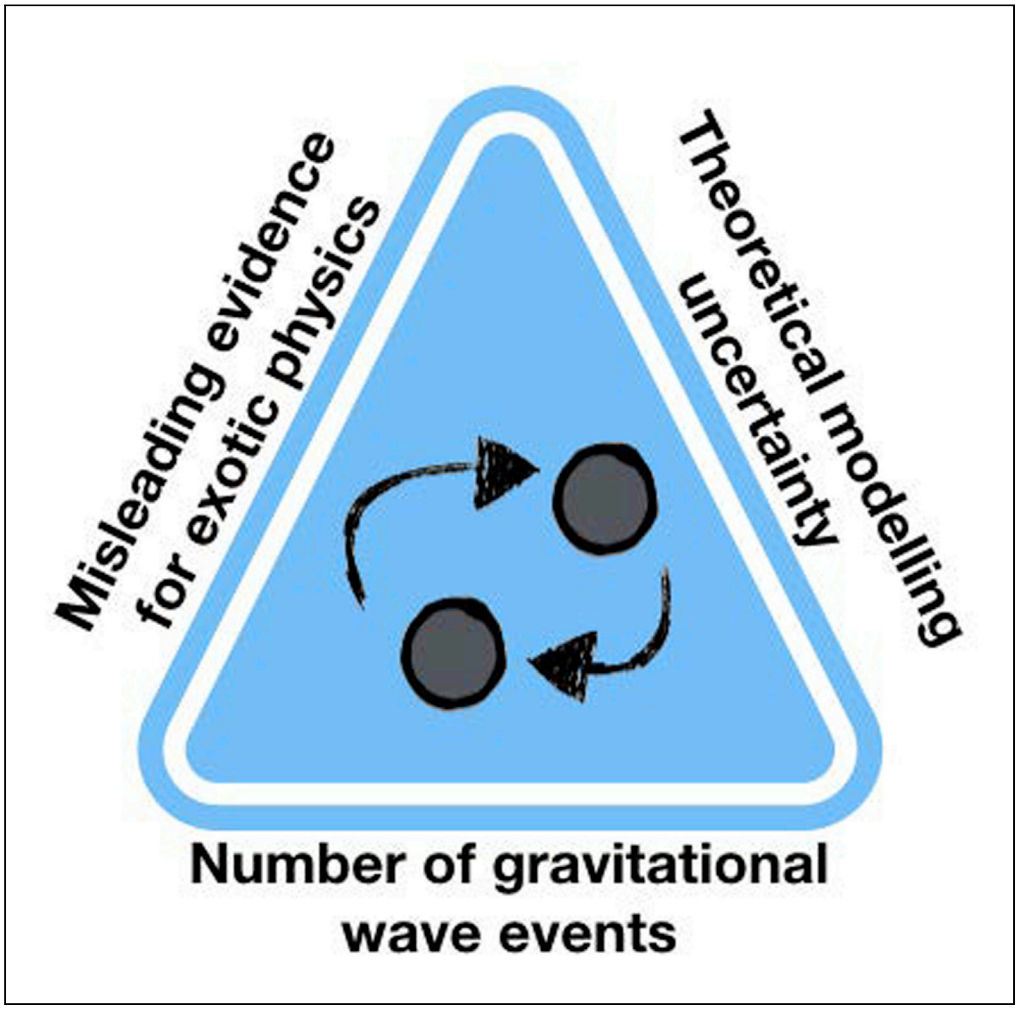
iScience

CellPress
OPEN ACCESS

Article

Testing general relativity with gravitational-wave catalogs: The insidious nature of waveform systematics



Christopher J. Moore, Eliot Finch, Riccardo Buscicchio, Davide Gerosa

moorecj@bham.ac.uk

Highlights

Gravitational waves provide a new way to test Einstein's theory of general relativity

We consider how such tests might be strengthened by combining information from different events in a gravitational-wave catalog

We urge caution-small theoretical errors in the models can build up in a catalog and lead to evidence for new physics, even when none exists

This is urgent-we show that incorrect evidence for a failure of general relativity might appear very soon

Moore et al., iScience *24*, 102577
June 25, 2021
https://doi.org/10.1016/j.isci.2021.102577

iScience

CellPress
OPEN ACCESS

Article

Testing general relativity with gravitational-wave catalogs: The insidious nature of waveform systematics

Christopher J. Moore,[1,2,*] Eliot Finch,[1] Riccardo Buscicchio,[1] and Davide Gerosa[1]

SUMMARY

Gravitational-wave observations of binary black holes allow new tests of general relativity (GR) to be performed on strong, dynamical gravitational fields. These tests require accurate waveform models of the gravitational-wave signal; otherwise waveform errors can erroneously suggest evidence for new physics. Existing waveforms are generally thought to be accurate enough for current observations, and each of the events observed to date appears to be individually consistent with GR. In the near future, with larger gravitational-wave catalogs, it will be possible to perform more stringent tests of gravity by analyzing large numbers of events together. However, there is a danger that waveform errors can accumulate among events: even if the waveform model is accurate enough for each individual event, it can still yield erroneous evidence for new physics when applied to a large catalog. This paper presents a simple linearized analysis, in the style of a Fisher matrix calculation that reveals the conditions under which the apparent evidence for new physics due to waveform errors grows as the catalog size increases. We estimate that, in the worst-case scenario, evidence for a deviation from GR might appear in some tests using a catalog containing as few as 10–30 events above a signal-to-noise ratio of 20. This is close to the size of current catalogs and highlights the need for caution when performing these sorts of experiments.

INTRODUCTION

The detection of gravitational waves (GWs) by LIGO (LIGO Collaboration, 2015) and Virgo (Virgo Collaboration, 2015) has made possible new tests of general relativity (GR) in the strong-field regime (Yunes and Siemens, 2013; Berti et al., 2015; Yunes et al., 2016; LIGO and Virgo Collaborations, 2019, 2020b). Numerous detections of binary coalescences have been made so far (mostly binary black holes) and in the coming years the size of this catalog of detections will continue to grow (LIGO and Virgo Collaborations, 2018). Furthermore, observations of GWs in different frequency bands will reveal new types of sources and enable other, complementary tests of GR. As the sensitivity of the instruments improve, the signal-to-noise ratios (SNRs) of the loudest individual events will increase and such tests will become increasingly stringent.

However, care must be taken when interpreting the results of these test or we risk incorrectly claiming evidence for new physics.

Among the wide range of possible tests of GR are parametric tests. These involve the introduction of additional degrees of freedom to the theory, which are described by one or more new parameters. These additional quantities are then measured along with the astrophysical source parameters. Frameworks for performing parametric tests include both the introduction of artificial coefficients to various terms in the waveform (for a review, see Will 2014) and extensions involving specific beyond-GR theories, such as those motivated by quantum gravity (Yunes et al., 2016). (See Chua and Vallisneri 2020 for a discussion of possible drawbacks of parametric tests.)

As most GW signals will have an SNR close to detection threshold (Schutz, 2011; Chen and Holz, 2014), combining information from multiple signals is an attractive avenue to perform stronger tests. The way different events are put together crucially depends on the test one wishes to perform (Zimmerman et al., 2019). For GR modifications with parameters that are thought to be common among all the events

[1]School of Physics and Astronomy & Institute for Gravitational Wave Astronomy, University of Birmingham, Birmingham B15 2TT, UK

[2]Lead contact

*Correspondence: moorecj@bham.ac.uk

https://doi.org/10.1016/j.isci.2021.102577

in the catalog (for instance, the mass of the graviton), one should multiply the individual likelihoods on the deviation parameters to find the combined, catalog-level likelihood on the deviation. If instead each event can have a different, independent deviation parameter value (for instance, addition black hole degrees of freedom, aka *hairs*) one should instead find the combined, catalog evidence by multiplying the individual event Bayes' factors. These possibilities represent the two extrema of a broader class of catalog tests where the non-GR parameters follow some distribution that can depend on the GR quantities (such as masses and spins), and can be tackled using a hierarchical Bayesian approach (Isi et al., 2019). Some of the analyses presented by LIGO and Virgo Collaborations, (2020b) were carried out within this framework.

