Impact of calibration uncertainties on Hubble constant measurements from gravitational-wave sources

Yiwen Huang,1,2, * Hsin-Yu Chen,1,2 Carl-Johan Haster,1,2 Ling Sun,3 Salvatore Vitale,4 and Jeff Kissel4

1LIGO Laboratory, Massachusetts Institute of Technology, 185 Albany St, Cambridge, MA 02139, USA
2Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA
3OzGrav-ANU, Centre for Gravitational Astrophysics, College of Science, The Australian National University, ACT 2601, Australia
4LIGO Hanford Observatory, Richland, WA 99352, USA

(Dated: April 8, 2022)

Gravitational-wave (GW) detections of electromagnetically bright compact binary coalescences can provide an independent measurement of the Hubble constant $H_0$. In order to obtain a measurement that could help arbitrating the existing tension on $H_0$, one needs to fully understand any source of systematic biases for this approach. In this study, we aim at understanding the impact of instrumental calibration errors (CEs) and uncertainties on luminosity distance measurements, $D_L$, and the inferred $H_0$ results. We simulate binary neutron star mergers (BNSs), as detected by a network of Advanced LIGO and Advanced Virgo interferometers at their design sensitivity. We artificially add CEs equal to exceptionally large values experienced in LIGO-Virgo’s third observing run (O3). We find that for individual BNSs at a network signal-to-noise ratio of 50, the systematic errors on $D_L$—and hence $H_0$—are still smaller than the statistical uncertainties. The biases become more significant when we combine multiple events to obtain a joint posterior on $H_0$. In the rather unrealistic case that the data around each detection is affected by the same CEs corresponding to the worst offender of O3, the true $H_0$ value would be excluded from the 90% credible interval after ~ 40 sources. If instead 10% of the sources suffer from severe CEs, the true value of $H_0$ is included in the 90% credible interval even after we combine 100 sources.

I. INTRODUCTION

The Hubble constant, $H_0$, is the current expansion rate of the Universe and serves an important role in our understanding of the cosmic expansion history. However, currently there exists a 4.4-$\sigma$ tension between the measurements of early and late Universe. Planck satellite’s observations of the cosmic microwave background anisotropies, assuming the standard flat cosmological model, lead to an inferred late-Universe measurement of $H_0 = 67.36 \pm 0.54$ km s$^{-1}$ Mpc$^{-1}$ [1, 2]. Direct measurements in the local universe lead to a different result: the SH0ES team measured $H_0 = 74.03 \pm 1.42$ km s$^{-1}$ Mpc$^{-1}$ using the Cepheid-supernova distance ladder [3, 4], consistent with results from H0LiCOW using lensed quasars [5, 6], or from the Carnegie-Chicago Hubble program using the Tip of the Red Giant Branch method to calibrate the distances [7–9].

Compact binary coalescences that are both electromagnetically (EM) bright and emit substantial gravitational waves (GWs) provide an independent method to measure $H_0$, and have the potential to resolve this tension [10–18]. Multi-messenger observations of a compact binary coalescence yield a measurement of the luminosity distance $D_L$ from the GW observations, and a measurement of the Hubble flow velocity $v_H$ from the EM observations. Combining these measurements leads to a direct estimation of $H_0$, a method often referred to as “bright sirens” measurement. Since EM-bright GW sources detected by the second-generation GW detectors will be relatively local, we can only use bright sirens to constrain $H_0$. Other cosmological parameters can be measured with methods that rely on neutron star (NS) equation-of-state [19–21], dark sirens [10, 17, 22–24], features in the mass distribution of binary neutron star mergers (BNSs) and binary black hole mergers (BBHs) [25–28], or the spatial clustering scale of GW sources with known galaxies [29, 30].

The observations of GW170817 by the LIGO and Virgo network [31–34], together with the kilonova AT2017gfo and the gamma-ray burst GRB170817A, served as the first demonstration of the bright siren approach, leading to a measurement of $H_0 = 70.0^{+12.0}_{-8.0}$ km s$^{-1}$ Mpc$^{-1}$ [35–39]. This method can be very powerful in the future, as we observe more such events. Ref. [14] presented a 5% precision in the $H_0$ measurement after 15 BNSs with EM counterparts, and 1% with 30 sources. However, GW observations may suffer from underlying systematic biases in their estimate of $D_L$, and one needs to understand these biases before interpreting the resulting $H_0$ measurements. One such bias is the systematic error in the production and calibration of the detectors’ primary data stream [40–42], referred to as calibration errors (CEs). Such errors are uncorrelated with the astrophysical event rate, evolve independently in each detector, and may be present in one or more detectors during several astrophysical events. CEs may bias the amplitude of the strain data in the same way over multiple observations, and lead to a biased inference of $H_0$.

* ywh@mit.edu
Previous studies have estimated the impact of typical values of CEs seen in the LIGO-Virgo’s first and second observing runs on individual events alone [43–45], or attempted to refine the estimates of CEs using detected [46] or expected astrophysical signals [47]. Our study instead uses large systematic errors as experienced during atypical times in the LIGO-Virgo’s third observing run (O3) to probe their worst-case impact on astrophysical parameter estimation (PE) for both the single-event characterization and the joint inference of $H_0$.

