Accumulating errors in tests of general relativity with the Einstein Telescope: overlapping signals and inaccurate waveforms

QIAN HU$^{1,2}$ AND JOHN VETICH$^{1}$

$^{1}$Institute for Gravitational Research, School of Physics and Astronomy, University of Glasgow, Glasgow, G12 8QQ, United Kingdom

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 (SNR) is high and the number of events is large. In this work, we investigate the impacts of overlapping signals and inaccurate waveform models on tests of general relativity. We simulate mock catalogs for Einstein Telescope and perform parametric tests of GR using waveform models with different levels of inaccuracy. We find the systematic error could accumulate towards false deviations of GR when combining results from multiple events, even though data from most events prefers GR. The waveform inaccuracies contribute most to the systematic errors, but a high merger rate could magnify the effects of systematics due to the incorrect removal of detected overlapping signals. We also point out that testing GR using selected events with high SNR is even more vulnerable to false deviations from GR. The problem of error accumulation is universal; we emphasize that it should be taken into consideration in future catalog-level data analysis, and further investigations, particularly in waveform accuracy, will be essential for third generation detectors.

1. INTRODUCTION

Detection of gravitational waves (GWs) from compact binary coalescences (CBCs) provides ideal tests of general relativity (GR) in the strong-field regime (Abbott et al. 2016, 2017a,b, 2019a,b, 2021a,b). The latest GW event catalogs contain nearly 100 CBC events (Abbott et al. 2021c,d), based on which various tests have been performed (Abbott et al. 2021a,b). No concrete evidence of deviation from GR has been found yet. In the next decades, the third-generation (3G) ground-based GW detectors are expected to detect $\mathcal{O}(10^5)$ CBC events per year, with signal-to-noise ratio (SNR) up to thousands (Maggiore & et al 2020; Himemoto et al. 2021; Samajdar et al. 2021; Relton & Raymond 2021; Oguri 2018). Since the statistical uncertainty shrinks when SNR increases or a catalog of events are combined, observations from 3G GW detectors are able to obtain much tighter constraints on gravity theories.

The inspiring enlarged detection catalog and higher SNR bring many difficulties to data analysis. For the purpose of testing GR (and any other theories), one needs to make sure the systematic errors are small so that the analysis will not favor the wrong theories. The parameterized tests of GR (Meidam & et al 2018) suffer from the same problems as in parameter estimation (PE), 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 (Hu & Veitch 2022; Williamson et al. 2017) and are expected to be more important in future high SNR detections (Cutler & Vallisneri 2007; Gamba et al. 2021; Purrer & Haster 2020). Additionally, the 3G detectors are able to detect multiple signals at the same time. Detected overlapping signals cannot be perfectly removed from the data, and could have non-negligible impact on PE when the merger times of overlapping signals are close (Himemoto et al. 2021; Samajdar et al. 2021; Relton & Raymond 2021; Antonelli et al. 2021; 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 types of 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 (Kunert et al. 2022; Moore et al. 2021).

q.hu.2@research.gla.ac.uk
John.Veitch@glasgow.ac.uk
In this work, we perform parameterized post-Newtonian (PPN) coefficient tests (Meidam & et al 2018) with our simulated event catalogs and inaccurate waveforms. Our results show systematic error does accumulate and could lead to a false measurement of deviation from GR. We find overlapping signals could magnify the effects of waveform systematics and cause wrong conclusions even if the waveform meets future accuracy requirement (Pürrer & Haster 2020). Even worse, the selected high-SNR events without known overlapping signals (so-called “golden events”) may be more vulnerable to biased conclusions.

