

# Accounting for the known unknowns: A parametric framework to incorporate systematic waveform errors in gravitational-wave parameter estimation

Sumit Kumar<sup>1,2,3,4,\*</sup>, Max Melching<sup>1,2,†</sup> and Frank Ohme<sup>1,2,‡</sup>

<sup>1</sup>*Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), D-30167 Hannover, Germany*

<sup>2</sup>*Leibniz Universität Hannover, D-30167 Hannover, Germany*

<sup>3</sup>*Institute for Gravitational and Subatomic Physics (GRASP), Utrecht University, Princetonplein 1, 3584 CC Utrecht, The Netherlands*

<sup>4</sup>*Nikhef—National Institute for Subatomic Physics, Science Park 105, 1098 XG Amsterdam, The Netherlands*



(Received 6 March 2025; accepted 31 October 2025; published 24 November 2025)

The parameter estimation (PE) for gravitational wave (GW) merger events relies on a waveform model calibrated using numerical simulations. Within the Bayesian framework, this waveform model represents the GW signal produced during the merger and is crucial for estimating the likelihood function. However, these waveform models may possess systematic errors that can differ across the parameter space. Addressing these errors in the current data analysis pipeline is an active area of research. We introduce parametrizations for the uncertainties in the amplitude and phase of the reference waveform model. When the error budget in the amplitude and phase of the waveform model, as a function of frequency, is known, it can be used as a prior distribution in the Bayesian framework. We also show that conservative priors can be used to quantify uncertainties in waveform modeling without any knowledge of waveform uncertainty error budgets. Through zero-noise injections and PE recoveries, we demonstrate that even 1%–2% of errors in relative phase to the actual waveform model, for a GW150914-like signal and advanced LIGO detector sensitivity, can introduce biases in the recovered parameters. These biases can be corrected when we account for waveform uncertainties within the PE framework. By analyzing a series of simulated signals from mergers with precessing orbits and recovering them using a nonspinning waveform model, we demonstrate that we can reduce the ratio of systematic errors to statistical errors. This approach allows us to address scenarios where specific physical effects are missing in waveform modeling. The code that implements our parametrization for performing PE is available as a Python package “pycbc\_wfererrors\_plugin”, compatible with the PyCBC open source GW analysis library.

DOI: [10.1103/nkq9-zh2y](https://doi.org/10.1103/nkq9-zh2y)

## I. INTRODUCTION

GW signals from compact binary merger events provide us with a unique opportunity to probe these highly energetic events. With the network of advanced GW detectors, including the LIGO detectors in Hanford and Livingston, USA, the Virgo detector in Italy, and the KAGRA detector in Japan, the detection of GW signals

has become routine [1–5]. Recent catalogs of GW mergers, compiled from the data analysis of the first three observation runs of the LIGO and Virgo detectors, compile over 90 merger events. These catalogs include the analysis from the LIGO-Virgo-KAGRA (LVK) Collaboration [6], as well as analyses conducted by other independent groups [7,8]. The emphasis of the GW community has now expanded from detection to precision science, such as constraining the theory of general relativity [9], inferring the astrophysical distribution of the population of compact binary mergers [7,10], inferring Hubble constant [11], constraining the equation of state of a neutron star [12–14], etc.

The detection of GWs and the precision of measurements of source properties relies on the accurate modeling of the signals from the merger of compact binary systems such as binary black holes (BBH), binary neutron stars (BNS), and neutron star-black hole (NSBH). These waveform models have been developed over the years with the help of

\*Contact author: [sumit.kumar@aei.mpg.de](mailto:sumit.kumar@aei.mpg.de)

†Contact author: [max.melching@aei.mpg.de](mailto:max.melching@aei.mpg.de)

‡Contact author: [frank.ohme@aei.mpg.de](mailto:frank.ohme@aei.mpg.de)

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Open access publication funded by the Max Planck Society.*

post-Newtonian (PN) analytical calculation for the early inspiral part of the signals [15], and they rely on calibration from numerical relativity (NR) simulations for merger and ringdown regimes [16–18]. The signal’s loudness and detector sensitivity determine the width of statistical uncertainties in the measurement, while the inaccuracies in the waveform models could lead to systematic errors. The statistical uncertainties go down as the detectors become more and more sensitive while systematic errors remain the same.

In the coming years, the planned upgrades to existing detectors [3,19] and the addition of new ones, such as LIGO India [20], are expected to enhance the sensitivity of the detector network. Efforts are currently underway to develop third-generation (3G) detectors, such as the Einstein Telescope (ET) [21,22] and Cosmic Explorer (CE) [23–25], over the next decade. These 3G detectors are anticipated to be significantly more sensitive—by an order of magnitude—than the current generation of detectors, and they are expected to enhance low-frequency sensitivity. As the GW detectors’ sensitivities keep improving, the waveform modeling community tries to keep up with the required accuracy to provide unbiased estimates.

One important question is how accurate waveform models need to be. One may not want to burden the waveform modeling and NR simulation community with generating waveform models much better than the analysis can resolve [26]. However, studies suggest that we are already reaching the point where we can not ignore the systematic bias in a fraction of loud events. In a recent study, [27], it is shown that the typical biases in the inferred source parameters will be roughly equivalent to the standard deviation for the design sensitivity of current-generation detectors, and these biases could increase significantly for third-generation (3G) detectors. While a small bias in the fraction of events may be irrelevant for some studies, it could be critical for those that depend on accurate estimates of source properties. This is particularly true for inferring population characteristics of BBH mergers [28], estimating the Hubble constant using localization volume, issuing premerger localization alerts [29], and conducting tests of general relativity [30,31], among others.

The impacts of the waveform modeling accuracy on estimating the source parameters have been extensively studied [26–28,32]. The standard approach to account for waveform systematics is combining the posterior samples from different waveform models by assigning appropriate weights [33–35] or using hyperparameters to sample over different waveform approximants at the likelihood evaluation level [36–41]. These approaches are expected to marginalize potential differences between different waveform approximants. However, if all the waveform models have similar systematic effects, their difference would not accurately represent the true systematics in each model. Another approach is to marginalize the uncertainties using prior distributions analytically. These priors can be constructed

using the Gaussian processes regression technique from a training set of accurate templates [42]. Similar techniques can also be applied to address the waveform systematics arising from calibration with NR simulations [43,44]. There are also approaches to provide probabilistic models for the waveform, which can be used in sampling with appropriate weights [45]. The authors in [46] marginalize over higher order PN terms to mitigate the systematic errors.

This work presents a general framework that accounts for the uncertainties in waveform modeling through parametric models. The additional parameters in PE analysis relate to the waveform’s amplitude and phase errors. When we know the expected amplitude and phase error distribution in the waveform as a function of frequency, we can use them as priors in PE analysis. These errors are expected to be a function of parameter space. In the absence of such knowledge, we can be agnostic and use wider priors to capture any possible deviation. We evaluate this framework by introducing arbitrary deviations into reference waveform models and recovering them through our data analysis pipeline. In addition to correcting for biases, we assess the method’s ability to quantify the nature of deviations from the true signal, particularly for the higher signal-to-noise ratio (SNR) “golden” binary systems.

This paper is structured as follows: Sec. II A reviews current techniques used in waveform model developments and the potential sources of systematic biases they might contain. Section II B visits the state-of-the-art PE techniques and also discusses current methods to deal with waveform systematics. Section III discusses modeling GW strain data from the detector and introduces the parametrizations that we use in this work to account for waveform errors in PE. The following section (Sec. IV) discusses the Fisher matrix approach to account for the systematic bias. We also discuss the limitations where the Fisher matrix approach is no longer valid, and we need to do a full PE analysis. In Sec. V, we present the detailed simulations followed by an injection-recovery campaign to test the validity of this method. We also discuss the scenario where we can use these parametric models to account for the missing physical effects not considered in the waveform model. In Sec. VI, we compare our method to the other approaches in the literature. In the end, we discuss the findings of the work and summarize them in Sec. VII. If the reader is already familiar with the PE techniques and the potential sources of errors, they may skip Sec. II B. Additionally, if the reader is acquainted with the Fisher matrix approach for correcting systematic biases and understands its limitations in specific scenarios, they can skip Sec. IV. The reader who is just interested in the methodology and results can read Secs. III, V, and VII in that order.

## II. TECHNICAL BACKGROUND

### A. Waveform modeling and sources of uncertainties

The GW signal of compact binary mergers can be predicted by solving Einstein’s equations for the two-body

problem in the general theory of relativity. However, due to the complexity of these equations, some form of approximations need to be employed to model binary systems. The most sophisticated models combine several modeling techniques. A review of those techniques and models is beyond the scope of this paper, but we give a brief overview of aspects that are relevant to our discussion of systematic waveform errors. Here, we focus on black hole binaries on quasicircular orbits, as they are the most frequent source of detected GWs. However, many considerations detailed below apply to more general binary sources as well.

A black hole binary and its emitted GW signal are described by several parameters, such as the BH component masses ( $m_1, m_2$ ), their spins ( $\vec{s}_1, \vec{s}_2$ ), luminosity distance ( $D_L$ ), inclination of binary plane with the line of sight ( $\iota$ ), polarization angle with respect to the detector ( $\psi$ ), coalescence phase of the binary ( $\phi$ ), right ascension (RA), and declination angle (dec) in the sky, and the coalescence time ( $t_c$ ).

When the binary constituents are widely separated and their velocities are small (compared to the speed of light), the system's evolution can be approximately modeled by a power series expansion. The most commonly used approach is the PN formalism that expands in the system's relative velocity (see [15] for a review). Alternative approaches exist, for example the post-Minkowskian formalism expands in the gravitational constant  $G$  [47,48]. In the highly relativistic regime, where the spacetime curvature is large and velocities are high, these approaches become inaccurate. More concretely, PN expansions are well suited to describe the early inspiral of a black hole binary, but as the black holes approach each other and their velocities increase, purely PN models become inaccurate and unable to describe the merger and postmerger stages.

While the postmerger ringdown of the remnant black hole can be well described by perturbation theory [49,50], the merger itself can only be modeled by numerically solving Einstein's equation. Since its breakthrough [16–18], NR has become an important, well established tool in GW astronomy. However, because of its computational complexity, NR simulation can only cover the late inspiral (typically, a few tens of orbits), merger, and ringdown of the binary coalescence.

Neither of the above mentioned approaches can model the entire GW signal of interest. PN models become inaccurate towards the merger and do not include any merger or ringdown portion. NR is computationally too expensive and cannot cover hundreds or even thousands of orbits before merger. Therefore, several model approaches have been designed to bridge the gap between numerical and analytical methods [51].

Three model families are regularly employed in the analysis of GW observations [33]. The phenomenological models [52–67] are based on a set of hybrid waveforms that smoothly connect analytically derived inspiral signals with

NR data. This dataset is then described by PN inspired phenomenological formulas with tuneable coefficients that fit across the binary's parameter space. The resulting model is a set of closed-form expressions that describe the amplitude and phase of the GW signal.

Another approach is based on the effective-one-body (EOB) formalism [68], in which the binary system is mapped to a Hamiltonian formulation of a particle orbiting in an effective potential. By introducing higher-order terms and tuneable coefficients that are fit to waveforms from NR simulations, the resulting EOBNR waveforms can describe the binary coalescence accurately up to the merger. Completed by a suitable ringdown attachment, several models of the EOBNR family have been developed [69–79] and applied to the analysis of GW events.

The third model family is called surrogate models. These models are built by decomposing waveform data from NR simulations into a suitable basis and interpolating between them [80]. NR surrogate models [81–84] are very faithful to the NR simulations they were built from, but they are limited by the length and parameter coverage of NR simulations.

All of the modeling approaches approximate the exact solutions of Einstein's equation. Therefore, they carry systematic uncertainties that are difficult to quantify in their entirety. Numerical simulations start with approximate initial data and discretize the spacetime. The resulting resolution error is often estimated by providing simulation results of the same system for different resolutions. All modeling approaches described above rely on input from numerical simulations, but they also need to interpolate between different simulations to cover the entire parameter space of interest. This interpolation introduces additional inaccuracies that are often less well quantified. Part of the problem is not only the interpolation (or even extrapolation) of existing NR data. The ansatz used to describe the data carries its own limitations, whether it is data driven (such as Gaussian process regression) or analytically motivated.

In summary, all modeling approaches have to balance representing known NR simulation data as accurately as possible while ensuring a smooth and robust interpolation across the parameter space. The techniques employed to accomplish this have advanced over the years, but they cannot be perfect. Therefore, any model's prediction has to be treated as a “best-guess” waveform for each set of parameters. While the faithfulness to NR data is often very impressive, one must not forget the uncertainties arising from each step of the modeling process.

Here, we discuss and test a framework that moves away from the “best-guess” paradigm to accounting for uncertain signal predictions in parameter-estimation analyses.

## B. Parameter estimation

In this section, we review the PE techniques for GW data analysis and potential errors in estimating source

properties. Readers who are familiar with this can skip to the next section. Time series strain data  $s(t)$  in each detector consists of noise  $n(t)$  and may contain a transient signal  $h(t)$ . It is assumed that when the signal is present, the noise and signal are additive to give the strain data, i.e.,

$$s(t) = h(t) + n(t). \quad (1)$$

The noise from the detector is modeled as Gaussian and stationary. In the absence of the signal, the detectors are assumed to contain only Gaussian and stationary noise. The Gaussian and stationary time series noise  $n(t)$  can be expressed as [85]

$$\mathbf{n} = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left\{\frac{1}{2}(n(t_i) - \hat{\mu})^T \Sigma_{ij}^{-1} (n(t_j) - \hat{\mu})\right\}, \quad (2)$$

where  $\mathbf{n}$  is a noise realization which is represented as a vector with discrete time samples  $n(t_i)$ ,  $\Sigma$  is a covariance matrix, and  $\Sigma_{ij}$  represent covariance between  $i$  and  $j$  time bins, and  $\mu$  is the expectation value of noise  $\mathbf{n}$ . The GW signal buried in strain data are modeled in terms of the parameters  $\vec{\Theta}$  described in Sec. II A as  $h(t; \Theta)$ . Under these assumptions, search pipelines use matched filter techniques to estimate the SNR as

$$\rho^2 = 4 \int_{f_{\text{low}}}^{f_{\text{high}}} \frac{|\tilde{s}(f)\tilde{h}(f, \Theta)|}{S_n(f)} df, \quad (3)$$

where  $\tilde{s}(f)$  and  $\tilde{h}(f)$  are the functions of frequency and are the Fourier transforms of the time domain quantities  $s(t)$  and  $h(t)$ , respectively.  $S_n(f)$  is the power spectral density, which quantifies the noise properties of the GW detector.  $f_{\text{low}}$  and  $f_{\text{high}}$  are the low and high-frequency cutoffs for the integration limit. A template bank-based approach is employed in data analysis pipelines for GW searches [86–88]. The SNR time series is evaluated for each point in the template bank. When the SNR surpasses a certain threshold, it is deemed a trigger and is saved for further analysis. Triggers from individual detectors that occur within the light travel time between the detectors are known as coincident triggers. In the next stage of GW searches, these coincident triggers are further examined, and their significance is estimated by comparing them with the expected distribution of background triggers. Statistically significant triggers are classified as GW merger events [85,87]. A more detailed analysis is done using PE techniques to determine the waveform model's parameters  $\vec{\Theta}$  [89,90].

In order to estimate the parameters of a GW merger signal, the standard method is to use a Bayesian framework. For the given data  $d$ , the likelihood function  $\mathcal{L}(d|\vec{\Theta}, I)$  describes the probability of obtaining the data for a given model with any other prior information  $I$ . The posterior

probability distribution  $p(\vec{\Theta}|d, I)$  is then calculated using Bayes' theorem,

$$P(\vec{\Theta}|d, I) = \frac{\mathcal{L}(d|\vec{\Theta}, I)\pi(\vec{\Theta}|I)}{p(d|I)}, \quad (4)$$

where  $\pi(\vec{\Theta}|I)$  represents the prior probability distribution for parameters  $\vec{\Theta}$  and  $p(d|I)$  describes the marginalised likelihood.

