The Astrophysical Journal, 945:103 (10pp), 2023 March 10 https://doi.org/10.3847/1538-4357/acbc18

Accumulating Errors in Tests of General Relativity with Gravitational Waves: Overlapping Signals and Inaccurate Waveforms

Qian Hu and John Veitch
Institute for Gravitational Research, School of Physics and Astronomy, University of Glasgow, Glasgow, G12 8QQ, UK; q.hu.2@research.gla.ac.uk, John.Veitch@glasgow.ac.uk
Received 2022 October 13; revised 2023 February 13; accepted 2023 February 13; published 2023 March 13

Abstract

Observations of gravitational waves (GWs) from compact binary coalescences provide powerful tests of general relativity (GR), but systematic errors in data analysis could lead to incorrect scientific conclusions. This issue is especially serious in the third-generation GW detectors in which the signal-to-noise ratio (S/N) is high and the number of detections is large. In this work, we investigate the impacts of overlapping signals and inaccurate waveform models on tests of GR. We simulate mock catalogs for Einstein Telescope and Cosmic Explorer and perform parametric tests of GR using waveform models with different levels of inaccuracy. We find that the systematic error in non-GR parameter estimates could accumulate toward a false deviation from GR when combining results from multiple events, although a Bayesian model selection analysis may not favor a deviation. Waveform inaccuracies contribute most to the systematic errors, but multiple overlapping signals could magnify the effects of systematics owing to the incorrect removal of signals. We also point out that testing GR using selected "golden binaries" with high S/N is even more vulnerable to false deviations from GR. The problem of error accumulation is universal; we emphasize that it must be addressed to fully exploit the data from third-generation GW detectors and that further investigations, particularly in waveform accuracy, will be essential.

Unified Astronomy Thesaurus concepts: Gravitational waves (678); General relativity (641)

1. Introduction

The observation of gravitational waves (GWs) from compact binary coalescences (CBCs) provides an ideal means of testing of general relativity (GR) in the strong-field regime (Abbott et al. 2016, 2017a, 2017b, 2019a, 2019b, 2021a, 2021b). The latest GW event catalogs contain nearly 100 CBC events (Abbott et al. 2021c, 2021d) based on which various tests of GR have been performed (Abbott et al. 2021a, 2021b). No concrete evidence of a deviation from GR has been found yet, but unprecedented constraints have been placed on possible violations of the theory. In the coming decades, the third-generation (3G) ground-based GW detectors (i.e., the Einstein Telescope; Punturo et al. 2010) and Cosmic Explorer (Reitze et al. 2019) are expected to detect $\mathcal{O}(10^5)$ CBC events per year, with signal-to-noise ratio (S/N) up to thousands (Oguri 2018; Maggiore et al. 2020; Himemoto et al. 2021; Relton & Raymond 2021; Samajdar et al. 2021). Since the statistical uncertainty of parameter estimates shrinks when the S/N increases and when a catalog of events is combined, observations from 3G GW detectors are expected to be able to obtain much tighter constraints on gravity theories.

However, this inspiring prospect of an enlarged detection catalog and higher S/Ns brings with it many difficulties in data analysis. For the purpose of testing GR (and any other theories), one needs to ensure that the systematic errors are small, so that the analysis will not favor the wrong theory and cause a false alarm (or false dismissal). Parameterized tests of GR (Meidam et al. 2018) suffer from the same problems as parameter estimation (PE) in general, which has been investigated in many works (e.g., Cutler & Vallisneri 2007; Antonelli et al. 2021). For instance, inaccurate waveform models may have already caused some tensions in current GW observations (Williamson et al. 2017; Hu & Veitch 2022) and are expected to be more important in future high-S/N detections (Cutler & Vallisneri 2007; Pürrer & Haster 2020; Gamba et al. 2021). Additionally, the 3G detectors with their improved low-frequency sensitivity are able to observe multiple signals at the same time. Detected overlapping signals cannot be perfectly removed from the data and could have nonnegligible impact on PE when the merger times of overlapping signals are close (Antonelli et al. 2021; Himemoto et al. 2021; Relton & Raymond 2021; Samajdar et al. 2021; Janquart et al. 2022; Pizzati et al. 2022). The undetected overlapping signals, i.e., the signals that are too faint to be detected, may also contribute to the systematic error (Antonelli et al. 2021; Reali et al. 2022). These errors are inevitable in 3G detectors, and repeated biased estimations for each event might end up with a wrong conclusion in the catalog-level analysis (Moore et al. 2021; Kunert et al. 2022).

Aforementioned works mainly focus on case studies for single events, or include only one type of systematic error. In this work, we aim to perform a more comprehensive investigation on systematic errors at the catalog level, including interactions between different types of systematics. We perform parameterized post-Newtonian (PPN) coefficient tests (Mishra et al. 2010; Cornish et al. 2011; Li et al. 2012) with our simulated event catalogs and inaccurate waveforms. Our simulations show that systematic errors can accumulate and could lead to an incorrect measurement of deviation from GR when results from multiple events are combined. We find that overlapping signals could magnify the effects of waveform systematics because of their imperfect subtraction from the data. Even worse, we find that the selected high-S/N events without known overlapping signals (so-called "golden events,"

which have been examined for GR tests in, e.g., Ghosh et al. 2016) may be more vulnerable to biased conclusions.

This paper is organized as follows. We introduce our methodology in Section 2, including the Fisher matrix formalism for error prediction in Section 2.1; configurations of PE, waveforms, and PPN tests in Section 2.2; catalog simulation and overlapping signals in Section 2.3; and methods of combining results in Section 2.4. Results are given in Section 3. We first demonstrate selected example events in Section 3.1, and then we move on to the catalog-level tests in Sections 3.2 and 3.3. Conclusions and discussions are given in Section 4.

2. Systematic Biases in PPN Tests

2.1. Estimating Systematic Errors

The generic formalism we use for estimating systematic errors in PE was first proposed in Cutler & Vallisneri (2007) and then generalized and validated against full PE by Antonelli et al. (2021). Let $\boldsymbol{\theta}$ be the parameters of a GW signal $h(\boldsymbol{\theta})$ (which may include more than one source, in the case of overlapping signals). The frequency domain data from a GW detector, denoted $d(\boldsymbol{\theta})$, are given by

$$d(\boldsymbol{\theta}) = h(\boldsymbol{\theta}) + n, \tag{1}$$

where n is noise. Under the assumption that this noise is stationary and Gaussian, the likelihood for GW PE is

$$L(\boldsymbol{\theta}) \propto e^{-\frac{1}{2}(d-h|d-h)} = e^{-\frac{1}{2}(n|n)}, \tag{2}$$

where $(...|...)$ is the inner product (Finn 1992), defined as

$$(a|b) = 4\Re \int_0^\infty \frac{a^*(f)b(f)}{S_n(f)} df, \tag{3}$$

where * means complex conjugate and $\Re$ denotes the real part. $S_n(f)$ is the noise power spectral density (PSD) of the detector. The optimal S/N is $\rho = \sqrt{(h|h)}$. For more than one data stream, the inner product's definition should be replaced by the sum of inner products calculated individually by each data stream.