2. SYSTEMATIC BIASES IN PPN TESTS

2.1. Estimating systematic errors

The generic formalism of estimating PE systematic errors is first proposed in Cutler & Vallisneri (2007) and then generalized and validated by Antonelli et al. (2021). Let \( \hat{\theta} \) be the parameters of GWs, the frequency domain data of a GW detector is denoted as \( d(\hat{\theta}) \). We simply have

\[
d(\hat{\theta}) = h(\hat{\theta}) + n,
\]

where \( n \) is the detector noise, and \( h(\hat{\theta}) \) is the GW detected by the detector. Under the assumption that the noise is stationary and Gaussian, the likelihood in GW PE is

\[
L(\hat{\theta}) \propto e^{-\frac{1}{2}(d-h)^{T}(d-h)} = e^{-\frac{1}{2}(n|n)},
\]

where \( (\ldots|\ldots) \) is the inner product (Finn 1992). The optimal SNR is \( \rho = \sqrt{\langle h|\hat{h} \rangle} \). For more than one data stream, the inner product 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 prior), the maximum point \( \hat{\theta}_{ML} \) satisfies

\[
\partial_{i} \ln L|_{\theta=\hat{\theta}_{ML}} = (\partial_{i} h|d-h) |_{\theta=\hat{\theta}_{ML}} = 0,
\]

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

\[
d - h = n + \delta H + \Delta \theta^{j} \partial_{j} h_{m}.
\]

The first term \( n \) is what \( d-h \) is supposed to be, the noise arising in the detector. The second term \( \delta H = h(\hat{\theta}_{\text{real}}) - h_{m}(\hat{\theta}_{\text{real}}) \) 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 linear expansion of \( h_{m}(\hat{\theta}_{\text{real}}) - h_{m}(\hat{\theta}_{ML}) \), and \( \Delta \theta^{j} \) is the error of the \( j \)th parameter from the maximum likelihood estimator, and we adopt Einstein notation to indicate the sum over parameters. Substitute Eq. 4 into Eq. 3 and approximate all derivatives at \( \hat{\theta}_{ML} \), we get

\[
\Delta \theta^{j} \approx (\Gamma^{-1})^{ij}(\partial_{j} h_{m}|n + \delta H) = \Delta \theta^{i}_{\text{stat}} + \Delta \theta^{i}_{\text{sys}},
\]

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

2.2. PPN formalism, choices of parameters and waveforms

The test of parameterized post-Newtonian coefficients (Meidam & et al 2018) is a generic formalism for testing deviations from GR. 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 \( \{\phi_{0}, \ldots, \phi_{r}\} \) and \( \{\phi_{\delta l}, \phi_{ed}\} \), phenomenological coefficients \( \{\delta_{0}, \ldots, \delta_{3}\} \), and merger-ringdown parameters \( \{\alpha_{0}, \ldots, \alpha_{3}\} \). Deviations \( p_{i} \rightarrow (1 + \delta p_{i}) p_{i} \) are introduced as the violations of GR; \( \delta p_{i} = 0 \) returns to GR. Testing GR is converted to estimating the testing parameters \( \delta p_{i} \). Although a specific modified gravity theory could create deviation in more than one testing parameter, previous works have shown that including one testing parameter at once is enough, and even more efficient to find violations from GR because it avoids the correlations between testing parameters and GR parameters (Meidam & et al 2018; Sampson et al. 2013). In this work, we choose \( \delta \phi_{0} \) as the example testing parameter. We assume GR is the correct theory and focus on whether the PPN test falsely indicates deviations of GR.

As a qualitative investigation, we only consider three GR parameters (chirp mass \( M \), mass ratio \( q \) and coalescence time \( t_{c} \)) in PE (i.e., in Fisher matrix), and other parameters are treated as perfectly known. This choice captures the parameter that appears in the leading PN term and the corresponding PPN modifications. The decisive parameter in the analysis of overlapping signals, \( t_{c} \), is also included. We induce a non-zero \( \delta \phi_{2} \) to mimic inaccurate waveform models. \( \delta \phi_{2} \) is a phenomenological coefficient for
the intermediate regime between inspiral and merger and has insignificant correlation with the testing parameter $\delta \bar{\theta}_0$. We assume $\delta \bar{\beta} \approx 0$ is our model waveform, while the “real” waveform could have $\delta \bar{\beta} = 0, 5 \times 10^{-2},$ or $5 \times 10^{-4}$. The first case means our model waveform is perfect, and all systematic errors will come from overlapping signals. The second case generates waveform mismatches around $10^{-4} - 10^{-3}$, which corresponds to the current waveform accuracy (Pratten et al. 2021; Ossokine et al. 2020). The last case produces mismatches around $10^{-7} - 10^{-6}$ and corresponds to the expectations for future waveform accuracy (Purrer & Haster 2020; Hu & Veitch 2022). The excess strains from inaccurate waveforms can be written as

