Improving gravitational-wave parameter estimation using Gaussian process regression

PHYSICAL REVIEW D **93,** 064001 (2016)

Improving gravitational-wave parameter estimation using Gaussian process regression

Christopher J. Moore,[1,*] Christopher P. L. Berry,[2,†] Alvin J. K. Chua,[1,‡] and Jonathan R. Gair[1,3,§]

[1]*Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, United Kingdom*
[2]*School of Physics & Astronomy, University of Birmingham, Birmingham B15 2TT, United Kingdom*
[3]*School of Mathematics, University of Edinburgh, King's Buildings, Edinburgh EH9 3JZ, United Kingdom*

(Received 14 September 2015; published 1 March 2016)

Folding uncertainty in theoretical models into Bayesian parameter estimation is necessary in order to make reliable inferences. A general means of achieving this is by marginalizing over model uncertainty using a prior distribution constructed using Gaussian process regression (GPR). As an example, we apply this technique to the measurement of chirp mass using (simulated) gravitational-wave signals from binary black holes that could be observed using advanced-era gravitational-wave detectors. Unless properly accounted for, uncertainty in the gravitational-wave templates could be the dominant source of error in studies of these systems. We explain our approach in detail and provide proofs of various features of the method, including the limiting behavior for high signal-to-noise, where systematic model uncertainties dominate over noise errors. We find that the marginalized likelihood constructed via GPR offers a significant improvement in parameter estimation over the standard, uncorrected likelihood both in our simple one-dimensional study, and theoretically in general. We also examine the dependence of the method on the size of training set used in the GPR; on the form of covariance function adopted for the GPR, and on changes to the detector noise power spectral density.

DOI: 10.1103/PhysRevD.93.064001

I. INTRODUCTION

The era of advanced ground-based interferometric gravitational-wave (GW) detectors is here. The Advanced LIGO detectors [1,2] in the USA started observing in September 2015, whilst the Advanced Virgo detector [3,4] in Europe is expected to come online shortly afterwards [5]. The principal target sources of GWs for these detectors are the coalescences of pairs of compact objects, either neutron stars or black holes. For all sources there is great uncertainty in the quoted event rate estimates, at least an order of magnitude in either direction [6], but regardless of the astrophysical uncertainty, the prospect of a first detection is imminent.[1]

The detection of GW signals is most efficient when we have accurate waveform models that can be matched to any signals in the (noisy) detector data.[2] For parameter estimation (PE), it is even more important that the template faithfully matches the true signal, as otherwise we could infer biased parameter values. Matching a template to GW data requires that the model waveform remains accurate over the entire duration of the signal; typically of the order of hundreds of seconds for neutron-star binaries and tens of seconds for black-hole binaries in the advanced-detector era (with frequency sensitivity down to 10 Hz). Although higher mass sources have shorter waveforms (in the detector band), these present more of a challenge for modeling as they have detectable merger and ringdown components. In contrast, binary neutron stars only have the (easier to model) inspiral part of the waveform in band. In this paper, we are concerned with problems that arise from inaccurate models; therefore, we focus our attention on black-hole binaries where the issue of waveform uncertainty is most acute. However, the techniques we develop could equally be applied to neutron-star binaries or any other uncertain signal. Inaccurate waveform models are known to cause significant systematic errors when recovering source parameters from observations with both ground-based [11] and space-based GW detectors [12].

There are two problems that arise when using inaccurate signal models: the *detection* problem, and the PE problem. The detection problem is that the inaccurate model does not perfectly match to the physical waveform, leading to a loss of signal-to-noise (SNR) ratio and, hence, a lower chance of detection (for the same false alarm probability). The PE problem, which is the focus of this paper, is that the model waveform which has the best overlap with the physical signal in the data generally has parameter values offset from the true source parameters, leading to a *systematic* error in any parameter estimates.

*cjm96@ast.cam.ac.uk
†cplb@star.sr.bham.ac.uk
‡ajkc3@ast.cam.ac.uk
§j.gair@ed.ac.uk

[1]While this paper was in proof, the first detection was announced [7].

[2]It is possible to detect GWs without templates by looking for coherent excess power in the detectors, e.g., [8–10]. This is effective for short-duration signals corresponding to high-mass binaries.

Recently, some of the authors proposed a novel method of improving the detection and PE prospects of complicated physical phenomena in noisy data [13]. The method applies generally to any situation where accurate models of the signal are available, but computational constraints mean that routine detection and PE tasks must be carried out with cheaper, less accurate, models.

In the case of binary black-holes, there are many different physical effects that should be included in waveform models, such as the merger and ringdown phases following the inspiral, the presence of generic spins and precession, eccentricity and higher-order modes. All these phenomena can, in principle, be simulated, thanks to recent rapid progress in numerical relativity (NR) [14–16]. However, NR simulations are extremely expensive, only a few hundred have been performed to date (see [17,18] and references therein), and these typically consist of only the final few tens of orbits (although, see [19]). Detection and PE are therefore currently performed using less expensive waveform approximants. The existence of NR waveforms has permitted the calibration of analytic inspiral–merger–ringdown approximants such as the effective-one-body–NR (EOBNR) [20–23] or IMRPhenom [24] families, with recent efforts concentrating on including the effects of precession in these [25–27].[3] For some recent PE work with inspiral merger and ringdown waveform models see [29–31]. Historically, PE has used models based on the post-Newtonian formalism [32], such as the TaylorF2 and TaylorT2 waveforms [33]. Despite these lacking some of the relevant features, they are sufficiently quick to calculate that they can be simply used in PE algorithms.

To include uncertainty in waveform templates whilst minimizing computational expense, we use Gaussian process regression (GPR) to estimate the effects of waveform errors. The method involves constructing a training set of the waveform differences between an expensive, accurate waveform and a cheaper, less accurate waveform. For the accurate waveforms it might be necessary to use some combination of NR and the NR calibrated approximants discussed above, depending on the numbers and length of the available NR simulations. The waveform difference is evaluated at a relatively small number of points in parameter space and stored for later use. GPR is then used to interpolate the difference across parameter space to give a best estimate and a corresponding uncertainty at a general point in parameter space. This interpolation provides a prior probability distribution on the waveform difference which is then used in marginalizing the likelihood over waveform uncertainty. The result is an expression for the likelihood in terms of the cheaper waveform model, but with corrections coming from the training set. This marginalized likelihood is negligibly more complicated or computationally expensive to evaluate than the standard expression, but provides a better estimate of the true likelihood surface (and hence the posterior), factoring in our imperfect knowledge of the waveform. Therefore, we have not only built a relatively inexpensive waveform approximant that can include additional physics, but we have also accounted for (marginalized over) the uncertainty in our new approximant.

If the standard likelihood with an approximate waveform is used for PE then, in general, biased parameter estimates are obtained.[4] It has recently been shown by some of the authors [36] that, under certain conditions, this bias is completely removed by the marginalized likelihood, and, more generally, that the bias is always reduced by the marginalized likelihood.

The technique of GPR assumes that the data in the training set have been drawn from a Gaussian process (GP) on the parameter space with a mean and a covariance function either specified *a priori* or estimated from the training set itself. The interpolation is then achieved by calculating the conditional probability for the GP at some new parameter point given the known training set values, the mean and the covariance. GPR provides a convenient nonparametric way to interpolate the waveform differences, and has the additional advantage that, by construction, it provides a Gaussian probability distribution for the unknown waveform difference which can be analytically marginalized over. This is important because it means no extra nuisance parameters are added to the PE task which would slow down an already expensive process.

The outline of this paper is as follows. In Sec. II the concept of the marginalized likelihood is introduced and the use of GPR in its construction is described in detail; we limit ourselves to interpolation across parameter space and not across frequency. The main choice made in implementing GPR is the specification of the covariance function; Sec. III discusses how the properties of the covariance function affect the properties of the corresponding GP, and the effects of different choices of covariance function are examined in a toy one-dimensional GPR problem. The marginalized likelihood possesses several properties which make it appealing for GW astronomy; Sec. IV presents proofs of these and discussions of their significance. In Sec. V the implementation of the marginalized likelihood is described for an illustrative one-dimensional example; here, properties of the interpolated waveforms are examined and PE results for the marginalized likelihood are also

[3]See [28] for a study of systematic error (or lack thereof) from using EOBNR waveforms with NR injections.

[4]This is commonly assessed in the GW literature using probability–probability (P–P) plots [34,35]; for a catalogue of events, these plot the cumulative fraction of events where the true parameter is found within the credible interval corresponding to a given probability. If the posteriors are well calibrated, then a proportion P should fall in the P credible interval, and the plot is a diagonal line. Introducing bias means that the line sags below the optimal diagonal.

presented. Additional material on the effect of changing the detector noise properties on the interpolated waveforms are considered in Appendix B. Finally, concluding remarks and a discussion of future directions for implementing the marginalized likelihood are presented in Sec. VI.

II. THE METHOD

In this section we detail how we incorporate waveform uncertainties into GW data analysis. The material presented is an expansion of that in [13]. In Sec. II A we introduce the standard likelihood function and show how model uncertainties can be treated like nuisance parameters that can be integrated out (marginalized over). Performing this integration requires that a prior probability distribution is specified for the model uncertainties, this is constructed using GPR. This is introduced in Sec. II B, where we briefly summarize some key results pertaining to GPR; further details can be found in standard textbooks (e.g., [37–39]). The result of the integration is the *marginalized likelihood* presented in Eq. (28) which accurately encodes our state of knowledge of the signal parameters, given our imperfect waveform models and the noisy data.

A. The marginalized likelihood

We consider the scenario where we can construct two different waveform models, one accurate but computationally expensive, the other less accurate but quick to calculate. We use the parameters vector $\vec{\lambda}$ to fully characterize the GW signal; Latin indices from the beginning of the alphabet $(a, b, \ldots)$ will be used to label the different components of this vector, and repeated indices should be summed over. The accurate waveforms will be referred to as the exact waveform $h(\vec{\lambda})$, although the method does not require that the accurate waveforms are perfect (see Sec. III C). The cheaper approximate waveform $H(\vec{\lambda})$ is related to $h(\vec{\lambda})$ by the waveform difference

$$H(\vec{\lambda}) = h(\vec{\lambda}) + \delta h(\vec{\lambda}). \tag{1}$$

The waveform templates may be calculated in either the time domain $h(t;\vec{\lambda})$ or the frequency domain $\tilde{h}(f;\vec{\lambda})$; the dependence of the waveform on time or frequency is suppressed in our notation for brevity.

In the context of modeling binary black-hole coalescences there are several highly accurate waveform approximants available, for example, NR waveforms [18] or spin EOBNR (SEOBNR) models [22,23,40]. There are also multiple possibilities for the approximate waveform family, for example, the Taylor family of approximants [33]. For the proof-of-principal numerical calculations in this paper, we need to be able to perform mock PE runs with both waveform families so that we can assess our marginalization technique does indeed offer a significant improvement. Therefore, we will pick both approximants to be quick to compute, rather than selecting on accuracy: our choice of waveform family is discussed in more detail in Sec. V A.

In a PE study, we wish to construct the posterior probability distribution for the signal parameters given the observed data (and any prior information we have about the source) $p(\vec{\lambda}|s)$. From Bayes' theorem, the posterior is given by

$$p(\vec{\lambda}|s) = \frac{L'(s|\vec{\lambda})\pi(\vec{\lambda})}{\mathcal{Z}'(s)}, \tag{2}$$

where (keeping the notation of [13]) $L'(s|\vec{\lambda})$ is the likelihood, $\pi(\vec{\lambda})$ is the prior distribution on the parameters and $\mathcal{Z}'(s)$ is the normalizing evidence

$$\mathcal{Z}'(s) = \int L'(s|\vec{\lambda})\pi(\vec{\lambda})\mathrm{d}\vec{\lambda}. \tag{3}$$

In a Bayesian analysis the evidence $\mathcal{Z}'(s)$ can be used as the detection statistic (by comparing it with the evidence for the null hypothesis to form the Bayes' factor) [41], and the positions and widths of peaks in the posterior $p(\vec{\lambda}|s)$ are used to give the parameter estimates and associated uncertainties [42]. For simplicity (although it is not necessary to do so), we assume throughout that $\pi(\vec{\lambda})$ is flat within the relevant region of parameter space. The single remaining challenge is to calculate the likelihood $L'(s|\vec{\lambda})$.

For a detector with stationary, Gaussian noise with power spectral density $S_n(f)$ [43], the likelihood is given by [44]

$$L'(s|\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\langle s - h(\vec{\lambda})|s - h(\vec{\lambda})\rangle\right). \tag{4}$$

Here the noise-weighted inner product has been defined as [45]

$$\begin{aligned}\langle x|y\rangle &= 4\Re\left\{\int_0^\infty \mathrm{d}f \frac{\tilde{x}(f)\tilde{y}(f)^*}{S_n(f)}\right\} \\ &= 4\Re\left\{\sum_{\kappa=1}^{M} \delta f \frac{\tilde{x}(f_\kappa)\tilde{y}(f_\kappa)^*}{S_n(f_\kappa)}\right\},\end{aligned} \tag{5}$$

where κ labels the M frequency bins with resolution δf. We define the norm of a waveform as

$$\|x\| = \sqrt{\langle x|x\rangle}, \tag{6}$$

for a signal this is equivalent to its SNR.

In practice it can be unfeasible to sample from the likelihood distribution in Eq. (4) because it is prohibitively expensive to calculate the exact waveforms $h(\vec{\lambda})$; instead, we must reply on the approximate waveforms to calculate an approximate likelihood,

$$L(s|\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\|s - H(\vec{\lambda})\|^2\right). \tag{7}$$

For a good approximant

$$L(s|\vec{\lambda}) \approx L'(s|\vec{\lambda}); \tag{8}$$

the natural way to improve this agreement is to construct (inevitably more expensive) approximants that have smaller waveform differences $\delta h(\vec{\lambda})$. Instead, the proposal of this paper is to replace $L(s|\vec{\lambda})$ with a new likelihood which accounts for the uncertainty in the waveforms. The alternative likelihood is

$$\mathcal{L}(s|\vec{\lambda}) \propto \int \mathrm{d}[\delta h(\vec{\lambda})] P[\delta h(\vec{\lambda})] \times \exp\left(-\frac{1}{2}\|s - H(\vec{\lambda}) + \delta h(\vec{\lambda})\|^2\right). \tag{9}$$

This new likelihood has marginalized over the uncertainty in the waveform difference using the (as yet unspecified) prior on the waveform difference $P[\delta h(\vec{\lambda})]$.

The prior on the waveform difference should include the information available from the limited number of available accurate waveforms and could also encode our prior expectations about the signal, for example, that the approximate waveforms are most accurate at early times (or equivalently at low frequencies) when the orbiting bodies are well separated [32], but gradually become inaccurate as the bodies inspiral. At most points in parameter space, an accurate waveform is not available, and so it is necessary to interpolate the waveform difference across parameter space while simultaneously accounting for the error this introduces. It would seem that the problem rapidly becomes complicated, and even if a suitable prior could be constructed the computational time needed to evaluate $\mathcal{L}(s|\vec{\lambda})$ would make it impractical in most contexts.

Fortunately, the technique of GPR provides a natural way to interpolate the waveform differences across parameter space, incorporating all necessary prior information. GPR also has the additional property that it naturally returns an expression for $P[\delta h(\vec{\lambda})]$ which is a Gaussian in $\delta h(\vec{\lambda})$. Since the exponential factor in Eq. (9) is also Gaussian in $\delta h(\vec{\lambda})$, the functional integral can be evaluated analytically. This gives an analytic expression for $\mathcal{L}(s|\vec{\lambda})$ which can be evaluated in approximately the same computational time as $L(s|\vec{\lambda})$.

Henceforth, for brevity, the s dependence will be suppressed in all likelihoods, i.e. $L'(\vec{\lambda}) \equiv L'(s|\vec{\lambda})$, $L(\vec{\lambda}) \equiv L(s|\vec{\lambda})$, and $\mathcal{L}(\vec{\lambda}) \equiv \mathcal{L}(s|\vec{\lambda})$.