More specifically, we simulate a collection of BNS detections, and introduce into the GW data stream artificial CEs that follow six cases of particularly large CEs from the Advanced LIGO detectors at Hanford (LHO) or Livingston (LLO) during O3 [41, 42]. We pick the instantiation that leads to the worst bias in the astrophysical measurements of $D_L$, apply it to varying fractions of 100 simulated BNSs and infer $H_0$ to explore the progressive impact of large CEs. In Sec. II, we will discuss the method we use to produce the miscalibrated data stream and to perform PE. In Sec. III, we report how CEs impact the $D_L$ results for individual events, as well as the $H_0$ results when we combine multiple sources.

II. METHOD

In Sec. II A, we summarize how we simulate the data stream; in Sec. II B how we define CEs and artificially add them to the data stream; and in Sec. II C how we perform PE and infer $H_0$.

A. Simulations

We simulate 5,000 BNSs with uniform-in-cosine inclinations and sky localizations uniformly distributed on the sky. Each event is assigned a $D_L$ randomly drawn from a uniform-in-comoving-volume distribution. The maximum $D_L$ is set at 600 Mpc, larger than the horizon$^1$ of a LIGO-Virgo network at the design sensitivity [31, 32]. We randomly draw 100 events with network SNRs above 12 to form our set of EM-bright GW events.

We simulate non-spinning BNSs with component masses $m_1 = 2 M_\odot$ and $m_2 = 1.5 M_\odot$, using the phenomenological waveform model IMRPhenomPv2 [48–51]. We use the same waveform model during PE in order to avoid any systematics caused by waveform mismatch$^2$. For all of the BNSs in our study, we do not include any NS tidal deformability in our analysis, as IMRPhenomPv2 does not model neutron star matter effects. We assume that an EM counterpart has been found and yielded an exact redshift measurement. Since the uncertainties on sky localization from the EM side are much smaller than the typical uncertainties on the GW side, we assume the sky position of the sources to be exactly known from the EM observations prior to the GW analysis. We also disregard effects from the peculiar velocity of the sources, as most of the events considered here are close to the horizon of GW detections.

We produce Gaussian noise n colored by the Advanced LIGO and Advanced Virgo design sensitivities [32]. We then project the signal, which is the sum of the modeled waveforms and the noise, to each detector of the LIGO-Virgo network to obtain the data stream.

B. Systematic Calibration Errors

Each of the current Advanced LIGO and Advanced Virgo detectors is a dual-recycled Fabry-Pérot Michelson laser interferometer (IFO) [55, 56], and its data stream $d_{IFO}$ is obtained from a voltage signal, $e_{IFO}$, measured from the output laser power incident on a photo detector, where the subscript “IFO” indexes the detector. The process of creating a model to convert $e_{IFO}$ into $d_{IFO}$ is referred to as calibration [57]. For each detector at any given time, the conversion from $e_{IFO}$ to $d_{IFO}$ is made by a complex-valued, frequency-dependent response function, $R(f; t)$,

$$
d_{IFO} = R_{IFO}(f; t) e_{IFO}^3. \tag{1}
$$

The model for the response function, $R_{model, IFO}(f; t)$, is constructed based on the expected behavior of the detectors, coupled with many supporting measurements of the parameters and components of the model as described in Ref. [58, 59]. Imperfections in $R_{model, IFO}(f; t)$, and thus in $d_{IFO}$, are referred to as calibration errors, or CEs [40–42]. The errors may be represented by the ratio of the true response function, $R_{true, IFO}(f; t)$, and the model, $R_{model, IFO}(f; t)$, by $^4$

$$
\eta_{IFO}(f; t) = R_{true, IFO}(f; t)/R_{model, IFO}(f; t). \tag{2}
$$

In the ideal case, $R_{model, IFO}(f; t)$ will be identical to $R_{true, IFO}(f; t)$, and $\eta_{IFO}(f; t)$ becomes a frequency-independent constant with unity magnitude and zero phase. In reality, $\eta_{IFO}(f; t)$ is usually a function of frequency and time. Since $R_{model, IFO}(f; t)$, $R_{true, IFO}(f; t)$ and $\eta_{IFO}(f; t)$ always depend on $f$ and $t$ in our case and evolve independently in every detector, we will drop the $(f; t)$ arguments and the IFO subscript henceforth unless we need to specify $f$, $t$, or the instrument.

While $R_{model}$ is known, $R_{true}$ may only be inferred by direct measurements that are invasive for observations

$^1$ For events with a network signal-to-noise ratio (SNR) of 12.

$^2$ Previous papers have looked into waveform systematics for BNSs, for example Ref. [52–54].

$^3$ $d_{IFO}$ and $e_{IFO}$ are also always functions of frequency $f$ and $t$, but we do not write their arguments out explicitly.