### C. Sources of errors in the parameter estimation

There can be systematic errors present in the PE analysis due to any of the following reasons:

- (i) *Data analysis artifacts due to mismodeling the noise:* For certain periods of strain data, the noise in the GW detector may not adhere to the assumptions of Gaussianity and stationarity. This deviation can occur due to environmental factors or issues with the instrument itself. In a study, Mozzon *et al.* [91] found that approximately 3% of the GW detector data does not conform to the assumptions of Gaussianity and stationarity. This represents a situation where the noise model fails. If left unaddressed, this could result in biased estimates of the properties of the binary source (systematic errors) or inaccurate statistical uncertainties (statistical errors) [29,92].
- (ii) *Waveform systematics:* With the improvements in the sensitivity of the GW detectors and with the addition of more physics in the description of waveform from the GW merger, such as the inclusion of higher modes, eccentricity, precision, the requirements for the accuracies of the waveform models are becoming stricter. We have already reached an era where the systematics of the waveform models can not be ignored. If the waveform model does not describe the underlying reality, and if we do not account for the possible errors, the estimates of source properties are expected to be biased.
- (iii) *Other unaccounted effects:* Even if our models for noise and the waveform describe the strain data perfectly, there can be rare cases such as modification of GW waveform by strong lensing by intermediate-mass black hole object or presence of subthreshold overlapping signals, etc. It can give rise to unmodeled effects, which can again bias the PE results.

It is crucial to consider any systematic errors that may arise from the above mentioned effects. This work focuses on the systematic and statistical errors arising from inaccuracies in waveform models. In PE, the systematic errors can be represented by  $|\Delta\Theta| = |\Theta_{\text{rec}} - \Theta_{\text{true}}|$ , where  $\Theta_{\text{true}}$  represents the true value of waveform parameters and  $\Theta_{\text{rec}}$  represents the recovered value of the parameters. For

individual waveform parameters, we define  $\Delta\Theta^i = |\Theta_{\text{rec}}^i - \Theta_{\text{true}}^i|$  where  $\Theta_{\text{rec}}^i$  is the median value of one dimensional marginalized posterior samples for  $i$ th waveform parameter. Similarly, the statistical error  $\sigma_i$  represents the standard deviation of marginalized posterior distribution of  $i$ th waveform parameter. For a given waveform model, the statistical errors depend on the signal loudness and detector sensitivity. We can ignore systematic errors when

$$|\Delta\Theta^i| \ll \sigma_i \quad \forall i. \quad (5)$$

Conversely, the systematic errors become important when

$$\frac{|\Delta\Theta^i|}{\sigma_i} \sim \mathcal{O}(1) \quad \text{for any } i. \quad (6)$$

### III. MODELING UNCERTAINTIES IN STRAIN

The description of transient signal  $h(t, \vec{\Theta})$  can be affected by following:

- (i) *Calibration uncertainties*: The GW detectors are calibrated periodically to map between input strain and time series output. GW detectors are not perfectly calibrated, and the calibration uncertainties are provided with detector characterization [93].
- (ii) *Inaccuracies in waveform modeling*: These inaccuracies might arise due to one of many reasons, such as (i) errors in calibrating to NR waveforms, (ii) not accounting for all the physics in the waveform modeling, (iii) inherent errors in NR simulations.

To model general uncertainties in the observed strain data, we go to the frequency domain where the signal,  $\tilde{h}_{\text{obs}}(f)$ , can be modeled as

$$\tilde{h}_{\text{obs}}(f) = \tilde{A}_{\text{obs}}(f) \exp\{i\tilde{\phi}_{\text{obs}}(f)\}, \quad (7)$$

where  $A_{\text{obs}}(f)$  and  $\phi_{\text{obs}}(f)$  are observed amplitude and phase of the signal. From now on, we will drop writing the explicit frequency dependence. Whenever we present a quantity in the frequency domain, with overhead  $\sim$ , it is assumed that it is generally an explicit function of frequency unless stated otherwise. Now we assume that the observed signal  $\tilde{h}_{\text{obs}}$  can be expressed in terms of “true” amplitude and phase with linear perturbation,

$$\begin{aligned} \tilde{h}_{\text{obs}} &= \tilde{A}_{\text{obs}} \exp\{i\tilde{\phi}_{\text{obs}}\} \\ &= (\tilde{A}_{\text{true}} + \delta\tilde{A}_{\text{abs}}) \exp\{i(\tilde{\phi}_{\text{true}} + \delta\tilde{\phi}_{\text{abs}})\} \end{aligned} \quad (8)$$

$$= \tilde{A}_{\text{true}}(1 + \delta\tilde{A}_{\text{rel}}) \exp\{i\tilde{\phi}_{\text{true}}(1 + \delta\tilde{\phi}_{\text{rel}})\}, \quad (9)$$

where  $\tilde{A}_{\text{true}}$  and  $\tilde{\phi}_{\text{true}}$  are true amplitude and phase of the signal.  $\delta\tilde{A}_{\text{abs}}$  and  $\delta\tilde{\phi}_{\text{abs}}$  are absolute errors in the true amplitude and phase, respectively, while  $\delta\tilde{A}_{\text{rel}}$  and  $\delta\tilde{\phi}_{\text{rel}}$  are relative deviation in true amplitude and phase respectively. The relation between absolute and relative deviation (or errors) are  $\delta\tilde{A}_{\text{rel}} = \frac{\delta\tilde{A}_{\text{abs}}}{\tilde{A}_{\text{true}}}$  and  $\delta\tilde{\phi}_{\text{rel}} = \frac{\delta\tilde{\phi}_{\text{abs}}}{\tilde{\phi}_{\text{true}}}$ .

#### A. Detector calibration uncertainties

When GW passes through the detector, it produces a differential arm displacement between two L-shaped arms relative to the original arm lengths. This differential arm displacement generates laser power fluctuations in the photodetector, reproduced as the time series data. The conversion to the digital strain data regarding time series from the differential arm displacement happens through the control loop and calibration pipeline [94,95]. The calibration of the strain data are modeled with systematic and statistical errors, which are then provided as calibration uncertainties as function of frequency [94–98]. Figure 1

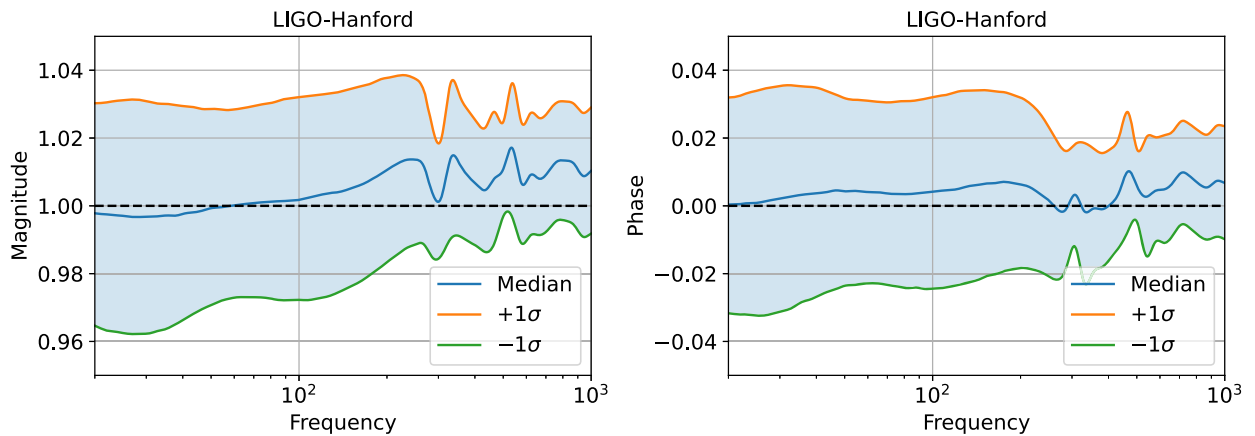


FIG. 1. The figure shows the calibration uncertainties as a function of frequency for the LIGO Hanford detectors around the time of the first GW detection: GW150914. The left panel shows the systematic error in the magnitude, while the right panel shows the systematic error in the phase angle (in radians). The shaded region represents the  $\pm 1\sigma$  band around the median systematic error. The systematic error is described by the ratio of the true response function to the modeled response function of the detector. The LIGO and Virgo calibration uncertainty files for O1, O2, and O3 observation runs are available at [93].

shows the calibration envelopes for the LIGO-Hanford detector around the event GW150914.

The calibration uncertainties are modeled as frequency dependent errors in phase and amplitude as follows [99,100]:

$$\tilde{h}_{\text{obs}} = \tilde{h}_{\text{true}}(1 + \delta\tilde{A}_{\text{cal}})\exp(i\delta\tilde{\phi}_{\text{cal}}), \quad (10)$$

In terms of the general parametrization we discussed above,  $\delta\tilde{A}_{\text{cal}}$  can be identified as relative error in amplitude while  $\delta\tilde{\phi}_{\text{cal}}$  is absolute error in the phase. In the calibration uncertainty model, Eq. (10), the exponential term can be expanded as a Taylor series and is rewritten as

$$\tilde{h}_{\text{obs}} = \tilde{h}_{\text{true}}(1 + \delta\tilde{A}_{\text{cal}})\frac{2 + i\delta\tilde{\psi}_{\text{cal}}}{2 - i\delta\tilde{\psi}_{\text{cal}}}, \quad (11)$$

where the ratio involving  $\delta\tilde{\psi}$  is designed to have a modulus of unity, thus being a pure phase [99]. In existing PE pipelines such as PYCBC-Inference [90] and BILBY [101], the projected strain in each detector is modeled with phase and amplitude correction using cubic-spline curves drawn from the distribution provided by calibration envelopes, as shown in Fig. 1.

## B. GW waveform model uncertainties

Motivated by how detector calibration uncertainties are handled, we suggest that uncertainties in waveform models can be treated similarly. Specifically, we propose parametrizing the uncertainties in the description of GW waveform in the frequency domain. Furthermore, we suggest that waveform developers provide uncertainty bands with their model. A similar approach was proposed in [32]. Here, we examine the technical details, feasibility, choices of parametrization, and degeneracy with the detector calibration-correction framework. We use the following parametrizations to account for waveform modeling errors:

- (i) *abs-phase*: Relative errors in amplitude and absolute errors in phase,

$$\begin{aligned} \tilde{h}_{\text{model}} &= \tilde{A}_0(1 + \delta\tilde{A})\exp\{i(\tilde{\phi}_0 + \delta\tilde{\phi})\}, \\ &= \tilde{h}_0(1 + \delta\tilde{A})\exp\{i\delta\tilde{\phi}\}, \end{aligned} \quad (12)$$

- (ii) *rel-phase*: Relative errors in amplitude and phase,

$$\begin{aligned} \tilde{h}_{\text{model}} &= \tilde{A}_0(1 + \delta\tilde{A})\exp\{i\tilde{\phi}_0(1 + \delta\tilde{\phi}_{\text{rel}})\}, \\ &= \tilde{h}_0(1 + \delta\tilde{A})\exp\{i\tilde{\phi}_0\delta\tilde{\phi}_{\text{rel}}\}, \end{aligned} \quad (13)$$

where  $\tilde{h}_0 = \tilde{A}_0 \exp(i\tilde{\phi}_0)$  represent the baseline (or reference) model, which aspires to be a “true” model. As no waveform model is perfect to arbitrary accuracy, we introduce additional terms to account for potential systematics.  $\delta\tilde{A}$  represent the fractional or relative error in the

amplitude of the signal, and  $\delta\tilde{\phi}$  ( $\delta\tilde{\phi}_{\text{rel}}$ ) represent absolute (relative/fractional) errors in phase.

The abs-phase parametrization described by Eq. (12) is also used in the correction for the detector calibration. Other studies also use this parametrization to quantify the waveform uncertainties [32,102]. Since  $\delta\tilde{A}$  is the relative change in amplitude, it is a dimensionless parameter in these parametrizations. The units of  $\delta\tilde{\phi}$  is radian in abs-phase parametrization while  $\delta\tilde{\phi}_{\text{rel}}$  is a dimensionless parameter.

A point of caution must be mentioned here for description of the relative phase error. The baseline phase  $\tilde{\phi}_0$  is only uniquely defined up to integer multiples of  $2\pi$ . However, the multiplication with  $\delta\tilde{\phi}_{\text{rel}}$  would give the phase error a different physical meaning depending on the chosen branch of the baseline phase. Therefore, a convention must be introduced to remove this ambiguity. We implement a simple method that ensures the baseline phase to be between  $-\pi$  and  $\pi$  at the starting frequency of the waveform generation. The rest of the phase evolution follows a standard unwrapping algorithm as implemented in NUMPY [103].

We note that alternative choices for removing this ambiguity exist. For example, one could parametrize the phase by an exact value  $\tilde{\phi}_{\text{ref}}$  at a given reference frequency  $f_{\text{ref}}$ . By this convention, the error would vanish at the reference frequency, and one could explicitly set

$$\tilde{\phi}_0(f)\delta\tilde{\phi}_{\text{rel}}(f) \mapsto [\tilde{\phi}_0(f) - \tilde{\phi}_0(f_{\text{ref}})]\delta\tilde{\phi}_{\text{rel}}(f) + \tilde{\phi}_{\text{ref}}. \quad (14)$$

In this formulation, phase errors would accumulate as one moves away (both in the positive and negative  $f$  direction) from  $f_{\text{ref}}$ .

From now on, we will drop the subscript “rel” from the  $\delta\tilde{\phi}$  parameter, and we will refer these parameters as the pair  $(\delta\tilde{A}, \delta\tilde{\phi})$ , which in general are functions of frequency. These parameters will represent either the abs-phase or rel-phase model, depending on the context of which parametrization is being discussed. We refer to these models as waveform error parametrization models or WF-error parametrization for short.

Following the calibration uncertainty framework, we use  $N$  nodal points (or knots) between the frequency interval  $f \in [f_{\text{min}}, f_{\text{max}}]$ . Here,  $f_{\text{min}}$  serves as the low-frequency cutoff, or the frequency for the first nodal point, while  $f_{\text{max}}$  represents the maximum frequency of our analysis which can be a frequency immediately following the merger. In this way, we can account for waveform errors throughout the signal in the detector band. The values of the WF-error parametrization parameters at the waveform nodal points can be utilized for cubic spline interpolation. This method employs piecewise polynomial curves that are cubic in order and have continuous second derivatives. We utilize the CubicSpline implementation from the publicly available SCIPY package [104] for our WF-error

parametrizations. Using  $f_i \in [f_1, f_2, \dots, f_N]$ , which are the nodes of the polynomial in the frequency, we construct  $\delta\tilde{A}(f)$  and  $\delta\tilde{\phi}(f)$  curves using

$$\delta\tilde{A}(f) = P_s^3(f; \{f_i, \delta\tilde{A}_i\}) \quad (15)$$

$$\delta\tilde{\phi}(f) = P_s^3(f; \{f_i, \delta\tilde{\phi}_i\}), \quad (16)$$

where  $P_s^3$  is the cubic spline polynomial. In the PE analysis, we provide the priors for  $\delta\tilde{A}_i$  and  $\delta\tilde{\phi}_i$  at each frequency node. Each realization of  $(\delta\tilde{A}_i, \delta\tilde{\phi}_i)$  points corresponds to the curves  $\delta\tilde{A}(f)$ , and  $\delta\tilde{\phi}(f)$ , which is used to modify the reference waveform model  $\tilde{h}_{\text{ref}}(f)$  in accordance to Eqs. (12) and (13).

We now turn to the question why we introduce two different phase parametrizations. The SNR (3) and the inner product (A2) used in PE obey a symmetry between the observed data and the GW model. Therefore, if one accounts for calibration uncertainties in the data, this should have the same effect as accounting for waveform systematics. Since the calibration uncertainties are modeled with absolute errors in phase, marginalizing over calibration uncertainties can account for waveform systematics if they are of the same order. However, it is worth pointing out that all GW transient catalogs (GWTCs) included the effect of uncertain detector calibration in the analysis of GW events. Despite marginalizing over amplitude and (absolute) phase uncertainty for each detector, the properties of some binaries showed significant systematic differences between the waveform models that have been employed for

the analysis (e.g., GW190412 [105], GW191109\_010717, GW191219\_163120, GW200129\_065458; see Sec. III E in [33]). Evidently, marginalizing over calibration uncertainties did not incorporate all waveform systematics.

This is not unexpected. Systematic uncertainties in waveform models can often appear as secular phase drifts caused by numerical errors, (unavoidably) incomplete series expansions, interpolation inaccuracies, or even missing physics. The resulting waveform errors can, therefore, become much more severe than random small phase variations. If one wants to allow for possibly significant waveform uncertainties, large absolute phase errors must be considered, which may lead to an overly pessimistic error estimate. Alternatively, a small relative phase error is likely more appropriate to model systematic waveform uncertainties at the leading order.