Tests of GR can be affected by inaccuracies in the GW signal models used to analyze the data (Lindblom et al., 2008; Pürrer and Haster, 2020). Waveform systematics can erroneously lead to evidence for a deviation from GR. In other words, even if GR is the correct description of nature, unmodelled waveform systematics can lead us to believe the opposite. Going beyond single events, in this paper we explore the role of waveform systematics when using multiple signals in a catalog to test GR. For each event, we employ a simplified linear analysis, similar to that by Cutler and Vallisneri (2007); Vallisneri and Yunes (2013); Gair and Moore (2015), and study how systematic waveform errors introduce biases in the beyond-GR parameters. By studying the two extreme cases highlighted above—GR deviations which are common and different among events—we extend the linear analysis to show how the effects of waveform systematics can accumulate as the size of the catalog grows. Even if the imperfect waveform model is good enough to safely analyze each of the events individually, it may induce evidence for beyond-GR physics with arbitrarily high confidence when applied to a large catalog. This serves to highlight the dangerous and insidious nature waveform systematics in GW catalogs.

This paper is organized as follows. We first derive the building blocks of our analysis by investigating the impact of systematics on single events. We then illustrate how those ingredients enter a catalog analysis. We then present the results of our findings applied to two sets of simulated event catalogs of increasing complexity. Finally, we draw our conclusions.

Linear signal analysis: single events

In order to establish how the evidence for new physics scales with the individual event SNR and the number of events in the catalog, we use a simplified, linearized signal analysis, in the spirit of a Fisher matrix calculation.

In the following analysis, it is assumed that GR is the correct description of nature. For each individual GW event the observed data contains a sum of instrumental noise, n, and a GW signal. Here, we restrict our analysis to the case when a single interferometer is involved in the observation; the extension to multiple interferometers is straightforward and does not significantly change the following arguments. The observed data, s, can be written

$$s = n + \underbrace{h(\alpha_{\rm Tr}=0;\theta_{\rm Tr}) + \Delta h(\theta_{\rm Tr})}_{\text{GW Signal}}. \qquad \text{(Equation 1)}$$

The true signal is (hopefully) close to our waveform model, $h(\alpha = 0;\theta)$, but will inevitably include some modeling error, denoted here by $\Delta h(\theta)$. Our model, $h(\alpha;\theta)$, is a function of the parameters that describe the source in GR, θ^i. This includes both intrinsic (masses, spins, etc.) and extrinsic (sky position, distance, etc.) quantities. The true source parameters (which are unknown *a priori*) are denoted by $\theta_{\rm Tr}$. The model error is also a function of θ (for example, regions of parameter space with asymmetric mass ratios and strong spin precession will typically have larger model errors; Pürrer and Haster 2020). As we are searching for parameterized deviations from GR, our model, $h(\alpha;\theta)$, is also a function of at least one modified gravity parameter, α. This parameter quantifies the deviation from GR and we assume it is defined such that GR is smoothly recovered in the limit $\alpha \to 0$.

The analysis of the single GW event in Equation 1 involves performing Bayesian parameter inference over the combined parameter space $\lambda \equiv (\alpha;\theta)$. Assuming the instrumental noise is Gaussian, the likelihood $\mathcal{L}(\alpha;\theta) \equiv P(s|\alpha;\theta)$ is given by

$$\begin{aligned} \log\mathcal{L}(\alpha;\theta) &= -\frac{1}{2}|s - h(\alpha;\theta)|^2 + c, \\ &= -\frac{1}{2}|n - \delta h(\alpha;\theta) + \Delta h(\theta_{\rm Tr})|^2 + c. \end{aligned} \qquad \text{(Equation 2)}$$

The norm is defined as $|\bullet|^2 = \langle \bullet | \bullet \rangle$, where angle brackets denote the usual signal inner product (Thorne, 1987; Moore et al., 2015). The constant c is an unimportant normalization, and we have defined $\delta h(\alpha;\theta) \equiv h(\alpha;\theta) - h(\alpha_{\rm Tr};\theta_{\rm Tr})$, where $\theta_{\rm Tr}$ and $\alpha_{\rm Tr} = 0$ denote the true parameters. For simplicity, we assume that the prior on λ is approximately flat within the range that $\mathcal{L}$ has significant support; therefore, the likelihood is proportional to the Bayesian posterior distribution.

If the SNR is large, then the posterior is expected to be strongly peaked in a relatively narrow region around $\theta_{\rm Tr}$ and $\alpha_{\rm Tr}$. We assume that this region is small enough that the model can be approximated as being linear in all parameters. We Taylor-expand our model about the true parameters as follows:

$$\begin{aligned}\delta h(\alpha;\theta) &\approx \left.\frac{\partial h}{\partial \alpha}\right|_{(0,\theta_{\rm Tr})} \alpha + \left.\frac{\partial h}{\partial \theta^i}\right|_{(0,\theta_{\rm Tr})} \delta\theta^i + \ldots \\ &= \left.\frac{\partial h}{\partial \lambda^\mu}\right|_{\lambda_{\rm Tr}} \delta\lambda^\mu + \mathcal{O}\left(\delta\lambda^2\right).\end{aligned} \qquad \text{(Equation 3)}$$

We have defined $\delta\theta = \theta - \theta_{\rm Tr}$, and $\delta\lambda = \lambda - \lambda_{\rm Tr}$ and all derivatives are evaluated at the true parameters (hereafter, this will be omitted from our notation). The index μ labels the components of the combined parameter vector, λ^μ. Hereafter, we retain only leading order terms in $\delta\lambda$.

The likelihood in Equation 2 is peaked at the maximum likelihood (ML) parameters, $\lambda_{\rm ML}$, which are defined implicitly by

$$\left.\frac{\partial \log \mathcal{L}}{\partial \lambda}\right|_{\lambda = \lambda_{\rm ML}} = 0. \qquad \text{(Equation 4)}$$

Using Equations 2 and 3, this can be solved to find

$$\lambda_{\rm ML} = \lambda_{\rm Tr} + \Delta\lambda_{\rm stat} + \Delta\lambda_{\rm sys}, \qquad \text{(Equation 5)}$$

where (using Einstein summation convention)

$$\Delta\lambda^\mu_{\rm stat} = \left(\Gamma^{-1}\right)^{\mu\nu} \left\langle n \middle| \frac{\partial h}{\partial \lambda^\nu} \right\rangle, \qquad \text{(Equation 6)}$$

$$\Delta\lambda^\mu_{\rm sys} = \left(\Gamma^{-1}\right)^{\mu\nu} \left\langle \Delta h(\theta_{\rm Tr}) \middle| \frac{\partial h}{\partial \lambda^\nu} \right\rangle, \qquad \text{(Equation 7)}$$

and $\Gamma_{\mu\nu}$ is the Fisher matrix,

$$\Gamma_{\mu\nu} = \left\langle \frac{\partial h}{\partial \lambda^\mu} \middle| \frac{\partial h}{\partial \lambda^\nu} \right\rangle. \qquad \text{(Equation 8)}$$

From Equation 5, it can be seen that the ML parameters are close to the true source parameters but shifted by both statistical and systematic errors. The statistical error $\Delta\lambda_{\rm stat}$ depends on the random noise realization, n, in the observed data. The systematic error $\Delta\lambda_{\rm sys}$ depends on the model error Δh.