$^4$ $\eta_{IFO}(f; t)$ is referred to as $\eta_R(f; t)$ in Ref. [41, 42].
and would reduce the duty cycle of the detector. Instead, \( \eta \) is numerically estimated by creating a probability distribution informed by those direct measurements. At a given reference time, \( T_0 \), the parameters of \( R_{\text{model}} \) are measured to create \( 10^4 \) samples of the probability distribution of \( \{ \eta(T_0) \} \), with relatively large uncertainty. Later, additional measurements and models are crafted to better inform or correct the estimated \( \{ \eta(T_0) \} \), in retrospect. In O3, the probability distribution of \( \eta \) for each detector was estimated regularly at discrete times \( \{ T \} \) with hourly cadence for times while the detector was stable in the observing mode.

The estimated \( \{ \eta \} \) may vary slowly in time, either continuously with drifts in the detector’s alignment or thermal state, or discretely with changes made in the detector’s control system between different configuration periods. In most cases, the slow time-dependent variation is appropriately tracked and accounted for when producing \( e_{\text{IFO}} \), and thus the time variation of \( \{ \eta \} \) is negligible for the majority of the observing time as long as the detector’s sensing and control configuration remain unchanged. Over such a particular observing period within static configuration, we can choose an arbitrary moment in time \( T_k \) such that \( \{ \eta(T_k) \} \) represents the typical CE distribution in that period, and refer to this distribution as \( \{ \eta^{\text{typ}}(T_k) \} \).

However, the detectors’ sensing and control configurations do not remain static for the entire observing run. There may be planned changes (typically to improve the detector sensitivity) or unplanned changes (due to issues with the hardware, electronics, or computers). Such changes may have either unexpected impact on the calibration, or get fixed in a temporary fashion such that the observation may resume (with an acceptable error) but may not be fully investigated and resolved until later. The entire detector’s response, \( R_{\text{model}} \), is checked by weekly measurements, and monitored continuously at a few select frequencies, to ensure coverage of any unexpected situations. When issues are found in the detector behavior, additional measurements may also made such that error may be modeled and included in \( \{ \eta \} \) retroactively at a later time. If at a time \( t_i \), \( \{ \eta(t_i) \} \) is modeled to be significantly different from \( \{ \eta^{\text{typ}}(T_i) \} \) at a nearby time \( T_i \), we call these outliers \( \{ \eta^{\text{out}}(t_i) \} \), henceforth \( \{ \eta^{\text{out}} \} \).

During O3, six such outliers have been identified in either LLO or LHO. We refer to the corresponding typical distribution as \( \{ \eta^{\text{typ}} \} \).

For example, in Fig. 1, we compare the median and 1-\( \sigma \) boundaries of the magnitude and the phase of \( \{ \eta^{\text{out}} \} \) against its corresponding \( \{ \eta^{\text{typ}} \} \). We include plots for the distributions of \( \{ \eta^{\text{out}} \} \) and \( \{ \eta^{\text{typ}} \} \) for \( i = 1 \ldots 5 \) in App. A. In reality, if there is an astrophysical signal at a time similar to one of the \( t_i \), its \( \{ \eta^{\text{out}} \} \) may not be readily available at the time of PE analyses, and \( \{ \eta^{\text{typ}} \} \) will be used instead. This is the scenario we investigate in this study: when the actual CEs are very different from the CEs known and used at the time of PE.

We introduce artificial CEs to the data streams \( d_{\text{LLO}} \) and \( d_{\text{LHO}} \). For all of the six cases, \( \{ \eta^{\text{out}} \} \), is only in one of the Advanced LIGO detectors. We will first select the worst realization from each \( \{ \eta^{\text{out}} \} \), denoted by \( \eta^{\text{min}} \), such that it maximizes its impact on the amplitude of the data, and hence on the estimation of the source distance. We define the impact, weighted by the detector sensitivity, and integrated over the bandwidth:

\[
D_{i,j} \equiv \int_{f_{\text{low}}}^{f_{\text{high}}} \frac{|\eta^{\text{typ}}| - |\eta^{\text{out}}_j|}{\sqrt{S_n(f)}} \, df,
\]

where \( j \) indexes the curves from each \( \{ \eta^{\text{out}} \} \) distribution. At each frequency, we take the difference between \( |\eta_{j,j}| \), the magnitude of the \( j \)-th sample of \( \{ \eta^{\text{out}} \} \), and \( |\eta^{\text{typ}}_i| \), the median of the magnitudes of samples in \( \{ \eta^{\text{typ}} \} \), (the dotted orange curve in Fig. 1). \( \sqrt{S_n(f)} \) is the amplitude spectral density of the detector’s sensitivity, for which we use aLIGO’s design sensitivity [31, 32]. The frequency limits of the integral, \( f_{\text{high}} \) and \( f_{\text{low}} \), have been chosen as 20 Hz and 1024 Hz, respectively, with a frequency resolution of 0.25 Hz. While the maximum difference may
correspond to a positive or negative $D_{i,j}$ value, we select the curve that maximizes $D_{i,j}$ in the negative direction\(^5\), denoted by $\eta_i^{\text{mis}}$,

$$\eta_i^{\text{mis}} \equiv \min_j D_{i,j}. \quad (4)$$

Fig. 2 shows the amplitude and phase of the selected $\eta_i^{\text{mis}}$ compared against $\{\eta^{\text{typ}}\}^6_{\text{LLO}}$ from Fig. 1. We can use this curve to miscalibrate the LLO data, the sum of the noise and the modeled waveform, as,

$$d_i^{\text{mis}} = \eta_i^{\text{mis}} \ d_i. \quad (5)$$

The noise, as part of $d$, will thus also be scaled by $\eta_i^{\text{mis}}$. The resulted amplitude spectral density of each detector is $\eta_i^{\text{mis}} \sqrt{S_n(f)}$.