Consider a maximum likelihood estimator (which is equivalent to Bayesian estimation with flat priors); the maximum point $\boldsymbol{\theta}_{\rm ML}$ satisfies

$$\partial_i \ln L|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\rm ML}} = (\partial_i h|d-h)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\rm ML}} = 0, \tag{4}$$

where ∂_i denotes the derivative with respect to the ith parameter. The data d are known, but real parameter $\boldsymbol{\theta}_{\rm real}$ and the GW signal in the detector $h(\boldsymbol{\theta}_{\rm real})$ are unknown. In practice, they are replaced by a waveform model $h_{\rm m}(\boldsymbol{\theta}_{\rm ML})$. By doing this, errors are introduced to $d-h$:

$$d - h = n + \delta H + \Delta\theta^j \partial_j h_{\rm m}. \tag{5}$$

The first term n is what $d-h$ is supposed to be: the noise in the detector. The second term $\delta H = h(\boldsymbol{\theta}_{\rm real}) - h_{\rm m}(\boldsymbol{\theta}_{\rm real})$ is the excess strain, which represents the difference between real signal(s) in the data and the model used to subtract signals. Inaccurate waveforms and overlapping signals can both contribute to this term. The third term comes from the imperfect measurement of signal parameters due to statistical noise and is given by the linear expansion of $h_{\rm m}(\boldsymbol{\theta}_{\rm real}) - h_{\rm m}(\boldsymbol{\theta}_{\rm ML})$, where $\Delta\theta^j$ is the statistical error of the jth parameter from the maximum likelihood estimator, and we adopt Einstein notation to indicate the sum over parameters. Substituting Equation (5) into Equation (4) and approximating all derivatives at $\boldsymbol{\theta}_{\rm ML}$, we get

$$\Delta\theta^i \approx (\boldsymbol{\Gamma}^{-1})^{ij}(\partial_j h_m|n + \delta H) = \Delta\theta^i_{\rm stat} + \Delta\theta^i_{\rm sys}, \tag{6}$$

where $\boldsymbol{\Gamma}_{ij} = (\partial_i h_m|\partial_j h_m)$ is the Fisher matrix (Cutler & Flanagan 1994). $\Delta\theta^i_{\rm stat} = (\boldsymbol{\Gamma}^{-1})^{ij}(\partial_j h|n)$ is the error induced by the detector noise. $\langle\Delta\theta^i_{\rm stat}\rangle = 0$, so the maximum likelihood estimator is unbiased if $\delta H = 0$, and $\langle\Delta\theta^i_{\rm stat}\Delta\theta^j_{\rm stat}\rangle = (\boldsymbol{\Gamma}^{-1})^{ij}$, which is consistent with the Fisher matrix formalism. The $\Delta\theta^i_{\rm sys} = (\boldsymbol{\Gamma}^{-1})^{ij}(\partial_j h_{\rm m}|\delta H)$ is the systematic error. Any effect that contributes to δH could be a source of systematic bias in PE. We will use $\sqrt{(\boldsymbol{\Gamma}^{-1})^{ii}}$ as statistical uncertainty and $\Delta\theta^i_{\rm sys}$ as the predicted systematic error.

2.2. PPN Formalism, Choices of Parameters, and Waveforms

The test of PPN coefficients is a generic formalism for finding deviations from GR, initially proposed by Mishra et al. (2010) and further developed for application with Bayesian inference (Li et al. 2012), and later applied to catalogs of real GW observations, most recently in Abbott et al. (2021b). We use the waveform model `IMRPhenomPv2` (Husa et al. 2016; Khan et al. 2016), whose phase is characterized by a set of parameters $\{p_i\}$, including inspiral phase parameters $\{\phi_0,...,\phi_7\}$ and $\{\phi_{5l}, \phi_{6l}\}$, phenomenological coefficients $\{\beta_0,...,\beta_3\}$, and merger-ringdown parameters $\{\alpha_0,...,\alpha_5\}$. Deviations $p_i \to (1 + \delta\hat{p}_i)p_i$ are introduced as the violations of GR; $\delta\hat{p}_i = 0$ reproduces GR. In this framework, testing GR is reduced to estimating the testing parameters $\delta\hat{p}_i$. Although a specific modified gravity theory could bring deviations in more than one testing parameter, previous works have shown that including one testing parameter at once is enough to detection violations. In fact, it can be more efficient to find violations from GR this way because it avoids the correlations between testing parameters and GR parameters (Sampson et al. 2013; Meidam et al. 2018). In this work, we choose $\delta\hat{\phi}_0$ as the example testing parameter. We assume that GR is the correct theory and focus on whether the PPN test falsely indicates deviations of GR.

We restrict our Fisher matrix analysis to a subset of the full signal parameters, to avoid computational issues. Parameterized deviations of the type we consider have a direct effect on the phasing of the signal, so in addition to $\delta\hat{\phi}$ we must include the other parameters that do the same: chirp mass $\mathcal{M}$ and mass ratio q, as well as the time of coalescence t_c. The full six-dimensional space of spin configurations is known to bring ill-conditioned Fisher matrices (Borhanian & Sathyaprakash 2022) owing to correlations between parameters, and because of the prior bounds on angular parameters, results can be misleading even when they can be computed. We therefore use only the effective spin $\chi_{\rm eff}$ to capture the dominant effect of (aligned) spin on the waveforms. We include this by forcing the two aligned spin components to contribute equally to $\chi_{\rm eff}$, which allows us to treat it as a single parameter. We neglect to include extrinsic parameters in the Fisher matrix, effectively assuming that they are measured precisely. Since these do not have a frequency-dependent effect on the phase, we do not expect them to be highly correlated with the intrinsic parameters. Our choice captures the parameters that appear in the leading PN term and the corresponding PPN

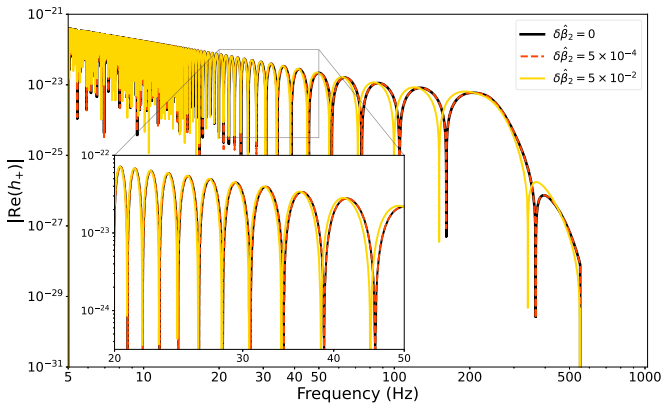

Figure 1. The absolute value of the real part of plus polarization from a nonspinning BBH with $\mathcal{M}_c = 30.69\,M_\odot$, $q = 0.88$, in frequency domain. Waveforms with different $\delta\hat{\beta}_2$ are shown in different colors and line styles. The intermediate region of this system starts around 50 Hz, which is consistent with where waveform difference appears in the plot.

modifications, as well as the decisive parameter in the analysis of overlapping signals, t_c. Other parameters are randomly generated (details in Section 2.3) but are treated as perfectly known. Setting parameters to their injection values excludes their contributions to both statistical and systematic errors in PE. For instance, if we removed the effective spin from our calculation, we would obtain tighter statistical and systematic errors because its correlation with mass parameters and the testing parameter is removed (Berti et al. 2005). Considering realistic PE in the future in which all parameters are included, correlation between parameters may make posteriors wider and systematic bias larger. However, due to the linear expression in Equation (6), we expect that the two changes are proportional and our conclusion will not change significantly under this simplification.

We induce a nonzero $\delta\hat{\beta}_2$ to mimic inaccurate waveform models based on the following considerations. To reduce potential correlations with $\delta\hat{\phi}_0$, we exclude testing parameters for the inspiral stage. Correlation between the testing parameter and the waveform systematic parameter may undermine the generality of the illustration. To make sure the testing parameter has enough influence on the waveform, we do not choose parameters for the merger-ringdown stage, which only includes the last few cycles. Therefore, we look for parameters in the intermediate region, which is described by $\delta\hat{\beta}_i$ (Khan et al. 2016). $\delta\hat{\beta}_0$ and $\delta\hat{\beta}_1$ bring global phase shift and time shift in this region, respectively, so $\delta\hat{\beta}_2$ is the dominant testing parameter that encodes physical (frequency-dependent) modifications.

We assume that $\delta\hat{\beta}_2 = 0$ is our model waveform, while the "real" waveform could have $\delta\hat{\beta}_2 = 0$, 5×10^{-2}, or 5×10^{-4}. The first case means that our model waveform is perfect, and all systematic errors will come from overlapping signals. The second case generates waveform mismatches around 10^{-4} to 10^{-3}, which corresponds to the current waveform accuracy (Ossokine et al. 2020; Pratten et al. 2021). The last case produces mismatches around 10^{-7} to 10^{-6} and corresponds to the expectations for future waveform accuracy (Pürrer & Haster 2020; Hu & Veitch 2022). A comparison of the three types of waveforms is shown in Figure 1. We show an example of a nonspinning binary black hole (BBH) merger with $\mathcal{M}_c = 30.69\,M_\odot$ (in the detector frame) and $q = 0.88$ whose intermediate region starts around 50 Hz. The mismatches are 3×10^{-7} and 2×10^{-3} between $\delta\hat{\beta}_2 = 0$ and $\delta\hat{\beta}_2 = 5\times10^{-4}$, 5×10^{-2}, respectively.

