frequency band, a physically motivated model of the response and control system are required.

The model is divided into two conceptual components. The first component, the *sensing function*, is an optomechanical description of the interferometer if it were free of control forces, photodiode signal processing electronics and the digital acquisition system. The second component, the *actuation function*, describes how the control system splits the single, digitally filtered, error signal among three stages of cascading actuators on the test mass quadruple suspension systems; incorporating those actuators' digital to analog converters and signal processing electronics. The actuation function also includes the complex displacement response of the test mass for those forces from each stage [28–30]. Both model components are frequency-dependent complex transfer functions that are mostly static in time, but each have slowly time-varying correction factor parameters to account for natural drifts in their behavior. The actuation function dominates the interferometer's response below ∼200 Hz, whereas the sensing function dominates elsewhere [18].

The calibrated strain signal $h$ is related to the differential arm error signal $d_{\mathrm{err}}$ (in the frequency domain) by the response function $R$,

$$h = \frac{1}{L} R d_{\mathrm{err}} = \frac{1}{L}\left(\frac{1 + CDA}{C}\right) d_{\mathrm{err}}. \tag{1}$$

Here, $C$ is the sensing function while $A$ is the actuation function. The variable $D$ describes the set of digital filters responsible for converting the error signal to the single differential arm control signal. The variable $L$ is the length of the interferometer arms. Equation (1) illustrates how systematic errors in $C$ and $A$ lead to an error in $h$.

Following [16,18], we employ the following model for the sensing function:

$$C(f|\Lambda) = \frac{\kappa_C(t) H_C}{1 + if/f_{CC}(t)} C_R(f) e^{-2\pi i f \tau_C} \times \frac{f^2}{f^2 + f_S^2 - i f f_S/Q}. \tag{2}$$

Here, $f$ is frequency, and $\{H_C, f_{CC}, \tau_C, f_S, Q\}$ are the parameters describing the optomechanical response (summarized in Table I), which are part of a larger set of

TABLE I. Sensing parameters. The optical gain describes the overall magnitude of the sensing function. The coupled cavity pole frequency describes the bandwidth of the interferometer arm and signal recycling cavity system. The time delay compensates for light travel time within the arms and computational delay in the photodiode analog-to-digital conversion process. The optical spring parameters describe the characteristic frequency and amount of detuning between the arm and signal recycling cavities.

| Symbol | Name | Units |
|---|---|---|
| $H_C$ | Optical gain | cts/m |
| $f_{CC}$ | Coupled cavity pole frequency | Hz |
| $\tau_C$ | Time delay | $\mu$s |
| $f_S$ | Optical spring frequency | Hz |
| $Q$ | Optical spring quality factor | ⋯ |

calibration model parameters $\Lambda$. The parameters $\kappa_C(t)$ and $f_{CC}(t)$ represent the time-dependent corrections needed to account for alignment drift in the suspended cavities. Detuning between the signal recycling cavity and arm cavities [31] is modeled as an optical spring with a characteristic frequency, $f_S$, and associated quality factor, $Q$. Finally, $C_R(f)$ is the digital acquisition response, which is measured *a priori* with high precision. The probability distributions for the time-independent parameters of the sensing function, $\Lambda$, are determined by fitting measurements from a single, representative reference time with the model outlined in Eq. (2) using Markov chain Monte Carlo (MCMC) sampling [32]. The probability distribution for the parameter $f_{CC}$ is determined both by an MCMC fit from the reference time measurement, and, as $\kappa_C(t)$, by the continuous high-precision tracking of its time dependence [33].

The actuation function is modeled as follows:

$$\begin{aligned} A(f|\Lambda) = {} & \kappa_U(t)F_U(f)H_U A_U(f)e^{-2\pi i f\tau_U} \\ & + \kappa_P(t)F_P(f)H_P A_P(f)e^{-2\pi i f\tau_P} \\ & + \kappa_T(t)F_T(f)H_T A_T(f)e^{-2\pi i f\tau_T}. \end{aligned} \tag{3}$$

Here, $\{H_U, \tau_U, H_P, \tau_P, H_T, \tau_T\}$ are the actuation calibration parameters summarized in Table II. The subscripts refer to the stage of the suspension system where actuation force is applied. These are the "upper-intermediate," $U$, "penultimate," $P$, and "test mass," $T$. The force-to-displacement response and the response of actuator electronics are incorporated in $A_i(f)$. The digital distribution filters, $F_i(f)$, and the scalar time-varying correction factors, $\kappa_i(t)$, are precisely known, and so do not appear in Table II. Again, the prior distributions for the actuation parameters are determined by MCMC sampling with data from single measurements of each stage's response. The values and uncertainties associated with time-dependent quantities are computed at a 1 hr cadence over the duration of an observing run.

TABLE II. Actuation parameters. The scalar gains applied to each of the stages calibrate the overall magnitude of the actuation. The time delays arise from computational delays in the digital-to-analog system.

| Symbol | Name | Units |
|---|---|---|
| $H_U$ | Upper intermediate stage gain | N/cts |
| $\tau_U$ | Upper intermediate stage delay | $\mu$s |
| $H_P$ | Penultimate stage gain | N/cts |
| $\tau_P$ | Penultimate stage delay | $\mu$s |
| $H_T$ | Test mass stage gain | N/cts |
| $\tau_U$ | Test mass stage delay | $\mu$s |

The final parameter within the physical calibration model is an overall scalar magnitude factor, $\eta_{\mathrm{PCAL}}$, whose probability distribution is derived from any systematic error and uncertainty in the photon calibrator systems (PCALs). The photon calibrator systems are used as fiducial displacement references for each detector [34,35]. Typically, the systematic error is negligibly different from unity, and only adds an overall magnitude uncertainty: coincidentally 0.79% for both LIGO detectors during the second observing run. This additional correction is applied as a multiplicative factor to the response function.