$$\delta H_{\text{wf}} = h(\bar{\theta}_{\text{real}}) - h_m(\bar{\theta}_{\text{ML}}) \approx h(\bar{\theta}_{\text{ML}}) - h_m(\bar{\theta}_{\text{ML}}),$$

where $\bar{\theta}$ 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. As a conservative estimation of errors in overlapping signals PE and following Antonelli et al. (2021), we ignore waveform systematic errors in $\bar{\theta}$, and adopt the lowest order approximation for its correlation with the main signal. Substituting it into Eq. 5, one obtains the covariance of the first term in systematic error

$$< \Delta \theta_{\text{DO}}^i \Delta \theta_{\text{DO}}^j > = \left( \Gamma_{\text{mix}}^{-1} \Gamma_1^{-1} \Gamma_2^{-1} (\bar{\Gamma}_{\text{mix}}^{-1})^T \Gamma^{-1} \right)_{ij},$$

where $(\Gamma_{\text{mix}})$ encodes the correlation between two signals and $\Gamma_1 = (\partial h/\partial h')$ is the Fisher matrix of the overlapping signal. The second term in Eq. 7 represents the inaccurate waveform model we use to subtract signals, and can be calculated the same way as the waveform systematics, yielding $\Delta \theta_{\text{DO}}^i = (\Gamma_1)^{-1} (\partial h_m/\partial h_m')$. In this work, systematic errors from detected overlapping signals is calculated as $\Delta \theta_{\text{DO}}^i$ plus a random sample drawn from a multivariate Gaussian distribution with covariance matrix Eq. 8 and zero mean. For more than one detected overlapping signal, Eq. 7 can be extended by defining $h'$ as the summation of all GWs in the data (Antonelli et al. 2021), which enlarges the dimension of $\Gamma_{\text{mix}}$. The undetected overlapping signal simply contributes to systematic error by $\delta H_{\text{UO}} = \sum_{\text{undetected}} h''(\bar{\theta}_{\text{real}})$. It is accessible in our simulation but unknown in real data analysis.

We consider BBH and BNS sources, and assume their distribution in redshift $z$ follows the analytical approximation (Oguri 2018; Samajdar et al. 2021)

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

which is then converted to observable event rate by multiplying a factor $\frac{1}{1+z} \frac{dz}{dV}$. Here $V_c$ is the comoving volume and we employ Planck15 cosmology (Ade et al. 2016). Note that “observable” GWs need to achieve an SNR of 8 to be “detectable”. $a_{1,2,3,4}$ are model parameters. We set $a_2 = 1.0, a_3 = 2.1, a_4 = 30$ to mimic a peak at $z \sim 2$. $a_1$ is scaled based on local merger rate given by Abbott et al. (2021c) ($R_{\text{BNS}} = 320^{+90}_{-40}$ and $R_{\text{BBH}} = 23.9^{+14.6}_{-6.6} \text{Gpc}^{-3} \text{yr}^{-1}$) such that $R_{\text{GW}}(z = 0) = R_{\text{BNS/BBH}}$. We choose three values for $a_1$ which corresponds 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. (2021c), while all BNS systems are set to be same: $1.45 + 1.4 \text{M}_\odot, \Lambda_1 = \Lambda_2 = 425$. IMRPhenomPv2\_NRTidal (Dietrich et al. 2019) is used to generate BNS waveforms with the same $\delta \bar{\beta}_2$ 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 zero spins, isotropically distributed inclination and source sky direction; and uniformly distributed coalescence time, phase, and polarization angle.

A summary of low, median, and high merger rates catalogs is shown in Tab. 1. It shows most BBH events will
not have an overlapping signal near their merger times, which implies overlapping signals contribute to systematic errors less frequently than waveform systematics. The undetected overlapping signal happens more often than the detected: the unnoticeable confusion background has drawn attention in recent works (Wu & Nitz 2022; Reali et al. 2022) and needs further investigation.