We now turn to incorporating WF-error parametrizations in PE analysis. Ideally, a waveform model shall consist of the best-estimated reference model  $\tilde{h}_{\text{ref}}$ , and priors for the parameters  $\delta\tilde{A}$  and  $\delta\tilde{\phi}$  for rel-phase or abs-phase parametrization to account for the waveform systematics. We call them WF-error priors or error budget. The WF-error priors should generally depend on the parameter space. We expect these priors to be narrower in the regime where many NR simulations are available, and waveform models are calibrated to greater accuracy. For the regions where NR simulations are sparse and theoretical waveform description is not so accurate, we expect them to be broader. For a given signal in parameter space, we also expect these priors to be broader for the frequencies close to the merger, where the analytical

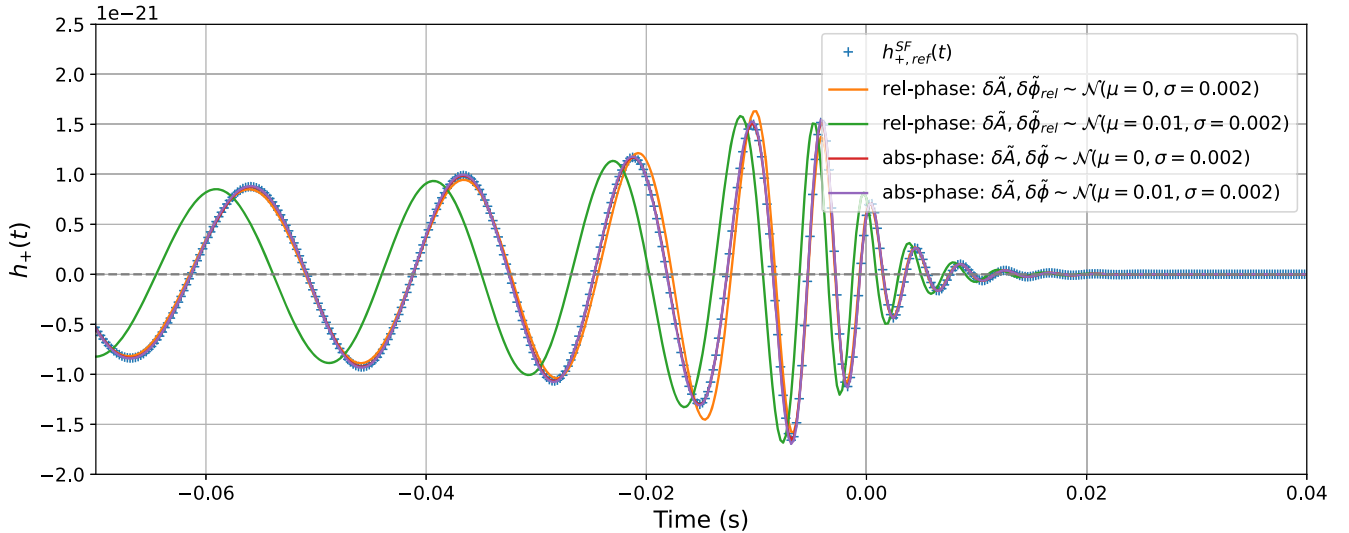


FIG. 2. We show the time domain waveform in a signal frame (before projecting it in the detector) for one of the polarization (+) for a GW150914-type signal. Apart from the reference waveform model IMRPhenomPv2  $h_{+,ref}^{SF}(t)$  (blue, denoted by “+” markers), we also show the modification to the reference signal by one of the parametrizations described in the text. Red and violet waveforms represent the abs-phase modification given by (12). We use cubic splines for the parameters  $(\delta\tilde{A}, \delta\tilde{\phi})$  with a realization from the normal distribution as shown in corresponding plot legends. Orange and green waveforms represent the modification with rel-phase parametrization given by Eq. (13).

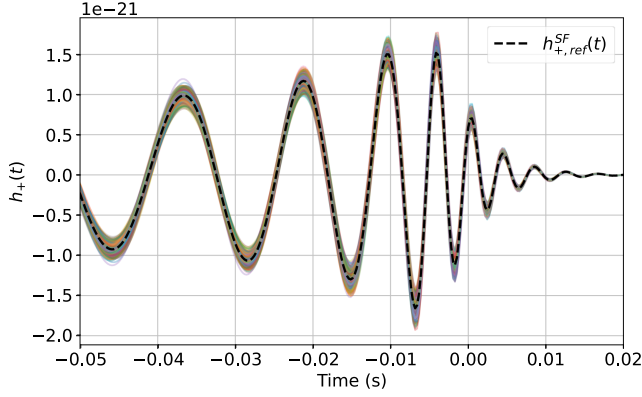


FIG. 3. We show the time domain waveform for (+) polarization for the GW150914 type signal. The reference waveform model IMRPhenomPv2:  $\tilde{h}_{+,ref}^{SF}(t)$  is shown as dashed black curve. Other curves are modified waveform with abs-phase modification, described by Eq. (12), with WF-error parameters sampled from the distribution  $\delta\tilde{A}, \delta\tilde{\phi} \sim \mathcal{N}(\mu = 0, \sigma = 0.05)$ .

models extending from the inspiral regime to the merger are calibrated with NR simulations.

In the absence of such priors from the waveform modeling side, we can use a conservative approach and use wide enough priors so that we expect to capture any systematic and use data to constrain the parameters related to waveform errors. It might lead to broader posterior distribution resulting in larger variance on signal parameters (especially for weak signals), but it is expected to be unbiased. We will test these assumptions in the Sec. V with simulated signals and recovery.

Figure 2 shows an example of waveform with GW150914 type signals and the modifications applied in the frequency domain using cubic splines. To generate the reference signal, we use the IMRPhenomPv2 waveform model with source frame masses  $(m_1^{\text{src}}, m_2^{\text{src}}) = (36M_\odot, 29M_\odot)$  at the luminosity distance of 500 Mpc. We used nonspinning injections for this example. For the modifications, we use both the parametrizations described previously with WF-error parameters chosen from a realization of the normal distribution described in the figure. For the same values of  $\delta\tilde{\phi}$ , the rel-phase parametrization significantly modifies the reference model compared to abs-phase parametrization. In Fig. 3, we show the reference waveform model for the GW150914 type signal described above along with the modified waveform curves with abs-phase modification. We use cubic splines with ten waveform nodal points in frequencies with values  $(\delta\tilde{A}, \delta\tilde{\phi})$  drawn from a normal distribution  $\mathcal{N}(\mu = 0, \sigma = 0.05)$ .

It is important to note that the WF-error parametrizations (abs-phase and rel-phase) can be transformed into one another using simple mathematical relations by evaluating the phase  $\tilde{\phi}_0$  of the reference waveform model  $\tilde{h}_{ref}$ . We can use one of the parametrizations and mathematical

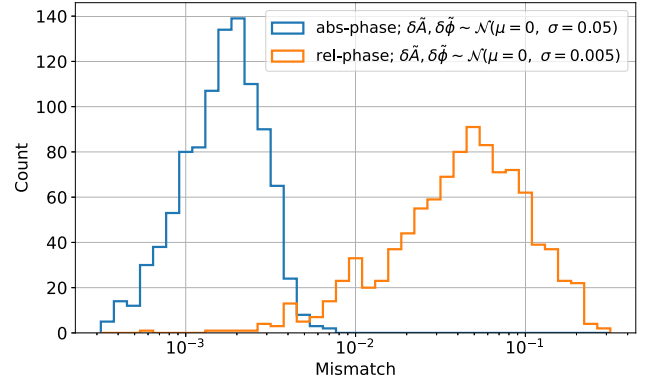


FIG. 4. On the x axis, we show the mismatch between the reference waveform model  $h_{ref}$  and the modified waveform model  $h_{mod}$ . The modified signal,  $h_{mod}$ , is generated by either abs-phase or rel-phase WF-error parametrization. The WF-error parameters are taken from the normal distribution, and we generate cubic spline curves to modify the reference model. The rel-phase model is generally well suited to model comparatively more significant mismatches.

transformations to convert it into the other according to the situation. However, using the correct parametrization is desirable for practical purposes in PE analysis. As an example, if the rel-phase parametrization is used for waveform systematics, and if one chooses abs-phase parametrization in PE, the priors for  $\delta\tilde{\phi}$  need to become wider for larger frequencies to account for the growing reference phase  $\tilde{\phi}_0$ .

We can use hints from the mismatch (see Appendix A for definition) studies to decide which WF-error parametrization is better suited for practical purposes. In Fig. 4, we show the distribution of mismatches between the reference waveform model and the modified one using both parametrizations. We use the Gaussian prior with zero mean and a fixed standard deviation across frequency range to generate the modified signal  $h_{mod}$ . As expected, even with the narrower Gaussian prior, the rel-phase parametrization produces significant deviation to the  $h_{ref}$  compared to the abs-phase model, hence more significant mismatches. When we are in the parameter space where larger mismatches are expected, we can set our PE analysis framework to use rel-phase parametrization.

### C. Degeneracy between detector calibration and waveform uncertainties

A comprehensive PE data analysis pipeline for PE should include (i) methods to incorporate uncertainties in waveform modeling, (ii) corrections due to detector calibration, and (iii) adjustments for possible deviations from Gaussian and stationary noise, whenever applicable. If we choose cubic-spline curves to mitigate waveform uncertainties and detector calibration uncertainties, a natural question arises: can these systematics be decoupled?

We will explore the specifics of this query in a future study. However, we would like to point out that there are fundamental differences in how these corrections are applied: While the calibration corrections are applied in the detector frame and are different for each detector, the WF-error parametrization corrections are applied before projecting the waveform into each detector. Therefore, the modification due to the calibration uncertainties and absolute phase waveform errors can be decoupled for a sufficiently loud signal.

#### IV. CAN FISHER MATRIX ESTIMATES BE USED FOR ACCOUNTING SYSTEMATIC BIASES?

A standard tool in the literature to quantify systematic biases between two waveform models is the Fisher matrix formalism [28,106]. In this section, we will apply it to see how significantly errors modeled as (13) change the recovery of source parameters  $\Theta^\mu$  and verify whether the formalism can account for the induced biases. If it can, this will help us predict the systematic parameter errors induced by externally provided estimates of waveform errors. Then we can assess the significance of these errors and determine if corrections are necessary. For the science case for next-generation GW detectors, many investigations focus on forecast studies using Fisher matrix formalism [107–109]. This section specifies the conditions under which corrections based on Fisher matrix analysis are applicable or not. Appendix A introduces the Fisher matrix formalism.

##### A. Setup of the analysis

In this part, we introduce methods to test how well the Fisher estimates work in trying to recover biases induced by certain choices of  $\delta\tilde{A}$ ,  $\delta\tilde{\phi}$ . We do that by injecting data generated from the signal model (13) for chosen values  $\delta\tilde{A}$ ,  $\delta\tilde{\phi}$ , and recovering using the baseline waveform model. The injection parameters chosen for these analyses are quoted in Table I. We focus on nonspinning binaries and keep the extrinsic parameters  $\psi$ ,  $\alpha$ ,  $\delta$  fixed, so in the end only  $\mathcal{M}^z$ ,  $q$ ,  $D_L$ ,  $\iota$  and time of coalescence  $t_0$ , global phase  $\phi_0$  are varied. Accordingly, the set of parameters constituting the Fisher matrices we calculate is

$$\{\mathcal{M}^z, q, D_L, \iota, t_c, \phi_0\}. \quad (17)$$

Please refer to Appendix A for details on the calculation.

There are two properties of the posteriors that we wish to compare to the respective Fisher matrix estimate:

- (1) First of all, the systematic error  $\Delta\Theta^\mu$ . From (A3), we see that it is formally defined as the difference between maximum *a posteriori* estimate  $\Theta_{\text{bf}}$  and injected parameters  $\Theta_{\text{tr}}$ .  
In practice, we follow the method chosen in the gravitational-wave transient catalogs [6,110–112] and use median values as proxies for the maximum *a posteriori* points.
- (2) Secondly, the posterior width. Ref. [113] shows that, in the high-SNR limit, it reads

$$\sigma_{\Theta^\mu} \simeq \sqrt{(\Gamma^{-1})^{\mu\mu}}, \quad (18)$$

and this number is compared to the standard deviation of the marginalized posterior for  $\Theta^\mu$ .

##### B. Results of comparing fisher matrix estimates with PE runs

In total, we have analyzed four scenarios: two different sets of error values of  $\delta\tilde{A} = 0.004$ ,  $\delta\tilde{\phi} = 0.004$  and  $\delta\tilde{A} = 0.01$ ,  $\delta\tilde{\phi} = 0.01$ , and for each of them, one moderate- and one high-SNR case (cf. Table I). All of the signals have been injected as a zero-noise injection into the Hanford detector with an O4 PSD. The results are summarized in Table II.

Since the chosen signal lies in the nonspinning part of the parameter space, the remaining intrinsic parameters are the masses parametrized by the tuple  $\mathcal{M}^z$ ,  $q$ . We will direct our attention toward these two, not discussing biases in the luminosity distance  $D_L$  and inclination  $\iota$  in detail. This is because the uncertainty in  $D_L$  is typically very large. Moreover, there is a known correlation between these two parameters, which makes their recovery and the assessment of their bias results more complicated. We have verified that this correlation has no significant impact on results for other parameters.

The first important finding is that the PE results are biased if a modified signal is recovered using the baseline, nonmodified model. This finding implies that the error model (13) is capable of parametrizing systematic errors in the way we intend it to. Moreover, we find that these biases can be recovered in the Fisher matrix framework, validating

TABLE I. Injection parameters for the base signal. In the high-SNR case, the luminosity distance is changed, now taking a value of 100 Mpc instead and accordingly, the SNR changes to 83.91. Note that the SNR values quoted here refer to the nonmodified injection, for the modified one the SNR will be slightly different (with deviations being on the order of 1%). We use priors that are uniform in component mass for  $\mathcal{M}^z$ ,  $q$ , uniform in  $\cos(\iota)$  for the inclination [i.e.,  $p(\iota) \propto \sin(\iota)$ ], uniform in comoving volume for the luminosity distance, and uniform in coalescence time  $t_c$  (no prior for  $\phi_0$  because we use a marginalized-phase likelihood).

|             | $m_1^z [M_\odot]$ | $m_2^z [M_\odot]$ | $D_L$ (Mpc) | $\iota$ (rad) | $f_{\text{ref}}$ (Hz) | $\text{ra}$ (rad) | $\text{dec}$ (rad) | $\psi$ (rad) | Detector | $t_{\text{gps}}$ [s] | SNR   |
|-------------|-------------------|-------------------|-------------|---------------|-----------------------|-------------------|--------------------|--------------|----------|----------------------|-------|
| Base signal | 20                | 10                | 500         | 0.2           | 20                    | 4.67              | 0.65               | 2.34         | Hanford  | 1192529720           | 16.78 |

TABLE II. Summary of results from full PE runs and Fisher matrix estimates. The numbers one should primarily compare the neighboring columns labeled “PE” and “Fisher”, for each of the four quantities that appear in the upper row. Moreover, when comparing them one should keep in mind that neither number is expected to be perfectly accurate, explaining certain small deviations; PE results are always subject to sampling uncertainties (which are noticeable, for instance, for  $\Delta q$ ), while limitations of the Fisher approach have been laid out at the beginning of this section and Appendix A.

| Configuration                                  |             | $\Delta\mathcal{M}^z [M_\odot]$ |        | $\sigma_{\mathcal{M}^z} [M_\odot]$ |        | $\Delta q$ |        | $\sigma_q$ |        |
|------------------------------------------------|-------------|---------------------------------|--------|------------------------------------|--------|------------|--------|------------|--------|
|                                                |             | PE                              | Fisher | PE                                 | Fisher | PE         | Fisher | PE         | Fisher |
| $\delta\tilde{A} = \delta\tilde{\phi} = 0.004$ | Base signal | −0.023                          | −0.026 | 0.040                              | 0.039  | −0.012     | 0.001  | 0.143      | 0.137  |
|                                                | High-SNR    | −0.028                          | −0.026 | 0.0077                             | 0.0078 | 0.007      | 0.001  | 0.0271     | 0.0273 |
| $\delta\tilde{A} = \delta\tilde{\phi} = 0.01$  | Base signal | −0.064                          | −0.04  | 0.040                              | 0.039  | −0.0048    | −0.057 | 0.143      | 0.137  |
|                                                | High-SNR    | −0.069                          | −0.039 | 0.0078                             | 0.0078 | 0.013      | −0.058 | 0.0275     | 0.0273 |

that the biases have a physical origin rather than being an artifact of flawed application of PE or an error in the implementation. The statistical bias is recovered very well in all four scenarios, with differences usually appearing in the second significant digit. This agreement is an interesting feature of the results because it indicates that the width estimate (A5) seems to be more robust to increasing waveform differences than the systematic bias estimate (A3) (at least for our applications), where only the first configuration  $\delta\tilde{A} = \delta\tilde{\phi} = 0.004$  shows consistency between Fisher matrix estimates and PE results.