While each detector's fiducial reference has its own systematic error and associated uncertainty, the collection of references for the entire network are seeded from a single global reference calibrated at the National Institute of Standards and Technology (NIST). This common error on the global reference, $\eta_{\mathrm{NIST}}$, is included in the uncertainty for each detector [35] and excluded as an independent error from this analysis as it is degenerate with the luminosity distance of a gravitational-wave source. The overall amplitude of the gravitational-wave strain scales as $\eta_{\mathrm{NIST}}/d_L$ and so no conclusions can be drawn about the absolute reference from gravitational waves because of degeneracy with the distance. This is contrary to the results from Ref. [36], where the luminosity distance is fixed—motivated by the detection of electromagnetic counterparts. The challenge we see with this proposal is that the uncertainty on redshift and cosmological parameters is likely to be large compared to the very small uncertainty on $\eta_{\mathrm{NIST}}$.

### B. The phenomenological model

While the physical model of the response function, $R(\Lambda)$, produces an approximately correct response, inspection of an ensemble of frequency-dependent residuals $R_{\mathrm{measured}}/R(\Lambda)$, constructed from sensing and actuation function measurements, shows that the model is incomplete, i.e., the residuals are not consistent with unity; see Fig. 11 from [18]. The authors of [18] build an additional phenomenological model for $C$ and each stage of $A$ on top of the physical model in order to estimate the residuals,

completing the error estimate with new phenomenological parameters.

The phenomenological model employs Gaussian process regression of the residuals, interpolating between 128 frequency points [37,38] and optimization of several hyperparameters constraining the covariance kernel between each frequency point. The correlation lengths between frequencies in the Gaussian process are frequency dependent [18], scaling as $\sim 5f$. Therefore, any information learned about the Gaussian process at a particular frequency, $f_i$, informs the distribution within a region $\sim 10 f_i$. The corrected sensing and actuation functions are given by

$$C'(\Lambda) = \eta_C(\Lambda) C(\Lambda), \tag{4}$$

$$A'(\Lambda) = \eta_A(\Lambda) A(\Lambda). \tag{5}$$

Here, $(C, A)$ are the physical sensing and actuation models while $(C', A')$ are the phenomenologically corrected models. They are included as a part of the frequency-dependent estimated distribution of calibration error. Since $\eta_C$ and $\eta_A$ for each stage are complex-valued functions described by a magnitude and phase, the phenomenological model introduces an additional $256 \times 4$ calibration parameters to $\Lambda$.

After applying both physical and phenomenological models to LIGO data, the authors of [18] find that the distribution of errors in the response $R$ completely explains $R_{\text{measured}}/R(\Lambda)$, and is dominated—in most frequency regions—by uncertainty from the Gaussian process fit. That is, the systematic error from imperfect design of the physical model is large compared to the uncertainty in its parameters. However, by introducing such a high-dimensional phenomenological model, the systematic error of the physical model is converted almost entirely into statistical uncertainty. With so many free parameters, we expect it should be possible to fit nearly any measured form of $R$.

## III. METHOD

Our goal is to estimate astrophysical parameters $\theta$ describing the gravitational waveform of a compact binary merger given strain data $h$ and marginalizing over the unknown calibration parameters, $\Lambda$. We follow style conventions from [39]. Assuming Gaussian noise, the likelihood is given by

$$\mathcal{L}(h_j|\theta, \Lambda) = \frac{1}{2\pi P_j} \exp\left(-2\Delta f \frac{|h_j - \lambda_j(\Lambda)\mu_j(\theta)|^2}{P_j}\right). \tag{6}$$

Here, $\mu_j(\theta)$ denotes the gravitational-wave model. In this manuscript, we utilize IMRPhenomPv2 [40,41] for our source model of binary black hole systems, and IMRPhenomPv2NRTidal [42] for binary neutron star mergers. The parameters of the compact binary coalescence, $\theta$, include intrinsic properties such as the masses and spins of the individual compact objects, and extrinsic parameters informing the orientation and location of the binary system. The calibration error is described by

$$\lambda(\Lambda) = \frac{R(\Lambda)}{R_\varnothing}, \tag{7}$$

the ratio of the model for the true response function $R(\Lambda)$, which depends on calibration parameters $\Lambda$, to the theoretical response function used to calibrate the data $R_\varnothing$. The calibration err, $\lambda(\Lambda)$, is denoted as $\eta_R$ in Ref. [18]. Note that the calibration error is applied directly to the gravitational-wave model [43]. The subscript $j$ refers to a single frequency bin, which is spaced by $\Delta f$. Since the noise in each bin is approximately independent, the combined likelihood is then simply

$$\mathcal{L}(h|\theta, \Lambda) = \prod_j \mathcal{L}(h_j|\theta, \Lambda). \tag{8}$$

The product over frequency bins is implied in subsequent equations.

Our *target distribution*, the one for which we want to generate posterior samples, is Eq. (8) marginalized over $\Lambda$:

$$\mathcal{L}_\Lambda(h|\theta) = \int d\Lambda \mathcal{L}(h|\theta, \Lambda)\pi(\Lambda). \tag{9}$$

Here $\pi(\Lambda)$ is our prior on the calibration parameters. The target distribution can be computationally expensive to sample from owing to the extra dimensionality associated with $\Lambda$. However, if the original calibration $R_\varnothing$ is at least approximately correct, and if the SNR of the event is not too large (we quantify how large momentarily), then we can employ importance sampling to avoid sampling in $\Lambda$.

Following [20], we define our *proposal distribution*,

$$\mathcal{L}_\varnothing(h_j|\theta) = \frac{1}{2\pi P_j} \exp\left(-2\Delta f \frac{|h_j - \mu_j(\theta)|^2}{P_j}\right), \tag{10}$$

corresponding to the likelihood we would use if we believed the original response function $R_\varnothing$ was perfectly accurate. We use the proposal distribution to generate samples in $\theta$ using the Bilby [44,45] implementation of Dynesty [46], a nested sampling algorithm [47]. Since we are not sampling in $\Lambda$, the proposal samples are computationally cheap to generate.