B. Gaussian process regression

Assume that we have access to accurate waveforms at a few values of the parameters $\{h(\vec{\lambda}_i)|i = 1, 2, \ldots, N\}$ and can cheaply compute approximate waveforms at the same parameter values. Our training set is the set of waveform differences

$$\mathcal{D} = \{(\vec{\lambda}_i, \delta h(\vec{\lambda}_i))|i = 1, 2, \ldots, N\}. \tag{10}$$

where necessary the Latin indices $i, j, \ldots$ will be used to label the different components of the training set (repeated indices are not summed over unless specified). It is now necessary to interpolate the training set to obtain the prior on the waveform difference first defined in Eq. (9),

$$P[\delta h] \equiv P(\delta h(\vec{\lambda})|\mathcal{D}, \mathcal{I}), \tag{11}$$

where $\mathcal{I}$ is any other prior information we possess about the waveforms. The simplest and most natural choice for such a prior is to assume that the waveform difference is a realisation of a GP (a Gaussian is the maximum-entropy distribution given that we know a characteristic range of variation [46]),

$$\delta h(\vec{\lambda}) \sim \mathcal{GP}(m(\vec{\lambda}), k(\vec{\lambda}, \vec{\lambda}')). \tag{12}$$

A GP can loosely be thought of as the generalisation of a Gaussian distribution to an infinite number of degrees of freedom. It is completely specified by the mean $m(\vec{\lambda})$ and covariance $k(\vec{\lambda}, \vec{\lambda}')$ functions in the same way as a Gaussian distribution is fully specified by a mean and variance. More formally, a GP is an infinite collection of variables, any finite subset of which are distributed as a multivariate Gaussian. For a set of parameter points $\{\vec{\lambda}_i\}$, including, but not limited to, the training set $\mathcal{D}$,

$$[\delta h(\vec{\lambda}_i)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}), \tag{13}$$

where the mean vector and covariance matrix of this Gaussian distribution are fixed by the corresponding functions of the GP,

$$[\boldsymbol{m}]_i = m(\vec{\lambda}_i), \qquad [\boldsymbol{K}]_{ij} = k(\vec{\lambda}_i, \vec{\lambda}_j), \tag{14}$$

with probability density function (here correcting the normalizing prefactor written in [13] which mistakenly included a square root)

$$P(\{\delta h(\vec{\lambda}_i)\}) = \frac{1}{(2\pi)^N |\boldsymbol{K}|} \exp\left(-\frac{1}{2}\sum_{i,j}[\boldsymbol{K}^{-1}]_{ij}(\delta h(\vec{\lambda}_i)|\delta h(\vec{\lambda}_j))\right). \quad (15)$$

The round brackets denote a new inner product with respect to some noise weighting $S'_n(f)$, which we leave unspecified for the moment;

$$(x|y) = 4\Re\left\{\int_0^\infty \mathrm{d}f \frac{\tilde{x}(f)\tilde{y}(f)^*}{S'_n(f)}\right\} = 4\Re\left\{\sum_{\kappa=1}^{M} \delta f \frac{\tilde{x}(f_\kappa)\tilde{y}(f_\kappa)^*}{S'_n(f_\kappa)}\right\}. \quad (16)$$

In writing down Eq. (15) and stipulating that the covariance function $k(\vec{\lambda}, \vec{\lambda}')$ has no dependence on frequency, we are effectively assuming that (i) the parameter space structure of the model errors is frequency independent; and (ii) the typical size of errors has a frequency dependence proportional to $\sqrt{S'_n(f)}$. Under the assumption that waveform model errors are uncorrelated in frequency, the normalizing factor in Eq. (15) should be raised to the power M; however, this assumption leads to model errors that average to zero over frequency and have only a small effect on the likelihood. The optimal means of incorporating frequency dependence would be to introduce an additional covariance function in frequency as well as the covariance in parameter space. This frequency covariance introduces a correlation length scale in frequency which can be learned from the training set in exactly the same manner as we describe below for correlations in $\vec{\lambda}$. This correlation length scale reduces the number of independent frequencies from M to some new effective number M_{eff}.

Performing this double GPR interpolation in f and $\vec{\lambda}$ is beyond the scope of the current paper. Instead, here we are in effect setting $M_{\mathrm{eff}} = 1$, giving the expression in Eq. (15); this is analogous to assuming that all the frequency bins of the noise-weighted waveform at a particular point in parameter space are perfectly correlated. Setting $M_{\mathrm{eff}} = 1$ gives the largest uncertainty of any fixed number of independent frequencies and is therefore a conservative choice. Despite these simplifications, our marginalized likelihood has many desirable properties (which we discuss and prove in Sec. IV), and performs well in the numerical example presented in Sec. V. We will return to the more general problem of performing the extended GPR including frequency in the future.

Specifying how we compute the mean and variance for the GP determines how the waveforms are interpolated and fixes our prior for waveform uncertainty across parameter space. Our GP has a zero mean as we have chosen to interpolate the waveform difference rather than the waveform directly. By first subtracting off an approximate model we leave a quantity which is uncertain, but has no known bias. If we had some additional prior knowledge that the approximate waveform was systematically wrong across parameter space, then this should be added into the approximate model so that the zero-mean assumption becomes valid. Identical results for the marginalized likelihood could also be obtained by directly interpolating the accurate waveforms using a GP with a mean equal to the approximate waveforms; however, we choose to interpolate waveform differences because zero-mean GPs are simpler to handle numerically.

Specifying the covariance function is central to GPR as it encodes our prior expectations about the properties of the function being interpolated. Possibly the simplest and most widely used choice for the covariance function is the squared exponential (SE) [38]

$$k(\vec{\lambda}_i, \vec{\lambda}_j) = \sigma_f^2 \exp\left[-\frac{1}{2} g_{ab}(\vec{\lambda}_i - \vec{\lambda}_j)^a(\vec{\lambda}_i - \vec{\lambda}_j)^b\right], \quad (17)$$

which defines a stationary, smooth GP. In Eq. (17), a scale σ_f and a (constant) metric g_{ab} for defining a modulus in parameter space have been defined. These are called *hyperparameters* and we denote them as $\vec{\theta} = \{\sigma_f, g_{ab}\}$, with Greek indices $\mu, \nu, \ldots$ to label the components of this vector. If the available accurate waveforms contain some uncertainty then this can also be included by adding a diagonal matrix $\boldsymbol{C}$ to Eq. (17), where the element C_{ii} (no summation) is the uncertainty in the accurate simulation at $\vec{\lambda}_i$; this is discussed further in Sec. III C.

The probability in Eq. (15) is referred to as the *hyperlikelihood*, or alternatively the *evidence* (as in [13]) for the training set; it is the probability that particular realization of waveform differences was obtained from a GP with a zero mean and specified covariance function. The hyperlikelihood depends only on the hyperparameters and the quantities in the training set, so we denote it as $Z(\vec{\theta}|\mathcal{D})$. The log hyperlikelihood is [5]

$$\ln Z(\vec{\theta}|\mathcal{D}) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\sum_{i,j}\mathrm{inv}[k(\vec{\lambda}_i, \vec{\lambda}_j)](\delta h(\vec{\lambda}_i)|\delta h(\vec{\lambda}_j)) - \frac{1}{2}\ln|\det[k(\vec{\lambda}_i, \vec{\lambda}_j)]|. \quad (18)$$

For all subsequent calculations the values of the hyperparameters are fixed to their optimum values $\vec{\theta}_{\mathrm{op}}$, defined as those which maximise the hyperlikelihood:

[5]Unless explicitly indicated otherwise all the logarithms used in this paper are natural logarithms.

$$\left.\frac{\partial Z(\vec{\theta}|\mathcal{D})}{\partial\theta^{\mu}}\right|_{\vec{\theta}=\vec{\theta}_{\rm op}}=0. \tag{19}$$

Maximizing the hyperlikelihood with respect to $\vec{\theta}$ is one of many approaches which could be taken. For example, a better motivated approach would be to consider the hyperparameters as nuisance parameters in addition to the source parameters $\vec{\lambda}$, and marginalize over them while sampling an expanded likelihood,

$$\Lambda_{\rm expanded}(\vec{\lambda},\vec{\theta}|\mathcal{D})\propto\mathcal{L}(\vec{\lambda}|\vec{\theta},\mathcal{D})Z(\vec{\theta}|\mathcal{D}). \tag{20}$$

The disadvantage of this approach is that the hyperlikelihood is *much* more expensive to compute than the standard approximate likelihood and the inclusion of extra nuisance parameters also slows down any PE. In contrast, our proposed method of maximizing the likelihood is a convenient heuristic which is widely used in other contexts [47–49] and allows all the additional computation to be done offline. It would be useful, in future work, to check explicitly that the different ways of dealing with the hyperparameters give consistent results in the context of GW source modeling.

Having fixed the properties of the covariance function by examining the training set, we can now move on to using the GP as a predictive tool. The defining property of the GP is that *any* finite collection of variables drawn from it is distributed as a multivariate Gaussian in the manner of Eq. (15). Therefore, the set of variables formed by the training set plus the waveform difference at one extra parameter point $\delta h(\vec{\lambda})$ is distributed as

$$\begin{bmatrix}\delta h(\vec{\lambda}_i)\\ \delta h(\vec{\lambda})\end{bmatrix}\sim\mathcal{N}(\mathbf{0},\Sigma),\qquad \Sigma=\begin{pmatrix}\boldsymbol{K} & \boldsymbol{K}_*\\ \boldsymbol{K}_*^T & \boldsymbol{K}_{**}\end{pmatrix}, \tag{21}$$

where $\boldsymbol{K}$ is defined in Eq. (14) and the vector $\boldsymbol{K}_*$ and scalar K_{**} are defined as

$$[\boldsymbol{K}_*]_i=k(\vec{\lambda}_i,\vec{\lambda}),\qquad K_{**}=k(\vec{\lambda},\vec{\lambda}). \tag{22}$$

On the right-hand side of Eq. (21) all the quantities are known because the hyperparameters have been fixed to their optimum values, and on the left-hand side all the quantities are known (from the training set) except for $\delta h(\vec{\lambda})$. Therefore, the conditional probability of the unknown waveform difference given the known differences in $\mathcal{D}$ can be found. This conditional probability is given by (e.g., [37,38])

$$P[\delta h(\vec{\lambda})]=\frac{1}{2\pi\sigma^2(\vec{\lambda})\prod_{\kappa=1}^{M}S_n'(f_\kappa)}\times\exp\left(-\frac{(\delta h(\vec{\lambda})-\mu(\lambda)|\delta h(\vec{\lambda})-\mu(\lambda))}{2\sigma^2(\vec{\lambda})}\right), \tag{23}$$

where the GPR mean and its associated error have been defined as

$$\mu(\vec{\lambda})=\sum_{i,j}[\boldsymbol{K}_*]_i[\boldsymbol{K}^{-1}]_{ij}\delta h(\vec{\lambda}_j), \tag{24}$$

$$\sigma^2(\vec{\lambda})=K_{**}-\sum_{i,j}[\boldsymbol{K}_*]_i[\boldsymbol{K}^{-1}]_{ij}[\boldsymbol{K}_*]_j. \tag{25}$$

Furnished with the expression for $P[\delta h(\vec{\lambda})]$, the marginalized likelihood in Eq. (9) can now be evaluated. The integrand in Eq. (9) is the product of two Gaussians and can be calculated analytically,

$$\mathcal{L}(\vec{\lambda})\propto\frac{1}{1+\sigma^2(\vec{\lambda})\prod_{\kappa=1}^{M}(S_n'(f_\kappa)/S_n(f_\kappa))}\times\exp\left(-\frac{1}{2}[s-H(\vec{\lambda})+\mu(\vec{\lambda})|s-H(\vec{\lambda})+\mu(\vec{\lambda})]\right). \tag{26}$$

The square brackets denote a third inner product with respect to the new noise weighting $S_n''(f)$, where $S_n''(f,\vec{\lambda})\equiv S_n(f)+\sigma^2(\vec{\lambda})S_n'(f)$,

$$[x|y]=4\Re\left\{\int_0^\infty {\rm d}f\frac{\tilde{x}(f)\tilde{y}(f)^*}{S_n''(f)}\right\}=4\Re\left\{\sum_{\kappa=1}^{M}\delta f\frac{\tilde{x}(f_\kappa)\tilde{y}(f_\kappa)^*}{S_n''(f_\kappa)}\right\}. \tag{27}$$

For the remainder of this paper, for simplicity, we take $S_n'(f)=S_n(f)$ so the three signal inner products we have defined become $\langle\cdot|\cdot\rangle=(\cdot|\cdot)=[\cdot|\cdot]/(1+\sigma^2(\vec{\lambda}))$ [13]. With this simplifying assumption, the marginalized likelihood becomes

$$\mathcal{L}(\vec{\lambda})\propto\frac{1}{1+\sigma^2(\vec{\lambda})}\exp\left(-\frac{1}{2}\frac{\|s-H(\vec{\lambda})+\mu(\vec{\lambda})\|^2}{1+\sigma^2(\vec{\lambda})}\right). \tag{28}$$

As mentioned earlier Eq. (15) issues that the waveform model errors are uncorrelated in frequency. The assumption that $S_n'(f)=S_n(f)$ additionally assumes that the typical size of the waveform error at a frequency f is given by $\sqrt{S_n(f)}$. This choice can be motivated to a certain extent by examining the hyperlikelihood in Eq. (18) which is used to train the GP. This hyperlikelihood contains the overlap matrix $(\delta h(\vec{\lambda}_i)|\delta h(\vec{\lambda}_j))$. Choosing $S_n'(f)=S_n(f)$ acts to

downweight the correlations at frequencies we are insensitive to (ignoring errors we cannot measure) and hence the resulting hyperparameters give an interpolant which is tuned to better represent the waveform correlations at the frequencies to which we are most sensitive: we weight waveform errors based upon their impact on the likelihood. The assumption of frequency-independent models errors gives a value for the GPR uncertainty σ^2 in Eq. (28) that is also frequency-independent. This can be shown to be a conservative choice in the sense that it gives broader and less informative posteriors.

In a follow-on study we will provide a proof of the conservative nature of this assumption and examine a number of different choices for the weighting function $S_n'(f)$, but we use the simplifying assumption $S_n'(f) = S_n(f)$ throughout the current paper. Despite these simplifying assumptions, we find that the resulting likelihood in Eq. (28) performs well. In Appendix B we examine the sensitivity of the method to small changes in the noise curve $S_n(f)$ which will occur in real experiments.

In Eq. (28) the best fit waveform has shifted by an amount $\mu(\vec{\lambda})$; this is the best estimate of the waveform difference returned by the GPR. The quantity $H(\vec{\lambda}) + \mu(\vec{\lambda})$ can be regarded as a new waveform approximant built from the accurate and approximate waveforms with the aid of GPR. However, a bonus of this way of including the training set directly into the likelihood is that the extra uncertainty associated with using the GPR as an interpolant is automatically included via the broadening of the posterior caused by $\sigma^2(\vec{\lambda}) \geq 0$.

In this section we have explained how uncertainty in waveform models can be included in PE through use of a marginalized likelihood. We defined such a likelihood in Eq. (9), but the marginalization requires a prior probability on the waveform uncertainty across parameter space. We construct this from a training set using GPR; the resulting prior is given in Eq. (23). Since this is of Gaussian form, we can marginalize analytically to produce the new likelihood Eq. (28). The properties of this marginalized likelihood are explored extensively throughout the remainder of this paper.

In Secs. III and IV we discuss theoretical properties of the GPR and marginalized likelihood respectively. A reader who is primarily interested in the PE results obtained with the likelihood in Eq. (28) may skip to Sec. V.

III. THE COVARIANCE FUNCTION

In the previous section we described how waveform uncertainties could be marginalized out using a prior constructed by using GPR on a training set. The only aspect of this that is not prescribed by the training data is the choice of the covariance function. This plays an important role in determining the properties of a GP. In this section, we discuss the properties of different choices of the covariance function in GPR. The properties of the covariance functions discussed in this section are known in the GPR literature, but are included here as they are not widely appreciated in the GW community. The material presented in this section on the covariance function will be used in the interpretation of our results in Sec. IV and Sec. V.

The only necessary requirements we have of a covariance function are that it is a positive definite; i.e. for *any* choice of points $\{\vec{\lambda}_i\}$ the covariance matrix $K_{ij} = k(\vec{\lambda}_i, \vec{\lambda}_j)$ is positive definite.

Throughout this paper, GPs are assumed to have zero mean, and therefore be fully specified by the covariance function $k(\vec{\lambda}_1, \vec{\lambda}_2)$. However, the proofs regarding continuity and differentiability of GPs discussed in this section, and proved in Appendix A, are done without recourse to the zero-mean assumption. The covariance encodes all information available about the properties of the function being interpolated by the GPR. It is central to the GPR and hence also to the marginalized likelihood.

The covariance function (and the corresponding GP) is said to be *stationary* if the covariance is a function only of $\vec{\tau} = \vec{\lambda}_1 - \vec{\lambda}_2$, furthermore it is said to be *isotropic* if it is a function only of $\tau \equiv |\vec{\tau}| = |\vec{\lambda}_1 - \vec{\lambda}_2|$.[6] Isotropy of a GP implies stationarity. All of the GPs used for numerical calculations in this paper are isotropic (and hence stationary) $k(\vec{\lambda}_1, \vec{\lambda}_2) \equiv k(\vec{\tau}) \equiv k(\tau)$, although the generalization to nonstationary GPs is briefly discussed in Sec. III B.