Those results motivate the following interpretation: in case of  $\delta\tilde{A} = \delta\tilde{\phi} = 0.004$ , the waveform differences introduced into the waveform are still well approximated by the linear expansion (A6), the LSA is justified; this clearly changes for  $\delta\tilde{A} = \delta\tilde{\phi} = 0.01$ , where the systematic bias estimates begin to diverge from the corresponding PE results—the LSA is not a faithful approximation anymore. This divergence happens in various ways, with both magnitude and sign being sources of inconsistency.

We have seen that Fisher estimates can work, but not always. First of all, it was expected that the LSA will not hold up to arbitrarily large values of  $\delta\tilde{A}, \delta\tilde{\phi}$ , simply because we use the relative error model where errors accumulate from low to high frequencies and thus will quickly result in large waveform differences. However, this being the case for a 1% error and the simplest setup we can think of (constant shifts, same deviations in both polarizations) has serious implications for our objectives of a practical application of the Fisher estimates: for realistic scenarios, where errors of this magnitude are expected, they are not a suitable prospect to predict systematic errors in intrinsic parameters induced by  $\delta\tilde{A}, \delta\tilde{\phi}$ . Of course, in other parts of the parameter space, other thresholds for the LSA validity might exist, but the existence of such a counterexample in a part of the parameter that is certainly relevant for detected signals is already sufficient to recognize that we should not rely on a faithful estimation of biases. The numbers are particularly concerning for the high-SNR signal with  $\delta\tilde{A} = \delta\tilde{\phi} = 0.01$  since this corresponds to the data from

next-generation GW detectors, where a tenfold increase in sensitivity is aimed for. Here, Fisher estimate and true recovery are basically fully separate. All of this plays into why we choose to not use Fisher matrix estimates for the rest of this paper. A suitable replacement to still account for these systematic errors by using waveform with an error model like (13) on top of them is discussed in the next section.

## V. SIMULATIONS AND TESTS OF PARAMETER ESTIMATION FRAMEWORK

This section uses simulations to test the WF-error parametrizations described in Sec. II A. Though we do not have access to the “true” waveform model, we can still check the validity of the PE framework we describe by considering a reference waveform model,  $h_{\text{ref}}$ , modifying the  $h_{\text{ref}}$ , and checking if our parametrization can pick up the introduced modification. We take a reference signal  $h_{\text{ref}}$  described by a waveform approximant and introduce errors in the model described by parametrizations in Eqs. (12) and (13). Now we have a modified signal  $h_{\text{mod}}$ , which shall have a systematic error compared to the  $h_{\text{ref}}$ . We consider  $h_{\text{mod}}$  as the ‘true’ signal and inject it in the strain data of LIGO detectors at Hanford (H1), Livingston (L1), and Virgo detector (V1). Then, we use the Bayesian PE data analysis pipeline to recover the source properties using  $h_{\text{ref}}$  as our waveform model with and without incorporating waveform uncertainties in PE. We use nonspinning signals (spin parameters in waveform approximant set to zero) with zero noise, except the last subsection where we deal with missing physics. In the subsequent subsections, we describe each injection-recovery campaign’s injected signals and PE scheme.

For PE, we use the publicly available toolkit PYCBC inference [90], and for sampling the likelihood function, we use the nested sampling algorithm implemented in the publicly available code DYNesty [114]. We also implement WF-error parametrizations in code that serves as a plugin [115] for PYCBC. We use the ALIGOZERODETHIGHPOWER PSD for the LIGO Hanford and Livingston detectors and the ADVIRGO PSD for the Virgo detector. These PSD

TABLE III. This table describes the priors used in the PE of simulated signals. The distribution  $\mathcal{U}(a, b)$  refers to a one-dimensional uniform distribution defined within the interval  $(a, b)$ . For the WF-error parameters, when we employ cubic spline curves with ten nodal points across the frequency range of (20, 500) Hz. The prior distributions for each  $\delta\tilde{A}_i$  and  $\delta\tilde{\phi}_i$  are specified accordingly. The same prior is used for the special test cases, when we do not use cubic spline but a constant shift in amplitude and phase with parameters  $\delta\tilde{A}, \delta\tilde{\phi}$ .  $\pi(\theta)$  denotes the prior distribution used for a parameter  $\theta$ .

| Recovery model            | Sampling parameters                  | Description                     | Prior                                                           |
|---------------------------|--------------------------------------|---------------------------------|-----------------------------------------------------------------|
| IMRPhenomPv2              | $\mathcal{M}^{\text{src}}$           | Source frame chirp mass         | Uniform in component masses                                     |
|                           | $q = \frac{m_2}{m_1} (m_1 \geq m_2)$ | Mass ratio                      |                                                                 |
|                           | $V_c$                                | Comoving volume                 | Uniform: $\pi(V_c) = \mathcal{U}(5 \times 10^3, 9 \times 10^9)$ |
|                           | $i$                                  | Inclination angle               | $\pi(i) \propto \sin(i)$                                        |
|                           | $t_c$                                | Time of arrival                 | Uniform: $\pi(t_c) = \mathcal{U}(t_c - 0.1, t_c + 0.1)$         |
|                           | RA, dec                              | Right ascension and declination | Uniform Sky                                                     |
|                           | $\psi$                               | Polarization Angle              | Uniform angle: $\pi(\psi) = \mathcal{U}(0, 2\pi)$               |
| IMRPhenomPv2+<br>WF error | $\delta A_i, \delta \phi_i$          | WF-error parameters             | $\mathcal{N}(\mu = 0, \sigma = 0.032)$                          |

functions are implemented in publicly available modules LALSIMULATION, which is a part of the LALSUITE code [116]. We use two sets of runs for each injection:

- (i) In first set of runs, we sample the parameter space with the following parameters: source frame chirp mass ( $\mathcal{M}^{\text{src}}$ ), mass ratio  $q (q = \frac{m_1}{m_2}, m_1 \geq m_2)$ , comoving volume, inclination angle, and geocentric time of arrival of the signal ( $t_c$ ). We fix right ascension (RA), declination (dec), and polarization angle ( $\psi$ ) for these runs.
- (ii) In the second set of runs, in addition to the parameters described above, we also vary ra, dec, and  $\psi$ .

We transform the sampling parameters to the waveform model parameters, such as detector frame masses ( $m_1^z, m_2^z$ ), and luminosity distance ( $D_L$ ) through standard transformations. We fix the spins to be zero for the study in this manuscript, except for in Sec. V C. For the PE runs described by WF-error parametrizations, we use additional parameters. We use two additional WF-error parameters for the constant shift models:  $\delta\tilde{A}$  and  $\delta\tilde{\phi}$ . For the cubic-spline model, we use ten frequency nodal points with log spacing between a minimum frequency of 20 Hz and a maximum frequency of 500 Hz.

In Table III, we describe the prior distribution for each of the abovementioned parameters. All the injections described in this section are zero-noise injections.

### A. Simplest modification: A constant shift in amplitude and phase with rel-phase parametrization

We consider a GW150915-like signal with source frame masses  $m_1^{\text{src}}, m_2^{\text{src}} = (36M_\odot, 29M_\odot)$  at a distance of 500 Mpc. We generate the corresponding reference signal using the IMRPhenomPv2 waveform model and introduce the errors in reference signal with rel-phase parametrization corresponding to

$$(\delta\tilde{A}, \delta\tilde{\phi}_{\text{rel}}) = (0.01, 0.01). \quad (19)$$

For this most straightforward case, a constant phase shift is expected to be significant for rel-phase parametrization. For abs-phase parametrization, a constant phase shift will not introduce any physical effect except a constant shift in time. We will consider the general modification with abs-phase parametrization in the following subsections. We perform PE with the injected signal in the detector network, HLV, with the following combination:

- (i) *Injection: standard, PE: standard* In this scenario, we inject the reference signal as described above and perform the PE with the same waveform model without introducing the WF-error parametrization.
- (ii) *Injection: modified, PE: standard* In this scenario, we inject the modified signal and perform the PE with the reference waveform model without introducing the waveform errors.
- (iii) *Injection: modified, PE: WF-errors* In this scenario, we inject the modified signal, perform the PE with the reference waveform model + WF-error parametrization.
- (iv) *Injection: standard, PE: WF-errors* Here, we inject the reference signal, perform the PE with the reference waveform model + WF-error.

In Fig. 5, we show the injection and recovery of various combinations. We show that when we have a modified injection (equivalent to a systematic error with rel-phase parametrization), and we perform PE with reference waveform model, we get a bias in the detector frame chirp mass  $\mathcal{M}^z$ . This bias is corrected when we use the right parametrization to account for waveform systematics. We also note that we can correct the bias when we use cubic splines for constant shift error in the reference signal. However, it is to be noted that when we use abs-phase parametrization to

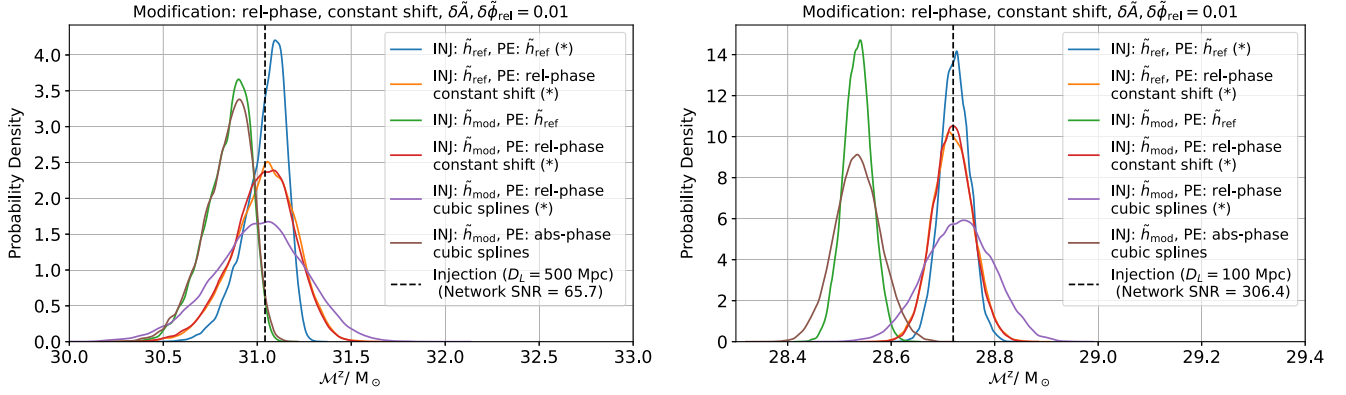


FIG. 5. Injection and recovery are shown for a nonspinning GW150914-like signal with source frame masses  $(m_1^{src}, m_2^{src}) = (36, 29)M_\odot$ , at a luminosity distance of 500 Mpc (left panel) and a distance of 100 Mpc (right panel). We also inject a modified signal with rel-phase modification with parameters  $(\delta\tilde{A}, \delta\tilde{\phi}) = (0.01, 0.01)$ . When we inject a modified signal and use the incorrect model to recover, we see a bias in the recovery of detector frame chirp mass  $\mathcal{M}^z$ . We also show that using the correct parametrization (in this case, rel-phase) can correct the bias and get a broader marginalized posterior sample. The bias is more prominent in a louder signal (right panel) than in a comparatively weaker signal (left panel). We mark that combination of injection and recovery by an asterisk (\*) where we do not expect a bias. The vertical dashed line represents the injected value of  $\mathcal{M}^z$ . For these PE runs, we fix the parameters RA, dec, and polarization to the injected values. The network SNR for reference injection is shown in each panel.

correct for the bias introduced by rel-phase parametrization, it fails to correct for the bias. These effects are more visible in a louder signal, which is simulated by injection at a closer luminosity distance (100) Mpc compared to one at 500 Mpc. We also show the network SNR,  $\rho^{\text{net}} = \sqrt{\sum_i \rho_i^2}$ , for the reference waveform injection and recovery, where  $\rho_i$  is the SNR for the  $i$ th detector. Figure 6 shows the recovery of parameter  $\delta\tilde{\phi}$  with rel-phase parametrization. We show that we can recover the injected value of  $\delta\tilde{\phi}$  with the correct parametrization. When there is no modification to the reference signal, the rel-phase parametrization recovers  $\delta\phi = 0$ . We fix the RA, dec, and  $\psi$  for the runs shown in Figs. 5 and 6. We find that the  $\delta\tilde{A}$  parameter returns the prior distribution and does not have a constraining power

for the PE runs described in this subsection. For simplistic cases, we expect  $\delta\tilde{A}$  to be degenerate with distance, i.e., a small constant shift in the amplitude can be modeled in the luminosity distance. At this stage, it can be argued that we can get rid of  $\delta\tilde{A}$  parameter in the WF-error parametrization and focus on  $\delta\tilde{\phi}$  parameter. However, we shall keep this parameter in our model for more general scenarios where errors in amplitude can be a function of the frequency. We do not find any bias in the other parameters considered in the simulations in this subsection, except for the detector frame chirp mass.

Through these simulations, we have established that for a simplistic model, a systematic error corresponding to  $\delta\tilde{\phi} = 0.01$  can lead to significant bias in the detector frame chirp

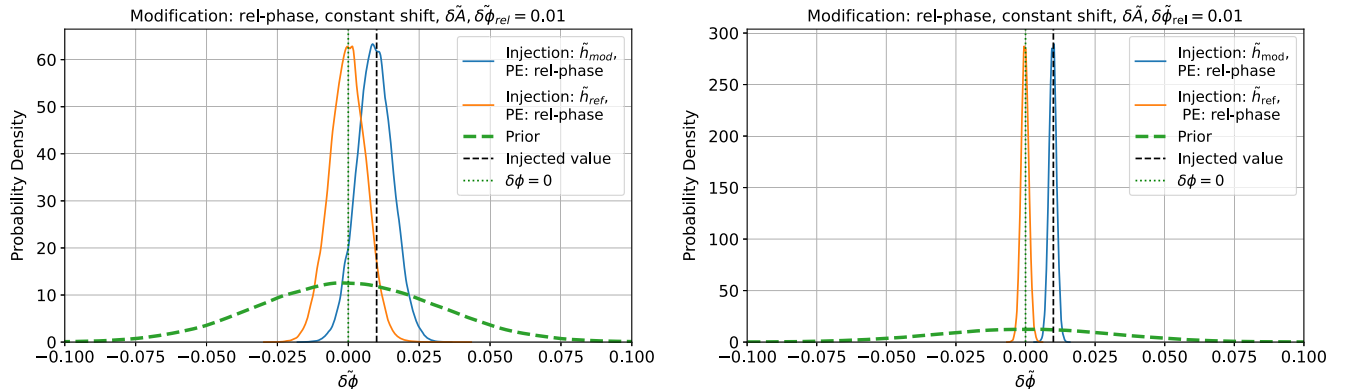


FIG. 6. Recovery of the parameter  $\delta\tilde{\phi}$  for rel-phase parametrization for various scenarios of injected parameters. The signal injected here is a nonspinning GW150914-like, as in Fig. 5. The left panel has an injection at 500 Mpc, while the right panel shows the case with an injection at 100 Mpc. The vertical dashed line shows the injected value for  $\delta\tilde{\phi}$  for the modified signal. The green dashed curve shows the prior distribution for the  $\delta\tilde{\phi}$  parameter. We can recover the injected value of  $\delta\tilde{\phi}$  with the correct parametrization for PE. We recover  $\delta\phi = 0$  for the reference injection. For these PE runs, we fix the parameters RA, dec, and polarization to the injected values.

mass for GW150914-type signal, with the detector sensitivity of advanced LIGO design sensitivity. We have also shown that, by using WF-error parametrization, we can account for the systematic errors introduced in the reference waveform. We can also recover the injected modification in the  $\delta\phi$  parameter with correct parametrization. However, when we modify the reference waveform with rel-phase parametrization, the PE recovery with abs-phase parametrization fails to correct the observed bias in  $\mathcal{M}^2$ . This is because the relative-phase parametrization can accommodate more significant deviations in the reference waveform than the absolute-phase parametrization for the same range of priors in  $\delta\tilde{\phi}$ , as previously discussed.

### B. A more realistic modification using cubic splines

Within the framework of WF-error parametrizations, a waveform model  $h_{\text{ref}}$  will deviate from the “true” model described by general functions of frequency:  $\delta\tilde{A}(f), \delta\tilde{\phi}(f)$ . We use cubic splines to generate them at specific frequency nodal points. In this subsection, we introduce a more general type of modification to our reference GW150914-like signal within the framework of both the parametrizations,

$$\delta\tilde{A}, \delta\tilde{\phi} \sim \mathcal{N}(\mu = 0, \sigma = 0.002) \quad (20)$$

$$\delta\tilde{A}, \delta\tilde{\phi} \sim \mathcal{N}(\mu = 0.01, \sigma = 0.002). \quad (21)$$

Figure 7 shows examples of cubic spline curves generated from the abovementioned distributions. We use a realization of  $(\delta\tilde{A}_i, \delta\tilde{\phi}_i)$  at nodal frequency points generated from the distributions described by the Eqs. (20) and (21).

