Gravitational wave tests of strong field general relativity with binary inspirals: Realistic injections and optimal model selection

Laura Sampson, Neil Cornish, and Nicolás Yunes

Phys. Rev. D **87**, 102001 — Published 17 May 2013

DOI: 10.1103/PhysRevD.87.102001

Gravitational Wave Tests of Strong Field General Relativity with Binary Inspirals: Realistic Injections and Optimal Model Selection

Laura Sampson, Neil Cornish, and Nicolás Yunes
Department of Physics, Montana State University, Bozeman, MT 59717, USA.

We study generic tests of strong-field General Relativity using gravitational waves emitted during the inspiral of compact binaries. Previous studies have considered simple extensions to the standard post-Newtonian waveforms that differ by a single term in the phase. Here we improve on these studies by (i) increasing the realism of injections and (ii) determining the optimal waveform families for detecting and characterizing such signals. We construct waveforms that deviate from those in General Relativity through a series of post-Newtonian terms, and find that these higher-order terms can affect our ability to test General Relativity, in some cases by making it easier to detect a deviation, and in some cases by making it more difficult. We find that simple single-phase post-Einsteinian waveforms are sufficient for detecting deviations from General Relativity, and there is little to be gained from using more complicated models with multiple phase terms. The results found here will help guide future attempts to test General Relativity with advanced ground-based detectors.

PACS numbers: 04.80.Cc,04.80.Nn,04.30.-w,04.50.Kd

I. INTRODUCTION

Einstein's General theory of Relativity (GR) has weathered an array of increasingly stringent tests since the theory first gained prominence in November 1919, when reports of Eddington's expedition appeared in newspapers around the world: "Revolution in science – New theory of the Universe – Newtonian ideas overthrown". Subsequent observations have continued to strengthen the case for Einstein's theory, though observations have yet to probe the dynamical, non-linear regime where the most revolutionary aspects of the theory take hold. For example, GR has passed all Solar System tests with flying colors, but these are based on stationary, weak, and linear gravitational fields, where characteristic velocities are small relative to the speed of light [1]. The theory has also passed all binary pulsar tests, but these systems have gravitational fields that are quasi-stationary and only moderately-strong, with characteristic velocities of $\sim 0.1\%$ the speed of light [1, 2]. In the near future, gravitational wave (GW) observations will test GR in a regime that has so-far evaded observation [3]: the *strong-field*, where the gravitational field is of order unity and velocities approach the speed of light.

Compact binary coalescences, the slow inspiral and merger of black holes (BHs) and/or neutron stars (NSs), will be strong sources of GWs, and these will be excellent tools for testing GR. During the inspiral phase, the binary components have orbital velocities ranging from 1% to $\sim 50\%$ the speed of light, which leads to strong and dynamically evolving gravitational fields. These GW signals evolve through thousands of radians of phase in the most sensitive band of ground-based detectors, such as aLIGO [4] and aVIRGO [5], with signal-to-noise ratios (SNRs) that will allow us to extract signal parameters with good accuracy. Thus, even small differences in the dynamics of the gravitational theory can lead to large accumulated effects in the waveform during the inspiral.

Despite their promise, GW tests of GR are, unfortunately, very difficult to carry out, for two main reasons. One reason is purely theoretical - we currently lack candidate alternative theories that are particularly appealing. Instead, we have many models that are either heavily constrained, like scalar-tensor theories [1], or that have theoretical issues, such as knowledge only of their effective, low-curvature form [6]. The other cause of difficulty lies in the data analysis. Most techniques for detecting and characterizing GW observations require accurate templates to identify weak signals buried in the instrument noise. Given the already large parameter dimensionality of the GR waveform models, and the wide variety of modified gravity theories [6–22], the construction of individual template banks for all possible non-GR models is simply not feasible.

A much more appealing alternative is to devise a generic non-GR template family with which to model the signals, and allow the data to select the appropriate model via Bayesian inference. The first such model was proposed by Arun *et al* [23–25], where the coefficients in the post-Newtonian (PN) expansion of the phase were independently fitted for. However, the structure of the PN series does not allow for all known modified gravity deviations, including potentially interesting ones such as the emission of dipolar radiation predicted in scalar-tensor theories. For this reason, Yunes and Pretorius [26] developed the so-called parameterized post-Einsteinian (ppE) framework, which allows for a wide range of deformations to the amplitude and phase of the waveform. In the inspiral phase, these can be represented through a polynomial in the GW frequency, with free constants, or ppE parameters, that represent the amplitude and the frequency exponent of the deformations [26]. The simplest ppE inspiral waveform in the Fourier domain has the form

$$\tilde{h}^{\rm ppE} = \tilde{h}^{\rm GR}\,(1+\alpha u^a)\,e^{i\beta u^b}\,, \qquad (1)$$

where $u = (\pi \mathcal{M} f)^{1/3}$ is a dimensionless velocity, $\mathcal{M} = \eta^{2/5} M$ is the chirp mass, $\eta = m_1 m_2/(m_1+m_2)^2$ is the symmetric mass ratio, $m_{1,2}$ are the component masses, and f the GW frequency. The Fourier transform of the GR waveform is here $\tilde{h}^{\rm GR}$, while (α, a, β, b) are ppE parameters. Clearly, in the limit $(\alpha, \beta) = (0, 0)$, one recovers the GR prediction, while for other values of the ppE parameters one recovers the leading-order waveforms of all known modified gravity theories.

The first data analysis implementation of the ppE framework was carried out by Cornish, *et al* [27], where ppE waveforms of the form of Eq. (1) were used both in the generation of the simulated signals, and in the extraction of the model parameters in a Bayesian model selection framework. This study was a proof-of-principle that the ppE framework can be successfully implemented to carry out tests of GR. A second study shortly followed [28] that confirmed the results of Cornish, *et al* and extended them to include lower SNR signals and multiple detections. While this study also used the simple one-phase ppE model of (1) for the signal injections, the models used to analyze the simulated data included more complicated ppE waveform models with multiple phase corrections.

In this paper we revisit the ppE framework and carry out a more realistic data analysis study. First, we examine the effect of more realistic non-GR injections that include modifications to several terms in the PN GR phase, instead of a single one. Generic deviations from GR will be characterized by an infinite number of phase corrections. Ground-based detectors will not be sensitive to all of them, just as they are not sensitive to GR signals to arbitrarily high PN order. The presence of the first few higher-order terms can affect our ability to test GR. We find that the presence of multiple phase modifications will improve our chances of detecting departures from GR if they are of the same sign. However, if the phase modifications are of alternating sign, they can cancel out to some degree, and make a non-GR signal appear to be well described by GR.

As something of an aside, we consider the issue of adding explicit noise realizations to the simulated signals, especially for low SNR signals. This is done because some concerns have been voiced about the conclusion of the Cornish *et al* [27] work due to the relatively high SNR of the signals used, and their technique of accounting for the noise solely through the weighting of the likelihood function. We analytically and numerically show that the conclusions of [27] remain unaffected when adding an explicit noise realization. We also show that these results scale linearly with SNR down to values close to the detection threshold, which for this source was SNR ~ 7.5.

We then tackle the problem of determining the *optimal* ppE model for detecting departures from GR. On the one hand, including additional phase terms will improve the fit and increase the likelihood. On the other hand, adding additional parameters to the model incurs an "Occam penalty". We find, on balance, that in almost all cases, templates with only one ppE parameter are preferred over those with multiple parameters. These suggests that the simple one-parameter ppE model may well be the ideal one to search for GR deviations in early data from advanced detectors.