For the other Advanced LIGO detector, in this case LHO, we will apply a sample curve $\eta_{i,\text{LHO}}^{\text{mis}}$ that lies within the 1-$\sigma$ credible interval (CI) from $\{\eta^{\text{typ}}\}^6_{\text{LHO}}$ to the data, in the same way as Eq. (5). No CEs are added to aVirgo data, since the full $\eta$ distributions of aVirgo are not available at the time of writing. In this case, $\{\eta_{6,\text{LLO}}, \eta_{6,\text{LHO}}, 1\}$ forms the set of curves to miscalibrate the data for scenario #6.

We also prepare a separate set of control runs in which we do not add any CE in any of the detectors. When comparing PE results of the miscalibrated runs and the control runs, we thus observe biases exclusively caused by the large added CEs in the former.

Next, we select one CE realization that leads to the most significant bias in the $D_L$ likelihood, and apply it to 100 segments of data, each containing a different BNS event, as described in Sec. II A. To select the desired CE realization, we first consider a single BNS event with network SNR of 50, inclination 30°, and located right above LLO in the sky. We add each of the six distinctive sets of CEs to a data chunk containing this BNS, and compare the resulting $D_L$ likelihoods. As we only consider single-event results in this part of the study, we work with zero noise [60] to eliminate random effects of noise realizations. We finally use the zero-noise realization where the noise $n$ has mean of 0 and standard deviation equal to the detector power spectral density. Lastly, we infer $H_0$ from the 100 events.

### C. Parameter Estimation and Inference of $H_0$

We perform Bayesian inference [61, 62] to obtain the likelihood $\mathcal{L}(d|\theta, H_0, \mathcal{H})$, where $\theta$ is the set of astrophysical parameters, under model $\mathcal{H}$. We write $H_0$ explicitly because it is a hyperparameter that we are interested in.

\[ \mathcal{L}(d|\theta, H_0, \mathcal{H}) = \int d\theta \mathcal{L}(d|\theta, H_0, \mathcal{H}) \pi(\theta|H_0, \mathcal{H}), \quad (6) \]

where $\pi(\theta|H_0, \mathcal{H})$ is the prior. We use a uniform prior for $H_0|\mathcal{H} = [20, 150] \text{ km s}^{-1} \text{ Mpc}^{-1}$ to obtain a posterior distribution $p(H_0|d, \mathcal{H})$ from the likelihood in Eq. (6). Since not all events have equal chances to be detected, we follow Ref. [15, 63] to account for selection effects, and call the function of detectable sources $\beta(H_0)$. Since all sources in our simulations have the same source-frame masses, the selection function only needs to be calculated once for all the events. We can finally obtain the posterior for $H_0$ as,

\[ p(H_0|d, \mathcal{H}) = \frac{\pi(H_0|\mathcal{H}) \mathcal{L}(d|H_0, \mathcal{H})}{\beta(H_0)}. \quad (7) \]

As all of the 100 BNSs are independent, their joint likelihood can be calculated by simply multiplying the likelihoods of each event [64],

\[ \mathcal{L}(d|H_0, \mathcal{H}) = \prod_{i=1}^{100} \mathcal{L}_i(d|H_0, \mathcal{H}). \quad (8) \]
One caveat is that even though our simulated events all have the same masses, our PE analysis uses a standard, uniform-in-individual-masses prior with additional bounds on the chirp mass $M \equiv (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$ and mass ratio. This may introduce a small bias since the function of event detectability in Eq. (7) is calculated assuming that all BNSs considered have the same masses. However, the bias will be present for both the miscalibrated runs and the control runs, thus not affecting their differential behaviors.

We can marginalize over CEs during PE by two approaches: the Spline interpolation method [65] and the PhysiCal method [44, 45]. Both methods treat the CEs in one detector as independent from another. The Spline method models $\eta$ by fitting a cubic spline polynomial at a small number of logarithmically-spaced frequencies $\{f_m\}$, for each detector independently. At each frequency, the prior on the magnitude and phase is a Gaussian distribution with the same mean and standard deviation as those of $\{\eta^{\text{typ}}(f_m)\}$. The recently developed PhysiCal method [44, 45] is more computationally efficient and physically motivated, and estimates the physical parameters in the models for $\eta$ along with $\theta$ during PE. PhysiCal directly uses draws from the distributions of $\{\eta^{\text{typ}}\}$ as priors.

In all PE analyses to-date, known CEs are marginalized over. In this study, however, we are interested in scenarios where large calibration errors $\eta^{\text{mis}}$ are not fully captured and marginalized over in PE. We assume that we only know distributions $\{\eta^{\text{typ}}\}_i$, rather than the true error distribution $\{\eta^{\text{true}}\}_i$ at the event times, and use $\{\eta^{\text{typ}}\}_i$ as the prior for PhysiCal. Similarly, we utilize the medians and bounds of 1-$\sigma$ CIs of $\{\eta^{\text{typ}}\}_i$ as priors for the Spline method. As mentioned in Sec. II B, since there the full $\{\eta\}$ distributions of aVirgo are not available, we only adopt the Spline method to marginalize CEs in aVirgo.