Next, for each posterior sample of the binary model parameters, drawn from the proposal distribution, $\theta_i$, we calculate a weight, which requires marginalizing over calibration parameters. Following [18], we carry out this calculation using a predetermined set of $N = 10^4$ calibration response curves, generated with random draws from the prior distribution $\{\Lambda_k\} \sim \pi(\Lambda)$. We define a doubly indexed weight relating the proposal likelihood to the target likelihood:

$$w_{ik} = \frac{\mathcal{L}(h|\theta_i, \Lambda_k)}{\mathcal{L}_{\varnothing}(h|\theta_i)}. \tag{11}$$

Here, $i$ indexes binary posterior samples for the parameter $\theta$ while $k$ indexes calibration prior samples for $\Lambda$. The calibration-marginalized weight is simply

$$w_i = \frac{1}{N}\sum_k^N w_{ik} = \frac{\mathcal{L}_{\Lambda}(h|\theta_i)}{\mathcal{L}_{\varnothing}(h|\theta_i)}. \tag{12}$$

Alternatively, we can marginalize over the gravitational-wave model parameters in order to obtain weights useful for constructing posteriors for the calibration parameters:

$$w_k = \frac{1}{n}\sum_i^n w_{ik} = \frac{\mathcal{L}_{\theta}(h|\Lambda_k)}{\mathcal{Z}_{\varnothing}(h)}, \tag{13}$$

where $\mathcal{Z}_{\varnothing}(h)$ is the normalization coefficient of the proposal posterior distribution, known as the Bayesian evidence. This procedure is similar to approaches for estimating neutron-star equations of state with Gaussian processes [48].

The weights quantify the relative importance of each sample in light of the fact that we are actually interested in the target distribution, not the proposal distribution. The weights can be input directly into routines for constructing corner plots. They may also be used to calculate the Bayesian evidence for the target distribution, $\mathcal{Z}_{\Lambda}(h)$, from the evidence for the proposal distribution. The ratio of the two evidences is simply the average weight,

$$\mathcal{B}^{\Lambda}_{\varnothing} = \bar{w} = \frac{\mathcal{Z}_{\Lambda}(h)}{\mathcal{Z}_{\varnothing}(h)}, \tag{14}$$

known as the Bayes factor which provides a measure of the preference for the calibration model in comparison to the null hypothesis $\varnothing$ that the data are already correctly calibrated. The process of constructing these weights is known as importance sampling [21,22]. This approach is not confined to the calibration model outlined in Sec. II, and allows for the application of improved models in the future. Furthermore, the method can equivalently be applied with other spline models [14,19] used for analyses in GWTC-1 [3].

The efficacy of importance sampling can be measured using an efficiency [20,49,50]:

$$\epsilon = \frac{n_{\rm eff}}{n} = \frac{1}{n}\frac{(\sum_i^n w_i)^2}{\sum_i^n w_i^2}. \tag{15}$$

Here, $n$ is the number of astrophysical samples generated using the proposal distribution while $n_{\rm eff} < n$ is the number of effective samples created from importance sampling. If the proposal distribution is close to the target distribution, the efficiency will be high. As a rule of thumb, $\epsilon > 50\%$ is "excellent" (providing a fast, reliable answer) while $\epsilon \approx$ 1%–50% is "good," providing adequate efficiency to make importance sampling clearly useful. Efficiencies $\lesssim$1% indicate that the proposal distribution is not necessarily a good approximation for the target distribution, and so reweighting begins to become inefficient, requiring a large number of initial samples and many evaluations of the target likelihood in order to obtain a reliable answer. The efficiency falls with increasing SNR, since louder events are characterized by progressively peaked likelihood functions. We verify that the efficiency is above 10% when SNR $\lesssim 40$. Noting that the distribution of network SNR approximately scales as SNR$^{-4}$ [7], ∼97% of all GW events are expected to lie within this regime. One can judge the convergence of the importance sampled result by considering the number of effective samples. The efficiency can also be used as a measure of the overall effect of the inclusion of a physical calibration model, though there are better measures. Pathological cases, where importance sampling fails due to multimodality, are unlikely to apply to our present application; see [20] for additional details.

One benefit of likelihood reweighting is its low computational cost. By directly executing Bayesian inference with the calibration-marginalized likelihood, the number of evaluations of the more computationally expensive model is orders of magnitude larger than the number of posterior samples produced. By utilizing likelihood reweighting, the proposal distribution is found with a cheap likelihood function before the expensive likelihood is used sparingly in postprocessing.

We can also use the astrophysical parameter-marginalized weights to construct posterior distributions for the calibration hyperparameters informed by an ensemble of events. We construct weights for the $k$th set of calibration curves informed by $M$ events as

$$w_k^{\rm tot} = \prod_{\nu}^{M} w_k^{\nu} = \prod_{\nu}^{M} \frac{\mathcal{L}_{\theta}(h^{\nu}|\Lambda_k)}{\mathcal{Z}_{\varnothing}(h^{\nu})}, \tag{16}$$

where $\nu$ indexes the different events, not to be confused with the additional implied product over frequency bins in Eq. (6). The average combined weight, $\bar{w}^{\rm tot}$, is the Bayes factor for the calibration error distribution compared to the null hypothesis that the calibration error is zero. Of course, in order to combine multiple events, we must take care to ensure that the interferometer is in the same state. Otherwise, the calibration parameters can be different for different events. Thus, one must ensure that events are only combined for a period during which the interferometer is maintained in a steady configuration.

## IV. SIMULATED EVENTS

We validate our method using simulated signals injected into Gaussian noise colored to match the Advanced LIGO design sensitivity noise curve [24]. We analyze two signals, both with properties consistent with GW150914 [1]. We focus on high-SNR events where calibration errors are

relatively more important. In one case, we adjust the distance to achieve an optimal SNR = 30, which is comparable to the loudest observed gravitational-wave signal, GW170817 [2] with SNR ≈ 32. In the second case, we set the distance to achieve SNR = 200. We use calibration envelopes equivalent to the calibration estimate at the time of GW170817 [51].

Starting with the SNR = 30 event, we compare the posterior distributions for binary parameters $\theta$ obtained three different ways: ignoring the calibration error, marginalizing over calibration error estimates with the importance sampling method described above, and with "direct sampling," in which we marginalize over the calibration with the $N = 10^4$ response curves at every step using Eq. (9) as the nested sampler explores the astrophysical parameter space. The direct sampling method is relatively slow compared to importance sampling (by a factor of ∼250) requiring the use of pBilby [52], a parallelized implementation of Dynesty [46].

