Gravitational waveform accuracy requirements for future ground-based detectors

Michael Pürrer [1,*] and Carl-Johan Haster [2,3,†]

[1] *Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, Potsdam 14476, Germany*

[2] *LIGO Laboratory, Massachusetts Institute of Technology, 185 Albany Street, Cambridge, Massachusetts 02139, USA*

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

(Received 20 December 2019; accepted 1 April 2020; published 11 May 2020)

Future third-generation (3G) ground-based gravitational wave (GW) detectors, such as the Einstein Telescope and Cosmic Explorer, will have unprecedented sensitivities enabling studies of the entire population of stellar mass binary black hole coalescences in the universe, while the A+ and Voyager upgrades to current detectors will significantly improve over advanced LIGO and Virgo design sensitivities. To infer binary parameters from a GW signal we require accurate models of the gravitational waveform as a function of black hole masses, spins, etc. Such waveform models are built from numerical relativity (NR) simulations and/or semianalytical expressions in the inspiral. We investigate the limits of the current waveform models and study at what detector sensitivity these models will yield unbiased parameter inference for loud "golden" binary black hole systems, what biases we can expect beyond these limits, and what implications such biases will have for GW astrophysics. For 3G detectors we find that the mismatch error for semianalytical models needs to be reduced by at least *three orders of magnitude* and for NR waveforms by *one order of magnitude*. We show that typical biases in units of standard deviations for the mass-ratio and effective aligned-spin will be of order unity for 2G design sensitivity and will reach several tens for 3G networks. In addition, we show that for a population of one hundred high mass precessing binary black holes, measurement errors sum up to a sizable population bias, about 10–30 times larger than the sum of 90% credible intervals for chirp mass, mass-ratio, effective aligned, and precessing spin parameters. Furthermore, we demonstrate that the residual signal between the GW data recorded by a detector and the best fit template waveform obtained by parameter inference analyses can have significant signal-to-noise ratio and can lead to Bayes factors as high as 10^{11} between a coherent and an incoherent wavelet model for the population events. This coherent power left in the residual could lead to the observation of erroneous deviations from general relativity. To address these issues and be ready to reap the scientific benefits of 3G GW detectors in the 2030s, waveform models that are significantly more physically complete and accurate need to be developed in the next decade along with major advances in efficiency and accuracy of NR codes.

DOI: 10.1103/PhysRevResearch.2.023151

I. INTRODUCTION

Observations of gravitational waves (GWs) from coalescing compact object binaries have revolutionized our knowledge about the universe and provided access to astrophysics previously outside our grasp [1]. These observations were made possible by the construction and operation of a network of GW detectors, Advanced LIGO [2], Advanced Virgo [3], and KAGRA [4]. As expected from a relatively young field of observational astrophysics, there are still significant technological improvements within reach [5–7], increasing the sensitivity of both current generation GW detectors over the next ∼5 years [8] in addition to paving the way for next-generation ground-based facilities [9–12] as well as space-based observatories [13,14], all planned to be operational in the early 2030s. This will allow for direct observation of all stellar-mass binary black hole (BBH) mergers throughout the cosmological history of the universe [15] and will enable unprecedented and unique science in extreme gravity and fundamental physics [16]. Whereas the vast majority of observed GW signals will have originated at large cosmological distances [17,18], binaries from the currently observable volume of the universe will, as the detector sensitivities improve (see Fig. 1), be observable with increasing fidelity. GW models are crucial for elucidating the astrophysical properties of compact binaries. For this increase in the information available for a typical BBH observation, these models need to satisfy stricter accuracy requirements. The currently available model waveforms, which approximates the solutions to the two-body problem in general relativity (GR), have been shown to be sufficiently accurate to not cause any systematic biases in the recovered parameters (masses, spins, location in the universe, etc.) for BBHs observed so far [19]. These "golden binaries," stellar-mass BBHs like GW150914 (the first direct

*michael.puerrer@aei.mpg.de
†haster@mit.edu

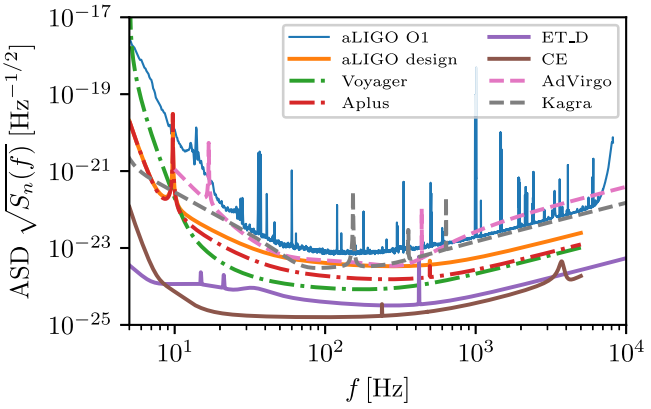

FIG. 1. Evolution of sensitivity of GW interferometric detectors. Text data files for the PSDs or amplitude spectral densitys (ASDs) can be found in References [45,46] under the names given in this table.

GW detection [20]) observed at high signal-to-noise ratio (SNR) and with high fidelity, have also allowed to perform previously inaccessible tests of GR [21] and the robustness of the waveform models used. As the GW detectors improve, the expected SNR for BBHs from the local universe will increase correspondingly, thus reducing the statistical uncertainties in the recovered source parameters. As the statistical uncertainties approach the inherent systematic uncertainties of the GW approximant models, parameter biases will eventually appear and reduce the reliability of future GW observations.

In this study we want to investigate the appearance and significance of these parameter biases, their connection to the accuracy of the GW models used and what relevance the biases will have on future astrophysical statements based on high SNR observations of BBH systems [22]. In addition to possible biases in the source parameters of the BBH, the use of GW observations as means to test GR puts even more stringent requirements on GW model accuracy [21,23–25]. If there are effects of beyond-GR theories embossed on the "raw" GR waveform, then these effects will be scrambled by any residual signal left by an inaccurate GW model and thus will limit the strength of the GR test. Even worse, inaccurate GW models may lead to erroneous results claiming deviations from GR. Many of these analyses, both parameter estimation (PE) studies and tests of GR, are strongly dependent on robust observation of the two polarization states of a GW signal as described by GR [26]. This is primarily done by requiring a coherent observation of a given GW signal in more than one detector [27–29], which is also crucial for accurate and reliable localization of the GW source in the universe. Whereas the BBHs investigated here are not expected to produce any observable counterpart [30–32], precise localization is crucial for cosmology studies [33–37] as well as for inferring the parameters of the BBHs in their source frame [38].

In addition to biases caused by inaccurate GW models, the data generated by the detectors themselves carry inherent uncertainties originating from the calibration process applied to the raw detector output [39] as well as imprecise modeling assumptions for the noise processes of a given detector system manifesting as inaccuracies in the estimated power spectral density (PSD) for the analysed data [40,41]. These types of uncertainties can however already be quantified, and thus incorporated into the PE infrastructure allowing their effects to be marginalised out from the final inferred parameter distributions [1,42–44]. The marginalization over calibration uncertainties, and similarly the marginalization over eventual uncertainties in the noise PSD estimation, is primarily expected to broaden the recovered posterior distributions thus effectively absorbing any misestimate in the GW amplitude or phase irrespective of whether it originated from uncertainties in the data itself or from the assumed waveform model.

The rest of the paper presents the details of our study. In Sec. II we describe the GW models used, together with details on the analysis methods. In Sec. III we report our findings on the analysis of individual "golden binary" BBH signals, including requirements on the accuracy of the GW models and the consequences any inaccuracies will entail. In Sec. IV we explore a population of BBH observations, and what effects GW model accuracy will have on the properties of the inferred population. Finally in Sec. V we discuss our findings and present an outlook for how to tackle the issues we have presented.

II. METHODOLOGY

In this section we introduce the methods we use to study the impact of waveform inaccuracies on measurements of compact binary parameters from GWs in Secs. III and IV. We first discuss common data analysis tools for GW waveforms in Sec. II A and numerical relativity (NR) waveforms, the most accurate waveforms we have available for the solutions of the two-body problem in GR, in Sec. II B. In addition to NR waveforms, we also use post-Newtonian–numerical relativity (PN-NR) hybrid waveforms as described in Sec. II C to more fully fill the band of more sensitive detectors to lower frequencies as mock signals in this study. In Sec. II D, we discuss fast, but approximate semianalytic models of the GWs emitted from compact binaries. These models are crucial to infer binary properties with Bayesian parameter estimation methods for single events and populations of binaries, as discussed in Sec. II E.

A. Gravitational waveforms and overlaps

Gravitational waveforms are often decomposed in a basis of spherical harmonics ${}_{-2}Y_{lm}$. The two GW polarizations can be expanded into modes as

$$h_+ - ih_\times = \sum_{\ell,m} h_{\ell m}\, {}_{-2}Y_{lm}. \tag{1}$$

We define the overlap, or match, between two waveforms h_1 and h_2 as

$$\mathcal{O}(h_1, h_2) := \frac{\langle h_1|h_2\rangle}{\sqrt{\langle h_1|h_1\rangle\langle h_2|h_2\rangle}}, \tag{2}$$

where

$$\langle h_1|h_2\rangle = 4\mathrm{Re}\int_{f_{\rm low}}^{f_{\rm high}} \frac{\tilde{h}_1(f)\tilde{h}_2^*(f)}{S_n(f)} df \tag{3}$$

TABLE I. Binary configurations studied in Sec. III. We indicate the SXS ID [47] of the SpEC NR simulations, the total mass and chirp mass in the source frame, the mass-ratio $q = m_2/m_1 \leqslant 1$, the dimensionless spin vectors $\vec{\chi}_i = \vec{S}_i/m_i^2$ of the BHs, the effective aligned spin and effective precession spin, and the inclination angle between the total angular momentum $\vec{J}$ and the line of sight $\vec{N}$. Signals are hybridized with `SpinTaylorT1`. Spin vectors are defined at a reference frequency of 30 Hz. We select the remaining common parameters to be ra $= 1.949725$, dec $= -1.261573$ (radians), a luminosity distance of $d_L = 562.59$Mpc (which corresponds to a redshift of about $z = 0.115$), a polarization angle $\psi = 1.4289$. The GPS time at the geocenter was 1126259642.413 s and coalescence phase $\phi_{\rm coa} = 0$.

Configuration	$M^{\rm src}_{\rm tot}/M_\odot$	$\mathcal{M}^{\rm src}/M_\odot$	q	$\vec{\chi}_1$	$\vec{\chi}_2$	$\chi_{\rm eff}$	$\chi_{\rm p}$	$\theta_{\rm JN}$
SXS_BBH_0308	66.4555	28.7443	0.8143	$(-0.1407, 0.0225, 0.3053)$	$(-0.2209, 0.3075, -0.5580)$	-0.0822	0.2994	2.7454
SXS_BBH_0104	66.4555	24.3406	0.3333	$(-0.0550, -0.0144, 0.4966)$	$(-0.2737, -0.4173, 0.0112)$	0.3753	0.1442	1.0839

represents a noise-weighted inner product with PSD S_n, and * denotes complex conjugation. We also maximize the overlap over a time and phase-shift between the two waveforms. We often quote the mismatch, $1 - \mathcal{O}(h_1, h_2)$ instead of the overlap.

B. Numerical relativity waveforms

We use two NR simulations of binary black hole coalescences by the SXS Collaboration [47] using the spectral Einstein code (SpEC) [48]. While there are now thousands of simulations available we pick a first simulation which is very representative of the mass-ratio and spin distribution of BBH systems observed by LIGO and Virgo so far [1], and a second binary with parameters that lie closer to the edge of the measured 90% credible regions, toward higher mass-ratio and effective spin. It is there that we expect current waveform models to disagree with each other more strongly than close to an equal-mass nonspinning binary. The first simulation, SXS_BBH_0308 [47,49], was performed at parameters inferred from the LIGO PE analysis of GW150914 with semianalytic waveform models [20,50] and was subsequently used to study possible effects of waveform systematics on the inferred parameters [19,51]. The waveform describes a nearly equal mass binary with small effective aligned spin and moderate precession (see Table I). The waveform accumulates 12.6 orbits and a length of $2822M$ in time before the formation of a common horizon. The mismatch between simulations at different resolutions at the total mass of GW150914 with aLIGO design sensitivity is $\sim 2 \times 10^{-4}$. We use the highest resolution available, Lev5.

The second simulation, SXS_BBH_0104 [52,53], is at mass-ratio 1:3 and has some effective aligned and precession spin. Systems at this mass-ratio still lie within the population posterior for the mass-ratio that has been found in the LIGO and Virgo O1 and O2 analysis [1]. The waveform accumulates 21.9 orbits and a length of $5192M$ before the formation of a common horizon. There is only a single resolution, Lev5, available for this simulation. An estimate for the mismatch for a simulation using similar technology (SXS_BBH_0053 [53,54]) gives $\sim 10^{-3}$ at the total mass of GW150914 with aLIGO design sensitivity.

These waveforms are for quasi-circular inspirals and mergers of BBHs. Since initial conditions are not exactly known, there is a low amount of residual eccentricity in these simulations. For SXS_BBH_0308 eccentricity at the relaxed time is estimated to be ~ 0.0005 while for SXS_BBH_0104 it is ~ 0.001. We do not consider the effect of eccentricity in the waveform in this study. The relevance of eccentricity for waveform systematics is currently not very well understood. While a few inspiral-merger-ringdown eccentric waveform models have been constructed so far for nonspinning [55,56] and aligned-spin binaries [57,58], only a single parameter estimation study [59] has been carried out so far. The construction and tuning to NR simulations of the model in Ref. [57] is unfortunately not quite up to date compared to current BBH models. No detailed systematics study has been carried out and it is currently not known how neglecting eccentricity would compete with other sources of systematics in terms of parameter bias.

C. Hybrid waveforms

We use an extension of the `GWFrames` [60,61] package to hybridize NR with post-Newtonian (PN) waveforms. First we read in an NR waveform and its horizon data (i.e., the spins and orbital track data computed from the apparent horizon finder). We generate a PN waveform at the physical parameters of the NR configuration and align it by shifting in time and attitude to match the NR waveform. The waveform modes and the quaternions describing the motion of the inertial frame are then blended over a hybridization region in time. More details about the procedure and the hybrid waveforms are given in the Appendix.

Estimates of the accuracy of hybrid PN-NR waveforms are difficult to obtain. Hybrid errors are expected to be significantly higher than for pure NR waveforms due to errors in the PN part of the waveform and additional errors from smoothly combining the PN and NR waveform modes over a blending window in time [62–65]. We show in the Appendix that hybridization errors are lower than NR error estimates for the simulations considered in this study. Semianalytical waveform models usually have good accuracy in the inspiral and are less accurate near merger. Therefore, the PN-NR hybrids used as mock signals in this study should be much more accurate than the semianalytic waveform models described in Sec. II D which we use as template waveforms.

D. Waveform models

In this study we use two fast frequency domain waveform models as template waveforms. These are the `IMRPhenomPv2` [66,67] and `SEOBNRv4_ROM` [68] inspiral-merger-ringdown (IMR) models.

`IMRPhenomPv2` uses the aligned-spin `IMRPhenomD` [69,70] model as a base waveform in the coprecessing frame and

twists up its $(2, \pm 2)$ modes with a PN prescription of the Euler angles that describe the motion of the inertial frame for precessing black hole binaries, thus generating all $\ell = 2$ modes [71,72]. The model also assumes that the opening angle of the precession cone is small [67] which make it most suitable for binaries with small to moderate precession and moderate mass-ratios. The model has been shown to be smooth [73] up to mass-ratio $q \sim 1/4$.

`SEOBNRv4_ROM` is a frequency domain reduced order model of the time domain `SEOBNRv4` effective-one-body model [68] using the methodology developed in Refs. [74,75]. The model describes the $(2, \pm 2)$ modes for nonprecessing binaries and can be used for a wide range in mass-ratio and BH spin magnitudes up to maximal spin.

Both `IMRPhenomD` which underlies `IMRPhenomPv2` and `SEOBNRv4` have been tuned to NR waveforms in the nonprecessing sector. While more complete models in terms of precession are available [76–78] we were not able to use them for this computationally demanding study because we could not obtain converged posterior distributions in time. Models that also include higher harmonics [79] or are computationally more efficient [80] are now becoming available.

In the population study described in Sec. IV we use `NRSur7dq2` to represent the population of astrophysical signals [81]. These signals were stochastically drawn and thus we could not use NR simulations which are only available at specific points in parameter space. The `NRSur7dq2` NR-surrogate model is however a very good approximation to NR waveforms. It describes generic precessing systems with mass-ratios up to $q = 1/2$ and spin magnitudes of 0.8. `NRSur7dq2` is built from multiple surrogates that model waveform mode combinations in the co-orbital frame, the averaged frequency of the $(2, \pm 2)$ modes in the coprecessing frame, and the frame motion through the right-hand sides of the precession equations [82–85]. We intended to also use `NRSur7dq2` as a template waveform for the study discussed in Sec. III, but, while being very accurate, this model has a limited length and this severely limits the mass space that can be explored to high mass systems and high starting frequencies.

E. Bayesian parameter estimation

The inference of the source parameters $\vec{\theta}$ of a GW signal is expressed as a posterior probability density function (PDF) $p[\vec{\theta}|d(t)]$ as part of a PE analysis given the data $d(t)$ recorded from the detectors. Through application of Bayes' theorem, $p[\vec{\theta}|d(t)]$ is directly proportional to the likelihood $\mathcal{L}[d(t)|\vec{\theta}]$ of observing the data given an assumed waveform model $h(t;\vec{\theta})$, in turn characterized by the source parameters $\vec{\theta}$, together with the prior probability $\pi(\vec{\theta})$.

For the analysis of the "golden binaries" in Sec. III this prior is defined to be uniform over the two-dimensional space defining the masses of the binary objects, m_1 and m_2 (with $m_1 \geqslant m_2$), as observed in the rest frames of the GW detectors. The dimensionless spins of the BHs are assumed to follow a prior uniform in spin magnitude (between 0 and 1) allowing for isotropic and uncorrelated directions of the two black hole spins. We also assume an isotropic prior for the location of the GW on the sky, and a distance prior corresponding to a homogenous rate density in the nearby universe. For these analyses, we disregard any cosmological corrections to the rate density which for the redshifts explored ($z \sim 0.1$) are expected to be negligible. The orientation of the binary follows a prior probability uniform in the polarization angle ψ and in the cosine of θ_{JN}, the angle between the total angular momentum $\mathbf{J}$ and the line of sight $\mathbf{N}$. The parameter space defined by $\pi(\vec{\theta})$ is, for the golden binaries analysed in Sec. III, explored stochastically using a Markov chain Monte Carlo code implemented as part of the LALINFERENCE package [43,86] available as part of the LSC Algorithm Library (LAL) [87].

For the analysis of the BBH population in Sec. IV the BILBY inference package was used [88,89] exploring the parameter space using the Nested Sampling algorithm DYNESTY [90]. Here, similar parametrizations and prior assumptions as for the analysis in Sec. III were made. The analyses however differ in their assumptions over BH masses, here using a prior uniform in the binary chirp mass $\mathcal{M} = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$ and the asymmetric mass ratio $q = m_2/m_1$ as well as assuming a prior on distance that is uniform in comoving volume. The different prior choices between Secs. III and IV are not expected to have significant impact on the recovered parameters, or on the conclusions about waveform accuracy requirements based on this inference.