The remainder of this paper is organized as follows. Section II builds non-GR injections and studies their effect on signal extraction and the detection of departures from GR. Section III considers the effects of adding explicit noise realizations to the signals, and how the strength of the signal affects our ability to test GR. Section IV studies different ppE waveform models to determine the optimal one for performing GW tests of GR. Section V concludes and points to future directions for research. Throughout this paper we use geometric units with $G = c = 1$.

II. REALISTIC SIGNAL INJECTIONS

The simplest ppE waveform family presented in the introduction is not sufficiently complex to represent a realistic alternative gravity theory. This is because modified gravity theories will differ from GR by an infinite series of terms in both the amplitude and the phase. We expect that an alternative theory of gravity will give rise to waveforms where the amplitude and phase depend on one or more fundamental coupling constants multiplied by functions of the system parameters. Thus, if one wishes to use a ppE-type template to inject non-GR signals, one must consider more complex ppE models, such as Eq. (46) in [26], namely Eq. (1) with the replacements [26]

$$\alpha u^a \to \sum_{i=0}^{N} \alpha_i u^{a_i}\,, \quad \beta u^b \to \sum_{i=0}^{N} \beta_i u^{b_i}\,, \tag{2}$$

where the α, β's depend on a universal coupling constant κ, and functions of the system parameters $\vec{\lambda}$:

$$\begin{aligned} \alpha_i(\kappa, \vec{\lambda}) &= \kappa \sum \phi_i(\vec{\lambda}) \\ \beta_i(\kappa, \vec{\lambda}) &= \kappa \sum \theta_i(\vec{\lambda})\,. \end{aligned} \tag{3}$$

The functions $\phi_i(\vec{\lambda}), \theta_i(\vec{\lambda})$ can be computed for specific theories, but their general form is unknown. So while κ takes a single value for a particular theory, the (α_i, β_i) constants will vary from detection to detection depending on the masses, spins and other parameters that describe the system. In some theories there will be more than one additional coupling constant κ, but here we will assume that one sector of the modified theory dominates and consider only a single series of correction terms. With a large number of high SNR detections, it may be possible to infer the functional form of $(\phi_i(\vec{\lambda}), \theta_i(\vec{\lambda}))$. However, since our immediate concern is in deciding if the data is consistent with the prediction of GR, we will argue that

it is best to use a much simpler waveform for the initial tests.

The ppE exponents (a_i, b_i) are real numbers that give the effective PN order at which the non-GR modification enters the signal, while the ppE amplitude parameters (α_i, β_i) are real numbers that indicate the strength of the modification, in turn controlled by the overall coupling strength κ. In principle, we could extend the sum in Eq. (2) to infinity, but in practice, realistic detectors are sensitive only to a finite number of terms in the phase and amplitude. The injected signals then consist of a GR waveform with its amplitude and phase modified by a series of ppE corrections.

Several simplifications can be made to the general waveform presented above. First, for quasi-circular inspiral signals, Chatziioannou, *et al* [29] have argued that analyticity demands that the exponents (a_i, b_i) take on integer values with possible logarithmic corrections (just as the PN expansion in GR comes in integer powers of u and products of integer powers of u with $\log u$, where recall here that u is related to the orbital velocity). Second, ground-based advanced detectors will be of limited sensitivity, rarely being sensitive to more than the first three terms in the PN expansion, and usually being much more sensitive to the phase evolution than they are to the amplitude evolution. Thus, we choose to simplify the analyses by truncating the sum at three terms and setting $\alpha_i = 0$. The injections are then given by Eq. (1) but with the replacement

$$\beta u^b \to \sum_{i=0}^{2} \beta_i u^{b+i} = \beta_b u^b + \beta_{b+1} u^{b+1} + \beta_{b+2} u^{b+2}\,, \quad (4)$$

and $\alpha = 0$, where in the last equality the Einstein summation convention is not assumed. Written in this way, β_b is always proportional to u^b for any b. Previous investigations have been restricted to signals with only one ppE correction injected, which reduces to Eq. (4) when one retains only the first term in the sum. As argued above, this is far from realistic for a modified gravity injection and we will show that the higher-order terms can have a significant effect on the analysis.

Ultimately the claim that a detection is in agreement (or conflict) with GR comes down to model selection. Does GR describe the data best or does another model do a better job? In Bayesian statistics [27, 30, 31], model selection is performed via the calculation of the Bayes factor, which is simply the "betting odds" of one model against another. For instance, if the Bayes factor between GR and a non-GR model is 100, and you originally gave both possibilities equal odds, then there is a 100:1 odds ratio that GR better describes the data than the other model. In this case, you would be well-advised to put your money on GR. There is no prescription for deciding what Bayes factor is required before we should consider one model "right" and another "wrong". However, in the case of a well-tested theory like GR being brought into question by, for instance, a GW signal, it is likely that the scientific community would require a detection that gives us a fairly high Bayes factor in favor of the non-GR model to overcome the prior belief in GR being the correct theory. In order to determine whether more ppE terms in an injection affect the detectability of a deviation from GR, we need to see how these different types of injections affect the Bayes factor.

Throughout this paper, Bayes factors are calculated using the Savage-Dicke density ratio [30, 32] and/or Reversible Jump Markov Chain Monte Carlo (RJMCMC) [30, 33]. In the Savage-Dicke method, the Bayes factor between two nested models, i.e. model X and model Y that differ only by the addition of a parameter to model Y, is calculated by comparing the weight of the marginalized posterior to the weight of the prior distribution for the "extra" parameter at the value that this parameter takes on for the lower-dimensional model:

$$B_{XY} \approx \frac{p(\kappa = 0|s)}{p(\kappa = 0)} \quad (5)$$

In our case, the extra parameter is the coupling strength κ, which has the GR limit $\kappa = 0$, and hence $\beta_i = 0$. To calculate the Bayes factors this way, we run a MCMC search using ppE templates in order to generate the posterior distribution for β_i, and then calculate the posterior weight in this distribution at $\beta_i = 0$. We then compare this posterior weight to the prior density at this point. We here use a flat prior distribution between -5.0 and 5.0 for all β values. The main advantage to this method over other possibilities is that it only requires exploration of the higher-dimensional space.

All tests in this section use GWs emitted by a NS-NS binary with $\approx 1.4 M_\odot$ component masses in the inspiral phase with SNR ~ 12. We model all waveforms with a quadrupolar, adiabatically quasi-circular waveform, with a 3.5PN-accurate phasing, but neglecting PN amplitude correction and spin effects, and truncating all evolution at the Schwarzschild test-particle innermost stable circular orbit. The waveforms are then described by nine source-parameters: the chirp and the reduced mass; the time and phase of coalescence; two sky-position angles; the inclination angle and the GW polarization angle; and the luminosity distance (see [27] for a similar waveform prescription). In addition to these we have the ppE phase parameters of Eq. (4). We consider a three detector network of second-generation detectors, such as aLIGO at Hanford, aLIGO at Livingston, and aVirgo, with identical broadband-configuration spectral densities, as in our previous paper [27], assuming the noise to be Gaussian and stationary. Table I shows the system parameters for all systems studied in this paper (masses are listed in solar masses, and luminosity distances are in megaParsecs).

In this paper, we examine two factors that influence the outcome - the signs of the different phase corrections, and their relative magnitude. We begin by exploring the effects of injecting phase corrections with the same or differing signs. In particular, let us study the effect that this relative sign has on the detectability of a non-GR

behavior. We will then explore the difference between non-GR phase corrections that either shrink in magnitude at higher PN order, stay at approximately the same magnitude, or grow in magnitude at higher PN order.

We begin by examine how the relative sign of the phase corrections affects the detectability of departures from GR. To do this, we consider three non-GR injections:

- **Case i.** A ppE waveform with a single ppE phase term ($b = -3$), with magnitude controlled by β_{-3}.
- **Case ii.** A ppE waveform with two ppE phase terms ($b = -3$ and $b = -2$), with β_{-3} and β_{-2} of the same sign.
- **Case iii.** A ppE waveform with two ppE phase terms ($b = -3$ and $b = -2$), with β_{-3} and β_{-2} of different sign.