All three methods produce nearly identical posterior distributions, which are difficult to distinguish by eye, illustrating that calibration error distribution has only a very small effect on our inferences about astrophysical parameters. This is also verified in Sec. V when analyzing all events from GWTC-1. In Table III, we present the maximum one-dimensional JensenShannon (JS) divergence [53] comparing the similarity of the posterior distributions obtained using each method. The JS divergence is a symmetric extension of the Kullback-Liebler (KL) divergence [54] which measures the divergence between 0 bit (no divergence) to 1 bit (maximal divergence). We obtain JS divergence values $\lesssim 6 \times 10^{-3}$ which are similar to those obtained from comparing the results obtained using different stochastic sampling codes to sample the same likelihood [44,45,55].

The SNR = 200 event allows us to study what is likely to be the maximum-SNR regime for second-generation gravitational-wave detectors. The Advanced LIGO/Virgo network at design sensitivity is expected to observe $\mathcal{O}(10^4)$ events over its operational lifetime. Assuming that the distribution of network SNR scales as $\mathrm{SNR}^{-4}$ [7], the number of events with a signal-to-noise ratio greater than 200 will be $\mathcal{O}(1)$ (see also [25]).

TABLE III. The largest one-dimensional JensenShannon (JS) divergence (bit) comparing the similarity of the posterior distributions obtained using different methods for a simulated SNR = 30 binary black hole signal. The small value in each cell indicates that the three methods produce similar distributions, which implies that calibration error distribution does not have a significant effect on astrophysical inference.

| | Direct | Importance |
|---|---|---|
| No error | $\mathrm{JS}_{\mathrm{RA}} = 8.07 \times 10^{-3}$ | $\mathrm{JS}_{\mathrm{RA}} = 5.16 \times 10^{-3}$ |
| Importance | $\mathrm{JS}_{a_1} = 3.94 \times 10^{-3}$ | |

The posterior distributions for the astrophysical parameters are presented in the top panels of Fig. 1. We see the largest differences in the extrinsic parameters (right panel). In particular, we highlight that the credible regions on the sky are approximately twice as large when marginalizing over calibration error estimates than when assuming no calibration error is present. Quantitatively, this difference corresponds to a JS divergence of 0.105 for right ascension. More modest changes are seen for the remaining parameters: with JS divergences $\leq 3.04 \times 10^{-2}$, qualitative astrophysical results are unchanged. The mean and 95% credible regions for the prior and posterior on the calibration error distribution are shown in the lower panels. We note that the widths of the posterior credible regions are approximately half that of the prior credible regions.

A GW150914-like event only has gravitational-wave signal content up to ∼350 Hz [1]. However, we see that our simulated event informs our model of the LIGO detectors' calibration error distributions up to 1000 Hz. This is because the physical calibration model has correlations encoded between lower and higher frequencies due to the Gaussian process (see Sec. II B). Once the signal can inform properties of the sensing function model, i.e., the signal frequency surpasses ∼200 Hz, then the correlation length of the calibration model exceeds ∼2000 Hz. Therefore any information gained about the calibration model at these frequencies will inform the remainder of the frequency domain. In contrast, as the spline calibration model was used for Virgo's calibration, we learn significantly less information about its calibration at the same higher frequencies.

The Bayes factor for the SNR = 200 injection is $\mathcal{B}^{\Lambda}_{\emptyset} = 2.04 \times 10^{-4}$, indicating a preference for the null hypothesis that there is no calibration error. We expect that the Bayes factor prefers the null hypothesis as no calibration error has been applied to the simulated data. However, this result also tells us something interesting about the calibration model. No calibration error ($\lambda = 1$) should be allowed as one possible realization of the calibration envelope. What does it mean, therefore, that the data so strongly prefer the null hypothesis for this injection? We suspect there are two factors at play. First, some of the preference is likely coming from a large physical calibration model parameter space. This results in a penalty known as an Occam factor, where simplified models with a smaller prior volume are preferred to models with a larger prior volume, provided the data are fit accurately. However, we suspect that there is a more important factor at play: the $10^4$ realizations of the calibration envelope may not be sufficient to adequately fit the zero-error data. If this is the case, it could be highlighting the limitations that arise when we represent a continuous response function with some finite number of curves. Additional work beyond our present scope would be useful to investigate these hypotheses.