In a multidetector PE analysis we project the signal and template waveforms on the interferometric GW detectors and compute the strain from the waveform polarizations ($+$ and $\times$) and their corresponding detector antenna pattern functions [91]

$$h(t;\vec{\theta}) = h_+(t;\vec{\theta})F_+(\mathrm{ra}, \mathrm{dec}, \psi) + h_\times(t;\vec{\theta})F_\times(\mathrm{ra}, \mathrm{dec}, \psi). \tag{4}$$

As the focus of this study is on effects of accuracy of the waveform themselves, the signal waveforms representing the true GW signals are added to a time series containing no noise, as the standard assumption of Gaussian noise could introduce random biases in the recovered parameters. The true GW strain $h(t;\vec{\theta})$ as emitted by the GW source may however differ from $h^M(t;\vec{\theta})$, the strain measured by the detectors, due to uncertainties in the calibration of the detectors and their recording of the GW strain [39,92]. We can model the relation between the measured and true strain as

$$\tilde{h}^M(f;\vec{\theta}) = \tilde{h}(f;\vec{\theta})[1 + \delta A(f;\vec{\theta}^{\mathrm{cal}})] \exp i\,\delta\phi(f;\vec{\theta}^{\mathrm{cal}}), \tag{5}$$

for $\tilde{h}(f;\vec{\theta})$ and $\tilde{h}^M(f;\vec{\theta})$ where the tilde denotes the Fourier transforms of the time-domain strain $h(t;\vec{\theta})$ and $h^M(t;\vec{\theta})$, respectively. The uncertainty in the strain amplitude and phase, caused by uncertainties in the detector calibration, are characterized by the terms $\delta A(f;\vec{\theta}^{\mathrm{cal}})$ and $\delta\phi(f;\vec{\theta}^{\mathrm{cal}})$ that are nominally expected to vary both across the bandwidth used for the observation as well as over time from observation to observation. The frequency-dependent correction factors are modelled as cubic splines with nodes spaced uniformly in $\log f$, each with an independent δA and $\delta\phi$ parameter [93] which are then numerically marginalised over. For this study, we assume zero-mean Gaussian priors on δA with a standard deviation of 1% (5% for the O1 analysis) and for $\delta\phi$ a standard deviation of 1° (5° for the O1 analysis). The magnitude of these amplitude and phase uncertainties are consistent with

the performance of the LIGO detectors during O1 [44,94,95], and the predicted calibration uncertainties for future detector configurations [96,97].

Hierarchical inference

For the population study detailed in Sec. IV, the BILBY inference package [88,89] was used for both the analysis of individual BBHs as well as the subsequent inference on their population parameters. Following the analysis of each individual GW signal assumed to be part of the observed population, their joint population properties, here a single parameter α, can be inferred as a hyperposterior [98],

$$p_{\rm tot}(\alpha|\vec{d}) = \frac{\mathcal{L}_{\rm tot}(\vec{d}|\alpha)\,\pi(\alpha)}{\int d\alpha\,\mathcal{L}_{\rm tot}(\vec{d}|\alpha)\,\pi(\alpha)}, \tag{6}$$

where $\mathcal{L}_{\rm tot}(\vec{d}|\alpha)$ is the hyperlikelihood, $\pi(\alpha)$ is the hyperprior, $\vec{d}$ is a collection of data for N independent events drawn from the injection distribution. We write the injection prior as $\pi(\theta|\alpha)$ and our goal is to estimate the hyperposterior which in turn relies on a hyperlikelihood that can be written as

$$\mathcal{L}_{\rm tot}(\vec{d}|\alpha) = \prod_i^N \frac{\mathcal{Z}_\emptyset(d_i)}{n_i} \sum_k^{n_i} \frac{\pi\left(\theta_i^k|\alpha\right)}{\pi\left(\theta_i^k|\emptyset\right)}, \tag{7}$$

where $\pi(\theta_i^k|\emptyset)$ denotes the default prior that is used to perform single event parameter estimation, k indexes samples from the posterior, and $\mathcal{Z}_\emptyset(d_i)$ is the evidence obtained for event i. The integral is then approximated in a Monte Carlo sense, using the single event posterior samples that have been obtained previously.

III. RESULTS FOR GOLDEN BINARIES

In this section we give predictions about parameter biases that would arise if we used current BBH semianalytic waveform models to infer the properties of high mass BBHs in a sequence of past, current and future ground-based detector networks.

We select two exceptionally loud "golden binaries": one binary with parameters mimicking GW150914 and one binary at mass-ratio 1:3. Both systems contain BHs with spins misaligned with the orbital angular momentum vector causing the systems to be moderately precessing. We hold the luminosity distance of the systems constant, so that more sensitive detector networks will observe them with higher SNRs and obtain more precise measurements. Parameters for these systems are given in Table I. As signal waveforms we use NR simulations from the SXS [47] catalog computed with the SpEC code [48], as described in Sec. II B. Since these waveforms are too short to fill the frequency band of future interferometers which extends well below 20 Hz, we hybridize the NR waveforms with PN approximants in the inspiral, including higher order modes up to $\ell = 8$. We use the effective precession spin IMR waveform model `IMRPhenomPv2` for our main results and also quote complementary results for the nonprecessing `SEOBNRv4_ROM` model. `IMRPhenomPv2` includes $\ell = 2, m = \pm 2$ modes in the coprecessing frame, and a PN description of the motion of the coprecessing frame with an approximation for small precession angles [66,67,69,70].

A. Indistinguishability

We want to find an estimate that predicts beyond which SNR a particular waveform model that is used as a template in PE yields biased posterior distributions for the above BBH signals. We can find the answer by calculating the posterior distribution using Bayesian inference. However, this method is fairly costly for the very sensitive future detectors (see Fig. 1) where the signals have SNRs up to several thousands. Therefore, we compare against and extend a simpler metric for predicting the presence of biases.

If two waveforms h_1 and h_2 fulfill the criterion [99–102]

$$1 - \mathcal{O}(h_1, h_2) < D/(2\rho^2) \tag{8}$$

for a given PSD and SNR ρ, then they are deemed *indistinguishable*, i.e., $\langle\delta h|\delta h\rangle < 1$ and the posterior PDF should be unbiased in the sense that systematic errors from waveform inaccuracies are smaller than $1-\sigma$ statistical errors.

While this criterion is simple to evaluate, there are several problems that affect its usefulness in practice: The criterion is only sufficient, but not necessary and as a result it tends to be too *conservative*. Namely, if it is violated, then biases *can* but *need not* arise. In addition, the prefactor D is not known precisely. It can be derived as the number of (intrinsic) parameters whose measurability is affected by model inaccuracy [102]. The criterion also applies only in the high SNR limit as is the case for the Fisher information matrix [103].

To enhance the usefulness of the indistinguishability criterion we use the following procedure to tune the prefactor D.

(1) We compute posterior distributions for a sequence of detector networks on the above synthetic signals.

(2) From the posterior distributions we compute statistical and systematic errors for key parameters (chirp-mass, mass-ratio, effective aligned spin, and effective precession spin).

(3) We estimate the network (balance) SNR ρ_b at which the computed systematic and statistical errors become comparable.

(4) We compute the mismatch $1 - \mathcal{O}(h_{\rm model}, h_{\rm true})(\theta_{\rm true})$ between the template waveform and the signal at signal parameters for a representative detector sensitivity.

(5) Finally, we calculate

$$D = 2\rho_b^2[1 - \mathcal{O}(h_{\rm model}, h_{\rm true})(\theta_{\rm true})]. \tag{9}$$

We present results of applying this procedure to the selected golden binaries in Sec. III B. First we discuss some assumptions we make in applying it.

When computing the balance SNR and the mismatch we have to assume a power spectral density (PSD). We find empirically that systematic and statistical errors become comparable at network SNRs of ∼60 for the above sources. This SNR is found at aLIGO design sensitivity for SXS_BBH_0308 and at about A+ sensitivity for SXS_BBH_0104. The mismatch is only sensitive to the shape of the PSD and the frequency range of the overlap integral. We pick aLIGO design sensitivity [46] as a reference PSD since this is close to the sensitivity where the balance SNR is found, and it is in its vicinity that the tuned indistinguishability criterion should be most accurate. In general, we expect that mismatches will degrade as we approach future detectors since they will be sensitive to lower frequencies and will have

significantly more waveform cycles in band. The network SNR determines the discerning power of a network of detectors since we analyze the signal coherently. We neglect that the interferometers that make up detector networks usually have different sensitivities and pick a representative PSD. We use this PSD to compute the single interferometer mismatch in the indistinguishability criterion.

We use mock signals as a proxy for the true waveform obtained from exactly solving the two body problem in General Relativity. Hence we also assume that GR is the correct theory of gravity. Ideally our mock signals would be pure NR waveforms. This is in general not feasible since the cost of computing BBH coalescences with NR simulations scales very steeply with the initial frequency, so that in practice only part of the detector band can be filled by the NR signal for high mass BBHs. Therefore, we hybridize NR signals with PN inspiral waveforms.

NR simulations are only approximations of true GR waveforms. NR accuracy depends on the choice of configuration (e.g., more unequal mass-ratios and higher spin systems are harder to simulate accurately as the size of the apparent horizon of the BHs decreases) and on the size of the grid used to discretize Einstein's equations. In reality, NR simulations use multiple domains and a particular discretization method (finite differences [104–106], multidomain spectral collocation methods [48], or more advanced methods, such as discontinuous Galerkin [107,108]). While we can obtain a good estimate of the NR waveform error by computing mismatches for the same physical configuration but different grid sizes to decrease the truncation error and wave extraction errors, it is difficult to estimate the error in a hybrid waveform. We discuss this further in the Appendix.

In the above procedure for estimating the prefactor D we need to find the SNR at which the systematic and statistical errors are comparable. We know that parameters are in general correlated and thus we should take these correlations into account when estimating these errors. The indistinguishability criterion also makes this assumption. When quoting parameter estimation results we rely on errors computed from one and two-dimensional marginal posterior distributions, which are straightforward to compute and present. Therefore, we also compute the statistical and systematic errors from 1D marginal posteriors. A more conservative measure of the error is to compute where the injection lies in the posterior distribution, or a marginal PDF thereof. We obtain the percentile of the credible level of the injected parameters in the full posterior by performing parameter estimation with all sampling parameters fixed, except for the time and phase of coalescence. Detailed measurements of the latter are of no astrophysical interest, and as they can very strongly affect the likelihood, we prefer to marginalize over them.

We also consider a third method where we take into account the correlations in a set of key parameters only. To do this, we compute a kernel density estimate (KDE) of the marginal posterior distribution in the parameters of interest, compute the posterior probability value at the injection parameters and find its credible level in the marginal posterior. We compute a Gaussian KDE $\mathcal{K}(\tilde{\theta}) = \mathrm{KDE}[p(\tilde{\theta}|d)]$ of the marginal posterior distribution $p(\tilde{\theta}|d)$ and then solve numerically the equation $Q[\mathcal{K}(\theta^{(i)}); p] = \mathcal{K}(\theta_s)$ to find at which percentile $100p$ the true parameters θ_s of the signal lie in the marginal posterior. Here Q is the quantile function $Q(\mathrm{PDF}; p) = \mathrm{CDF}^{-1}(p)$ for a given PDF and its cumulative distribution function (CDF). In practice we work with the logarithm of the PDF to reduce the dynamic range. We discuss results from these procedures in the next section.

B. Predicted waveform accuracy requirements

We now apply the procedure presented in Sec. III A to posterior probability distributions and mismatches obtained for the two mock BBH signals shown in Table I for a series of detector networks. The networks are defined by the positions of the detectors on the Earth and their PSDs as listed in Table II.

Figure 2 shows the main results. According to Eq. (8) the general takeaway is that as long as the mismatch for a given semianalytical waveform model against the mock signal (red lines) lies below the *tuned* indistinguishability curve (light or dark blue lines) we do not expect parameter recovery to be biased. One can think of the indistinguishability curve showing the "acceptable error" for a waveform model for a particular SNR. Without tuning, the predicted SNR above which we would see biases (assuming that six intrinsic model parameters are affected) is about 25 for the SXS_BBH_0308 NR signal (and SNR 11 for the hybrid). For SXS_BBH_0104 it is predicted to be an SNR of ∼6. As we will see in Sec. III C these predictions are certainly way too conservative for the hybrid signals when compared with the parameter estimation results and the assumption that six parameters are biased is not correct either.

A first observation is that semianalytic models (here the representative `IMRPhenomPv2` and `SEOBNRv4_ROM` models) were sufficiently accurate to analyze GW150914 during aLIGO's first observing run. This is hardly a surprise and has been studied in depth by comparing against NR simulations and waveform models by the LVC [19,51]. Figure 2 also predicts that semianalytical models will lead to biased parameter recovery at and beyond HLVK sensitivity for SXS_BBH_0308 and at and beyond the A+ network for SXS_BBH_0104. Moreover, current NR waveforms will not be guaranteed to be sufficiently accurate for unbiased parameter recovery beyond the Voyager network (where the dark blue line intersects the dark green line). Clearly then current waveform models will not be accurate enough for 3G ground-based detectors such as ET and Cosmic Explorer (CE), which are currently being planned. We will require waveform models to be at least *three orders of magnitude* more accurate and improvements of *one order of magnitude* for NR waveforms.

Figure 2 presents a simplified picture to convey the main message that current waveform models are not accurate enough for planned 3G detectors. We now come back to some of the assumptions we have mentioned in Sec. III A and shed some light on details. The shape of the PSDs and the range in frequency over which particular interferometers are sensitive varies with the networks and influences the value of the mismatch that enters the indistinguishability criterion. The horizontal lines shown in Fig. 2 provide a simplified representative measure of the error. For SXS_BBH_0308

TABLE II. List of ground-based detector networks used in this study. The networks are defined by the positions of the detectors on the Earth and their PSDs in parentheses. We also indicate the frequency $f_{\rm low}$ at which we start integrating the likelihood integral and the network SNR of the PN-NR hybrid signals in these networks (see Table I for the parameters). Detector locations are indicated by: H1, LIGO Hanford; L1, LIGO Livingston; V1, Virgo; K1, KAGRA; I1, LIGO India; E1, E2, and E3, the interferometers of the triangular Einstein Telescope (ET) detector [109]. Text data files for the PSDs or ASDs can be found in Refs. [45,46] under the names given in this table.

			Network SNR	
Network	List of Interferometers and PSDs	$f_{\rm low}$[Hz]	0308	0104
O1	H1, L1 (`O1`)	30	25.4	11.2
HLVK	H1, L1 (`aLIGODesign_2018_T1800044`), V1 (`AdVirgo`), K1 (`KAGRA`)	10	88.9	41.6
A+	H1, I1 (`A+`)	10	125.7	57.5
Voyager	H1, L1 (`Voyager`)	10	276.3	128.4
ET	[E1, E2, E3] (`ET_D`)	5	950.9	466.3
ET-CE	[E1, E2, E3] (`ET_D`), H1 (`CE`)	5	2598.8	1205.2

mismatches against `IMRPhenomPv2` range from 0.002 (aLIGO O1) to 0.02 (CE) for the pure NR signal, which is in band from 20 Hz and above, and from 0.002 (aLIGO O1) to 0.008 (CE) for the hybrid signal. Starting frequencies are given in Table II. Mismatches for aLIGO, AdVirgo, KAGRA, and A+ are very similar to those for the aLIGO O1 results. For the nonprecessing `SEOBNRv4_ROM` model mismatches range from 0.003 (aLIGO O1) to 0.02 (CE) for the pure NR signal and from 0.005 (aLIGO O1) to 0.03 (CE) for the hybrid signal. For the SXS_BBH_0104 hybrid signal the mismatches against `IMRPhenomPv2` range from 0.06 (CE) to 0.09 (aLIGO O1, aLIGO design). Here, mismatches against `SEOBNRv4_ROM` are surprisingly slightly better 0.04 (CE) to 0.07 (AdvVirgo).

We want to stress that the mismatches depend very sensitively on the inclination angle under which the signal is seen. If we were to change the inclination for SXS_BBH_0308 from near face-off, 2.7454, which is compatible with GW150914, to $\pi/3$ which emphasizes more harmonics content beyond the dominant $(2, \pm 2)$ mode, then the mismatch is about an order of magnitude worse. If biases were to appear at the same SNR for this changed inclination, then this would make D an order of magnitude larger and the left panel of Fig. 2 would look markedly different and have stronger implications for how much waveform models need to be improved.

We have indicated in Fig. 2 the estimated accuracy of waveform models and NR simulations for the particular binary configurations by colored regions that are independent of the detector networks. These regions are supposed to give a rough sense of how accurate currently available models or codes are in the neighborhood of the BBH configurations considered here. Similar considerations as for the mismatches quoted above apply for the bounds of these regions. For

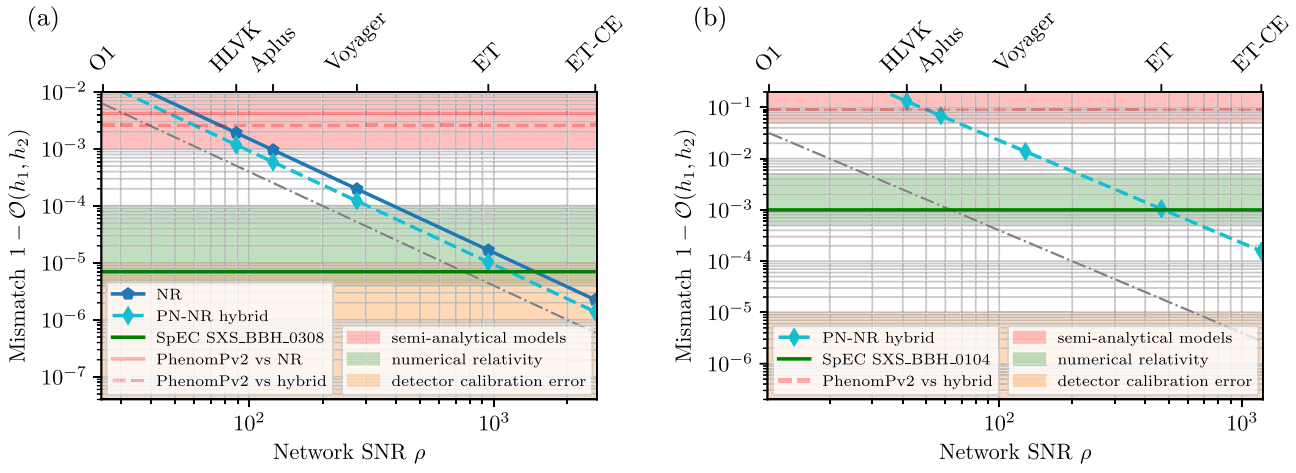

FIG. 2. Predicted waveform accuracy requirements for second and third generation ground-based detector networks. We show results for two binaries: (a) SXS_BBH_0308 and (b) SXS_BBH_0104 (see Table I). Each panel shows mismatch against network SNR and on the top x axis the detector network (see Table II) in which the signal had the SNR shown in the bottom x axis. Solid lines indicate results for pure NR signals, while dashed lines come from NR signals hybridized against PN waveforms in the inspiral. The blue lines and data points show how the mismatch falls with rising SNR according to the indistinguishability criterion Eq. (8) with the prefactor D tuned according to Eq. (9), as $D/(2\rho^2)$. The dash-dotted gray line shows the prediction of Eq. (8) with $D = 8$. Horizontal red lines show the mismatch of the signal against the `IMRPhenomPv2` template waveform at the signal parameters (also called "unfaithfulness") for aLIGO design sensitivity. The horizontal green line shows the mismatch between NR waveforms obtained for different grid resolutions for the same signal configuration. Shaded regions provide rough estimates of the accuracy of current semianalytic waveform models and current NR waveforms for the particular binary systems and the level of expected detector calibration error in terms of mismatch. Waveform error estimates are higher for the more challenging unequal mass spinning SXS_BBH_0104 configuration compared to SXS_BBH_0308.

simplicity we have bounded these very rough estimates by constant mismatch. The extents of the horizontal bands shown in Fig. 2 are as follows. Left panel: dominant-mode semianalytic models $[10^{-3}, 10^{-2}]$, NR: $[4 \times 10^{-6}, 10^{-4}]$. Right panel: dominant-mode semianalytic models $[5 \times 10^{-2}, 0.2]$, NR: $[5 \times 10^{-4}, 5 \times 10^{-3}]$. Finally, detector calibration error depends on the detector network and is expected to improve over time up to a level that is believed to be attainable from current understanding. In Fig. 2 the estimate of the mismatch error due to detector calibration errors uses a realistic estimate for future detectors and assumes 1% relative error in amplitude, 1° error in phase [96,97]. We assume that the functional form of the dephasing from detector calibration errors can be modeled by a polynomial of degree two or higher which starts at 1° at the low frequency cutoff and decreases toward zero dephasing in the ringdown regime. We find that detector calibration errors only contribute below 10^{-5}. Ultimately, the noise floor that comes from detector calibration error will only become problematic for 3G detector networks if we are not otherwise dominated by waveform errors.

