Reconstructing the sky location of gravitational-wave detected compact binary systems: Methodology for testing and comparison

T. Sidery,[1,*] B. Aylott,[1] N. Christensen,[2] B. Farr,[3,1] W. Farr,[3,1] F. Feroz,[4] J. Gair,[5] K. Grover,[1] P. Graff,[6] C. Hanna,[7] V. Kalogera,[3] I. Mandel,[1] R. O'Shaughnessy,[8] M. Pitkin,[9] L. Price,[10] V. Raymond,[10] C. Röver,[11,12] L. Singer,[10] M. van der Sluys,[13] R. J. E. Smith,[1] A. Vecchio,[1] J. Veitch,[14] and S. Vitale[15]

[1]*School of Physics and Astronomy, University of Birmingham, Birmingham B15 2TT, United Kingdom*
[2]*Physics and Astronomy, Carleton College, Northfield, Minnesota 55057, USA*
[3]*Department of Physics and Astronomy, Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA), Northwestern University, Evanston, Illinois 60208, USA*
[4]*Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, United Kingdom*
[5]*Institute of Astronomy, University of Cambridge, Cambridge CB3 0HA, United Kingdom*
[6]*NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA*
[7]*Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada*
[8]*Center for Gravitation and Cosmology, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53211, USA*
[9]*SUPA, School of Physics and Astronomy, University of Glasgow, University Avenue, Glasgow G12 8QQ, United Kingdom*
[10]*LIGO, California Institute of Technology, Pasadena, California 91125, USA*
[11]*Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Callinsraßa 38, 30167 Hannover, Germany*
[12]*Department of Medical Statistics, University Medical Center, Göttingen, 37073 Göttingen, Germany*
[13]*Radboud University Nijmegen, P.O. Box 9010, NL-6500 GL Nijmegen, The Netherlands*
[14]*Nikhef, Science Park 105, Amsterdam 1098 XG, The Netherlands*
[15]*Massachusetts Institute of Technology, 185 Albany Street, Cambridge, Massachusetts 02139, USA*

(Received 19 December 2013; published 18 April 2014)

The problem of reconstructing the sky position of compact binary coalescences detected via gravitational waves is a central one for future observations with the ground-based network of gravitational-wave laser interferometers, such as Advanced LIGO and Advanced Virgo. Different techniques for sky localization have been independently developed. They can be divided in two broad categories: fully coherent Bayesian techniques, which are high latency and aimed at in-depth studies of all the parameters of a source, including sky position, and "triangulation-based" techniques, which exploit the data products from the search stage of the analysis to provide an almost real-time approximation of the posterior probability density function of the sky location of a detection candidate. These techniques have previously been applied to data collected during the last science runs of gravitational-wave detectors operating in the so-called initial configuration. Here, we develop and analyze methods for assessing the self consistency of parameter estimation methods and carrying out fair comparisons between different algorithms, addressing issues of efficiency and optimality. These methods are general, and can be applied to parameter estimation problems other than sky localization. We apply these methods to two existing sky localization techniques representing the two above-mentioned categories, using a set of simulated inspiral-only signals from compact binary systems with a total mass of $\leq 20M_{\odot}$ and nonspinning components. We compare the relative advantages and costs of the two techniques and show that sky location uncertainties are on average a factor $\approx$20 smaller for fully coherent techniques than for the specific variant of the triangulation-based technique used during the last science runs, at the expense of a factor $\approx$1000 longer processing time.

DOI: 10.1103/PhysRevD.89.084060 PACS numbers: 04.30.-w, 97.60.-s, 97.80.-d, 95.75.Pq

I. INTRODUCTION

Ground-based gravitational-wave (GW) laser interferometers—Laser Interferometer Gravitational Wave Observatory (LIGO) [1], Virgo [2] and GEO-600 [3]—have completed science observations in 2010 (S6/VSR2-3) [4] in the so-called initial configuration, and are currently being upgraded with the plan to start running again from 2015 at a significantly improved sensitivity [5,6]. No detection was achieved during this initial period of observations; however, the expectations are that by the time the instruments reach design "advanced" sensitivity they shall routinely

*tsidery@star.sr.bham.ac.uk

detect gravitational-wave signals. One of the most promising candidate sources for detection are coalescing binary systems of compact objects containing neutron stars and black holes [7].

One of the key pieces of information to extract is the source location in the sky. Once a detection candidate is identified by search pipelines, the location parameters that describe the source are reconstructed using a number of techniques, both high and low latency [8,9]. In contrast to traditional telescopes, gravitational-wave instruments are all-sky monitors and the source location in the sky is reconstructed *a posteriori*. Information about the source geometry is primarily encoded in the relative time of arrival of GW radiation at the different detector sites, together with the relative amplitude and phase of the GWs as seen in different detectors. Constraining the source location on the sky will be an important element of the analysis because it allows for follow-ups of the relevant portion of the sky with electromagnetic instruments, possibly over a wide spectral range, and could offer information about the environment of a GW-detected binary [10–12]. The electromagnetic signatures associated with the merger of the compact objects are expected to be transient, so the time scale over which the sky location information becomes available from the gravitational-wave ground-based network is also important.

For this reason the problem of reconstructing the sky position of GW sources with the network of ground-based laser interferometers is an area of active work in preparation for advanced instruments [13–18]. By the end of observations with instruments in initial configuration, two main implementations had been used to determine the sky localization uncertainty region of a coalescing binary candidate [8,9]:

(i) LALInference [19], a library of fully coherent Bayesian analysis algorithms, computes the posterior probability density function (PDF) on the sky location and other parameters on the time scale of hours to several weeks, depending on the specific signal. Using two classes of stochastic sampling techniques, Markov-Chain Monte Carlo [20–22] and nested sampling [23–25], LALInference coherently analyzes the data from all the interferometers in the network and generates the multidimensional PDF on the full set of parameters needed to describe a binary system before marginalizing over all parameters other than the sky location (a binary in circular orbit is described by 9 to 15 parameters, depending on whether spins of the binary components are included in the model).

(ii) A much faster low-latency technique, that we will call Timing++ [8], uses data products from the search stage of the analysis, and can construct sky maps on (sub)minute time scales by using primarily time-delay information between different detector sites. In particular, the masses, time and phase of arrival, and the amplitude of the signal are searched for in each detector separately and the masses and time of arrival are checked for consistency [26]. The time of arrival and amplitude of the signal in each detector are the intermediate data products used by Timing++ to construct the PDF of the sky location.

These two approaches were initially designed to serve different purposes: a thorough parameter reconstruction and a low-latency sky localization technique, trading off accuracy for computational speed.