Now that we have found the location of the maximum likelihood, we may evaluate the second derivatives of Equation 2, $\partial_\mu \partial_\nu \log\mathcal{L}$, at the ML parameters and expand the log likelihood to second order about this point. Doing this, we find

$$\log\mathcal{L}(\lambda) \approx c' - \frac{1}{2}\Gamma_{\mu\nu}(\lambda - \lambda_{\rm ML})^\mu (\lambda - \lambda_{\rm ML})^\nu, \qquad \text{(Equation 9)}$$

where c' is another unimportant normalization constant. Within the approximations that have been made, the likelihood (and the posterior) is approximately a multivariate Gaussian on the parameters λ with mean vector $\lambda_{\rm ML} \equiv (\alpha_{\rm ML}, \theta_{\rm ML})$ and covariance matrix Γ^{-1}.

We wish to use the observed data to test GR. Therefore, we investigate the 1D marginalized posterior on the α parameter to see if it is peaked away from the GR value, $\alpha = 0$. Because the full posterior in Equation 9 is a multivariate Gaussian, the 1D marginalization integral can be carried out analytically. The 1D marginalized posterior on the α parameter reads

$$P(\alpha) = \int \mathrm{d}\theta\, \mathcal{L}(\theta,\alpha) = \frac{\exp\left[-\frac{(\alpha-\alpha_{\rm ML})^2}{2\sigma_\alpha^2}\right]}{\sqrt{2\pi\sigma_\alpha^2}}, \qquad \text{(Equation 10)}$$

where $\alpha_{\rm ML} = \alpha_{\rm stat} + \alpha_{\rm sys}$, and

$$\alpha_{\rm stat} = \left(\Gamma^{-1}\right)^{0\nu}\left\langle n\left|\frac{\partial h}{\partial \lambda^{\nu}}\right\rangle, \qquad \text{(Equation 11)}$$

$$\alpha_{\rm sys} = \left(\Gamma^{-1}\right)^{0\nu}\left\langle \Delta h(\theta_{\rm Tr})\left|\frac{\partial h}{\partial \lambda^{\nu}}\right\rangle, \qquad \text{(Equation 12)}$$

$$\sigma_{\alpha}^{2} = \left(\Gamma^{-1}\right)^{00} = \left[\Gamma_{00} - \Gamma_{0i}\left(\gamma^{-1}\right)^{ij}\Gamma_{j0}\right]^{-1}. \qquad \text{(Equation 13)}$$

where $\gamma_{ij} = \Gamma_{ij}$ is the lower-right block of the Fisher matrix, and in the final equality we have used the block-matrix inversion formula.

The optimal SNR, defined as $\rho(\lambda) = |h(\lambda)|$, is a convenient measure of the strength of the signal. In order to investigate the scaling with the SNR, ρ, it will be convenient to separate it from the other parameters by defining the normalized model $\widehat{h} = h/\rho$, with $|\widehat{h}| = 1$. We also define the normalized Fisher matrix $\widehat{\Gamma}_{\mu\nu} = \Gamma_{\mu\nu}/\rho^{2}$ and the normalized model error $\Delta\widehat{h} = \Delta h/\rho$.

It will also be convenient to rescale the deviation parameter (i.e. redefine $\alpha \to \kappa\alpha$ where κ is a constant) such that $\widehat{\Gamma}_{00} \equiv \left|\partial\widehat{h}/\partial\alpha\right|^{2} = 1$. We are always free to perform such a rescaling and this does not interfere with our earlier choice of placing flat priors on all parameters.

We also assume that $\widehat{\Gamma}_{0i} \equiv <\partial\widehat{h}/\partial\alpha|\partial\widehat{h}/\partial\theta^{i}> = 0$ where $i \neq 0$, i.e. the deviation parameter induces waveform changes which are orthogonal to those arising from changes in all the GR parameters. Although this is probably rarely true in practice, it is a conservative assumption in the sense that it makes the problem of waveform systematics as severe as possible by minimizing the estimate for σ_{α} [see Equation 13], while keeping $\alpha_{\rm sys}$ fixed, thereby maximizing the chances that the model errors lead us to erroneously claim to have seen a deviation from GR. It is this worst-case scenario, which we choose to study here in order to better understand when we need to worry about waveform systematics.

Under these simplifying assumptions and conventions, the statistical fluctuations in the deviation, given in Equation 11, are distributed as a Gaussian random variable, $\alpha_{\rm stat} = z/\rho$ where $z \sim \mathcal{N}(0,1)$. Furthermore, the expression for the standard deviation of the distribution, given in Equation 13, simplifies to $\sigma_{\alpha} = 1/\rho$. This just leaves the systematic offset in the deviation parameter in Equation 12, which can be written as

$$\alpha_{\rm sys} = (\widehat{\Gamma}^{-1})^{00}\left\langle \Delta\widehat{h}(\theta_{\rm Tr})\left|\frac{\partial\widehat{h}}{\partial\alpha}\right\rangle = \left|\Delta\widehat{h}(\theta_{\rm Tr})\right|\cos\iota\,, \qquad \text{(Equation 14)}$$

where the first equality follows from Equation 12 and our assumption that $\widehat{\Gamma}_{0i} = 0$, the second equality follows from our renormalization of α such that $\left|\partial\widehat{h}/\partial\alpha\right| = 1$, and ι is defined as the angle between the signals $\Delta\widehat{h}(\theta_{\rm Tr})$ and $\partial\widehat{h}/\partial\alpha$. The quantity ι has the interpretation of an angle if the signals (which are discretely sampled time series) are thought of as being very high-dimensional vectors in some signal space, $\mathcal{S}\,\mathbb{R}^{\rm high\ dim}$. The angle ι encodes information on how the model error couples with the deviation parameter. The worst-case scenario is when $|\cos\iota|$ is maximal and occurs when $\iota = 0$ or π; therefore, we set $\cos\iota = \pm 1$ in the following. The norm of the model error, $|\Delta\widehat{h}|$, is related to the mismatch which is commonly defined in GW applications as (e.g. Lindblom et al., 2008)

$$\mathcal{M} = 1 - \frac{\left\langle \widehat{h} + \Delta\widehat{h}\middle|\widehat{h}\right\rangle}{\left|\widehat{h}\right|\left|\widehat{h} + \Delta\widehat{h}\right|} = 1 - \cos\phi \approx \frac{\phi^{2}}{2}\,, \qquad \text{(Equation 15)}$$

where ϕ is the generalized angle between the signals $\widehat{h}$ and $\widehat{h} + \Delta\widehat{h}$. Provided the $\Delta\widehat{h}$ is small, the angle ϕ will also be small and is bounded above by $\phi < |\Delta\widehat{h}|$. The exact value of ϕ will depend on the details of the model error and can be considered to be quasi-random. If the signal space dimensionality is large, and if $\Delta\widehat{h}$ is a random vector, then the distribution of ϕ-values will be peaked near the maximum value. Therefore, we set $\phi = |\Delta\widehat{h}|$ and, using the small angle approximation in Equation 15, obtain

$$\mathcal{M} \approx \frac{\left|\Delta\widehat{h}\right|^2}{2}. \quad \text{(Equation 16)}$$

Finally, using Equation 16 to eliminate $\Delta\widehat{h}$ from Equation 14, the systematic error is $\alpha_{\rm sys} = \sqrt{2\mathcal{M}}\cos\iota$. In summary, the 1D marginalized posterior on the GR deviation parameter α is given by Equation 10, with

$$\alpha_{\rm stat} = \frac{z}{\rho}\,, \quad \text{(Equation 17)}$$

$$\alpha_{\rm sys} = \sqrt{2\mathcal{M}}\cos\iota\,, \quad \text{(Equation 18)}$$

$$\sigma_\alpha = \frac{1}{\rho}\,, \quad \text{(Equation 19)}$$

where $z \sim \mathcal{N}(0,1)$ is a random number associated with the noise realization and $\cos\iota = \pm 1$ is a random choice of sign associated with the model error, $\Delta\widehat{h}$.