Balancing accuracy using the full posterior

The above results used 1D marginal posterior distributions to calculate statistical and systematic errors and find the SNR at which they are comparable. We now discuss results where we take into account correlations between binary parameters and how they compare to the above. Irrespective of how many parameters we choose to include in the marginal posterior distribution we can always ask the question at which credible level the injection lies in the posterior distribution. We want to estimate when this is close to the 68th percentile. Since we only have data for fixed networks we need to interpolate the percentile values to estimate the SNR at which errors are balanced.

For the SXS_BBH_0308 PN-NR hybrid signal we find the injection in the full posterior at the 2nd percentile for the O1 network and at the 100th percentile for HLVK, marginalizing over relative time and phase. For the marginal posterior in $(\mathcal{M}, q, \chi_{\rm eff}, \chi_{\rm p})$ we find the injection at the 12th and 100th percentile in O1 and HLVK, respectively. For the 1D marginal distributions in these parameters we find that the injection lies between the 4th to 50th percentile for O1 and between the 78th and 99th percentile in HLVK. Therefore, for this configuration we find a similar balance SNR of 60 for these different ways of computing the error balance. This estimate is somewhat uncertain, since we do not have any datapoints in between the O1 and HLVK networks. In terms of the prefactor D, we would expect to have $D \sim 8$ if the key parameters are biased, but we find $D \sim 20$ if the errors are balanced at SNR 60. We note that the chirp mass and the effective precession spin are quite biased for this signal. For the NR only signal we find $D \sim 30$ because the mismatch is worse in the late inspiral and merger part. For this signal the rapid increase of the percentile at which the signal is found in the posterior with increasing detector sensitivity is mostly due to the strong bias in $\chi_{\rm p}$ that appears when going from O1 to HLVK sensitivity as can be seen in Fig. 3.

For the SXS_BBH_0104 source we find the injection in the full posterior at the 7th and 100th percentiles for the O1 and HLVK networks, respectively. For the marginal posterior in $(\mathcal{M}, q, \chi_{\rm eff}, \chi_{\rm p})$ we find the injection below the 40th percentile for O1, HLVK, A+, and Voyager, and it lies at the 100th percentile for the ET and ET-CE networks. For the 1D marginal distributions in these parameters we find that the injection lies between the 3rd to 43rd percentile for O1 and between the 12th and 40th percentile in HLVK. The balance SNR is then estimated to be ∼250. In combination with the large mismatch between the signal and template waveforms, it results in an enormous prefactor of $D \sim 10^4$. The reason is that there is almost no bias in $(\mathcal{M}, q, \chi_{\rm eff}, \chi_{\rm p})$, as can be seen in Fig. 4 discussed in Sec. III C. If we add inclination and distance parameters, then there is a noticeable bias and we find the injection at the 50th percentile for O1 and at the 100th percentile for the HLVK network and beyond. This results in a more reasonable balance SNR of ∼22 and a prefactor of $D \sim 90$. Using the 1D marginal errors we find a balance SNR of roughly 50 and a prefactor of $D \sim 450$. The naive indistinguishability criterion with $D = 8$ predicts biased recovery at SNR 10, which is close to the SNR of the signal in the O1 network.

C. Parameter estimation results

We now turn to looking directly at posterior distributions for the analysis of the two mock BBH signals from Table I for a series of detector networks. Histograms and 90% credible regions for key parameters are shown in Figs. 3 and 4 for the SXS_BBH_0308 and SXS_BBH_0104 sources, respectively. Here we show `IMRPhenomPv2` posteriors since this model includes approximate precession effects, in contrast to the nonprecessing `SEOBNRv4_ROM` model.

1. Results

The posterior distributions of the detector-frame chirp mass shown in the top left panel of Figs. 3 and 4 become progressively tighter as we go to more sensitive networks, their widths scaling roughly inversely with the SNR. This is the expected behavior for a multimodal Gaussian which the posterior distribution is expected to follow in the high SNR limit, although the 90% credible regions for some marginal 2D posteriors shown in the other panels are clearly not Gaussian. Only part of the chirp mass posterior is shown for O1 sensitivity so that we can more clearly see the posteriors for networks operating at higher sensitivities. The measurement precision in chirp mass in terms of the width of the 90% credible interval increases from $\sim 5 M_\odot$ (O1), to $0.4 M_\odot$ (HLVK), and $0.004 M_\odot$ (ET+CE). The massive increase in precision for 3G detectors is expected due to the improved sensitivity and the significantly larger number of waveform cycles in the detector frequency band. For instance, for SXS_BBH_0308 there are 64 cycles in band from 10 Hz, compared to 217 cycles from 5 Hz and 1025 cycles from 2 Hz. For SXS_BBH_0308 the O1 posterior is unbiased, with the true chirp mass value (red dashed line) near its peak. For the HLVK network and beyond the posteriors peak away from the true chirp mass. The true chirp mass is found at the 98th percentile for HLVK and for the Voyager network and beyond at the 100th percentile. For ET and ET-CE the chirp mass is underestimated by $0.18 M_\odot$. For SXS_BBH_0104, there is again no visible bias at O1

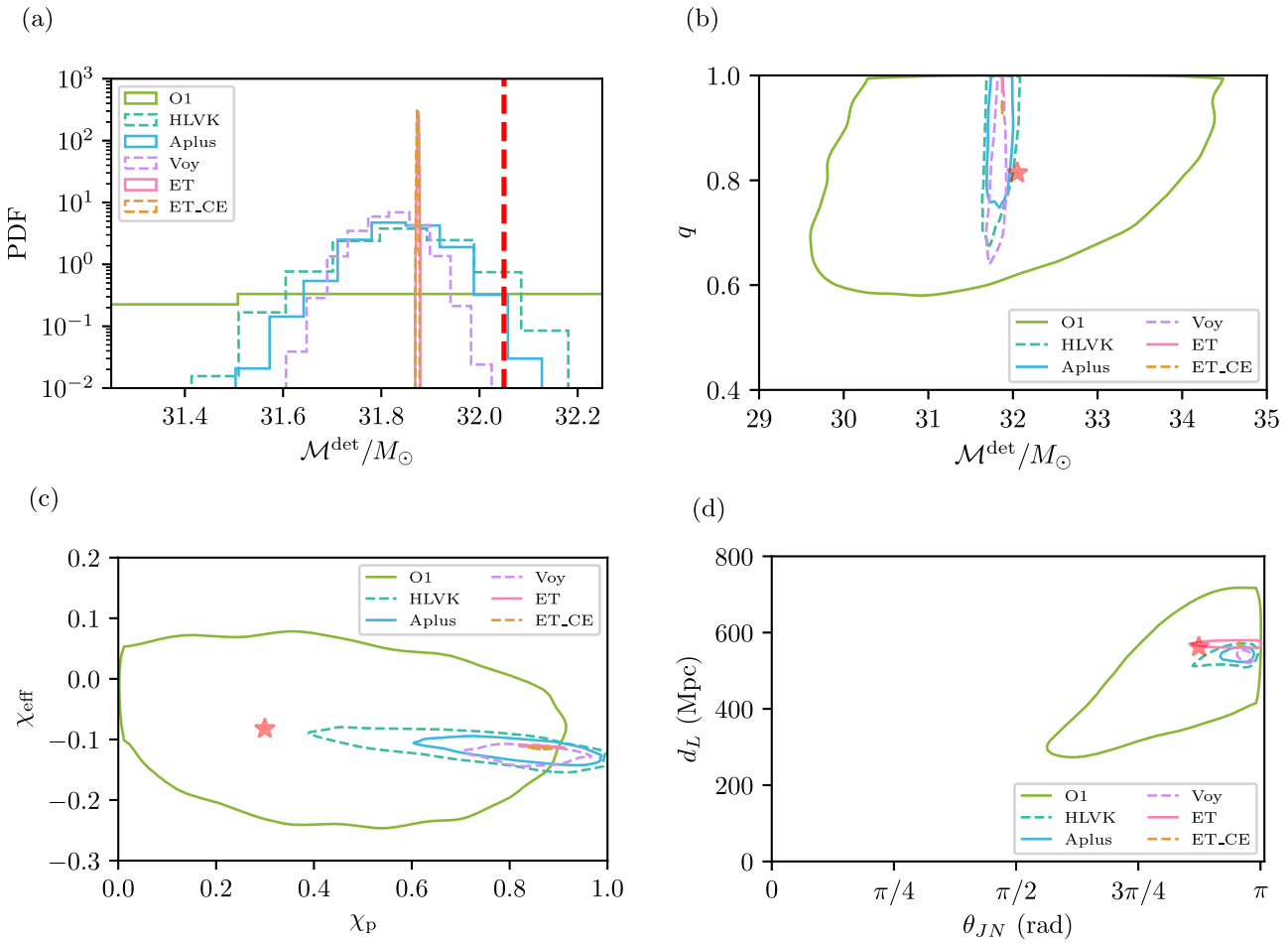

FIG. 3. Posterior PDFs for the SXS_BBH_0308 PN-NR hybrid signal (see Table I) using `IMRPhenomPv2` as the template waveform for a sequence of detector networks (see Table II). We show either histograms of one-dimensional PDFs or contours indicating 90% credible regions for 2-dimensional PDFs. (a) Histograms for detector-frame chirp mass, (b) contours for chirp-mass and mass-ratio, (c) contours for effective precession spin $\chi_{\rm p}$ and effective aligned spin $\chi_{\rm eff}$, (d) contours for inclination angle $\theta_{\rm JN}$ and luminosity distance d_L. True parameter values of the source binary are indicated as red dashed lines or a red asterisk.

sensitivity. For HLVK, the true value lies at one sigma away from the peak, and at the 92nd percentile for the Voyager network. Recovery is very accurate for the ET and ET-CE networks with a bias of $-0.01M_\odot$.

In the remaining panels of Figs. 3 and 4 we show 90% credible regions for marginal 2D posteriors for several key parameters, to give a sense of the correlations between binary parameters, starting with chirp mass and mass-ratio. Compared to chirp mass, the mass-ratio is much more difficult to measure, resulting in very wide posteriors. This is especially true for the near equal mass SXS_BBH_0308 source. There, the one-sided 10% percentile of the mass-ratio PDFs is roughly at 0.7 for 2G detectors. For 3G detectors the measurement is much more precise, again due to more inspiral cycles being observable, but in this case the mass-ratio is estimated to be too close to equal-mass with a bias $q_{\rm true} - q_{\rm MAP} \approx -0.15$. For the unequal mass SXS_BBH_0104 source, the mass-ratio is much better measured. The measurement precision in terms of the 90% interval increases from 0.5 (O1) to 0.1 (HLVK) and 0.01 (ET-CE). Biases only appear for 3G detectors, where they are about -0.05.

While there are 6 spin degrees of freedom in a generic precessing BBH, most of them are very difficult to measure. The aligned-spin degrees of freedom, in particular a mass-weighted linear combination called $\chi_{\rm eff}$ is the best measured spin parameter which is also degenerate with the mass-ratio [19,28,91,110–115]. For SXS_BBH_0308 the 90% interval for $\chi_{\rm eff}$ shrinks from 0.26 (O1) to 0.05 (HLVK), and 0.003 (ET-CE), while for SXS_BBH_0104 it shrinks from 0.45 (O1) to 0.05 (HLVK), and 0.004 (ET-CE). Beyond O1 sensitivity the biases are below 0.05 for SXS_BBH_0308 and 0.02 for SXS_BBH_0104, only becoming significant for 3G detectors.

During LIGO and Virgo's O1 and O2 observing runs precession effects have so far eluded measurement from compact binaries [1]. In terms of the effective precession spin parameter $\chi_{\rm p}$ [66,72] the posterior distributions shown in GWTC-1 have not provided new information compared to the prior distribution. We expect this situation to change with the improved sensitivity of future detectors [27,28] and the analysis of these sources is a case in point that we will be able to measure precession effects with future detectors. For SXS_BBH_0308

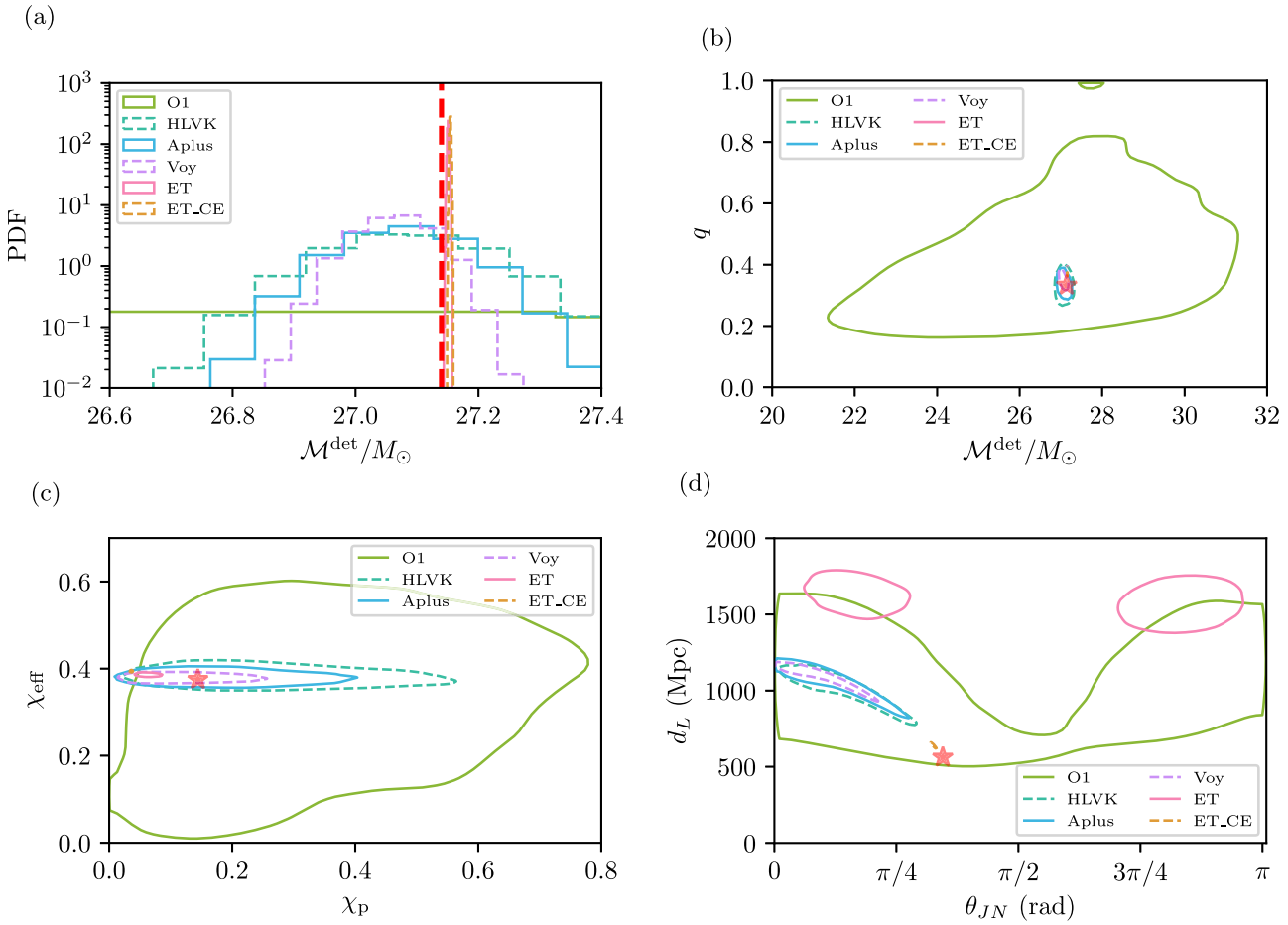

FIG. 4. Posterior PDFs for the SXS_BBH_0104 PN-NR hybrid signal (see Table I) using `IMRPhenomPv2` as the template waveform for a sequence of detector networks (see Table II). We show either histograms of one-dimensional PDFs or contours indicating 90% credible regions for two-dimensional PDFs. (a) histograms for detector-frame chirp mass, (b) contours for chirp-mass and mass-ratio, (c) contours for effective precession spin $\chi_{\rm p}$ and effective aligned spin $\chi_{\rm eff}$, (d) contours for inclination angle $\theta_{\rm JN}$ and luminosity distance d_L. True parameter values of the source binary are indicated as red dashed lines or a red asterisk.

the 90% interval for $\chi_{\rm p}$ shrinks from 0.7 (O1) to 0.5 (HLVK), 0.2 (Voyager) and 0.04 (ET-CE), while for SXS_BBH_0104 it shrinks from 0.6 (O1) to 0.4 (HLVK), 0.2 (Voyager) and 0.004 (ET-CE). Beyond O1 sensitivity where the measurement is uninformative, SXS_BBH_0308 posteriors are severely biased, overestimating $\chi_{\rm p}$ by about 0.6. The system is thus seen as close to maximally precessing while the averaged in-plane spin is only ∼0.3. For SXS_BBH_0104 the $\chi_{\rm p}$ measurements are much more reliable and only offset by ∼0.1.

Finally we show results for the marginal posteriors in luminosity distance d_L and the inclination angle $\theta_{\rm JN}$ between the total angular momentum J of the binary and the line of sight vector N under which the binary is seen from the detector network. These two parameters are especially degenerate in how they affect the amplitude of the source and the 2D posteriors are in general not Gaussian which limits the usefulness of 1-dimensional interval estimates and biases. The inclination posterior can have a single mode as for SXS_BBH_0308 which is seen close to the face-off inclination of the source, with some overestimation in $\theta_{\rm JN}$ and underestimation in the distance, or it can be bi-modal as for SXS_BBH_0104 for networks with (close to) colocated detectors (O1, ET) which have a harder time constraining it to the correct mode. Networks with better coverage of the Earth (HLVK, A+, Voyager, ET-CE) obtain the correct mode, but the inclination angle is substantially underestimated along with overestimating the distance by a factor of about 2. The only network to recover the inclination and distance with good accuracy is the ET-CE network.

2. Discussion

In this section we provide a comparison between results obtained from two different waveform models, the agreement between these models and the source PN-NR waveforms and discuss the importance of limitations of the models in interpreting the parameter estimation results.

In this study we perform parameter estimation on signals in *zero noise*. This is a particular noise realization that can be interpreted as the average over all possible Gaussian noise realizations. It is an appropriate choice when one wants to focus on the effect of waveform systematics on posterior distributions. Therefore, any discrepancy we see between the posterior estimates and the true source parameters of the mock signals must be due to disagreements between the source and template waveforms or due to prior effects. Given that we use

TABLE III. Medians and 90% credible intervals for selected source parameters for the O1 network (top), the HLVK network (middle), and the the ET-CE network (bottom). We show the detector-frame chirp mass $\mathcal{M}^{\rm det}$, the mass-ratio $q = m_2/m_1 \leqslant 1$, the effective aligned spin $\chi_{\rm eff}$, effective precession spin $\chi_{\rm p}$, the luminosity distance d_L, and the inclination angle $\theta_{\rm JN}$ between the total angular momentum of the binary and the line of sight. Only `IMRPhenomPv2` provides a posterior for the effective precession spin $\chi_{\rm p}$, since `SEOBNRv4_ROM` is a nonprecessing model.