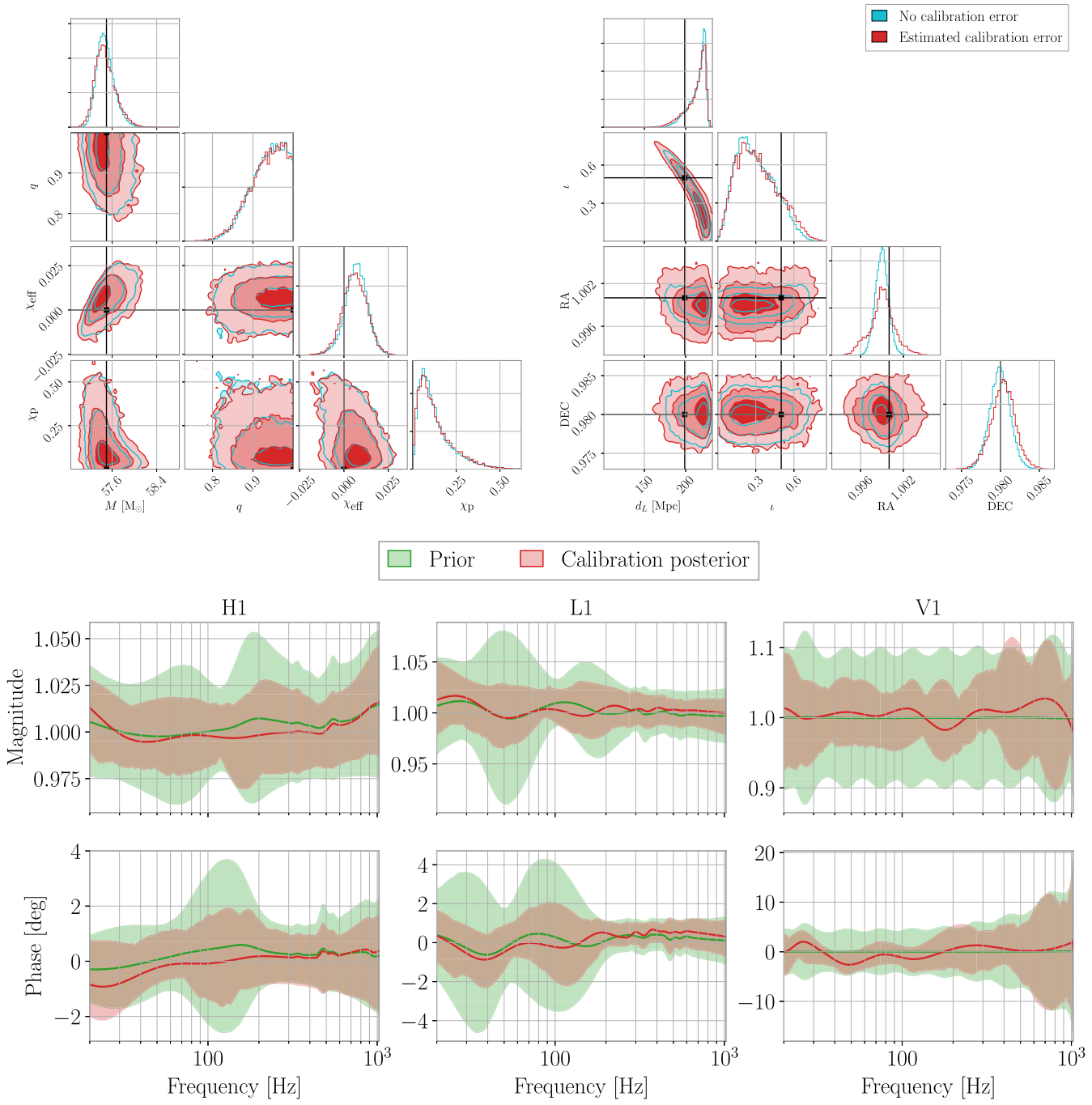


FIG. 1. Posterior distributions for an SNR = 200 GW150914-like event. The top panels show astrophysical parameters. The different shaded regions are $1\sigma$, $2\sigma$, and $3\sigma$ credible intervals. The red contours include the calibration error estimate while the blue contours do not. The black lines correspond to the injected properties of the source. The bottom panels show the reconstructed response function $R$. The red curves show the response curves averaged over calibration hyperparameters. The green curves show the response functions averaged over prior samples. The 95% credible intervals are indicated with translucent shading. The inclusion of the calibration envelope broadens the majority of astrophysical parameters by a modest amount. The sky localization of the event broadens noticeably with the inclusion of calibration error distribution, expanding by a factor of $\approx 2$ in a solid angle. This indicates the possibility that even for the loudest events observed, the calibration error estimate may not play a major role in the inferences made about the intrinsic properties of the source. It is interesting to note that constraining calibration model parameters at lower frequencies where the gravitational-wave signal is detected can inform the calibration model at higher frequencies where no signal is present.

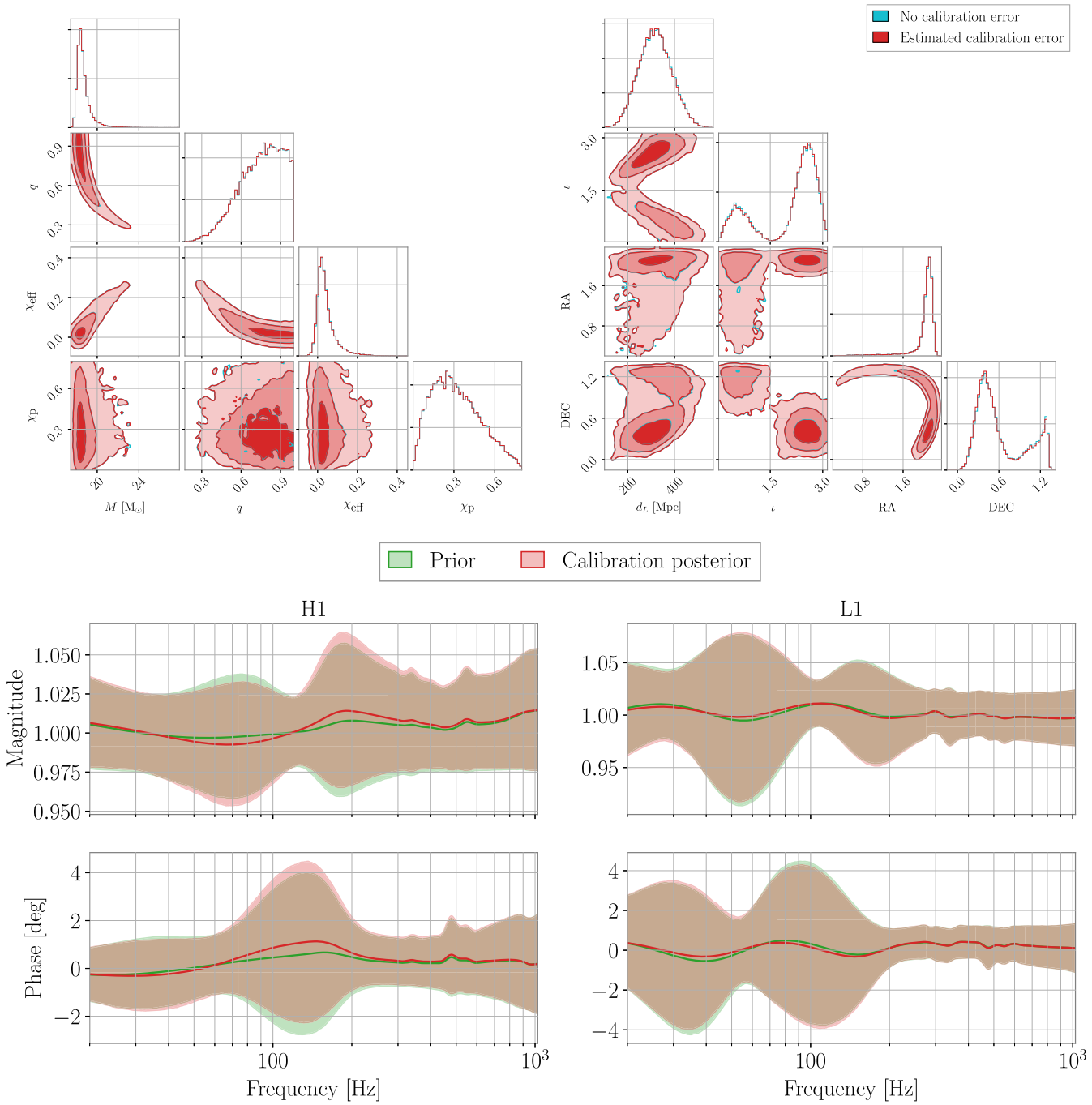


FIG. 2. Posterior distributions for GW170608. The top panels show astrophysical parameters. The different shaded regions are $1\sigma$, $2\sigma$, and $3\sigma$ credible intervals. The red contours include the calibration error estimate while the blue contours do not. The inclusion of uncertainty in the calibration error leads to only marginal changes. The bottom panels show the reconstructed response function $R$. The red curves show the response curves averaged over calibration hyperparameters. The green curves show the response functions averaged over prior samples. The 95% credible intervals are indicated with translucent shading. The data are marginally informative about the calibration parameters.