### III. RESULT

We first present $D_L$ likelihoods when we apply the six sets of $\eta^{\text{mis}}$ (selected using the method described in Sec. II B) to a single BNS event with a network SNR of 50. The likelihoods for the miscalibrated runs are plotted as green kernel density plots in Fig. 3, and the results for the corresponding control run are plotted in orange. The different set-ups for miscalibrated runs and control runs are detailed in Sec. II B. We report the difference between the true value $D_{L,\text{true}}$ and the median of the recovered $D_L$ likelihoods, normalized by the true value: $\Delta D_L = (D_{L,\text{med}} - D_{L,\text{true}})/D_{L,\text{true}}$, in Tab. I.

The results from using the PhysiCal and Spline methods for all three detectors agree very well. Thus, we only show the results from the PhysiCal methods in Fig. 3 and Tab. I, and the ones from the Spline method in Appendix B.

Comparing with the control runs, where the offsets are between $-2.1\%$ to $-2.6\%$, the CE realization #6 leads to the most significant differences in the $D_L$ likelihoods.

Thus, we apply it to the data of a hundred BNSs. We note the control runs all show a negative $\Delta D_L$; that can be explained by the well-known correlation between $D_L$ and inclination [66]. Since all sources in Fig. 3 have an inclination, $i$, of 30$^\circ$, and the inclination prior goes like $\sin i$, we expect an offset towards larger inclination values, where the prior is larger, and thus smaller distances. This offset will no longer be present in the analysis of 100 events, since the 100 events have inclinations drawn from a uniform-in-cosine distribution (effectively $\sin i dx$), the same as the prior.

In Fig. 4, we vary the fraction of BNSs whose data are miscalibrated with CE realization #6 and plot the posterior distribution of the dimensionless $h_0$, defined as $H_0 = h_0 100 \text{ km s}^{-1} \text{Mpc}^{-1}$. We assign CE #6 to $x\%$ of the BNS events, while the other events do not suffer from

![FIG. 3: $D_L$ likelihood for the six scenarios, miscalibrated (green) vs. control (orange) runs, the vertical dashed lines mark the 25%, 50% and 75% percentiles.](image)

| Label | Miscalibrated | Control |
|-------|--------------|---------|
| #1    | $-1.0\%$    | $-2.5\%$ |
| #2    | $-2.3\%$    | $-2.3\%$ |
| #3    | $-1.2\%$    | $-2.1\%$ |
| #4    | $-0.9\%$    | $-2.3\%$ |
| #5    | $-1.5\%$    | $-2.5\%$ |
| #6    | 0.5\%       | $-2.6\%$ |

**TABLE I:** $\Delta D_L$ in the likelihoods for PhysiCal runs with and without large CEs.
any miscalibration. The joint $h_0$ posterior starts to shift towards smaller values as we increase the fraction of BNSs affected by CE #6, and the posteriors exclude the true value of $h_0 = 0.679$ from the 90% CI when the data of more than 50% of the 100 BNSs are miscalibrated.

Next, we vary the total number of detected events besides varying the percentages of miscalibrated events. For $n$ detected events, we randomly miscalibrate $x\%$ of the $n$ events, and repeat the random draw for $\lceil 100/n \rceil$ trials (i.e. rounding down $100/n$ to an integer) for each pair of $\{n, x\}$. The number of trials is reduced as $n$ increases in order to keep each trial independent, and no event appears in more than one trial. We calculate the median and the bounds of 90% CI for each trial. In Fig. 5, the results are plotted for each pair of $\{n, x\}$; the green points are medians of the $\lceil 100/n \rceil$ medians from each trial, and the error bars are the medians of the 90% CI bounds from each trial. The orange points and error bars indicate the results for the control runs.

If 100% of the detected events suffer from miscalibration, the joint $h_0$ posterior excludes the true value from its 90% CI after 40 detections or more. When 50% of the detected BNSs suffer from miscalibration, the posterior only start to exclude the true $h_0$ after more than 90 events, and when 10% of the detected BNSs suffer from miscalibration, the posterior includes the true $h_0$ even after 100 detections. In the case of 100 detections, the results in Fig. 5 (the rightmost sets of points in each subplot) represent the same scenario as Fig. 4.

### IV. CONCLUSION

GW observations of EM-bright compact binary coalescences provide an independent way to measure the Hubble constant and to potentially break the existing tension between the early and late universe $H_0$ measurements. As we observe more of such events, the resulting $H_0$ posterior will be increasingly constraining, making it important to thoroughly control and understand potential systematic
biases.

In this study, we artificially added large CEs to the GW data stream, and investigated their effects on the inference of $H_0$. Our analysis is constructed not to contain any systematic errors or statistical uncertainties from the EM observations like peculiar velocities [67, 68], or viewing angles [69], for our inference of $H_0$. Since we simulate and recover the GW signals with the same waveform family (IMRPhenomPv2), there are no waveform systematics present in our study either.