Event	Waveform model	$\mathcal{M}^{\rm det}/M_\odot$	q	$\chi_{\rm eff}$	$\chi_{\rm p}$	d_L/Mpc	$\theta_{\rm JN}$
SXS_BBH_0308	IMRPhenomPv2	$31.959^{+1.838}_{-2.043}$	$0.842^{+0.141}_{-0.216}$	$-0.072^{+0.116}_{-0.145}$	$0.37^{+0.45}_{-0.29}$	514^{+132}_{-191}	$2.698^{+0.337}_{-1.365}$
	SEOBNRv4_ROM	$31.524^{+2.077}_{-2.290}$	$0.831^{+0.151}_{-0.237}$	$-0.076^{+0.122}_{-0.163}$	N/A	469^{+160}_{-200}	$2.501^{+0.469}_{-1.242}$
SXS_BBH_0104	IMRPhenomPv2	$26.703^{+3.599}_{-3.427}$	$0.352^{+0.425}_{-0.120}$	$0.330^{+0.209}_{-0.236}$	$0.30^{+0.38}_{-0.21}$	1040^{+485}_{-452}	$0.815^{+2.083}_{-0.640}$
	SEOBNRv4_ROM	$27.897^{+4.086}_{-3.622}$	$0.406^{+0.423}_{-0.154}$	$0.390^{+0.249}_{-0.209}$	N/A	1088^{+591}_{-516}	$0.939^{+1.899}_{-0.732}$
SXS_BBH_0308	IMRPhenomPv2	$31.845^{+0.169}_{-0.182}$	$0.894^{+0.095}_{-0.174}$	$-0.113^{+0.027}_{-0.027}$	$0.75^{+0.17}_{-0.31}$	$540.6^{+21.0}_{-29.8}$	$2.970^{+0.109}_{-0.224}$
	SEOBNRv4_ROM	$32.060^{+0.136}_{-0.132}$	$0.793^{+0.141}_{-0.102}$	$-0.071^{+0.021}_{-0.017}$	N/A	$519.8^{+66.7}_{-109.3}$	$2.636^{+0.359}_{-0.339}$
SXS_BBH_0104	IMRPhenomPv2	$27.083^{+0.188}_{-0.184}$	$0.331^{+0.056}_{-0.047}$	$0.383^{+0.026}_{-0.027}$	$0.23^{+0.27}_{-0.14}$	1038^{+93}_{-215}	$0.407^{+0.417}_{-0.271}$
	SEOBNRv4_ROM	$27.045^{+0.165}_{-0.175}$	$0.374^{+0.046}_{-0.044}$	$0.389^{+0.038}_{-0.065}$	N/A	949^{+188}_{-328}	$0.607^{+0.435}_{-0.442}$
SXS_BBH_0308	IMRPhenomPv2	$31.874^{+0.002}_{-0.002}$	$0.939^{+0.010}_{-0.010}$	$-0.114^{+0.001}_{-0.001}$	$0.86^{+0.02}_{-0.02}$	$569.6^{+3.8}_{-3.6}$	$3.008^{+0.004}_{-0.005}$
	SEOBNRv4_ROM	$31.894^{+0.002}_{-0.002}$	$0.770^{+0.005}_{-0.005}$	$-0.100^{+0.001}_{-0.001}$	N/A	$461.2^{+44.6}_{-23.0}$	$2.405^{+0.152}_{-0.074}$
SXS_BBH_0104	IMRPhenomPv2	$27.154^{+0.002}_{-0.002}$	$0.369^{+0.005}_{-0.004}$	$0.394^{+0.002}_{-0.002}$	$0.036^{+0.002}_{-0.002}$	639^{+17}_{-16}	$1.025^{+0.014}_{-0.016}$
	SEOBNRv4_ROM	$27.179^{+0.002}_{-0.002}$	$0.415^{+0.003}_{-0.003}$	$0.421^{+0.002}_{-0.002}$	N/A	179^{+3}_{-3}	$1.731^{+0.003}_{-0.003}$

high accuracy NR or PN-NR waveforms as the signal, which are good approximations of GR waveforms, and we analyze high SNR events these disagreements are assumed to come from approximations to GR waveforms made in the waveform models we use as templates.

We performed the parameter estimation analyses with the `IMRPhenomPv2` and `SEOBNRv4_ROM` IMR models. The assumptions made in these models are described in Sec. II. In Sec. III C 1 we presented results from the effective precessing `IMRPhenomPv2` model. Here we juxtapose these results against the posterior distributions obtained for the aligned-spin `SEOBNRv4_ROM` model. In Table III we show medians and 90% credible intervals for selected source parameters and the two BBH sources for the O1, HLVK, and ET-CE networks. To gauge measurement accuracy we show absolute biases divided by the standard deviation in Table IV. We find that the two models give overall similar results for the parameter estimates. Noticeable differences are as follows: The chirp mass for the HLVK network is recovered more accurately for `SEOBNRv4_ROM` for SXS_BBH_0308 compared to `IMRPhenomPv2`. Similarly, `SEOBNRv4_ROM` recovers the mass-ratio, effective aligned spin, luminosity distance, and inclination angle with better accuracy than `IMRPhenomPv2` for SXS_BBH_0308 in the HLVK network. For SXS_BBH_0104 in the HLVK network, `SEOBNRv4_ROM` does not recover the chirp mass very accurately, but finds the other selected source parameters with better accuracy than `IMRPhenomPv2`. For the O1 network all parameters except distance and inclination are unbiased. At HLVK sensitivity several parameters exceed unity in the modulus of the normalized bias, which indicates that the difference between true and MAP parameter value is larger than one standard deviation. The largest biases are found in the luminosity distance and inclination for SXS_BBH_0104 recovered by the `IMRPhenomPv2` model and for the effective precession spin $\chi_{\rm p}$ for SXS_BBH_0308 found by `IMRPhenomPv2`.

Turning toward the ET-CE network we see in the size of the 90% intervals that measurement precision has increased dramatically, for instance the chirp mass is measured to $\pm 0.002 M_\odot$, two orders of magnitude more accurately than for HLVK. The precision for the mass-ratio has increased by about one order of magnitude to roughly ± 0.005 and similarly the effective aligned spin is measured to ± 0.002 and the effective precession spin better than ± 0.02. In the ET-CE network all parameters shown here have normalized biases exceeding unity in their absolute value. All of these parameters are estimated to lie outside one standard deviation for the two waveform models employed here, making it clear that waveform models need to be improved for analyses with 3G detectors.

The PN-NR signal waveforms we used to represent the GWs emitted by the source binaries contain higher harmonics beyond $\ell = 2$, but the waveform models used as templates only include the dominant quadrupolar modes. In fact, the models do also not include all of the m modes at $\ell = 2$, but merely the $\ell = 2, m = \pm 2$ contributions in the coprecessing frame. This begs the question how much the missing higher modes affect the analyses. In terms of SNR ρ for SXS_BBH_0308, 99.5% of ρ^2 is found in the $(2, \pm 2)$ mode (ignoring precession), while for SXS_BBH_0104 95.6% of the total ρ^2 is found in the $(2, \pm 2)$ mode, and 3.8% in the $(3, \pm 3)$ mode. The above percentages are stated in terms of ρ^2 as SNR adds in quadrature. The overlap between a signal with and without higher harmonics at the signal parameters is 0.9997 for SXS_BBH_0308 and 0.96 for SXS_BBH_0104, which illustrates that higher modes only become important

TABLE IV. Normalized biases from 1D marginal posteriors for selected source parameters for the O1 network (top), the HLVK network (middle), and the ET-CE network (bottom). We show the bias $\theta_s - \theta_{\mathrm{MAP}}$ in the binary parameter θ divided by the standard deviation of $p(\theta|d)$ s refers to the value for the mock source and maximum a posteriori (MAP) is the maximum a posteriori value of the posterior distribution, i.e., $\max_\theta p(\vec{\theta}|d)$.

Event	Waveform model	$\mathcal{M}^{\mathrm{det}}$	q	χ_{eff}	χ_{p}	d_L	θ_{JN}
SXS_BBH_0308	IMRPhenomPv2	−0.16	−0.05	−0.23	−0.47	−0.63	−0.38
	SEOBNRv4_ROM	−0.34	0.27	−0.55	N/A	1.04	0.09
SXS_BBH_0104	IMRPhenomPv2	0.80	0.02	0.76	−0.88	−2.19	0.88
	SEOBNRv4_ROM	−0.25	0.09	0.05	N/A	−2.39	1.12
SXS_BBH_0308	IMRPhenomPv2	2.21	−1.28	2.73	−4.32	1.10	−2.30
	SEOBNRv4_ROM	0.98	−0.23	0.62	N/A	−0.02	−0.37
SXS_BBH_0104	IMRPhenomPv2	−0.08	0.93	−1.09	−1.29	−5.91	3.91
	SEOBNRv4_ROM	1.37	−0.82	0.18	N/A	−1.52	1.10
SXS_BBH_0308	IMRPhenomPv2	154.80	−20.21	39.16	−45.78	−2.73	−98.98
	SEOBNRv4_ROM	153.90	15.59	55.09	N/A	4.56	4.26
SXS_BBH_0104	IMRPhenomPv2	−10.34	−13.23	−13.51	83.57	−7.23	6.02
	SEOBNRv4_ROM	−30.67	−42.55	−48.59	N/A	215.28	−333.11

for higher mass-ratios. To compute these numbers we used the `SEOBNRv4_ROM` [68,74,75] and a `SEOBNRv4HM_ROM` [116] waveform models and aLIGO design sensitivity with a starting frequency of 10 Hz. Computing the detector response Eq. (4) and optimizing over the polarization angle for the template waveform while keeping sky location fixed yields overlaps of 0.9993 and 0.96, instead. For SXS_BBH_0308 (SXS_BBH_0104), the overlap between an NR waveform that includes all $\ell = 2$ modes versus a waveform that only includes the $(2, \pm 2)$ modes in the coprecessing frame is 0.99996 (0.99992) or 0.9992 when optimizing over the polarization angle. This shows that the for these configurations the $(2, \pm 1)$ modes in the coprecessing frame are very weak.

In Sec. III B we quoted a range of overlaps between the PN-NR signals and the waveform models `IMRPhenomPv2` or `SEOBNRv4_ROM` at the signal parameters. For aLIGO design sensitivity they are as follows: 0.97 (0.91) for the SXS_BBH_0308 hybrid signal and 0.91 (0.94) for SXS_BBH_0104 using `IMRPhenomPv2` (`SEOBNRv4_ROM`). In contrast, the overlaps between signals with and without higher harmonics which we just computed above are 0.9997 and 0.96 for SXS_BBH_0308 and SXS_BBH_0104, respectively, for the `SEOBNRv4_ROM` and `SEOBNRv4HM_ROM` models. From this we see that the overlaps between the PN-NR hybrids and `IMRPhenomPv2` or `SEOBNRv4_ROM` are significantly lower than the ones the ones between signals with and without higher harmonics. This indicates that the a sizable part of the disagreement between the signal and template waveforms comes from modeling error in the coprecessing frame $(\ell, m) = (2, \pm 2)$ mode. Some disagreement could also come from the approximate description of precession.

For all analyses presented in this study, we have assumed that the zero noise data still carries an inherent uncertainty in its calibration, as described in Sec. II E. This uncertainty is modelled as a cubic spline, enforcing a smooth variation across the bandwidth of the analysis. In principle, by allowing this additional degree of freedom which could absorb some of the mismatch between the PN-NR hybrids and the approximate waveform models used in the PE analysis the observed biases could be expected to be reduced. Comparing the 1D posterior distributions shown in Fig. 3 to distributions from analyses where the marginalization over calibration uncertainties has been disabled, the observed biases remain. It should be noted that the analyses which includes marginalization over calibration uncertainties systematically recovers a slightly higher SNR accumulated over the detector network, but as this increase is of order $\lesssim 1/1000$ this is not expected to affect the conclusions with any significance.

IV. POPULATION STUDY

We have seen in Sec. III that in the HLVK design network we already expect biases with current waveform models for loud BBHs such as GW150914. Even small biases found for weaker single events could still manifest themselves when estimating properties of the population of BBHs [117–120]. In this section we perform a PE analysis for a population consisting of one hundred high mass precessing BBH events. On the one hand, we study the distribution and correlation of parameter biases and compute the overall bias over the population. On the other hand, we analyze the residual between the signal and the best matching template waveform, in terms of its SNR, power in the time frequency plane, and in terms of Bayes factors between analyses assuming coherent and incoherent signals across the detector network as implemented with BAYESWAVE [121,122]. Finally, we compute the population posterior for the power law index of the primary mass of the source binaries.

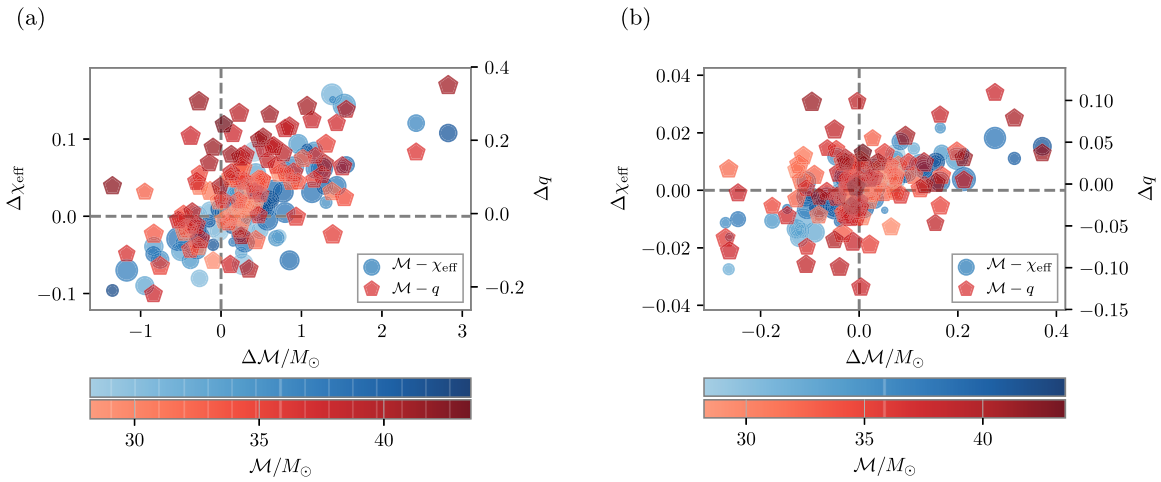

FIG. 5. Absolute biases $\theta_s - \theta_{\rm MAP}$ between chirp-mass $\mathcal{M}/M_\odot$ and effective aligned-spin $\chi_{\rm eff}$ (blue circles) and chirp-mass and mass-ratio q (red pentagons). (a) `NRSur7dq2` signals recovered with `IMRPhenomPv2`. The Pearson correlation coefficients are $R_{\mathcal{M},\chi_{\rm eff}} \sim 0.8$ and $R_{\mathcal{M},q} \sim 0.5$. (b) `IMRPhenomPv2` signals recovered with `IMRPhenomPv2`. The Pearson correlation coefficients are $R_{\mathcal{M},\chi_{\rm eff}} \sim 0.8$ and $R_{\mathcal{M},q} \sim 0.3$. We indicate the signal chirp mass by the luminosity of the symbol colors and the absolute value of the effective aligned-spin and mass-ratio of the signal by the symbol area to give a sense of how the biases are distributed over the parameter space. The symbol area was calculated as $200\sqrt{|\chi_{\rm eff}|}$ and $100\sqrt{q}$.

A. Setup

The events in this population study were drawn from the following distribution of source parameters: The primary mass has a PDF $p(m_1) \propto m_1^{-\alpha}$ with $\alpha = 1.3$ and $m_1 \in [45, 50]M_\odot$ and the mass-ratio is distributed as $q \sim U(0.5, 1)$. The chirp mass $\mathcal{M} = M_{\rm tot}\eta^{3/5}$ is computed from (m_1, q), where $M_{\rm tot} = m_1 + m_2$ and $\eta = q/(1+q)^2$. The remaining parameters are distributed as follows: spin magnitudes $a_i \sim U(0, 0.8)$, spin tilts $\cos t_i \sim U(-1, 1)$, the azimuthal angle between the spin vectors $\phi_{12} \sim U(0, 2\pi)$, the angle between the total and orbital angular momentum $\phi_{\rm JL} \sim U(0, 2\pi)$, the inclination angle $\cos\theta_{\rm JN} \sim U(-1, 1)$, the polarization angle $\psi \sim U(0, \pi)$. The luminosity distance, geocenter time, sky location, and phase were fixed at the parameters given in Table I.

Since NR waveforms are only available at isolated points in parameter space and thus cannot well represent the above distribution we choose the `NRSur7dq2` NR surrogate model for the signal waveforms [81]. This choice implies restrictions to mass-ratio $q \geqslant 1/2$, spin magnitudes $a_i \leqslant 0.8$ and, due to the relatively short waveform length, the constraint on the primary mass $m_1 \geqslant 45M_\odot$, so that waveforms representing the BBH population start at or below 20 Hz. We perform PE analyses with the BILBY code [88,89] with signals in zero noise and `IMRPhenomPv2` templates for the HLVK network.

B. Bias

We can learn about how population parameters will be affected by studying correlations between biases in key source parameters for events drawn from a population and to what degree single event biases average out over the population. In Fig. 5 we show absolute biases, defined as the difference between the true source parameters θ_s^i and a point estimate of the posterior distribution $\theta_{\rm p}^i$ for event i,

$$\mathcal{B}_i := \theta_s^i - \theta_{\rm p}^i. \tag{10}$$

As a default we use the MAP value of the posterior distributions as the point estimate.

We see that biases are large when the signal is represented by `NRSur7dq2` waveforms and the template by the `IMRPhenomPv2` model. In contrast, when the signal and template are represented by the same `IMRPhenomPv2` waveforms there is only a small discrepancy between the MAP and the true signal parameters which is expected to arise from stochastic sampling and prior effects. Here the posterior distribution is dominated by the likelihood since the signal SNRs are high. We find that log-likelihood values come close, but are a bit lower than, the peak value of the log-likelihood at the signal parameters. While the MAP (or equivalently maxL) parameters are a bit different than the signal parameters, the deviations in the waveform are tiny and the SNR in the residual is on the order of one. The spread is a factor 7 smaller in chirp mass, a factor 4 smaller in effective aligned spin and a factor 2 smaller in mass-ratio. For both types of signals we observe pronounced correlations between these parameters which we expect on physical grounds due to how these parameters enter the inspiral waveform [19,28,91,110–115]. We find Pearson correlation coefficients of $R_{\mathcal{M},\chi_{\rm eff}} \sim 0.8(0.8)$ and $R_{\mathcal{M},q} \sim 0.5(0.3)$ for `NRSur7dq2` (`IMRPhenomPv2`) signals. For NR-surrogate signals the chirp mass shows a clear tendency to be *overestimated*. This is also true for effective spin and mass-ratio. In contrast, for `IMRPhenomPv2` signals the distribution of the single event biases is more symmetrical. We also see that for NR-surrogate signals heavy binaries are prone to *overestimation* of $\chi_{\rm eff}$ as indicated by the luminosity of the red pentagons.