An example of how the properties of the covariance function relate to the properties of the GP, and hence the properties of the resulting interpolant, is given by considering the *mean-square* (MS) continuity and differentiability of GPs. It can be shown that the first n_d MS derivatives of a GP are MS continuous (the GP is said to be n_d-times MS differentiable) if and only if the first $2n_d$ derivatives of the covariance function are continuous at the diagonal point $\vec{\lambda}_1 = \vec{\lambda}_2 = \vec{\lambda}_*$. For a stationary GP this condition reduces to checking the $2n_d$ derivatives of $k(\vec{\tau})$ at $\vec{\tau} = \vec{0}$, and for an isotropic GP checking the $2n_d$ derivatives of $k(\tau)$ at $\tau = 0$. A proof of this result, following [39], is given in Appendix A. It is the smoothness properties of the covariance function at the origin that determine the differentiability of the GP. This result is used in Sec. III B when discussing different functional forms of covariance for use in GPR.

In this section, the effect of the choice of covariance function on the GPR are explored. We consider three aspects that enter the definition of the covariance function:

(A) specifying the distance metric in parameter space g_{ab};

[6] We have yet to define a metric on parameter space with which to take the norm of this vector (see Sec. III A), but all that is required here is that a suitably smooth metric exists.

(B) specifying the functional form of the covariance with distance $k(\tau)$,
(C) and whether or not to include errors σ_n on the training set points.

Stages A and B cannot be completely separated; there exists an arbitrary scaling, α of the distance $\tau \to \alpha\tau$ which can be absorbed into the definition of the covariance, $k(\tau) \to k(\tau/\alpha)$. However, provided the steps are tackled in order, there is no ambiguity.

A. The metric g_{ab}

The first stage involves defining a distance τ between two points in parameter space. One simple way of doing this, and the way used in the SE covariance function in Eq. (17), is to define $\tau^2 = g_{ab}(\vec{\lambda}_1 - \vec{\lambda}_2)^a(\vec{\lambda}_1 - \vec{\lambda}_2)^b$, where g_{ab} are constant hyperparameters. This distance is obviously invariant under a simultaneous translation of $\vec{\lambda}_1 \to \vec{\lambda}_1 + \vec{\Delta}$ and $\vec{\lambda}_2 \to \vec{\lambda}_2 + \vec{\Delta}$; therefore, this defines a stationary GP. For a D-dimensional parameter space, this involves specifying $D(D+1)/2$ hyperparameters g_{ab}.

More complicated distance metrics (with a larger number of hyperparameters) are possible if the condition of stationarity is relaxed, i.e. $g_{ab} \to g_{ab}(\vec{\lambda})$. It was demonstrated by [50] how, given a family of stationary covariance functions, a nonstationary generalization can be constructed. A stationary covariance function can be considered as a kernel function centered at $\vec{\lambda}_1$; $k(\vec{\lambda}_1, \vec{\lambda}_2) \equiv k_{\vec{\lambda}_1}(\vec{\lambda}_2)$. Allowing a different kernel function to be defined at each point $\vec{\lambda}_1$, a new, nonstationary covariance function is $k(\vec{\lambda}_1, \vec{\lambda}_2) = \int \mathrm{d}\vec{u} k_{\vec{u}}(\vec{\lambda}_1) k_{\vec{u}}(\vec{\lambda}_2)$.[7] Applying this procedure to a D-dimensional SE function generates a nonstationary analogue [50]

$$k(\vec{\lambda}_i, \vec{\lambda}_j) = \sigma_f |\mathcal{G}^i|^{1/4} |\mathcal{G}^j|^{1/4} \left| \frac{\mathcal{G}^i + \mathcal{G}^j}{2} \right|^{-1/2} \exp\left(-\frac{1}{2} \mathcal{Q}_{ij} \right), \tag{29}$$

where

$$\mathcal{Q}_{ij} = (\vec{\lambda}_i - \vec{\lambda}_j)^a (\vec{\lambda}_i - \vec{\lambda}_j)^b \left(\frac{\mathcal{G}^i_{ab} + \mathcal{G}^j_{ab}}{2} \right)^{-1}, \tag{30}$$

and $\mathcal{G}^i_{ab} = \mathrm{inv}[g_{ab}(\vec{\lambda}_i)]$ is the inverse of the parameter-space metric at position $\vec{\lambda}_i$. Provided that the metric $g_{ab}(\vec{\lambda})$ is smoothly parametrized this nonstationary SE function retains the smoothness properties discussed earlier.

[7] To see that k is a valid covariance function consider an arbitrary series of points $\{\vec{\lambda}_i\}$, and the sum over training set points $I = \sum_{i,j} a_i a_j k(\vec{\lambda}_i, \vec{\lambda}_j)$; for k to be a valid covariance it is both necessary and sufficient that $I \geq 0$. Using the definition of k gives $I = \int \mathrm{d}\vec{u} \sum_{i,j} a_i a_j k_{\vec{u}}(\vec{\lambda}_i) k_{\vec{u}}(\vec{\lambda}_j) = \int \mathrm{d}\vec{u} (\sum_i a_i k_{\vec{u}}(\vec{\lambda}_i))^2 \geq 0$.

For the interpolation of waveform differences, it is easy to imagine the potential benefits of using nonstationary GPs. For example, in the case of the spin parameter, it could be imagined that the waveform difference considered as a function of the effective spin of the compact objects $\delta h(\chi)$ would vary on long length scales in χ for small values of the spin, but on much shorter scales for larger values of the spin.

The generalization in Eq. (29) involves the inclusion of a large set of additional hyperparameters to characterize how the metric changes over parameter space; for example one possible parameterisation would be the Taylor series

$$g_{ab}(\vec{\lambda}) = g_{ab}(\vec{\lambda}_0) + (\vec{\lambda}^c - \vec{\lambda}^c_0) \left. \frac{\partial g_{ab}(\vec{\lambda})}{\partial \lambda^c} \right|_{\vec{\lambda}=\vec{\lambda}_0} + \ldots \tag{31}$$

with the hyperparameters $g_{ab}(\vec{\lambda}_0)$, $\partial g_{ab}(\vec{\lambda})/\partial\lambda^c$, and so on. As we see below, the inclusion of even a single extra hyperparameter can incur a significant Occam penalty [37] which pushes the training set to favor a simpler choice of covariance function. For this reason we only consider stationary GPs. However, the generalization to a nonstationary GP (perhaps in only a limited number of parameters, e.g., spin) should be investigated further in the future. In making this generalization, one would have to be guided significantly by the prior expectations of which parameters to include and how to parametrize the varying metric.

An alternative to considering nonstationary metrics is instead to try and find new coordinates $\tilde{\lambda} \equiv \tilde{\lambda}(\lambda)$ such that the metric in these coordinates becomes (approximately) stationary. There could be hope for this approach, as a similar problem has been tackled in the context of template placement for GW searches [51]. Here the problem is to find coordinates such that waveform templates placed on a regular grid in these coordinates have a constant overlap with each other. The waveform match can be viewed as defining a metric in parameter space, and hence the desired coordinates make this metric stationary. For a post-Newtonian inspiral signal, a set of chirp-time coordinates were proposed by [52] which make the metric nearly stationary. Metrics have also been calculated for inspiral–merger–ringdown models, for example IMRPhenomB [53]. While it could be possible to adapt the parameter-space metrics already calculated for different approximants for use in template placement algorithms to help in constructing our GPR training sets, we do not consider this approach further here.

Throughout the remainder of this paper the metric components g_{ab} are treated as constant hyperparameters fixed to their optimum values, as discussed in Sec. II.

B. The functional form of $k(\tau)$

The second stage of specifying the covariance function involves choosing the function of distance $k(\tau)$. In general whether a particular function $k(\tau)$ is positive definite (and

hence is a valid covariance function) depends on the dimensionality D of the underlying space (i.e. $\vec{\lambda} \in \mathbb{R}^D$); however, all the functions considered in this section are positive definite for all D. Several choices for $k(\tau)$ are particularly common in the literature. These include the SE covariance function (which has already been introduced), given by

$$k_{\rm SE}(\tau) = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^2\right). \tag{32}$$

The *power-law exponential* (PLE) covariance function is given by

$$k_{\rm PLE}(\tau) = \sigma_f^2 \exp\left(-\frac{1}{2}\tau^\eta\right), \tag{33}$$

where $0 < \eta \leq 2$. The PLE reduces to the SE in the case $\eta = 2$. The *Cauchy* function is given by

$$k_{\rm Cauchy}(\tau) = \frac{\sigma_f^2}{(1+\tau^2/2\eta)^\eta}, \tag{34}$$

where $\eta > 0$. This recovers the SE function in the limit $\eta \to \infty$. And finally, the *Matérn* covariance function is given by [54]

$$k_{\rm Mat}(\tau) = \frac{\sigma_f^2 2^{1-\eta}}{\Gamma(\eta)}(\sqrt{2\eta}\tau)^\eta K_\eta(\sqrt{2\eta}\tau), \tag{35}$$

where $\eta > 1/2$, and K_η is the modified Bessel function of the second kind [55]. In the limit $\eta \to \infty$, the Matérn covariance function also tends to the SE.

Figure 1 shows the functional forms of the covariance functions. They have similar shapes: they all return a finite covariance at zero distance which decreases monotonically with distance and tends to zero as the distance becomes large. In the case of interpolating waveform differences this indicates that the errors in the approximate waveform at two nearby points in parameter space are closely related, whereas the errors at two well separated points are nearly independent. The PLE, Cauchy and Matérn function can all be viewed as attempts to generalize the SE with the inclusion of one extra hyperparameter η, to allow for more flexible GP modeling. All three alternative functions are able to recover the SE in some limiting case, but the Matérn is the most flexible of the three. This can be seen from the discussion of the MS differentiability of GPs given at the beginning of this section.

The SE covariance function is infinitely differentiable at $\tau = 0$, and so the corresponding GP is infinitely MS differentiable. The PLE function is infinitely differentiable at $\tau = 0$ for the SE case when $\eta = 2$, but for all other cases it is not at all MS differentiable. In contrast, the Cauchy function is infinitely differentiable at $\tau = 0$ for all choices of the hyperparameter η. The Matérn function, by contrast, has a variable level of differentiability at $\tau = 0$, controlled via the hyperparameter η [54]. The GP corresponding to the Matérn covariance function in Eq. (35) is n_d-times MS differentiable if and only if $\eta > n_d$. This ability to adjust the differentiability allows the same covariance function to successfully model a wide variety of data. In the process of maximizing the hyperlikelihood for the training set over hyperparameter η, the GP *learns* the (non)smoothness properties favored by the data, and the GPR returns a correspondingly (non)smooth function.

C. The inclusion of noise σ_n

Even the most accurate waveform models $h(\vec{\lambda})$ still contain some error with respect to the unknown true physical signal $h'(\vec{\lambda})$. This could be because the waveform model does not include all of the physics or because it is calculated using a method with finite accuracy. We can account for the error in our training set points by adding a noise variance term $\sigma_f^2\sigma_{n,i}^2$ in the covariance function,

$$k(\vec{\lambda}_i, \vec{\lambda}_j) \to k(\vec{\lambda}_i, \vec{\lambda}_j) + \sigma_f^2\sigma_{n,i}^2\delta_{ij}, \tag{36}$$

which alters the covariance matrix in Eq. (14) correspondingly, but not the expressions in Eq. (22). Here $\sigma_{n,i}$ is the *fractional* error $\|h - h'\|/\|\delta h\|$ in each training set point, where the norm is taken with respect to the inner product in

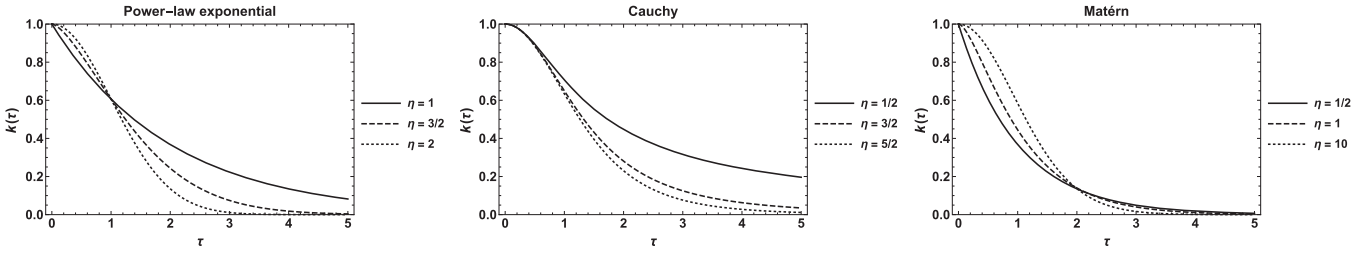

FIG. 1. Plots of the different generalizations of the SE covariance function discussed in Sec. III B. The left-hand panel shows the PLE function, the center panel shows the Cauchy function, and the right-hand panel shows the Matérn function; in all cases the value of σ_f was fixed to unity. In each panel the effect of varying the additional hyperparameter is shown by the three curves. For the PLE covariance the case $\eta = 2$ recovers the SE covariance, while for the Cauchy and Matérn covariances the case $\eta \to \infty$ recovers the SE covariance.

Eq. (5) and $\delta h = H - h$. This ensures that σ_f^2 is still an overall scale for the covariance function.

We do not maximize the hyperlikelihood over $\sigma_{n,i}^2$; this is because $\sigma_{n,i}$ is related to $\|h - h'\|$, which cannot be learned from a training set containing the differences δh. The noise error is instead fixed at some overall error estimate for the accurate model, which is a conservative approach. We consider the simple case $\sigma_{n,i} = \sigma_n$ in this paper; however, it is not necessary for all training set points to have the same error, as a training set might comprise different families of waveform models (e.g., a mix of different variants of (S)EOBNR or IMRPhenom waveforms, or NR waveforms with different numerical resolutions).

If the overall noise error is $\sigma_f \sigma_n$, then the GPR uncertainty at a training set point $\sigma(\vec{\lambda}_i)$ satisfies,

$$\sigma(\vec{\lambda}_i) \leq \sigma_f \sigma_n, \quad \forall i \in \{1, 2, \ldots, N\}. \tag{37}$$

This is because the different points in the training set are assumed to come from a correlated GP, and so nearby measurements also act to decrease the error.

There is also a more practical motivation for the inclusion of noise. Inversion of the covariance matrix in Eq. (14) can pose issues of numerical stability for large training sets. In general, as the number of points in the training set increases, the determinant of the covariance matrix decreases rapidly toward zero, such that the matrix is nearly singular (and hence the matrix is difficult to invert). The solution to this is to add a small fixed noise $\sigma_n^2 = J \ll 1$, or *jitter*, to each training set point as per Eq. (36). The eigenvalues of the new covariance matrix are then (approximately) the eigenvalues of the original matrix plus J. This prevents the determinant, the product of the eigenvalues, from becoming vanishingly small and dramatically improves the stability of the inversion. In effect, we are no longer requiring our interpolant to pass through every training set point; instead, we only ask it to pass close to each point (with the proximity determined by the value of J).

D. Compact support and sparseness

All of the covariance functions considered up until this point have been strictly positive;

$$k(\tau) > 0 \quad \forall\ \tau \in [0, \infty). \tag{38}$$

When evaluating the covariance matrix for the training set K_{ij} this leads to a matrix where all entries are positive; i.e. a dense matrix. When performing the GPR it is necessary to maximize the hyperlikelihood for the training set with respect to the hyperparameters. This process involves inverting the dense matrix K_{ij} at each iteration of the optimization algorithm. Although this procedure is carried out offline, it still can become prohibitive for large training sets. A related problem, as pointed out in Sec. III C is that for large training sets the determinant of the covariance matrix is typically small which also contributes to making the covariance matrix hard to invert.

One potential way around these issues is to consider a covariance function with compact support,

$$\begin{aligned} k(\tau) &> 0 \quad \forall\ \tau \in [0, T], \\ k(\tau) &= 0 \quad \forall\ \tau \in (T, \infty), \end{aligned} \tag{39}$$

where T is some threshold distance beyond which we assume that the waveform differences become uncorrelated. This leads to a sparse, band-diagonal covariance matrix, which is much easier to invert. Care must be taken when specifying the covariance function to ensure that the function is still positive definite (which is required of a GP): if the SE covariance function is truncated, then the matrix formed from the new covariance function is not guaranteed to be positive definite.

