Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data

Cody Messick,[1,2,*] Kent Blackburn,[3] Patrick Brady,[4] Patrick Brockill,[4] Kipp Cannon,[5,6] Romain Cariou,[7]
Sarah Caudill,[4] Sydney J. Chamberlin,[1,2] Jolien D. E. Creighton,[4] Ryan Everett,[1,2] Chad Hanna,[1,8,2]
Drew Keppel,[9] Ryan N. Lang,[4] Tjonnie G. F. Li,[10] Duncan Meacher,[1,2] Alex Nielsen,[9] Chris Pankow,[11]
Stephen Privitera,[12] Hong Qi,[4] Surabhi Sachdev,[3] Laleh Sadeghian,[4] Leo Singer,[13] E. Gareth Thomas,[14]
Leslie Wade,[15] Madeline Wade,[15] Alan Weinstein,[3] and Karsten Wiesner[9]

[1]*Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*
[2]*Institute for Gravitation and the Cosmos, The Pennsylvania State University,
University Park, Pennsylvania 16802, USA*
[3]*LIGO Laboratory, California Institute of Technology, MS 100-36, Pasadena, California 91125, USA*
[4]*Leonard E. Parker Center for Gravitation, Cosmology, and Astrophysics,
University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53201, USA*
[5]*Canadian Institute for Theoretical Astrophysics, 60 St. George Street, University of Toronto,
Toronto, Ontario M5S 3H8, Canada*
[6]*RESCEU, University of Tokyo, Tokyo 113-0033, Japan*
[7]*Département de physique, École Normale Supérieure de Cachan, 94230 Cachan, France*
[8]*Department of Astronomy and Astrophysics, The Pennsylvania State University,
University Park, Pennsylvania 16802, USA*
[9]*Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, D-30167 Hannover, Germany*
[10]*Department of Physics, The Chinese University of Hong Kong,
Shatin, New Territories, Hong Kong, China*
[11]*Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA)
and Department of Physics and Astronomy, Northwestern University,
2145 Sheridan Road, Evanston, Illinois 60208, USA*
[12]*Albert-Einstein-Institut, Max-Planck-Institut für Gravitationsphysik, D-14476 Potsdam-Golm, Germany*
[13]*NASA/Goddard Space Flight Center, Greenbelt, Maryland 20771, USA*
[14]*University of Birmingham, Birmingham, B15 2TT, United Kingdom*
[15]*Department of Physics, Hayes Hall, Kenyon College, Gambier, Ohio 43022, USA*

(Received 14 June 2016; published 7 February 2017)

We describe a stream-based analysis pipeline to detect gravitational waves from the merger of binary neutron stars, binary black holes, and neutron-star–black-hole binaries within ~1 min of the arrival of the merger signal at Earth. Such low-latency detection is crucial for the prompt response by electromagnetic facilities in order to observe any fading electromagnetic counterparts that might be produced by mergers involving at least one neutron star. Even for systems expected not to produce counterparts, low-latency analysis of the data is useful for deciding when not to point telescopes, and as feedback to observatory operations. Analysts using this pipeline were the first to identify GW151226, the second gravitational-wave event ever detected. The pipeline also operates in an offline mode, in which it incorporates more refined information about data quality and employs acausal methods that are inapplicable to the online mode. The pipeline's offline mode was used in the detection of the first two gravitational-wave events, GW150914 and GW151226, as well as the identification of a third candidate, LVT151012.

DOI: 10.1103/PhysRevD.95.042001

I. INTRODUCTION

The field of gravitational-wave astronomy has come to life in a spectacular way, with the first detections of gravitational waves on September 14, 2015 [1] and December 26, 2015 [2] by the two detectors of the Laser Interferometer Gravitational-wave Observatory (LIGO) [3]. These detectors are currently undergoing further commissioning and will reach design sensitivity in the next few years. Additionally, they will be joined by a network of gravitational-wave observatories that include Advanced Virgo [4], KAGRA [5], and a third LIGO observatory in India [6]. We expect this network to bring more observations of binary black hole mergers [7], as well as binary neutron star (BNS) and neutron-star–black-hole (NSBH) mergers [8].