Note that the systematic offset does not scale with SNR. Therefore, there always exists a critical SNR above which we are in danger of erroneously claiming a deviation from GR. When analyzing a single GW event for a deviation from GR, we are safe from the effects of model errors if $\alpha_{\rm sys} \ll \sigma_\alpha$. From Equations 17, 18, and 19, we see that the average size statistical error equals the systematic error when $\rho = 1/\sqrt{2\mathcal{M}}$; therefore, we are safe from the effects of model errors if $\rho \ll 1/\sqrt{\mathcal{M}}$.

Because the posterior in Equation 2 is Gaussian, it is possible to evaluate the Bayesian evidence integral analytically. Doing so, and letting $k = \dim(\theta)$ [hence $\dim(\lambda) = k+1$)], gives

$$Z_{\rm nonGR} \equiv \int {\rm d}\lambda\, \mathcal{L}(\lambda) = e^{c'}\sqrt{\frac{(2\pi)^{k+1}}{\det\Gamma_{\mu\nu}}}. \quad \text{(Equation 20)}$$

Because our waveform model $h(\alpha,\theta)$ is an extension of GR, it includes GR as a sub-model. The GR sub-model is the hypersurface $\alpha = 0$ of the full model. Using the same assumptions described above for the full model, the GR likelihood (and hence the posterior) on this hypersurface can be found from Equation 2 and is given by

$$\log\mathcal{L}_{\rm GR}(\theta) = c' - \frac{1}{2}\Gamma_{00}\alpha_{\rm ML}^2 - \frac{1}{2}(\theta-\theta_{\rm ML})^i\Gamma_{ij}(\theta-\theta_{\rm ML})^j\,, \quad \text{(Equation 21)}$$

where $i, j \in \{1, 2\ldots, k\}$ label the components of θ. The evidence for the GR sub-model can also be evaluated analytically and reads

$$Z_{\rm GR} \equiv \int {\rm d}\theta\, \mathcal{L}_{\rm GR}(\theta) = e^{c'-\Gamma_{\alpha\alpha}\alpha_{\rm ML}^2/2}\sqrt{\frac{(2\pi)^k}{\det\Gamma_{ij}}}. \quad \text{(Equation 22)}$$

The odds ratio (or Bayes' factor) in favor of a deviation from GR is defined as

$$\mathcal{B} \equiv \frac{\Pi}{A}\frac{Z_{\rm nonGR}}{Z_{\rm GR}}\,, \quad \text{(Equation 23)}$$

where Π is the prior odds ratio in favor of a deviation from GR and $A = \alpha_{\rm max} - \alpha_{\rm min}$ is the prior range on α which must be included to account for the differing prior volumes between the models. Computing the Bayes' factor in this manner between nested models is known as the Savage-Dickey density ratio (Dickey, 1971). Under our conservative assumption that $\widehat{\Gamma}_{0i} = 0$, the Fisher matrix has a block diagonal structure and $\det\Gamma_{\mu\nu} = \Gamma_{00}\det\Gamma_{ij}$. The Bayes' factor simplifies to

$$\mathcal{B} = \frac{\Pi}{A}\sqrt{\frac{2\pi}{\Gamma_{00}}}\exp\left(\frac{1}{2}\Gamma_{00}\alpha_{\rm ML}^2\right). \quad \text{(Equation 24)}$$

Recalling that $\alpha_{\rm ML} = \alpha_{\rm stat} + \alpha_{\rm sys}$ and $\Gamma_{00} = \sigma_\alpha^{-2}$ and using the results in Equations 17, 18, and 19 gives

$$\log\mathcal{B} = \log\left(\frac{\Pi}{A}\frac{\sqrt{2\pi}}{\rho}\right) + \frac{\left(z+\rho\sqrt{2\mathcal{M}}\cos\iota\right)^2}{2}\,. \quad \text{(Equation 25)}$$

Note that the first term in Equation 25, known as the *Occam penalty*, decays slowly with increasing SNR. However, the second term in Equation 25 grows rapidly. This reveals again, in another guise,

the existence of a critical SNR above which we are in danger of erroneously claiming a deviation from GR due to model errors. From the final term in Equation 25, we again conclude that when analyzing individual GW events for deviations from GR we are safe from the effects of waveform systematics if $\rho \ll 1/\sqrt{\mathcal{M}}$.

Linear signal analysis: event catalogs

The previous section considered tests of GR with a single GW event and concluded that we expect our analysis to be robust against the effects of waveform systematic errors provided $\rho \ll 1/\sqrt{\mathcal{M}}$. When we are in the situation that no single event in the catalog shows a clear deviation from GR, it is desirable to combine all the observed events to "dig deeper" and perform more stringent tests of GR. This section extends the linearized analysis of the previous section and investigates the impact of waveform systematics on such catalog tests.

A GW catalog contains N events, indexed by $m \in \{1, 2, \ldots, N\}$. As described in the previous section, each event provides us with an independent measurement of the deviation parameter and the likelihoods for these measurements, $P_m(\alpha)$, are all Gaussian of the form in Equation 10 with parameters given by Equations 11, 12, and 13, or Equations 17, 18, and 19. We replace $z \to z_m$, $\rho \to \rho_m$, $\cos\iota \to \cos\iota_m$ and $\mathcal{M} \to \mathcal{M}_m$ to distinguish different events.

There are different ways of combining the information from multiple events. Two particularly simple ways are (i) multiplying the 1D marginalized posteriors on α and (ii) multiplying the odds ratios (Zimmerman et al., 2019). These approaches can been seen as two extrema of a more generic hierarchical-inference strategy (Isi et al., 2019). In particular, the former assumes that the deviation parameter takes the same value for each event in the catalog while latter assumes that the parameter takes an independent value for each event. We consider each approach in turn.

Multiplying likelihoods

Under the assumption that the deviation takes the *same* value in each event of the catalog, the combined posterior on the deviation parameter is given by the product of the independent likelihoods in each of the N events:

$$P_{\rm same}(\alpha) = \prod_{m=1}^{N} P_m(\alpha) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\left(\alpha - \alpha_{\rm ML}^{\rm same}\right)^2}{2\left(\sigma_\alpha^{\rm same}\right)^2}\right]. \qquad \text{(Equation 26)}$$

The product of several Gaussian distributions with known means and variances is another Gaussian, the mean and variance of which are given by

$$\alpha_{\rm ML}^{\rm same} = \left(\sigma_\alpha^{\rm same}\right)^2 \sum_{m=1}^{N} \frac{\alpha_{{\rm ML},m}}{\sigma_{\alpha,m}^2}, \qquad \text{(Equation 27)}$$

$$\sigma_\alpha^{\rm same} = \left(\sum_{m=1}^{N} \sigma_{\alpha,m}^{-2}\right)^{-1/2}. \qquad \text{(Equation 28)}$$

Using the combined catalog posterior on the deviation parameter in Equation 26, we can now compute combined Bayes' factor in favor of a deviation from GR. This is computed using the Savage-Dickey density ratio and reads (see e.g. Sivia and Skilling 2006)

$$\mathcal{B}_{\rm same} = \frac{\Pi}{A}\sqrt{2\pi}\sigma_\alpha^{\rm same} \exp\left[\frac{1}{2}\left(\frac{\alpha_{\rm ML}^{\rm same}}{\sigma_\alpha^{\rm same}}\right)^2\right]. \qquad \text{(Equation 29)}$$

Multiplying Bayes' factors

Each catalog event provides some evidence for or against a deviation from GR which is quantified by the Bayes' factor $\mathcal{B}_i$ in Equation 25. Under the assumption that the deviation takes independent values in each event, the combined Bayes' factor in favor of a deviation from GR is given by the product

$$\mathcal{B}_{\rm diff} = \prod_{m=1}^{N} \mathcal{B}_m. \qquad \text{(Equation 30)}$$

Simple event catalogs

In this section, we perform Monte-Carlo simulations of highly simplified, mock GW catalogs. It is assumed throughout that GR is the correct description of nature, but that our GR waveforms contain modeling errors. The purpose of these simulations is to understand under what situations model errors might lead us to mistakenly think that we have observed a deviation from GR.