Nevertheless, it is possible to construct covariance functions which have the requisite properties and satisfy the compact support condition in Eq. (39). These are typically based on polynomials. We consider a series of polynomials proposed by [56], which we will refer to as the *Wendland* polynomials. These have the property that they are positive definite in $\mathbb{R}^D$ and are $2q$-time differentiable at the origin. Therefore the discrete parameter q is in some sense analogous to the η hyperparameter of the Matérn covariance function in that it controls the smoothness of the GP. Defining β to be

$$\beta = \left\lfloor \frac{D}{2} \right\rfloor + q + 1 \tag{40}$$

and using $\Theta(x)$ to denote the Heaviside step function, the first few Wendland polynomials $k_{D,q}(\tau)$ are given by,

$$k_{D,0}(\tau) = \sigma_f^2 \Theta(1-\tau)(1-\tau)^{\beta}, \tag{41}$$

$$k_{D,1}(\tau) = \sigma_f^2 \Theta(1-\tau)(1-\tau)^{\beta+1}[1 + (\beta+1)\tau], \tag{42}$$

$$\begin{aligned} k_{D,2}(\tau) = \frac{\sigma_f^2}{3} \Theta(1-\tau)(1-\tau)^{\beta+2}[3 + (3\beta+6)\tau \\ + (\beta^2 + 4\beta + 3)\tau^2], \end{aligned} \tag{43}$$

$$\begin{aligned} k_{D,3}(\tau) = \frac{\sigma_f^2}{15} \Theta(1-\tau)(1-\tau)^{\beta+3}[15 + (15\beta+45)\tau \\ + (6\beta^2 + 36\beta + 45)\tau^2 \\ + (\beta^3 + 9\beta^2 + 23\beta + 15)\tau^3]. \end{aligned} \tag{44}$$

These are plotted in Fig. 2. Other types of covariance function with compact support have also been proposed and explored in the literature (e.g., [57–59]), but we do not consider them in this paper.

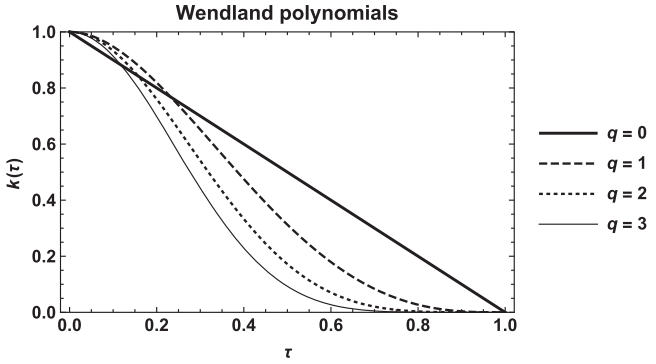

FIG. 2. Plots of the first few Wendland polynomial covariance functions. All these functions have compact support, $k(\tau) = 0$ for $\tau > 1$. As the value of q increases the functions become smoother near the origin.

IV. PROPERTIES OF THE METHOD

In this section proofs of several useful features of the marginalized likelihood are presented. In Sec. IV A we derive the PE error in a linearized formalism, recovering results of [12] as well as new results for our marginalized likelihood; in Sec. IV B we use these results to show that the marginalized likelihood should not exclude the true parameter values even at large SNR, and in Sec. IV C, we derive other limits of the marginalized likelihood at specific points in parameter space.

A. The error at linear order

A more detailed understanding of the theoretical error problem, and the solution offered by the marginalized likelihood can be gained by examining the behavior of the likelihoods in the vicinity of a maximum.

The *exact likelihood*, from Eq. (4), is given by

$$L'(\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\|s - h(\vec{\lambda})\|^2\right), \tag{45}$$

and has a maximum at the best fit parameters, $\vec{\lambda}_{\rm bf}$, which satisfy the equation

$$\langle \partial_a h(\vec{\lambda}_{\rm bf}) | s - h(\vec{\lambda}_{\rm bf})\rangle = 0. \tag{46}$$

The measured data consist of noise and the physical signal with the true parameters, $\vec{\lambda}_{\rm tr}$, that is $s = n + h(\vec{\lambda}_{\rm tr})$. Therefore Eq. (46) becomes

$$\langle \partial_a h(\vec{\lambda}_{\rm bf}) | n + h(\vec{\lambda}_{\rm tr}) - h(\vec{\lambda}_{\rm bf})\rangle = 0. \tag{47}$$

Expanding the difference in the signals to leading order in $\Delta\vec{\lambda} = \vec{\lambda}_{\rm bf} - \vec{\lambda}_{\rm tr}$ gives

$$\langle \partial_a h(\vec{\lambda}_{\rm bf}) | n - \Delta\vec{\lambda}^b \partial_b h(\vec{\lambda}_{\rm bf})\rangle = 0, \tag{48}$$

whence

$$\Delta\vec{\lambda}^a = (\Sigma^{-1})^{ab}\langle n | \partial_b h(\vec{\lambda}_{\rm bf})\rangle, \tag{49}$$

where $\Sigma_{ab} = \langle \partial_a h(\vec{\lambda}_{\rm bf}) | \partial_b h(\vec{\lambda}_{\rm bf})\rangle$. Therefore, at leading order, the shift between the best fit and true parameters for the exact likelihood consists of one term proportional to n; we call this the noise error. The matrix Σ_{ab} is the usual Fisher information matrix (FIM) which characterizes the random errors at leading order [60].

The *approximate likelihood*, from Eq. (7), is given by

$$L(\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\|s - H(\vec{\lambda})\|^2\right), \tag{50}$$

and has a maximum at the best fit parameters which satisfy the equation

$$\langle \partial_a H(\vec{\lambda}_{\rm bf}) | s - H(\vec{\lambda}_{\rm bf})\rangle = 0. \tag{51}$$

Using $s = n + h(\vec{\lambda}_{\rm tr})$ in Eq. (51) and expanding to leading order in $\Delta\vec{\lambda}$ gives

$$\langle \partial_a H(\vec{\lambda}_{\rm bf}) | n - \delta h(\vec{\lambda}_{\rm tr}) - \Delta\vec{\lambda}^b \partial_b H(\vec{\lambda}_{\rm bf})\rangle = 0, \tag{52}$$

thus

$$\Delta\vec{\lambda}^a = (\Gamma^{-1})^{ab}\langle n | \partial_b H(\vec{\lambda}_{\rm bf})\rangle - (\Gamma^{-1})^{ab}\langle \delta h(\vec{\lambda}_{\rm tr}) | \partial_b H(\vec{\lambda}_{\rm bf})\rangle, \tag{53}$$

where $\Gamma_{ab} = \langle \partial_a H(\vec{\lambda}_{\rm bf}) | \partial_b H(\vec{\lambda}_{\rm bf})\rangle$. Therefore, at leading order the shift between the best fit and true parameters for the approximate likelihood consists of two terms: the noise error as before (except with the FIM evaluated with the approximate model) and what we call the model error,

$$\Delta_{\rm model}\vec{\lambda}^a = -(\Gamma^{-1})^{ab}\langle \delta h(\vec{\lambda}_{\rm tr}) | \partial_b H(\vec{\lambda}_{\rm bf})\rangle. \tag{54}$$

The model error is independent of the noise realization, and hence represents a systematic error in the PE associated with using inaccurate models.

The above treatment of the exact and approximate likelihoods is a brief summary of part of the analysis done by [12]. We now apply the same type of analysis to the new marginalized likelihood to see how this reduces or removes the model error.

The *marginalized likelihood* is given in Eq. (9). From Eq. (25) it can be seen that the interpolated waveform difference $\mu(\vec{\lambda})$ is a linear combination of $\delta h(\vec{\lambda}_i)$ from the training set. We will assume, for this calculation only, that the waveform difference is also a small quantity in the sense that $\|\delta h\| \ll \|h\|$ with the norm from Eq. (6). Therefore, $\mu = \mathcal{O}(\delta h)$ and $\sigma_f = \mathcal{O}(\delta h)$. We shall keep contributions up to $\mathcal{O}(\delta h)$.

Under the twin assumptions that $\Delta\vec{\lambda}$ and $\|\delta h\|$ are small, the marginalized likelihood is approximately given by

$$\mathcal{L}(\vec{\lambda}) \approx \exp\left(-\frac{1}{2}\|s - H(\vec{\lambda}) + \mu(\vec{\lambda})\|^2\right). \quad (55)$$

This has a maximum at the best fit parameters $\vec{\lambda}_{\rm bf}$ which satisfy the equation

$$\langle \partial_a(\mu(\vec{\lambda}_{\rm bf}) - H(\vec{\lambda}_{\rm bf}))|s - H(\vec{\lambda}_{\rm bf}) + \mu(\vec{\lambda}_{\rm bf})\rangle = 0. \quad (56)$$

Using $s = n + h(\vec{\lambda}_{\rm tr})$, and expanding to leading order in $\Delta\vec{\lambda}$ and δh gives

$$\langle -\partial_a(\mu(\vec{\lambda}_{\rm bf}) - H(\vec{\lambda}_{\rm bf}))|n + h(\vec{\lambda}_{\rm tr}) - H(\vec{\lambda}_{\rm bf}) + \mu(\vec{\lambda}_{\rm bf})\rangle = 0, \quad (57)$$

$$\langle -\partial_a(\mu(\vec{\lambda}_{\rm bf}) - H(\vec{\lambda}_{\rm bf}))|n - \delta h(\vec{\lambda}_{\rm tr}) - \Delta\vec{\lambda}^b\partial_b H(\vec{\lambda}_{\rm bf}) + \mu(\vec{\lambda}_{\rm bf})\rangle = 0. \quad (58)$$

This expression can be rearranged to find $\Delta\vec{\lambda}$, dropping all terms second order in small quantities,

$$\begin{aligned}\Delta\vec{\lambda}^a = {} & (\Gamma^{-1})^{ab}\langle n|\partial_b(H(\vec{\lambda}_{\rm bf}) - \mu(\vec{\lambda}_{\rm bf}))\rangle \\ & - (\Gamma^{-1})^{ab}\langle \delta h(\vec{\lambda}_{\rm tr})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle \\ & + (\Gamma^{-1})^{ab}\langle \mu(\vec{\lambda}_{\rm bf})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle. \end{aligned} \quad (59)$$

Therefore, at leading order, the shift between the best fit and true parameters for the marginalized likelihood consists of three terms: the noise and model errors from before, and a new shift arising from the marginalization, $\Delta_{\rm marg}\vec{\lambda}^a = (\Gamma^{-1})^{ab}\langle \mu(\vec{\lambda}_{\rm bf})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle$. The expression for the model and marginalization errors are similar and appear with opposite signs (as would be hoped since the marginalized likelihood was designed to remove the model error) so the remaining model error is proportional to $\delta h(\vec{\lambda}_{\rm bf}) - \mu(\vec{\lambda}_{\rm bf})$ (integrated inside the inner product).

If the training set is dense in the region of the peak, and the hyperparameters have been correctly estimated, it is reasonable to assume that the GPR interpolant of the waveform difference performs well, and we have $\langle \delta h(\vec{\lambda}) - \mu(\vec{\lambda})|\cdot\rangle \approx 0$. Under these conditions the marginalized likelihood removes the systematic model error from the parameter estimates. In reality the interpolation is not perfect, and the method is limited by the available information in the training set, so that a residual model error proportional to $\langle \delta h(\vec{\lambda}_{\rm bf}) - \mu(\vec{\lambda}_{\rm bf})|\cdot\rangle$ remains.

B. The limit of large SNR

As first pointed out by [12], the systematic error associated with the inaccurate model used in the approximate likelihood is independent of the SNR, whereas the random error associated with the noise realization decreases with increasing SNR. Therefore, there exists a critical SNR for the approximate likelihood above which the systematic model error dominates the random noise error. If the approximate likelihood is used to infer the parameters of a source with an SNR close to or above this critical value then the inferred parameters are significantly and systematically biased. In this section we examine the behavior of all three likelihood functions for large SNR and show that the marginalized likelihood does *not* suffer from this problem even in the limit of infinite SNR. Therefore, parameter estimates obtained using the marginalized likelihood can always be trusted.

In this section in order to ease the process of taking the limit of large SNR all waveforms are understood to be normalized such that $\|h(\vec{\lambda})\| = 1$, and the amplitude is taken out as a prefactor, so the full signal is $Ah(\vec{\lambda})$. In addition we will assume for simplicity that the measured value of A is equal to the true value for the signal; this has no effect on our final result.

The *exact likelihood* Eq. (4) is given by

$$L'(\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\|s - Ah(\vec{\lambda})\|^2\right). \quad (60)$$

The measured data is given by $s = n + Ah(\vec{\lambda}_{\rm tr})$, and the exact likelihood is peaked at $\vec{\lambda}_{\rm bf} = \vec{\lambda}_{\rm tr} + \Delta\vec{\lambda}$, where [see Eq. (49)]

$$\Delta\vec{\lambda}^a = \frac{1}{A}(\Sigma^{-1})^{ab}\langle n|\partial_b h(\vec{\lambda}_{\rm bf})\rangle. \quad (61)$$

In this section, the FIM Σ_{ab} is defined in terms of the normalized waveforms, i.e. Σ_{ab} is independent of A; this is done so that all of the dependence on A remains explicit. The exact likelihood evaluated on the true parameters is given by

$$L'(\vec{\lambda}_{\rm tr}) \propto \exp\left(-\frac{1}{2}\|n\|^2\right). \quad (62)$$

The exact likelihood evaluated on the best-fit parameters is given by

$$L'(\vec{\lambda}_{\rm bf}) \propto \exp\left[-\frac{1}{2}\|n + A(h(\vec{\lambda}_{\rm tr}) - h(\vec{\lambda}_{\rm bf}))\|^2\right]. \quad (63)$$

The ratio of these two likelihood values is denoted $R_{\rm exact} = L'(\vec{\lambda}_{\rm tr})/L'(\vec{\lambda}_{\rm bf})$. Expanding the difference $h(\vec{\lambda}_{\rm tr}) - h(\vec{\lambda}_{\rm bf})$ in the above equation to leading order in $\Delta\vec{\lambda}$ gives

$$\ln R_{\rm exact} = -\frac{1}{2}(\Sigma^{-1})^{ab}\langle n|\partial_a h(\vec{\lambda}_{\rm bf})\rangle\langle n|\partial_b h(\vec{\lambda}_{\rm bf})\rangle. \quad (64)$$

The quantity $R_{\rm exact}$ is the factor by which the likelihood of the true parameters is suppressed with respect to the peak likelihood. From Eq. (64) it can be seen that this factor is a random variable dependent on the noise realization n; the expectation of this random variable is given by [44]

$$\overline{\ln R_{\rm exact}} = -\frac{1}{2}. \quad (65)$$

Both Eqs. (65) and Eq. (64) are independent of the signal amplitude A, and hence are unchanged by taking the limit of large SNR, $A \to \infty$. Therefore (as one might have expected) the exact likelihood evaluated at $\vec{\lambda}_{\rm tr}$ remains finite in this limit and the true parameters are never completely excluded from the posterior at any value of the SNR.

The *approximate likelihood* Eq. (7) is given by

$$L(\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\|s - AH(\vec{\lambda})\|^2\right), \quad (66)$$

The approximate likelihood is peaked at $\vec{\lambda}_{\rm bf} = \vec{\lambda}_{\rm tr} + \Delta\vec{\lambda}$, where [see Eq. (53)]

$$\Delta\vec{\lambda}^a = \frac{1}{A}(\Gamma^{-1})^{ab}\langle n|\partial_b H(\vec{\lambda}_{\rm bf})\rangle - (\Gamma^{-1})^{ab}\langle \delta h(\vec{\lambda}_{\rm tr})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle. \quad (67)$$

The FIM Γ_{ab} is also here defined to be independent of A. The approximate likelihood evaluated on the true parameters is given by

$$L(\vec{\lambda}_{\rm tr}) \propto \exp\left[-\frac{1}{2}\|n + A(h(\vec{\lambda}_{\rm tr}) - H(\vec{\lambda}_{\rm tr}))\|^2\right] \propto \exp\left(-\frac{1}{2}\|n - A\delta h(\vec{\lambda}_{\rm tr})\|^2\right). \quad (68)$$

The approximate likelihood evaluated on the best fit parameters is given by

$$L(\vec{\lambda}_{\rm bf}) \propto \exp\left[-\frac{1}{2}\|n - A(h(\vec{\lambda}_{\rm tr}) - H(\vec{\lambda}_{\rm bf}))\|^2\right] \propto \exp\left[-\frac{1}{2}\|n + A(\delta h(\vec{\lambda}_{\rm tr}) - \Delta\vec{\lambda}^a\partial_a H(\vec{\lambda}_{\rm bf}))\|^2\right], \quad (69)$$

where, as before, the waveform difference has been expanded to leading order in $\Delta\vec{\lambda}$. The ratio of the two likelihood $R_{\rm approx} = L(\vec{\lambda}_{\rm tr})/L(\vec{\lambda}_{\rm bf})$ can be evaluated from Eq. (68) and Eq. (69), and taking the limit of large SNR gives

$$\lim_{A\to\infty} \ln R_{\rm approx} = -\frac{A^2}{2}(\Gamma^{-1})^{ab}\langle \delta h(\vec{\lambda}_{\rm tr})|\partial_a H(\vec{\lambda}_{\rm bf})\rangle \times \langle \delta h(\vec{\lambda}_{\rm tr})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle. \quad (70)$$

Unlike $R_{\rm exact}$, this ratio does not depend on n. This is because in the limit of large SNR, only the terms from the exponents of Eq. (68) and Eq. (69) proportional to A^2 contribute, and the noise-dependent terms are all proportional to A. Also unlike $R_{\rm exact}$, this ratio does depend on the amplitude and $R_{\rm approx} \to 0$ in the limit of large SNR. Therefore, as anticipated above, the approximate likelihood excludes the true source parameters with complete certainty in the limit of large SNR (unless $\langle \delta h(\vec{\lambda}_{\rm tr})|\cdot\rangle = 0$, in which case the exact likelihood is recovered).