Several simplifications have been adopted in our mock catalog: we regard BNS as a background and use only BBH as the test source; we ignore neutron star-black hole (NSBH) mergers and other possible types of sources; we use an analytical merger rate peaks at $z \sim 2$, ignoring binaries from Pop 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 and expect the real situation would lie somewhere between our lowest and highest estimates.

Signals are injected into the 3rd generation GW detector Einstein Telescope with ET-B PSD (Punturo et al. 2010) and Gaussian noise realization. The frequency band is set to 10-4096Hz.

3. RESULTS

3.1. Single events

We first present an example event. The main signal is from a BBH with $M_c = 32M_\odot$, $q = 0.9$ and network SNR of 50.3. The overlapping signal is an equal mass BBH with $M_c = 20M_\odot$. We scale its SNR from ~37 down to $\lesssim 8$ to make it detectable or undetectable. We vary the merger time difference (by 0.1s 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 $\delta \phi_0$. We define the error ratio as the absolute value of the ratio of systematic error and statistical uncertainty. The PPN coefficient test is subject to false deviations due to the repeating alignments and misalignments of phases of the two GWs. Overlap error is not symmetric around $\Delta t = 0$ because the waveforms of the two sources are not symmetric, but the peak is always located in the region $|\Delta t| \leq 1s$, which means the overlapping signal only produces a large influence when two mergers are very close. These characteristics are consistent with previous works (Himemoto et al. 2021; Samajdar et al. 2021; Relton & Raymond 2021; Antonelli et al. 2021). When $|\Delta t|$ is large, it is waveform inaccuracy that dominates the systematic error.

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. This implies different types of systematic errors are correlated and could be a magnifying factor for each other, as expected from Eq. 7. However, we want to emphasize it is also possible for undetected signals to produce significant systematic errors in our simulation.

The statistical error $\Delta \theta_{\text{stat}} = (\Gamma^{-1})^{ij}(\partial_i h|n) \approx 1/|h| \approx 1/$SNR, while the systematic error $\Delta \theta_{\text{stat}} = (\Gamma^{-1})^{ij}(\partial_j h|\delta H)$ does not necessarily shrink when the SNR increases. An example is waveform systematics of the main signal. Therefore, systematic error may dominate in high SNR scenario. We calculate systematic errors for each BBH event in our mock catalog, and plot the absolute error and error ratio with SNR in Fig. 2. The error ratio could exceed one for the “current” waveform, and it happens more often when SNR > 30 despite the fact that high SNR events are rarer. Error ratios for the perfect waveform and “future” waveform are usually below one. However, as pointed out by Moore et al. (2021), false deviation could be achieved even though estimations for individual events are generally accurate.

![Figure 1](image-url). The error ratio of $\delta \phi_0$ varies with merger time difference. The main signal has $M_c = 32M_\odot$, $q = 0.9$, and SNR of 50.33. The overlapping signal is an equal mass BBH with $M_c = 20M_\odot$. SNR of the overlapping signal is adjusted by changing luminosity distance: detected overlap is shown in upper panel, and the undetected is the lower one. We use three kinds of waveforms mentioned in Sec. 2.2: perfect waveform (solid line), “current” waveform (dashed line), and “future waveform” (faint dotted-dashed line).
We will investigate this in more detail in the next subsection.

3.2. Error accumulation in a catalog

In this subsection, we will combine the results from all BBHs in a catalog and show how systematic error in testing GR accumulates. There are several ways of combining results from multiple events (Zimmerman et al. 2019; Isi et al. 2019). We employ two straightforward methods: multiplying likelihoods (equivalently, multiplying posteriors if priors are flat) and multiplying Bayes factors. The former assumes the modification parameter is the same for all events, while the latter allows the modification parameter to vary across events.