The goal of this paper is twofold. Several parameter estimation approaches have been, and continue to be, developed in preparation of the advanced instruments coming online in 2015. Algorithms may be tuned in specific ways to serve different purposes. The first goal of this paper is to provide fair and rigorous methods to compare different approaches in order to inform future developments. One of the most actively investigated parameter estimation aspects is sky localization reconstruction. It is therefore natural to apply these comparison methods to the algorithms used up to now to check the consistency of the results, quantify relative benefits and identify the areas that need the most attention in the future. The second goal of this paper is to provide the first rigorous comparison of the two sky localization techniques described above. We examine the sky location PDFs for a large number of simulated signals from coalescing compact binaries with total masses of up to $20M_{\odot}$ in simulated stationary, Gaussian noise. Although our signal distribution is not astrophysically motivated, it allows us to statistically examine the self consistency of both techniques by testing whether the claimed uncertainty regions match the actual probability that the source is found at those sky locations. Furthermore, by comparing the uncertainties in sky location across the code outputs we gain an understanding of the systematic behavior of each technique. Many of these comparison methods have now become the routine test bed in the development effort for gravitational-wave data analysis and may have applicability in other areas of astronomy.

The paper is organized as follows. In Sec. II we describe two techniques used to determine the sky location of a candidate coalescing compact binary. In Sec. III, we evaluate the correctness of the two techniques using a simulated population of binaries over a wide range of the parameter space, comparing their sky localization capabilities and latency time scales. Sec. IV contains our conclusions and pointers to future work.

II. LOCATION RECONSTRUCTION METHODS

Gravitational-wave interferometers are, by design, sensitive to sources across much of the sky. Because of this, position reconstruction estimates rely largely on time delays between sites in a multiple detector network, i.e., triangulation. Using only time-delay information, there is generally a degeneracy in the position reconstructed. For a two-detector network, this degeneracy is a conical surface of constant time delay around the line connecting the two detectors, whose projection onto the sky plane yields a ring.

For a three-detector network, this degeneracy is broken into two regions symmetric about the plane defined by the detectors: the intersections of two rings on the sky. A four- (or more) detector network will generally identify a single region in the sky. However, time delays are not the only source of sky location information. Though the observed amplitude of gravitational waves depends only weakly on the source location, it typically helps to break these degeneracies in two and three detector networks; further information is contained in the relative phasing between detectors [18]. In this section we outline the two methods considered so far for position reconstruction.

We can formalize the problem we want to address as follows. The data,

$$d_j(t) = n_j(t) + h_j(t;\vec{\theta}), \tag{1}$$

from each gravitational-wave interferometer in the network $j = 1, \ldots, N$, where N is the number of instruments, is a sum of the noise $n_j(t)$ and any signal $h_j(t;\vec{\theta})$, where $\vec{\theta}$ is a vector that describes the set of unknown parameters that characterize the GW source. For this study we consider coalescing binaries of compact objects with approximately circular orbits and negligible spins; $\vec{\theta}$ is a nine-dimensional parameter vector: two mass parameters (the two component masses $m_{1,2}$, or an alternative combination of these, e.g., the symmetric mass ratio $\eta = m_1 m_2/(m_1+m_2)^2$ and the chirp mass $\mathcal{M} = \eta^{3/5}(m_1+m_2)$), the distance to the source D, the source location in the sky (described by two angles that identify the unit vector $\vec{\Omega}$, e.g., right ascension α and declination δ), the orientation of the binary (polarization ψ and inclination of the orbital plane ι) and the reference phase ϕ_0 and time t_0. To simplify notation, we define

$$\vec{\theta} = \{\vec{\Omega}, \vec{\beta}\}, \tag{2}$$

where $\vec{\beta}$ is the parameter vector that *does not* contain the sky location parameters, right ascension and declination. Regardless of the specific technique that one decides to adopt, the goal is to evaluate $p(\vec{\Omega}|d)$, the marginalized joint posterior density function of the sky location parameters given the observations.

A straightforward application of Bayes' theorem allows us to calculate the posterior probability density for a model with parameters $\vec{\theta}$ given the data, d, using

$$p(\vec{\theta}|d) = \frac{p(d|\vec{\theta})p(\vec{\theta})}{p(d)}. \tag{3}$$

The prior probability density, $p(\vec{\theta})$, encapsulates all our *a priori* information about the expected distribution of sources in distance, masses or other parameters in the model. The likelihood $p(d|\vec{\theta})$ is the probability of generating the data set d given an assumed signal with parameters $\vec{\theta}$. The evidence $p(d)$ is used to normalize the integral of the posterior over the entire parameter space to unity.

A. LALInference

The evaluation of $p(\vec{\theta}|d)$ is notoriously difficult in high-dimensional problems with complex likelihood functions, as is the case for coalescing compact binaries in a network of laser interferometers. We have developed a set of sampling algorithms within the LSC Algorithms Library (LAL) [27], collected under LALInference [19], specifically for the analysis of gravitational-wave data, and for what is relevant here, coalescing binary signal models. The library contains two main stochastic parameter-space exploration techniques: Markov-Chain Monte Carlo (lalinference_mcmc [22]), and nested sampling (lalinference_nest [24] and lalinference_bambi [28]). Different algorithms are included to validate results during the development stage and to explore a range of schemes to optimize the run time. These techniques have been used to analyze a set of hardware and software injections as well as detection candidates during the last LIGO/Virgo science runs [9]; a technical description of the algorithms will be reported elsewhere [19].

The output of a LALInference run is a list of "samples," values of $\vec{\theta}$ drawn from the joint posterior probability density function. The density of samples in a region of parameter space is proportional to the value of the PDF. For the specific sky localization problem we are considering here, the marginalized posterior probability density function on the sky location is simply

$$p(\vec{\Omega}|d) = \int p(\vec{\Omega}, \vec{\beta}|d) d\vec{\beta}, \tag{4}$$

where $p(\vec{\Omega}, \vec{\beta}|d) \equiv p(\vec{\theta}|d)$ is derived using Eq. (3). If we could extract an infinite number of samples then we would be able to map out the PDF perfectly; however, these are computationally intensive algorithms, see Sec. III D for more details, and we typically have ~1000 *independent* samples. The finite number of samples can introduce both stochastic and systematic bias, and so we have implemented a two-step kD-tree binning process to estimate the PDF that removes the systematic issues [29].

The fully coherent Bayesian analysis takes into account the search stage of the analysis only to set the prior range for the arrival time of a gravitational wave around the observed detection candidate. However, the matched-filtering stage of a search already offers processed information that can be used to generate approximate posterior density functions $p(\vec{\Omega}|d)$. This is the approach taken in Timing++.