The *marginalized likelihood* Eq. (9) is given by

$$\mathcal{L}(\vec{\lambda}) \propto \exp\left(-\frac{1}{2}\frac{\|s - AH(\vec{\lambda}) + A\mu(\vec{\lambda})\|^2}{1 + A^2\sigma^2(\vec{\lambda})}\right). \quad (71)$$

The marginalized likelihood is peaked at $\vec{\lambda}_{\rm bf} = \vec{\lambda}_{\rm tr} + \Delta\vec{\lambda}$, where, by comparison with Eq. (59),

$$\Delta\vec{\lambda}^a = \frac{1}{A}(\Gamma^{-1})^{ab}\langle n|\partial_b(H(\vec{\lambda}_{\rm bf}) - \mu(\vec{\lambda}_{\rm bf}))\rangle - (\Gamma^{-1})^{ab}\langle \delta h(\vec{\lambda}_{\rm tr})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle + (\Gamma^{-1})^{ab}\langle \mu(\vec{\lambda}_{\rm bf})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle. \quad (72)$$

The values of the marginalized likelihood evaluated on the true and best-fit parameters are given by Eq. (73) and Eq. (74), and the ratio of these two likelihoods is denoted $R_{\rm marg}$,

$$\mathcal{L}(\vec{\lambda}_{\rm tr}) \propto \exp\left(-\frac{1}{2}\frac{\|n - A\delta h(\vec{\lambda}_{\rm tr}) + A\mu(\vec{\lambda}_{\rm tr})\|^2}{1 + A^2\sigma^2(\vec{\lambda}_{\rm tr})}\right); \quad (73)$$

$$\mathcal{L}(\vec{\lambda}_{\rm bf}) \propto \exp\left(-\frac{1}{2}\frac{\|n - A\delta h(\vec{\lambda}_{\rm tr}) - A\Delta\vec{\lambda}^a\partial_a H(\vec{\lambda}_{\rm bf}) + A\mu(\vec{\lambda}_{\rm bf})\|^2}{1 + A^2\sigma^2(\vec{\lambda}_{\rm bf})}\right); \quad (74)$$

$$\lim_{A\to\infty} \ln R_{\rm marg} = -\frac{1}{2\sigma^2(\vec{\lambda}_{\rm bf})}[(\Gamma^{-1})^{ab}\langle \delta h(\vec{\lambda}_{\rm tr}) - \mu(\vec{\lambda}_{\rm bf})|\partial_a H(\vec{\lambda}_{\rm bf})\rangle\langle \delta h(\vec{\lambda}_{\rm tr}) - \mu(\vec{\lambda}_{\rm bf})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle$$
$$- \|\mu(\vec{\lambda}_{\rm bf})\|^2 + \|\mu(\vec{\lambda}_{\rm tr})\|^2 + 2\langle \delta h(\vec{\lambda}_{\rm tr})|\mu(\vec{\lambda}_{\rm tr}) - \mu(\vec{\lambda}_{\rm bf})\rangle] \tag{75}$$

$$\approx -\frac{1}{2\sigma^2(\vec{\lambda}_{\rm bf})}(\Gamma^{-1})^{ab}\langle \delta h(\vec{\lambda}_{\rm tr}) - \mu(\vec{\lambda}_{\rm bf})|\partial_a H(\vec{\lambda}_{\rm bf})\rangle\langle \delta h(\vec{\lambda}_{\rm tr}) - \mu(\vec{\lambda}_{\rm bf})|\partial_b H(\vec{\lambda}_{\rm bf})\rangle. \tag{76}$$

The approximation made in going from Eq. (75) to Eq. (76) involves dropping terms which are products of small quantities. Because the FIM is a symmetric, positive-definite matrix, the numerator in Eq. (76) is a negative number, and hence $R_{\rm marg} < 1$ as required to ensure $\vec{\lambda}_{\rm bf}$ is the peak of the likelihood.

As was the case with $R_{\rm approx}$, this expression for $R_{\rm marg}$ does not depend on the noise. However, unlike $R_{\rm approx}$ the expression for $R_{\rm marg}$ also does not depend on the amplitude A. Therefore, in the limit that the SNR becomes large $R_{\rm marg}$ tends to a constant value which depends quadratically on $\langle \delta h(\vec{\lambda}_{\rm tr}) - \mu(\vec{\lambda}_{\rm bf})|\cdot\rangle$. As the SNR increases, the true parameters are not excluded from the marginalized likelihood, instead the likelihood distribution tends to a constant distribution (i.e. no dependence on n), and the ratio by which the true parameters are disfavored compared to the best fit parameters is set by the ability of the GPR to recover the true waveform difference.

Intuitively, the reason the marginalized likelihood is able to achieve this useful behavior, even if the true waveform difference is not perfectly recovered by the GPR interpolation (i.e. $\langle \delta h(\vec{\lambda}_{\rm tr}) - \mu(\vec{\lambda}_{\rm bf})|\cdot\rangle \neq 0$), is due to the way the hyperparameters in the covariance function are chosen. The hyperparameters were fixed to their optimum values by maximizing the hyperlikelihood for the training set (as described in Sec. II). During this process the overall scale hyperparameter σ_f gains a dependence on the amplitude proportional to A^2. Hence the GPR uncertainty $\sigma^2(\vec{\lambda})$ is also proportional to A^2. As can be seen from Eq. (71), in the limit of large SNR the amplitude dependence cancels in the exponential and the marginalized likelihood tends to a constant distribution. Therefore, the marginalized likelihood never excludes the true source parameters from the final posterior with complete certainty.

C. Limits of the marginalized likelihood across parameter space

In this section we examine the behavior of the marginalized likelihood in the limit of being far from any training points and being at a training point.

First we examine the behavior of the marginalized likelihood in the former case, at a large distance ($\tau^2 \gg 1$) from any of the points in the training set. From Eq. (25) it can be seen that well outside of the training set $\mu(\vec{\lambda}) \to 0$ and $\sigma^2(\vec{\lambda}) \to \sigma_f^2$. Therefore, from Eq. (28), the log marginalized likelihood tends to

$$\ln \mathcal{L}(\vec{\lambda}) \to \frac{\ln L(\vec{\lambda})}{1 + \sum_{i,j} \boldsymbol{K}_{ij}\langle \delta h(\vec{\lambda}_i)|\delta h(\vec{\lambda}_j)\rangle}. \tag{77}$$

Well outside of the training set the marginalized likelihood $\ln \mathcal{L}(\vec{\lambda})$ recovers the standard, approximate likelihood $L(\vec{\lambda})$ up to a constant factor. This constant factor is one plus a linear combination of the overlap integrals of all the waveform differences in the training set. Since the denominator in Eq. (77) is always greater than unity (this is ensured by the positive-definite property of the covariance matrix), it broadens any peak in the likelihood outside of the training set. The amount of the broadening is set by the magnitude of the waveform differences in the training set via the overlap matrix $\langle \delta h(\vec{\lambda}_i)|\delta h(\vec{\lambda}_j)\rangle$. This is the behavior that would be expected; in the absence of any accurate waveforms the parameter uncertainties obtained from the approximate waveforms should be multiplied by a constant factor depending upon our level of belief in the accuracy of the approximate waveform model. In turn, our level of belief in the accuracy of the approximate waveform is learned from the training set in the process of training the GP.

We now consider the behavior of the marginalized likelihood evaluated at one of the training set points $\vec{\lambda}_\ell$. First, consider the case where $\sigma_n = 0$. In this case, the interpolated waveform difference, from Eq. (24), at $\vec{\lambda}_\ell$ recovers the true waveform difference, and the GPR uncertainty, from Eq. (25), vanishes at $\vec{\lambda}_\ell$;

$$\mu(\vec{\lambda}_\ell) = \delta h(\vec{\lambda}_\ell), \tag{78}$$

$$\sigma^2(\vec{\lambda}_\ell) = 0. \tag{79}$$

Therefore the marginalized likelihood in Eq. (28) recovers the exact likelihood with no additional broadening.

$$\mathcal{L}(\vec{\lambda}_\ell) = L'(\vec{\lambda}_\ell). \tag{80}$$

This is also the behavior that would be expected; at a point in parameter space where the accurate waveform is known, the accurate likelihood is recovered.

If $\sigma_n \neq 0$, then Eq. (78) and Eq. (79) become

$$\mu(\vec{\lambda}_\ell) = \delta h(\vec{\lambda}_\ell) - \sigma_n^2 \sum_i k(\vec{\lambda}_i, \vec{\lambda}_\ell)\delta h(\vec{\lambda}_i) + \mathcal{O}(\sigma_n^4),$$
$$\sigma^2(\vec{\lambda}_\ell) = \sigma_n^2 \sum_i k(\vec{\lambda}_i, \vec{\lambda}_\ell)k(\vec{\lambda}_i, \vec{\lambda}_\ell) + \mathcal{O}(\sigma_n^4). \quad (81)$$

In this case any peak in the marginalized likelihood will be slightly shifted and broadened relative to the peak in the accurate likelihood by an amount consistent with the uncertainty σ_n in the accurate waveform model.

V. IMPLEMENTATION

In this section we present an illustrative implementation of our GPR approach. As a simple example, we consider estimating a single parameter; a full multidimensional application that would be appropriate for actual GW data analysis will be investigated in future work. We begin in Sec. V A by introducing the waveforms we use for this study. In Sec. V B we describe the placement of the training set points for the GPR; in order to investigate the effect of training set on the GPR interpolant two sets were constructed with different numbers of points and grid spacings. In Sec. V C we present results for maximizing the hyperlikelihood to find the optimum hyperparameters, $\vec{\theta}_{\rm op}$, for the interpolation; this is done for a range of different covariance functions on each of the training sets described in Sec. V B. In Sec. V D we interpolate the waveforms across parameter space for the different training sets and different covariance functions described and compare the interpolated waveforms $H(\vec{\lambda}) - \mu(\vec{\lambda})$ to the accurate waveforms $h(\vec{\lambda})$. In Sec. V E we present results for the GPR uncertainty, $\sigma^2(\vec{\lambda})$, for the different training sets and different covariance functions considered. Finally in Sec. V F we present results for the marginalized likelihood $\mathcal{L}(\vec{\lambda})$, and compare with the results obtained using the approximate likelihood $L(\vec{\lambda})$, and the exact likelihood $L'(\vec{\lambda})$.

A. Model waveforms

In order to implement the GPR, a choice has to be made regarding which waveform models to use. The method uses two waveform approximants; the accurate $h(\vec{\lambda})$ and the approximate $H(\vec{\lambda})$ waveforms. The accurate waveform should be the most accurate available at a computational cost that permits the offline construction of the training set $\mathcal{D}$. The criteria for choosing the approximate waveform is less clear, a balance needs to be struck between accuracy and speed. If the model is computationally cheap but not accurate enough the waveform difference, $\delta h(\vec{\lambda}) = H(\vec{\lambda}) - h(\vec{\lambda})$, will be large and vary on short length scales over parameter space; these are the situations which will cause the GPR to perform worst. On the other hand an accurate model which is too computationally expensive could slow down any PE to such an extent that there ceases to be any benefit in using the marginalized likelihood instead of the accurate likelihood.

We used two waveform models implemented in the LIGO Scientific Collaboration Algorithm Library (LAL).[8] As our intention here is to provide a proof of principle, we choose the IMRPhenomC approximant [61] as the accurate waveform and the widely used TaylorF2 approximant [33,62,63] as the approximate waveform; both of these models are sufficiently fast to evaluate that we can compute and then compare the three likelihoods [accurate $L'(\vec{\lambda})$, approximate $L(\vec{\lambda})$, and marginalized $\mathcal{L}(\vec{\lambda})$] and directly assess the performance of the GPR.

Both of the approximants we have chosen to use here are frequency-domain models, i.e. they naturally return the waveform in the Fourier domain $\tilde{h}(f)$.[9] The IMRPhenomC waveform includes inspiral, merger and ringdown, while the TaylorF2 waveform only includes the inspiral.

We investigate the merger of nonspinning circular binaries. This limits the number of intrinsic parameters describing the system to two, the masses of the two component objects, $\vec{\lambda} = \{m_1, m_2\}$. To further simplify the problem we place training set points only along a one-dimensional subspace, which we choose to be a surface of constant mass ratio, $Q = m_2/m_1$ (with $m_1 \geq m_2$), parametrised by the value of the chirp mass $\mathcal{M}_{\rm c} = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$. This keeps the size of the training set small, and hence the computational complexity of the GPR to a minimum. This allows us to instead focus our attention on the novel features of the marginalized likelihood, and explore the effect of changing various features of the method.

B. The training set

For simplicity we restrict the range of the coordinates which we search over to reduce the computational complexity. This is again to allow us to focus our attention on the novel features of the method. The training sets cover the chirp mass in the range $\mathcal{M}_{\rm c} \in [5, 5.6]M_\odot$ and the mass ratio is fixed to the (nearly equal mass) $Q = 0.75$. The placement of training set points was done as a regular grid in chirp mass with a step size between points of $\Delta\mathcal{M}_{\rm c}$.

[8] http://www.lsc-group.phys.uwm.edu/lal

[9] In previous work [13] the marginalized likelihood has been implemented with time-domain approximants. The method works equally with frequency-domain or time-domain models without the need to transform between them. In the offline stage the waveforms enter only via the overlap matrix $\langle \delta h(\vec{\lambda}_i) | \delta h(\vec{\lambda}_j) \rangle$, and in the online stage the waveforms enter only in the linear combination for $\mu(\vec{\lambda})$ in Eq. (24), which commutes with the operation of taking the Fourier transform.

TABLE I. The properties describing the positions of the template waveforms for each of the three training sets used.

	$\Delta\mathcal{M}_c$	N
$\mathcal{D}_0$	$1.0 \times 10^{-2} M_\odot$	60
$\mathcal{D}_1$	$5.0 \times 10^{-3} M_\odot$	120

The chirp-mass range has been chosen to demonstrate the properties of the method. For lower masses, the signal is dominated by the inspiral where both approximants agree well. Therefore, interpolating these small differences would not be a robust test. At higher masses, where the signal is just merger and ringdown, the two approximants are completely different; we get no useful information from the TaylorF2 waveform and may as well interpolate IMRPhenomC directly. We do not anticipate that in practice we would consider waveform differences as significant as the complete absence of merger and ringdown; hence, this example, although only one-dimensional, should be a rigorous test of what waveform uncertainties can be successfully interpolated. Understanding if this continues to be true for the interpolation of a more intricate waveform difference across a higher-dimensional parameter space, must wait for further studies to be completed.

To allow us to explore the effect that the density of points in the training set has on the marginalized likelihood two different values for $\Delta\mathcal{M}_c$ were considered. This leads to two different training sets whose total number n of points are different; the properties of these two training sets are summarized in Table I. It is expected that the GPR interpolation, and hence the marginalized likelihood, will perform better when using the denser set $\mathcal{D}_1$.

Once the training set points $\{\vec{\lambda}_i\}$ were specified, both the approximate $H(\vec{\lambda}_i)$ and accurate $h(\vec{\lambda}_i)$ waveforms discussed in Sec. V A were evaluated at each point, and the waveform differences $\{\delta h(\vec{\lambda}_i)\}$ stored for use during the GPR interpolation. The matrix of waveform difference overlaps $M_{ij} = \langle \delta h(\vec{\lambda}_i) | \delta h(\vec{\lambda}_j) \rangle$ was also evaluated and stored for use during the hyperlikelihood maximization procedure.

C. The hyperparameters

Initially the training sets described in Sec. V B were interpolated using the SE covariance function in Eq. (32). This covariance function has just two hyperparameters, $\vec{\theta} = \{\sigma_f, g_{\mathcal{M}_c\mathcal{M}_c}\}$. The one-dimensional metric $g_{\mathcal{M}_c\mathcal{M}_c}$ can be exchanged for a length scale in the chirp mass parameter $\delta\mathcal{M}_c \equiv 1/\sqrt{g_{\mathcal{M}_c\mathcal{M}_c}}$. A fixed noise term with $\sigma_n^2 = 10^{-4}$ was used for all the covariance functions in this section, to make the inverse of the covariance function numerically stable as discussed in Sec. III C. The hyperlikelihood for the training set $\mathcal{D}_0$ was maximized with respect to these two hyperparameters. The optimum values for the hyperparameters were found to be

$$\sigma_f = 3.49 \times 10^4, \tag{82}$$

$$\delta\mathcal{M}_c = 1.11 \times 10^{-2} M_\odot. \tag{83}$$

The hyperlikelihood is shown in Fig. 3. The hyperlikelihood was also maximized for the training set $\mathcal{D}_1$ using the same SE covariance function and those results are also shown in Fig. 3. For the denser training set $\mathcal{D}_1$ the optimum length scale was found to be smaller, $\delta\mathcal{M}_c = 6.31 \times 10^{-3} M_\odot$. For both training sets, in the limit that the length scale becomes much larger than the total width of the training set ($0.6 M_\odot$) or much smaller than the grid point spacing ($\Delta\mathcal{M}_c$), the hyperlikelihood tends to a constant value. This behavior can be understood by examining the expression for the hyperlikelihood in Eq. (18).