For simplicity, in this section it is assumed that all events in the catalog have the same SNR, ρ. It is also assumed that all events have the same amount of modeling error and that this leads to a mismatch value of $\mathcal{M} = 10^{-3}$ in each event. Finally, in this section the effects of instrumental noise are also neglected; i.e. it is assumed that the specific noise realization in each observed GW event is $n = 0$, which corresponds to setting $z = 0$ for each event. All of these assumptions are relaxed in the later sections where more realistic catalogs are considered.

We simulate a mock catalog by first choosing the number of events to be considered, N. We then choose the value of the SNR, ρ. All that remains is to choose the value of $\cos\iota = \pm 1$ for every event; this is done in two ways described below.

We can then compute the evidence in favor of a deviation from GR either under the assumption that the deviation parameter takes the same value for each event [$\mathcal{B}_{\text{same}}$, see Equation 29; i.e. multiplying likelihoods] or else under the assumption that the deviation takes independent values in each event [$\mathcal{B}_{\text{diff}}$, see Equation 30; i.e. multiplying Bayes' factors]. We consider these two cases in turn.

Multiplying likelihoods

Each individual event, labeled by $m \in \{1, 2, \ldots, N\}$, gives a measurement of the deviation parameter. Under the assumptions described above, and neglecting the statistical fluctuations due to the noise, the likelihood on α from this measurement is a 1D Gaussian with a mean $\alpha_{\text{ML},m} = \alpha_{\text{sys},m}$ given by Equation 18 and a standard deviation $\sigma_{\alpha,m}$ given by Equation 19.

Multiplying these likelihood functions together gives a single, combined catalog measurement of the deviation parameter. The likelihood from this combined measurement is also a 1D Gaussian with a mean and standard deviation given by Equations 27 and 28 respectively. These expressions simplify further to give

$$\alpha_{\text{ML}}^{\text{same}} = \frac{\sqrt{2\pi}}{N} \sum_{m=1}^{N} (\cos\iota_m) \,, \qquad \text{(Equation 31)}$$

$$\sigma_{\alpha}^{\text{same}} = \frac{1}{\sqrt{N}\rho}. \qquad \text{(Equation 32)}$$

The Bayes' factor in favor of a deviation from GR that comes from this combined catalog measurement of α was derived in Equation 29 and simplifies further here to give

$$\log\mathcal{B}_{\text{same}} = \log\left(\frac{\Pi}{A}\frac{\sqrt{2\pi}}{\sqrt{N}\rho}\right) + \frac{\mathcal{M}\rho^2}{N}\left(\sum_{m=1}^{N}\cos\iota_m\right)^2. \qquad \text{(Equation 33)}$$

The heat maps in Figure 1 show the numerical, Monte-Carlo results for the Bayes' factor $\mathcal{B}_{\text{same}}$ under two possible scenarios. The left panel of Figure 1 illustrates a case where the model errors differ among events such that they are equally likely to favor positive and negative value for α. We mimic this scenario by randomly selecting either $\cos\iota = 1$ or $\cos\iota = -1$ for each event. The right panel of Figure 1 illustrates the case where the model errors are such that they always tend to favor a deviation of α with the same sign. We mimic this scenario by always choosing $\cos\iota = +1$ in every event. In reality, the situation is likely to be somewhere in between these two extreme possibilities. The real distribution of $\cos\iota$ will depend on the astrophysical population of sources and any detection biases. Unless we are very unlucky, the modeling error is unlikely to always resemble exactly the same type of deviation from GR. However, because we analyze all GW events using the same waveform model the modeling errors are also not independent between events. For simplicity, the results in Figure 1 are scaled to $\Pi = A = 1$.

It is possible to understand analytically the distinctly different scaling of $\log\mathcal{B}_{\text{same}}$ observed in the two panels of Figure 1. First, we consider a case where the model errors are such that they always tend

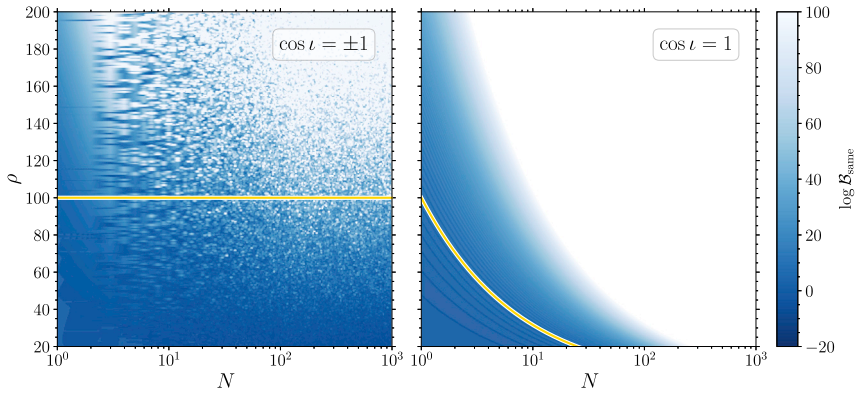


Figure 1. The log Bayes' factor, $\log\mathcal{B}_{\rm same}$, in favor of a deviation from GR under the assumptions that the deviation parameter takes the same value in all events

These results were obtained for the highly simplified mock GW event catalogs described in the main text; in particular, this assumes that each of the N events has the same SNR, ρ, and the same model mismatch error, $\mathcal{M} = 10^{-3}$, and we neglect noise fluctuations by setting $z = 0$. Left panel: the model errors are equally likely to favor positive or negative α ($\cos\iota = \pm 1$ randomly in all events); in this case the model errors do not accumulate strongly when combining the catalog events. The speckled pattern comes from the random, Monte-Carlo choices for $\cos\iota$ in each event and would tend to average out if we simulated multiple catalog realizations. Right panel: the model errors always favor positive α ($\cos\iota = 1$ in all events); in this case the model errors accumulate rapidly as the number of events increases and $\mathcal{B}_{\rm same}$ increases with N. The yellow lines shows the analytic prediction of the threshold $\log\mathcal{B}_{\rm same} \approx 10$ above which model errors might cause us to erroneously claim to have detected a deviation from GR; the horizontal line in the left hand comes from Equation 35, while the line in the right hand figure comes from Equation 34. In both panels the analytic predictions for the threshold follow the contours of the heatmap.

to favor a deviation of α with the same sign (right panel). In this case $\cos\iota_m = +1$ for every event and we simply have that $\sum_m \cos\iota = N$. For large catalogs, the expression for the Bayes' factor in Equation 33 now becomes

$$\log\mathcal{B}_{\rm same} \approx \mathcal{M} N \rho^2 \,. \qquad \text{(Equation 34)}$$

The logarithm term is neglected as we are mainly interested in the limiting behavior for large N and ρ.

If we choose an arbitrary threshold Bayes' factor (say, $\mathcal{B}_{\rm threshold} = e^{10}$) above which we will claim to have seen evidence for a deviation from GR, then rearranging Equation 34 gives an expression for the threshold SNR as a function of catalog size. This is plotted as the yellow curve in the right panel of Figure 1, where it can be seen to follow the contours of the heatmap. We see that even if $\rho \ll 1/\sqrt{\mathcal{M}}$, and our waveform model is comfortably good enough to analyze each event individually, there always exists a critical catalog size about which the Bayes' factor in favor of a deviation from GR exceeds any threshold. In this case, as the catalog size increases there is a growing danger of erroneously claiming to detect a deviation from GR due to the model error.