The excess strain from inaccurate waveforms can be written as

$$\delta H_{\rm wf} = h(\boldsymbol{\theta}_{\rm real})|_{\delta\hat{\beta}_2\neq0} - h(\boldsymbol{\theta}_{\rm real})|_{\delta\hat{\beta}_2=0}. \quad (7)$$

We can use the approximation $h(\boldsymbol{\theta}_{\rm real}) - h_{\rm m}(\boldsymbol{\theta}_{\rm real}) \approx h(\boldsymbol{\theta}_{\rm ML}) - h_{\rm m}(\boldsymbol{\theta}_{\rm ML})$, as the error would be a higher-order term.

2.3. Overlapping Signals and Mock Catalogs

When multiple signals come into data, they may have impacts on the analysis of each other (Antonelli et al. 2021; Himemoto et al. 2021; Relton & Raymond 2021; Samajdar et al. 2021). It is known that the correlation between signals is not strong unless the merger times are very close (typically <1 s); in this work we regard two signals as "overlapping" only if the merger time difference $|\Delta t| < 4$ s, which captures the most influential neighbors of a signal.

Overlapping signals can be classified into two types: detected signals and undetected signals (confusion signals). The former is strong enough to be detected and should be subtracted from data in the analysis for other signals (or the "main" signal).[1] The latter, however, is too faint to be recognized by the detection pipeline and may have an unnoticed impact on PE. In this work, the network S/N threshold for detection is set to 8, under which GWs are assumed to be undetected.

If a signal is detected, it will still contribute excess strain since we cannot perfectly remove it from the data. The excess strain after imperfect removal is

$$\delta H_{\rm DO} = h'(\boldsymbol{\theta}_{\rm real}) - h_m{}'(\boldsymbol{\theta}_{\rm ML}) \approx \Delta\theta'^i\partial_i h'_m + \delta H'_{\rm wf}, \quad (8)$$

where $'$ denotes variables of the detected overlapping signal. The first term arises from the inaccurate estimation of parameters for the overlapping signal, which is random since the error is partly caused by the random noise, although other factors, such as waveform inaccuracies and overlapping signals, also contribute to it. As a conservative estimation and following Antonelli et al. (2021), we ignore waveform systematic errors in $\Delta\theta'^i$ (i.e., assuming that $\Delta\theta'^i$ is merely caused by noise, which tends to underestimate it) and adopt the lowest-order approximation for its correlation with the main signal. Substituting it into Equation (6), one obtains the covariance of the first term in the systematic error Equation (8),

$$\langle\Delta\theta^i_{\rm DO1}\Delta\theta^j_{\rm DO1}\rangle = (\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}^{-1}_{\rm mix}\boldsymbol{\Gamma}'^{-1}(\boldsymbol{\Gamma}^{-1}_{\rm mix})^{\rm T}(\boldsymbol{\Gamma}^{-1})^{\rm T})_{ij}, \quad (9)$$

where $(\boldsymbol{\Gamma}_{\rm mix})_{ij} = (\partial_i h|\partial_j h')$ encodes the correlation between two signals and $\boldsymbol{\Gamma}' = (\partial_i h'|\partial_j h')$ is the Fisher matrix of the overlapping signal. The second term in Equation (8) represents the inaccurate waveform model we use to subtract signals and can be calculated the same way as the waveform systematic, yielding $\Delta\theta^i_{\rm DO2} = (\boldsymbol{\Gamma}^{-1})^{ij}(\partial_j h_m|\delta H'_{\rm wf})$. In this work, the systematic error from detected overlapping signals is calculated as $\Delta\theta^i_{\rm DO2}$ plus a random sample drawn from a multivariate Gaussian distribution with covariance matrix Equation (9) and zero mean. For more than one detected overlapping signal, Equation (8) can be extended by defining h' as the summation

[1] It is also possible to do a joint PE for all existing signals; see Janquart et al. (2022).

Table 1
A Summary of Three Mock Catalogs

	No. of Observable Binaries		Detected Overlaps on BBH Events		Undetected Overlaps on BBH Events	
	BBH	BNS	No. of Overlaps	No. (Fraction) of Events	No. of Overlaps	No. (Fraction) of Events
Low	56,526	286,088	0	53,118 (95%)	0	54,067 (96%)
			1	2847 (5.1%)	1	1936 (3.5%)
			2	74 (0.13%)	2	37 (0.066%)
			3	2 (0.0040%)	3	1 (0.0018%)
Median	88,300	1,144,354	0	73,200 (84%)	0	76,270 (87%)
			1	13,125 (15%)	1	10,461 (12%)
			2	1093 (1.2%)	2	721 (0.82%)
			3	67 (0.077%)	3	35 (0.040%)
			4	2 (0.0023%)		
High	143,349	2,896,647	0	92,692 (65%)	0	100,862 (71%)
			1	39,450 (28%)	1	34,519 (24%)
			2	8559 (6.0%)	2	5940 (4.2%)
			3	1208 (0.85%)	3	673 (0.47%)
			4	131 (0.092%)	4	58 (0.041%)
			5	20 (0.014%)	5	7 (0.0049%)
					6	1 (0.00070%)

Note. From left to right, the table shows catalog type, observable BBH and BNS per year (note that this is not detectable), and distributions of numbers of overlapping signals among BBH events. For example, in the median merger rate catalog, there are 13,125 detected BBH events (15% of all detected BBH events) coming with one detected overlapping GW signal and 10,461 detected BBHs coming with one undetected overlapping GW signal. The overlapping signal can be BBH or BNS, and two signals are defined as overlapped if their merger time difference $\Delta t < 4$ s.

of all GWs in the data (Antonelli et al. 2021), which enlarges the dimension of $\mathbf{\Gamma}_{\rm mix}$ and $\mathbf{\Gamma}'$.

The undetected overlapping signal simply contributes to systematic error by $\delta H_{\rm UO} = \sum_{\rm undetected} h''(\boldsymbol{\theta}_{\rm real})$. It is accessible in our simulation but unknown in real data analysis.

We consider BBH and binary neutron star (BNS) sources and assume that their distribution in redshift z follows the analytical approximation (Oguri 2018)

$$R_{\rm GW}(z) = \frac{a_1 e^{a_2 z}}{e^{a_3 z} + a_4}\ {\rm Gpc}^{-3}\ {\rm yr}^{-1}, \tag{10}$$

which is then converted to observable event rate by multiplying by a factor $\frac{1}{1+z}\frac{dV_c}{dz}$. Here V_c is the comoving volume and we employ Planck15 cosmology (Ade et al. 2016). Note that "observable" GWs need to achieve a network S/N of 8 to be "detectable." $a_{\{1,2,3,4\}}$ are model parameters. We set $a_2 = 1.6$, $a_3 = 2.1$, $a_4 = 30$ to mimic a peak at $z \sim 2$. a_1 is scaled based on local merger rate given by Abbott et al. (2021e) ($\mathcal{R}_{\rm BNS} = 320^{+490}_{-240}$ and $\mathcal{R}_{\rm BBH} = 23.9^{+14.3}_{-8.6}\ {\rm Gpc}^{-3}\ {\rm yr}^{-1}$) such that $R_{\rm GW}(z=0) = \mathcal{R}_{\rm BNS/BBH}$. We choose three values for a_1, corresponding to lower, median, and higher estimation of local merger rate, respectively.

The masses of BBHs are generated by the PowerLaw + Peak model in Abbott et al. (2021e), while all BNS systems are set to be same: $1.45 + 1.4\ M_\odot$, $\Lambda_1 = \Lambda_2 = 425$. The effective spin follows the Gaussian distribution in Abbott et al. (2021e), with mean of 0.06 and standard deviation of 0.12. `IMRPhenomPv2_NRTidal` (Dietrich et al. 2019) is used to generate BNS waveforms with the same $\delta\hat{\beta}_2$ as BBH. We will perform tests of GR with all BBH events and use BNS events as a background: BNS events are only involved in the calculation as overlapping signals. We assume isotropically distributed inclination and source sky direction and uniformly distributed coalescence time, phase, and polarization angle.