In order to explore the effect that the choice of covariance function has on the marginalized likelihood, the training sets were also interpolated using the Matérn covariance function in Eq. (35). This covariance function has an additional hyperparameter, $\vec{\theta} = \{\sigma_f, g_{\mathcal{M}_c\mathcal{M}_c}, \eta\}$. The hyperlikelihood for training set $\mathcal{D}_0$ was maximized for this covariance function. It was found that the hyperlikelihood surface did not possess a peak, instead a ridge was found tending to a maximum at a value $\eta \to \infty$, and values of σ_f and $g_{\mathcal{M}_c\mathcal{M}_c}$ were found to be the same as for the SE covariance function. In Fig. 4 we plot the log-hyperlikelihood (maximized over σ_f) against chirp-mass length scale and the additional hyperparameter η.

As the Matérn covariance function recovers the SE function in the limit $\eta \to \infty$, there will be no difference in the performance of the interpolants for this training set when using the Matérn or SE covariance functions. If the volume under the hyperlikelihood surface (the hyperevidence) is used as a figure-of-merit for which covariance function the data favors, then in this case the data is equally well described by either covariance function, but the SE covariance function is favored over the Matérn due to the smaller prior volume (the Occam penalty).

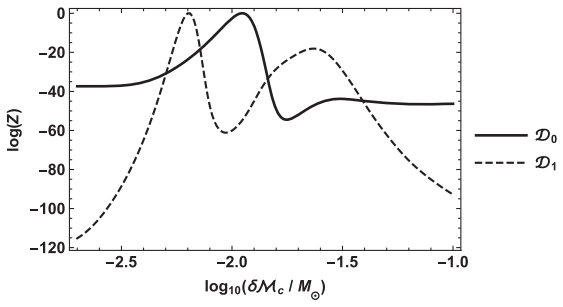

FIG. 3. The hyperlikelihood, from Eq. (18), for the SE covariance function, maximized over the scale hyperparameter σ_f, plotted against the chirp mass length scale $\delta\mathcal{M}_c$. The hyperlikelihood is shown for both of the training sets (normalized to a peak value of 1). The denser training set $\mathcal{D}_1$ was found to favor smaller length scales.

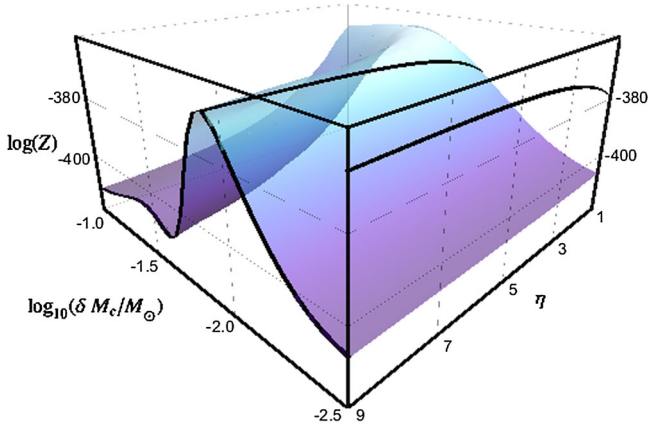

FIG. 4. The hyperlikelihood, from Eq. (18), surface for the training set $\mathcal{D}_0$ using the Matérn covariance, maximized over the hyperparameter σ_f, plotted against the chirp mass length scale $\delta\mathcal{M}_c$ and the hyperparameter, η. The hyperlikelihood does not show a clear peak, instead a ridge in the hyperparameter space favors the limiting case $\eta \to \infty$, in which limit the Matérn covariance function is equal to the SE covariance function. On the near-side faces of the plot box we show the hyperlikelihood sliced parallel to the coordinate axes though the point $(\delta\mathcal{M}_c = 10^{-1.9}M_\odot, \eta = 10)$. The solid black line on the near, left-hand face of the box very closely matches the solid black curve in Fig. 3 (up to an arbitrary additive constant).

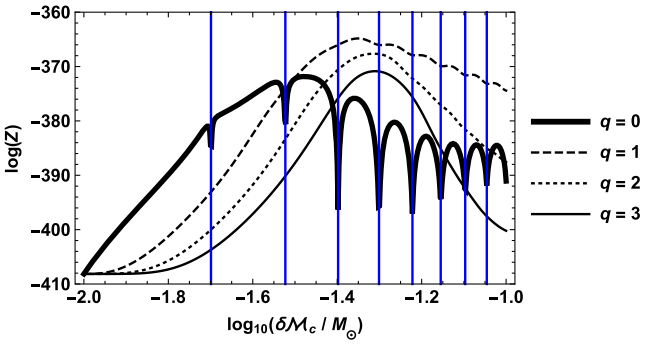

FIG. 5. The hyperlikelihood, from Eq. (18), for the training set $\mathcal{D}_0$ using the Wendland polynomial covariance functions, maximized over the scale hyperparameter σ_f, plotted against the chirp-mass length scale $\delta\mathcal{M}_c$. The vertical blue lines indicate multiples of the training-set grid spacing $\Delta\mathcal{M}_c$.

The hyperlikelihood was also calculated for both training sets $\mathcal{D}_0$ and $\mathcal{D}_1$ using the PLE covariance, see Eq. (33), and the Cauchy covariance, see Eq. (34), considered in Sec. III. In both cases a similar behavior was observed. For the PLE covariance, a peak in the hyperlikelihood was found at $\eta = 2$, where the PLE covariance equals the SE covariance. For the Cauchy covariance, a ridge in the hyperlikelihood was found tending to a maximum for $\eta \to \infty$ (similar to the Matérn case shown in Fig. 4), in which limit the Cauchy covariance also recovers the SE covariance. As with the Matérn covariance, if the hyperlikelihood is used as a figure-of-merit for selecting the covariance function then the SE covariance is favored over both the PLE and Cauchy functions due to the Occam penalty.

It is clear that interpolations of the training sets $\mathcal{D}_0$ and $\mathcal{D}_1$ using any of the PLE, Cauchy, or Matérn covariance functions, evaluated at the hyperlikelihood-maximizing hyperparameters, would yield identical results to an interpolation using the simpler SE covariance. For this reason, in the following sections we do not use the PLE, Cauchy, or Matérn functions further and instead focus on the SE covariance function. We will, however, also consider using the Wendland polynomial function in the following sections as it reduces the computational cost.

The hyperlikelihood for the compact support Wendland polynomial covariance functions are shown in Fig. 4, for the cases $q = 0, 1, 2, 3$. The compact-support functions can develop multiple-peaks in the hyperlikelihood surface associated with the length-scale of the training set: multiples of the training-set grid spacing are indicated with vertical blue lines in Fig. 4. These subsidiary peaks occur in the $\delta\mathcal{M}_c$ hyperparameter because as the size of the compact-support region grows, the (integer) number of training set points it contains changes discontinuously.

From Fig. 5 it can be seen that for the training set $\mathcal{D}_0$, a value of $q = 1$ is favored with a length-scale $\delta\mathcal{M}_c = 4.37 \times 10^{-2}M_\odot$. In the following sections we will use Wendland covariance function with all values of q (and their associated peak hyperlikelihood length scales) to interpolate $\mathcal{D}_0$.

The optimum hyperparameters depend on the detector noise power spectral density via the overlap matrix $\langle\delta h(\vec{\lambda}_i)|\delta h(\vec{\lambda}_j)\rangle$. In Appendix B, an investigation of the sensitivity of the optimum hyperparameters to small changes in the detector noise properties is described. It was found that for any realistic changes to the noise curve, the optimum hyperparameters were changed by an amount too small to have any noticeable effect on the interpolant.

D. The interpolated waveforms

The GPR waveform $H(\vec{\lambda}) - \mu(\vec{\lambda})$ could be viewed as a new waveform approximant formed from the approximant waveforms and the use of GPR on the training set of accurate waveforms. It is then natural to ask how this new approximant compares to the original ones. This can be assessed by calculating the overlap between the different waveforms, where the overlap is defined by

$$\text{overlap}(a, b) = \frac{\langle a|b\rangle}{\|a\|\|b\|}, \tag{84}$$

using the inner product defined in Eq. (5).

Only considering the overlap misses the important extra benefit which the marginalized likelihood approach brings. Our method is not just supplying a new waveform approximant, but also providing a way of modifying the posterior to account for the uncertainties known to be in the approximant. This extra information which modifies the

likelihood surface is included through $\sigma(\vec{\lambda})$. Nonetheless, it is still informative to temporarily treat $H(\vec{\lambda}) - \mu(\vec{\lambda})$ as if it were a new waveform approximant and see how it compares with the approximants $h(\vec{\lambda})$ and $H(\vec{\lambda})$ from which it was built. Figure 6 shows the waveform overlap between the interpolated waveform $H(\vec{\lambda}) - \mu(\vec{\lambda})$ and the accurate waveform $h(\vec{\lambda})$ as a function of chirp mass near the edge of the training set. Also shown in the dotted curve is the overlap between the approximate waveform $H(\vec{\lambda})$ and the accurate waveform $h(\vec{\lambda})$. The interpolated waveforms have a much higher overlap than the approximate waveforms, as would be expected. Within the training set the overlap is increased from ∼0.35 to no less than ∼0.985 even for the sparser training set D_0. For the denser training set $\mathcal{D}_1$ overlaps no worse than ∼0.999 were found inside the range of the training set. Outside the training set the interpolated waveform tends rapidly to the approximate waveform $H(\vec{\lambda})$.

The training set waveforms were also interpolated using the Wendland compact support covariance functions discussed in Sec. III D. The cases $q = 0$, 1, 2, 3 were considered separately. The waveform overlap using these

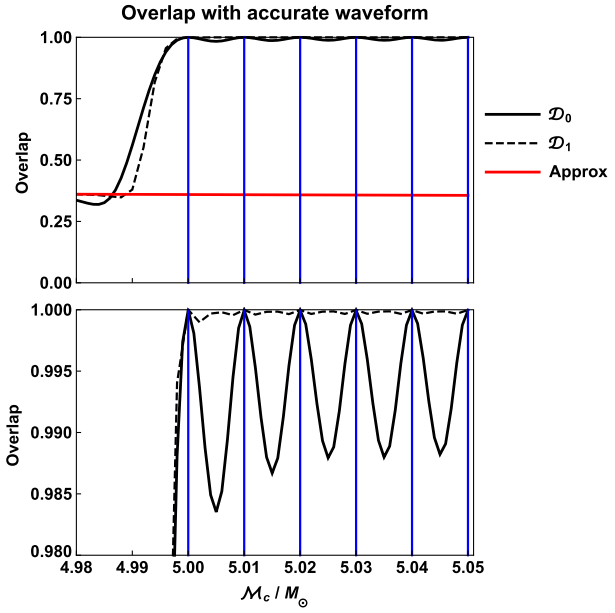

FIG. 6. A plot of the overlap between the interpolated waveform $H(\vec{\lambda}) - \mu(\vec{\lambda})$ and the accurate waveform $h(\vec{\lambda})$ as a function of the chirp mass $\mathcal{M}_c$. The bottom panel is the same plot with a different ordinate axis scale. The two black lines show the overlap using both training sets, $\mathcal{D}_0$ and $\mathcal{D}_1$, interpolated using the SE covariance function. The red line shows the overlap between the approximate waveform $H(\vec{\lambda})$ and the accurate waveform $h(\vec{\lambda})$ for comparison. The vertical blue lines show the position of the training set points for $\mathcal{D}_0$. In the bottom panel, it can be seen that, for either interpolant, the overlap becomes one when evaluated at the training set points.

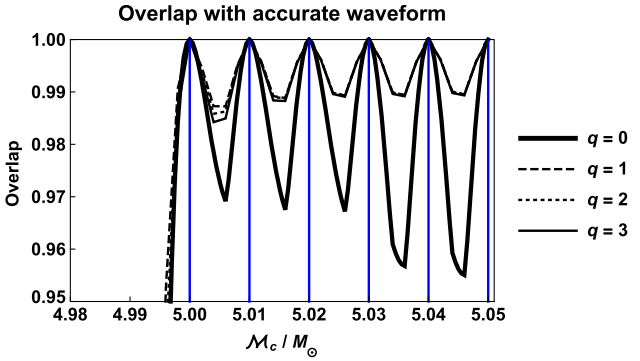

FIG. 7. A plot of the overlap (or overlap) between the interpolated waveform $H(\vec{\lambda}) - \mu(\vec{\lambda})$ and the accurate waveform $h(\vec{\lambda})$ as a function of the chirp mass $\mathcal{M}_c$. The different curves correspond to using the Wendland polynomial covariance functions with different values of q to interpolate the training set $\mathcal{D}_0$. The vertical blue lines show the position of the training set points for $\mathcal{D}_0$.

interpolants is plotted in Fig. 7. The performance of these interpolants should be compared with the results using the SE covariance function in Fig. 6.

The least smooth of the Wendland polynomials, the $q = 0$ case, performs noticeably worse than the SE covariance; inside the training set the overlap drops as low as ∼0.955 compared to ∼0.985 for the SE. However, even a overlap of ∼0.955 is still a great improvement over the overlap of ∼0.35 for the approximate waveform alone. For the $q = 0$ Wendland polynomial the interpolant has a discontinuous first derivative, which can be seen in Fig. 7 (this is expected and was discussed in Sec. III and in detail in Appendix A). The higher values of q have discontinuities in the higher ordered derivatives, but these curves look smooth to the eye. The smoother Wendland polynomials, with $q > 0$, all perform very similarly to the SE covariance function; inside the training set the overlap drops as low as ∼0.985 for the $q = 2$ interpolant.

E. The GPR uncertainty

The GPR performs an interpolation of the points in the training set and naturally returns a Gaussian error $\sigma(\vec{\lambda})$, see Eq. (25), for each interpolated point. In our present one-dimensional interpolation this is simply a function of $\mathcal{M}_c$. A small section of this curve taken from the edge of the training set is shown in Fig. 8. Inside the training set, the error surface has a regular, periodic pattern with minima at the training set points and maxima in between. This regularity is because the GP used for the interpolation is stationary, the training-set points used are regularly spaced, and each point has an identical error (a jitter $J = 10^{-4}$). If these conditions were to be relaxed, then the error surface would become more complicated. In general, a larger $\sigma(\vec{\lambda})$ indicates greater theoretical uncertainty and highlights regions where we would benefit from additional accurate

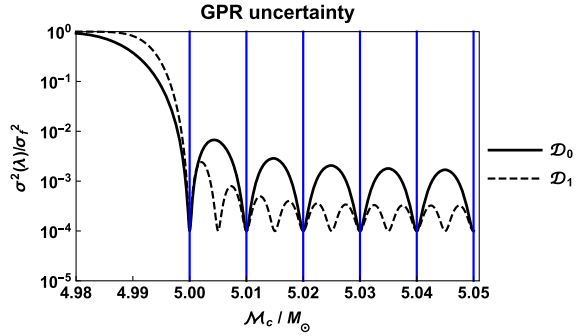

FIG. 8. A plot of the GPR uncertainty $\sigma^2(\vec{\lambda})$ as a function of the chirp mass parameter for both of the training sets, using the SE covariance function. The vertical blue lines show the position of the training set points for $\mathcal{D}_0$. Outside of the training set the uncertainty tends to a constant σ_f^2. Inside the training sets the error is approximately periodic with minima at the training set points. The maximum uncertainty inside the training set is smaller for the denser training sets.

waveforms (e.g., where it would be beneficial to perform more NR simulations).

Near the edge of the training set the behavior becomes less regular and well outside of the training set the error tends to a constant value, $\sigma^2(\vec{\lambda}) \to \sigma_f$ as $\vec{\lambda} \to \infty$. This behavior is seen in Fig. 8 for all three training sets. The training sets with smaller grid spacings have smaller uncertainties everywhere in parameter space.

The GPR uncertainty was also calculated using the Wendland polynomial covariance functions to interpolate the training set $\mathcal{D}_0$; these are shown in Fig. 9. The GPR uncertainty, expressed as a fraction of σ_f^2, is largest for the smallest values of q; this can be traced back to the optimum length scale for the Wendland polynomials increasing with q (see Fig. 5). This means that the uncertainty grows more slowly as the interpolating point moves away from the training set points, and hence reaches a smaller maximum value between training set points. The smoother ($q > 0$) Wendland polynomials perform similarly to the SE covariance function, in the sense that both the GPR interpolants

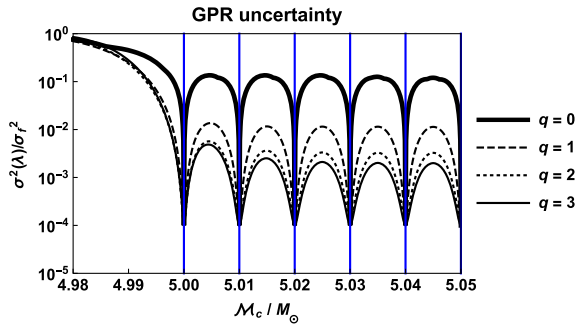

FIG. 9. A plot of the GPR uncertainty $\sigma^2(\vec{\lambda})$ as a function of the chirp mass parameter for the training set $\mathcal{D}_0$, using the Wendland polynomial covariance functions. The vertical blue lines show the position of the training set points.