We choose these values of b because, for $b < -5$, β_b is already well-constrained by binary pulsar observations, as demonstrated in [27, 34]. The $b = -3$ terms correspond to non-GR corrections at the first post-Newtonian (1PN) level, and the $b = -2$ terms correspond to a 1.5PN correction. Case (i) is the type of injection that has been explored in previous work. Cases (ii) and (iii) include higher-order phase corrections, but differ in their relative sign.

Figure 1 shows the Bayes factors between GR and a one-parameter ppE template family with $b = -3$ and ppE parameter β_{-3} for the three injections discussed above. The error bars in this figure are estimated by calculating the Bayes factors using multiple MCMC runs with different random seeds. The spread in the calculated values are reflected in the error bars. Observe that when the injection contains ppE corrections of the same sign (dotted, magenta curve), these add up to make the signal more discernible from GR. In this case, the Bayes factor becomes larger than 10, i.e. crosses our threshold for detectability, for the smallest value of β_{-3}. Therefore, if (β_b, β_{b+1}) share the same sign, we can detect deviations from GR with lower strengths than if there were only one phase correction. On the other hand, observe how when the non-GR signal contains alternating sign GR modifications (dashed blue line), these have the effect of partially canceling the non-GR effect out. In this case, the Bayes factor crosses 10 for a much larger value of β_{-3}. Therefore, if the corrections have alternating signs, e.g. if (β_b, β_{b+1}) have different signs, then our ability to detect departures from GR is reduced. The sign of the ppE amplitude exponent also affects the PDFs of the recovered β_i parameters, as we will see below.

The relative magnitudes of the terms also affects the analysis. Concentrating on the multi-term ppE models of Eq. (4), we define three cases, depending on the relative magnitude of these exponents in the series expansion:

- **Sub-Critical Case**: Injections where the ppE terms get smaller as the PN order increases, i.e. $\beta_b > \beta_{b+1} > \beta_{b+2}$.

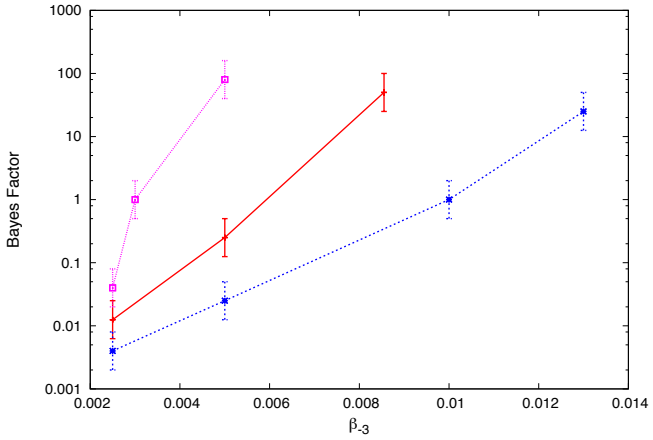

FIG. 1: (Color Online) Bayes factors between a GR model and a one-parameter ppE model for three different ppE signal injections. The dotted (magenta) line corresponds to an injection with the two positive ppE terms $\beta_{-3} > 0$ and $\beta_{-2} > 0$ (case ii), the solid (red) line corresponds to the single, positive ppE term $\beta_{-3} > 0$ (case i), and the dashed (blue) line corresponds to the two ppE terms of alternating sign $\beta_{-3} > 0$ and $\beta_{-2} < 0$ (case iii). System parameters for the systems studied here are listed in Table I. As expected, the signal with ppE terms of alternating sign is harder to distinguish from GR, as evidenced by its Bayes factor growing the slowest with the magnitude of β_{-3}.

- **Critical Case**: Injections where the ppE terms remain of about the same size as the PN order increases, i.e. $\beta_b \sim \beta_{b+1} \sim \beta_{b+2}$.
- **Super-Critical Case**: Injections where the ppE terms get bigger as the PN order increases, i.e. $\beta_b < \beta_{b+1} < \beta_{b+2}$.

Obviously, there are an infinite number of ways to choose how large the β_i constants are relative to each other, but the classification defined above provides a useful summary. More concretely, we here define *sub-critical*cases as those where the ppE terms injected have $\beta_{n+1} < (u_{\rm max})^{-b_n}$, where $u_{\rm max} = \pi \mathcal{M} f_{\rm max}$. Similarly, *critical* cases are defined such that $\beta_{n+1} \approx (u_{\rm max})^{-b_n}$, while *super-critical* cases have $\beta_{n+1} > (u_{\rm max})^{-b_n}$.

An alternative and roughly equivalent way to define these three different cases is by the number of *useful cycles* of phase [35] that accumulate during the signal for each correction to the phase. The number of useful cycles is defined via

$$N_{useful} \equiv \left(\int_{F_{\rm min}}^{F_{\rm max}} \frac{df}{f} \frac{a^2(f)}{S_n(f)} \frac{d\phi}{2\pi df} \right) \left(\int_{F_{\rm min}}^{F_{\rm max}} \frac{df}{f} \frac{a^2(f)}{f S_n(f)} \right)^{-1} \quad (6)$$

where $|\tilde{h}(f)|^2 = N(f) a^2(f)/f^2$ is the squared modulus of the frequency domain GW signal, and $N(f) = (f/2\pi)(d\phi/df)$. This quantity tells us about the phase ac-

Signal	α	ϕ_L	ϕ_c	$m_1(M_\odot)$	$m_2(M_\odot)$	$\log(D_L)$(Mpc)	t_c	δ	θ_L	β_{-3}	β_{-2}	β_{-1}
One ppE Term	1.0	4.76	1.9	1.62	1.73	3.96	5.58	0.77	-0.43	0.01	0.0	0
Alternating Sign	1.0	4.76	1.9	1.62	1.73	3.96	5.58	0.77	-0.43	0.01	-0.04	0
Same Sign	1.0	4.76	1.9	1.62	1.73	3.96	5.58	0.77	-0.43	0.01	0.04	0
Sub-Critical	1.0	4.76	1.9	1.62	1.73	3.96	5.58	0.77	-0.43	0.01	0.005	0
Critical	1.0	4.76	1.9	1.62	1.73	3.96	5.58	0.77	-0.43	0.01	0.08	0
Super-Critical	1.0	4.76	1.9	1.62	1.73	3.96	5.58	0.77	-0.43	0.01	0.25	0
GR Source	3.95	4.14	0.68	1.45	1.43	0.9	3.41	-0.66	0.76	0	0	0

TABLE I: Source parameters for sources used in Fig. 1 (top), Fig. 2 (middle) and Figs. 3, 4, and 6 (bottom).

cumulated from each PN (or ppE) term during the course of the signal, weighted by the sensitivity of the detector to different parts of frequency space. Tables of the number of useful cycles of phase for each system analyzed in this paper are included in this section. "Sub-Critical" signals are those for which the number of useful cycles due to the non-GR phase corrections decreases at higher order. "Critical" signals have roughly the same number of useful cycles at each order. "Super-critical" signals have larger numbers of useful cycles from the non-GR phase at higher orders.

Signal	ϕ_{-3}	ϕ_{-2}	ϕ_{-1}
Convergant	0.312	0.012	0
Critical	0.312	0.194	0
Super-Critical	0.312	0.607	0

TABLE II: Number of useful cycles from the different injected ppE terms - Fig 1 and Fig 2.