Second, we consider the case where the model errors differ among events such that they are equally likely to favor positive and negative value for α. This scenario was mimicked in our toy model by choosing $\cos\iota_m = \pm 1$ randomly. Therefore, the term $\sum_m \cos\iota$ is a new random variable, and in the limit of large catalog size (i.e. as $N \to \infty$) the central limit theorem implies that this will be normally distributed as $\sum_m \cos\iota \sim \mathcal{N}(0, \sqrt{N})$. It follows that the combination $(\sum_m \cos\iota)^2/N$ appearing in Equation 33 is now distributed as a χ^2 random variable with 1 degree of freedom and has an expectation value of 1. Therefore, the expectation value for the Bayes' factor in Equation 33 becomes

$$\log\mathcal{B}_{\rm same} \approx \mathcal{M}\rho^2 \,. \qquad \text{(Equation 35)}$$

Again, we neglect the logarithm term as it is unimportant in the limit of large ρ. This expression can be rearranged to find the threshold SNR above which $\mathcal{B}_{\rm same}$ exceeds the threshold; this is plotted as the horizontal yellow line in the left panel of Figure 1. Note the very different scaling from that in Equation 34; in this

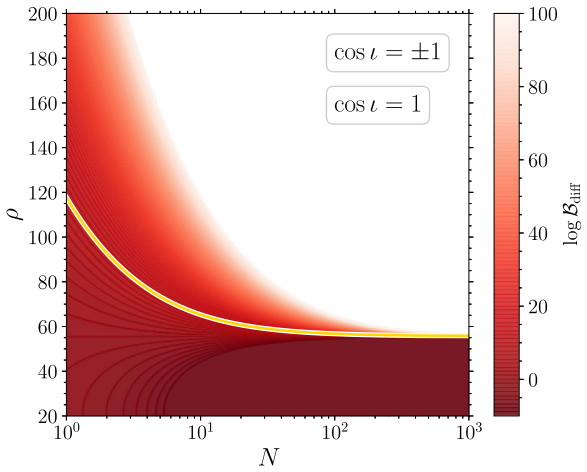


Figure 2. The log Bayes' factor, $\log\mathcal{B}_{\rm diff}$, in favor of a deviation from GR under the assumption that the deviation parameter takes independent values in each event

These results were obtained for the highly simplified mock GW event catalogs described in the main text. In this situation, the two cases where the model errors equally favor positive and negative α ($\cos\iota = \pm 1$) and where they always favor positive α ($\cos\iota = 1$) give identical results. The yellow lines shows the analytic prediction in Equation 36 of the threshold $\log\mathcal{B}_{\rm diff} = 10$ above which model errors might cause us to erroneously claim to have detected a deviation from GR; this prediction closely follows the contours of the heatmap. Here, the model errors accumulate only if the SNR in each individual event is above a critical value, $\rho_i \gtrsim \rho_* = 55.68$ [see Equation 37]. If the individual event SNRs are below this critical value then the Bayes' factor actually decreases with increasing catalog size leading us to (correctly) favor the GR hypothesis.

case, the model errors do not accumulate as the catalog size increases and the danger of erroneously claiming a deviation from GR does not increase with N.

Multiplying Bayes' factors

Figure 2 shows the results of another Monte-Carlo analysis, this time combining the catalog events under the assumption that the GR deviation parameter takes independent values in each event. As discussed above, this corresponds to multiplying together the Bayes' factors for each individual catalog event in order to obtain the combined $\mathcal{B}_{\rm diff}$ catalog Bayes' factor in favor of a GR deviation.

Again, we consider a simplified GW catalog containing N events each at the same SNR, ρ. As before, we further assume that each event has the same mismatch, $\mathcal{M} = 10^{-3}$, due to modeling errors and we neglect the statistical fluctuations due to noise by setting $z = 0$ for each event. Again, we could consider both a scenario where the model errors equally favor positive and negative α (i.e. randomly selecting $\cos\iota_m \pm 1$) and a scenario where the model errors always favor positive α (i.e. always setting $\cos\iota_m = 1$). However, in this worst-case scenario, these two possibilities give identical results (inspecting Equation 25 we see that, when setting $z = 0$, the individual event Bayes' factor $\mathcal{B}_m$ depends only on $\cos\iota_m^2$).

The heatmap in Figure 2 shows the numerical, Monte-Carlo results for the Bayes' factor $\mathcal{B}_{\rm diff}$. As before, it is possible to understand analytically the observed scaling of $\log\mathcal{B}_{\rm diff}$. The Bayes' factor for each individual event is given by Equation 25 (with $z = 0$, as we are neglecting statistical noise fluctuations in this section). The combined log Bayes' factor $\mathcal{B}_{\rm diff}$ is simply the sum of the individual log Bayes' factors and is given by

$$\log\mathcal{B}_{\rm diff} \approx N\left[\log\left(\frac{\Pi}{A}\frac{\sqrt{2\pi}}{\rho}\right) + \mathcal{M}\rho^2\right]. \qquad \text{(Equation 36)}$$

This expression can be rearranged to find the SNR at which the Bayes' factor exceeds the threshold for claiming evidence for a deviation from GR. This predicted threshold SNR is plotted as a yellow line in Figure 2 for the choice $\mathcal{B}_{\rm threshold} = e^{10}$.

In this case, we see a qualitatively new behavior as the catalog size, N, increases. Whenever a new event is added to the catalog, there is a competition between the model error [second term in Equation 36] which

tends to increase the Bayes' factor in favor of a deviation from GR and the Occam penalty [first term in Equation 36] which tends to do the opposite. Which effect ends up winning depends on the SNR. There exists a critical SNR, ρ_*, above which the Bayes' factor increases with N and this is given by the solution to

$$\log\left(\frac{A\rho_*}{\sqrt{2\pi\Pi}}\right) = \mathcal{M}\rho_*^2 \,, \qquad \text{(Equation 37)}$$

which in our example where $\Pi = A = 1$ and $\mathcal{M} = 10^{-3}$ is $\rho_* = 55.68$. Below this critical SNR we are safe from model systematics and the Bayes' factor in favor of a deviation from GR actually decreases as the catalog grows. In the large-N limit and within the assumption of this model, this implies that evidence against GR grows (is suppressed) in catalogs made of events with SNR $\rho > \rho_*$ ($\rho < \rho_*$).

More realistic event catalogs

The GW catalogs considered in the previous section were rather unrealistic. The SNR of each event was the same; the mismatch was the same for every waveform, and the statistical fluctuations due to individual noise realizations were ignored. In this section, we relax these assumptions and perform Monte-Carlo simulations of more realistic catalogs.

We simulate catalogs of N events where the SNR of individual events are drawn from a $P(\rho) \propto \rho^{-4}$ distribution, which is the expected distribution for a population of sources in a Euclidean universe with no cosmological evolution in the merger rate (Schutz, 2011; Chen and Holz, 2014). The lower (upper) cutoffs in the SNR distribution where chosen to be $\rho_{\rm low} = 20$ ($\rho_{\rm high} = 200$). Our results are somewhat sensitive to the lower cutoff of the SNR distribution; the value of 20 used here is larger than the usual LIGO/Virgo detection threshold $\rho \to 8$ because: (i) we do not want to invalidate the assumptions behind the linearized analysis which are only expected to hold for large SNR, and (ii) it is reasonable to expect that delicate analyzes such as tests of GR will only be performed on a subset of loud events. This was done, for example, in the recent analysis by LIGO and Virgo Collaborations, (2020b) where none of the marginal triggers with false alarm rate $>10^{-3}{\rm yr}^{-1}$ were investigated.