As we enter the era of gravitational-wave astronomy, the need for low-latency analyses becomes critical. Gravitational waves from BNS and NSBH mergers are

[*] Cody.Messick@ligo.org

expected to be paired with electromagnetic emission and neutrinos [9–11]. Gravitational-wave-triggered electromagnetic observations may lead to the detection of prompt short gamma-ray bursts and high-energy neutrinos within seconds, followed by x-ray, optical, and radio afterglows days to years later. Multimessenger observations will aid in our understanding of astrophysical processes and increase our search sensitivity [9,12]. Additionally, even in the absence of a counterpart, the rapid identification of gravitational waves has a number of benefits. Low-latency detection allows us to provide feedback to commissioners when search sensitivity drops unexpectedly, helping to return the detector to its nominal state [13]. Furthermore, upon identification of a candidate, we can submit timely requests to minimize detector changes in order to gather enough data to reliably estimate the search background and perform follow-up calibration measurements.

In this work, we present the GstLAL-based inspiral pipeline, a gravitational-wave search pipeline based on the GstLAL library [14], and derived from GStreamer [15] and the LIGO Algorithm Library [16]. The pipeline can operate in a low-latency mode to ascertain whether a gravitational-wave signal is present in data, provide point estimates for the binary parameters, and estimate event significance. Analysts running the low-latency mode of this pipeline were the first to identify the second gravitational wave event detected, GW151226 [2]. The pipeline can also operate in an "offline" configuration that can be used to process archival gravitational-wave data with additional background statistics and data quality information. The offline configuration was used in the detection of GW150914, LVT151012 [17], and GW151226 [2].

The GstLAL-based inspiral pipeline expands on the parameter space covered by previous low-latency searches [18–21]. In addition, it extends many of the techniques used in prior searches for compact binary coalescences [22,23] to operate in a fully parallel, stream-based mode that allows for the identification of candidate gravitational-wave events within seconds of recording the data. The key differences include the following: (1) time-domain [24] rather than frequency-domain [25] matched filtering, (2) time-domain rather than frequency-domain [26] signal consistency tests to reject nonstationary noise transients, (3) a multidimensional likelihood ratio ranking statistic to robustly identify gravitational-wave candidates in a way that automatically adjusts to the properties of the noise [27], and (4) a background estimation technique that relies on tracking noise distributions to allow rapid evaluation of significance of identified candidates [28]. For a discussion of performance differences between time-domain and frequency-domain matched filtering, the reader is referred to Ref. [24].

This paper is organized as follows: In Sec. II, we discuss inputs to the low-latency and offline analyses, the online acquisition of data, measurement of the power spectral density (PSD), and whitening and conditioning of the data for matched filtering. We also present the basic offline and low-latency workflows in Figs. 1 and 2, respectively. In Sec. III, we discuss the matched-filter algorithm and our procedure for producing a list of ranked candidate events.