Figure 2 shows the PDFs of the recovered β_{-3} parameter for a one ppE parameter template family, with injections given by sub-critical, critical and super-critical versions of cases (ii) and (iii). These PDFs are computed using a MCMC approach. The top panel of this figure shows the PDFs for β_{-3} given an super-critical injection, the middle panel given a critical injection, and the bottom panel given a sub-critical injection. The left and right panels correspond to injections with the same (left) or alternating (right) signs. When there is as much or more weight at $\beta_{-3} = 0$ in the PDF's as there was in the prior probability density, this indicates that GR is the preferred model. In our case, the prior probability for β_{-3} is flat between -5.0 and 5.0, and so the prior probability density at all points, including $\beta_{-3} = 0$, is 0.1. When the posterior density at $\beta_{-3} = 0$ is less than 0.1, an alternative model is preferred.

Figure 2 reveals several interesting facts. First, observe that in the sub- and super-critical injection cases, the sign of the βs is irrelevant: in both cases most of the weight is outside $\beta_{-3} = 0$. Second, observe that in the sub-critical case, the second ppE term ($b = -2$) is very sub-dominant to the first term, and so its sign has little impact on the results. Third, observe that in the critical injection case, when the βs have alternating signs, the modified gravity effects partially cancel out, yielding a β_{-3} PDF with non-negligible weight at the GR value. It is clear from these studies that neglecting higher-order phase corrections can seriously bias our assessment of our ability to test GR with GW signals. For the "Critical" case, our ability to detect departures from GR is enhanced if the terms have the same sign, and diminished if the signs alternate.

III. NOISE MODELING AND SIGNAL STRENGTH

Most of our studies have been conducted on signals that do not have a noise realization explicitly added to the signal injection, although all analyses incorporate the noise spectrum of the detectors in the likelihood calculation. We chose not to include an explicit noise realization in order to expedite the calculation of the likelihood [36], which then allows us to produce long Markov chains that fully explore the high dimensional parameter spaces. Unfortunately, our use of this technique has led some to question the reliability of our results [28, 37]. Here we show that those concerns are unfounded.

The inclusion of noise in our signals has little effect on the conclusions we drew in our previous paper, as can be seen in Fig. 3. In this figure, we plot the (3σ)-bounds that we could place on the ppE phase parameters, if one has detected a NS-NS inspiral with SNR 15 that has no GR deviation. To calculate these bounds, we inject a GR signal and try to recover it using a single parameter ppE template, ie. Eq. (4) with a single β. For any given value of b, we integrate out over all other parameters and take the standard deviation of the β PDF as a 1σ bound. In other words, the curves show the upper limit of the magnitude a ppE parameter could be found to have, and still have the signal be consistent with GR. This plot shows that the bounds placed on the ppE parameters from a signal that includes an explicit noise realization are consistent with those found when no noise is added to the signal. That is, including an explicit noise realization does not affect the conclusions derived from a cheap-bound calculation with noise accounted for only through the detectors' noise spectrum in the likelihood.

To understand this result, it is useful to look at Figure 4, which shows the recovered PDF's for the β parameter from three different runs, each including noise

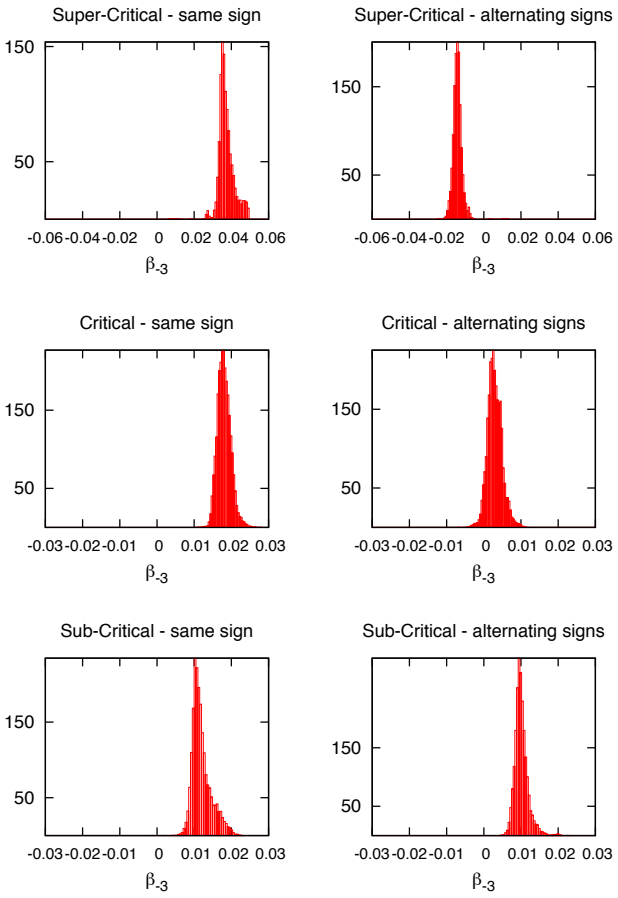

FIG. 2: The PDF's for β_{-3} in a one-parameter ppE template recovered from MCMC searches on injections containing two ppE parameters ($b = -3$ and $b = -2$). In all injections, $\beta_{-3} = 0.01$, but the value of β_{-2} varies between cases. The plots on the left are for injections containing two ppE parameters of the same sign, and on the right of opposite signs. The more weight in the PDF at $\beta = 0$, the lower the Bayes factor in favor of a non-GR signal. In the critical case, we find that alternating signs in the phase corrections can cause a non-GR signal to be indistinguishable from a GR one. In the sub- and super-critical cases, this does not occur. System parameters for this figure are the same as in Figure 1, also listed in Table I, and the useful cycles of phase are in listed in Table II.

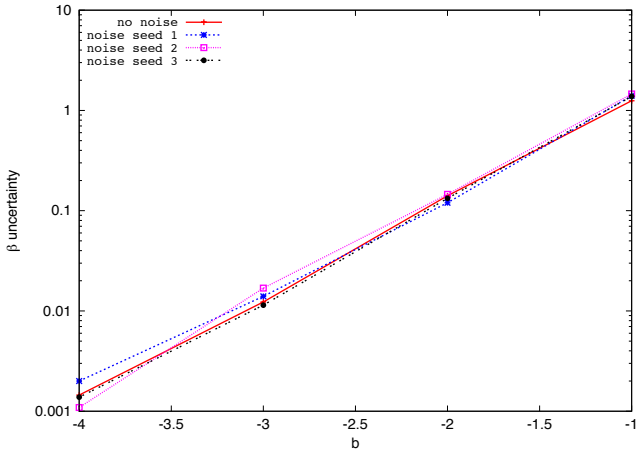

FIG. 3: (Color Online) (3σ)-bounds on β that can be inferred for different values of b, calculated from the PDF's of β generated by recovering a GR signal with a ppE template. This plot shows the bounds for both a signal with no noise, and three that include Gaussian noise, generated from three different random seeds. The results are essentially identical. The signal parameters for this injection are in Table I.

generated with a different random seed. Since the injected signal was a GR NS-NS inspiral waveform with SNR 15, we would expect the β PDF's to peak at zero. It is clear from this figure that, although the peak of the PDF is shifted by the inclusion of noise, the uncertainty in the recovery of this parameter, i.e. the spread of the distribution, is not affected. This concept has been explored before, in [38] and [37]. In [38], the authors argue that when discussing our ability to measure system parameters in general, and not for a particular case, what we really want to do is examine the *noise-averaged* uncertainties in these parameters. That is, we are interested in how well we can measure parameters when averaged over many specific realizations of the noise. The authors show that the noise-averaged uncertainties are the same as the uncertainties calculated with zero noise injected into the signal. In [37] it is argued that the specific noise realization will affect our parameter estimation, and while this is technically true, we have shown in this section that the overall effect is minimal. In any case, for the type of analysis that we want to do in the rest of this paper, the reasoning of [38] applies, and so we do not inject an explicit noise realization for any of our analyses in the other sections. It has also been claimed in [28] that simulated data that only includes a signal injection, ie. that does not include a noise realizations, will necessarily lead to posterior distributions for the system parameters that are Gaussian. This is patently false, as can easily be demonstrated by analytically calculating the posterior distribution for a signal of the form $(d_0/d)\cos(2\pi ft)$, which leads to a highly non-Gaussian distribution in the distance d.