Instead of fixing the mismatch at a single value $\mathcal{M} = 10^{-3}$ for all events, we now allow the mismatch to differ between events by drawing this from a distribution with lower (upper) cutoffs of $\mathcal{M}_{\rm low} = 10^{-4}$ ($\mathcal{M}_{\rm high} = 10^{-2}$). The choice of these cutoffs is roughly motivated by the accuracy of existing models and the results from Figure 13 of Blackman et al. (2017). The shape distribution of $\mathcal{M}$ between these limits is difficult to predict as it will depend on the waveform models used, on where in parameter space this model perform best/worst, and on the distribution of the event properties such as mass ratio and spins presented to us nature in the catalog. All of these are difficult to predict. However, given existing observations, we expect most events will be nearly equal mass and with low spins (LIGO and Virgo Collaborations, 2020a), where our waveform models perform relatively well (although a small number of more exotic events should be expected). Therefore, the distribution of $\mathcal{M}$ will be skewed toward low values. Here, we consider two possibilities: a bad case $P(\mathcal{M}) \propto \mathcal{M}^{-1}$ and a good case $P(\mathcal{M}) \propto \mathcal{M}^{-2}$.

The log Bayes' factors obtained from the catalogs assuming the deviation takes the same value in every event (i.e. $\mathcal{B}_{\rm same}$; multiplying likelihoods) are shown in Figure 3. We consider the same two cases for $\cos\iota$ as was done for the simple mock catalogs. In the left hand panel we see that the Bayes' factor does not scale strongly with the size of the catalog; this agrees with the results in the left panel of Figure 1 obtained using the simpler catalogs. In the right hand panel, we see that the Bayes' factor in favor of a deviation from GR increases rapidly with the size of the catalog. This is also in agreement with the results in the right panel of Figure 1 obtained using the simpler catalogs.

The results for the log Bayes' factors obtained assuming that the deviation parameter takes independent values in each event (i.e. $\mathcal{B}_{\rm diff}$; multiplying the individual Bayes' factors) are shown in Figure 4. Again, we consider both $\cos\iota = \pm 1$ and $\cos\iota = 1$. We see very similar behavior in both cases (consistent with the identical results found for the simple catalogs). In both panels we see that the Bayes' factor scales strongly with the size of the catalog but that it can either increase or decrease depending on the distribution of the mismatches. This behavior can be understood from the results in Figure 2 obtained using the simpler catalogs. If $P(\mathcal{M}) \propto \mathcal{M}^{-2}$, most events have very small mismatches and therefore have $\rho < \rho_*$ (i.e. below the yellow line in Figure 2) and, as the catalog size increases, the increasing Occam penalty dominates over the effect of the model error and GR is favored. On the other hand, if $P(\mathcal{M}) \propto \mathcal{M}^{-1}$, more

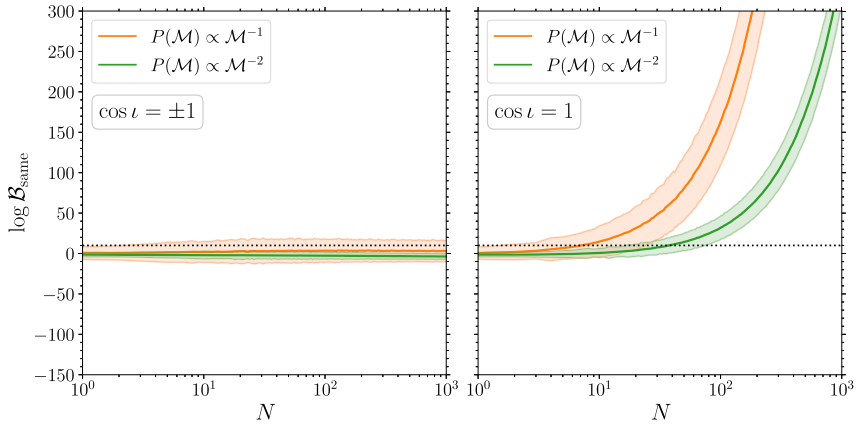


Figure 3. The log Bayes' factor, $\log\mathcal{B}_{\rm same}$, in favor of a deviation from GR under the assumption that the deviation parameter takes the same value in all events

The solid lines indicates the mean value of $\log\boldsymbol{\mathcal{B}}_{\rm diff}$ obtained from 10^4 realisations of the more realistic simulated catalogs described in the main text while the shaded region between the two paler lines indicates the $\pm 1\sigma$ spread in this set of simulated catalogs. The dashed horizontal line denotes the threshold $\log\boldsymbol{\mathcal{B}}_{\rm same} = 10$: above this line there is a risk that model errors cause us to incorrectly claim a deviation from GR, while below this line we correctly conclude that GR is favored. Left panel: the model errors are equally likely to favor positive or negative α ($\cos\iota = \pm 1$ randomly in all events); in this case the model errors do not accumulate strongly when combining the catalog events. Right panel: the model errors always favor positive α ($\cos\iota = 1$ in all events); in this case the model errors accumulate rapidly as the number of events increases and the evidence for a deviation from GR grows with the size of the catalog. Depending on the distribution of the model errors, misleading evidence for a deviation from GR can appear with catalogs with as few as ≈ 10 events above the minimum SNR of $\rho > 20$.

events have larger mismatches and $\rho > \rho_*$ (i.e. above the yellow line in Figure 2) and, as the catalog size increases, the accumulating model errors overcome the Occam penalty and a deviation from GR is favored.

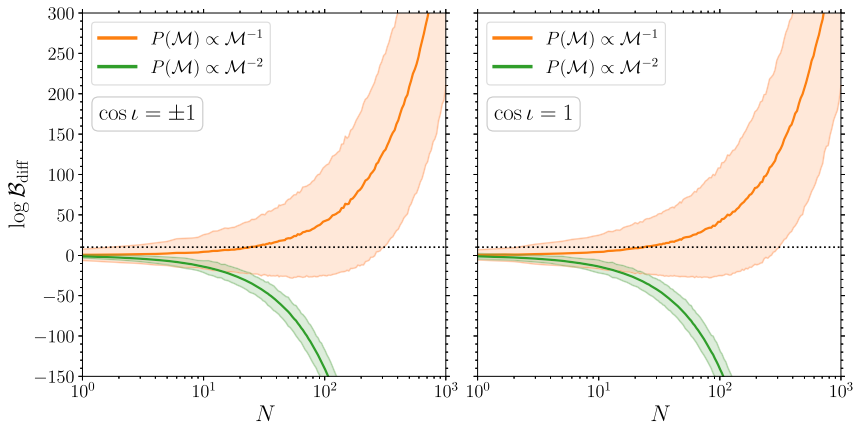


Figure 4. The log Bayes' factor, $\log\mathcal{B}_{\rm diff}$, in favor of a deviation from GR under the assumption that the deviation parameter takes independent values in all events

The solid lines indicate the mean value of $\log\boldsymbol{\mathcal{B}}_{\rm diff}$ obtained from 10^4 realisations of the simulated catalog while the shaded region between the two paler lines indicates the $\pm 1\sigma$ spread in this set of simulated catalogs. The dashed horizontal line denotes the threshold $\log\boldsymbol{\mathcal{B}}_{\rm same} = 10$. Left panel: the model errors are equally likely to favor positive or negative α ($\cos\iota = \pm 1$ randomly in all events). Right panel: the model errors always favor positive α ($\cos\iota = 1$ in all events). In both cases we see the evidence for a deviation from GR grows rapidly with the size of the catalog if our waveform models are bad (i.e. $P(\boldsymbol{\mathcal{M}}) \propto \boldsymbol{\mathcal{M}}^{-1}$) but decreases rapidly if our models are good (i.e. $P(\boldsymbol{\mathcal{M}}) \propto \boldsymbol{\mathcal{M}}^{-2}$). In the worst case, misleading evidence for a deviation from GR can appear with catalogs containing as few as ≈ 30 events above the minimum SNR of $\rho > 20$.