## V. RESULTS FROM GWTC-1

We analyze the 11 binary merger events identified in GWTC-1 [3,56] using the method described in Sec. III. Strain data are utilized from the open data release [56], while noise power-spectral densities are used from Ref. [57] produced with BayesWave [58,59]. Calibration error distributions are estimated for LIGO detectors in the first observing run and Virgo using the spline method [19,60]. Observations during the second observing run directly utilize the physical calibration model presented in

TABLE IV. Summary of results from GWTC-1 and the two injections described in Sec. IV. The efficiency, $\epsilon$, is defined in Eq. (15). The Bayes factor, $\mathcal{B}^{\Lambda}_{\varnothing}$, compares the likelihood obtained marginalizing over the calibration envelope to the marginal likelihood obtained ignoring any calibration error estimates. The Jensen-Shannon (JS) divergence measures the change in the posterior distribution when we include the calibration error estimate. For the SNR = 200 injection, no efficiency is given as the results were obtained via direct sampling of the marginalized likelihood.

| Event | $\epsilon$ [%] | $\mathcal{B}^{\Lambda}_{\varnothing}$ | Max. JS divergence (bit) |
|---|---|---|---|
| GW150914 | 78.2 | 0.97 | $\mathrm{JS}_{t_c} = 1.55 \times 10^{-3}$ |
| GW151012 | 99.7 | 0.97 | $\mathrm{JS}_{\mathcal{M}} = 5.05 \times 10^{-5}$ |
| GW151226 | 99.4 | 0.96 | $\mathrm{JS}_{\mathcal{M}} = 1.71 \times 10^{-4}$ |
| GW170104 | 98.7 | 0.96 | $\mathrm{JS}_{\phi_{\mathrm{JL}}} = 2.87 \times 10^{-5}$ |
| GW170608 | 97.4 | 1.12 | $\mathrm{JS}_{\mathrm{DEC}} = 2.98 \times 10^{-4}$ |
| GW170729 | 99.2 | 0.93 | $\mathrm{JS}_{\mathcal{M}} = 1.12 \times 10^{-4}$ |
| GW170809 | 99.3 | 0.91 | $\mathrm{JS}_{t_c} = 1.28 \times 10^{-4}$ |
| GW170814 | 98.0 | 1.08 | $\mathrm{JS}_{\mathrm{RA}} = 2.93 \times 10^{-4}$ |
| GW170817 | 64.9 | 1.93 | $\mathrm{JS}_{d_L} = 8.90 \times 10^{-4}$ |
| GW170818 | 98.9 | 1.06 | $\mathrm{JS}_{\phi_{\mathrm{JL}}} = 2.88 \times 10^{-4}$ |
| GW170823 | 98.9 | 0.97 | $\mathrm{JS}_{\mathrm{RA}} = 2.89 \times 10^{-5}$ |
| SNR = 30 inj | 90.3 | 0.78 | $\mathrm{JS}_{\mathrm{RA}} = 5.16 \times 10^{-3}$ |
| SNR = 200 inj | ⋯ | $2.04 \times 10^{-4}$ | $\mathrm{JS}_{\mathrm{RA}} = 1.05 \times 10^{-1}$ |

Sec. II. To illustrate the typical effect of the inclusion of the physical calibration model, we first consider GW170608. In the top panels of Fig. 2, we show the posterior distributions for the astrophysical parameters for GW170608. The red contours include marginalization over calibration error estimates while the blue contours do not. While there are small differences between the red and blue contours—we encourage the reader to squint at the posterior distributions for declination, DEC, and inclination, $\iota$—it is clear that the inclusion of uncertainty in the calibration error has a very small effect on the size and shape of the astrophysical posterior distributions. In the bottom panels of