Obviously, signals with high SNR will be better for testing GR, as they are better for any type of GW data analysis. When discussing how well GR can be tested using GW detections, the highest-SNR events are the ones that will lead to the strongest constraints. In our previous paper, we analyzed signals with SNR ~ 20, which would be considered a high SNR detection by the LIGO detectors. It is irrelevant, however, that most signals will probably have SNRs in the low 10s. There will always be one signal with highest SNR, and this is likely to be above 15. It is therefore still useful to study GR tests assuming detections with SNRs ~ 20, as it is not a hopeless propo-

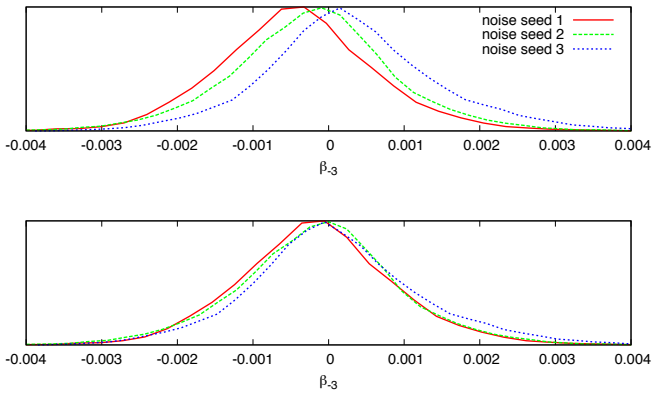

FIG. 4: (Color Online) The top panel shows posterior distributions of β recovered from three ppE injections, including noise in the injection. Each of the three signals was generated using a different random seed for the noise, but the same system parameters. The lower panel shows the same distributions, now with the best-fit value of β subtracted. This illustrates that, although noise affects the peak of the posterior distribution for a given parameter, it does not affect the uncertainty in that parameter. Thus the *cheap bounds* of [27] are unaffected by the inclusion of noise.

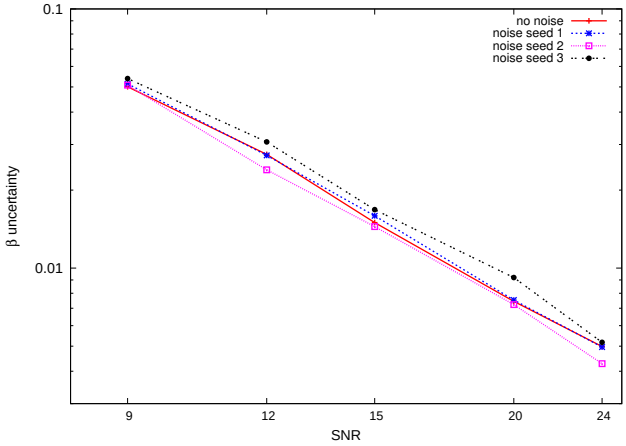

FIG. 5: (Color Online) (3σ)-bounds on β for $b = -1.0$, calculated from the PDF's of β generated by recovering a GR signal with a ppE template. This plot shows the linear relationship between the bounds on β and the SNR of the signal. There are four lines shown - one for a signal that had no noise injected, and three for signals that had noise injected, each with a different random seed. The results are essentially identical. The signal parameters for this injection are in Table I.

sition that we will have this type of event in our GW catalog. Throughout the rest of this paper, however, we have taken a more pessimistic tack, and restricted ourselves to analyzing signals with SNR $\sim 10 - 12$. The results follow the theoretical linear scaling with SNR [36] down to values of the SNR that are close to the detection threshold. This scaling is shown in Figure 5.

IV. OPTIMAL MODEL SELECTION

We have seen that it is important to consider multi-term ppE signal injections when assessing the bounds we will be able to place on alternative gravity theories. The question still remains, however, as to what type of templates we should use to recover such signals. In this section we address this question by showing first that adding too many parameters to the templates is counterproductive. Then we determine the optimal ppE template family to detect departures from GR described by the more realistic multi-term ppE signal injection model.

A. Overfitting

One may consider using a ppE template with many ppE phase and amplitude terms in the sums of Eq. (2). For example, one could include as many ppE phase terms as there are in the GR PN series, but this is far from ideal. The reason is clear: if we include the same number of free ppE parameters in our phase model as we have phase terms that are functions of system parameters, then there is no way to constrain any of them. In other words, the ppE phase terms will have a 100% correlation with the standard GR system parameters that form the coefficients of the GR PN phase.

As a simple example, consider the possibility of detecting a non-GR signal that includes ppE corrections at $b = -5$ (a so-called *Newtonian* ppE correction) and $b = -3$ (a 1PN ppE correction). We will truncate our injection at 1PN order for this example, which implies that the GW phase contains two standard PN terms that are functions of the system parameters, and two free ppE terms. Figure 6 shows that there is a 100% correlation between these PN and ppE parameters.

These types of correlations are commonly encountered in GW data analysis, but they may not be widely appreciated by theoretical model builders. We can understand this correlation analytically as follows. Let us write the simplified ppE template Fourier phase $\Psi_{\rm ppE}(f)$ as follows

$$\Psi_{\rm ppE}(f) = \left[\frac{3}{128}\left(\pi\mathcal{M}\right)^{-5/3} + \beta_{-5}\left(\pi\mathcal{M}\right)^{-5/3}\right] f^{-5/3} + \left[\frac{3}{128\eta^{2/5}\pi\mathcal{M}}\left(\frac{3715}{756}+\frac{55}{9}\eta\right) + \frac{\beta_{-3}}{\pi\mathcal{M}}\right] f^{-1}\,. \tag{7}$$

where we have expanded out the definition of u. Clearly, we can rescale β_{-5} by a constant and β_{-3} by a function of η, and then also adjust $\mathcal{M}$ and η, to recover the same value of the Fourier phase. This shows a direct correlation between these parameters. Figure 6 demonstrates how such a correlation manifests itself in the posterior

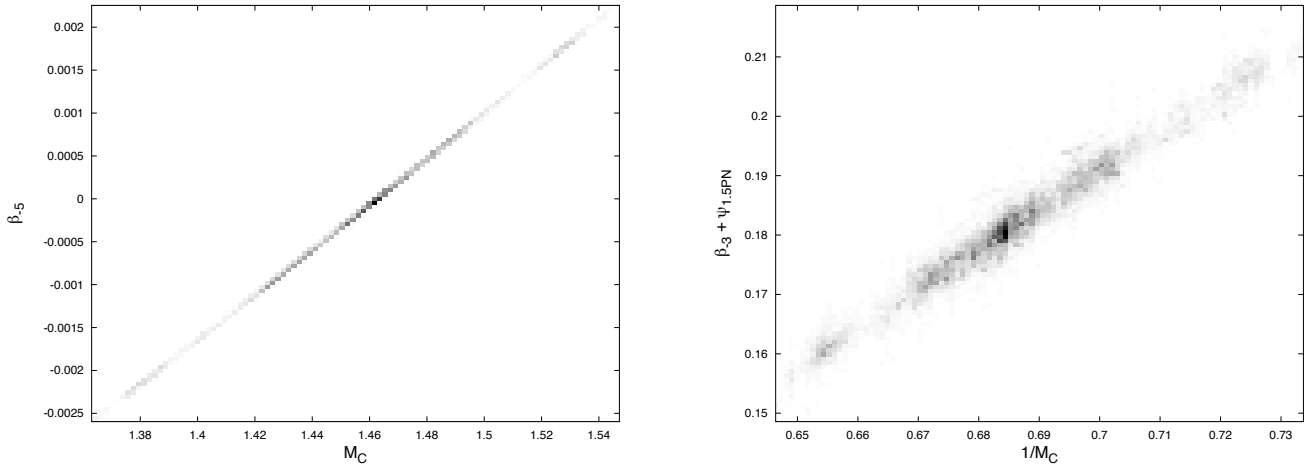

FIG. 6: Correlation between the β_{-5} ppE parameter and the chirp mass (left panel) and the β_{-3} parameter and the inverse chirp mass (right panel) for an injection including two PN phase terms as well as two ppE phase corrections. The parameters are restricted only by their prior ranges.

distributions.