From the results in Figures 3 and 4, in four of the eight "realistic" scenarios considered here the misleading evidence in favor of a deviation from GR due to the modeling errors accumulates rapidly with increasing catalog size. This occurs even if the waveform model is good enough to safely analyze each event in the catalog individually. These results highlight the potentially insidious effects of waveform systematics when performing testing of GR with catalogs of GW events.

DISCUSSION

Developing waveform models is a challenging task that inevitably involves some approximations, simplifications, and modeling errors. These include truncating post-Newtonian series at some high order, neglecting certain physical effects (e.g. tidal terms, subdominant spin effects and orbital eccentricity) and the finite accuracy in numerical-relativity simulations. If the resulting models are interpreted at a face value, these systematic offsets can mimic the effect of new physics beyond GR.

This is a rather generic effect that has long been known about at the level of individual events. In this paper, we show how this extends to the case when a catalog of events is analyzed for signs of a deviation from GR. Using a simple, linearized analysis we have studied whether and how fast the modeling errors accumulate and have shown that it depends on:

1. the alignment of the model errors with the particular deviation from GR under consideration (i.e. does the modeling error always tend push α in one direction, or does it vary across parameter space and tend to average out across many different events);
2. how the catalog events are combined to give a test of GR (i.e. whether the deviation is assumed to take the same value in each event [multiplying likelihoods], independent values [multiplying Bayes' factors], or some intermediate case);
3. the distribution of waveform modeling errors (i.e. mismatches $\mathcal{M}$) across catalog events, which in turn depends on the waveform models used and the location of new events in parameter space.

Furthermore, our idealized calculation shows that this is a rather urgent problem. Erroneous evidence for new physics from waveform systematics *might* occur with as few as $10-30$ events at SNR $\gtrsim 20$. Although this is a conservative estimate and reflects the worst-case scenario, it is dangerously close to the size of current catalogs.

Going forward, our Fisher-like analysis needs to be backed up by injection and recovery campaigns. This will address more realistically the details of how current waveform models perform when used for a selection of parameterized tests of GR on catalogs of various sizes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHODS DETAILS

ACKNOWLEDGMENTS

We thank Antoine Klein, Geraint Pratten, Elinore Roebber, Patricia Schmidt, Lucy Thomas, and Alberto Vecchio for discussions. D.G. is supported by European Union's H2020 ERC Starting Grant No. 945155–GWmining, Leverhulme Trust Grant No. RPG-2019-350, and Royal Society Grant No. RGS-R2-202004. Computational work was performed on the University of Birmingham BlueBEAR cluster.

AUTHOR CONTRIBUTIONS

All authors contributed equally to writing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 30, 2021
Revised: April 23, 2021
Accepted: May 17, 2021
Published: June 16, 2021

REFERENCES

Berti, E., Barausse, E., Cardoso, V., Gualtieri, L., Pani, P., Sperhake, U., Stein, L.C., Wex, N., Yagi, K., Baker, T., et al. (2015). Testing general relativity with present and future astrophysical observations. CQG *32*, 243001. arXiv:1501.07274.

Blackman, J., Field, S.E., Scheel, M.A., Galley, C.R., Hemberger, D.A., Schmidt, P., and Smith, R. (2017). A Surrogate model of gravitational waveforms from numerical relativity simulations of precessing binary black hole mergers. PRD *95*, 104023. arXiv:1701.00550.

Chen, H.Y., and Holz, D.E. (2014). The loudest gravitational wave events. arXiv, 1409.0522.

Chua, A.J.K., and Vallisneri, M. (2020). On parametric tests of relativity with false degrees of freedom. arXiv, 2006.08918.

Cutler, C., and Vallisneri, M. (2007). LISA detections of massive black hole inspirals: parameter extraction errors due to inaccurate template waveforms. PRD *76*, 104018. arXiv:0707.2982.

Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann. Math. Stat. 204–223.

Gair, J.R., and Moore, C.J. (2015). Quantifying and mitigating bias in inference on gravitational wave source populations. PRD *91*, 124062. arXiv:1504.02767.

Isi, M., Chatziioannou, K., and Farr, W.M. (2019). Hierarchical test of general relativity with gravitational waves. PRL *123*, 121101. arXiv:1904.08011.

LIGO and Virgo Collaborations (2018). Prospects for observing and localizing gravitational-wave transients with Advanced LIGO, Advanced Virgo and KAGRA. LRR *21*, 3. arXiv:1304.0670.

LIGO and Virgo Collaborations (2019). Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1. PRD *100*, 104036. arXiv:1903.04467.

LIGO and Virgo Collaborations (2020a). Population properties of compact objects from the second LIGO-virgo gravitational-wave transient catalog. arXiv, 2010.14533.

LIGO and Virgo Collaborations (2020b). Tests of general relativity with binary black holes from the second LIGO-virgo gravitational-wave transient catalog. arXiv, 2010.14529.

LIGO Collaboration (2015). Advanced LIGO. CQG *32*, 074001. arXiv:1411.4547.

Lindblom, L., Owen, B.J., and Brown, D.A. (2008). Model waveform accuracy standards for gravitational wave data analysis. PRD *78*, 124020. arXiv:0809.3844.

Moore, C.J., Cole, R.H., and Berry, C.P.L. (2015). Gravitational-wave sensitivity curves. CQG *32*, 015014. arXiv:1408.0740.

Pürrer, M., and Haster, C.J. (2020). Gravitational waveform accuracy requirements for future ground-based detectors. PRR *2*, 023151. arXiv:1912.10055.

Schutz, B.F. (2011). Networks of gravitational wave detectors and three figures of merit. CQG *28*, 125023. arXiv:1102.5421.

Sivia, D., and Skilling, J. (2006). Data Analysis: A Bayesian Tutorial (Oxford University Press).

Thorne, K.S. (1987). Gravitational radiation. In Three Hundred Years of Gravitation, S.W. Hawking and W. Israel, eds. (Cambridge University Press), pp. 330–458.

Vallisneri, M., and Yunes, N. (2013). Stealth bias in gravitational-wave parameter estimation. PRD *87*, 102002. arXiv:1301.2627.

Virgo Collaboration (2015). Advanced Virgo: a second-generation interferometric gravitational wave detector. CQG *32*, 024001. arXiv:1408.3978.

Will, C.M. (2014). The confrontation between general relativity and experiment. LRR *17*, 4. arXiv:1403.7377.

Yunes, N., and Siemens, X. (2013). Gravitational-wave tests of general relativity with ground-based detectors and pulsar-timing arrays. LRR *16*, 9. arXiv:1304.3473.

Yunes, N., Yagi, K., and Pretorius, F. (2016). Theoretical physics implications of the binary black-hole mergers GW150914 and GW151226. PRD *94*, 084002. arXiv:1603.08955.

Zimmerman, A., Haster, C.J., and Chatziioannou, K. (2019). On combining information from multiple gravitational wave sources. PRD *99*, 124044. arXiv:1903.11008.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software		
Python3	https://www.python.org	
Numpy	https://numpy.org	version 1.20.0
Scipy	https://www.scipy.org/index.html	version 1.6.3

RESOURCE AVAILABILITY

Lead contact

Christopher J Moore (moorecj@bham.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This study did not generate any unique dataset.

METHODS DETAILS

All results contained in this paper were all obtained from the Monte-Carlo (MC) simulations described in the body of the paper. These MC calculations involve drawing, from the distributions described, the following random numbers for each GW event in the simulated mock catalogs: the signal-to-noise ratio, mismatch, and model error alignment angle (cosine iota). All calculations were performed using Python and the standard libraries referenced below. All information necessary to repeat these calculations is clearly described in the main body of the paper.