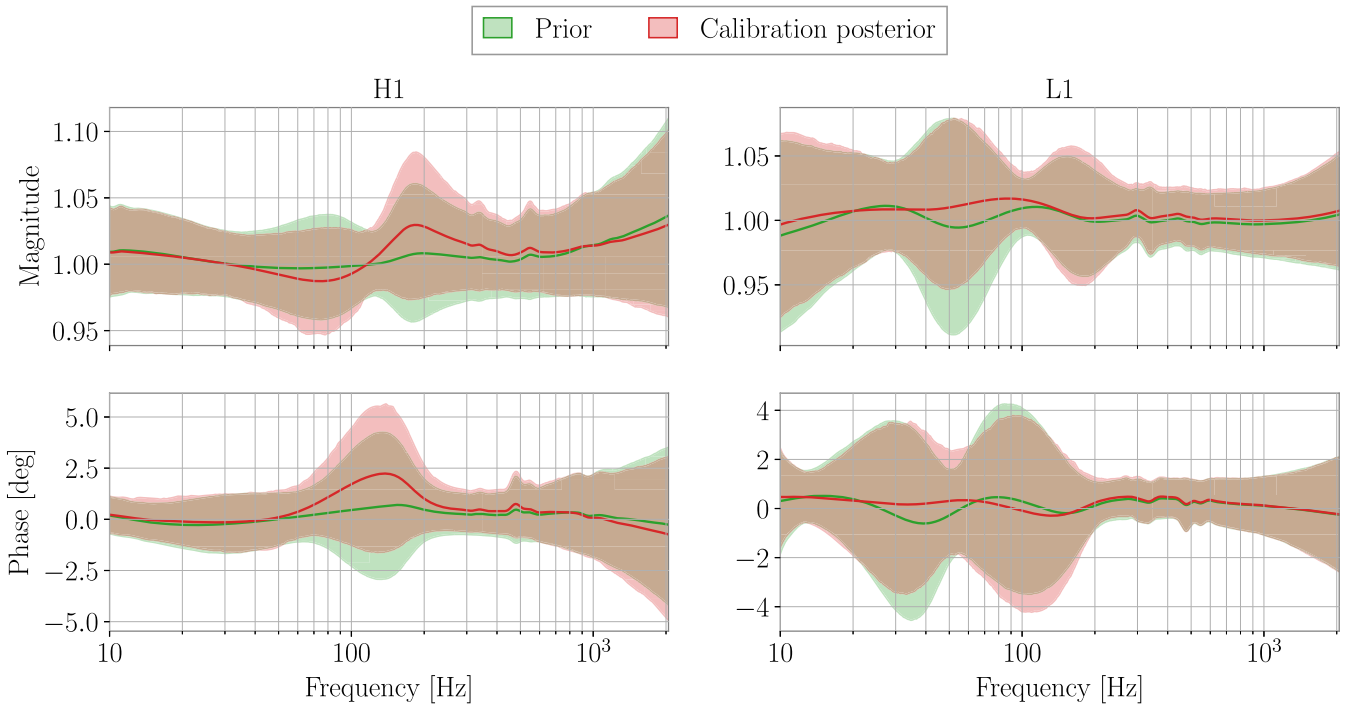


FIG. 3. Calibration response curve posterior distributions for both LIGO detectors informed by all events during the second observing run evaluated at the time of GW170729. Marginal shifts in the calibration error distributions are observed. The Bayes factor marginally favors the inclusion of the calibration error distribution over the zero-error hypothesis by 2.33.

Fig. 2 we show the reconstructed calibration response function. The thick red curve is averaged over draws from the calibration parameter posterior distribution while the green curve is averaged over draws from the prior. The slight difference between the red and green credible intervals shows a (small) change in the mean and 95% confidence intervals of the calibration envelope.

We also present the efficiencies and JS divergences for all events in GWTC-1 in Table IV. The efficiency for obtaining calibration-marginalized samples is $\epsilon = 78.2\%$ for GW150914, and $\epsilon = 64.9\%$ for GW170817. The nonunity efficiency for these two events is due to their larger network SNR. For other events in GWTC-1, we obtain efficiencies of $\epsilon = 97.4\%$–$99.7\%$. Visual inspection of the posterior distributions for the other events in GWTC-1 confirm that the effect of uncertainty in the calibration error is negligible for events in GWTC-1. This is further verified by JS divergences $\lesssim 1.5 \times 10^{-3}$, which are comparable to values found between different implementations of stochastic sampling algorithms [45] and smaller than differences due to differences in waveform models [3]. The full analysis of GWTC-1 results are available for download [61].

Finally, we conclude by determining the calibration envelope using events from the second observing run (O2) as standard sirens. We only use events from O2 to ensure that the time-independent calibration parameters are identical. We compute the combined weights for the calibration response curves following Eq. (16). With eight events in O2, the combined calibration envelope is only marginally informed by the gravitational-wave signals. The reconstructed envelope, evaluated at the time of GW170729, is presented in Fig. 3. We observe only a modest change from the prior. The total Bayes factor comparing the calibration error distribution hypothesis to the zero-error hypothesis is likewise modest: $\mathcal{B}^{\Lambda}_{\varnothing} = 2.33$. More events are required to meaningfully inform the calibration error estimate. However, with the requirement to periodically update the calibration model parameters as improvements to the detectors are made [18], the required number of events may not be achievable in the foreseeable future. This is also concluded within Ref. [23], where they comment that due to the periodic model updates, astrophysical calibration may never be competitive.

## VI. CONCLUSIONS

We have presented a calibration-marginalized likelihood for astrophysical parameters employing a physically informed model for the calibration error as presented in [16,18]. Within the signal-to-noise ratio regime of previously observed events and estimates of calibration errors at the levels reported in GWTC-1, we find the effect of calibration error is at the same level as the effect of stochastic sampling errors and less than other known systematics. Recent work from Ref. [36] has also investigated similar marginalization using direct sampling of the calibration error curve index, instead of importance sampling. The conclusions drawn within Ref. [36] are consistent with those drawn here. We also demonstrated that, if calibration errors remain as low as in GWTC-1, even future loud events will incur only modest changes in the estimates of astrophysical parameters, with the potential exception of increased uncertainty in the sky location. We also demonstrated the improved inference of calibration parameters using the collection of events from GWTC-1 as standard sirens. Our findings are consistent with [23], where it is found that using gravitational-wave events to improve the estimate of calibration errors beyond that determined from *in situ* measurements requires thousands of detections.

## ACKNOWLEDGMENTS

We thank Salvatore Vitale, Carl-Johan Haster, Lilli Sun, Ben Farr, and Evan Goetz for insightful comments and sharing an early version of their manuscript. We thank Nikhil Sarin and Rory Smith in providing guidance for the use of pBilby. We thank Greg Mendell and Rick Savage for thoughtful discussions, and Reed Essick for helpful comments on the manuscript. This work is supported through Australian Research Council (ARC) Centre of Excellence CE170100004. E. P. acknowledges the support of the LSC Fellows program. P. D. L. is supported through ARC Future Fellowship FT160100112, and ARC Discovery Project DP180103155. E. T. is supported through ARC Future Fellowship FT150100281. This is document LIGO-P2000294. This research has made use of data, software, and/or web tools obtained from the Gravitational Wave Open Science Center [62], a service of LIGO Laboratory, the LIGO Scientific Collaboration, and the Virgo Collaboration. 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. Computing was performed with computing clusters at California Institute of Technology (LIGO Laboratory) and Swinburne University of Technology (OzSTAR). We 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.