(which we quantify via the overlap) and the GPR uncertainties are almost identical. Hence, in the following sections we will only consider using the SE covariance function; the high q Wendland polynomials would yield identical results.

F. The likelihood

Finally we put together the interpolated waveform $H(\vec{\lambda}) - \mu(\vec{\lambda})$ and the GPR uncertainty $\sigma^2(\vec{\lambda})$ to give the marginalized likelihood in Eq. (28). We compare the performance of the marginalized likelihood $\mathcal{L}(\vec{\lambda})$ to the approximate likelihood $L(\vec{\lambda})$ and the accurate likelihood $L'(\vec{\lambda})$. For the injected signal we use the accurate waveform $h(\vec{\lambda})$. We also consider the case where the noise realization is zero (the most likely realization), this makes comparisons easier.

We injected a signal at a chirp mass of $\mathcal{M}_c = 5.045 M_\odot$; this is inside the training set $\mathcal{D}_0$ and midway between training set points. Injecting the signal midway between the points is conservative as this is the point at which we would expect the marginalized likelihood to perform worst. The three different likelihoods were evaluated as a function of chirp mass (all other parameters set to the injected values). This was done at a range of SNRs and the results are shown in Fig. 10. The top row of panels in Fig. 10 show the likelihoods renormalized to a peak value of one, this makes the relative positions of the peaks clear and easy to compare. The bottom row of panels shows the log-likelihood without any renormalization, this illustrates how the approximate likelihood is suppressed relative to the true likelihood (the detection problem discussed in Sec. I).

The exact likelihood $L'(\vec{\lambda})$ is always peaked at the injected value of the chirp mass (because the injected noise realization is zero) and the width of the peak decreases with increasing SNR. The approximate likelihood $L(\vec{\lambda})$ is peaked way from the true value, indicating a systematic error of $\Delta_{\rm sys}\mathcal{M}_c = 5.2 \times 10^{-3} M_\odot$. The width of the approximate likelihood peak also decreases with increasing SNR and for SNR $\gtrsim 12$ (which is also roughly the detection threshold [5,35]) the true parameters are excluded at increasing significance. The bottom row of panels in Fig. 10 shows that the approximate likelihood is suppressed by a significant amount, for a typical SNR of 16 it is suppressed by 80 in log relative to the exact likelihood; this reduces the Bayesian evidence for a detection. The factor by which the approximate likelihood is suppressed increases exponentially with SNR. Finally, the marginalized likelihood is peaked much closer to the exact likelihood: the systematic error is reduced to $\Delta_{\rm sys}\mathcal{M}_c = 9.0 \times 10^{-4} M_\odot$. However, as discussed in Sec. IV B, the peak in the marginalized likelihood does not continually narrow as the SNR increases; for SNR $\gtrsim 30$ the width becomes constant. Consequently, the true

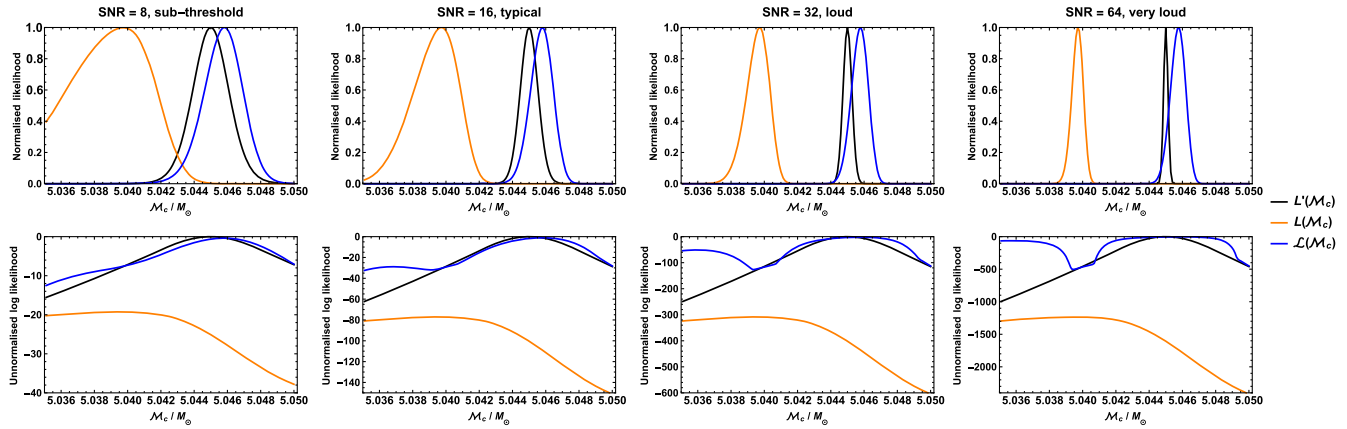

FIG. 10. A plot of the different likelihoods for a variety of SNRs. Vertical lines indicate the position of training set points. The top row of panels show the likelihood normalized to the same peak value; this makes the peak positions clear and shows how the marginalized likelihood tackles the parameter estimation problem associated with the inaccurate models. The bottom row of panels shows the log-likelihood; this makes the suppression of the peak value of the approximate likelihood clear and shows how the marginalized likelihood could be used to tackle the detection problem.

parameters are never excluded at high significance; in the limit of infinite SNR the true parameters lie at the $\sim 1\sigma$ level. The bottom panel of Fig. 10 shows that the marginalized likelihood is not suppressed relative to the exact likelihood in the vicinity of the peak.

Comparing the properly normalized likelihoods, we see that the marginalized and exact likelihoods roughly agree at low SNR. As the SNR is increased, the marginalized likelihood deviates from the exact likelihood and develops oscillatory behavior with period equal to the training set point spacing. In the limit of low SNR, all of the parameter estimation uncertainty comes from the noise, but as the SNR increases, the relative size of this statistical uncertainty becomes smaller and at high SNR we are dominated by model uncertainty. The marginalized likelihood correctly encapsulates this behavior, as can be seen in the sequence from left to right in Fig. 10.

VI. SUMMARY

In [13], some of the authors suggested GPR as a means of incorporating theoretical uncertainty into GW data analysis. We have now thoroughly investigated the properties of this method, elucidating considerations for a practical implementation. A detailed derivation of the marginalized likelihood, and the use of GPR to interpolate model error was presented in Sec. II. GPR is nonparametric, in the sense that only the functional form of the covariance function is specified by hand, with its hyperparameters then learned from the training set, making it well suited to modeling theoretical uncertainty. The expression for the marginalized likelihood derived in Sec. II made some assumptions about the frequency covariance of the waveforms in the training set; in particular it was assumed that at a particular point in parameter space, the model error is highly correlated in frequency. These assumptions may prove to be too restrictive in the future, and could be relaxed by simultaneously performing GPR interpolation in frequency (as well as in parameter space) on the training set waveforms; this would have the added advantage of allowing the inclusion of waveforms with different frequency samplings, which may be beneficial when using multiple waveform approximants and NR waveforms from multiple sources.

The choice of covariance function is central to GPR as it encodes our prior beliefs about the function space that we are interpolating. We discussed various choices of covariance function in Sec. III. We have found that the simple SE covariance function (as used in [13]) performs as well as more complicated alternatives, at least for the relatively small one dimensional training sets considered here. The compact-support Wendland covariance functions with large q were found to perform comparably to the SE, but offer the additional advantage of reduced computational cost. This makes them appealing for future work involving larger training sets.

We proved a number of properties for our marginalized likelihood in Sec. IV, in particular its limiting behavior for large signal amplitude (where the theoretical errors are known to be most significant [12]) and its limiting behavior both far from and near a point in the training set. In the discussion of the latter, the linearized results previously obtained in [12] were recovered. All of these properties demonstrate the suitability of GPR for making robust inferences. The marginalized likelihood successfully describes our belief in our inferences, including our uncertainty in waveform templates.

In Sec. V, we presented a one-dimensional implementation of our marginalized likelihood and demonstrated that it offers an improvement in PE accuracy. For this, we chose two inexpensive waveforms to aid computation; in real data analysis situations, we expect more accurate waveforms

(including those calculated using NR) to be used in the training. However, this choice of waveforms does illustrate the efficacy of the new marginalized likelihood. In particular we find that even when using qualitatively different waveforms (inspiral-only TaylorF2 compared to inspiral–merger–ringdown IMRPhenomC), waveform matches as high as ∼98.5% can be obtained in the mass range we considered. Restricting ourselves to the simple case of one-dimensional PE, we explored various possibilities for GPR. In particular, the effect of different training set sizes was examined; as expected, the performance of the marginalized likelihood is improved by using denser training sets. Additionally, the impact of varying the SNR of the injected waveform was studied. In the standard likelihood model, errors become more severe as the SNR is increased, but we confirmed that even in the limit of large SNR, the marginalized likelihood remained consistent with the injected parameters. We expect these results to carry over to a full multidimensional analysis, which is the next step in developing this technique.

A possibly complementary use for the GPR approximant is as a less expensive alternative to the accurate waveform for more expedient PE; however, there exist other means of constructing computationally inexpensive waveforms, such as reduced-order modeling [64,65]. The advantage of GPR is that it not only supplies an interpolant, but also gives an uncertainty, which can be used to gauge accuracy away from training points. A second, related feature is that GPR naturally allows for uncertainty to be included in the accurate models that the interpolant is calibrated against. It is the ability of GPR to include theoretical uncertainty that makes it attractive for GW astronomy, that this can be done without significant online cost is a welcome bonus.

In conclusion, marginalizing over waveform uncertainty is a robust and effective method of accounting for theoretical error in both PE and detection problems. GPR is a natural and effective means of performing this marginalization. The marginalized likelihood is naturally inferior to a likelihood calculated with more accurate (but inevitably more computationally expensive) waveforms, but it offers significantly improved performance over the standard likelihood calculated with cheap waveforms. In addition, the marginalized likelihood is almost as quick to evaluate online as the standard likelihood, although there is additional offline computation required to construct the training set and train the Gaussian process.

ACKNOWLEDGMENTS

C. J. M. and C. P. L. B. are supported by the STFC. A. J. K. C.'s work is supported by the Cambridge Commonwealth, European and International Trust. J. R. G.'s work is supported by the Royal Society. We thank Will Farr, Alberto Vecchio, Ilya Mandel, Ben Farr, Daniel Holz and the Compact Binary Coalescence PE Group for useful discussions. This work made use of computational resources provided by the Leonard E. Parker Center for Gravitation, Cosmology and Astrophysics at University of Wisconsin–Milwaukee. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation and operates under cooperative agreement PHY-0757058. This document has been assigned LIGO document reference LIGO-P1500162.

APPENDIX A: CONTINUITY AND DIFFERENTIABILITY OF GPS

In this appendix we give proofs of the results stated in Sec. III concerning the continuity and differentiability of GPs, following the approach of [39]. Let $\vec{\lambda}_1, \vec{\lambda}_2, \vec{\lambda}_3 \ldots$ be a sequence of points in parameter space which converges to a point $\vec{\lambda}_*$, in the sense $\lim_{\ell\to\infty}|\vec{\lambda}_\ell - \vec{\lambda}_*| = 0$, where, as in Sec. III, $|\vec{x}|$ denotes the norm with respect to the metric on parameter space, as discussed in Sec. III A. The GP $Y(\vec{\lambda})$ is said to be MS continuous at $\vec{\lambda}_*$ if

$$\lim_{\ell\to\infty}\mathbb{E}[(Y(\vec{\lambda}_\ell) - Y(\vec{\lambda}_*)|Y(\vec{\lambda}_\ell) - Y(\vec{\lambda}_*))] = 0, \tag{A1}$$

where $\mathbb{E}[\ldots]$ denotes the expectation of the enclosed quantity over realisations of the GP. For notational convenience, we denote this MS limit as

$$Y(\vec{\lambda}_*) = \underset{\ell\to\infty}{\text{l.i.m.}} Y(\vec{\lambda}_\ell), \tag{A2}$$

where l.i.m. stands for limit in mean [66]. MS continuity implies continuity in the mean,

$$\lim_{\ell\to\infty}\mathbb{E}[Y(\vec{\lambda}_\ell) - Y(\vec{\lambda}_*)] = 0. \tag{A3}$$

This follows from considering the variance of the quantity $Y(\vec{\lambda}_\ell) - Y(\vec{\lambda}_*)$, and the fact that variance is non-negative. There are other notions of continuity of GPs used in the literature, but the notion of MS continuity relates most easily to the covariance. The mean and the covariance of a GP are defined as

$$\begin{aligned} m(\vec{\lambda}) &= \mathbb{E}[Y(\vec{\lambda})], \\ k(\vec{\lambda}_1, \vec{\lambda}_2) &= \mathbb{E}[(Y(\vec{\lambda}_1) - m(\vec{\lambda}_1)|Y(\vec{\lambda}_2) - m(\vec{\lambda}_2))]. \end{aligned} \tag{A4}$$

Using these, Eq. (A1) can be written as

$$\begin{aligned} \lim_{\ell\to\infty}\{&k(\vec{\lambda}_*, \vec{\lambda}_*) - 2k(\vec{\lambda}_\ell, \vec{\lambda}_*) + k(\vec{\lambda}_\ell, \vec{\lambda}_\ell) \\ &+ (m(\vec{\lambda}_*) - m(\vec{\lambda}_\ell)|m(\vec{\lambda}_*) - m(\vec{\lambda}_\ell))\} = 0, \end{aligned} \tag{A5}$$

and using the continuity of the mean in Eq. (A3) gives

$$\lim_{\ell\to\infty}[k(\vec{\lambda}_*, \vec{\lambda}_*) - 2k(\vec{\lambda}_\ell, \vec{\lambda}_*) + k(\vec{\lambda}_\ell, \vec{\lambda}_\ell)] = 0. \tag{A6}$$

This condition is satisfied if the covariance function is continuous at the point $\vec{\lambda}_1 = \vec{\lambda}_2 = \vec{\lambda}_*$. Therefore, we arrive at the result that if the covariance function is continuous in the usual sense at some point $\vec{\lambda}_*$, then the corresponding GP is MS continuous at this point. In fact, a GP is continuous in MS if *and only if* the covariance function is continuous [39], although this is not proved here. In the special case of stationary covariance this reduces to checking continuity of $k(\vec{\tau})$ at $\vec{\tau} = 0$, and in the special case of isotropic covariance, continuity of $k(\tau)$ at $\tau = 0$.

We now move on from continuity to consider differentiability. In the spirit of Eq. (A1), the notion of taking the MS derivative of a GP is defined as

$$\frac{\partial Y(\vec{\lambda})}{\partial \vec{\lambda}^a} = \underset{\epsilon\to 0}{\text{l.i.m.}} X_a(\vec{\lambda}, \epsilon), \tag{A7}$$

$$\text{where } X_a(\vec{\lambda}, \epsilon) = \frac{Y(\vec{\lambda} + \epsilon \hat{e}_a) - Y(\vec{\lambda})}{\epsilon} \tag{A8}$$

with parameter-space unit vector $\hat{e}_a$. This definition can be easily extended to higher-order derivatives [39]. The MS derivative of a GP is also a GP; this follows simply from the fact that the sum of Gaussians is also distributed as a Gaussian. The covariance of $X_a(\vec{\lambda}, \epsilon)$ is given by

$$K_\epsilon(\vec{\lambda}_1, \vec{\lambda}_2) = \mathbb{E}[(X_a(\vec{\lambda}_1, \epsilon) - \Xi(\vec{\lambda}_1, \epsilon)| X_a(\vec{\lambda}_2, \epsilon) - \Xi(\vec{\lambda}_2, \epsilon))] \tag{A9}$$

where $\Xi_a(\vec{\lambda}, \epsilon) = \mathbb{E}[X_a(\vec{\lambda}, \epsilon)]$. It then follows that

$$K_\epsilon(\vec{\lambda}_1, \vec{\lambda}_2) = \frac{k(\vec{\lambda}_1 + \epsilon, \vec{\lambda}_2 + \epsilon) - k(\vec{\lambda}_1, \vec{\lambda}_2 + \epsilon)}{\epsilon^2} - \frac{k(\vec{\lambda}_1 + \epsilon, \vec{\lambda}_2) - k(\vec{\lambda}_1, \vec{\lambda}_2)}{\epsilon^2}. \tag{A10}$$

Substituting this into Eq. (A8), the limit in MS becomes a normal limit, and the result is obtained that the MS derivative of a MS continuous GP with covariance function $k(\vec{\lambda}_1, \vec{\lambda}_2)$ is a GP with covariance function $\partial^2 k(\vec{\lambda}_1, \vec{\lambda}_2)/\partial \vec{\lambda}_1^a \partial \vec{\lambda}_2^a$. In general the covariance function of the n_d-times MS differentiated GP

$$\frac{\partial^{n_d} Y(\vec{\lambda})}{\partial \vec{\lambda}^{a_1} \partial \vec{\lambda}^{a_2} \ldots \partial \vec{\lambda}^{a_{n_d}}}, \tag{A11}$$

is given by the $2n_d$-times differentiated function

$$\frac{\partial^{2n_d} k(\vec{\lambda}_1, \vec{\lambda}_2)}{\partial \vec{\lambda}_1^{a_1} \partial \vec{\lambda}_2^{a_1} \partial \vec{\lambda}_1^{a_2} \partial \vec{\lambda}_2^{a_2} \ldots \partial \vec{\lambda}_1^{a_{n_d}} \partial \vec{\lambda}_2^{a_{n_d}}}. \tag{A12}$$

From the above results relating the MS continuity of GPs to the continuity of the covariance function at $\vec{\lambda}_1 = \vec{\lambda}_2 = \vec{\lambda}_*$, it follows that the n_d-times MS derivative of the GP is MS continuous (the GP is said to be n_d-times MS differentiable) if the $2n_d$-times derivative of the covariance function is continuous at $\vec{\lambda}_1 = \vec{\lambda}_2 = \vec{\lambda}_*$ [54]. So it is the smoothness properties of the covariance function along the diagonal points that determines the differentiability of the GP. (It can also be shown that if a covariance function is continuous at all diagonal points $\vec{\lambda}_1 = \vec{\lambda}_2$ then it is everywhere continuous.)