To get a better sense of how much these biases matter we discuss the distribution of the ratio $\mathcal{R}$ of absolute biases and

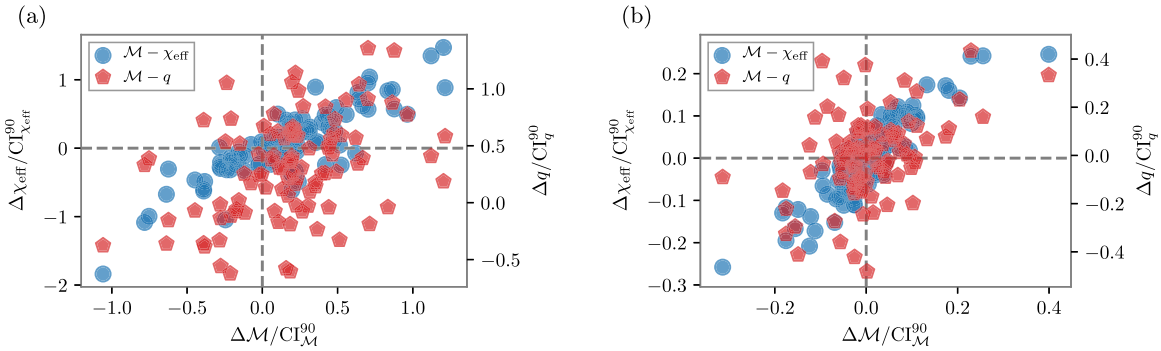

FIG. 6. Ratio of absolute biases and 90% credible intervals $\mathcal{R}(\theta, d)$ as defined in Eq. (11) between chirp-mass $\mathcal{M}/M_\odot$ and effective aligned-spin $\chi_{\rm eff}$ and chirp-mass and mass-ratio q. $\mathcal{R} = \pm 1/2$ when the true value is found at the boundaries of the 90% interval. (a) `NRSur7dq2` signals recovered with `IMRPhenomPv2`. (b) `IMRPhenomPv2` signals recovered with `IMRPhenomPv2`.

90% credible intervals

$$\mathcal{R}_i := \frac{\theta^i_s - \theta^i_{\rm MAP}}{{\rm CI}_{90}[p(\theta^i|d^i)]} \tag{11}$$

shown in Fig. 6, where we divide the bias by the extent of the 90% credible interval for each event and parameter. For NR-surrogate signals $|\mathcal{R}|$ reaches unity for the chirp-mass, takes values up to 2 for the effective aligned spin, and about 1.5 for the mass-ratio which indicates that the parameter recovery is strongly biased. The choice of comparing to the 90% interval is more conservative than to $1-\sigma$ which is assumed in the indistinguishability criterion. In contrast, for `IMRPhenomPv2` signals $|\mathcal{R}|$ is smaller than 0.4 for all parameters and the majority of events are found with very good accuracy $|\mathcal{R}| \lesssim 0.1$.

In Fig. 7 we see that bias in the luminosity distance tends to be negative, and with the definition in Eq. (10) this implies that the distance is overestimated in inference as a rule. The distance bias is reduced by about half for `IMRPhenomPv2` signals, compared to `NRSur7dq2` signals, but it is still sizable. In contrast we find relatively small biases of about 10° in the inclination angle.

We show the overall bias over the population in Table V. For NR-surrogate signals the largest population bias is seen for the MAP. Using the mean or median as a point estimate the overall bias is significantly lower than when using the MAP. This is not the case for `IMRPhenomPv2` signals, where the largest bias is found for the mean. We also show the sum of ratios of the biases over the 90% intervals, $\sum_i \mathcal{R}_i$. The size of this quantity shows more clearly how severe the biases are overall averaged over the population. Again the sum of the biases is much larger for the NR-surrogate signals, about 10–30 times larger than the sum of 90% credible intervals for key astrophysical parameters. The magnitude of $\sum_i \mathcal{R}_i$ for `IMRPhenomPv2` signals is about ∼2, indicating that there is no significant bias when combining all events in the population. We will revisit the question of how population estimates are affected in Sec. IV D.

C. Residuals

We previously discussed biases found for events in the BBH population study. The biases stem from a disagreement between the signal NR-surrogate waveforms $h_s(t;\theta_s)$ and `IMRPhenomPv2` template waveforms $h_{\rm m}(t;\theta_s)$ used in the analysis at the source parameters θ_s. This disagreement will also lead to some residual power being left over after subtracting the data containing the signal from the best fit template

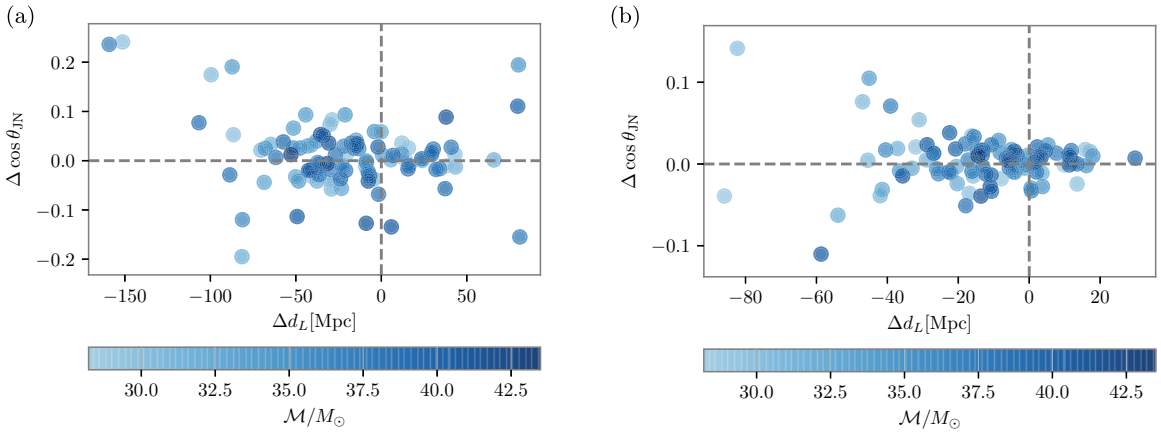

FIG. 7. Absolute biases $\theta_s - \theta_{\rm MAP}$ between luminosity distance and the cosine of the inclination angle. (a) `NRSur7dq2` signals recovered with `IMRPhenomPv2`. (b) `IMRPhenomPv2` signals recovered with `IMRPhenomPv2`. As in Fig. 5 we indicate the signal chirp mass by the luminosity of the symbol colors.

TABLE V. Population biases $\sum_i \mathcal{B}_i$, and $\sum_i \mathcal{R}_i$ [see Eqs. (10) and (11)] for chirp mass, mass-ratio, effective aligned spin, and effective precession spin. The point estimate θ_p is either the mean, median, or MAP. We show biases for `NRSur7dq2` and `IMRPhenomPv2` signals.

Quantity	Signal waveform	Point estimate	$\mathcal{M}^{\rm det}/M_\odot$	q	$\chi_{\rm eff}$	$\chi_{\rm p}$
$\sum_i \mathcal{B}_i$	`NRSur7dq2`	Mean	17.22	4.36	0.49	5.49
		Median	16.45	4.67	0.41	6.20
		MAP	34.01	7.78	1.12	5.37
$\sum_i \mathcal{B}_i$	`IMRPhenomPv2`	Mean	−4.68	−0.44	−0.26	−0.38
		Median	−3.87	−0.25	−0.25	0.02
		MAP	0.90	−0.13	0.04	−1.27
$\sum_i \mathcal{R}_i$	`NRSur7dq2`	Mean	9.92	18.94	3.36	27.73
		Median	9.58	20.18	2.79	30.19
		MAP	17.89	30.82	7.39	28.78
$\sum_i \mathcal{R}_i$	`IMRPhenomPv2`	Mean	−2.39	−1.80	−2.56	−1.33
		Median	−2.30	−1.23	−2.78	0.20
		MAP	0.86	−0.24	0.31	−5.10

waveform, $h_{\rm m}(t;\theta_{\rm MAP})$. Here we discuss how this residual power can be characterized in terms of SNR and power in the time frequency plane. We also perform an analysis with BAYESWAVE.

In Fig. 8 we show the network SNRs found in the signal strain $h_s(t;\theta_s)$ and in the residual strain $h_s(t;\theta_s) - h_{\rm m}(t;\theta_{\rm MAP})$. In each detector of the network we compute the strain by projecting the waveform polarizations on the detector as defined in Eq. (4). We observe that residuals reach SNRs of about 12, expect for one event with residual SNR ∼18.37. Parameters for this event are shown in Table VI. We find residual SNRs up to 30% of the signal SNR. The log-likelihood at MAP is highest for events where the agreement between the signal and template waveforms is good and thus the residual is small, and it drops substantially for events where the residual contains a sizable fraction of the signal SNR. For the event with the highest residual SNR the biases are only moderate $\Delta\mathcal{M} = -0.27M_\odot$, $\Delta q = 0.11$, $\Delta\chi_{\rm eff} = -0.02$, and $\Delta\chi_{\rm p} = -0.04$, but it has a high signal SNR 87.91.

Next we take a look at the power in the time frequency plane and compare the loudest residual against a chirp signal. In Fig. 9 we plot the Q transform [123] using PyCBC [124] of the residual with the highest SNR. As shown in the left panel, the power in the residual in LIGO Hanford traces out a chirp signal and agrees well with the overlaid time frequency evolution of the waveform emitted by the source. The right panel shows that the coherent power in the residual (taking into account time-shifts for each detector) is about a factor 5 larger than the power of the loudest single detector residual. Most of the coherent residual power is concentrated near merger where the GW signal is most nonlinear.

Finally, we analyze the residual strain across the detector network using the BAYESWAVE [121] code, assuming no predefined signal model apart from constraining signals coherent across the detector network to an elliptical polarization. Here, the waveform is reconstructed directly, through a superposition of Morlet-Gabor, or sine-Gaussian, wavelets [121,125], where the number, placement, and properties of the wavelets

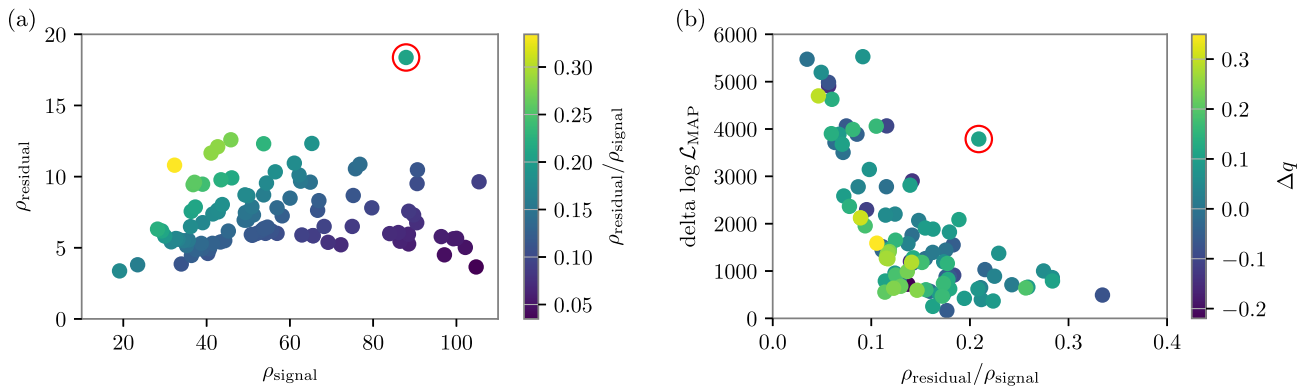

FIG. 8. Fractional network SNR in strain residuals $h_s(t;\theta_s) - h_{\rm m}(t;\theta_{\rm MAP})$ between the `NRSur7dq2` signal strains $h_s(t;\theta_s)$ and `IMRPhenomPv2` MAP template strain $h_{\rm m}(t;\theta_{\rm MAP})$ for each detector in the HLVK network compared to the SNR of the respective signal strain. (a) Residual SNR as a function of signal (injection) SNR with fractional SNR indicated by color for the events in the population study. (b) Delta log-likelihood $\langle d|h\rangle - 0.5\langle h|h\rangle$ at MAP against fractional SNR colored by the mass-ratio bias. The bias in mass-ratio is indicated by color. The event with the highest residual SNR is indicated by a red circle around the marker. The parameters for this event are given in Table VI.

TABLE VI. Source parameters for BBH with the loudest residual, shown in Fig. 9.

$M_{\rm tot}^{\rm src}/M_\odot$	$\mathcal{M}^{\rm src}/M_\odot$	q	$\vec{\chi}_1$	$\vec{\chi}_2$	$\chi_{\rm eff}$	$\chi_{\rm p}$	$\theta_{\rm JN}$
81.35	34.64	0.68	(−0.23, −0.64, 0.31)	(0.21, 0.53, 0.30)	0.31	0.68	2.63

are themselves variables in the analysis. For this study, we compare two competing models for the observed residual data [122]. The coherent model assumes a common waveform across the entire network, as originating from a point in the sky and projected onto each detector assuming standard antenna pattern functions for the two tensorial polarization modes as defined in Sec. II E [126]. The incoherent model assumes complete independence between the observed signals across the network. Instead of the data being represented through a common set of wavelets projected onto the detectors this model constructs a separate waveform for each detector where the placement and structure of the wavelets is independent from other detectors and no phase and time coherence across the network is required. The two models [127] can then be directly compared through a Bayes factor for each set of residual strains as shown in Fig. 10. As BAYESWAVE is constructed, it has a strong dependance on signal complexity, as opposed to simply depending on signal strength only, to make observational claims such as for example preferring a coherent description of the signal over an incoherent one [128]. This means that the Bayes' factors inherently incorporate the Occam factor between the two models, where the incoherent model can require a larger number of wavelets (and hence a larger number of signal parameters) to reconstruct the data across the network as it does not need to consider extrinsic parameters (sky position and two angles describing the polarization and ellipticity of the gravitational wave). For the set of residual strains in this study, we often find the incoherent model incapable of capturing the signal in an individual detector, with a median number of 0 wavelets per detector. The coherent signal, however, always captures the common signal, but even here the median number of wavelets is "only" 1. We interpret this as BAYESWAVE being consistently able to determine that there is *something* originating from a common coherent source in the data, but due to the relatively low SNR we are not generally able to make strong inference on the physical description of what this coherent signal would be. Even so, we argue that this type of analysis will be a valuable tool in determining the power and accuracy of future modelled inference [21,129], and can ensure that all of the observable signal can be captured and characterized. Note that the analysis here is performed in a noise-free set of data, assuming a known and fixed set of detector sensitivities shown in Fig. 1. For "real" data, the presence of time-varying random Gaussian noise [130], as well as actual detector glitches [131], is expected to reduce the fidelity of this category of tests, however BAYESWAVE is already capable of accounting for such variance [40,128]. The level to which variations in data will affect a study of residual recovery will be left for future investigation.

D. Population inference

We follow the hierarchical Bayesian inference method described in Sec. II E 1. We show the hyperposterior for α, the power-law index of the primary BH mass in Fig. 11, where we assumed a hyperprior $\pi(\alpha) \sim U(1, 2)$. Unfortunately, the PDFs of the power-law distribution for the true value and the boundary values of α are rather similar over the narrow mass interval considered here. This is probably due to the rather tight lower mass bound which is set by the finite length of the `NRSur7dq2` waveform model. Given that there is not much information in the hyperposterior we ask the question whether we prefer $\alpha = 1$ or $\alpha = 2$. Clearly, $\alpha = 1$ is preferred by the hyperposterior. This agrees with the observation that in the single event posterior PDFs we overestimate the masses (see Fig. 5).

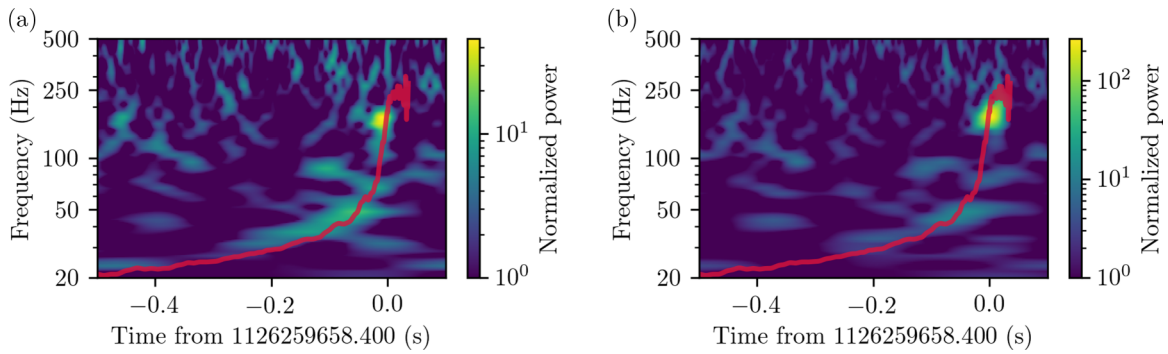

FIG. 9. Q transform of strain residuals for the event with the loudest residual (optimal network SNR 18.372). The parameters for this event are given in Table VI. The strain residuals were computed by subtracting the `IMRPhenomPv2` MAP template from the `NRSur7dq2` signal for each detector. Gaussian colored noise was added to the residual before computing the Q transform. (a) Residual in the interferometer where the residual is loudest: LIGO Hanford, SNR 11.5. (b) Coherent sum of the time-shifted residuals in LIGO Hanford, LIGO Livingston, Virgo, and KAGRA. The normalized power is shown in color. The chirp-trace of the injected `NRSur7dq2` signal is shown in crimson.

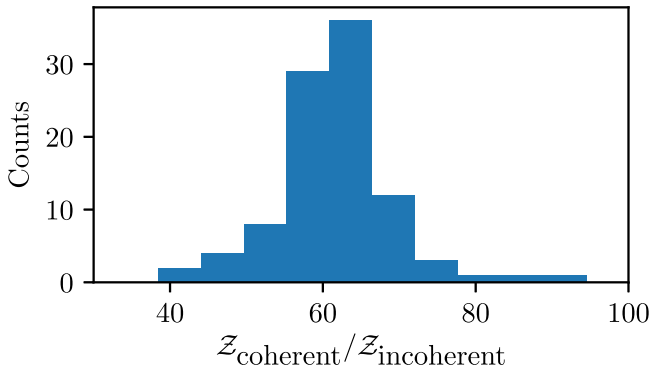

FIG. 10. Histogram of the Bayes factors, for all residuals (except the event highlighted in red in Fig. 8 and Table VI), between a coherent (assuming the same incoming signal in all participating detectors) and an incoherent (where the signal in each detector is assumed independent) model [122]. Both the coherent and incoherent models are constructed from a superposition of Morlet-Gabor wavelets, where the number of wavelets is in itself a variable, as implemented in BAYESWAVE [121]. This shows unequivocally that although the modelled analysis, where a known GR waveform approximants attempts to match the signal that best matches what is observed across the detector network, there is a significant fraction of observable coherent signal left. The properties of this left-over signal are not strongly constrained by this analysis however, as expected by the typical SNR ∼12 for these residual signals. The excluded event, with properties listed in Table VI, has a Bayes factor of $\sim 4 \times 10^{11}$.

V. DISCUSSION

A. Summary of results

In this study we have looked at the impact of inaccuracies in models of the GW waveform on inferring parameters for single loud events and for populations of binary black holes BBH. In Sec. III we presented results from parameter estimation analyses with current waveform models (`IMRPhenomPv2`, `SEOBNRv4_ROM`) for two simulated PN-NR signals at fixed luminosity distance for a series of detector networks. These "golden binaries" are therefore observed with increasingly high SNR as we look toward future detectors which are about a hundred times more sensitive than the current ones. From the posterior distributions we calculated systematic and statistical errors and produced a tuned version of the indistinguishability criterion [see Eqs. (8) and (9)]. In Fig. 2 we show the resulting "acceptable error" as a function of SNR. The main result of this paper shows that current waveform models used as templates in our PE analyses need to be improved for aLIGO design sensitivity and beyond: For 3G detectors such as Cosmic Explorer and the Einstein Telescope, the mismatch error for semianalytical models needs to be reduced by *three orders of magnitude* and by *one order of magnitude* for NR waveforms.