[1] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **116**, 061102 (2016).
[2] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **119**, 161101 (2017).
[3] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. X **9**, 031040 (2019).
[4] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Nature (London) **551**, 85 (2017).
[5] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Astrophys. J. Lett. **882**, L24 (2019).
[6] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **100**, 104036 (2019).
[7] B. F. Schutz, Classical Quantum Gravity **28**, 125023 (2011).
[8] S. Vitale, Phys. Rev. D **94**, 121501 (2016).
[9] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), arXiv:1904.03187.
[10] M. Pürrer and C.-J. Haster, Phys. Rev. Research **2**, 023151 (2020).
[11] S. Biscoveanu, C.-J. Haster, S. Vitale, and J. Davies, Phys. Rev. D **102**, 023008 (2020).
[12] C. Talbot and E. Thrane, arXiv:2006.05292.
[13] 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).
[14] S. Vitale, W. D. Pozzo, T. G. F. Li, C. V. D. Broeck, B. Aylott, and J. Veitch, Phys. Rev. D **85**, 064034 (2012).
[15] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. D **95**, 062003 (2017).
[16] C. Cahillane *et al.*, Phys. Rev. D **96**, 102001 (2017).
[17] F. Acernese *et al.* (Virgo Collaboration), Classical Quantum Gravity **35**, 205004 (2018).
[18] L. Sun *et al.*, Classical Quantum Gravity **37**, 225008 (2020).
[19] W. M. Farr, B. Farr, and T. Littenberg, Modelling calibration errors in CBC waveforms, CERN Tech. Report No. LIGO-T1400682, 2014.
[20] E. Payne, C. Talbot, and E. Thrane, Phys. Rev. D **100**, 123017 (2019).
[21] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. (Springer, New York, 2004).
[22] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, 1st ed. (Springer, New York, 2004).
[23] R. Essick and D. E. Holz, Classical Quantum Gravity **36**, 125002 (2019).
[24] J. Aasi *et al.* (LIGO Scientific Collaboration), Classical Quantum Gravity **32**, 074001 (2015).
[25] H.-Y. Chen and D. E. Holz, arXiv:1409.0522.
[26] F. Acernese *et al.* (Virgo Collaboration), Classical Quantum Gravity **32**, 024001 (2015).
[27] K. Izumi and D. Sigg, Classical Quantum Gravity **34**, 015001 (2017).
[28] N. Robertson *et al.*, Classical Quantum Gravity **19**, 4043 (2002).
[29] S. Aston *et al.*, Classical Quantum Gravity **29**, 235004 (2012).
[30] L. Carbone *et al.*, Classical Quantum Gravity **29**, 115005 (2012).
[31] O. Miyakawa *et al.*, Phys. Rev. D **74**, 022001 (2006).
[32] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, Publ. Astron. Soc. Pac. **125**, 306 (2013).
[33] The analysis undertaken in this manuscript does not incorporate the probability distribution from MCMC sampling for $f_{CC}(t)$ following an implementation error in the analysis undertaken in Ref. [18]. This omission has only a marginal effect on the calibration error estimate.
[34] S. Karki *et al.*, Rev. Sci. Instrum. **87**, 114503 (2016).
[35] D. Bhattacharjee, Y. Lecoeuche, S. Karki, J. Betzwieser, V. Bossilkov, S. Kandhasamy, E. Payne, and R. Savage, arXiv:2006.00130.
[36] S. Vitale, C.-J. Haster, L. Sun, B. Farr, E. Goetz, J. Kissel, and C. Cahillane, arXiv:2009.10192.
[37] M. Seeger, Int. J. Neural Systems **14**, 69 (2004).
[38] F. Pedregosa *et al.*, J. Mach. Learn. Res. **12**, 2825 (2011).
[39] E. Thrane and C. Talbot, Pub. Astron. Soc. Aust. **36**, E010 (2019).
[40] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016).
[41] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).
[42] T. Dietrich, S. Bernuzzi, and W. Tichy, Phys. Rev. D **96**, 121501 (2017).
[43] To understand the application of the calibration error model to the gravitational-wave template, we note that the measured strain is

$$h = \frac{1}{L} R_{\text{true}} d_{\text{err}}, \tag{17}$$

where $R_{\text{true}}$ is the true response function of the interferometer. By writing the measured strain in terms of our model for the true response function, $R(\Lambda)$, the measured strain is

$$h = \frac{R(\Lambda)}{R_{\varnothing}} h_{\varnothing}, \tag{18}$$

where $R_{\varnothing}$ and $h_{\varnothing}$ are the theoretical response function and strain, respectively. Therefore, in order for the gravitational-wave model to be accurately subtracted from the measured strain in the Gaussian likelihood (6), a correction to the detector response using $R(\Lambda)/R_{\varnothing}$ must be propagated into the template. When $R(\Lambda) = R_{\varnothing}$, the measured strain precisely corresponds to the theoretical strain expected.
[44] G. Ashton *et al.*, Astrophys. J. Suppl. Ser. **241**, 27 (2019).
[45] I. M. Romero-Shaw *et al.*, Mon. Not. R. Astron. Soc. **499**, 3295 (2020).
[46] J. S. Speagle, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).
[47] J. Skilling, AIP Conf. Proc. **735**, 395 (2004).
[48] P. Landry and R. Essick, Phys. Rev. D **99**, 084049 (2019).
[49] L. Kish, *Survey Sampling*, 3rd ed. (Wiley-Interscience, Oxford, England, 1995).
[50] V. Elvira, L. Martino, and C. P. Robert, arXiv:1809.04129.
[51] The Virgo observatory uses a spline-based calibration error model [19].
[52] R. J. E. Smith, G. Ashton, A. Vajpeyi, and C. Talbot, Mon. Not. R. Astron. Soc. **498**, 4492 (2020).
[53] J. Lin, IEEE Trans. Inf. Theory **37**, 145 (1991).

[54] S. Kullback and R. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[55] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).
[56] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), arXiv:1912.11716.
[57] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), https://dcc.ligo.org/LIGO-P1900011/public.
[58] N. J. Cornish and T. B. Littenberg, Classical Quantum Gravity **32**, 135012 (2015).
[59] T. B. Littenberg and N. J. Cornish, Phys. Rev. D **91**, 084034 (2015).
[60] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), https://dcc.ligo.org/LIGO-P1900040/public.
[61] E. Payne, C. Talbot, P. Lasky, E. Thrane, and J. Kissel, https://git.ligo.org/ethan.payne/gwtc1_calibration_reweighting (2020).
[62] https://www.gw-openscience.org.