We assume the posterior follows a multivariate Gaussian distribution with covariance matrix $\Gamma^{-1}$ and mean equal to injection values plus systematic errors. From the first event in a catalog, we multiply the posterior from new events one by one and calculate the error ratio. Note that, the multiplication of two Gaussian distributions is still Gaussian whose mean (systematic error) is a linear combination of two original means. Therefore, a plus systematic error and a minus systematic error could neutralize if we combine events by multiplying likelihood. Bayes factor, on the other hand, takes the form Moore et al. (2021)

$$B \sim \frac{Z_{\text{nonGR}}}{Z_{\text{GR}}} = \frac{\int d\hat{\theta}_{\text{nonGR}} L_{\text{nonGR}}}{\int d\hat{\theta}_{\text{GR}} L_{\text{GR}}}$$

$$= \sqrt{2\pi e^{\frac{1}{2} \Gamma_{\hat{\phi}_0 \hat{\phi}_0} \Delta \theta_{\text{sys}}^2}} \sqrt{\frac{\det \Gamma_{\text{GR}}}{\det \Gamma_{\text{nonGR}}}},$$

(10)

where $\Gamma_{\text{nonGR}}$ is the Fisher matrix including the testing parameter, while $\Gamma_{\text{GR}}$ only includes GR parameters. $\Gamma_{\hat{\phi}_0 \hat{\phi}_0} = (\partial h/\partial \hat{\phi}_0)(\partial h/\partial \hat{\phi}_0)$. $\Delta \theta_{\text{sys}}$ is the systematic error of $\hat{\phi}_0$. Here the systematic error appears as a quadratic term, so errors with different signs can accumulate.

Considering the arbitrary sequence of events, we permute the sequence 200 times and extract the ensemble average and 68% confidence interval of the error ratio and Bayes factor. The results are shown Fig. 3.

Let $N_{\text{event}}$ be the number of events. When multiplying likelihoods, the statistical uncertainty shrinks as $1/\sqrt{N_{\text{event}}}$. The absolute error of the testing parameter also decreases, but at a slower pace due to the perturbations from newly coming systematic errors. It also follows $1/\sqrt{N_{\text{event}}}$ if there were no systematic errors and 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

|               | # of observable binaries | Detected overlaps on BBH events | Undetected overlaps on BBH events |
|---------------|--------------------------|---------------------------------|----------------------------------|
|               | BBH  | BNS  | # of overlaps | # (fraction) of events | # of overlaps | # (fraction) of events |
| Low           | 56526 | 286088 | 0            | 48380 (98%)       | 0            | 45991 (93%)       |
|               |      |      | 1            | 937 (1.9%)        | 1            | 3217 (6.5%)        |
|               |      |      | 2            | 11 (0.022%)       | 2            | 119 (0.24%)        |
| Median        | 88300 | 1144354 | 0           | 73224 (95%)       | 0            | 58921 (77%)        |
|               |      |      | 1            | 3574 (4.6%)       | 1            | 15658 (20%)        |
|               |      |      | 2            | 74 (0.096%)       | 2            | 2108 (2.7%)        |
|               |      |      | 3            | 1 (0.0010%)       | 3            | 174 (0.23%)        |
| High          | 143349 | 2896647 | 0           | 112745 (90%)      | 0            | 63932 (51%)        |
|               |      |      | 1            | 11496 (9.2%)      | 1            | 42931 (34%)        |
|               |      |      | 2            | 589 (0.47%)       | 2            | 14143 (11%)        |
|               |      |      | 3            | 19 (0.015%)       | 3            | 3208 (2.6%)        |
|               |      |      | 4            | 45 (0.037%)       | 4            | 540 (0.43%)        |
|               |      |      | 5            | 86 (0.069%)       | 5            | 86 (0.069%)        |
|               |      |      | 6            | 6 (0.0050%)       | 6            | 6 (0.0050%)        |
|               |      |      | 7            | 3 (0.0020%)       | 7            | 3 (0.0020%)        |

Table 1. A summary of three mock catalogs. From left to right, it shows catalog type, observable BBH and BNS per year (note this is not detectable), number of detected overlapping signals and number of (and fraction in) detected BBH events that have this number of detected overlapping signals, and the same statistic for undetected overlapping signals.
Figure 2. Absolute error (first column) and error ratio (second column) of $\delta \hat{\phi}_0$ vs SNR for low (uppermost row), median (median row), high (bottom row) merger rate catalogs. Each point represents a BBH event. Blue points are for perfect waveform and all systematic errors come from overlapping signals; red points stand for “current waveform” case and yellow points for “future waveform” case. Grey points in the first column are statistical errors. This plot shows error ratios greater than 1 are mostly from “current waveform” and high SNR events.

keeps contributing to the systematic errors. The slower decay of systematic error results in a climb of error ratio. At some point (typically $\sim 10^3$ events) it leads to a false deviation of GR for the “current” waveform. For the better waveform, the error ratio climbs as well, but it keeps below the statistical level until $10^5 - 10^6$ events.