B. Timing++

Timing++ [8] takes the parameters of the waveform that best fit the data in each detector, as found by the initial

search [26], and assumes that the posterior of interest is only a function of the arrival times in each detector, $t^{(i)}$, and the amplitude of the signal in each detector, $A^{(i)}$. That is, we write

$$p(\vec{\Omega}|d) \approx p(\vec{\Omega}|t^{(i)}, A^{(i)}), \tag{5}$$

where $\vec{\Omega}$ is the location on the sky. We further assume that the information in the arrival times and amplitudes can each be replaced by a single quantity so that

$$\begin{aligned} p(\vec{\Omega}|d) \approx p(\vec{\Omega}|t^{(i)}, A^{(i)}) &\propto f(\Delta t_{\mathrm{rss,sc}}(\vec{\Omega}), \Delta A_{\mathrm{rss}}(\vec{\Omega})) \\ &\equiv f(\Delta t_{\mathrm{rss,sc}}, \Delta A_{\mathrm{rss}}), \end{aligned} \tag{6}$$

where $f(\Delta t_{\mathrm{rss,sc}}, \Delta A_{\mathrm{rss}})$ is an empirically derived distribution function and $\Delta t_{\mathrm{rss,sc}}$ and ΔA_{rss} are described in the following. For a source at position $\vec{\Omega}$, the arrival time at detector i allows us to predict the arrival time at any other fiducial point, which, for the sake of simplicity, we choose to be the geocenter. In the absence of noise, the predicted geocentric arrival times, computed separately from each detector's measured arrival time, should coincide. The summed squared differences of the predicted arrival times at the geocenter between detector pairs give us a measure of how far we expect to be from the true location:

$$\Delta t_{\mathrm{rss}} = \sqrt{\sum_{i>j} ((t^{(i)}_{\mathrm{ref}} - t^{(i)}_{\mathrm{geo}}(\vec{\Omega})) - (t^{(j)}_{\mathrm{ref}} - t^{(j)}_{\mathrm{geo}}(\vec{\Omega})))^2}, \tag{7}$$

where $t^{(i)}_{\mathrm{geo}}(\vec{\Omega})$ is the difference between the arrival time of a signal from $\vec{\Omega}$ at detector (i) and at the geocenter, and $t^{(i)}_{\mathrm{ref}}$ is the time the signal crosses a *reference frequency* in the band of detector *i*. This vanishes in the idealized case of no noise for the true location. By appropriately choosing the reference frequency we minimize the correlation between the determined mass and phase in the waveform and the recovered time of arrival [30]. This is important since the parameters of the waveform are determined separately in each detector. Moreover, we expect that these errors in timing will scale inversely with the signal-to-noise ratio (SNR) of the system in the high-SNR regime:

$$\Delta t_{\mathrm{rss}} = \Delta t_{\mathrm{rss,sc}} \frac{10}{\rho}, \tag{8}$$

where $\rho = \sqrt{\sum_i \rho_i^2}$ is the *combined* SNR, ρ_i is the SNR measured in detector i, and the factor of 10 is chosen as a fiducial SNR. We use the SNR-corrected $\Delta t_{\mathrm{rss,sc}}$ in place of Δt_{rss} to remove this dependence on SNR.

Incorporating the amplitude of the signal is more complicated. The SNR is a function not only of sky location but also of luminosity distance, inclination and polarization of the signal. Because this method is designed for low-latency sky localization, a somewhat *ad hoc* measure of amplitude consistency between detectors is used. The starting point is the fact that

$$\rho_i \propto \frac{1}{D^{(i)}_{\mathrm{eff}}}, \tag{9}$$

where D_{eff} is an effective distance, defined by

$$D_{\mathrm{eff}} = D\left[F_+^2\left(\frac{1+\cos^2\iota}{2}\right)^2 + F_\times^2\cos^2\iota\right]^{-1/2}, \tag{10}$$

and $F_{+,\times} = F_{+,\times}(\vec{\Omega}, \psi)$ are the antenna beam pattern functions; see Eqs. B9 and B10 of Ref. [31]. While the matched filter detection pipeline produces an estimate of D_{eff} separately in each detector, it is not invertible to obtain any of the quantities in Eq. (10) directly. With that in mind, we define

$$A^2 \equiv \frac{1}{F_+^2(\vec{\Omega}, \psi = 0) + F_\times^2(\vec{\Omega}, \psi = 0)}, \tag{11}$$

and use

$$\Delta A_{\mathrm{rss}} = \sqrt{\sum_{i>j}\left(\frac{D^{(i)2}_{\mathrm{eff}} - D^{(j)2}_{\mathrm{eff}}}{D^{(i)2}_{\mathrm{eff}} + D^{(j)2}_{\mathrm{eff}}} - \frac{A^{(i)2} - A^{(j)2}}{A^{(i)2} + A^{(j)2}}\right)^2} \tag{12}$$

as a measure of the consistency of the calculated and observed difference in response functions between each detector pair. In contrast to Eq. (7), this quantity is typically not zero in the absence of noise as $A^2 = D_{\mathrm{eff}}/D$ only when inclination and polarization are both 0. However, the use of amplitude reconstruction in this manner has been determined empirically to improve position reconstruction estimates. In contrast to $\Delta t_{\mathrm{rss,sc}}$, there is no adjustment for SNR in ΔA_{rss}. Grover *et al.* [18] showed that phase consistency between detectors can provide additional information on sky location and significantly reduce sky localization uncertainty; however, phase consistency was not included in TIMING++.

Putting together our previous assumptions,

$$\begin{aligned} p(\vec{\Omega}|d) &\approx p(\vec{\Omega}|t^{(i)}, A^{(i)}) \\ &\propto p(\vec{\Omega}) f(\Delta t_{\mathrm{rss,sc}}, \Delta A_{\mathrm{rss}}) \\ &\approx p(\vec{\Omega}) f_t(\Delta t_{\mathrm{rss,sc}}) f_A(\Delta A_{\mathrm{rss}}), \end{aligned} \tag{13}$$

where $p(\vec{\Omega})$ is the prior on the sky location and in the third line we have assumed that $f(\Delta t_{\mathrm{rss,sc}}, \Delta A_{\mathrm{rss}})$ can be written as the product of two other empirical distributions, $f_t(\Delta t_{\mathrm{rss,sc}})$ and $f_A(\Delta A_{\mathrm{rss}})$. In this work we assume isotropic priors on the sky location. In the low-latency search for compact binaries and associated electromagnetic counterparts for which TIMING++ was designed, a restrictive prior that limited consideration to only areas of the sky

containing galaxies was imposed, as described in [8]. In practice, $f_t(\Delta t_{\mathrm{rss,sc}})$ and $f_A(\Delta A_{\mathrm{rss}})$ are measured beforehand using simulations, where $\Delta t_{\mathrm{rss,sc}}$ and ΔA_{rss} are computed from the recovered arrival times and effective distances, respectively, and the true (known) sky location, $\vec{\Omega}_{\mathrm{true}}$. This amounts to evaluating $\Delta t_{\mathrm{rss,sc}}$ (ΔA_{rss}) according to Eq. (7) [Eq. (12)] at $\vec{\Omega}_{\mathrm{true}}$ using the time of arrival (effective distance) from the matched filter pipeline. A kernel density estimator is then used to estimate the distribution of these quantities. When a candidate is found, $\Delta t_{\mathrm{rss,sc}}$ and ΔA_{rss} are computed across a fixed grid on the sky, and the likelihood is taken from the previously simulated distributions and the result is normalized, leading to an inherently fast method.

III. TESTING

The goal of this study is to compare the relative performances in terms of sky localization of TIMING++ and LALINFERENCE and in doing so to develop a set of criteria and general tools that can be applied to many parameter estimation problems in which different techniques are considered. The tests should ensure that each algorithm separately is self consistent, and then provide fair methods of making comparisons.