We found that the $H_0$ posterior does not exclude the true value from its 90\% CI, corresponding to a 2–3\% systematic error, unless we are inferring $H_0$ with more than 40 BNS detections that all suffer from the same large CEs. The total number of detections required increases to $\sim 90$ BNSs when 50\% of them suffer from miscalibration. When 10\% of BNSs are miscalibrated, the true value is not excluded even after 100 BNS detections. For comparison, systematic errors due to the EM observation side, for example kilonova viewing angles, will be 2\% after 50 BNS detections [69].

All of the outliers $\{\eta_{\text{out}}\}$, that motivate our study are based on modeled systematic error surrounding times of real physical changes or malfunctions in the O3 detectors. These events are generally rare and relatively short-lived; typically $<1\%$ of the time over any few-month observing period – the typical duration of stable detector configurations during an observing run [41, 42]. In our analysis, we assume we only know and use $\{\eta^{\text{typ}}\}$, to marginalize over CEs during PE, while $\{\eta_{\text{out}}\}$, although resent through some fraction of detected BNS events, is assumed to be unknown and uncharacterized. There is a very low probability that a large CE remains uncharacterized over a period of many months during which dozens of BNSs are detected, given the current estimate of astrophysical event rate, $320^{+490}_{−240}$ Gpc$^{-1}$ yr$^{-1}$ [70] vs. the few-month duration of stable detector configurations and the frequency of measurements during those periods.

Our results imply that CEs are not going to be significant concern in the measurement of the Hubble constant with the bright sirens method for the next many years. In the most realistic case, where large instances of CE like the ones described in this paper affect a small percent of the sources, CE will not become the limiting factor until more than 100 BNSs, each with an EM counterpart, have been found. Since the bright siren method is likely to provide the smallest statistical uncertainties, other approaches to constrain $H_0$ using distance measurements from GW sources are going to be even less sensitive to CEs.

In this work we have focused on the effect of CEs on inference results of BNSs. We will report on other types of compact binary coalescences, such as neutron star black hole mergers and binary black holes, as well as on the posteriors for calibration parameters from PhysiCal, in a forthcoming paper.

V. ACKNOWLEDGMENTS

Software: The authors acknowledge the use of the LIGO Algorithm Library [71], and specifically of the LALInference inference package [61], as released through conda anaconda. Plots were produced with matplotlib [72]. The authors acknowledge use of iPython [73], NumPy [74] and SciPy [75]. The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants PHY-0757058 and PHY-0823459.

The authors would like to thank Jonathan Gair, Evan Goetz, Martin Hendry, Daniel Holz, Suadip Mukherjee, Ethan Payne Jameson Graef Rollins, Nicola Tamanini, Paxton Turner, Madeline Wade, Alan J. Weinstein for useful discussions. Y. H., C.-J. H., S. V. acknowledge support of the National Science Foundation, and the LIGO Laboratory. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under cooperative agreement PHY-0757058. S. V. and Y. H. acknowledge support of the National Science Foundation trough the award PHY-2045740. Virgo is funded by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale della Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by Polish and Hungarian institutes. H.-Y. C. is supported by NASA through NASA Hubble Fellowship grants No. HST-HF2-51452.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. L. S. acknowledges the support of the Australian Research Council Centre of Excellence for Gravitational Wave Discovery (OzGrav), Project No. CE170100004. This is LIGO Document Number DCC-P2100350.
Appendix A: $\eta^{\text{mis}}$, {$\eta^{\text{out}}$}, and {$\eta^{\text{typ}}$} for other five scenarios

Here, in Fig. 6–10, we show the $\eta^{\text{mis}}$, {$\eta^{\text{out}}$}, and {$\eta^{\text{typ}}$} for the other five calibration outlier cases identified in O3.

![Graph showing amplitude and phase](image_url)

FIG. 6: Time #1 of large amplitude and phase errors, {$\eta^{\text{out}}$} (blue), compared to the corresponding typical distribution, {$\eta^{\text{typ}}$} (orange), both showing the edges of the 1-σ CIs in each frequency bin. Also plotted in green is the $\eta^{\text{mis}}$ for this time.

In Fig. 11, we show the distance likelihoods for BNSs at an SNR of 50, where the green and pink shaded distributions are obtained from the runs with the PhysiCal and Spline methods, respectively, both effected by the same $\eta^{\text{mis}}$. We report $\Delta D_L$ in Tab. II. The differences between the results are quite small compared to those between the PhysiCal runs with and without CEs.