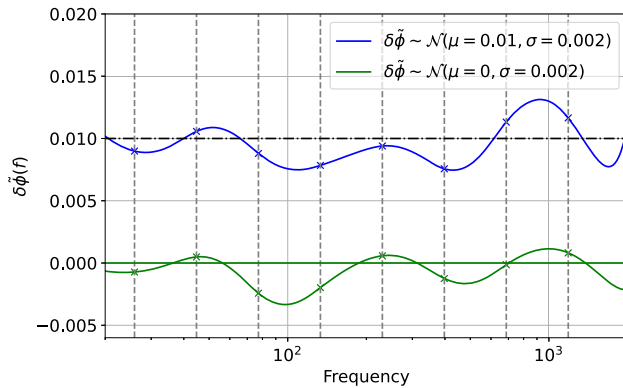


FIG. 7. We show a realization of cubic spline curves with ten frequency nodal points (vertical dashed curves) for  $\delta\tilde{\phi}_i$  taken from two normal distributions: one with zero mean (green curve) and the another with nonzero mean  $\mu = 0.01$  (blue curve). The points marked with a cross (x) are the realization of the normal distribution. In reality, these curves will represent a general function of frequency that captures the waveform systematics or deviation from the “true” waveform model.

In Fig. 8, we show the recovery of  $\delta\tilde{\phi}$  parameter at each frequency nodal point for the modified injection generated using rel-phase parametrization. We demonstrate that the WF-error parametrization framework can recover the  $\delta\tilde{\phi}$  curves used to modify the reference waveform model. For a relatively louder signal, corresponding to the source located at  $D_L = 100$  Mpc, we observe that the posterior samples at nodal points align well with the injected values of  $\delta\tilde{\phi}$ . These examples demonstrate that if the reference waveform model deviates from reality, which is described by rel-phase parametrization, the framework presented here has the potential to account for that. In other words, in the absence of true knowledge of WF-error priors, we can use a data-driven approach to determine if there is a deviation from the reference waveform model, especially for loud signals.

We now focus on the other parameters, such as chirp mass  $\mathcal{M}$ . In Fig. 9, we present various combinations of injections and PE recovery. Our findings can be summarized as follows:

- (i) When the injection uses the reference waveform model  $h_{\text{ref}}$  and PE recovery is performed with the same model, we achieve the expected recovery with no bias.
- (ii) If we enable the waveform error parametrizations (rel-phase or abs-phase) while using the correct recovery model, we observe a broadening of the posterior samples as anticipated. The broadening of the posterior distribution is more pronounced with the rel-phase parametrization than the abs-phase parametrization.
- (iii) In the scenario when the injection is performed by modifying the reference signal and we use the reference signal to recover the source parameters, we get bias in the chirp mass. As expected, the bias is more prominent for a large SNR signal.
- (iv) In the case of modified injections, using WF-error parametrization in conjunction with the  $h_{\text{ref}}$  recovery model broadens the posteriors. Additionally, employing the correct parametrization leads to a correction in bias. The bias introduced by rel-phase parametrization can only be corrected using rel-phase parametrization. The abs-phase parametrization does not correct the biases introduced by rel-phase parametrization.
- (v) For the abs-phase parametrization, the moderate deviation introduced by the cubic spline curves generated from the abovementioned distribution does not introduce significant bias for a GW150914-like signal with advanced LIGO detector sensitivity. In order to see any significant bias for abs-phase parametrization, we either need a comparatively large deviation in phase or a very high SNR, such as in the case of 3G detectors.

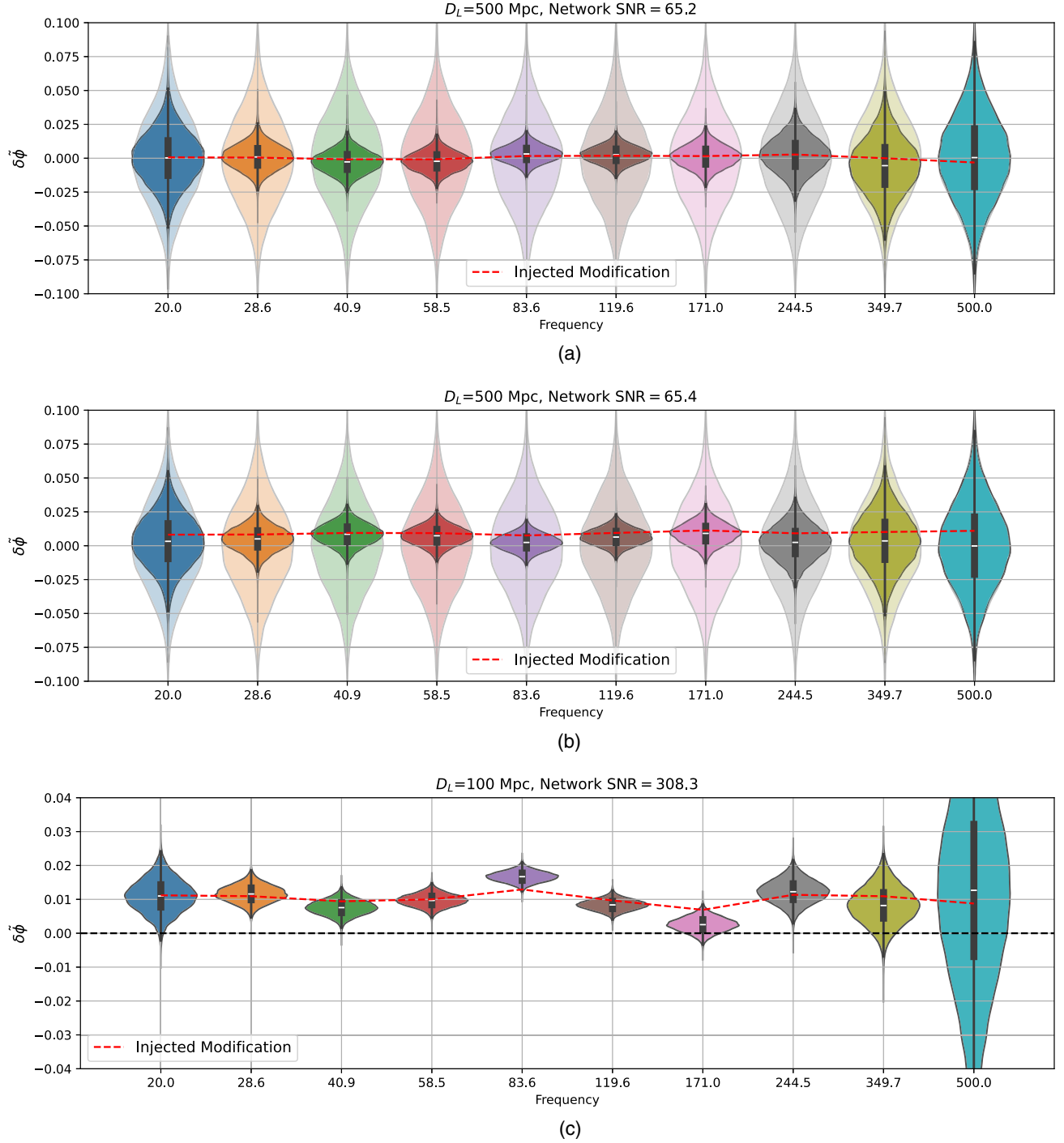


FIG. 8. The recovery of WF-error parameter  $\delta\tilde{\phi}_i$  for rel-phase parametrization is shown for each frequency nodal point used in PE. The light violin plot represents the prior distribution for each  $\delta\tilde{\phi}_i$ , and the dark-shaded region represents the one-dimensional marginalized posterior samples. The last frequency nodal point at  $f = 500$  Hz returns the prior because the signal merges before this frequency. The constraints are also weaker at  $f_i = 20$  Hz because this frequency bin's signal-to-noise ratio is comparatively low. (a) The modification to the reference waveform is applied using the rel-phase parametrization with a cubic spline realization generated from the distribution  $\delta\tilde{A}_i, \delta\tilde{\phi}_i \sim \mathcal{N}(\mu = 0, \sigma = 0.002)$ . Panels (b) and (c) represent the modification from the distribution  $\delta\tilde{A}_i, \delta\tilde{\phi}_i \sim \mathcal{N}(\mu = 0.01, \sigma = 0.002)$ . The network SNR and luminosity distance of the injected (GW150914-like) signal are shown at the top of each panel. For these PE runs, we fix the parameters RA, dec, and polarization to the injected values.

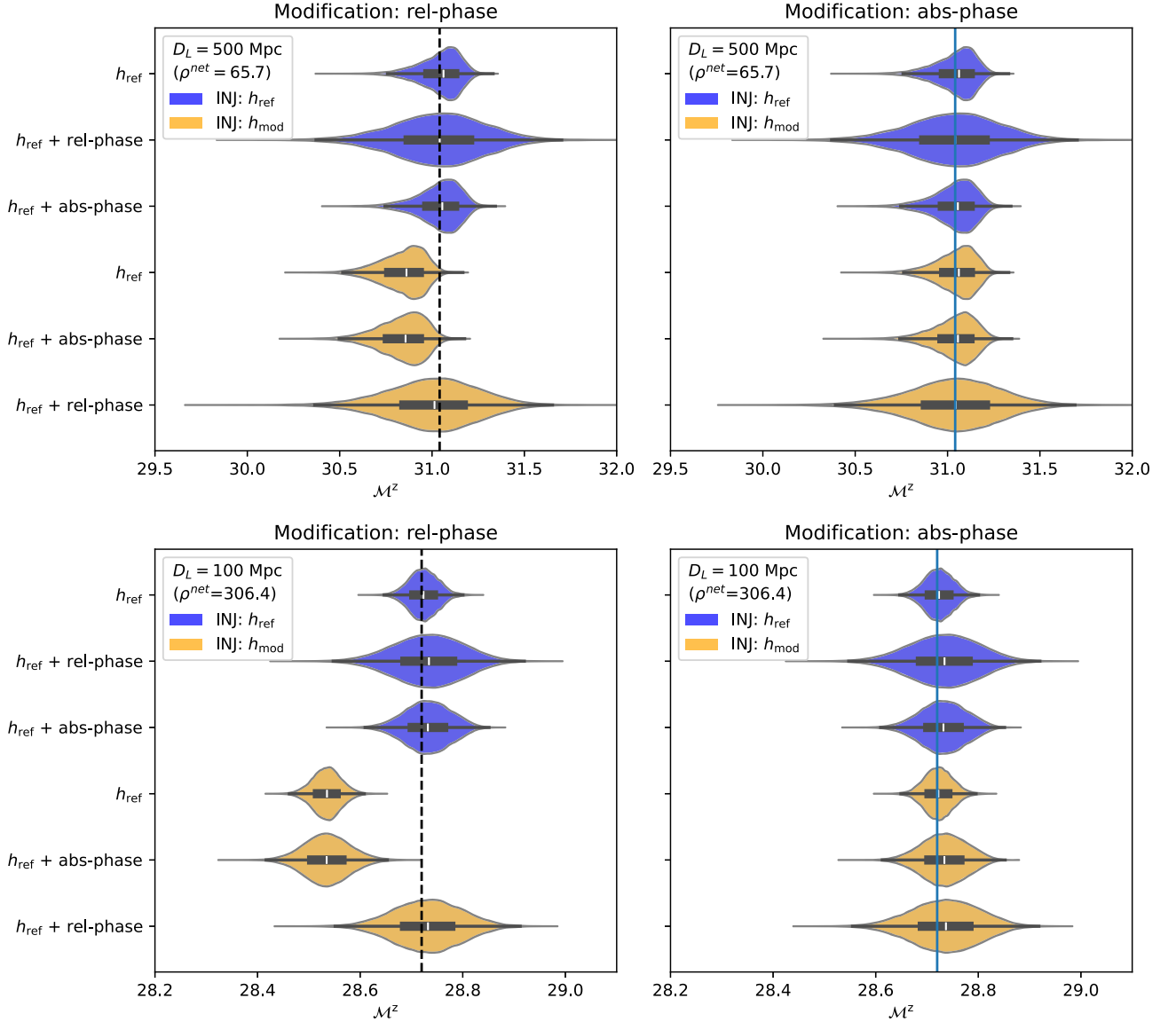


FIG. 9. A GW150914-like signal is injected and recovered using two WF-error formalisms. The blue violin plots show the cases with reference injection  $h_{ref}$  while orange violin plots show the modified injections. The left panels represent the cases where the modification is done with rel-phase modification, while the right panel shows abs-phase modification. The top rows represent the injections with the source placed at a luminosity distance of 500 Mpc, while the bottom panels show the injection corresponds to the same source placed at 100 Mpc. Labels on the y axis show the waveform models used in PE runs. For these PE runs, we fix the parameters RA, dec, and polarization to the injected values. The network SNR,  $\rho^{net}$ , for reference injection is shown in each panel.

Through these simulations, we conclude that if the “true” waveform model deviates from the reference waveform model in a manner described by the parametrizations in Eqs. (12) and (13), our framework can correct any bias in source parameters, if present. Additionally, a sufficiently loud signal can recover the frequency dependence of WF-error parameters. We still need to ensure that the priors are sufficiently broad to capture the deviations. Additionally, the selection of knots or nodal points for cubic splines should accurately reflect the true nature of  $\delta\tilde{\phi}$  as a function of frequency. In these examples, we use binning in log

frequency and use a total of ten nodal points between 20 Hz and 500 Hz.

### C. Incorporating missing physics

One use of WF-error parametrization can be to account for the missing physics. For BBH mergers, the most up-to-date waveform model includes effects such as higher-order harmonics and system precession, and the waveform developers are working on systems with eccentricity [117–120]. The most general description of a BBH merger can include other environmental effects such as matter

accretion, another binary or heavy object nearby, or a merger near a supermassive BH near the host galaxy's center.

However, the waveform model used in PE might not fully describe the reality. We want to see if the WF-error parametrization can capture the deviation from the actual model because the missing effects are not considered in the waveform model. In order to test this scenario, we injected a set of precessing signals with components coming from the uniform distribution  $s_{1,2}^{x,y,z} \in \mathcal{U}(-0.1, 0.1)$ . We use the IMRPhenomPv2 model and other parameters described in the previous section. We call the injected model  $h_{\text{prec}}$ . We expect bias in the measured parameters for a precessing injection if we use a recovery model that does not include spins in the waveform description:  $h_{\text{no-spins}}$ , where we set spin parameters to zero in the IMRPhenomPv2 waveform model. For the PE analysis, we use the priors described in Table III. For the precessing PE runs, we use isotropic spin priors for all the spin components  $s_{1,2}^{x,y,z}$ . Our selection of spin parameters aims to modify the reference waveform model, which includes the missing physics. Our goal is not to model deviations based on an actual astrophysical distribution, but rather to investigate whether we can incorporate such deviations for individual signals, if they are present.