For the specific problem considered in this paper, TIMING++ and LALINFERENCE both evaluate the posterior probability density function $p(\vec{\Omega}|d)$; see Eqs. (4) and (13). For a given model assumption and data realization, there is an exact PDF of which the algorithms produce an approximation. There are many effects that can distort the recovered PDF from the true one. They can be grouped in two broad categories.

Irrespective of the algorithm that is used, the assumptions on the elements that enter the PDF calculation may differ from the actual problem, and therefore produce a bias in the results. For the problem at hand, they can be summarized as follows: (i) the model waveform family does not describe the actual signal contained in the data; (ii) the noise model is incorrect; and (iii) the choice of priors does not match the actual ones that describe the problem, and, in the specific case considered here, the priors from which the source parameters have been drawn. Each of these enter the calculation of the PDF; see Eq. (3). In the test described here, the signal model (the waveform family) is exactly known, and the same waveform family is used for the signal generation and the likelihood calculation. The statistical properties of the noise—Gaussian and stationary drawn from a known distribution—are also known. It is, however, important to emphasize that in the case of LALINFERENCE the noise power spectral density (PSD) is estimated from the data surrounding the signal, and as a consequence it does not exactly describe the distribution from which the noise is drawn. For the TIMING++ analysis, on the other hand, the noise PSD is taken to be exactly the one used to generate the noise realizations.

A different set of effects that can affect the recovered PDF are more fundamentally intrinsic to the algorithms: (i) the assumptions that go into the likelihood calculation are not perfect, (ii) there are algorithmic issues that produce errors, and (iii) PDFs cannot be reconstructed perfectly from a finite number of samples (postprocessing). The likelihood calculation makes assumptions about the form of the noise and so is linked to the previously mentioned noise issue. For TIMING++, the likelihood is calculated using a mix of approximations and simulated runs. This is a point of possible bias entering the results of the TIMING++ runs; measuring its extent is part of our investigation.

As well as the obvious statement that the algorithm must be working correctly, it was found with LALINFERENCE that the way that the results are processed to create a continuous PDF from discrete samples from the posterior can also introduce noticeable distortions. This is linked to the finite sampling issues mentioned previously and fixed with two-stage kD-trees [29].

While in theory the sources of bias due to the test itself are straightforward to control, any erroneous results may either be due to code issues or a failure to properly treat the setup issues, both of which may give very similar distortions in the final PDF. This leads to a cycle of code checking and test setup checking while codes are being developed. This is particularly true of the LALINFERENCE-type algorithms that, with the correct setup, should precisely recover the PDF, creating a stringent checking mechanism for the codes' self consistency.

A. Test population

To set up a rigorous comparison test bed we have considered 360 mock inspiralling compact binary signals from a population of binary sources and "injected" the waveforms into Gaussian and stationary noise representing observations with the three-detector network consisting of the two LIGO detectors at Hanford, Washington and Livingston, Louisiana and the Virgo detector near Pisa, Italy. The power spectrum of the noise was chosen to mimic the LIGO sensitivity achieved during the last science run [4] and was the same for all the instruments of the network, including Virgo. A subset of this population has been recently used for other parameter estimation studies; see Refs. [18,32]. The noise data were generated with the infrastructure used for the NINJA-2 project [33]. The low-frequency cutoff was set to 40 Hz.

The source distribution was chosen to test these two sky localization approaches over a large range of signal-to-noise ratios and physical parameters that describe stellar mass binary systems, rather than being astrophysically motivated. The mass distribution was uniform in component masses with $1M_{\odot} \leq m_{1,2} \leq 15M_{\odot}$ and a cutoff on the total mass $m_1 + m_2 \leq 20M_{\odot}$. The sky position and orientation of the systems with respect to the interferometers

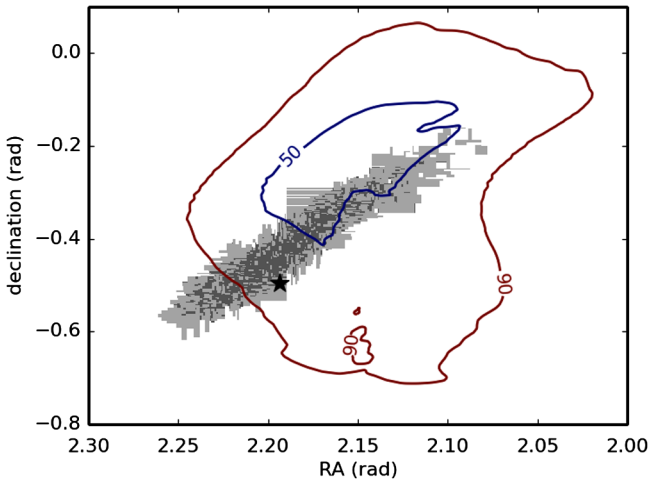

FIG. 1 (color online). An output PDF of the sky position from the two codes. The contour lines label the 50% and 90% credible regions for TIMING++ while the light and dark shaded regions show the 50% and 90% credible regions, respectively, for LALINFERENCE. The star indicates the source location.

were distributed uniformly. For distances between 10 and 40 Mpc the logarithm of the distance was uniformly distributed in order to give a broad range of network SNRs above the detection threshold.

The waveforms used to generate, and then analyze, the signal are restricted post-Newtonian approximations of the inspiral phase, with spins of the binary components set to zero. The time-domain TaylorT3 and TaylorT4 approximants of the LAL [27] at second post-Newtonian order in phase, in which the differential equations that describe the evolution of a characteristic orbital velocity and phase of the system are Taylor expanded in terms of the characteristic velocity of the two inspiralling objects [34], were used for TIMING++ and LALINFERENCE, respectively. The precise forms of the two families of waveforms have phase differences from post$^{2.5}$-Newtonian order and above, which has no effect for the purpose of these comparisons; the crucial factor for these tests was that each code used the same waveform family for injection and subsequent recovery of the signal. It was necessary to use different waveforms in each code due to compatibility issues of the implementations.

The synthetic data containing GW signals added to noise were processed using the standard matched-filter search pipeline IHOPE [26] used in the LIGO/Virgo analyses in this parameter range; see, e.g., Ref. [35] and references therein. LALINFERENCE was run on all the 360 injections, with a flat prior on the time of arrival over a range of ± 100 milliseconds around the time of the injection. TIMING++ uses an additional criterion that the SNR must be greater than 5.5 in each of three detectors; 243 candidates passed this cut. Figure 1 gives an example output PDF from one of the runs. For the self-consistency tests described in Sec. III B we used all the results available for each algorithm. For the comparisons between the codes in Sec. III C, we only used those data sets for which results from both methods are available.