In Sec. III C 2 we saw that waveform inaccuracies can come from a combination of factors: errors in the dominant $(2, \pm 2)$ modes in the coprecessing frame, approximate modeling of the precessing reference frame of the binary, and from missing higher harmonics in the waveform. Better semianalytical models that include more physics are becoming available [79,80,116,132–135].

It stands to reason that if inferred binary parameters for single events are affected by inaccuracies in waveform models, then these deficiencies will also impact the analysis of populations of compact binaries. In populations, many events will be significantly weaker than the loud "golden binaries" we have considered before. Still, many small errors may sum up to give a sizable effect that can impact analyses. Therefore, in Sec. IV we presented a study for one hundred high mass BBHs mock signals (either `NRSur7dq2` or `IMRPhenomPv2`) drawn from an astrophysically motivated distribution in the intrinsic parameters. We again performed PE with the semianalytical `IMRPhenomPv2` model for the aLIGO-Virgo-KAGRA design sensitivity network.

In Fig. 6 we find that parameter biases between key parameters such as chirp mass, mass-ratio, and effective spin are strongly correlated, the population sum of these biases is nonzero for the `NRSur7dq2` signals, and the largest parameter biases lie outside 90% credible intervals. Posteriors for `IMRPhenomPv2` signals still show the correlations, but as shown in Table V the population sum of their biases is close to zero.

The residual between the GW data recorded by a detector and the best fit template waveform obtained from PE can be analysed further. If the waveform template cannot capture all

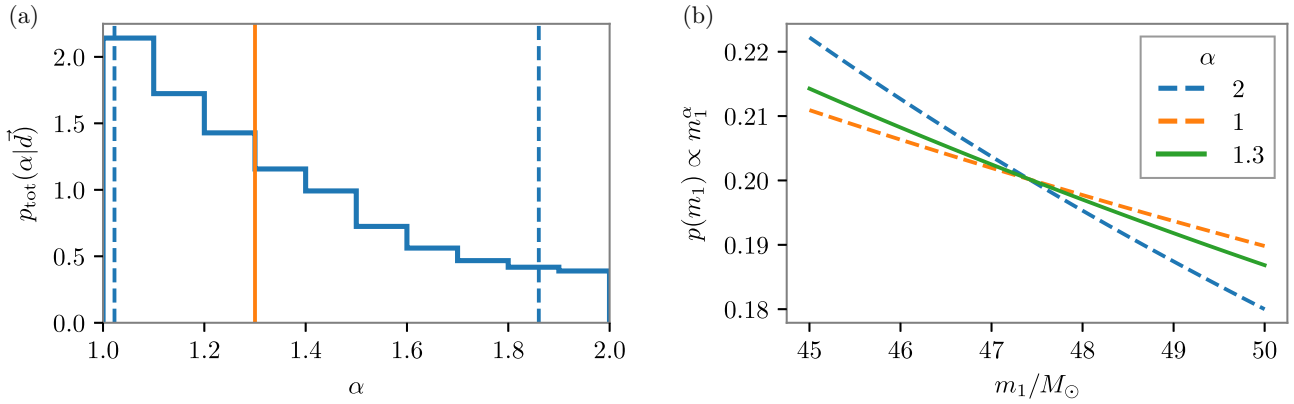

FIG. 11. (a) Hyperposterior $p_{\rm tot}(\alpha|\vec{d}|)$ for power-law index α for m_1. (b) Power-law PDFs for m_1 for $\alpha = 1.3$ (true value), and the bounds of the uniform hyperprior on α: $\alpha = 1, 2$.

the features in the signal, then the residuals (for the detectors in a network) will contain some coherent power (i.e., the residuals are not just due to random noise fluctuations in each detector). We show that this is the case for a `NRSur7dq2` event in our population study which has significant SNR in its residual (see Fig. 8), and significant power in the time frequency plane (see Fig. 9). We have also carried out a BAYESWAVE analysis and show in Fig. 10 the Bayes factors between a coherent and an incoherent wavelet model for the population events. Most events have a residual SNR of about 12 and Bayes factors of about 60 in favor of the coherent model while the event with the loudest residual has a residual SNR of 18 and BF of 4×10^{11}.

We have computed the population hyperposterior of the power-law index of the larger BH, the only free parameter in the distribution of source parameters. Due to the shortness of the signals in time, the hyperposterior is not very informative, but it shows preference for the lower bound of the prior $\alpha = 1$ over then upper bound $\alpha = 2$, and is thus closer to the true value $\alpha = 1.3$.

B. Outlook

Let us discuss several further implications of systematic errors in measured binary parameters caused by inaccuracies in waveform models. They concern the astrophysical relevance of biases. the future of waveform modeling and NR simulations, and how tests of GR will be affected.

In this study we have reported extensively about biases in inferred binary parameters. How much should we care about these biases? Beyond the simple statement that parameter biases will matter more when they are large, we would like to point out particular situations when biases are especially important and can severely impact the interpretation of GW observations. Severe biases could cause a misidentification of the class of a compact binary, e.g., confusing BNS, NSBH, and BBH sources near the lower mass gap [136,137]. Large biases in spin parameters such as the effective precession spin $\chi_{\rm p}$ could lead to a misidentification of formation channel of a binary. This could also happen if the effective aligned spin parameter $\chi_{\rm eff}$ was heavily biased, but in general $\chi_{\rm eff}$ measurements are a lot more robust since this parameter is connected with the length of the inspiral signal [115]. We have seen in Fig. 3 that $\chi_{\rm p}$ can indeed be significantly underestimated, especially if the precession modulations are suppressed when the binary is viewed nearly face-on or face-off.

Extrinsic parameters are in general less affected by waveform systematics and we do not expect their measurement errors to have a big impact. Sky location parameters enter in the detector pattern functions and should not be affected. We expect luminosity distance measurements to be affected mainly through their correlation with the binary's inclination angle. The latter can be better measured [132,133,138–140] when the waveform includes higher harmonics beyond the dominant $(2, \pm 2)$ modes. Amplitude errors should play a lesser role than phase errors, which could lead to us to misestimate $\mathcal{M}$ and thus bias the recovered distance. If we misestimate $\mathcal{M}$ due to phase errors, then that will also bias the recovered distance. Through this correlation mis-estimation of distance can lead to additional bias on the source-frame masses. This can be significant for very distant binaries. Finally, based on the discussion in Sec. IV B we expect that for population analyses parameters characterizing the mass and spin distributions will be affected to some degree since the events making up the population will suffer some amount of parameter biases.

How can waveform models be improved and made ready for the planned future 3G detectors, such as Cosmic Explorer and Einstein Telescope? On the one hand, the accuracy in the inspiral regime needs to be improved. This requires a higher order and more complete PN description and further work on re-summation and effective-one-body theory to extend the validity of the inspiral to higher frequencies. The inclusion of self force terms into effective-one-body (EOB) could help accuracy for large mass ratios [141]. Post-Minkowskian results obtained with modern scattering amplitude methods could be useful to improve the accuracy, if pushed to higher order [142,143]. PN calculations have being made at 4PN order for nonspinning BBHs [144–147] and were recently extended to 5PN order [148,149]. Practical semianalytic inspiral-merger-ringdown waveform models for BBHs, whether they are phenomenological or EOB models, require more NR waveforms covering larger parts of the binary parameter space and ultimately higher NR accuracy. Especially for unequal mass-ratios we will also require longer NR simulations in time to be able to combine NR waveforms with PN or EOB inspirals to form highly accurate hybrid waveforms [62–64], and to better determine inspiral coefficients in the construction of EOB models [68]. So far semianalytic models have been tuned only in the non-precessing sector. Extending calibration as more precessing NR simulations are becoming available will be essential to improve their accuracy. In addition, a novel NR-independent, analytical approach for modeling the merger has been put forward [150]. The accuracy of this approach beyond NR accuracy could be assessed with constraints on waveforms obtained from balance laws at future null infinity [151].

Surrogate and reduced order models of NR waveforms [81,152–155] and of EOB waveform models [68,74,75,156,157] have come to prominence in the past several years. They preserve the accuracy of the training set waveforms they are constructed from and are orders of magnitude faster to evaluate making them crucial for data analysis applications. They depend on their input data and so their accuracy is limited by the accuracy of the training set waveforms, and the requirement that the training data is sufficiently dense in the parameter space, since they need to fit or interpolate waveform coefficients over parameter space.

Waveform models should also include all physical effects that will leave a measurable trace in the emitted GW signal. This includes spin effects (aligned and precessing spins), higher harmonics beyond the dominant $(2, \pm 2)$ modes in the waveform, imprints of eccentricity, and tidal effects if the binary contains at least one neutron star. As the number of waveform parameters increases it becomes harder to carry out enough NR simulations to accurately tune models. A further desirable improvement for waveform models is to also model internal errors in waveform models and marginalize over these parameters in PE, which can be achieved with Gaussian process regression (GPR) [55,158–162]. Posterior distributions

obtained with such models should be more accurate (reduced bias) but somewhat less precise.

We have seen that NR waveforms are central for IMR waveform modeling. According to our results, NR waveforms will have to be improved in the future along three different dimensions: First, accuracy; second, length; third improved parameter space coverage. It turns out that each of these aspects will make simulations more expensive. Regarding length, the cost of an NR simulation is at least proportional to the time-to-merger; hence, the cost will increase as $1/\eta\ (M\Omega_i)^{-8/3}$, so that starting at one half the initial (orbital) frequency $M\Omega_i$ increases cost by at least a factor of 5. This scaling of the time-to-merger already indicates that making the mass-ratio more unequal will also increase the computational cost: The time-to-merger (and this computational cost) will increase at least as $1/\eta$. In addition, current NR codes use explicit time integration and are therefore limited by the Courant–Friedrichs–Lewy condition [163], so that each time-step can cover at most a time-interval $\propto q$ (for $q \leqslant 1$), giving a second power of the mass-ratio. Regarding accuracy, it is difficult to predict how the achieved accuracy scales with computational cost; one estimate for SpEC is that the cost goes as $\epsilon^{-1/3}$ [164], where ϵ is the NR error. Therefore, reducing the mismatch-error by a factor of 10—at the same parameters and length of the simulation—increases computational cost by about 50% for SpEC since the mismatch error goes as the square of the NR error [99,101,165]. Finally, both higher spins and higher mass-ratio make NR simulations more expensive, with the mass-ratio dependence most pronounced.

The accuracy and number of NR simulations have improved dramatically since the breakthrough in 2005 [104,166,167]. What improvements can we expect for the future that can deliver the simulations needed to be ready for 3G science? We can no longer rely on Moore's law to deliver massive improvements of CPU clock speeds. Instead advances in CPU development have shifted to increasing the number of cores and to exploit that NR codes need better parallelization and scaling. New codes are being developed to address these accuracy and performance issues. The SpECTRE code [168,169] from the SXS Collaboration uses task-based parallelism combined with the discontinuous Galerkin method to significantly increase the efficiency and scalability of relativistic astrophysics simulations. Work to significantly reduce computational cost for NR simulations is also under way for finite difference codes [170]. These approaches could lead to a two order of magnitude improvement in efficiency and bring us closer to solving the problems we have pointed out here. In addition to the truncation error which results from the finite degree polynomial approximations to continuum derivatives in Einstein's equations, errors are made when extracting the GW waveform on computer grids extending finite distances away from the merging binary. Traditionally, the waves are extracted (ideally on spherical shells) at several radii as far away from the origin as possible and the ideal waveform at future null infinity is extrapolated from that data. The Cauchy characteristic extraction method [171–174] can compute the emitted GWs with higher accuracy and should be available for future NR simulations. Combining waveforms from SpEC and finite difference codes by hybridization is a promising technique for especially challenging configurations [175].

Our study on the impact of waveform inaccuracies should be extended to tests of GR which we expect to be especially susceptible to systematic effects which could be misinterpreted as genuine deviations from GR. All of the current tests of GR [21] should be scrutinized. This includes tests on the distribution of the SNR of residuals in detector noise, testing whether the final mass and spin inferred from the low and high frequency parts of the GW signal are consistent, computing posterior distributions of deviations in, e.g., PN waveform coefficients, computing posteriors on parameters in phenomenological dispersion relations and tests that put constraints on alternative GW polarizations. Ultimately, tests of GR should be done by estimating parameters of waveform models for alternative theories of gravity, along with Bayesian model comparisons. Work is under way to identify well-posed alternative theories of gravity [176–180] and to numerically compute what the emitted GW will look like in the strong field regime [181].

Finally, we expect that LISA analyses of massive BBHs, which are should have SNRs of hundreds to thousands, will be affected in similar ways as demonstrated here for 3G ground-based detectors [22]. Updated estimates for current IMR waveform models will need to be explored in future studies.

ACKNOWLEDGMENTS

This work was stimulated by the Gravitational Wave International Committee (GWIC) 3G science-case study [182]. The authors thank Harald Pfeiffer, Katerina Chatziioannou, Will Farr, Sergei Ossokine, John Veitch, Alessandra Buonanno, Mark Hannam, Salvatore Vitale, Frank Ohme, Sascha Husa, Badri Krishnan, and Bruce Allen for useful discussions. We thank Patricia Schmidt for help with the LVC NR injection infrastructure, Tito dal Canton for help with PyCBC's qscans, and Sergei Ossokine and Stas Babak for the code to compute the max-max overlap. C.-J.H. acknowledges support of the MIT physics department through the Solomon Buchsbaum Research Fund, the National Science Foundation, and the LIGO Laboratory. The authors acknowledge usage of LIGO Data Grid clusters and AEI's Slarti computer. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under Cooperative Agreement No. PHY-0757058. This is LIGO Document No. DCC-1900377.

APPENDIX A: HYBRIDIZATION PROCEDURE

We construct hybrid waveforms by combining multimodal precessing PN and NR waveforms using the `GWFrames` [60] code. The code first reads the NR waveform data and transforms it to the corotating frame [61] and shifts it in time so that the merger lies at $t = 0$. Data for the evolution of the positions, masses, and spin vectors of the BHs as determined by locating their apparent horizons in the NR code is read in as well. Next, the separation vector between the two BHs and the orbital frequency are computed, along with the rotor of the

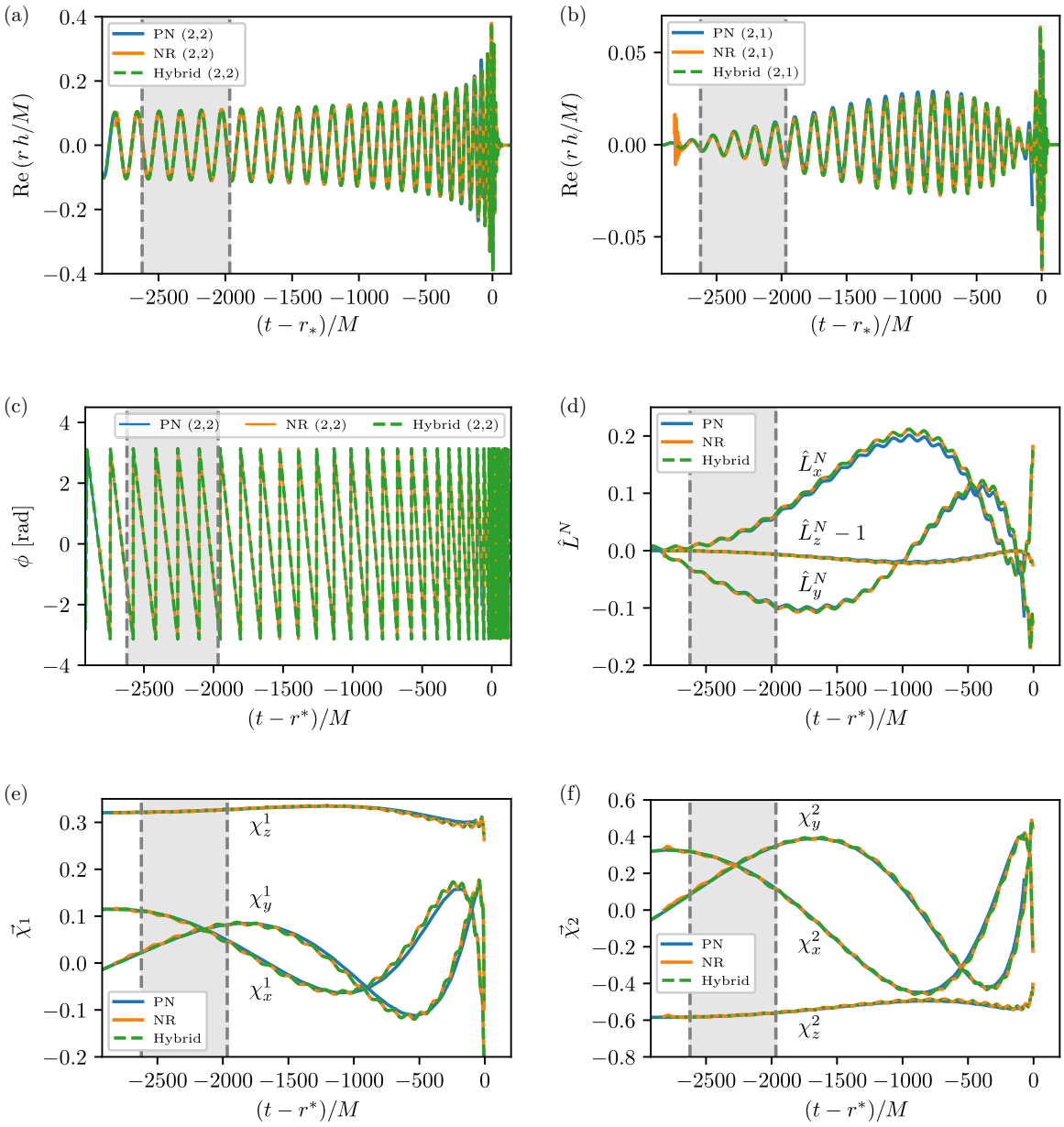

FIG. 12. PN-NR hybrid for SXS_BBH_0308 with the `SpinTaylorT1` approximant. For each quantity we show the PN, NR, and hybrid data as a function of retarded time before merger. The waveforms have been blended together in the gray shaded hybridization region. (a) Real part of the $(2, 2)$ and (b) $(2, 1)$ modes in the inertial frame. (c) (Wrapped) phase of the $(2, 2)$ mode in the inertial frame. (d) Cartesian components of the Newtonian orbital angular momentum unit vector in the inertial frame (e) and (f) dimensionless spin vectors of the BHs.

reference frame at the relaxed time (after the junk radiation has passed).

We compute a PN waveform from the `PNWaveform` package included in the `GWFrames` code. The PN implementation includes nonspinning orbital binding up to 4 pN [183]. The 5 pN term is set to zero. Spin-orbit terms in the angular momentum are included up to 3.5 pN [84]. Nonspinning flux terms are included up to 3.5 PN [183], and higher-order terms from [184] up to 6 PN along with absorption terms from [185]. Spin-spin and spin-orbit squared terms at 2 PN order are included [83,186,187] and spin-orbit terms in the flux are included up to 4.0 PN [188]. Precession of the orbital angular velocity and spins follows [83,84,189]. Expressions for waveform modes are taken from [190–193]. We use the `SpinTaylorT1` and `SpinTaylorT4` implemented in this code which are simply called `TaylorT1` and `TaylorT4` there, but we add the prefix `Spin` to make it clear that they support precession.

Initial data for the PN integration is set at the NR relaxed time and the PN equations are also evolved backwards in time to the desired starting orbital frequency $M\Omega_i$. The PN waveform is then transformed to the corotating frame. To prepare for hybridization, the PN and NR waveforms are aligned by minimizing the distance between their rotors in their corotating frames. The aligned waveforms are then blended and hybridized.