A summary of low, median, and high merger rate catalogs is shown in Table 1. It shows that most BBH events will not have an overlapping signal near their merger time, which implies that overlapping signals contribute to systematic errors less frequently than waveform systematics. With our ET+CE configuration, the numbers of the two kinds of overlaps are close. However, if the number of detectors is less than assumed, or detector sensitivities are lower than designed, some of the detected overlaps would become undetected, and vice versa. The unnoticeable confusion background has drawn attention in recent works (Reali et al. 2022; Wu & Nitz 2022) and needs further investigation. Compact binaries formed by Population III stars (which we have ignored) could also contribute to the confusion background. However, according to the model in Oguri (2018), the numbers of observable Population III binaries of B17 and K16 models per year are roughly 40,000 and 180,000, respectively, which is much lower than the BNS background.

Several simplifications have been adopted in our mock catalog: we regard BNS as a background and use only BBH as

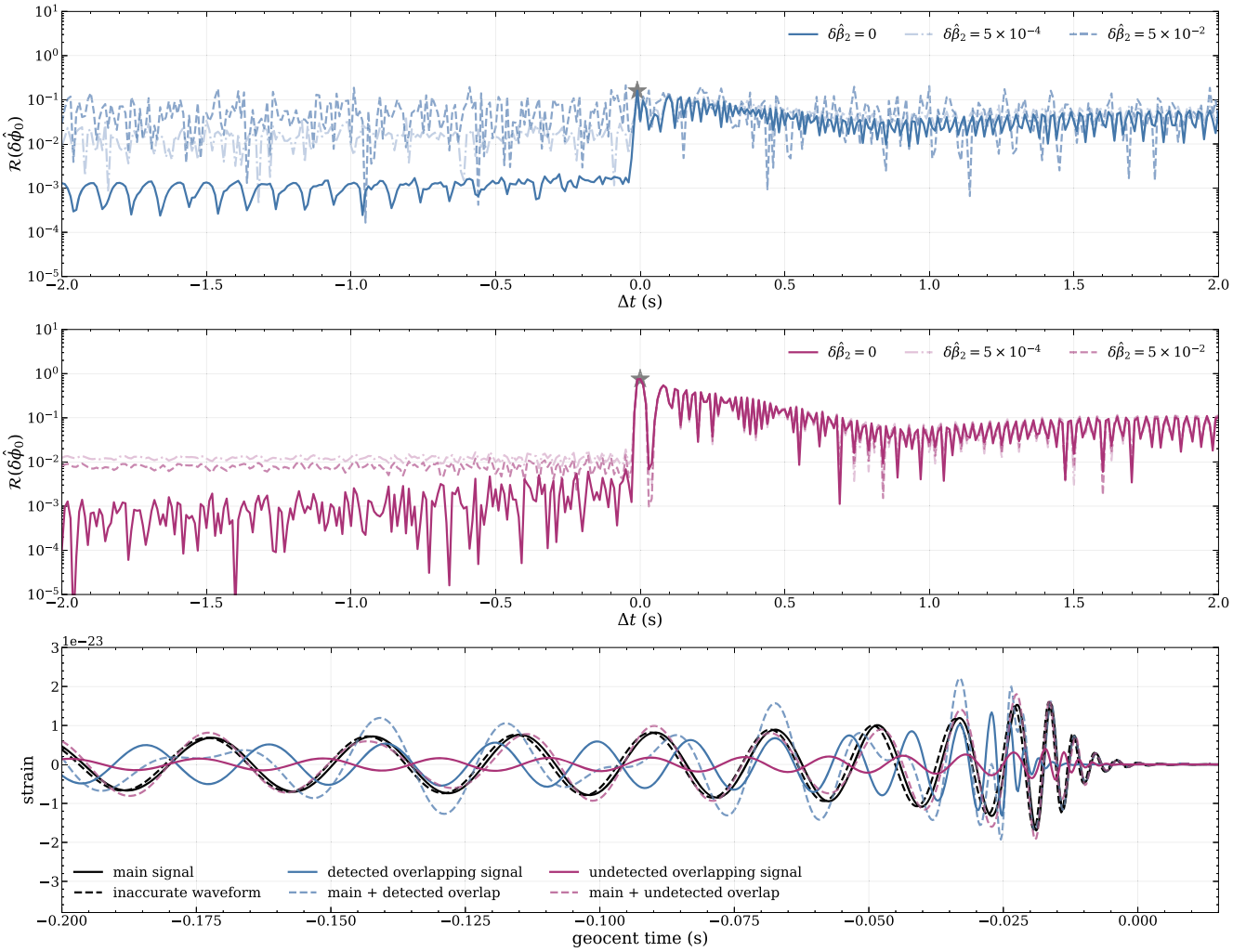

Figure 2. Top and middle rows: the error ratio of $\delta\hat{\phi}_0$ varies with merger time difference. The main signal has (detector frame) $\mathcal{M}_c = 32\,M_\odot$, $q = 0.9$, $\chi_{\rm eff} = 0.2$, and S/N of 27. The overlapping signal is an equal mass BBH with $\mathcal{M}_c = 20\,M_\odot$ and $\chi_{\rm eff} = 0.1$. The S/N of the overlapping signal is adjusted by changing its luminosity distance: the detected overlap is shown in the top panel, and the undetected one in the middle panel. We use three kinds of waveforms explained in Section 2.2: perfect waveform (solid line), "current" waveform (dashed line), and "future waveform" (faint dotted–dashed line). Bottom row: waveforms of the main and overlapping signals and their superposition. Merger times of overlapping signals are chosen to maximize their influences, as marked by gray stars in the first two rows. Inaccurate waveform in the $\delta\hat{\beta}_2 = 5\times10^{-2}$ case is also plotted for comparison.

the test source; we ignore neutron star–black hole mergers and other possible types of sources; we use an analytical merger rate that peaks at $z\sim2$, ignoring compact binaries from Population III stars. Our catalogs aim to generate an appropriate merger rate for the study of systematic error accumulation, rather than accurately modeling the astrophysical population. To achieve this, we also adjust the merger rate to different levels, expecting that the real situation will lie somewhere between our lowest and highest estimates.

Signals are injected into the 3G GW detector Einstein Telescope with ET-D PSD (Punturo et al. 2010) located at the Cascina site of the current Virgo detector and Cosmic Explorer located at the LIGO Hanford site with the sensitivity curve proposed by Abbott et al. (2017c). The frequency band used for the analysis is 5–2048 Hz.

2.4. Combining Results

There are several ways of combining results from multiple events (Isi et al. 2019; Zimmerman et al. 2019). We employ two straightforward methods: multiplying likelihoods (equivalently, multiplying posteriors if priors are flat) and multiplying Bayes factors. The former assumes that the modification parameter is the same for all events, while the latter allows the modification parameter to vary across events.

We assume a flat prior distribution and that the posterior follows a multivariate Gaussian distribution with covariance matrix $\mathbf{\Gamma}^{-1}$ and mean $\boldsymbol{\mu}$ equal to injection values $\boldsymbol{\theta}_{\rm inj}$ plus systematic errors $\underline{\boldsymbol{\Delta\theta}_{\rm sys.}}$ The statistical uncertainty of a parameter is $\sigma_i = \sqrt{(\mathbf{\Gamma}^{-1})_{ii}}$. We define the error ratio between systematic and statistical errors as

$$\mathcal{R}(\theta_i) = |\Delta\theta_{i,\rm sys}/\sigma_i|. \tag{11}$$

We consider that the PPN test coefficient is subject to false deviations from GR when $\mathcal{R}(\delta\hat{\phi}_0) > 1$.

In order to combine results from multiple events, one would multiply the posterior distributions of the testing parameter for each. Multiplication of Gaussian distributions results in another Gaussian distribution whose mean (systematic error) is a linear combination of the original means. From the first event in a catalog, we multiply the posterior of new events one by one

and calculate the error ratio. Considering the arbitrary sequence of events, we permute the sequence 200 times and extract the ensemble average and 68% confidence interval.