B. Self-consistency checks

We describe the PDF via credible levels (CL): the integrated probability, in our case $p(\vec{\Omega}|d)$, over a given region of the parameter space. In particular we consider the smallest region, or minimum credible region ($\mathrm{CR}_{\mathrm{min}}$), for a given CL; in our case, this corresponds to the smallest region in the sky that contains the given probability that the source is in that location. More formally, for a given CL, any credible region (CR) must satisfy

$$\mathrm{CL} = \int_{\mathrm{CR}} p(\vec{\Omega}|d) d\vec{\Omega}. \tag{14}$$

We can then find the smallest region such that this still holds, which we call $\mathrm{CR}_{\mathrm{min}}$. By considering the full range of probabilities we can map out the PDF with a set of contours that bound each $\mathrm{CR}_{\mathrm{min}}$.

While the analysis of a single GW signal will not tell us very much about the correctness of the analysis, considering how CL and $\mathrm{CR}_{\mathrm{min}}$ are related over a large number of GW signals gives us statistical information: Does a given credible level really correspond to the probability of finding the source in that location? For each run and a given CL we can check if the injection's parameter coordinates fall within the associated $\mathrm{CR}_{\mathrm{min}}$; if there are no sources of bias in the analysis, this should happen with probability CL in order for the credible regions to be meaningful. We can plot a cumulative figure, over all injected signals and the full range of CLs, of the proportion of injections found within a given CL's $\mathrm{CR}_{\mathrm{min}}$. We expect this to be diagonal, up to statistical fluctuations arising from a finite number of injections. Deviations from the diagonal indicate that the parameter estimation algorithm does not correctly evaluate the PDF, or other sources of bias are present, e.g., the priors used in the analysis do not match the distribution of the injected source population.

The results of this test from all the signals detected out of the 360 injections in each of LALINFERENCE and TIMING++ is shown in Fig. 2. The error bars are calculated from the expected variance in the number of injections that fall within a given CR. For a CL of p, and n runs, the variance on the number of sources found within $\mathrm{CR}_{\mathrm{min}}$ is $np(1-p)$ if the fraction of injections that fall within a given $\mathrm{CR}_{\mathrm{min}}$ is really described by the binomial distribution, as expected. The error bars on the fraction of injections found within a given $\mathrm{CR}_{\mathrm{min}}$ are given by the standard deviation normalized by the number of runs, $\sqrt{p(1-p)/n}$.

We can see here that LALINFERENCE produces results that indeed follow the expected relation; we can therefore conclude that the algorithm is self consistent. During the LALINFERENCE development, parallel to this investigation, this test was used as one of the primary tools to check the

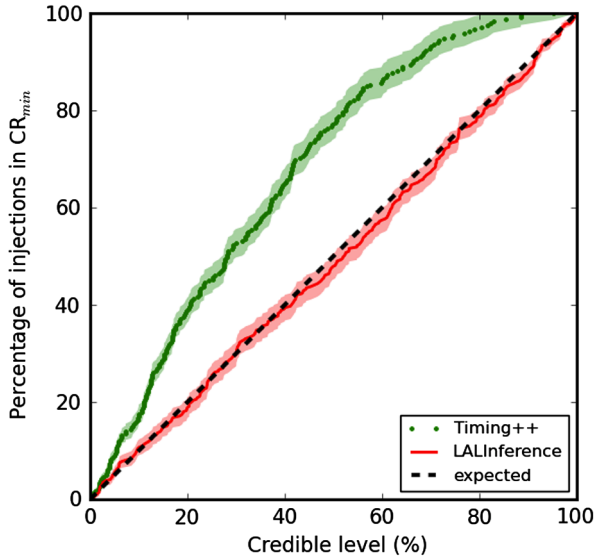

FIG. 2 (color online). For each CL we plot the number of injections that fall within the associated minimum credible region CR_{min} for all the signals analyzed with LALINFERENCE, bottom (red) curve, and TIMING++, top (green) curve. The error bars correspond to the binomial error; see text for more details. A self-consistent algorithm gives results that lie along the diagonal line of this plot. Results that fall above the expected line, as is the case for TIMING++, highlight an algorithm that is overcautious in its estimation of CR_{min}.

algorithm. As well as checking sky location, this test was done in each of the model parameters separately, though rather than using the minimal CL it is easier and sufficient to use a connected credible region whose lower bound is the lowest value of the parameter being investigated.

On the other hand, the results obtained with TIMING++ show a significant deviation from the expected behavior: the calculated CRs for TIMING++ do not represent the "true" CL. As the results are *above* the expected behavior, the sky regions are too large. This shows that TIMING++ is not "self-consistent." This is not necessarily unexpected because TIMING++ is purposefully an approximation in favor of speed; it is useful to note that TIMING++ is over conservative.

From these results it also follows that we need to be cautious when designing comparisons between TIMING++ and LALINFERENCE applied to the same GW signal. We consider these comparisons in the next section.

C. Comparisons

We can now turn to comparisons between TIMING++ and LALINFERENCE, and we consider two different figures of merit for this.

For a self-consistent code, the CR_{min} of a chosen CL is a natural metric of the ability of the algorithm to localize the source. This is equivalent to stating the expected smallest region of the sky that needs to be scanned by a follow-up observation to have a given probability that the actual source location is covered. Here, we will consider the 50% minimum credible region, and therefore set $CL = 0.5$. While this is natural for the fully coherent Bayesian codes, the same is not true of TIMING++. We saw in the previous section that TIMING++ is not self consistent: it does not provide the correct CRs at a given CL but actually overstates it.

It is, however, still interesting and possible to know the size of the CR_{min} that relates to the true CL. From the self-consistency test we have a relation between the output CRs and the true CLs from TIMING++. This means we can compare the output areas of the minimal credible regions of the true 50% CL by using the quoted 23% CR_{min} from TIMING++ and the 50% from LALINFERENCE. In other words, we are correcting for the lack of self consistency of TIMING++ and can produce a fair comparison of the two methods.

Figure 3 shows the fraction of signals whose 50% CR_{min}s were smaller than a given area. We can see that even after the corrections to the CLs are implemented, TIMING++ gives significantly larger CR_{min}s. This happens because the PDFs returned from TIMING++ are not quite the same shape as the "correct" PDFs that LALINFERENCE is returning; the differences are not simply a rescaling of the width of the peak.

While this test was quite natural from the Bayesian framework point of view, another piece of information that would be passed to follow-up telescopes would be a list of the most likely "pixels" on the sky. One can easily consider a follow-up strategy in which these tiles are observed by telescopes in order, until a possible counterpart of the GW-detected source is imaged (or one runs out of pointings). This searched area is equivalent to the size of the CR_{min}

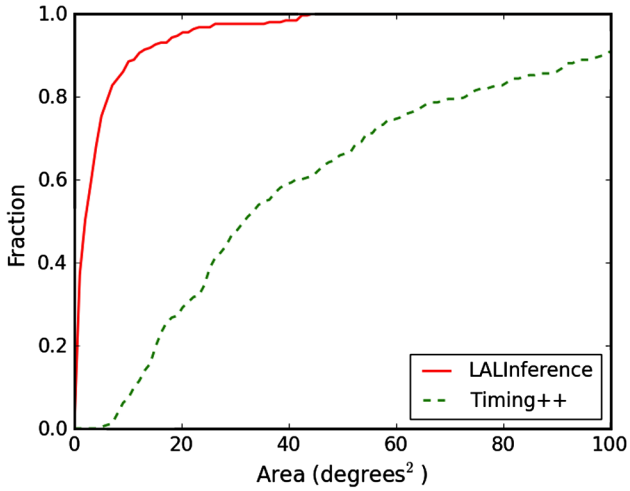

FIG. 3 (color online). The fraction of detected signals whose associated true (corrected) 50% CR_{min} covers less than a given area on the sky. We can see that LALINFERENCE gives much tighter constraints than TIMING++ on the location of a source.