In order to test if cubic-spline corrections to the  $h_{\text{no-spin}}$  work, we simulate 500 realizations of the GW strain  $h_{\text{prec}}$ . We then estimate the amplitude and phase differences  $[\delta\tilde{A}_{\text{rel}}(f), \delta\tilde{\phi}_{\text{rel}}(f)]$  with respect to the nonspinning waveform template  $\hat{h}_{\text{no-spin}}$ . We then choose a different number of nodal points,  $N_p$ , and apply the correction to the nonspinning template, using cubic splines generated using these nodal points. We define the mismatch between two normalized templates  $\hat{h}_1$  and  $\hat{h}_2$  as

$$\text{MM}(\hat{h}_1, \hat{h}_2) = 1 - \langle \hat{h}_1 | \hat{h}_2 \rangle, \quad (22)$$

where  $\hat{h}$  is a normalized template such that the inner product  $\langle \hat{h} | \hat{h} \rangle = 1$ . We estimate the mismatch between the waveform strain before and after the correction, i.e.,  $\text{MM}_1 = \text{MM}(\hat{h}_{\text{prec}}, \hat{h}_{\text{no-spin}})$ , and  $\text{MM}_2 = \text{MM}(\hat{h}_{\text{prec}}, \hat{h}_{\text{no-spin}}^{\text{corrected}})$ . When the cubic spline corrections to  $h_{\text{no-spin}}$  bring the model closer to the injected signal  $h_{\text{prec}}$ , the ratio  $\frac{\text{MM}_2}{\text{MM}_1}$  is less than unity, or the logarithm of this ratio is less than 0. In Fig. 10, we show the distribution of the ratio  $\frac{\text{MM}_2}{\text{MM}_1}$  for different number of nodal points used for cubic spline curves. For  $N_p = 8$ , we get this ratio less than one for  $> 70\%$  simulated precessing signals. For  $N_p = 10$ , more than 99% of cases show the improvement. This number was 100% for  $N_p \geq 12$ . This indicate that, in principle, using  $\geq 10$  nodal points to apply corrections to baseline nonspinning waveform model should be enough

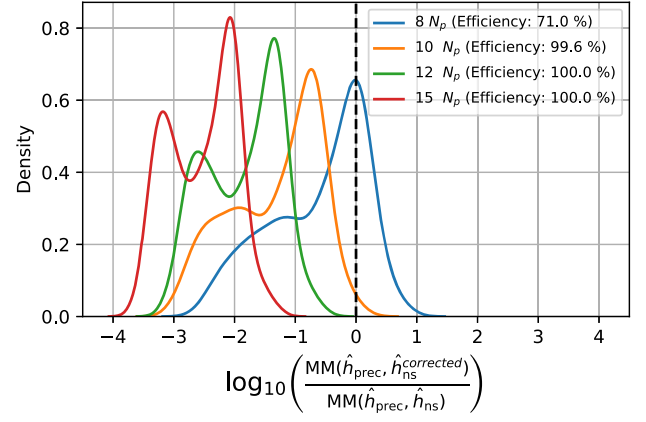


FIG. 10. We demonstrate the efficiency of the cubic spline method based on the number of nodal points ( $N_p$ ) used to correct the baseline model. The x axis represents the logarithm of the ratio of mismatches: the mismatch after correction compared to the mismatch before correction. The function  $\text{MM}(h_1, h_2)$  is defined in Eq. (A2). Here,  $h_{\text{prec}}$  represents the injected waveform strain with precessing signals, while  $h_{\text{ns}}$  refers to the strain generated with the same parameters except the spin parameters set to zero.  $h_{\text{ns}}^{\text{corrected}}$  indicates the corrected strain using waveform-error parametrization with  $N_p$  frequency nodal points. A logarithmic mismatch ratio less than one indicates an improvement, meaning that the corrected model is closer to the injected waveform  $h_{\text{prec}}$ . The vertical dashed line serves as a reference. The efficiency is defined as the percentage of simulation points where an improvement is achieved (i.e., where the logarithmic mismatch ratio is less than 0).

to incorporate missing physics considered in this specific study.

We now focus on PE runs done in this study with ten realizations of precessing signals generated from the distribution described earlier in this subsection. Figure 11 shows that when we use a nonspinning waveform model  $h_{\text{no-spins}}$  to recover precessing injections, we get bias in the recovered chirp mass parameter  $\mathcal{M}^z$  in detector frame. However, when we use WF-error parametrizations along with  $h_{\text{no-spins}}$  waveform model, we notice that, for rel-phase parametrization, it results in broader marginalized posterior samples. However, abs-phase parametrization cannot make their posterior samples broad enough to correct the biases for this specific example. We observe that for injections with a significant mismatch ( $\geq 1\%$ ) to the reference model  $h_{\text{ref}}$ , the correction remains inadequate to account for the missing physics fully. We will revisit this issue later in this subsection.

We assess the performance of the WF-error parametrization using the ratio of systematic errors to statistical errors, expressed as  $|\Delta\mathcal{M}^z|/\sigma_{\mathcal{M}^z}$ , with two runs consist of one that employs the  $h_{\text{no-spin}}$  waveform model and another that applies the same model with the rel-phase parametrization. As demonstrated in Eq. (6), systematic

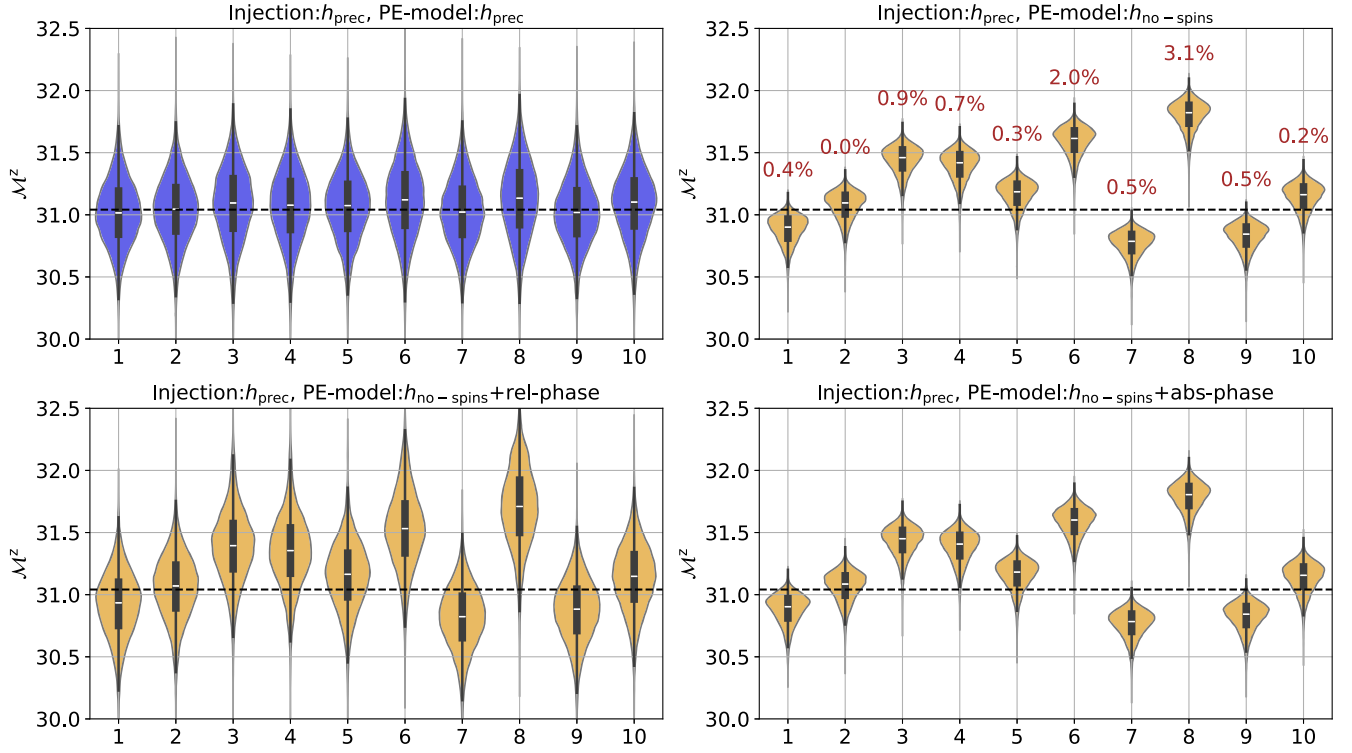


FIG. 11. We use ten precessing injections in the HLV network. The top left panel shows the posterior samples from each PE run where we use the precessing recovery model:  $h_{\text{prec}}$ . We use the nonspinning model  $h_{\text{no-spins}}$  for PE in the top right panel. The bottom panels use a nonspinning model along with the WF-error framework. The bottom left plot shows recovery with rel-phase parametrization, and the bottom right model uses abs-phase parametrization. The horizontal dashed line represents the injected chirp mass ( $\mathcal{M}^z$ ) value in the detector frame. The x axes represent the injection simulation IDs. We use all the parameters listed in Table III for these PE runs. We also use component spin parameters  $s_{1,2}^{x,y,z}$  for the precessing recovery. We use the same priors for  $(\delta\tilde{A}_i, \delta\tilde{\phi}_i)$  parameters for both the WF-error parametrizations in bottom panels. The numbers above the data points in the top right panel represent mismatch (%) between the injected template waveform  $h_{\text{prec}}$  and the corresponding nonspinning template  $h_{\text{no-spin}}$ .

errors become important when this ratio is of the order of unity. In Fig. 12, on the x axis, we show the mismatch between two templates:  $h_{\text{prec}}$  and  $h_{\text{no-spins}}$ , keeping the values of nonspinning parameters the same. It quantifies how much overlap there is between the waveform templates. On the y axis, we show the quantity  $|\Delta\mathcal{M}^z|/\sigma_{\mathcal{M}^z}$  (left panel), and the bias  $\mathcal{M}^z$  (right panel). The bias is estimated as the difference between the injected value and the median of the 1D marginalized posterior samples. The ratio  $|\Delta\mathcal{M}^z|/\sigma_{\mathcal{M}^z}$  increases when there is a larger mismatch between the two templates. It shows a significant drop in the ratio when the rel-phase WF-error parametrization is used with the  $h_{\text{no-spins}}$  waveform model. This drop is a combination of both the effects: broadening of posterior samples (larger  $\sigma$ ), as well as a decrease in bias  $\Delta\mathcal{M}^z$  (See right panel of Fig. 12). In this set of PE runs, we do not find significant biases in any other parameter except  $\mathcal{M}^z$ .

The functional form of the parameters  $(\delta\tilde{A}, \delta\tilde{\phi})$  for the simulation set considered in this section is shown in the Appendix B. In our PE runs, though we notice

consistent decrease in the bias value  $|\Delta\mathcal{M}^z|$  and the ratio  $|\Delta\mathcal{M}^z|/\sigma_{\mathcal{M}^z}$ , we focus on two injections where the original mismatch was more than 1%. The current framework fails to fully mitigate the systematic bias (ideally  $|\Delta\mathcal{M}^z| \rightarrow 0$ ). Upon further investigation, we find that this could be because of a combination of (i) A most generalized waveform modification should be applied on both polarizations ( $h^+, h^\times$ ) independently, and (ii) a large prior ranges for  $\delta\tilde{A}_i, \delta\tilde{\phi}_i$  are needed to incorporate significant deviations for large mismatch. We leave the details of such an investigation for a follow-up study.

We want to emphasize that this is an extreme example of the proof of concept. However, it indicates that we can switch on appropriate WF-error parametrization in PE analysis whenever we are unsure if the waveform model captures the realistic description. If the “true” model differs from the reference model in such a way that the WF-error parametrizations can capture it, we should be able to account for potential biases in the source parameters of the binary system.

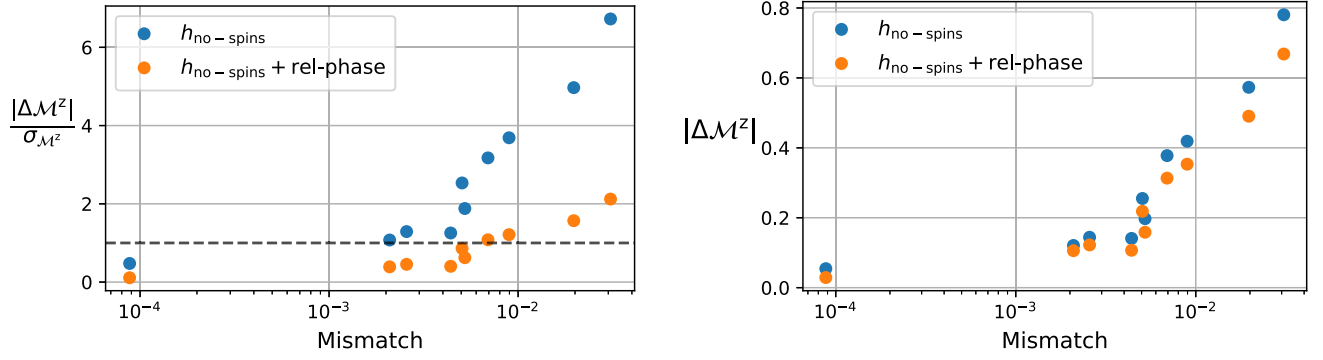


FIG. 12. For the injections and recovery shown in Fig. 11, on the x axis, we show the mismatch between the waveform generated with nonspinning and precessing parameter values. We kept all other parameters the same (see text). Left panel: The y axis shows the ratio of the absolute value of the bias ( $\Delta\mathcal{M}^z = \mathcal{M}_{\text{median}}^z - \mathcal{M}_{\text{inj}}^z$ ) to the standard deviation of the 1D marginalized posterior samples of  $\mathcal{M}^z$ . Horizontal dashed line represent the line where  $\Delta\mathcal{M}^z = \sigma_{\mathcal{M}^z}$ . Right panel: The y axis shows the absolute value of the bias  $\Delta\mathcal{M}^z$ .

## VI. COMPARISON WITH OTHER APPROACHES IN THE LITERATURE

This section compares our framework with existing approaches in the literature regarding waveform systematics. We categorize the types of systematic errors that waveform modeling may encounter, focusing on at least two primary categories:

- (i) *Type-A errors*: These errors are realized when GW waveform morphology of a true signal  $h_{\text{true}}$  can not be achieved by the waveform model  $h_{\text{model}}(\Theta)$ . Mathematically,

$$\langle \hat{h}_{\text{true}} | \hat{h}(\Theta) \rangle < 1 \quad \forall \quad \Theta, \quad (23)$$

where  $\hat{h}$  is the normalized strain such that  $\langle \hat{h} | \hat{h} \rangle = 1$ .

- (ii) *Type-B errors*: The waveform model can generate the true signal's morphology, but for the wrong parameters  $\Theta$ . These errors are challenging to address, particularly if we rely on only one waveform model or if all available waveform models exhibit similar systematic errors.

Current approaches to mitigate the waveform systematics can be categorized in several ways, with some methods fitting into multiple categories. However, we will list them under their primary category:

- (i) *Account for differences between waveform models*: In some regions of parameter space, such as high spins or highly unequal mass ratios, waveform models fall short of the accuracy requirements [121], and different waveform models may give inconsistent results in PE [6]. Several approaches try to account for these differences, e.g., by combining the posterior samples with appropriate weighting [33–35] or using hyperparameters [36–41] to sample across different waveform models. It is important to understand that if all waveform models share

common systematics, particularly type-A errors, then these methods will not be able to address those systematics. However, if one of the waveform models exhibits type-B errors, this approach can help mitigate the impact of waveform systematics by diluting the contribution from that model.

- (ii) *Numerical relativity calibration errors*: Other approaches include using Gaussian process regression techniques to model deviation from the reference waveform [42]. The deviations can be modeled by interpolating from differences found with accurate templates, e.g., numerical relativity simulations, or by modeling the numerical relativity calibration parameters used in waveform models based on effective one-body (EOB) approaches [43,44]. These approaches may fail to mitigate type-A errors if the NR simulations used for the waveform model calibration themselves contain the systematics, e.g., thermal systematics in binary neutron star waveforms [122,123].
- (iii) *Probabilistic waveform models*: Another suggested approach is that waveform developers provide the probabilistic models across the parameter space. For each waveform call, a bunch of curves with different weights associated with the model or an error band as a function of time or frequency can be provided [45].

Our framework provides a generic, data-driven approach in which the cubic splines can account for various systematic errors, including waveform systematics. While providing the prior ranges, we expect them to contain error budgets from waveform systematics, missing physics, and other data analysis artifacts. Examples of the data analysis artifacts include residual power from glitch removal near the signal or nonstationary noise. We can integrate other approaches in the literature with our framework to provide error budgets for waveform

systematics, thus serving as a complementary method. Another unique feature of our framework is the ability to mitigate type-A systematic errors. If these types of systematic errors are present, they can show up in the posterior distribution of the nodal points as deviating away from  $(\delta\tilde{A}_i, \delta\tilde{\phi}_i) = 0$ . It can guide determining that type-A systematic errors are present in the data, as shown in panel (c) of Fig. 8. For type-B errors mitigation, we can combine our framework with the methods that account for differences between waveform models.

## VII. DISCUSSION AND SUMMARY

The systematic errors in the PE analysis of a GW merger can arise due to waveform modeling errors, data analysis artifacts, or missing physics. As the GW detectors become more sensitive, the statistical errors become smaller, and we are reaching an era where systematic errors can not be ignored. Several approaches in the literature deal with accounting potential systematic errors in PE. To account for potential differences between existing waveform models, posterior samples are combined [33–35], or hyperparameter models are used to sample likelihood samples [36–39,41]. Other approaches include marginalizing the potential differences using prior distribution analytically [42]. These priors can be constructed by estimating the NR calibration errors [43,44] or using PN-based methods to mitigate the errors [46].

In this work, we present a general PE framework to account for the errors in the waveform modeling of GW sources. We assume that the reference signal,  $h_{\text{ref}}$ , used in the recovery of the GW source parameters, is different from the “true” signal  $h_{\text{true}}$ ; the parameters can model the differences  $(\delta\tilde{A}, \delta\tilde{\phi})$ , as differences in the amplitude and the phase of the signal. We present two WF-error parametrizations: one corresponds to the relative phase, while the other corresponds to the absolute phase difference from the so-called true phase.