Treating GR as a submodel of the non-GR theory, the Bayes factor can be calculated analytically with the Gaussian posterior (Moore et al. 2021). Denoting systematic error of $\delta\hat{\phi}_0$ as $\Delta\theta_{\rm sys}$, we have

$$L_{\rm GR}(\boldsymbol{\theta}_{\rm GR}) = L_{\rm nonGR}(\boldsymbol{\theta}_{\rm nonGR})|_{\delta\hat{\phi}_0=\Delta\theta_{\rm sys}}. \tag{12}$$

The Bayes factor is then calculated as

$$\begin{aligned}\mathcal{B}^{\rm nonGR}_{\rm GR} \sim \frac{Z_{\rm nonGR}}{Z_{\rm GR}} &= \frac{\int d\boldsymbol{\theta}_{\rm nonGR} L_{\rm nonGR}}{\int d\boldsymbol{\theta}_{\rm GR} L_{\rm GR}} \\ &= \sqrt{2\pi}\, e^{\frac{1}{2}(\Gamma_{\delta\hat{\phi}_0\delta\hat{\phi}_0} - \boldsymbol{\nu}^{\rm T}(\boldsymbol{\Gamma}_{\rm GR})^{-1}\boldsymbol{\nu})\Delta\theta^2_{\rm sys}} \sqrt{\frac{\det\boldsymbol{\Gamma}_{\rm GR}}{\det\boldsymbol{\Gamma}_{\rm nonGR}}}, \end{aligned} \tag{13}$$

where $\boldsymbol{\Gamma}_{\rm nonGR}$ is the Fisher matrix including the testing parameter, while $\boldsymbol{\Gamma}_{\rm GR}$ only includes GR parameters. $\boldsymbol{\nu}_i = (\partial h/\partial\boldsymbol{\theta}_i|\partial h/\partial\delta\hat{\phi}_0)$ represents the correlation between GR and non-GR parameters. $\boldsymbol{\Gamma}_{\delta\hat{\phi}_0\delta\hat{\phi}_0} = (\partial h/\partial\delta\hat{\phi}_0|\partial h/\partial\delta\hat{\phi}_0)$. The exponential term in the Bayes factor accounts for the deviation of GR, while the determinant ratio term usually favors GR since modified theories introduce extra parameters to explain the data. We also note that the correlation term $\boldsymbol{\nu}^{\rm T}(\boldsymbol{\Gamma}_{\rm GR})^{-1}\boldsymbol{\nu}$ mitigates the deviation of GR. Ignoring this term may overestimate the Bayes factor (e.g., Moore et al. 2021). When combining events, Bayes factors are numerically multiplied, with the same permutation mentioned before. We consider a false deviation from GR to be achieved when $\ln\mathcal{B}^{\rm nonGR}_{\rm GR} > 8$. We reemphasize that Bayes factors are first computed for each event and then combined across the catalog, rather than calculated after different posteriors are multiplied. This analysis should be interpreted as not assuming that the testing parameter is the same for all events. In this sense it is less sensitive to violations of GR when there is a common underlying deviation parameter, so we would expect it to be less vulnerable to simulated false violations. While error ratio accumulation is decided by errors from each event, Bayes factor accumulation is more sensitive to the fraction of correct analyses in the catalog. The two methods of combining results are independent and do not necessarily lead to the same conclusion. More details are given in Section 3.2.

3. Results

3.1. Single Events

We first present an example event, investigating the effect of a detected or undetected overlapping signal. The main signal is from a BBH with $\mathcal{M}_c = 32\,M_\odot$ (in the detector frame), $q = 0.9$, $\chi_{\rm eff} = 0.2$, and network S/N of 27. The overlapping signal is an equal-mass BBH with $\mathcal{M}_c = 20\,M_\odot$ and $\chi_{\rm eff} = 0.1$. We scale its S/N from ∼26 down to $\lesssim$8 to make it detectable or undetectable. We vary the merger time difference (by 0.01 s per step) and calculate the total systematic error with different waveform models. Note that, throughout this section, the "systematic error" refers to that of the testing parameter $\delta\hat{\phi}_0$ and is denoted as $\Delta\theta_{\rm sys}$.

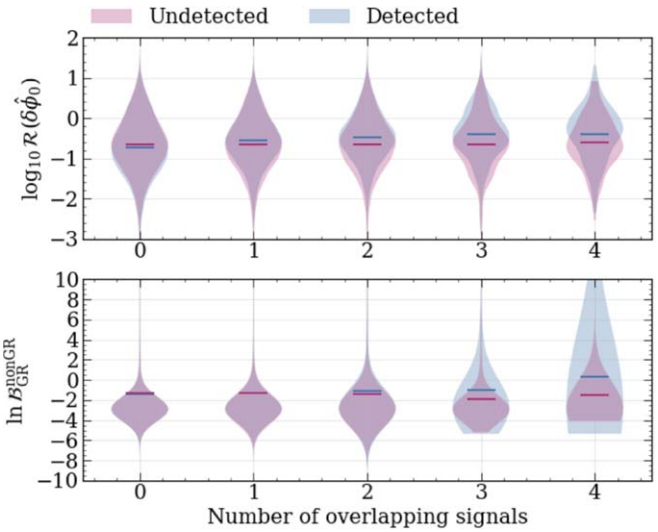

Figure 3. Distributions of systematic errors and Bayes factors in the high merger rate catalog with $\delta\hat{\beta}_2 = 5\times10^{-2}$ waveform, classified by number and type of overlapping signals. Bars denote the mean value. The number of overlapping signals is cut at 4 because of the insufficient number of events coming with >4 overlapping signals. The difference in the increase of mean values shows that detected overlapping signals could magnify the effects of inaccurate waveform models.

The error ratio for this example event is shown in Figure 2, including an illustration of the waveforms. The error from the overlapping signal oscillates when Δt changes owing to the repeating alignments and misalignments of phases of the two GWs. The overlap error is not symmetric around $\Delta t = 0$ because the two waveforms are not symmetric, but the peak is always located in the region $|\Delta t| \leqslant 1$ s, meaning that the overlapping signal only produces a large influence when two mergers are very close. Waveforms in the last row show how the main signal is modulated by overlapping signals. Around $\Delta t \sim 0$, the confusion signal has larger impacts than waveform systematics, so it dominates the systematic error. The detected signal changes the signal significantly, but it is then subtracted from data and therefore produces less residual strains. When $|\Delta t|$ is large, it is waveform inaccuracy that dominates the systematic error. These characteristics are consistent with previous works (Antonelli et al. 2021; Himemoto et al. 2021; Relton & Raymond 2021; Samajdar et al. 2021).

It is possible for undetected signals to produce significant systematic errors in our simulation. However, comparing the detected and undetected overlapping signal, the former produces larger systematic error when the waveform is not accurate because the waveform systematic is also involved in signal subtraction. One can see this from the more intense perturbations in the $\delta\hat{\beta}_2 \neq 0$ case in Figure 2 for detected overlaps (errors are directly added, so they may constructively or destructively interfere). This implies that different types of systematic errors are correlated and could be a magnifying factor for each other, as expected from Equation (8). A more direct comparison is given in Figure 3. We calculate systematic errors for each BBH event in our mock catalog and show systematic errors and Bayes factors caused by different numbers of overlapping signals for the high merger rate catalog. With the increase of the number of overlaps, detected overlaps tend to produce larger errors, while errors from confusion background signals make smaller incremental changes.