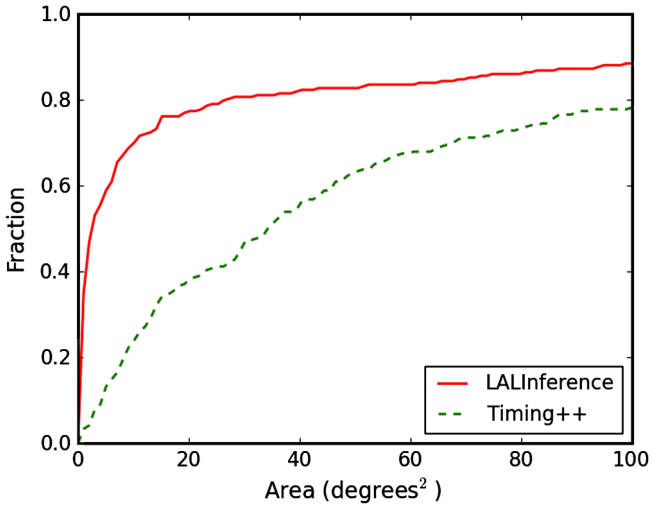

FIG. 4 (color online). The fraction of sources where the injection would have been imaged after searching less than the given area in a telescope greedy algorithm.

whose boundary passes through the source's true location on the sky. Furthermore, by considering this area for both approaches we bypass the need to correct for the true relation between probability and CL. Figure 4 shows the fraction of sources that would be imaged after only the given area is searched over, for each source, using the $\mathrm{CR}_{\mathrm{min}}$s as discussed above. We can see that there is a significant difference between the two sky localization approaches; for example, 76% of sources would be found after searching 20 deg^2 if we followed the output of LALINFERENCE, whereas we would only have found 38% of the injections by following TIMING++.

To gain a better feel for the difference in the calculated areas for the two methods, we compared the areas injection by injection. We plot the areas of the true (corrected) 50% CR found by each code where the injections are sorted by SNR (Fig. 5). For the LALINFERENCE results we can see the expected scaling of the area $\propto 1/\mathrm{SNR}^2$. We also plot the ratio of the 50% $\mathrm{CR}_{\mathrm{min}}$ areas determined by the two codes in Fig. 6. We can see that there is significant spread around the typical factor of 20 difference between the calculated $\mathrm{CR}_{\mathrm{min}}$ areas.

These results should not be taken as a statement on the expected sky localization accuracy as the underlying injection distribution is not astrophysical. The set of injections was chosen to test and compare the codes over a wide region of parameter space and should be treated as such.

D. Run time

TIMING++ has been set up with speed in mind and so the run time to extract the sky location after data is received is on the order of minutes [8]. Prior to the analysis, the distributions $p(\Delta t_{\mathrm{rss,sc}}|\vec{\Omega})$ and $p(\Delta A_{\mathrm{rss}}|\vec{\Omega})$ need to be generated, and this is done with large scale simulations. Despite being computationally expensive—the simulations require on the order of days to weeks—this step is done prior to the actual analysis and therefore has no impact on the latency of the online analysis.

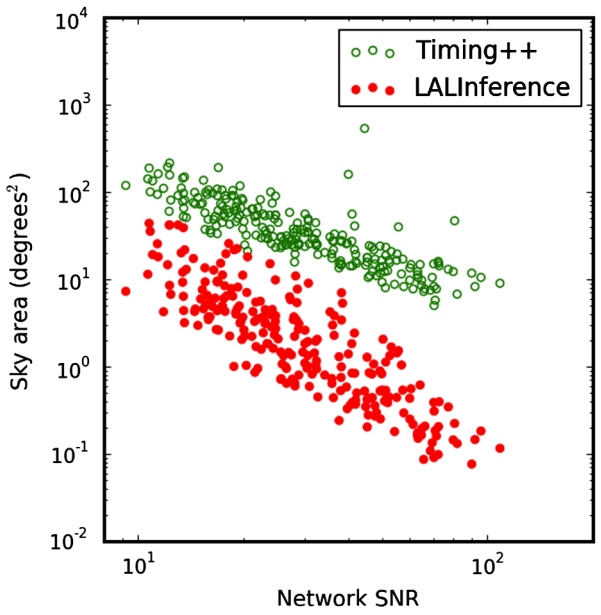

FIG. 5 (color online). The sky area of the 50% true (corrected) minimum credible region for each of the sources as a function of the optimal network SNR of the signal. While there is some scatter, the areas from LALINFERENCE [solid (red) dots] scale as $\propto 1/\mathrm{SNR}^2$, as one would expect, while the areas from TIMING++ [open (green) circles] are closer to $\propto 1/\mathrm{SNR}$.

While considering code speed, we need to specify the specific sampler used in LALINFERENCE. Here, we report results for LALINFERENCE_MCMC, the sampling method that was used for this study. A comparison between different samplers in LALINFERENCE will be reported elsewhere.

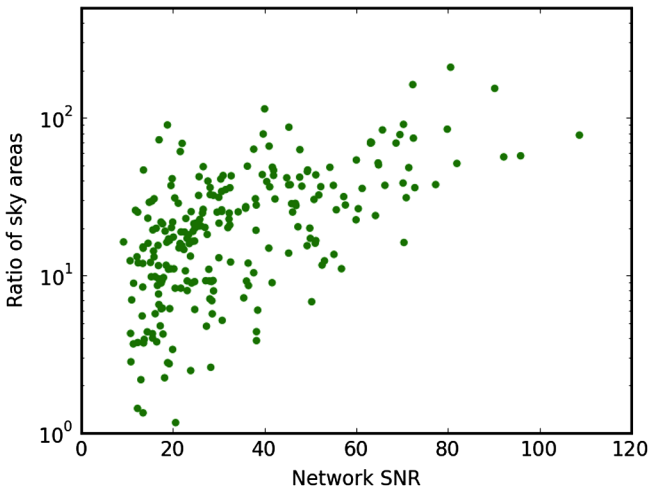

FIG. 6 (color online). The ratio of recovered areas of the 50% true (corrected) CRs using LALINFERENCE as the baseline. While there is some scatter, LALINFERENCE is consistently producing smaller areas than TIMING++ by a factor which is roughly 10 for low SNRs and approximately scales with SNR.