We use the cubic-spline method to modify the reference waveform model. We developed a Python code as a plugin that can be easily integrated with the publicly available GW analysis tool PYCBC. The source code can be downloaded from [115]. We also release the frame files for the simulated injections and configuration files of all the PE runs done in this analysis [124].

If the WF-error budgets for the parameters  $(\delta\tilde{A}, \delta\tilde{\phi})$  are available for a given waveform model, these error budgets can be used as a prior in WF-error parametrization. Without such error budgets, we can use wider priors, potentially making our errors in estimating other parameters broader. This code and analysis framework can be utilized for the following purposes: (i) to account for systematic errors in waveform models by incorporating appropriate priors for  $(\delta\tilde{A}, \delta\tilde{\phi})$ , such as calibration errors from numerical relativity (NR), or (ii) to address the potential bias that may

arise from missing physical effects in the waveform description.

We use Fisher matrix formalism to study its abilities and limitations in accounting for the bias induced by  $(\delta\tilde{A}, \delta\tilde{\phi})$ . In the LSA regime, where the modified model is very close to the reference model, we can use the Fisher matrix formalism to account for systematic biases. However, in the non-LSA regime, where the overlap between the waveform models is smaller, we can no longer trust the Fisher matrix approach to correctly predict the systematic bias.

Our findings indicate that even a 1% phase error in the rel-phase parametrization can introduce bias in the observed chirp mass. The bias in the chirp mass could affect the conclusion of follow-up studies that rely on the correct distribution of source parameters, such as inferring the properties of an intrinsic population of BBH mergers. For the abs-phase parametrization, the moderate deviation in  $\delta\phi$  of the order of  $\sim\mathcal{O}(0.01)$  radians, with cubic splines, do not introduce significant bias, at least in current generation detectors. We need a more significant deviation or a very high SNR signal to notice any bias. However, even in the absence of significant bias, we observe a broadening of the posterior samples by up to 20% ( $\frac{\sigma}{\sigma_{\text{ref}}} \approx 0.2$ ), where  $\sigma$  is the standard deviation of 1D marginalized posterior of parameter  $\mathcal{M}$  using the abs-phase parametrization, and  $\sigma_{\text{ref}}$  is the standard deviation with the reference waveform model.

We make a case for waveform developers to provide error budgets alongside the waveform models. It is necessary to accurately account for the WF-error parametrization in the PE analysis. These error budgets can be frequency dependent  $(\delta\tilde{A}, \delta\tilde{\phi})$  regions of  $1\sigma$  uncertainties. We call them WF-error envelopes. These WF-error envelopes are expected to be a function of parameter space. These envelopes can be used as priors for  $(\delta\tilde{A}, \delta\tilde{\phi})$  parameters in WF-error parametrization. However, if such envelopes are unavailable, we can use wider priors to account for potential waveform systematics, allowing the signal’s loudness and data to inform us about the constraints on the WF-error parameters. In this scenario, we may face a penalty in the form of broader posterior samples due to the significantly larger prior volume.

One of the use cases of WF-error parametrization is to see if the waveform model in use is missing a description of reality. The next generation of waveform models are expected to include combined effects such as eccentric orbits, precessing systems, and higher-order modes. If any of the above-mentioned effects are missing from the description of the waveform model, we might introduce bias in inferred parameters such as chirp mass. Assuming the true model differs from the reference waveform model used for the PE, we can use WF-error parametrization to determine if the data favor this hypothesis. At the very least, we expect the posterior samples to be

broad enough to bring down the bias and standard deviation ratio. In order to test this, we used ten random realizations of mildly processing systems and performed PE analysis with the nonspinning model. We find expected bias in this scenario. However, when we switch on rel-phase parametrization, the ratio  $|\Delta\mathcal{M}^z|/\sigma$  comes down, up to, by a factor of 3. In this case, the abs-phase parametrization could not account for correcting the bias with given prior range.

The method we present here is based on certain assumptions that need to be tested in broader scenarios. We assume that the reference waveform differs from the true waveform in a manner that can be parametrized using one of the WF-error parametric models described by Eqs. (12) or (13). Additionally, we assume that the number of knots or nodal points selected within the frequency bins are sufficient to produce cubic spline curves that accurately model the deviation.

As we approach an era where loud signals are more common, we expect systematic biases to become amplified. It is important to account for the WF-error parametrization or any other scheme that can address potential systematics in the reference waveform model. The pipeline we present can work in scenarios where the types of systematics are unknown. We can select our preferred waveform model, and the parameterization best describes the deviations, allowing the data to indicate any possible systematic errors. At the same time, this approach results in broader posterior samples, and such an outcome is anticipated.

The authors plan to explore the use of WF-error parametrization to correct for potential biases in source parameters due to missing physics in more detail in the follow-up studies. For example, the next generation waveform models are expected to include the effects such as spin-precession, eccentric orbits, as well as higher modes. Until this is achieved, there might always be a question of whether there is inherent bias in estimating source properties due to a missing description of one of the above effects. In our follow-up studies, we will use existing NR simulation and state-of-the-art waveform models to study if any potential biases can be accounted for, and corrected if possible.

We make certain assumptions for the method presented here, and these assumptions need to be tested in broader scenarios. We assume that the reference waveform differs from the true waveform in a manner that can be parametrized using one of the WF-error parametric models described by Eqs. (12) or (13). Additionally, we assume that the number of knots or nodal points selected within the frequency bins are sufficient to produce cubic spline curves that accurately model the deviation. In our future studies, we will thoroughly test these assumptions for more general scenarios such as the systems with precession and/or eccentricity. Moreover, if the amplitude and phase corrections are highly oscillating functions of frequency, which

can not be captured by the cubic-spline curves and a small number of fixed nodal points, we need to test other schemes, such as flexible nodal points and/or more complicated interpolation schemes.

## ACKNOWLEDGMENTS

We acknowledge the Max Planck Gesellschaft and the Max Planck Independent Research Group Program, through which this work was supported. S.K. also acknowledges support by the research programme of the Netherlands Organisation for Scientific Research (NWO). We thank the computing team from AEI Hannover for their significant technical support. We also thank Binary Merger Observations and NR group members at AEI for their feedback and valuable comments. We further thank Francesco Jimenez Forteza for in depth discussions and feedback. We are also grateful to Krishnendu NV, Prayush Kumar, and Chris Van Den Broeck for discussions and valuable inputs. We thank Michael Pürrer for valuable comments on an earlier version of this manuscript. We thank the anonymous referee for their valuable feedback. We used the Holodeck computing cluster and the Atlas cluster at AEI Hannover for all the computation.

## DATA AVAILABILITY

The data that support the findings of this article are openly available [115,124], embargo periods may apply.

## APPENDIX A: FISHER MATRIX FORMALISM

In Sec. IV, we have applied the Fisher matrix formalism to quantify systematic biases, but we have assumed a familiarity with the topic. This section serves as an introduction to the formalism itself.

The framework in which Fisher matrix estimates are derived is based on a geometric perspective on PE. Using a metric on the parameter space, the Fisher matrix,

$$\Gamma_{\mu\nu} = \langle \partial_\mu h, \partial_\nu h \rangle, \quad (\text{A1})$$

one can derive estimates about various errors in the process.<sup>1</sup> Here,  $h$  is some fiducial waveform model, and  $\partial_\mu$  refers to the derivative with respect to the parameter  $\Theta^\mu$ . Equation (A1) is defined in terms of the noise-weighted inner product,

$$\langle a, b \rangle = 4\Re \int_{f_{\text{low}}}^{f_{\text{high}}} \frac{\tilde{a}(f)\tilde{b}^*(f)}{S_n(f)} df, \quad (\text{A2})$$

<sup>1</sup>As far as we can see, Ref. [125] was the first one to use these estimates. The related geometric interpretation was first applied to GW data analysis by Ref. [126].

which has already been used in Eq. (3). It is this inner product that several important notions in GW data analysis are based on. Specifically, we will use “match” to refer to the inner product optimized over relative time and phase shifts  $t_0, \phi_0$  between  $a, b$  and “overlap” to refer to the normalized match, where the result is divided by the product  $\|a\| \cdot \|b\|$ . A complementary notion is the “mis-match”, which is defined as  $1 - \text{match}$ .

The Fisher matrix formalism is a very popular tool because it allows to estimate biases in the source parameters  $\Theta$ , which we will denote as  $\Delta\Theta = \Theta_{\text{bf}} - \Theta_{\text{tr}}$ . The two parameters  $\Theta_{\text{tr}}, \Theta_{\text{bf}}$  are to be understood in the following context: say we have data that contains a signal produced by a model  $h_1$  (and potentially some noise on top of that), and we wish to run PE on this data using another model  $h_2$ ; then we call the parameters that  $h_1$  is evaluated in the “true parameters”  $\Theta_{\text{tr}}$  and the parameters obtained from PE “best-fitting” parameters  $\Theta_{\text{bf}}$ . What the Fisher matrix formalism does for us is that it allows to obtain an estimate of the parameter difference between true and recovered parameters without having to perform a full PE run. This difference will have two contributions: the systematic bias due to  $h_1 \neq h_2$  can be estimated as [125]

$$\Delta\Theta_{\text{sys}}^\mu = \sum_\nu (\Gamma^{-1})^{\mu\nu} \langle h_1 - h_2, \partial_\nu h_2 \rangle, \quad (\text{A3})$$

and the corresponding measurement uncertainty caused by noise (which is also called statistical bias) as [125]

$$\Delta\Theta_{\text{stat}}^\mu = \sum_\nu (\Gamma^{-1})^{\mu\nu} \langle n, \partial_\nu h_2 \rangle. \quad (\text{A4})$$

In reality, where  $n$  is not known, it is more common to work with the corresponding standard deviation [125],

$$\sigma_{\Delta\Theta_{\text{stat}}^\mu} = \sqrt{(\Gamma^{-1})^{\mu\mu}}. \quad (\text{A5})$$

These estimates are very valuable because they can be calculated without having to do an extra PE run. The right-hand side of both equations can be evaluated either in the true, injected parameters  $\Theta_{\text{tr}}$  or in the best-fitting parameters  $\Theta_{\text{bf}}$  that would be the result of a PE run using  $h_2$  on the signal  $h_1(\Theta_{\text{tr}})$ . To calculate the Fisher matrix in this context,  $h_2$  must be used.

However, there is also a caveat. The whole approach depends crucially on the validity of the linear signal approximation (LSA),

$$h_2(\Theta_{\text{tr}}) \simeq h_2(\Theta_{\text{bf}}) + \partial_\mu h_2(\Theta_{\text{bf}})(\Theta_{\text{tr}} - \Theta_{\text{bf}})^\mu. \quad (\text{A6})$$

Besides numerical issues with condition numbers of the Fisher matrix, which are also common problems, this is the major bottleneck of Fisher matrix estimates. If  $h_1$  and  $h_2$  exhibit large differences, they might produce

maximum-*a-posteriori* estimates  $\Theta_{\text{tr}}, \Theta_{\text{bf}}$  for which (A6) is not a good approximation. In this case, the accuracy of later applied estimates such as (A3) may be compromised.

Recently, the authors of [28] introduced a scheme to improve this issue, which we have adopted in this paper as well. The idea, which we only briefly review, is to replace the waveform difference  $h_1(\Theta_{\text{tr}}) - h_2(\Theta_{\text{tr}})$  by another difference  $h_1(\Theta_{\text{tr}}) - h_2(\tilde{\Theta}_{\text{tr}})$  in a procedure called alignment.<sup>2</sup> The new point  $\tilde{\Theta}_{\text{tr}}$  is defined in such a way that most of its components still coincide with the one of  $\Theta_{\text{tr}}$ ; only the ones that belong to parameters from a set  $S$  are changed in order to minimize the normalized inner product,

$$\frac{\langle h_1(\Theta_{\text{tr}}), h_2(\tilde{\Theta}_{\text{tr}}) \rangle}{\sqrt{\langle h_1(\Theta_{\text{tr}}), h_1(\Theta_{\text{tr}}) \rangle \langle h_2(\tilde{\Theta}_{\text{tr}}), h_2(\tilde{\Theta}_{\text{tr}}) \rangle}}. \quad (\text{A7})$$

In principle,  $S$  can contain an arbitrary combination of parameters because there is no restriction as to which parameters can be optimized over in (A7). From the motivation, it is natural to select parameters for which conventional differences might occur and as a basic (yet effective) version of  $S$ , we choose to include relative time and phase shifts  $t_0, \phi_0$  [so that (A7) coincides with our definition of the “overlap”]. This also allows a computationally very efficient optimization, exploiting that (A2) can be written as an inverse Fourier transform [127]. Of course, the estimate (A3) has to be changed accordingly, and the details of how to do this are presented in Sec. III. B. of Ref. [28] (the basic structure remains the same, which is why we do not cover it explicitly here).

## APPENDIX B: ADDITIONAL DETAILS OF SIMULATIONS AND PE RESULTS

In the examples we considered in Sec. V, we find that it is primarily the  $\delta\phi$  parameter (for both the parametrizations) that is constrained by the data. We obtain the prior distribution only for the  $\delta\tilde{A}$ . It can be understood because, in our simulations, we introduce 1%–2% relative errors in amplitude. It seems too low for the simulations to pick up variation of that order in the amplitude parameter. Moreover, a constant shift of  $\delta\tilde{A}$  is expected to be absorbed in distance posteriors. For the 2G PE analysis, we can use WF-error parametrizations (12), and (13) which uses only  $\delta\tilde{\phi}$  parameter. We choose to keep the  $\delta\tilde{A}$  in our parametrizations to account for most general cases of deviation.

In Fig. 13, we show the relative amplitude and phase errors between  $h_{\text{no-spin}}$  and 500 simulated  $h_{\text{prec}}$  GW strain

<sup>2</sup>Here, we assume estimate (A3) to be evaluated in  $\Theta_{\text{tr}}$ , without loss of generality (the discussion also applies when using  $\Theta_{\text{bf}}$ ).

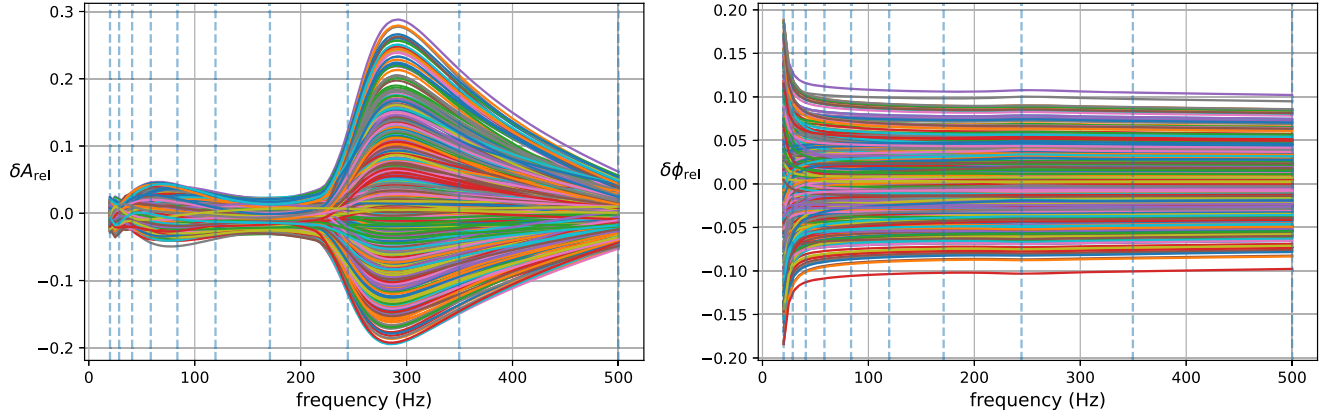


FIG. 13. For the simulations done in Sec. V C, we show the relative errors in amplitude and phase (between precessing and nonspinning waveform template) as function of frequency. For reference, we also show ten waveform nodal points as dashed vertical lines.

for the GW150914-like system. The details of the waveform parameters are given in Sec. V C. In PE, we use ten frequency nodal points with logarithmic

spacing in the range  $f \in [20, 500]$ . Distribution of  $\delta \tilde{A}(f_i), \delta \tilde{\phi}(f_i)$  at given frequency nodal point  $f_i$  can be used as prior distribution.