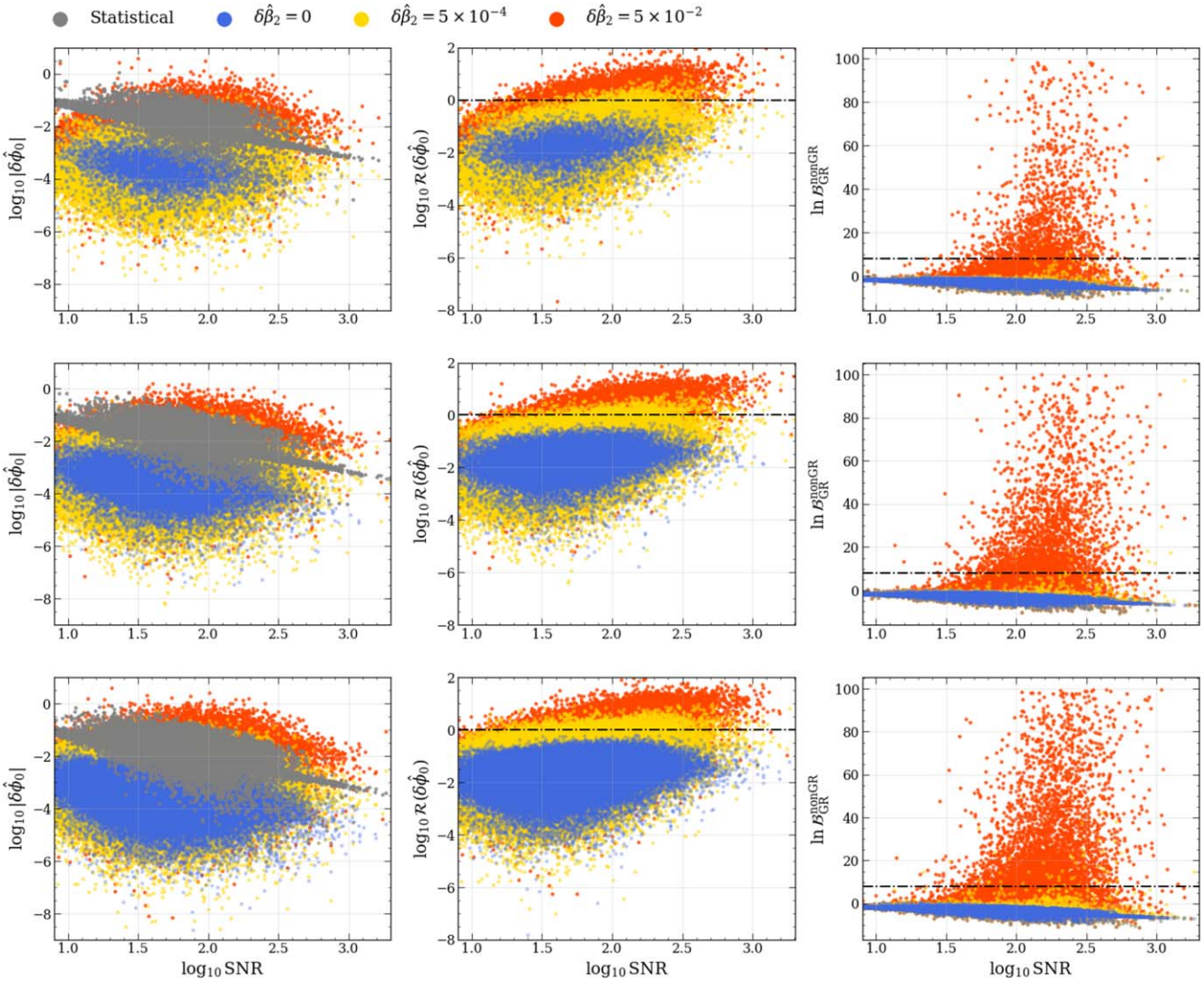

Figure 4. Relation between S/N (*x*-axes) and absolute error (first column), error ratio (second column) of $\delta\hat{\phi}_0$, and Bayes factor (third column) for low (top row), median (middle row), and high (bottom row) merger rate catalogs. Each point represents a BBH event. Blue points are for the "perfect waveform" case, where all systematic errors come from overlapping signals; red points stand for the "current waveform" case, and yellow points are for the "future waveform" case. Gray points in the first column are statistical errors. Dashed lines in the second and third columns are the threshold above which GR is mistakenly disfavored. This plot shows that $\mathcal{R}(\delta\hat{\phi}_0) > 1$ and $\ln\mathcal{B}^{\rm nonGR}_{\rm GR} > 8$ are mostly from "current waveform" and high-S/N events.

The statistical error $\Delta\theta^i_{\rm stat} = (\mathbf{\Gamma}^{-1})^{ij}(\partial_j h|n) \approx 1/|h| \approx 1/{\rm S/N}$, while the systematic error $\Delta\theta^i_{\rm stat} = (\mathbf{\Gamma}^{-1})^{ij}(\partial_j h|\delta H)$ does not necessarily shrink when the S/N increases, for example, waveform systematics of the main signal. Therefore, systematic errors may dominate in the high-S/N scenario. We plot the absolute error, error ratio, and Bayes factor with S/N in Figure 4. We find that the error ratio could exceed 1 for the "current" waveform, and this happens more often when ${\rm S/N} > 30$ despite the fact that high-S/N events are rarer. Error ratios for the "future" waveform simulations are usually below 1, but a certain number of exceptions exist. For the Bayes factor analysis we find a similar situation, although there are a smaller fraction of more extreme values. There are roughly 0.8% and 18% events producing $\mathcal{R}(\delta\hat{\phi}_0) > 1$ for $\delta\hat{\beta}_2 = 5\times10^{-4}$ and 5×10^{-2}, respectively, while for $\ln\mathcal{B}^{\rm nonGR}_{\rm GR} > 8$ the fractions are 0.02% and 3%. As pointed out by Moore et al. (2021), false deviations could be achieved even though estimations for individual events are generally accurate. We will investigate this in more detail in the next subsection.

3.2. Error Accumulation in a Catalog

As mentioned in Section 2.4, we combine all BBH events by multiplying likelihoods or Bayes factors. The results are shown in Figure 5. Let $N_{\rm event}$ be the number of events. When multiplying likelihoods, the statistical uncertainty shrinks as $1/\sqrt{N_{\rm event}}$. The absolute error of the testing parameter also decreases, but at a slower pace owing to the perturbations from newly accumulating systematic errors. It also follows $1/\sqrt{N_{\rm event}}$ if there were no systematic errors—we observe that the test with the perfect waveform in a low merger rate catalog is approximately doing so. In most simulations it is the waveform inaccuracy that keeps contributing to the systematic errors. The slower decay of systematic error results in a climbing error ratio as the number of events increases. At some point (typically $\sim10^3$ events, considering error bars) it leads to a false deviation of GR for the "current" waveform. For the better waveform, the