---

[1] T. M. C. Abbott et al. (DES), Mon. Not. Roy. Astron. Soc. 480, 3879 (2018), arXiv:1711.00403 [astro-ph.CO].
[2] N. Aghanim et al. (Planck), Astron. Astrophys. 641, A6 (2020), arXiv:1807.06209 [astro-ph.CO].
[3] A. G. Riess et al., Astrophys. J. 826, 56 (2016), arXiv:1604.01424 [astro-ph.CO].
[4] A. G. Riess et al., Astrophys. J. 861, 126 (2018), arXiv:1804.10655 [astro-ph.CO].
[5] S. Birrer et al., Mon. Not. Roy. Astron. Soc. 484, 4726 (2019), arXiv:1809.01274 [astro-ph.CO].
[6] K. C. Wong et al., Mon. Not. Roy. Astron. Soc. 498, 1420 (2020), arXiv:1907.04869 [astro-ph.CO].
[7] I. S. Jang and M. G. Lee, Astrophys. J. 836, 74 (2017), arXiv:1702.01118 [astro-ph.CO].
[8] W. L. Freedman et al., (2019), 10.3847/1538-4357/ab2f73, arXiv:1907.05922 [astro-ph.CO].
[9] Y. J. Kim, J. Kang, M. G. Lee, and I. S. Jang, Astrophys. J. 905, 104 (2020), arXiv:2010.01364 [astro-ph.CO].
[10] B. F. Schutz, Nature 323, 310 (1986).
[11] D. E. Holz and S. A. Hughes, Astrophys. J. 629, 15 (2005), arXiv:astro-ph/0504616.
[12] N. Dalal, D. E. Holz, S. A. Hughes, and B. Jain, Phys. Rev. D 74, 063006 (2006), arXiv:astro-ph/0601275.
[13] S. Nissanke, D. E. Holz, S. A. Hughes, N. Dalal, and J. L. Sievers, Astrophys. J. 725, 496 (2010), arXiv:0904.1017
FIG. 7: Time #2 of large amplitude and phase errors, \(\{\eta^{\text{out}}\}\) (blue), compared to the corresponding typical distribution, \(\{\eta^{\text{typ}}\}\) (orange), both showing the edges of the 1-\(\sigma\) CIs in each frequency bin. Also plotted in green is the \(\eta^{\text{mis}}\) for this time.

FIG. 8: Time #3 of large amplitude and phase errors, \(\{\eta^{\text{out}}\}\) (blue), compared to the corresponding typical distribution, \(\{\eta^{\text{typ}}\}\) (orange), both showing the edges of the 1-\(\sigma\) CIs in each frequency bin. Also plotted in green is the \(\eta^{\text{mis}}\) for this time.

| Label | PhysiCal | Spline |
|-------|----------|--------|
| #1    | -0.8%    | -1.5%  |
| #2    | -2.0%    | -2.6%  |
| #3    | -1.6%    | -1.4%  |
| #4    | -0.8%    | -0.6%  |
| #5    | -1.7%    | -1.2%  |
| #6    | 0.2%     | 1.0%   |

TABLE II: \(\Delta D_L\) in the likelihoods for PhysiCal vs. Spline results.

[astro-ph.CO].

[14] S. Nissanke, D. E. Holz, N. Dalal, S. A. Hughes, J. L. Sievers, and C. M. Hirata, (2013), arXiv:1307.2638 [astro-ph.CO].

[15] H.-Y. Chen, M. Fishbach, and D. E. Holz, Nature **562**, 545 (2018), arXiv:1712.06531 [astro-ph.CO].

[16] K. Hotokezaka, E. Nakar, O. Gottlieb, S. Nissanke, K. Masuda, G. Hallinan, K. P. Mooley, and A. T. Deller, Nature Astron. **3**, 940 (2019), arXiv:1806.10596 [astro-ph.CO].

[17] M. Soares-Santos et al. (DES, LIGO Scientific, Virgo), Astrophys. J. Lett. **876**, L7 (2019), arXiv:1901.01540 [astro-ph.CO].

[18] S. M. Feeney, H. V. Peiris, S. M. Nissanke, and D. J. Mortlock, Phys. Rev. Lett. **126**, 171102 (2021), arXiv:2012.06593 [astro-ph.CO].

[19] C. Messenger and J. Read, Phys. Rev. Lett. **108**, 091101 (2012), arXiv:1107.5725 [gr-qc].

[20] W. Del Pozzo, T. G. F. Li, and C. Messenger, Phys. Rev. D **95**, 043502 (2017), arXiv:1506.06590 [gr-qc].

[21] T. Dietrich, M. W. Coughlin, P. T. H. Pang, M. Bulla, J. Heinzel, L. Issa, I. Tews, and S. Antier, Science **370**, 1450 (2020), arXiv:2002.11355 [astro-ph.HE].

[22] W. Del Pozzo, Phys. Rev. D **86**, 043011 (2012), arXiv:1108.1317 [astro-ph.CO].

[23] B. P. Abbott et al. (LIGO Scientific, Virgo), Astrophys. J. **909**, 218 (2021), arXiv:1908.06060 [astro-ph.CO].

[24] R. Gray et al., Phys. Rev. D **101**, 122001 (2020), arXiv:1908.06050 [gr-qc].

[25] S. R. Taylor, J. R. Gair, and I. Mandel, Phys. Rev. D **85**, 023535 (2012), arXiv:1108.5161 [gr-qc].

[26] W. M. Farr, M. Fishbach, J. Ye, and D. Holz, Astrophys. J. Lett. **883**, L42 (2019), arXiv:1908.09084 [astro-ph.CO].
FIG. 9: Time #4 of large amplitude and phase errors, \( \{\eta^{\text{typ}}\}_{LLO} \) (blue), compared to the corresponding typical distribution, \( \{\eta^{\text{typ}}\}_{LHO} \) (orange), both showing the edges of the 1-\( \sigma \) CIs in each frequency bin. Also plotted in green is the \( \eta^{\text{mis}} \) for this time.