APPENDIX B: THE EFFECT OF SMALL CHANGES IN THE NOISE PSD ON THE GPR INTERPOLANT

In the offline stage of the method, the GP was trained using the hyperlikelihood in Eq. (18). The result of this process was an interpolant which enabled fast online PE. However, this splitting into offline and online stages also has a potential problem, because the training process makes use of the overlap matrix $M_{ij} = \langle \delta h(\vec{\lambda}_i) | \delta h(\vec{\lambda}_j) \rangle$ which, in turn, depends upon the detector noise PSD $S_n(f)$. The noise PSD is not constant; it changes on short timescales as the noise drifts in the instrument (e.g., [67]), on longer timescales it changes more dramatically as the instrument is gradually upgraded [5]. There are also differences between different detectors, for example between the aLIGO and AdV instruments (or even between the two aLIGO interferometers). It would be a significant drawback if the offline training stage of the process had to be repeated for every single candidate signal because of small differences in the detector PSD.

We do not expect small changes in the noise curve to have a significant effect on the resulting interpolant. First, the noise can be rescaled by an overall constant and have no effect on the position of the peak in the hyperlikelihood; this can be seen from Eq. (18). Second, the peak in the hyperlikelihood is typically wide, and using the hyperparameters from anywhere in the vicinity of the peak still gives reasonable, if not perfect, interpolation. Accordingly, when the PSD changes, some of the difference can be absorbed by an overall scaling, which has no effect on the results, and the remaining change shifts the peak of the hyperlikelihood away from the previously optimised values, but not enough to limit their applicability. If this is the case, then GPs trained on slightly different noise PSDs perform nearly identically to each other and there is no need to retrain for the new PSD.

To assess the sensitivity of our results to changes in the noise curve, we considered three different noise curves chosen to represent the range of possibilities in the advanced-detector era. These are: an estimate of the observing run 1 (O1) aLIGO sensitivity (the early curve of [68]); the zero-detuned high-power (ZDHP) design sensitivity of aLIGO [2,69], and the design sensitivity of AdV [3,70]. As an additional check, we also considered an inverted top-hat noise curve. All of these noise curves

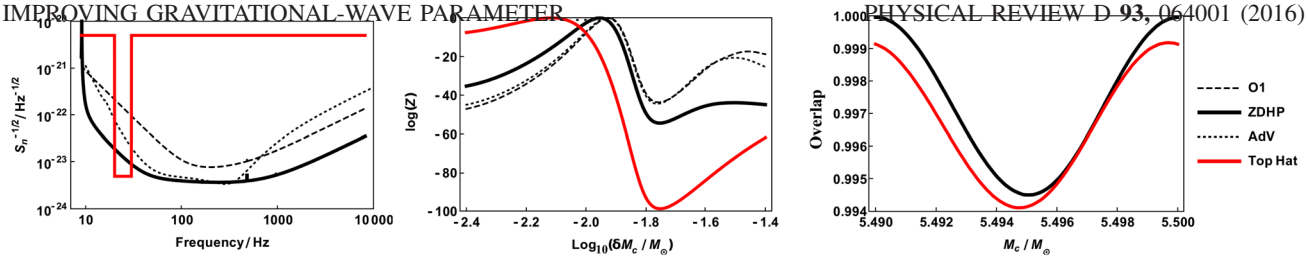

FIG. 11. The left-hand panel shows three different noise curves for ground-based detectors in the advanced era. Also shown in red is an unrealistic noise which we use for comparison. The centre-panel shows the hyper-likelihood surface for the training set $\mathcal{D}_0$ using the SE covariance function and with the overlap matrix calculated using each of the noise curves in the left-hand panel. The right-hand panel shows the waveform overlap between the accurate and the interpolated waveform evaluated for parameter values between two training set points. The interpolants based on the realistic noise curves perform equally well (the curves lie on top of each other). The unrealistic noise curve performs worst, but still gives overlaps greater than 0.994.

are plotted in the left-hand panel of Fig. 11. We then took the training set $\mathcal{D}_0$ and trained the SE GP to find the optimum hyperparameters. Shown in the center panel of Fig. 11 is the hyperlikelihood surface as a function of chirp mass length scale for the different noise curves. As expected, for the range of realistic noise curves the peak in the hyperlikelihood only shifts by a small amount. Finally we used the optimum hyperparameters from each of these hyperlikelihood surfaces to interpolate the training set and calculated the overlap to the accurate waveforms using the ZDHP noise curve; the results of this are shown in the right-hand panel of Fig. 11. For the range of realistic noise curves, the overlap is equally good (cf. [54]). Although the inverted top-hat noise curve gives noticeably lower overlaps, even in that case the drop in the overlap is still less than 0.1%, which is smaller than the difference between the approximate and GPR likelihoods.

This suggests it is safe to train a GP with a fixed noise curve (typical for the instruments considered). The resulting interpolants can be used to analyze all signals without worrying about small drifts in the noise.

[1] G. M. Harry (The LIGO Scientific Collabortation), Classical Quantum Gravity **27**, 084006 (2010).

[2] J. Aasi, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, K. Ackley, C. Adams, T. Adams, P. Addesso *et al.*, Classical Quantum Gravity **32**, 115012 (2015).

[3] F. Acernese, M. Alshourbagy, F. Antonucci *et al.* (The Virgo Collaboration), Virgo Technical Report No. VIR-0027A-09, 2009, URL https://tds.ego-gw.it/ql/?c=6589.

[4] F. Acernese, M. Agathos, K. Agatsuma, D. Aisa, N. Allemandou, A. Allocca, J. Amarni, P. Astone, G. Balestri, G. Ballardin *et al.*, Classical Quantum Gravity **32**, 024001 (2015).

[5] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams *et al.* (The LIGO Scientific–Virgo Collabortation), arXiv:1304.0670.

[6] J. Abadie, B. P. Abbott, R. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, P. Ajith, B. Allen *et al.*, Classical Quantum Gravity **27**, 173001 (2010).

[7] B. P. Abbott *et al.*.(The LIGO Scientific Collabortation and Virgo Collaboration), Phys. Rev. Lett. **116**, 061102 (2016).

[8] S. Klimenko, I. Yakushin, A. Mercer, and G. Mitselmakher, Classical Quantum Gravity **25**, 114029 (2008).

[9] N. J. Cornish and T. B. Littenberg, Classical Quantum Gravity **32**, 135012 (2015).

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

[11] P. Canitrot, Phys. Rev. D **63**, 082005 (2001).

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

[13] C. J. Moore and J. R. Gair, Phys. Rev. Lett. **113**, 251101 (2014).

[14] F. Pretorius, Phys. Rev. Lett. **95**, 121101 (2005).

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

[16] J. G. Baker, J. Centrella, D.-I. Choi, M. Koppitz, and J. van Meter, Phys. Rev. Lett. **96**, 111102 (2006).

[17] A. H. Mroué, M. A. Scheel, B. Szilágyi, H. P. Pfeiffer, M. Boyle, D. A. Hemberger, L. E. Kidder, G. Lovelace, S. Ossokine, N. W. Taylor *et al.*, Phys. Rev. Lett. **111**, 241104 (2013).

[18] J. Aasi, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, T. Accadia, F. Acernese, K. Ackley, C. Adams, T. Adams *et al.* (The LIGO Scientific–Virgo Collabortation & The NINJA-2 Collaboration), Classical Quantum Gravity **31**, 115004 (2014).

[19] B. Szilgyi, J. Blackman, A. Buonanno, A. Taracchini, H. P. Pfeiffer, M. A. Scheel, T. Chu, L. E. Kidder, and Y. Pan, Phys. Rev. Lett. **115**, 031102 (2015).

[20] A. Buonanno and T. Damour, Phys. Rev. D **59**, 084006 (1999).
[21] A. Buonanno and T. Damour, Phys. Rev. D **62**, 064015 (2000).
[22] Y. Pan, A. Buonanno, M. Boyle, L. T. Buchman, L. E. Kidder, H. P. Pfeiffer, and M. A. Scheel, Phys. Rev. D **84**, 124052 (2011).
[23] A. Taracchini, A. Buonanno, Y. Pan, T. Hinderer, M. Boyle, D. A. Hemberger, L. E. Kidder, G. Lovelace, A. H. Mroué, H. P. Pfeiffer *et al.*, Phys. Rev. D **89**, 061502 (2014).
[24] L. Santamaría, F. Ohme, P. Ajith, B. Brügmann, N. Dorband, M. Hannam, S. Husa, P. Mösta, D. Pollney, C. Reisswig *et al.*, Phys. Rev. D **82**, 064016 (2010).
[25] P. Schmidt, M. Hannam, and S. Husa, Phys. Rev. D **86**, 104063 (2012).
[26] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, Phys. Rev. Lett. **113**, 151101 (2014).
[27] P. Schmidt, F. Ohme, and M. Hannam, Phys. Rev. D **91**, 024043 (2015).
[28] T. B. Littenberg, J. G. Baker, A. Buonanno, and B. J. Kelly, Phys. Rev. D **87**, 104003 (2013).
[29] A. Ghosh, W. Del Pozzo, and P. Ajith, arXiv:1505.05607.
[30] J. Veitch, M. Pürrer, and I. Mandel, Phys. Rev. Lett. **115**, 141101 (2015).
[31] P. B. Graff, A. Buonanno, and B. S. Sathyaprakash, Phys. Rev. D **92**, 022002 (2015).
[32] L. Blanchet, Living Rev. Relativity **17**, 2 (2014).
[33] A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, Phys. Rev. D **80**, 084043 (2009).
[34] T. Sidery, B. Aylott, N. Christensen, B. Farr, W. Farr, F. Feroz, J. Gair, K. Grover, P. Graff, C. Hanna *et al.*, Phys. Rev. D **89**, 084060 (2014).
[35] C. P. L. Berry, I. Mandel, H. Middleton, L. P. Singer, A. L. Urban, A. Vecchio, S. Vitale, K. Cannon, B. Farr, W. M. Farr *et al.*, Astrophys. J. **804**, 114 (2015).
[36] J. R. Gair and C. J. Moore, Phys. Rev. D **91**, 124062 (2015).
[37] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003).
[38] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006).
[39] R. J. Adler, The Geometry of Random Fields, *Wiley Series in Probability and Mathematical Statistics* (Wiley, New York, 1981).
[40] A. Taracchini, Y. Pan, A. Buonanno, E. Barausse, M. Boyle, T. Chu, G. Lovelace, H. P. Pfeiffer, and M. A. Scheel, Phys. Rev. D **86**, 024011 (2012).
[41] J. Veitch and A. Vecchio, Phys. Rev. D **81**, 062003 (2010).
[42] J. Veitch, V. Raymond, B. Farr, W. Farr, P. Graff, S. Vitale, B. Aylott, K. Blackburn, N. Christensen, M. Coughlin *et al.*, Phys. Rev. D **91**, 042003 (2015).
[43] C. J. Moore, R. H. Cole, and C. P. L. Berry, Classical Quantum Gravity **32**, 015014 (2015).
[44] C. Cutler and É. E. Flanagan, Phys. Rev. D **49**, 2658 (1994).
[45] L. S. Finn, Phys. Rev. D **46**, 5236 (1992).
[46] C. E. Shannon, Bell Syst. Tech. J. **27**, 623 (1948).
[47] D. J. C. MacKay, Neural Comput. **11**, 1035 (1999).
[48] E. Snelson and Z. Ghahramani, in *Advances in Neural Information Processing Systems*, edited by Y. Weiss, B. Schölkopf, and J. C. Platt (MIT Press, Cambridge, MA, 2006), vol. 18, p. 1257.
[49] J. Quińonero-Candela, C. E. Rasmussen, and C. K. I. Williams, in *Large-scale Kernel Machines*, edited by L. Bottou, O. Chapelle, D. DeCoste, and J. Weston (MIT Press, Cambridge, MA, 2007), Chap. 9, p. 203.
[50] C. J. Paciorek and M. J. Schervish, in *Advances in Neural Information Processing Systems 16*, edited by S. Thrun, L. Saul, and B. Schölkopf (MIT Press, Cambridge, MA, 2004).
[51] B. J. Owen, Phys. Rev. D **53**, 6749 (1996).
[52] B. S. Sathyaprakash and S. V. Dhurandhar, Phys. Rev. D **44**, 3819 (1991).
[53] C. Kalaghatgi, P. Ajith, and K. G. Arun, Phys. Rev. D **91**, 124042 (2015).
[54] M. L. Stein, Interpolation of Spatial Data, *Springer Series in Statistics* (Springer–Verlag, New York, NY, 1999).
[55] G. N. Watson, A Treatise on the Theory of Bessel Functions, 2nd ed., *Cambridge Mathematical Library* (Cambridge University Press, Cambridge, England, 1995).
[56] H. Wendland, Scattered Data Approximation, *Cambridge Monographs on Applied and Computational Mathematics* (Cambridge University Press, Cambridge, 2004).
[57] T. Gneiting, J. Multivariate Anal. **83**, 493 (2002).
[58] A. Melkumyan and F. Ramos, in *IJCAI'09 Proceedings of the 21st international joint conference on Artifical intelligence* Vol. 9 (Morgan Kaufmann Publishers Inc., San Francisco, CA, 2009), p. 1936.
[59] M. Liang and D. Marcotte, Stoch. Environ. Res. Risk Assess. 15 (2015).
[60] M. Vallisneri, Phys. Rev. D **77**, 042001 (2008).
[61] L. Santamaría, F. Ohme, P. Ajith, B. Brügmann, N. Dorband, M. Hannam, S. Husa, P. Mösta, D. Pollney, C. Reisswig *et al.*, Phys. Rev. D **82**, 064016 (2010).
[62] T. Damour, B. R. Iyer, and B. S. Sathyaprakash, Phys. Rev. D **63**, 044023 (2001).
[63] T. Damour, B. R. Iyer, and B. S. Sathyaprakash, Phys. Rev. D **66**, 027502 (2002).
[64] P. Canizares, S. E. Field, J. Gair, V. Raymond, R. Smith, and M. Tiglio, Phys. Rev. Lett. **114**, 071104 (2015).
[65] M. Pürrer, Classical Quantum Gravity **31**, 195010 (2014).
[66] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series* (John Wiley & Sons, New York, 1949).
[67] J. Aasi, J. Abadie, B. P. Abbott, R. Abbott, T. Abbott, M. R. Abernathy, T. Accadia, F. Acernese, C. Adams, T. Adams *et al.*, Classical Quantum Gravity **32**, 115012 (2015).
[68] L. Barsotti and P. Fritschel (The LIGO Scientific Collaboration), Technical Report No. LIGO-T1200307-v4, 2012, URL https://dcc.ligo.org/LIGO-T1200307/public.
[69] D. Shoemaker (The LIGO Scientific Collaboration), LIGO Report No. LIGO-T0900288-v3, 2010, URL https://dcc.ligo.org/LIGO-T0900288/public.
[70] T. Accadia, F. Acernese, M. Agathos *et al.* (The Virgo Collaboration), Virgo Technical Report No. VIR-0128A-12, 2012, URL https://tds.ego-gw.it/ql/?c=8940.