This argument can be extended to whatever PN order we choose. If we include the same number of ppE terms as PN terms in our model, then we will not be able to place bounds on *any* parameter, let alone use the results as a test of GR. It is also true, however, that ppE models that include more ppE terms will be able to achieve a better overall fit of whatever signal we happen to detect, just as any model with extra parameters can typically fit data better than a simpler model. In the next section we explore the tradeoff between these two effects. We also attempt to determine what types of signals are best to analyze using more complex ppE models, and what types are better served with a simple ppE model.

B. Inclusion of Spin

There are many potential effects, both astrophysical and purely gravitational, that will make it more difficult to test GR. For instance, the presence of accretion disks [39, 40], the presence of a third companion [41], the unknown effects of the neutron star equations of state [42, 43], etc. To illustrate how these types of effects can hinder our ability to test the nature of gravity, in Figure 7 we have plotted the Bayes factors between a $b = -3$ ppE model and GR applied to critical ppE injections. In the top panel, we included the standard PN terms for aligned spins in the phase for both the GR and ppE waveform models, which introduces two new parameters. The correlation between these spin parameters and the β_{-3} ppE parameter causes the detection threshold for β_{-3} to be larger by a factor of ~ 20 compared to the case in which the spins are held fixed to zero.

The inclusion of spin effects when testing GR has been explored before in the context of particular theories of gravity. Using systems with aligned spins degrades the bounds due to correlation between the spin parameters and the alternative theory parameters [10]. Including spin precession effects [11] restores the bounds to levels closer to what is found for systems without spin [9], as recently explained in [3]. Using waveforms that include additional structure such as higher harmonics of the the orbital frequency can also improve the bounds on alternative theory parameters [44]. Thus, the situation shown in Fig. 7 should be considered a worst-case scenario. Throughout the rest of this paper, we hold spins fixed to zero. This means that our actual bounds on the ppE strength parameters are probably a little optimistic, but it does not change the conclusions we draw about model selection.

C. Parsimonious Fitting: Detecting and Characterizing non-GR signals

Let us now study what type of ppE templates are best suited for detecting a GR deviation. In particular, let us examine whether using one-term or two-term ppE templates works better. For this analysis, we inject ppE signals containing three phase terms, and attempt to recover them using one- and two-parameter ppE templates. We calculate Bayes factors between the ppE models against the GR model to see which model is best suited to detecting departures from GR. Because of our strong prior belief in the validity of GR, a Bayes factor significantly greater than unity would be necessary to convince us that a new theory of gravity is needed.

Let us then consider three different ppE injections, starting at 1 PN order ($b = -3, -2, -1$), a sub-critical, a critical and a super-critical one, each for a NS-NS inspiral, with parameters listed in Table III . We explore

Source	α	ϕ_L	ϕ_c	$m_1(M_\odot)$	$m_2(M_\odot)$	$\log(D_L)$(Mpc)	t_c	δ	θ_L	β_{-3}	β_{-2}	β_{-1}
Sub-Critical	1.42	2.5	0.8	1.42	1.73	3.83	3.5	0.87	0.43	0.003	0.003	0.003
Critical	1.42	2.5	0.8	1.52	1.33	3.9	3.5	0.87	0.43	0.0006	0.018	0.54
Super-Critical	1.42	2.5	0.8	2.04	1.34	3.86	3.5	0.87	0.43	0.0007	0.07	7.0

TABLE III: Source parameters for Figures 8 and 9. The β_b values listed are for a particular case - the ratio between different β_b values was kept constant for each injected signal. The ratio for sub-critical was $\times 1.0$, critical was $\times 30$, and super-critical was $\times 100$.

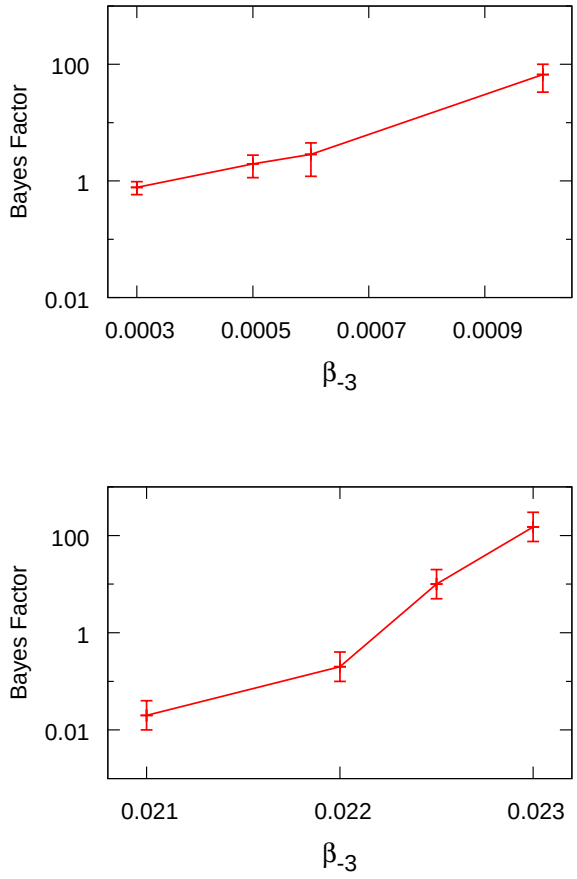

FIG. 7: Bayes factors of a $b = -3$ ppE model versus GR. The injected signals in both cases were un-spinning, critical ppE injections, with the value of β_{-3} plotted on the x axis. The upper panel shows Bayes factors for templates that include aligned spin parameters, and the lower panel is for templates with no spin parameters. The degeneracy between the 1.5PN spin term and the β_{-3} ppE amplitude parameter significantly weakens the bounds.

these simulated signals with a MCMC algorithm, using a one- and a two-term ppE model. The one-term ppE models are allowed to choose between phase exponents $b = -3$ and $b = -2$, while the two-term models are allowed to choose between the pairs $(-3, -2)$ and $(-2, -1)$ - *i.e.* the two terms must differ by a single power of u, and models with exponents $(-3, -1)$ are not allowed.

The Bayes factors between the one-term ppE model and GR (red solid curve) and between the two-term ppE model and GR (blue dashed curve) are shown in Fig. 8 as a function of the injected value of β_{-3} for a sub-critical (top-left panel), critical (top-right panel) and super-critical (bottom panel) injection. These Bayes Factors are again calculated using the Savage-Dicke density ratio. Calculating the posterior density at a $\beta_i = 0$ from a Markov chain involves counting the number of points in the chain that fall within the histogram bin containing $\beta_i = 0$, and so the error bars reflect the counting error involved in this process, as well as the spread in BF values calculated from multiple MCMC runs on the same signal but with different random seeds. Observe that the only injections for which two-term ppE templates consistently outperform one-term ppE templates are the critical ones. Even in this case, however, the preference is not large; the curves track each other very well in all cases. Therefore, our results indicate that the one-term ppE templates are sufficient for searching for deviation from GR in GW data.