FIG. 10: Time #5 of large amplitude and phase errors, \( \{\eta^{\text{typ}}\}_{LLO} \) (blue), compared to the corresponding typical distribution, \( \{\eta^{\text{typ}}\}_{LHO} \) (orange), both showing the edges of the 1-\( \sigma \) CIs in each frequency bin. Also plotted in green is the \( \eta^{\text{mis}} \) for this time.

[27] S. Mastrogiavanni, K. Leyde, C. Karathanasis, E. Chassande-Mottin, D. A. Steer, J. Gair, A. Ghosh, R. Gray, S. Mukherjee, and S. Rinaldi, Phys. Rev. D 104, 062009 (2021), arXiv:2103.14663 [gr-qc].

[28] A. Finke, S. Foffa, F. Iacovelli, M. Maggiore, and M. Mancarella, JCAP 08, 026 (2021), arXiv:2101.12660 [astro-ph.CO].

[29] S. Mukherjee, B. D. Wandelt, S. M. Nissanke, and A. S. L. Barsotti, Phys. Rev. D 103, 043520 (2021), arXiv:2007.02943 [astro-ph.CO].

[30] C. C. Diaz and S. Mukherjee, (2021), arXiv:2107.12787 [astro-ph.CO].

[31] B. P. Abbott et al. (KAGRA, LIGO Scientific, VIRGO), Living Rev. Rel. 21, 3 (2018), arXiv:1304.0670 [gr-qc].

[32] L. Barsotti, P. Fritschel, M. Evans, and S. Gras, Updated Advanced LIGO sensitivity design curve, Tech. Rep. LIGO-T1800044 (LIGO Laboratory, 2018).

[33] T. L. S. Collaboration, Classical and Quantum Gravity 32, 074001 (2015).

[34] M. Tse, H. Yu, N. Kijbunchoo, A. Fernandez-Galiana, P. Dupej, L. Barsotti, C. D. Blair, D. D. Brown, S. E. Dwyer, A. Effler, M. Evans, P. Fritschel, V. V. Frolov, A. C. Green, G. L. Mansell, F. Matichard, N. Mavalvala, D. E. McClelland, L. McCuller, T. McRae, J. Miller, A. Mullavey, E. Oelker, I. Y. Phinney, D. Sigg, B. J. J. Slagmolen, T. Vo, R. L. Ward, C. Whittle, R. Abbott, C. Adams, R. X. Adhikari, A. Amavasay, S. Appert, K. Arai, J. S. Areeda, Y. Asali, S. M. Aston, C. Austin, A. M. Baer, M. Ball, S. W. Ballmer, S. Banagiri, D. Barker, J. Bartlett, B. K. Berger, J. Betzwieser, D. Bhattacharjee, G. Billingsley, S. Biscans, R. M. Blair, N. Bode, P. Booker, R. Bork, A. Bramley, A. F. Brooks, A. Buikema, C. Cahillane, K. C. Cannon, X. Chen, A. A. Ciobanu, F. Clara, S. J. Cooper, K. R. Corley, S. T. Countryman, P. B. Covas, D. C. Coyne, L. E. H. Datrier, D. Davis, C. Di Fronzo, J. C. Driggers, T. Etzel, T. M. Evans, J. Feicht, P. Fulda, M. Fyffe, J. A. Giaime, K. D. Giardina, P. Godwin, E. Goetz, S. Gras, C. Gray, R. Gray, A. Gupta, E. K. Gustafson, R. Gustafson, J. Hanks, J. Hanson, T. Hardwick, R. K. Haskew, M. C. Heintze, A. F. Helmling-Cornell, N. A. Holland, J. D. Jones, S. Kandhasamy, S. Karki, M. Kasprzack, K. Kawabe, P. J. King, J. S. Kissel, R. Kumar, M. Landry, B. B. Lane, B. Lantz, M. Laxen, Y. K. Lee, E. Logue, J. Leviton, J. Liu, M. Lormand, A. P. Lundgren, R. Macas, M. MacInnis, D. M. Macleod, S. Márka, Z. Márka, D. V. Martynov, K. Mason, T. J. Massinger, R. McCarthy, S. McCormick, J. McIver, G. Mendell, K. Merfeld, E. L. Merilh, F. Meylahn, T. Mistry, R. Mittelman, G. Moreno, C. M. Mowlowry, S. Mozzon, T. J. N. Nelson, P. Nguyen, L. K. Nut-
miscalibrated, PhysiCal (green) vs. Spline (pink) runs, the vertical dashed lines mark the 25%, 50% and 75% percentiles.

\[ D_L/Mpc \]

\[ #1 \]
\[ #2 \]
\[ #3 \]
\[ #4 \]
\[ #5 \]
\[ #6 \]

Calibration Error Realization

True value
physiCal
Spline

FIG. 11: \( D_L \) likelihoods for the six scenarios, miscalibrated, PhysiCal (green) vs. Spline (pink) runs, the vertical dashed lines mark the 25%, 50% and 75% percentiles.