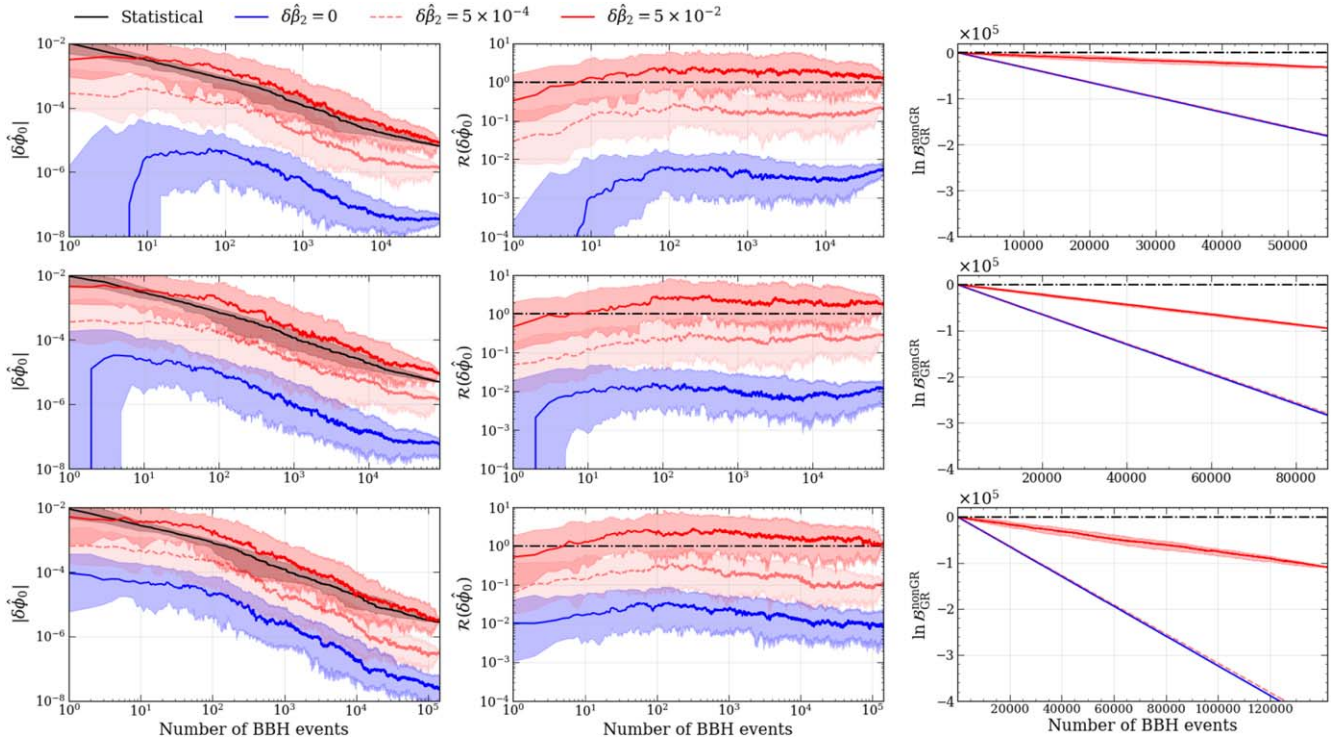

Figure 5. Systematic error accumulates with the increase of number of events. The first column shows the absolute error of $\delta\hat{\phi}_0$, and the second column shows the error ratio. The third column is the Bayes factor. Solid red lines are the ensemble average for the "current waveform," dashed red lines are for the "future waveform," and blue lines stand for the perfect waveform. The shading along lines is the 68% confidence interval. The first, second, and third rows are for low, median, and high catalogs, respectively. The black dotted–dashed line is the threshold above which a false deviation of GR is claimed. False deviations can be achieved with the increase in the number of events by multiplying posteriors, but multiplying Bayes factors does not give wrong conclusions in these correct-analyses-dominated catalogs.

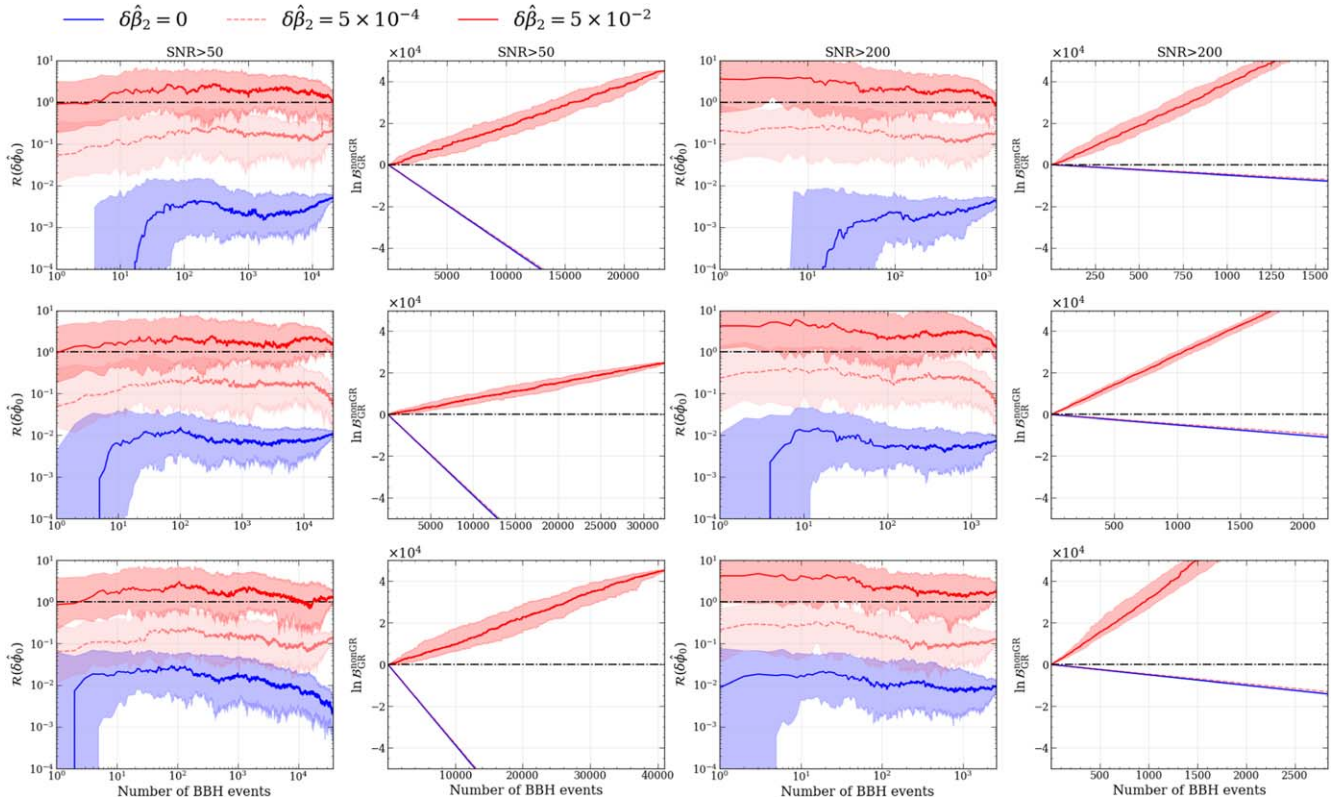

Figure 6. Similar to Figure 5, the error ratio and Bayes factor accumulation. The left two columns show results from $S/N > 50$ events; the right two columns are for $S/N > 200$ events. Compared with Figure 5, it shows that tests with high-S/N events are more likely to make a false deviation from GR.

error ratio climbs as well, but it remains below the statistical level until 10^5–10^6 events.

Multiplying Bayes factors is a direct addition of $\ln \mathcal{B}^{\rm nonGR}_{\rm GR}$. "Correct analyses" can effectively decrease the combined Bayes factor so that a correct-analyses-dominated catalog leads to correct conclusions. Since there are only 3% of events with $\ln \mathcal{B}^{\rm nonGR}_{\rm GR} > 8$ (furthermore, only 7% of events with $\ln \mathcal{B}^{\rm nonGR}_{\rm GR} > 0$) for the current waveform, the sum of all Bayes factors is negative; thus, false deviation is not achieved in this case. In contrast, multiplying likelihoods linearly adds systematic errors: for Gaussian distributions f and g, the mean of their product is $\mu_{fg} = \frac{\mu_f \sigma_g^2 + \mu_g \sigma_f^2}{\sigma_f^2 + \sigma_g^2}$. Correct analysis and different sign of errors could diminish systematic error a bit, but it is never guaranteed for the error to be held around 0. Moreover, statistical uncertainty also shrinks during event stacking, so the error ratio shows a clear increase.

3.3. Golden Events

We have combined all the detected BBH events in the above subsection. It is also interesting to test GR with only the "golden events," i.e., the GW events with high S/N and clean data that contribute to most of the information in the whole catalog test. This idea is widely used in many works, such as recent GWTC-3 tests of GR (Abbott et al. 2021b) and cosmology (Abbott et al. 2021f). Since the noise is Gaussian in our simulation, we select the golden events with only two criteria: S/N above a chosen threshold (50 or 200) and that there are no detected overlapping signals.

Results for the error ratio and Bayes factor are shown in Figure 6: high-S/N events are more vulnerable to systematic errors. Fewer events are needed to create a false deviation for the "current" waveform model, and the "future" waveform is closer to false deviation in all three catalogs. Moreover, the golden events catalog consists of more incorrect analyses ($\mathcal{R}(\delta\hat{\phi}_0) > 1$ or $\ln \mathcal{B}^{\rm nonGR}_{\rm GR} > 8$), and it causes the Bayes factor of current waveform to incorrectly favor the non-GR theory.

As mentioned in Section 3.1, statistical uncertainty decreases as 1/S/N while systematics do not as long as waveform is imperfect. The false deviation for golden events is not surprising from this angle, but it does need more attention and an appropriate solution for future data analysis.

4. Conclusions and Discussions

We have investigated how systematic errors in testing GR accumulate under the influence of overlapping signals and inaccurate waveforms. We have considered different levels of waveform inaccuracies and event rates and employed two approaches to combining the results.

We confirm that systematic errors could accumulate when combining multiple events and could lead to incorrectly disfavoring GR in some cases. Since overlapping signals do not always occur, it is waveform inaccuracies that keep contributing to the systematic error in the catalog tests. An accurate waveform model is effective at preventing false deviations in most cases, while a worse one could lead to biased conclusions. We additionally find that overlapping signals can enlarge the effect of waveform systematics. By increasing the number of overlaps, we tend to achieve a greater systematic error and a Bayes factor that leans more toward the non-GR model. One can avoid this correlated error by selecting events with no detected overlapping signals and, if one prefers, with high S/N as well. However, we have shown that these events produce biases much faster because waveform systematics dominate in the high-S/N scenario.