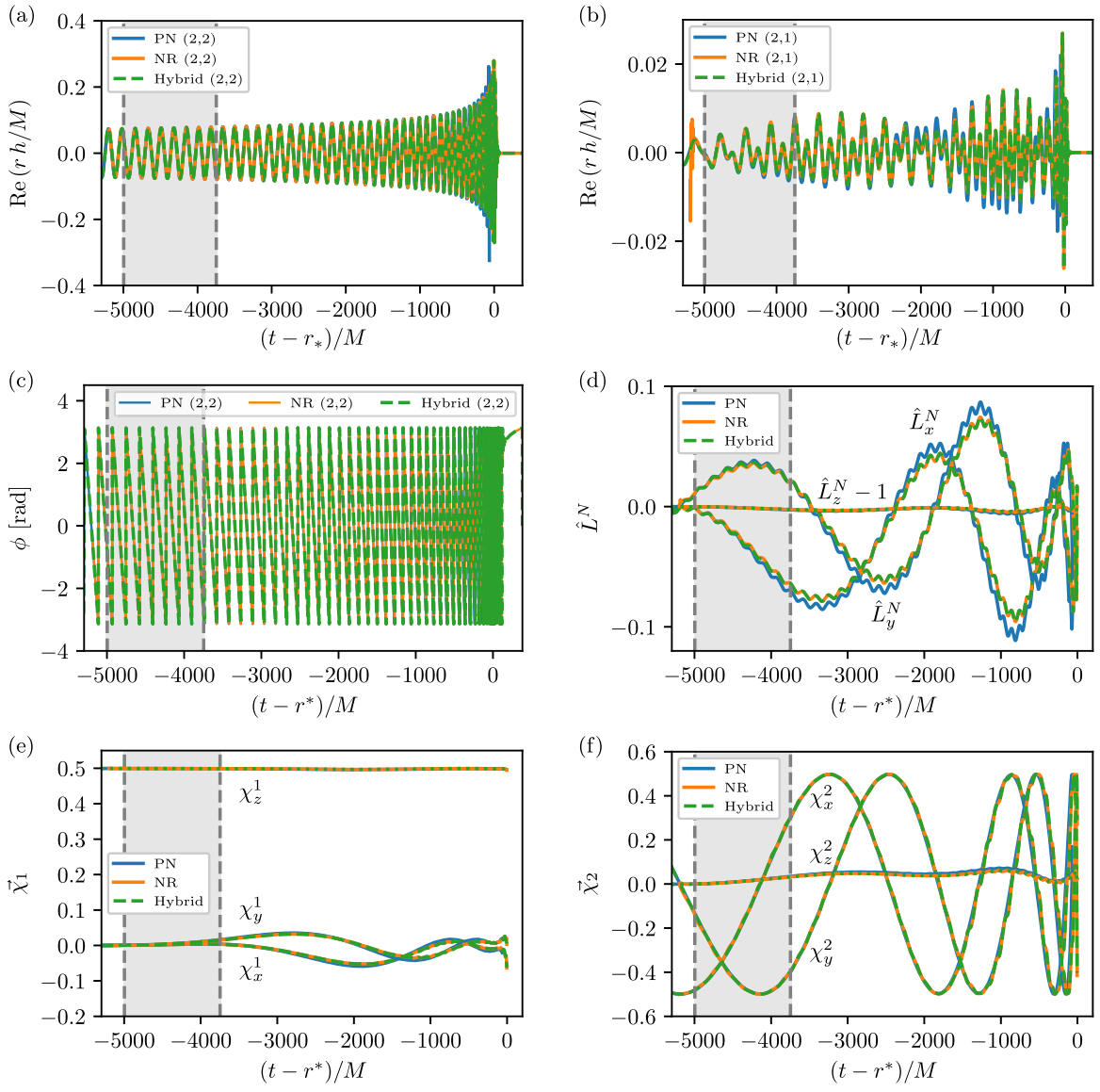

FIG. 13. PN-NR hybrid for SXS_BBH_0104 with the `SpinTaylorT1` approximant. For each quantity we show the PN, NR, and hybrid data as a function of retarded time before merger. The waveforms have been blended together in the gray shaded hybridization region. (a) Real part of the $(2, 2)$ and (b) $(2, 1)$ modes in the inertial frame. (c) (Wrapped) phase of the $(2, 2)$ mode in the inertial frame. (d) Cartesian components of the Newtonian orbital angular momentum unit vector in the inertial frame (e) and (f) dimensionless spin vectors of the BHs.

In this study we choose $M\Omega_i = 0.002$ due to computational restrictions. This corresponds to $f_{\rm GW} \approx 1.74\,$Hz for the (2,2) mode. Higher (ℓ, m) modes in the waveform enter the frequency band at $m/2$ times the frequency at which the (2,2) mode enters. Therefore, some of the higher harmonics are truncated at low frequencies but this effect is minor because they are very small compared to the dominant modes.

To use the LVC NR-injection infrastructure [194] we also hybridize dynamics quantities, namely the spin vectors, orbital frequency, the Newtonian orbital angular momentum vector, the vector $\hat{n}$ pointing from one BH to the other, and the position vectors of the BHs. This allows us to define the spin vectors at a particular reference frequency and to output the result in "LVC NR" format.

Figures 12 and 13 show selected waveform modes, the phase of the (2,2) mode, the orbital angular momentum vector and the spin vectors for the two configurations used in this study. These plots demonstrate the good blending between the PN and NR data in the hybridization time region (gray shaded). The absolute value of inertial and coprecessing frame modes for the hybrids are shown in Fig. 14. Higher harmonics are stronger for the more unequal mass SXS_BBH_0104 configuration, whereas precession effects that give rise to modes like the (2,1) and (3,2) mode are stronger for SXS_BBH_0308.

For SXS_BBH_0308 there is a disagreement in the corotating frame (2,1) mode between PN and NR. This mode is weak since the system is almost equal mass which likely exacerbates the disagreement. In contrast, we find excellent agreement in the same mode for SXS_BBH_0104. The effect of this discrepancy for SXS_BBH_0308 is very small. The mismatch between the hybrid with and without the corotating

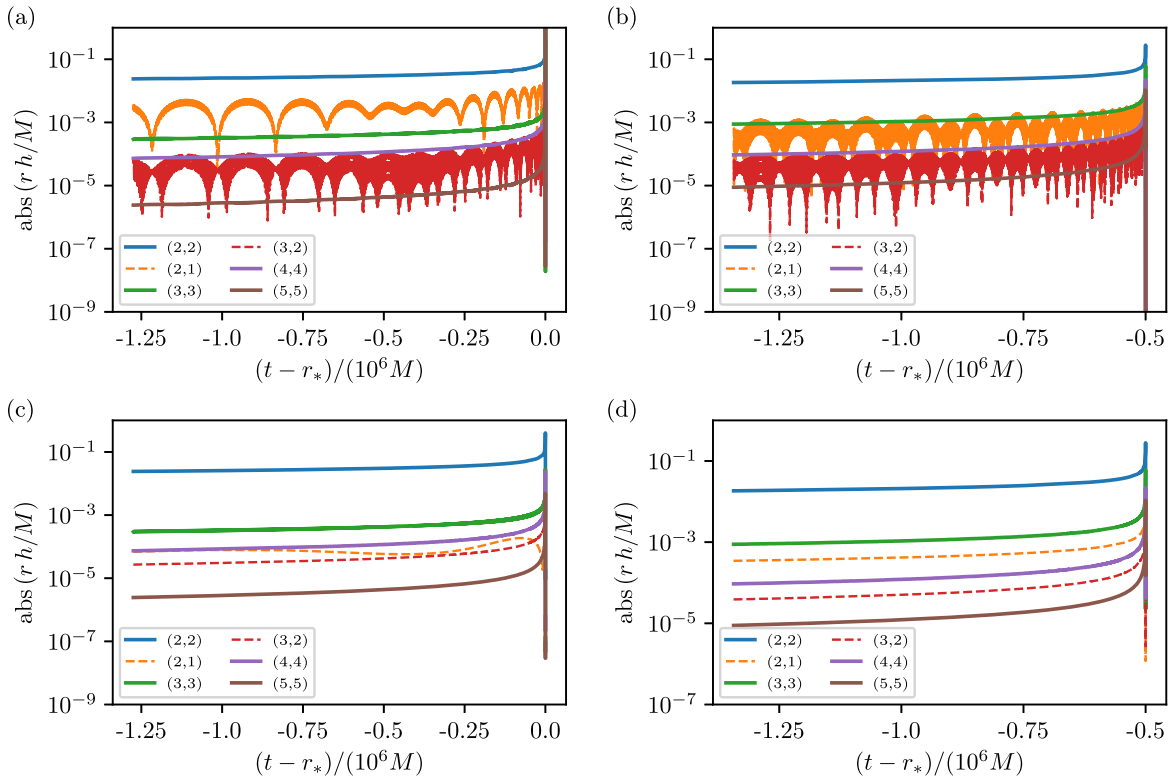

FIG. 14. Selected modes for PN-NR hybrids with the `SpinTaylorT1` approximant. Waveform modes are shown in the inertial frame [panels (a) and (b)] and coprecessing frame [panels (c) and (d)]. The hybrids start at an orbital frequency of $M\Omega = 0.002$. Results are shown for configurations SXS_BBH_0308 (left column) and SXS_BBH_0104 (right column). Modes in the coprecessing frame are close to nonprecessing waveforms, while inertial modes are modulated by the precession of the orbital plane.

frame $(2, \pm 1)$ modes is on the order of hybridization error and NR error, about $\sim 10^{-5}$.

To study the error introduced by hybridizing PN and NR waveforms the optimal test would be to compute the mismatch of a hybrid against a very long high accuracy NR waveform that fills the detector band. Since this is in practice not possible we perform the following experiments to make sure that hybridization errors are subdominant. We compute

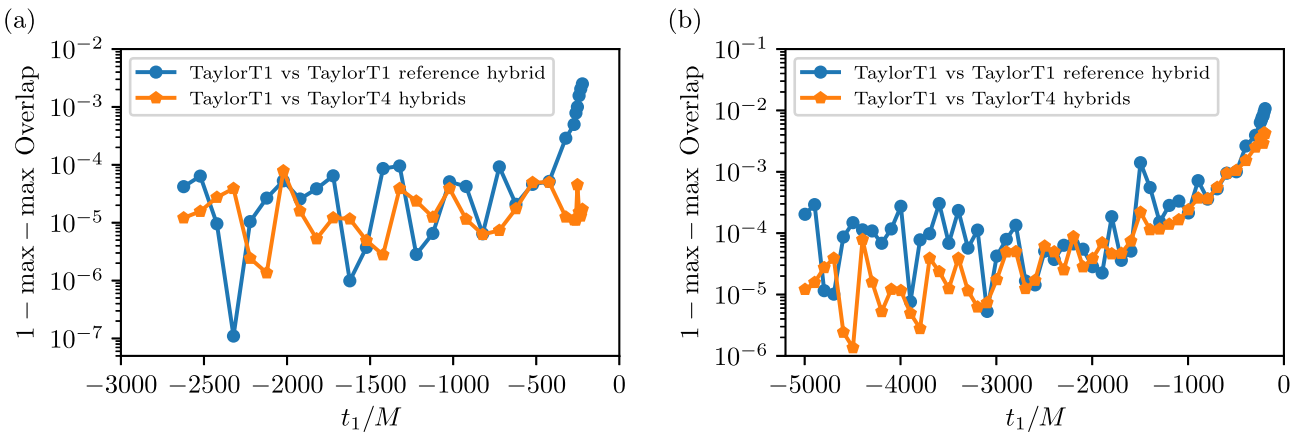

FIG. 15. Mismatch between PN-NR hybrid waveforms, shown for (a) SXS_BBH_0308 and (b) SXS_BBH_0104. The mismatch is computed as one minus the max - max overlap including higher modes up to $\ell = 8$ in the NR waveforms. Here, hybrids are constructed from an orbital frequency of $M\Omega = 0.01$ and the overlap integral starts at 10 Hz and uses the aLIGO design PSD. (Therefore, higher harmonics will be incomplete in the frequency band, but in a consistent manner. Degradation of the mismatch will come from high frequencies close to merger.) The blue curves show mismatches between `SpinTaylorT1`-NR hybrids constructed with a $100M$ hybridization window as a function of the start time t_1 of this window before merger against a reference hybrid with a broader window at $t = [200, 800]M$ measured from the beginning of the NR waveform, and after the relaxation time. The binaries merge at $2822.24M$ and $5196.2M$ from the beginning of the NR simulations, respectively, for SXS_BBH_0308 and SXS_BBH_0104. The orange curves show mismatches between `SpinTaylorT1`-NR and `SpinTaylorT4`-NR hybrids constructed with the same $100M$ hybridization window as a function of the start time t_1.

overlaps between (a) a reference hybrid in the time window $t = [200, 800]M$, measured from the beginning of the NR waveform, and "sliding hybrids," a series of hybrids blended with $100M$ long time windows that approach the merger in discrete steps. We also compute (b) overlaps between sliding hybrids for the same window starting time between the `SpinTaylorT1` and `SpinTaylorT4` PN approximants. The reference hybrids are used as a signal waveforms for PE in the main study of the paper. We compute the max-max overlaps as defined in Appendix B of Ref. [77]. We show the resulting overlaps for SXS_BBH_0308 and SXS_BBH_0104 in Fig. 15. Both curves show that if one hybridizes early the mismatch is small and noisy. These mismatches are lower than the mismatches between different NR resolutions quoted in Sec. II B. Therefore hybridization errors are subdominant for these configurations. For SXS_BBH_0308 the mismatch only rises beyond 10^{-4} for windows that start within $500M$ of the merger, while for SXS_BBH_0104 the mismatches approach 10^{-3} already $1000M$ before merger. In this regime PN waveforms become inaccurate compared to NR and differences between PN approximants grow.

We also want to briefly mention additional sources of errors. Spin vectors are defined differently in PN and NR [195–197], and therefore, using the same spin values for both waveforms at the same time as we do in the hybrid construction will introduce an additional error that we do not quantify here. The $m = 0$ "memory" modes may not be accurate without using Cauchy characteristic extraction (CCE) [172]. The waveforms used in this study, SXS_BBH_0308 and SXS_BBH_0104 do not use CCE.

The configurations considered in this study are fairly easy to hybridize and one should not infer a general behavior of hybridization errors from them. For more challenging configurations (higher mass-ratios and spins) PN and NR are expected to show discrepancies further away from merger. How long NR waveforms need to be so that hybridization errors are subdominant requires detailed study [62–65].

[1] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. X **9**, 031040 (2019).

[2] J. Aasi *et al.* (LIGO Scientific), Classic. Quant. Grav. **32**, 115012 (2015).

[3] F. Acernese *et al.* (VIRGO), Classic. Quant. Grav. **32**, 024001 (2015).

[4] Y. Aso, Y. Michimura, K. Somiya, M. Ando, O. Miyakawa, T. Sekiguchi, D. Tatsumi, and H. Yamamoto (KAGRA), Phys. Rev. D **88**, 043007 (2013).

[5] LIGO Scientific Collaboration, Instrument Science White Paper 2018, https://dcc.ligo.org/T1800133/public (2018), Technical Report LIGO-T1800133.

[6] B. P. Abbott *et al.*, Living Rev. Rel. 21 (2018).

[7] National Science Foundation, The A+ Upgrade to Advanced LIGO, https://www.nsf.gov/awardsearch/showAward?AWD_ID=1834382 (2018).

[8] Funding has already been awarded to upgrade the LIGO detectors as part of the A+ project [7].

[9] M. Punturo *et al.*, Proceedings of the 14th Workshop on Gravitational Wave Data Analysis (GWDAW'14), Classic. Quant. Grav. **27**, 194002 (2010).

[10] B. P. Abbott *et al.* (LIGO Scientific), Classic. Quant. Grav. **34**, 044001 (2017).

[11] D. Reitze *et al.*, Bull. Am. Astron. Soc. **51**, 035 (2019), arXiv:1907.04833.

[12] D. Reitze *et al.*, Bull. Am. Astron. Soc. **51**, 141 (2019), arXiv:1903.04615.

[13] P. Amaro-Seoane *et al.* (LISA), arXiv:1702.00786.

[14] J. Luo *et al.* (TianQin), Classic. Quant. Grav. **33**, 035010 (2016).

[15] V. Kalogera *et al.*, arXiv:1903.09220.

[16] B. S. Sathyaprakash *et al.*, arXiv:1903.09221.

[17] S. Vitale, W. M. Farr, K. Ng, and C. L. Rodriguez, Astrophys. J. **886**, L1 (2019).

[18] W. M. Farr, M. Fishbach, J. Ye, and D. Holz, Astrophys. J. **883**, L42 (2019).

[19] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Classic. Quant. Grav. **34**, 104002 (2017).

[20] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 061102 (2016).

[21] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. D **100**, 104036 (2019).

[22] C. Cutler and M. Vallisneri, Phys. Rev. D **76**, 104018 (2007).

[23] N. Yunes and F. Pretorius, Phys. Rev. D **80**, 122003 (2009).

[24] E. Berti, K. Yagi, and N. Yunes, Gen. Relativ. Gravit. **50**, 46 (2018).

[25] E. Berti, K. Yagi, H. Yang, and N. Yunes, Gen. Relativ. Gravit. **50**, 49 (2018).

[26] D. M. Eardley, D. L. Lee, and A. P. Lightman, Phys. Rev. D **8**, 3308 (1973).

[27] S. Vitale and M. Evans, Phys. Rev. D **95**, 064052 (2017).

[28] S. Vitale, R. Lynch, V. Raymond, R. Sturani, J. Veitch, and P. Graff, Phys. Rev. D **95**, 064053 (2017).

[29] S. Vitale and C. Whittle, Phys. Rev. D **98**, 024029 (2018).

[30] B. P. Abbott *et al.* (LIGO Scientific, Virgo, ASKAP, BOOTES, DES, Fermi GBM, Fermi-LAT, GRAWITA, INTEGRAL, iPTF, InterPlanetary Network, J-GEM, La Silla-QUEST Survey, Liverpool Telescope, LOFAR, MASTER, MAXI, MWA, Pan-STARRS, PESSTO, Pi of the Sky, SkyMapper, Swift, C2PU, TOROS, VISTA), Astrophys. J. **826**, L13 (2016).

[31] Z. Doctor *et al.* (DES), Astrophys. J. **873**, L24 (2019).

[32] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. **875**, 161 (2019).

[33] B. F. Schutz, Nature **323**, 310 (1986).

[34] B. P. Abbott *et al.* (LIGO Scientific, Virgo), arXiv:1908.06060.

[35] M. Soares-Santos *et al.* (DES, LIGO Scientific, Virgo), Astrophys. J. **876**, L7 (2019).

[36] R. Gray *et al.*, arXiv:1908.06050.

[37] H.-Y. Chen, M. Fishbach, and D. E. Holz, Nature **562**, 545 (2018).

[38] GW observations measure a signal that has been redshifted during its propagation, thus changing the recovered BH masses. The source-frame parameters are inferred by taking the measured luminosity distance, assuming a cosmology

and applying the appropriate redshift factor to the measured detector-frame parameters.

[39] C. Cahillane *et al.* (LIGO Scientific), Phys. Rev. D **96**, 102001 (2017).

[40] T. B. Littenberg and N. J. Cornish, Phys. Rev. D **91**, 084034 (2015).

[41] K. Chatziioannou, C.-J. Haster, T. B. Littenberg, W. M. Farr, S. Ghonge, M. Millhouse, J. A. Clark, and N. Cornish, Phys. Rev. D **100**, 104004 (2019).

[42] C. Röver, R. Meyer, and N. Christensen, Classic. Quant. Grav. **28**, 015010 (2011).

[43] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).

[44] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. X **6**, 041015 (2016); **8**, 039903(E) (2018).