- 
- [1] J. Aasi *et al.* (LIGO Scientific Collaboration), *Classical Quantum Gravity* **32**, 074001 (2015).
  - [2] A. Buikema *et al.* (aLIGO Collaboration), *Phys. Rev. D* **102**, 062003 (2020).
  - [3] B. P. Abbott *et al.* (KAGRA, LIGO Scientific, and Virgo Collaborations), *Living Rev. Relativity* **19**, 1 (2016).
  - [4] F. Acernese *et al.* (Virgo Collaboration), *Classical Quantum Gravity* **32**, 024001 (2015).
  - [5] F. Acernese *et al.* (Virgo Collaboration), *Phys. Rev. Lett.* **123**, 231108 (2019).
  - [6] R. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration), *Phys. Rev. X* **13**, 041039 (2023).
  - [7] A. H. Nitz, S. Kumar, Y.-F. Wang, S. Kasta, S. Wu, M. Schäfer, R. Dhurkunde, and C. D. Capano, *Astrophys. J.* **946**, 59 (2023).
  - [8] D. Wadekar, J. Roulet, T. Venumadhav, A. K. Mehta, B. Zackay, J. Mushkin, S. Olsen, and M. Zaldarriaga, *arXiv:2312.06631*.
  - [9] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), *arXiv:2112.06861*.
  - [10] R. Abbott *et al.* (KAGRA, Virgo, and LIGO Scientific Collaborations), *Phys. Rev. X* **13**, 011048 (2023).
  - [11] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), *Astrophys. J.* **949**, 76 (2023).
  - [12] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Phys. Rev. Lett.* **121**, 161101 (2018).
  - [13] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Phys. Rev. X* **9**, 011001 (2019).
  - [14] C. D. Capano, I. Tews, S. M. Brown, B. Margalit, S. De, S. Kumar, D. A. Brown, B. Krishnan, and S. Reddy, *Nat. Astron.* **4**, 625 (2020).
  - [15] L. Blanchet, *Living Rev. Relativity* **17**, 2 (2014).
  - [16] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower, *Phys. Rev. Lett.* **96**, 111101 (2006).
  - [17] F. Pretorius, *Phys. Rev. Lett.* **95**, 121101 (2005).
  - [18] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, *Phys. Rev. Lett.* **96**, 111102 (2006).
  - [19] B. P. Abbott *et al.*, *Phys. Rev. D* **93**, 112004 (2016); **97**, 059901(A) (2018).
  - [20] M. Saleem *et al.*, *Classical Quantum Gravity* **39**, 025004 (2022).
  - [21] M. Punturo *et al.*, *Classical Quantum Gravity* **27**, 084007 (2010).
  - [22] S. Hild *et al.*, *Classical Quantum Gravity* **28**, 094013 (2011).
  - [23] M. Evans *et al.*, *arXiv:2109.09882*.
  - [24] V. Srivastava, D. Davis, K. Kuns, P. Landry, S. Ballmer, M. Evans, E. D. Hall, J. Read, and B. S. Sathyaprakash, *Astrophys. J.* **931**, 22 (2022).
  - [25] M. Evans *et al.*, *arXiv:2306.13745*.
  - [26] L. Lindblom, B. J. Owen, and D. A. Brown, *Phys. Rev. D* **78**, 124020 (2008).
  - [27] M. Pürrer and C.-J. Haster, *Phys. Rev. Res.* **2**, 023151 (2020).
  - [28] A. Dhani, S. Völkel, A. Buonanno, H. Estelles, J. Gair, H. P. Pfeiffer, L. Pompili, and A. Toubiana, *Phys. Rev. X* **15**, 031036 (2025).
  - [29] S. Kumar, A. H. Nitz, and X. J. Forteza, *Astrophys. J.* **982**, 67 (2025).

- [30] A. Gupta *et al.*, *SciPost Phys. Commun. Rep.* **5**, 1 (2025).
- [31] R. S. Chandramouli, K. Prokupa, E. Berti, and N. Yunes, *Phys. Rev. D* **111**, 044026 (2025).
- [32] J. S. Read, *Classical Quantum Gravity* **40**, 135002 (2023).
- [33] R. Abbott *et al.* (KAGRA, Virgo, and LIGO Scientific Collaborations), *Phys. Rev. X* **13**, 041039 (2023).
- [34] G. Ashton and S. Khan, *Phys. Rev. D* **101**, 064037 (2020).
- [35] A. Z. Jan, A. B. Yelikar, J. Lange, and R. O'Shaughnessy, *Phys. Rev. D* **102**, 124069 (2020).
- [36] G. Ashton and T. Dietrich, *Nat. Astron.* **6**, 961 (2022).
- [37] C. Hoy, *Phys. Rev. D* **106**, 083003 (2022).
- [38] A. Puecher, A. Samajdar, G. Ashton, C. Van Den Broeck, and T. Dietrich, *Phys. Rev. D* **109**, 023019 (2024).
- [39] A. Samajdar and T. Dietrich, *Phys. Rev. D* **98**, 124030 (2018).
- [40] A. Samajdar and T. Dietrich, *Phys. Rev. D* **100**, 024046 (2019).
- [41] C. Hoy, S. Akcay, J. Mac Uilliam, and J. E. Thompson, *Nat. Astron.* **9**, 1256 (2025).
- [42] C. J. Moore and J. R. Gair, *Phys. Rev. Lett.* **113**, 251101 (2014).
- [43] L. Pompili, A. Buonanno, and M. Pürrer, *Phys. Rev. D* **112**, 044062 (2025).
- [44] R. Bachhar, M. Pürrer, and S. R. Green, *Phys. Rev. D* **111**, 084050 (2025).
- [45] S. Khan, *Phys. Rev. D* **109**, 104045 (2024).
- [46] C. B. Owen, C.-J. Haster, S. Perkins, N. J. Cornish, and N. Yunes, *Phys. Rev. D* **108**, 044018 (2023).
- [47] S. L. Detweiler and L. H. Brown, Jr., *Phys. Rev. D* **56**, 826 (1997).
- [48] Z. Bern, C. Cheung, R. Roiban, C.-H. Shen, M. P. Solon, and M. Zeng, *Phys. Rev. Lett.* **122**, 201603 (2019).
- [49] C. V. Vishveshwara, *Nature (London)* **227**, 936 (1970).
- [50] S. Chandrasekhar and S. L. Detweiler, *Proc. R. Soc. A* **344**, 441 (1975).
- [51] F. Ohme, *Classical Quantum Gravity* **29**, 124002 (2012).
- [52] P. Ajith *et al.*, *Phys. Rev. D* **77**, 104017 (2008); **79**, 129901 (E) (2009).
- [53] P. Ajith *et al.*, *Phys. Rev. Lett.* **106**, 241101 (2011).
- [54] L. Santamaria *et al.*, *Phys. Rev. D* **82**, 064016 (2010).
- [55] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044007 (2016).
- [56] S. Khan, K. Chatziioannou, M. Hannam, and F. Ohme, *Phys. Rev. D* **100**, 024059 (2019).
- [57] S. Khan, F. Ohme, K. Chatziioannou, and M. Hannam, *Phys. Rev. D* **101**, 024056 (2020).
- [58] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044006 (2016).
- [59] L. London, S. Khan, E. Fauchon-Jones, C. García, M. Hannam, S. Husa, X. Jiménez-Forteza, C. Kalaghatgi, F. Ohme, and F. Pannarale, *Phys. Rev. Lett.* **120**, 161102 (2018).
- [60] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, *Phys. Rev. Lett.* **113**, 151101 (2014).
- [61] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, G. Pratten, A. Ramos-Buades, M. Mateu-Lucena, and R. Jaume, *Phys. Rev. D* **102**, 064002 (2020).
- [62] G. Pratten *et al.*, *Phys. Rev. D* **103**, 104056 (2021).
- [63] G. Pratten, S. Husa, C. García-Quirós, M. Colleoni, A. Ramos-Buades, H. Estelles, and R. Jaume, *Phys. Rev. D* **102**, 064001 (2020).
- [64] H. Estellés, A. Ramos-Buades, S. Husa, C. García-Quirós, M. Colleoni, L. Haegel, and R. Jaume, *Phys. Rev. D* **103**, 124060 (2021).
- [65] H. Estellés, M. Colleoni, C. García-Quirós, S. Husa, D. Keitel, M. Mateu-Lucena, M. d. L. Planas, and A. Ramos-Buades, *Phys. Rev. D* **105**, 084040 (2022).
- [66] S. Ghosh, P. Kolitsidou, and M. Hannam, *Phys. Rev. D* **109**, 024061 (2024).
- [67] J. E. Thompson, E. Hamilton, L. London, S. Ghosh, P. Kolitsidou, C. Hoy, and M. Hannam, *Phys. Rev. D* **109**, 063012 (2024).
- [68] A. Buonanno and T. Damour, *Phys. Rev. D* **59**, 084006 (1999).
- [69] Y. Pan, A. Buonanno, M. Boyle, L. T. Buchman, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, *Phys. Rev. D* **84**, 124052 (2011).
- [70] Y. Pan, A. Buonanno, A. Taracchini, L. E. Kidder, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi, *Phys. Rev. D* **89**, 084006 (2014).
- [71] A. Taracchini *et al.*, *Phys. Rev. D* **89**, 061502 (2014).
- [72] A. Bohé *et al.*, *Phys. Rev. D* **95**, 044028 (2017).
- [73] R. Coste, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, *Phys. Rev. D* **98**, 084028 (2018).
- [74] S. Ossokine *et al.*, *Phys. Rev. D* **102**, 044055 (2020).
- [75] L. Pompili *et al.*, *Phys. Rev. D* **108**, 124035 (2023).
- [76] A. Gamboa *et al.*, *Phys. Rev. D* **112**, 044038 (2025).
- [77] T. Damour, A. Nagar, and S. Bernuzzi, *Phys. Rev. D* **87**, 084035 (2013).
- [78] R. Gamba, S. Bernuzzi, and A. Nagar, *Phys. Rev. D* **104**, 084058 (2021).
- [79] A. Nagar, S. Bernuzzi, D. Chiaramello, V. Fantini, R. Gamba, M. Panzeri, and P. Retegno, *Phys. Rev. D* **111**, 064050 (2025).
- [80] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, *Phys. Rev. X* **4**, 031006 (2014).
- [81] J. Blackman, S. E. Field, M. A. Scheel, C. R. Galley, D. A. Hemberger, P. Schmidt, and R. Smith, *Phys. Rev. D* **95**, 104023 (2017).
- [82] J. Blackman, S. E. Field, M. A. Scheel, C. R. Galley, C. D. Ott, M. Boyle, L. E. Kidder, H. P. Pfeiffer, and B. Szilágyi, *Phys. Rev. D* **96**, 024058 (2017).
- [83] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, L. E. Kidder, and H. P. Pfeiffer, *Phys. Rev. D* **99**, 064045 (2019).
- [84] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, D. Gerosa, L. C. Stein, L. E. Kidder, and H. P. Pfeiffer, *Phys. Rev. Res.* **1**, 033015 (2019).
- [85] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Classical Quantum Gravity* **37**, 055002 (2020).
- [86] R. Prix, *Classical Quantum Gravity* **24**, S481 (2007).
- [87] S. A. Usman *et al.*, *Classical Quantum Gravity* **33**, 215004 (2016).
- [88] B. Allen, *Phys. Rev. D* **104**, 042005 (2021).
- [89] J. Veitch *et al.*, *Phys. Rev. D* **91**, 042003 (2015).
- [90] C. M. Biwer, C. D. Capano, S. De, M. Cabero, D. A. Brown, A. H. Nitz, and V. Raymond, *Publ. Astron. Soc. Pac.* **131**, 024503 (2019).

- [91] S. Mozzon, L. K. Nuttall, A. Lundgren, T. Dent, S. Kumar, and A. H. Nitz, *Classical Quantum Gravity* **37**, 215014 (2020).
- [92] O. Edy, A. Lundgren, and L. K. Nuttall, *Phys. Rev. D* **103**, 124061 (2021).
- [93] LIGO-T2100313-v3: LIGO and Virgo Calibration Uncertainty (O1, O2 and O3)—[dcc.ligo.org](https://dcc.ligo.org/T2100313/public), <https://dcc.ligo.org/T2100313/public> (Accessed 10-02-2025).
- [94] C. Cahillane, J. Betzwieser, D. A. Brown, E. Goetz, E. D. Hall, K. Izumi, S. Kandhasamy, S. Karki, J. S. Kissel, G. Mendell, R. L. Savage, D. Tuyenbayev, A. Urban, A. Viets, M. Wade, and A. J. Weinstein, *Phys. Rev. D* **96**, 102001 (2017).
- [95] L. Sun *et al.*, *Classical Quantum Gravity* **37**, 225008 (2020).
- [96] L. Sun *et al.*, [arXiv:2107.00129](https://arxiv.org/abs/2107.00129).
- [97] F. Acernese *et al.* (Virgo Collaboration), *Classical Quantum Gravity* **39**, 045006 (2022).
- [98] F. Acernese *et al.* (Virgo Collaboration), *Classical Quantum Gravity* **35**, 205004 (2018).
- [99] W. M. Farr, B. Farr, and T. Littenberg, Modelling calibration errors in CBC waveforms, Technical Report, LIGO-T1400682, LIGO Project, 2015, <https://dcc.ligo.org/LIGO-T1400682/public>.
- [100] C. Cahillane *et al.* (LIGO Scientific Collaboration), *Phys. Rev. D* **96**, 102001 (2017).
- [101] G. Ashton *et al.*, *Astrophys. J. Suppl. Ser.* **241**, 27 (2019).
- [102] B. Edelman *et al.*, *Phys. Rev. D* **103**, 042004 (2021).
- [103] C. R. Harris *et al.*, *Nature (London)* **585**, 357 (2020).
- [104] P. Virtanen *et al.* (SciPy 1.0 Contributors), *Nat. Methods* **17**, 261 (2020).
- [105] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), *Phys. Rev. D* **102**, 043015 (2020).
- [106] U. Dupletsa, J. Harms, K. K. Y. Ng, J. Tissino, F. Santoliquido, and A. Cozzumbo, *Phys. Rev. D* **111**, 024036 (2025).
- [107] F. Iacovelli, M. Mancarella, S. Foffa, and M. Maggiore, *Astrophys. J.* **941**, 208 (2022).
- [108] F. Iacovelli, M. Mancarella, S. Foffa, and M. Maggiore, *Astrophys. J. Suppl. Ser.* **263**, 2 (2022).
- [109] Y. Li, I. S. Heng, M. L. Chan, C. Messenger, and X. Fan, *Phys. Rev. D* **105**, 043010 (2022).
- [110] B. P. Abbott *et al.*, *Phys. Rev. X* **9**, 031040 (2019).
- [111] R. Abbott *et al.*, *Phys. Rev. X* **11**, 021053 (2021).
- [112] R. Abbott *et al.* (The LIGO Scientific Collaboration and the Virgo Collaboration), *Phys. Rev. D* **109**, 022001 (2024).
- [113] M. Vallisneri, *Phys. Rev. D* **77**, 042001 (2008).
- [114] J. S. Speagle, *Mon. Not. R. Astron. Soc.* **493**, 3132 (2020).
- [115] S. Kumar, M. Melching, and F. Ohme, PyCBC plugin for waveform errors, [https://github.com/gwastro/pycbc\\_wferrors\\_plugin](https://github.com/gwastro/pycbc_wferrors_plugin) (2025).
- [116] LIGO Scientific Collaboration, LIGO Algorithm Library—LALSuite, free software (GPL) (2018).
- [117] K. Paul and C. K. Mishra, *Phys. Rev. D* **108**, 024023 (2023).
- [118] Q. Henry and M. Khalil, *Phys. Rev. D* **108**, 104016 (2023).
- [119] A. Ramos-Buades, A. Buonanno, M. Khalil, and S. Ossokine, *Phys. Rev. D* **105**, 044035 (2022).
- [120] K. Paul, A. Maurya, Q. Henry, K. Sharma, P. Satheesh, Divyajyoti, P. Kumar, and C. K. Mishra, *Phys. Rev. D* **111**, 084074 (2025).
- [121] Q. Hu and J. Veitch, *Phys. Rev. D* **106**, 044042 (2022).
- [122] P. Hammond, I. Hawke, and N. Andersson, *Phys. Rev. D* **104**, 103006 (2021).
- [123] F. Gittins, R. Matur, N. Andersson, and I. Hawke, *Phys. Rev. D* **111**, 023049 (2025).
- [124] S. Kumar, M. Melching, and F. Ohme, Raw data and configuration files for analysis in [arXiv:2502.17400](https://arxiv.org/abs/2502.17400), accounting for waveform systematic errors (2025), <https://zenodo.org/records/14975178>.
- [125] C. Cutler and M. Vallisneri, *Phys. Rev. D* **76**, 104018 (2007).
- [126] R. Balasubramanian, B. S. Sathyaprakash, and S. V. Dhurandhar, *Phys. Rev. D* **53**, 3033 (1996).
- [127] B. Allen, *Phys. Rev. D* **71**, 062001 (2005).