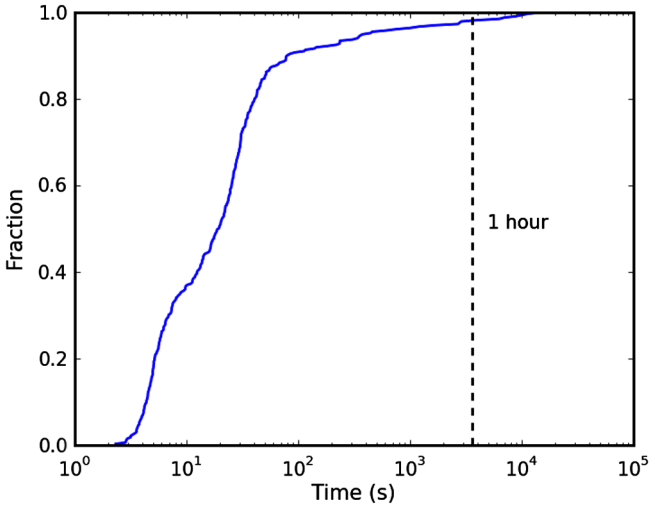

FIG. 7 (color online). The cumulative distribution of wall times for LALINFERENCE_MCMC to output a new independent sample across the runs performed to generate the results reported in this paper. With 10 cores used for each run, CPU times were a factor of 10 larger.

There are two main metrics of computational cost that we consider here: the so-called "wall time" (the time an analysis job takes from start to finish), and the total processing (CPU) time. LALINFERENCE_MCMC is designed to take advantage of multiple cores and runs in parallel on different processors. The parallel chains explore likelihoods at different contrast levels ("temperatures"). We find that roughly 10 chains are optimal for improving sampling and convergence for the data sets considered in this study; therefore, CPU times are a factor of ten larger than wall times.

The important quantity to report for LALINFERENCE is the time required to output a new independent sample of the posterior PDF. The precise number of samples that we deem necessary to describe the PDF is a balance between speed and precision; as mentioned earlier, finite sample-size issues are a concern for postprocessing, and we have found that we require at least 1000 independent samples.

In Fig. 7 we show the fraction of the analysis runs that output a single independent sample within a given wall time. This quantity was derived by dividing the total wall time of each injection run by the number of independent samples generated in that run. From this graph we can see that 90% of the runs had output 1000 independent samples in ∼14 hours of wall time. The runs were done on nodes composed of Intel Nehalem E5520 processors (2.26 GHz) with Infiniband double data rate interconnects.

IV. DISCUSSION

In this paper we have considered two sky localization algorithms, LALINFERENCE and TIMING++, used during the final science run of the LIGO and Virgo instruments in initial configuration. Our goal was to assess the relative benefits and costs of the two approaches, and to develop a strategy as well as practical tools to evaluate the consistency of the results and inform the future direction of development. We are now applying these tools to a number of parameter estimation research projects.

For the study presented in this paper we have considered a synthetic data set representing a three-detector network. GW signals generated during the inspiral phase of the coalescence of binary systems with a total mass smaller than $20M_\odot$ and nonspinning components were added to Gaussian and stationary noise representative of the sensitivity of initial LIGO. We have chosen the range of source parameters in order to best explore the performance of the algorithms. This is important for testing purposes, but one cannot draw conclusions about the actual performance of the GW instruments in future observations from these simulations. To address that question, one would need to consider an astrophysically motivated population of sources, e.g., binaries distributed uniformly in volume, and then consider sky localization only for those signals that pass a detection threshold of the search pipeline.

As discussed in Sec. III, posteriors can be systematically biased because of incorrect models, inaccurate priors, insufficient sampling or improper postprocessing to estimate credible regions.

Incorrect models are always a concern in parameter estimation. Our likelihood model, $p(d|\vec{\theta}, H)$, could be incorrect because of inaccuracies in the waveform models, noise models or calibration errors. Waveforms may not include certain features (e.g., in this study, we did not allow for spinning binary components) or are affected by limitations in the accuracy of waveform models; efforts are under way to develop more accurate and complete models [36,37] and to account for waveform uncertainty directly in parameter estimation. Real detector noise is neither stationary nor Gaussian; promising strides have been made in accounting for noise nonstationarity [38], shifts in spectral lines and even glitches in the noise. The impact of calibration errors on parameter estimation was analyzed in the context of advanced detectors [39]. In this study, our models were correct by construction, as we used stationary, Gaussian noise, assumed perfect calibrations and employed the same waveform families for injections and templates.

In this paper, we explicitly made sure that the priors assumed by LALINFERENCE were identical to the injection distribution to guarantee that inaccurate priors did not introduce a bias in the results, and our code development efforts and thorough testing ensured that insufficient sampling was not a concern.

We did find early in our studies that our initial approach to postprocessing could lead to systematically understated posterior credible regions. We addressed this by developing a more sophisticated postprocessing procedure (see below and [29]).

There is an important difference between self consistency and optimality of the results. Self consistency is a

requirement of any code that claims to provide reliable credible regions: the credible regions corresponding to a given confidence level must include the true source parameters for a fraction of signals equal to that confidence level. Optimality refers to an algorithm's ability to return the smallest credible region among all self-consistent credible regions. A self-consistent algorithm need not be optimal. When it comes to our ability to optimize, we must consider both the main algorithm and the postprocessing of the results.

As has been shown here, the proportion of available information that is utilized in the analysis can significantly affect the accuracy of parameter estimation. LALInference uses the data taken from all detectors coherently and thereby recovers small credible regions while staying self consistent. Timing++, on the other hand, purposefully makes simplifications, using intermediate data products from the incoherent analysis of individual detector data, and hence the recovered credible regions, even after a correction for self consistency, are much larger. The trade-off lies in the runtime of the analyses: Timing++ returns a sky location within minutes of the completion of the search, whereas LALInference takes approximately half a day (wall time) for the specific waveform family and network considered here.

Optimality is also important for the postprocessing of the algorithms' output to generate marginalized PDFs and credible regions. A binning scheme is traditionally applied in which the parameter space is split into a uniform grid and the average density of samples in each region found. Using a greedy approach based on this scheme to calculate optimal credible regions (CR_{min}), self consistency is broken [29]. For LALInference we have therefore implemented a more sophisticated way of setting up the initial bins known as a kD-tree so that the resolution of bins follows the density of the samples. A two-stage approach to ordering bins and estimating their contribution to the posterior is required to satisfy self consistency while managing to get close to optimality. This method will be described in full elsewhere [29]. While we have successfully applied this to two-dimensional posteriors in this study, we cannot currently extend this scheme to higher dimensions: the number of LALInference output samples required for accurate kD-tree PDF interpolation grows exponentially with the number of dimensions and so the runs become impractically long.

While we have outlined the procedure for testing that an algorithm and its implementation report self-consistent results, it is difficult to check for optimality. One approach is to set up runs where the posterior PDFs are known, which was indeed done as part of the LALInference testing and validation [19]. By design these will be simple analytic functions and there is no general prescription that will test for all circumstances.