Once a deviation from GR has been definitively detected, the next step is to learn as much about the signal as possible, in order to give theorists as much guidance as possible in their attempts to build an alternative theory of gravity. The information we could hope to extract from the type of analysis we have described in this paper is the structure of the series of phase corrections - do they enter at a certain PN power and then fade away? Or do they enter at that power and grow more important at higher orders in the expansion? Figure 9 plots the posterior distribution of the five models under consideration derived using a RJMCMC [30] analysis. In RJMCMC, moves are proposed between models of different dimensionality according to the Metropolis-Hastings ratio:

$$\alpha = \min\left\{1, \frac{p(\vec{\lambda})_Y p(s|\vec{\lambda}_Y) q(\vec{u}_Y)}{p(\vec{\lambda})_X p(s|\vec{\lambda}_X) q(\vec{u}_X)} |\mathbb{J}|\right\} \tag{8}$$

Here, model X and model Y differ by some number of parameters, $q(\vec{u})$ is the distribution for random numbers chosen to generate the extra parameters, and $|\mathbb{J}|$ is the Jacobian of the two sets of parameters, which compensates for the difference in dimensionality. When using this Hastings ratio as an acceptance probability, we can allow our chains to explore the full space of allowed ppE models, both one- and two-term families, and use these to generate PDF's for the models themselves. The ratio of the heights of the PDF for model X and model Y is equal to the Bayes Factor between X and Y.

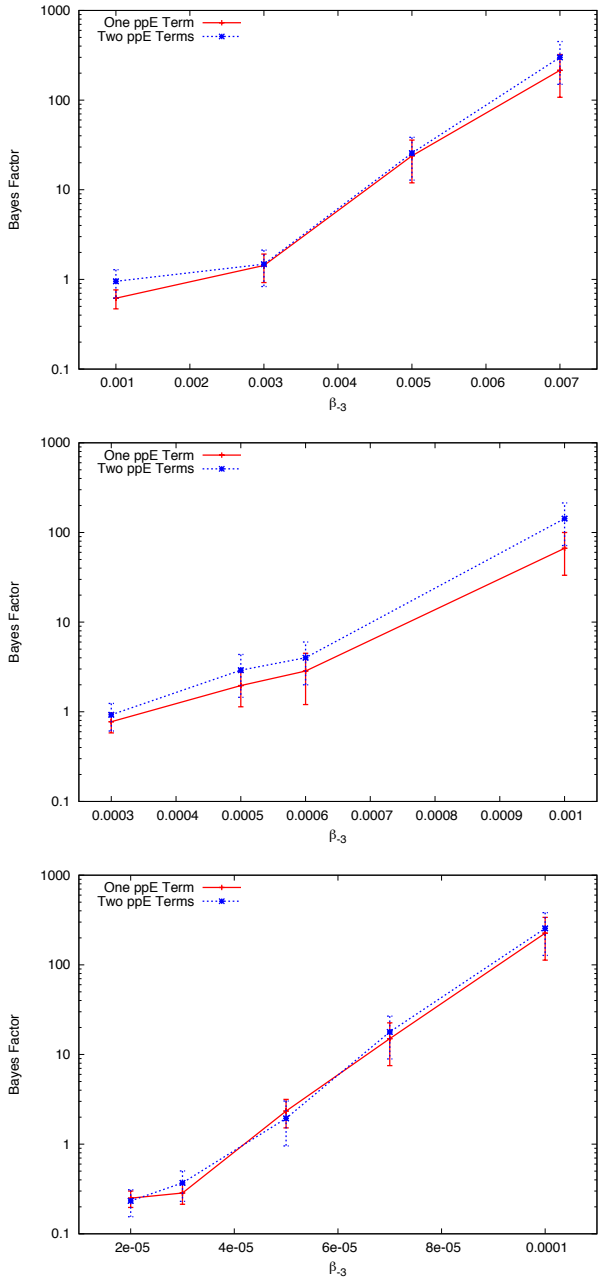

FIG. 8: (Color Online) Bayes factors for one-term (solid red) and two-term (dashed blue) ppE templates for a sub-critical (top-left), critical (top-right) and super-critical (bottom) ppE injection as a function of the injected value of β_{-3}. System parameters are listed in Table III, and useful cycles of phase in Table IV. In the sub- and super-critical cases, both models perform equally well at detecting a deviation from GR. In the critical case, the two-term model slightly out-performs the one-term model.

To generate Figure 9, we have run a RJMCMC search on three different types of signals - one sub-critical, one critical, and one super-critical - and plotted the number of iterations that the chains spent in each of the five different models. These five models include two ppE models with only one phase correction, ($b = -3$ and $b = -2$), two ppE models with two phase corrections, ($b = -3 + b = -2$ and $b = -2 + b = -1$), and GR. We find that, although there are some slight differences between the different models, in all cases we cannot draw meaningful distinctions between the different ppE models. The strongest Bayes Factor between two models is in the sub-critical case, where the Bayes Factor between the $b = -3$ only model and the $b = -2$ only model is ≈ 5. While this does show some preference for the first model, it is not a strong preference, and so we would not want to use this result to draw conclusions about the underlying theory of gravity. In summary - even though these signals are clearly differentiable from GR (all have Bayes Factors of ≈ 100), the four different ppE models perform almost as well in fitting the signal. This means that if we hope to gain more information about the underlying nature of an alternative gravity theory, we would need higher SNR signals and/or multiple detections. On a more hopeful note, it means that our ability to detect a deviation from GR is not strongly dependent on which particular ppE template we choose to use in our analysis.

Signal	ϕ_{-3}	ϕ_{-2}	ϕ_{-1}
Convergant	0.109	0.008	0.0005
Critical	0.024	0.051	0.085
Super-Critical	0.024	0.181	1.047

TABLE IV: Number of useful cycles from the different injected ppE terms - Fig. 8

V. FUTURE DIRECTIONS

In this paper, we have investigated the effects of using more realistic non-GR injections to investigate our ability to test GR using GW signals. We have found that the inclusion of noise in our analysis does not significantly affect our results, but that the failure to include higher-order deviations from GR in the phase of the injected signal can bias them. We have also determined that one-parameter ppE template families are best for detecting deviations from GR, at least for the simple cases investigated here.

The main direction of future work will be in determining how analyzing more *astrophysically* realistic systems affects our ability to test GR. That is, systems that incorporate not only more complicated deviations from GR, such as we examined in this paper, but that also include some of the messiness we know will exist in real systems in our universe. For instance, if we were analyzing systems that merge within the aLIGO frequency band, we would need to include the merger and ringdown parts of the waveforms in our injections. If we then performed a Bayesian model selection between ppE and GR *inspiral-only* templates, it is entirely possible that the ppE templates would win over the GR ones, simply by being able to fit more of the power in the non-inspiral signal. It is