[45] M. Evans, R. Sturani, S. Vitale, and E. Hall, Unofficial sensitivity curves (ASD) for aLIGO, Kagra, Virgo, Voyager, Cosmic Explorer, and ET, Tech. Rep. (2018), https://dcc.ligo.org/LIGO-T1500293/public.

[46] L. Barsotti, P. Fritschel, M. Evans, and S. Gras, Updated Advanced LIGO sensitivity design curve, Tech. Rep. (2018), https://dcc.ligo.org/LIGO-T1800044/public.

[47] M. Boyle *et al.*, Classic. Quant. Grav. **36**, 195006 (2019).

[48] The SXS Collaboration, SpEC, https://www.black-holes.org/code/SpEC.html (2019).

[49] S. Ossokine, L. Kidder, H. Pfeiffer, M. Scheel, M. Boyle, D. Hemberger, G. Lovelace, and B. Szilagyi, *Binary black-hole simulation sxs:bbh:0308* (Zenodo, 2018), https://doi.org/10.5281/zenodo.1215583.

[50] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 241102 (2016).

[51] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. D **94**, 064035 (2016).

[52] L. Kidder, G. Lovelace, D. Hemberger, H. Pfeiffer, M. Boyle, M. Scheel, A. Mroue, N. Taylor, B. Szilagyi, and A. Zenginoglu, *Binary black-hole simulation sxs:bbh:0104* (Zenodo, 2018), https://doi.org/10.5281/zenodo.1213396.

[53] A. H. Mroue *et al.*, Phys. Rev. Lett. **111**, 241104 (2013).

[54] A. Mroue, M. Boyle, G. Lovelace, B. Szilagyi, H. Pfeiffer, A. Zenginoglu, L. Kidder, N. Taylor, D. Hemberger, and M. Scheel, *Binary black-hole simulation sxs:bbh:0053* (Zenodo, 2019), https://doi.org/10.5281/zenodo.3311880.

[55] E. A. Huerta *et al.*, Phys. Rev. D **97**, 024031 (2018).

[56] I. Hinder, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. D **98**, 044015 (2018).

[57] Z. Cao and W.-B. Han, Phys. Rev. D **96**, 044028 (2017).

[58] D. Chiaramello and A. Nagar, arXiv:2001.11736.

[59] I. M. Romero-Shaw, P. D. Lasky, and E. Thrane, Mon. Not. R. Astron. Soc. **490**, 5210 (2019).

[60] M. Boyle, Gwframes, https://github.com/moble/GWFrames (2019).

[61] M. Boyle, Phys. Rev. D **87**, 104006 (2013).

[62] M. Hannam, S. Husa, F. Ohme, and P. Ajith, Phys. Rev. D **82**, 124052 (2010).

[63] I. MacDonald, S. Nissanke, H. P. Pfeiffer, and H. P. Pfeiffer, Theory meets data analysis at comparable and extreme mass ratios. Proceedings of the NRDA/CAPRA 2010 Conference, Classic. Quant. Grav. **28**, 134002 (2011).

[64] M. Boyle, Phys. Rev. D **84**, 064013 (2011).

[65] P. Ajith *et al.*, Gravitational waves. Numerical relativity-data analysis. Proceedings of the 9th Edoardo Amaldi Conference, Amaldi 9, and meeting, NRDA 2011, Classic. Quant. Grav. **29**, 124001 (2012) [Erratum: **30**, 199401 (2013)].

[66] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).

[67] A. Bohé, M. Hannam, S. Husa, F. Ohme, M. Pürrer, and P. Schmidt, PhenomPv2—Technical notes for the LAL implementation, Tech. Rep. (2016), https://dcc.ligo.org/LIGO-T1500602/public.

[68] A. Bohé *et al.*, Phys. Rev. D **95**, 044028 (2017).

[69] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044007 (2016).

[70] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, Phys. Rev. D **93**, 044006 (2016).

[71] P. Schmidt, M. Hannam, and S. Husa, Phys. Rev. D **86**, 104063 (2012).

[72] P. Schmidt, F. Ohme, and M. Hannam, Phys. Rev. D **91**, 024043 (2015).

[73] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, Phys. Rev. D **94**, 044031 (2016).

[74] M. Pürrer, Classic. Quant. Grav. **31**, 195010 (2014).

[75] M. Pürrer, Phys. Rev. D **93**, 064041 (2016).

[76] 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).

[77] S. Babak, A. Taracchini, and A. Buonanno, Phys. Rev. D **95**, 024010 (2017).

[78] T. D. Knowles, C. Devine, D. A. Buch, S. A. Bilgili, T. R. Adams, Z. B. Etienne, and S. T. Mcwilliams, Classic. Quant. Grav. **35**, 155003 (2018).

[79] S. Ossokine *et al.*, arXiv:2004.09442.

[80] S. Khan, K. Chatziioannou, M. Hannam, and F. Ohme, Phys. Rev. D **100**, 024059 (2019).

[81] 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).

[82] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne, Phys. Rev. D **49**, 6274 (1994).

[83] L. E. Kidder, Phys. Rev. D **52**, 821 (1995).

[84] A. Bohe, S. Marsat, G. Faye, and L. Blanchet, Classic. Quant. Grav. **30**, 075017 (2013).

[85] S. Ossokine, M. Boyle, L. E. Kidder, H. P. Pfeiffer, M. A. Scheel, and B. Szilágyi, Phys. Rev. D **92**, 104028 (2015).

[86] LALInference, https://git.ligo.org/lscsoft/lalsuite/tree/master/lalinference (2019).

[87] LIGO Scientific Collaboration, LIGO Algorithm Library-LALSuite, free software (GPL) (2019).

[88] G. Ashton *et al.*, Astrophys. J. Suppl. **241**, 27 (2019).

[89] Bilby, https://git.ligo.org/lscsoft/bilby (2019).

[90] J. S. Speagle, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).

[91] C. Cutler and E. E. Flanagan, Phys. Rev. D **49**, 2658 (1994).

[92] A. Viets *et al.*, Classic. Quant. Grav. **35**, 095015 (2018).

[93] W. M. Farr, B. Farr, and T. Littenberg, Modeling calibration errors in CBC waveforms, Tech. Rep. LIGO-T1400682 (LIGO Project, 2015).

[94] B. P. Abbott *et al.* (LIGO Scientific), Phys. Rev. D **95**, 062003 (2017).

[95] For LIGO's second observing run, the detectors had calibration uncertainties of $\delta A \sim 3\%$ and $\delta\phi \sim 2^\circ$ [1].

[96] J. Kissel, Calibration Considerations for the 3G Detector Era, in *Proceedings of the Gravitational Wave Data Analysis Workshop* (2017), https://dcc.ligo.org/LIGO-G1700810/public.

[97] J. Kissel, Absolute Calibration of a Galactic Gravitational Wave Detector Network, in *Proceedings of the Gravitational Wave Data Analysis Workshop* (2018), https://dcc.ligo.org/LIGO-G1800953/public.

[98] E. Thrane and C. Talbot, Publ. Astron. Soc. Aust. **36**, 10 (2019).

[99] E. E. Flanagan and S. A. Hughes, Phys. Rev. D **57**, 4566 (1998).

[100] L. Lindblom, B. J. Owen, and D. A. Brown, Phys. Rev. D **78**, 124020 (2008).

[101] S. T. McWilliams, B. J. Kelly, and J. G. Baker, Phys. Rev. D **82**, 024014 (2010).

[102] K. Chatziioannou, A. Klein, N. Yunes, and N. Cornish, Phys. Rev. D **95**, 104004 (2017).

[103] M. Vallisneri, Phys. Rev. D **77**, 042001 (2008).

[104] M. Campanelli, C. O. Lousto, P. Marronetti, and Y. Zlochower, Phys. Rev. Lett. **96**, 111101 (2006).

[105] B. Bruegmann, J. A. Gonzalez, M. Hannam, S. Husa, U. Sperhake, and W. Tichy, Phys. Rev. D **77**, 024027 (2008).

[106] S. Husa, J. A. Gonzalez, M. Hannam, B. Bruegmann, and U. Sperhake, Classic. Quant. Grav. **25**, 105006 (2008).

[107] J. S. Hesthaven and T. Warburton, *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis and Applications*, Texts in Applied Mathematics, Vol. 54 (Springer, Berlin, 2008), pp. xiv+500, algorithms, analysis, and applications.

[108] T. Vincent, H. P. Pfeiffer, and N. L. Fischer, Phys. Rev. D **100**, 084052 (2019).

[109] LALDetectors, https://git.ligo.org/lscsoft/lalsuite/blob/master/lal/lib/tools/LALDetectors.h (2019).

[110] E. Poisson and C. M. Will, Phys. Rev. D **52**, 848 (1995).

[111] E. Baird, S. Fairhurst, M. Hannam, and P. Murphy, Phys. Rev. D **87**, 024035 (2013).

[112] M. Pürrer, M. Hannam, and F. Ohme, Phys. Rev. D **93**, 084042 (2016).

[113] M. Pürrer, M. Hannam, P. Ajith, and S. Husa, Phys. Rev. D **88**, 064007 (2013).

[114] S. Vitale, R. Lynch, J. Veitch, V. Raymond, and R. Sturani, Phys. Rev. Lett. **112**, 251101 (2014).

[115] K. K. Y. Ng, S. Vitale, A. Zimmerman, K. Chatziioannou, D. Gerosa, and C.-J. Haster, Phys. Rev. D **98**, 083007 (2018).

[116] R. Cotesta, S. Marsat, and M. Pürrer, arXiv:2003.12079.

[117] E. D. Kovetz, I. Cholis, P. C. Breysse, and M. Kamionkowski, Phys. Rev. D **95**, 103010 (2017).

[118] D. Wysocki, J. Lange, and R. O'Shaughnessy, Phys. Rev. D **100**, 043012 (2019).

[119] C. Talbot and E. Thrane, Astrophys. J. **856**, 173 (2018).

[120] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. **882**, L24 (2019).

[121] N. J. Cornish and T. B. Littenberg, Classic. Quant. Grav. **32**, 135012 (2015).

[122] T. B. Littenberg, J. B. Kanner, N. J. Cornish, and M. Millhouse, Phys. Rev. D **94**, 044050 (2016).

[123] J. C. Brown, J. Acoust. Soc. Am. **89**, 425 (1991).

[124] A. Nitz, I. Harry, D. Brown, C. M. Biwer, J. Willis, T. D. Canton, C. Capano, L. Pekowsky, T. Dent, A. R. Williamson, M. Cabero, S. De, G. Davies, D. Macleod, B. Machenschalk, P. Kumar, S. Reyes, T. Massinger, F. Pannarale, M. Tápai, dfinstad, S. Fairhurst, S. Khan, A. Nielsen, shasvath, S. Kumar, I. Dorrington, L. Singer, H. Gabbard, and B. U. V. Gadre, *gwastro/pycbc: PyCBC Release v1.14.2* (Zenodo, 2019), https://doi.org/10.5281/zenodo.3483184.

[125] D. Gabor, J. Inst. Electr. Eng., Part 3 **93**, 429 (1946).

[126] This is also the model generally assumed in LALINFERENCE and BILBY.

[127] The nomenclature used internally by BAYESWAVE is to call the coherent model *signal* and the incoherent model *glitch*.

[128] J. B. Kanner, T. B. Littenberg, N. Cornish, M. Millhouse, E. Xhakaj, F. Salemi, M. Drago, G. Vedovato, and S. Klimenko, Phys. Rev. D **93**, 022002 (2016).

[129] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 221101 (2016) [Erratum: **121**, 129902 (2018)].

[130] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Class. Quantum Grav. **37**, 055002 (2020).

[131] M. Zevin *et al.*, Classic. Quant. Grav. **34**, 064003 (2017).

[132] 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).

[133] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, Phys. Rev. D **98**, 084028 (2018).

[134] S. Khan, F. Ohme, K. Chatziioannou, and M. Hannam, Phys. Rev. D **101**, 024056 (2020).

[135] B. Gadre, M. Pürrer, S. Field, and S. Ossokine (unpublished).

[136] I. Mandel, C.-J. Haster, M. Dominik, and K. Belczynski, Mon. Not. R. Astron. Soc. **450**, L85 (2015).

[137] T. B. Littenberg, B. Farr, S. Coughlin, V. Kalogera, and D. E. Holz, Astrophys. J. **807**, L24 (2015).

[138] P. B. Graff, A. Buonanno, and B. S. Sathyaprakash, Phys. Rev. D **92**, 022002 (2015).

[139] C. Kalaghatgi, M. Hannam, and V. Raymond, Phys. Rev. D **101**, 103004 (2020).

[140] F. H. Shaik, J. Lange, S. E. Field, R. O'Shaughnessy, V. Varma, L. E. Kidder, H. P. Pfeiffer, and D. Wysocki, arXiv:1911.02693.

[141] A. Antonelli, M. van de Meent, A. Buonanno, J. Steinhoff, and J. Vines, Phys. Rev. D **101**, 024024 (2020).

[142] A. Antonelli, A. Buonanno, J. Steinhoff, M. van de Meent, and J. Vines, Phys. Rev. D **99**, 104004 (2019).

[143] Z. Bern, C. Cheung, R. Roiban, C.-H. Shen, M. P. Solon, and M. Zeng, Phys. Rev. Lett. **122**, 201603 (2019).

[144] T. Damour, P. Jaranowski, and G. Schäfer, Phys. Rev. D **89**, 064058 (2014).

[145] T. Marchand, L. Bernard, L. Blanchet, and G. Faye, Phys. Rev. D **97**, 044023 (2018).

[146] S. Foffa, P. Mastrolia, R. Sturani, and C. Sturm, Phys. Rev. D **95**, 104009 (2017).

[147] J. Blümlein, A. Maier, P. Marquard, and G. Schäfer, arXiv:2003.01692.

[148] S. Foffa, P. Mastrolia, R. Sturani, C. Sturm, and W. J. Torres Bobadilla, Phys. Rev. Lett. **122**, 241605 (2019).

[149] J. Blümlein, A. Maier, and P. Marquard, Phys. Lett. B **800**, 135100 (2020).

[150] S. T. McWilliams, Phys. Rev. Lett. **122**, 191102 (2019).

[151] A. Ashtekar, T. De Lorenzo, and N. Khera, arXiv:1906.00913.

[152] J. Blackman, S. E. Field, C. R. Galley, B. Szilágyi, M. A. Scheel, M. Tiglio, and D. A. Hemberger, Phys. Rev. Lett. **115**, 121102 (2015).

[153] 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).
[154] V. Varma, S. E. Field, M. A. Scheel, J. Blackman, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. D **99**, 064045 (2019).
[155] 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).
[156] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Phys. Rev. X **4**, 031006 (2014).
[157] B. D. Lackey, S. Bernuzzi, C. R. Galley, J. Meidam, and C. Van Den Broeck, Phys. Rev. D **95**, 104036 (2017).
[158] C. J. Moore and J. R. Gair, Phys. Rev. Lett. **113**, 251101 (2014).
[159] C. J. Moore, C. P. L. Berry, A. J. K. Chua, and J. R. Gair, Phys. Rev. D **93**, 064001 (2016).
[160] Z. Doctor, B. Farr, D. E. Holz, and M. Pürrer, Phys. Rev. D **96**, 123011 (2017).
[161] B. D. Lackey, M. Pürrer, A. Taracchini, and S. Marsat, Phys. Rev. D **100**, 024002 (2019).
[162] D. Williams, I. S. Heng, J. Gair, J. A. Clark, and B. Khamesra, Phys. Rev. D **101**, 063011 (2020).
[163] R. Courant, K. Friedrichs, and H. Lewy, Math. Ann. **100**, 32 (1928).
[164] H. Pfeiffer (private communication).
[165] F. Ohme, Gravitational waves. Numerical relativity-data analysis. Proceedings of the 9th Edoardo Amaldi Conference, Amaldi 9, and meeting, NRDA 2011, Classic. Quant. Grav. **29**, 124002 (2012).
[166] F. Pretorius, Phys. Rev. Lett. **95**, 121101 (2005).
[167] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, Phys. Rev. Lett. **96**, 111102 (2006).
[168] L. E. Kidder *et al.*, J. Comput. Phys. **335**, 84 (2017).
[169] The SXS Collaboration, SpECTRE, https://spectre-code.org (2019).
[170] Z. B. Etienne and I. Ruchlin, Blackholes@home, https://blackholesathome.net (2019).
[171] M. C. Babiuc, B. Szilagyi, J. Winicour, and Y. Zlochower, Phys. Rev. D **84**, 044057 (2011).
[172] N. W. Taylor, M. Boyle, C. Reisswig, M. A. Scheel, T. Chu, L. E. Kidder, and B. Szilágyi, Phys. Rev. D **88**, 124010 (2013).
[173] C. J. Handmer, B. Szilágyi, and J. Winicour, Classic. Quant. Grav. **32**, 235018 (2015).
[174] C. J. Handmer, B. Szilágyi, and J. Winicour, Classic. Quant. Grav. **33**, 225007 (2016).
[175] I. Hinder, S. Ossokine, H. P. Pfeiffer, and A. Buonanno, Phys. Rev. D **99**, 061501 (2019).
[176] J. Healy, T. Bode, R. Haas, E. Pazos, P. Laguna, D. M. Shoemaker, and N. Yunes, Classic. Quant. Grav. **29**, 232002 (2012).
[177] E. Berti, V. Cardoso, L. Gualtieri, M. Horbatsch, and U. Sperhake, Phys. Rev. D **87**, 124020 (2013).
[178] E. Berti *et al.*, Classic. Quant. Grav. **32**, 243001 (2015).
[179] N. Yunes, K. Yagi, and F. Pretorius, Phys. Rev. D **94**, 084002 (2016).
[180] H. Witek, L. Gualtieri, P. Pani, and T. P. Sotiriou, Phys. Rev. D **99**, 064035 (2019).
[181] M. Okounkova, L. C. Stein, J. Moxon, M. A. Scheel, and S. A. Teukolsky, arXiv:1911.02588.
[182] Gwic 3g science case, https://gwic.ligo.org/3Gsubcomm/.
[183] L. Blanchet, Living Rev. Relativ. **9**, 4 (2006).
[184] R. Fujita, Prog. Theor. Phys. **128**, 971 (2012).
[185] K. Alvi, Phys. Rev. D **64**, 104020 (2001).
[186] C. M. Will and A. G. Wiseman, Phys. Rev. D **54**, 4813 (1996).
[187] K. G. Arun, A. Buonanno, G. Faye, and E. Ochsner, Phys. Rev. D **79**, 104023 (2009) [Erratum: **84**, 049901 (2011)].
[188] S. Marsat, A. Bohé, L. Blanchet, and A. Buonanno, Classic. Quant. Grav. **31**, 025023 (2014).
[189] E. Racine, Phys. Rev. D **78**, 044021 (2008).
[190] L. Blanchet, G. Faye, B. R. Iyer, and S. Sinha, Classic. Quant. Grav. **25**, 165003 (2008) [Erratum: **29**, 239501 (2012)].
[191] G. Faye, S. Marsat, L. Blanchet, and B. R. Iyer, Classic. Quant. Grav. **29**, 175004 (2012).
[192] G. Faye, L. Blanchet, and B. R. Iyer, Classic. Quant. Grav. **32**, 045016 (2015).
[193] A. Buonanno, G. Faye, and T. Hinderer, Phys. Rev. D **87**, 044009 (2013).
[194] P. Schmidt, I. W. Harry, and H. P. Pfeiffer, arXiv:1703.01076.
[195] A. Ashtekar and B. Krishnan, Living Rev. Relativ. **7**, 10 (2004).
[196] K. Kyrian and O. Semerak, Mon. Not. R. Astron. Soc. **382**, 1922 (2007).
[197] L. Santamaria *et al.*, Phys. Rev. D **82**, 064016 (2010).