We should point out that GR is assumed to be the true theory to describe the data in this work, which is not necessarily correct. The inverse problem, namely, what happens to detection and PE when we use GR waveform for data analysis but GR is wrong (stealth bias), is investigated in previous works (Cornish et al. 2011; Vallisneri & Yunes 2013; Vitale & Del Pozzo 2014). The core idea of our work and that of stealth bias are the same: using an incorrect model in data analysis can lead to biased results. Stealth bias emphasizes the importance of assuming the correct theory, while our work points out that even if the assumed fundamental theory is correct, waveform modeling and overlapping signals are still able to corrupt the results.

We reemphasize that systematic errors can accumulate when combining multiple events and lead to incorrect scientific conclusions. This problem is universal: in addition to tests of GR, any analysis based on a GW catalog is faced with this issue, such as constraints on cosmological models, neutron star models (Kunert et al. 2022), and astrophysical population inference. Furthermore, there are more sources of systematic errors than those investigated in this work: instrumental calibration (Hall et al. 2019; Sun et al. 2020), glitches (Pankow et al. 2018; Powell 2018), missing physical effects (Pang et al. 2018; Saini et al. 2022), and so forth. A full analysis of these contributions and their relative importance will be essential in designing analysis strategies for 3G detectors. An obvious solution to these issues is continuing improvements to waveform model accuracy and instrumental stability, but we believe that more efforts are needed from the angle of data analysis. A proper estimate of confusion background may be necessary (Reali et al. 2022), and new techniques might be needed, such as accounting for waveform systematic errors during PE (Moore & Gair 2014), performing specific analysis of residual strain (Dideron et al. 2022), and so forth.

The authors would like to thank Chris Messenger and Christian Chapman-Bird for helpful discussions and suggestions. We are grateful for computational resources provided by Cardiff University and funded by STFC grant ST/I006285/1. Q.H. is supported by CSC. J.V. is supported by STFC grant ST/V005634/1.

ORCID iDs

Qian Hu https://orcid.org/0000-0002-3033-6491
John Veitch https://orcid.org/0000-0002-6508-0713

References

Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2016, PhRvL, 116, 221101
Abbott, B. P., Abbott, R, Abbott, T. D., et al. 2017a, PhRvL, 119, 141101
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017b, PhRvL, 118, 221101
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017c, CQGra, 34, 044001
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2019a, PhRvL, 123, 011102
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2019b, PhRvD, 100, 104036
Abbott, R., Abbott, T. D., Abraham, S., et al. 2021a, PhRvD, 103, 122002
Abbott, R., Abbott, T. D., Abraham, S., et al. 2021b, ApJL, 913, L7
Abbott, R., Abbott, T. D., Abraham, S., et al. 2021c, arXiv:2111.03606
Abbott, R., Abbott, T. D., Abraham, S., et al. 2021d, arXiv:2108.01045
Abbott, R., Abbott, T. D., Abraham, S., et al. 2021e, ApJL, 913, L7
Abbott, R., Abe, H., Acernese, F., et al. 2021f, arXiv:2111.03604
Ade, P. A. R., Aghanim, N., Arnaud, M., et al. 2016, A&A, 594, A13
Antonelli, A., Burke, O., & Gair, J. R. 2021, MNRAS, 507, 5069

Berti, E., Buonanno, A., & Will, C. M. 2005, PhRvD, 71, 084025
Borhanian, S., & Sathyaprakash, B. S. 2022, arXiv:2202.11048
Cornish, N., Sampson, L., Yunes, N., & Pretorius, F. 2011, PhRvD, 84, 062003
Cutler, C., & Flanagan, I. E. 1994, PhRvD, 49, 2658
Cutler, C., & Vallisneri, M. 2007, PhRvD, 76, 104018
Dideron, G., Mukherjee, S., & Lehner, L. 2022, arXiv:2209.14321
Dietrich, T., Khan, S., Dudi, R., et al. 2019, PhRvD, 99, 024029
Finn, L. S. 1992, PhRvD, 46, 5236
Gamba, R., Breschi, M., Bernuzzi, S., Agathos, M., & Nagar, A. 2021, PhRvD, 103, 124015
Ghosh, A., Ghosh, A., Johnson-McDaniel, N. K., et al. 2016, PhRvD, 94, 021101
Hall, E. D., Cahillane, C., Izumi, K., Smith, R. J. E., & Adhikari, R. X. 2019, CQGra, 36, 205006
Himemoto, Y., Nishizawa, A., & Taruya, A. 2021, PhRvD, 104, 044010
Hu, Q., & Veitch, J. 2022, PhRvD, 106, 044042
Husa, S., Khan, S., Hannam, M., et al. 2016, PhRvD, 93, 044006
Isi, M., Chatziioannou, K., & Farr, W. M. 2019, PhRvL, 123, 121101
Janquart, J., Baka, T., Samajdar, A., Dietrich, T., & Van Den Broeck, C. 2022, arXiv:2211.01304
Khan, S., Husa, S., Hannam, M., et al. 2016, PhRvD, 93, 044007
Kunert, N., Pang, P. T. H., Tews, I., Coughlin, M. W., & Dietrich, T. 2022, PhRvD, 105, L061301
Li, T. G. F., Del Pozzo, W., Vitale, S., et al. 2012, PhRvD, 85, 082003
Maggiore, M., Van Den Broeck, C., Bartolo, N., et al. 2020, JCAP, 2020, 050
Meidam, J., Tsang, K. W., Goldstein, J., et al. 2018, PhRvD, 97, 044033
Mishra, C. K., Arun, K. G., Iyer, B. R., & Sathyaprakash, B. S. 2010, PhRvD, 82, 064010
Moore, C. J., Finch, E., Buscicchio, R., & Gerosa, D. 2021, iSci, 24, 102577
Moore, C. J., & Gair, J. R. 2014, PhRvL, 113, 251101
Oguri, M. 2018, MNRAS, 480, 3842
Ossokine, S., Buonanno, A., Marsat, S., et al. 2020, PhRvD, 102, 044055
Pang, P. T. H., Calderón Bustillo, J., Wang, Y., & Li, T. G. F. 2018, PhRvD, 98, 024019
Pankow, C., Chatziioannou, K., Chase, E. A., et al. 2018, PhRvD, 98, 084016
Pizzati, E., Sachdev, S., Gupta, A., & Sathyaprakash, B. 2022, PhRvD, 105, 104016
Powell, J. 2018, CQGra, 35, 155017
Pratten, G., Garcia-Quiros, C., Colleoni, M., et al. 2021, PhRvD, 103, 104056
Punturo, M., Abernathy, M., Acernese, F., et al. 2010, CQGra, 27, 194002
Pürrer, M., & Haster, C.-J. 2020, PhRvR, 2, 023151
Reali, L., Antonelli, A., Cotesta, R., et al. 2022, arXiv:2209.13452
Reitze, D., Adhikari, R. X., Ballmer, S., et al. 2019, BAAS, 51, 35
Relton, P., & Raymond, V. 2021, PhRvD, 104, 084039
Saini, P., Favata, M., & Arun, K. G. 2022, PhRvD, 106, 084031
Samajdar, A., Janquart, J., Van Den Broeck, C., & Dietrich, T. 2021, PhRvD, 104, 044003
Sampson, L., Cornish, N., & Yunes, N. 2013, PhRvD, 87, 102001
Sun, L., Goetz, E., Kissel, J. S., et al. 2020, CQGra, 37, 225008
Vallisneri, M., & Yunes, N. 2013, PhRvD, 87, 102002
Vitale, S., & Del Pozzo, W. 2014, PhRvD, 89, 022002
Williamson, A. R., Lange, J., O'Shaughnessy, R., et al. 2017, PhRvD, 96, 124041
Wu, S., & Nitz, A. H. 2022, arXiv:2209.03135
Zimmerman, A., Haster, C.-J., & Chatziioannou, K. 2019, PhRvD, 99, 124044