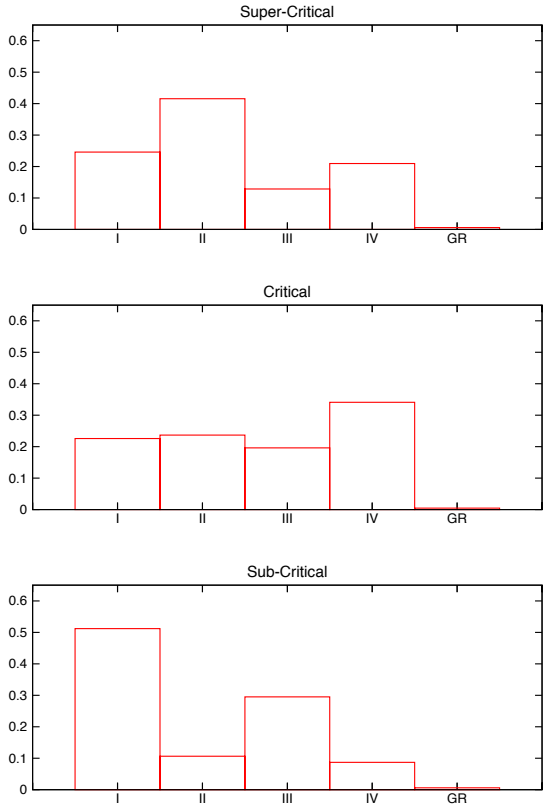

FIG. 9: Posterior distributions for the four different ppE models, generated by RJMCMC. The top two panels show the distribution for a sub-critical injection, the middle two for a critical injection, and the bottom two for an super-critical injection. All systems are NS-NS binaries with Bayes Factors of 100 favoring ppE over GR. System parameters are in Table III. Model I has $b = -3$, model II has $b = -2$, model III has $b = -3$ *and* $b = -2$, and model IV has $b = -2$ *and* $b = -1$. The y axis shows the percentage of iterations that the chain spent in each model, and the Bayes Factors between two models are simply the ratios of the percentages. Because the Bayes Factors are not large enough, these results indicate that we would not be able to make confident statements about the type of non-GR signal we had observed with this type of analysis.

also possible that the presence of accretion disks in real BH systems could alter the GW signature of the systems enough that we could mistakenly claim to have detected a deviation from GR. We will examine these potential sources of systematic error in a future paper.

Acknowledgments

We thank Michelle Vallisneri for his useful comments and suggestions. N. C. and L. S. acknowledge support from the NSF Award PHY-1205993 and NASA grant NNX10AH15G. N. Y. acknowledges support from NSF grant PHY-1114374 and NASA grant NNX11AI49G, under sub-award 00001944.

[1] C. M. Will, Living Reviews in Relativity **9** (2006), URL `http://www.livingreviews.org/lrr-2006-3`.
[2] M. Kramer et al., Science **314**, 97 (2006), astro-ph/0609417.
[3] N. Yunes and X. Siemens (2013), 1304.3473.
[4] B. P. Abbott, R. Abbott, R. Adhikari, P. Ajith, B. Allen, G. Allen, R. S. Amin, S. B. Anderson, W. G. Anderson, M. A. Arain, et al., Reports on Progress in Physics **72**, 076901 (2009), 0711.3041.
[5] T. Accadia, F. Acernese, F. Antonucci, P. Astone, G. Ballardin, F. Barone, M. Barsuglia, A. Basti, T. S. Bauer, M. Bebronne, et al., Classical and Quantum Gravity **28**, 114002 (2011).
[6] S. Alexander and N. Yunes, Phys. Rept. **480**, 1 (2009), 0907.2562.
[7] C. M. Will, Phys. Rev. **D50**, 6058 (1994), gr-qc/9406022.
[8] P. D. Scharre and C. M. Will, Phys. Rev. **D65**, 042002 (2002), gr-qc/0109044.
[9] C. M. Will and N. Yunes, Class. Quant. Grav. **21**, 4367 (2004), gr-qc/0403100.
[10] E. Berti, A. Buonanno, and C. M. Will, Class. Quant. Grav. **22**, S943 (2005), gr-qc/0504017.
[11] K. Yagi and T. Tanaka (2009), 0906.4269.
[12] C. M. Will, Phys. Rev. **D57**, 2061 (1998), gr-qc/9709011.
[13] A. Stavridis and C. M. Will, Phys. Rev. **D80**, 044002 (2009), 0906.3602.

[14] K. G. Arun and C. M. Will, Class. Quant. Grav. **26**, 155002 (2009), 0904.1190.

[15] D. Keppel and P. Ajith, Phys. Rev. **D82**, 122001 (2010), 1004.0284.

[16] S. Alexander, L. S. Finn, and N. Yunes, Phys. Rev. D **78**, 066005 (2008), 0712.2542.

[17] N. Yunes, R. O'Shaughnessy, B. J. Owen, and S. Alexander, Phys. Rev. **D82**, 064017 (2010), 1005.3310.

[18] N. Yunes and F. Pretorius, Physical Review D (Particles, Fields, Gravitation, and Cosmology) **79**, 084043 (pages 14) (2009), URL `http://link.aps.org/abstract/PRD/v79/e084043`.

[19] C. F. Sopuerta and N. Yunes, Physical Review D (Particles, Fields, Gravitation, and Cosmology) **80**, 064006 (pages 24) (2009), URL `http://link.aps.org/abstract/PRD/v80/e064006`.

[20] K. Yagi, N. Yunes, and T. Tanaka (2012), 1208.5102.

[21] N. Yunes, F. Pretorius, and D. Spergel (2009), 0912.2724.

[22] J. D. Bekenstein, Phys. Rev. **D70**, 083509 (2004), astro-ph/0403694.

[23] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Phys. Rev. **D74**, 024006 (2006), gr-qc/0604067.

[24] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Class. Quant. Grav. **23**, l37 (2006), gr-qc/0604018.

[25] C. K. Mishra, K. G. Arun, B. R. Iyer, and B. S. Sathyaprakash (2010), 1005.0304.

[26] N. Yunes and F. Pretorius, Phys. Rev. **D80**, 122003 (2009), 0909.3328.

[27] N. Cornish, L. Sampson, N. Yunes, and F. Pretorius, Phys.Rev. **D84**, 062003 (2011), 1105.2088.

[28] T. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, et al., Phys.Rev. **D85**, 082003 (2012), 1110.0530.

[29] K. Chatziioannou, N. Yunes, and N. Cornish, Phys.Rev. **D86**, 022004 (2012), 1204.2585.

[30] N. J. Cornish and T. B. Littenberg, Physical Review D (Particles, Fields, Gravitation, and Cosmology) **76**, 083006 (pages 11) (2007), URL `http://link.aps.org/abstract/PRD/v76/e083006`.

[31] W. Del Pozzo, J. Veitch, and A. Vecchio, ArXiv e-prints (2011), 1101.1391.

[32] J. M. Dickey, Ann. Math. Statist. **42**, 204 (1971).

[33] M. Sambridge, K. Gallagher, A. Jackson, and P. Rickwood, Geophysical Journal International **16**, 528 (2006).

[34] N. Yunes and S. A. Hughes, Phys. Rev. D **82**, 082002 (2010), 1007.1995.

[35] T. Damour, B. R. Iyer, and B. S. Sathyaprakash, Phys. Rev. D **62**, 084036 (2000), arXiv:gr-qc/0001023.

[36] N. J. Cornish (2010), 1007.4820.

[37] M. Vallisneri, Phys.Rev.Lett. **107**, 191104 (2011), 1108.1158.

[38] S. Nissanke, D. E. Holz, S. A. Hughes, N. Dalal, and J. L. Sievers, Astrophys.J. **725**, 496 (2010), 0904.1017.

[39] N. Yunes, B. Kocsis, A. Loeb, and Z. Haiman, Phys.Rev.Lett. **107**, 171103 (2011), 1103.4609.

[40] B. Kocsis, N. Yunes, and A. Loeb, Phys.Rev. **D84**, 024032 (2011), 1104.2322.

[41] N. Yunes, M. Coleman Miller, and J. Thornburg, Phys.Rev. **D83**, 044030 (2011), 1010.1721.

[42] K. Yagi and N. Yunes (2013), 1303.1528.

[43] K. Yagi and N. Yunes (2013), 1302.4499.

[44] K. G. Arun and C. M. Will, Classical and Quantum Gravity **26**, 155002 (2009), 0904.1190.