Multiplying Bayes factors, however, is more sensitive to systematics errors in PE because of the quadratic term. The “current” waveform could lead to a strong false deviation of GR with only tens of events. Intriguingly, this method differs from multiplying likelihoods in the behavior of tests with “future” waveform in the high merger rate catalog. Multiplying likelihoods does not lead to false deviation in this case, but multiplying Bayes factors finally claims deviation from GR after some oscillations at $\sim 5000$ events. Compared with the lower merger rates, we conclude the frequent inaccurate subtractions of detected overlapping signals amplify the effects of waveform systematics, even if we have used a relatively accurate waveform model with mismatches around $10^{-7} - 10^{-6}$.

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 SNR 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: SNR above a chosen threshold (50 or 200) and there is no detected overlapping signals. Results for the error ratio and Bayes factor are shown in Fig. 4.

It turns out that the high SNR 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 also able to produce false deviations in all three catalogs. There is no qualitative difference between results of different merger rates, because we have removed events with detected overlapping signals which magnifies waveform systematic effects. As mentioned in Sec. 3.1, statistical uncertainty decreases as 1/SNR while systematics do not as long as waveform is not perfect. The false deviation for golden events is not surprising from this angle, but it
Figure 3. Systematic error accumulates with the increase of number of events. The first column shows the absolute error of $\delta \phi_0$ and the second column shows the error ratio. The third column is the Bayes factor. Zoom-in windows show the behavior of the first several events. Solid red lines are ensemble average for “current waveform”, dashed red lines are for “future waveform”, and blue lines stand for the perfect waveform. The shadow along lines are 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. In both methods, false deviations can be achieved with the increase in the number of events, but multiplying the Bayes factor does it faster due to the quadratic term. The “future waveform” in the third column shows that a high overlap rate could be an amplifying factor of waveform inaccuracies.

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. 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 merger rate (and therefore the number of overlaps), we can achieve a false deviation of GR which could not happen at a lower merger rate. One can avoid this correlated error by selecting events with no detected overlapping signals, and, if one prefers, with high SNR as well. However, we have showed these events produce biases much faster because waveform systematics dominate in high SNR scenario.

We re-emphasize 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 (Sun et al. 2020; Hall et al. 2019), glitches (Powell 2018; Finkow et al. 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
Figure 4. Similar to Fig. 3, the error ratio and Bayes factor accumulation. The left two columns show results from SNR > 50 events, the right two columns are for SNR > 200 events. Compared with Fig. 3, it shows that tests with high SNR events are more likely to make a false deviation from GR.

detectors. An obvious solution to these issues is continuing to improve waveform model accuracy and instrument stability, but we believe 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).

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. QH is supported by CSC. JV is supported by STFC grant ST/V005634/1.

REFERENCES
Abbott, B. P., et al. 2016, Phys. Rev. Lett., 116, 221101, doi: 10.1103/PhysRevLett.116.221101
—. 2017a, Phys. Rev. Lett., 119, 141101, doi: 10.1103/PhysRevLett.119.141101
—. 2017b, Phys. Rev. Lett., 118, 221101, doi: 10.1103/PhysRevLett.118.221101
—. 2019a, Phys. Rev. Lett., 123, 011102, doi: 10.1103/PhysRevLett.123.011102
—. 2019b, Phys. Rev. D, 100, 104036, doi: 10.1103/PhysRevD.100.104036
Abbott, R., et al. 2021a, Phys. Rev. D, 103, 122002, doi: 10.1103/PhysRevD.103.122002
—. 2021b, Tests of General Relativity with GWTC-3. https://arxiv.org/abs/2112.06861
—. 2021c, GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run. https://arxiv.org/abs/2111.03606