The work that we have reported here, and the tools that we have developed and described, have already been important in the further development of LALInference. A new low-latency sky localization pipeline has also been developed [40]. It is important for future work that while we strive to improve on our methods in both speed and accuracy, we continue to validate these methods against the tests described here in order to have a reliable analysis when the next generation of detectors begins collecting data. As we move toward simultaneous and targeted electromagnetic observations of gravitational-wave sources, it is ever more important that sky localization be performed accurately and self consistently.

ACKNOWLEDGMENTS

J. V. was supported by the research program of the Foundation for Fundamental Research on Matter (FOM), which is partially supported by the Netherlands Organization for Scientific Research (NWO). N. C. was supported by the NSF Grant No. PHY-1204371. P. G. was supported by a NASA postdoctoral fellowship from the Oak Ridge Associated Universities. B. F., W. F. and V. K. were supported by the NSF Grant No. PHY-1307020, and B. F. was also supported by the NSF Grant No. DGE-0824162. V. R. was supported by a prize postdoctoral fellowship from the California Institute of Technology division of Physics, Mathematics & Astronomy and LIGO Laboratory. R. O. S. was supported by the NSF Grant No. PHY-0970074 and the UWM Research Growth Initiative. S. V. acknowledges the support of the National Science Foundation and the LIGO Laboratory. 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 no. PHY-0757058.

[1] B. Abbott *et al.* (LIGO Scientific Collaboration), Rep. Prog. Phys. **72**, 076901 (2009).
[2] T. Accadia *et al.*, JINST **7**, P03012 (2012).
[3] H. Grote and the LIGO Scientific Collaboration, Classical Quantum Gravity **25**, 114043 (2008).
[4] The LIGO Scientific Collaboration and the Virgo Collaboration, arXiv:1203.2674.
[5] G. M. Harry and the LIGO Scientific Collaboration, Classical Quantum Gravity **27**, 084006 (2010).

[6] Virgo Collaboration, Report No. VIR-0027A-09, 2009, https://tds.ego-gw.it/itf/tds/file.php?callFile=VIR-0027A-09.pdf.
[7] J. Abadie *et al.*, The LIGO Scientific Collaboration, the Virgo Collaboration, Classical Quantum Gravity **27**, 173001 (2010).
[8] J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia, F. Acernese, C. Adams, R. Adhikari, C. Affeldt *et al.*, Astron. Astrophys. **541**, A155 (2012).
[9] J. Aasi, J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia, F. Acernese *et al.*, The LIGO Scientific Collaboration, the Virgo Collaboration, Phys. Rev. D **88**, 062001 (2013).
[10] B. D. Metzger and E. Berger, Astrophys. J. **746**, 48 (2012).
[11] S. Nissanke, J. Sievers, N. Dalal, and D. Holz, Astrophys. J. **739**, 99 (2011).
[12] L. Z. Kelley, E. Ramirez-Ruiz, M. Zemp, J. Diemand, and I. Mandel, Astrophys. J. **725**, L91 (2010).
[13] F. Cavalier, M. Barsuglia, M.-A. Bizouard, V. Brisson, A.-C. Clapson, M. Davier, P. Hello, S. Kreckelbergh, N. Leroy, and M. Varvella, Phys. Rev. D **74**, 082004 (2006).
[14] J. Veitch, I. Mandel, B. Aylott, B. Farr, V. Raymond, C. Rodriguez, M. van der Sluys, V. Kalogera, and A. Vecchio, Phys. Rev. D **85**, 104045 (2012).
[15] S. Nissanke, M. Kasliwal, and A. Georgieva, Astrophys. J. **767**, 124 (2013).
[16] M. M. Kasliwal and S. Nissanke, arXiv:1309.1554.
[17] S. Fairhurst, Classical Quantum Gravity **28**, 105021 (2011).
[18] K. Grover, S. Fairhurst, B. F. Farr, I. Mandel, C. Rodriguez, T. Sidery, and A. Vecchio, Phys. Rev. D **89**, 042004 (2014).
[19] B. Farr *et al.* (to be published).
[20] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics Series (Chapman & Hall//CRC 1995, London, 1996), ISBN 9780412055515.
[21] N. Christensen, R. Meyer, and A. Libson, Classical Quantum Gravity **21**, 317 (2004).
[22] M. van der Sluys, V. Raymond, I. Mandel, C. Röver, N. Christensen, V. Kalogera, R. Meyer, and A. Vecchio, Classical Quantum Gravity **25**, 184011 (2008).
[23] J. Skilling, Bayesian Analysis **1**, 833 (2006).
[24] J. Veitch and A. Vecchio, Phys. Rev. D **81**, 062003 (2010).
[25] F. Feroz and M. P. Hobson, Mon. Not. R. Astron. Soc. **384**, 449 (2008).
[26] S. Babak, R. Biswas, P. Brady, D. Brown, K. Cannon *et al.*, Phys. Rev. D **87**, 024033 (2013).
[27] LSC Algorithm Library software packages LAL, LALWRAPPER, and LALAPPS, http://www.lsc-group.phys.uwm.edu/lal.
[28] P. Graff, F. Feroz, M. P. Hobson, and A. Lasenby, Mon. Not. R. Astron. Soc. **421**, 169 (2012).
[29] T. Sidery, J. Gair, I. Mandel, and W. Farr (to be published).
[30] F. Acernese, P. Amico, M. Alshourbagy, F. Antonucci, S. Aoudia, P. Astone, S. Avino, D. Babusci, G. Ballardin, F. Barone *et al.*, Classical Quantum Gravity **24**, S617 (2007).
[31] W. G. Anderson, P. R. Brady, J. D. Creighton, and É. É. Flanagan, Phys. Rev. D **63**, 042003 (2001).
[32] C. L. Rodriguez, B. Farr, W. M. Farr, and I. Mandel, Phys. Rev. D **88**, 084013 (2013).
[33] J. Aasi *et al.* (to be published).
[34] A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, Phys. Rev. D **80**, 084043 (2009).
[35] J. Abadie, B. P. Abbott, R. Abbott, T. D. Abbott, M. Abernathy, T. Accadia *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Phys. Rev. D **85**, 082002 (2012).
[36] M. Hannam, P. Schmidt, A. Bohé, L. Haegel, S. Husa, F. Ohme, G. Pratten, and M. Pürrer, arXiv:1308.3271.
[37] A. Taracchini, A. Buonanno, Y. Pan, T. Hinderer, M. Boyle, D. A. Hemberger, L. E. Kidder, G. Lovelace, A. H. Mroue, H. P. Pfeiffer *et al.*, Phys. Rev. D **89**, 061502 (2014).
[38] T. B. Littenberg, M. Coughlin, B. Farr, and W. M. Farr, Phys. Rev. D **88**, 084044 (2013).
[39] S. Vitale, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, I. Mandel, B. Aylott, and J. Veitch, Phys. Rev. D **85**, 064034 (2012).
[40] L. Singer and L. Price (to be published).