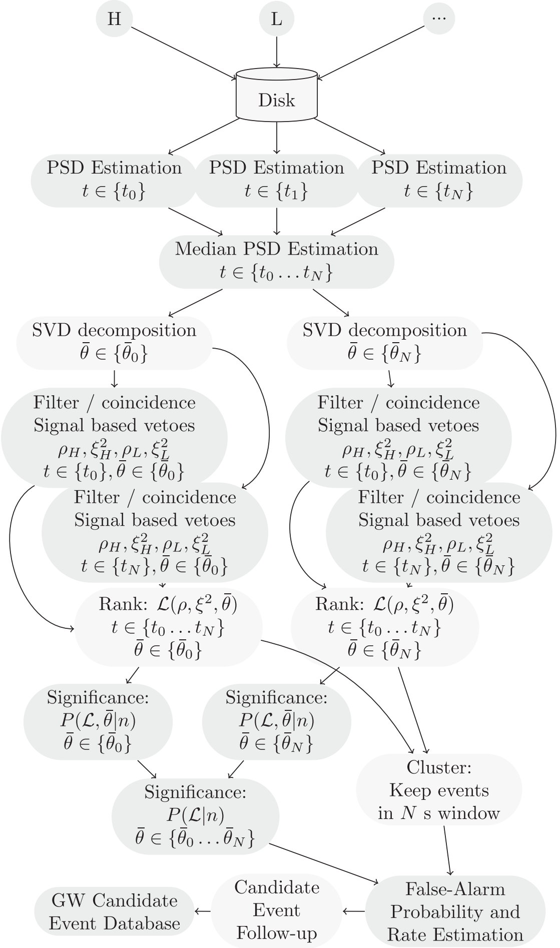
H
L
...
Disk
PSD Estimation $t \in \{t_0\}$
PSD Estimation $t \in \{t_1\}$
PSD Estimation $t \in \{t_N\}$
Median PSD Estimation $t \in \{t_0 \dots t_N\}$
SVD decomposition $\bar{\theta} \in \{\bar{\theta}_0\}$
SVD decomposition $\bar{\theta} \in \{\bar{\theta}_N\}$
Filter / coincidence Signal based vetoes $\rho_H, \xi^2_H, \rho_L, \xi^2_L$ $t \in \{t_0\}, \bar{\theta} \in \{\bar{\theta}_0\}$
Filter / coincidence Signal based vetoes $\rho_H, \xi^2_H, \rho_L, \xi^2_L$ $t \in \{t_0\}, \bar{\theta} \in \{\bar{\theta}_N\}$
Filter / coincidence Signal based vetoes $\rho_H, \xi^2_H, \rho_L, \xi^2_L$ $t \in \{t_N\}, \bar{\theta} \in \{\bar{\theta}_0\}$
Filter / coincidence Signal based vetoes $\rho_H, \xi^2_H, \rho_L, \xi^2_L$ $t \in \{t_N\}, \bar{\theta} \in \{\bar{\theta}_N\}$
Rank: $\mathcal{L}(\rho, \xi^2, \bar{\theta})$ $t \in \{t_0 \dots t_N\}$ $\bar{\theta} \in \{\bar{\theta}_0\}$
Rank: $\mathcal{L}(\rho, \xi^2, \bar{\theta})$ $t \in \{t_0 \dots t_N\}$ $\bar{\theta} \in \{\bar{\theta}_N\}$
Significance: $P(\mathcal{L}, \bar{\theta}|n)$ $\bar{\theta} \in \{\bar{\theta}_0\}$
Significance: $P(\mathcal{L}, \bar{\theta}|n)$ $\bar{\theta} \in \{\bar{\theta}_N\}$
Significance: $P(\mathcal{L}|n)$ $\bar{\theta} \in \{\bar{\theta}_0 \dots \bar{\theta}_N\}$
Cluster: Keep events in N s window
False-Alarm Probability and Rate Estimation
Candidate Event Follow-up
GW Candidate Event Database

FIG. 1. Diagram of the offline search mode of the GstLAL-based inspiral pipeline. First, data are transferred from each observatory $(H, L, \ldots)$ to a central computing cluster (Sec. II A). Next, data are read from the disk, and the PSD is estimated (Sec. II B) in chunks of time $t_0, t_1, \ldots, t_N$ for each observatory. The median over the entire analysis time of each observatory PSD estimate is computed. The input template bank, which is generated upstream of the analysis, is split into regions of similar parameters $\bar{\theta}_0, \bar{\theta}_1, \ldots, \bar{\theta}_N$ (Sec. II D) and then decomposed into a set of orthonormal filters weighted by the median PSD for each observatory. The data are filtered to produce a series of triggers characterized by the signal-to-noise ratio (SNR), ρ, signal consistency check, ξ^2 (Secs. III A, III B, and III C), and coalescence time. Coincident triggers between detectors are identified and promoted to the status of events (Sec. III D). Events are ranked according to their relative probability of arising from signal versus noise (Sec. III E). The data are then reduced to the most highly ranked event in 8 s windows (Sec. III F). In parallel, triggers not found in coincidence are used to construct the probability of obtaining a given event from noise, $P(\Lambda|n)$. Finally, the event significance and false-alarm rate are estimated (Sec. IV A). Note that the arrows drawn between nodes in this diagram do not imply the output of one node is the input of the next node; they simply indicate the order in which tasks are performed.

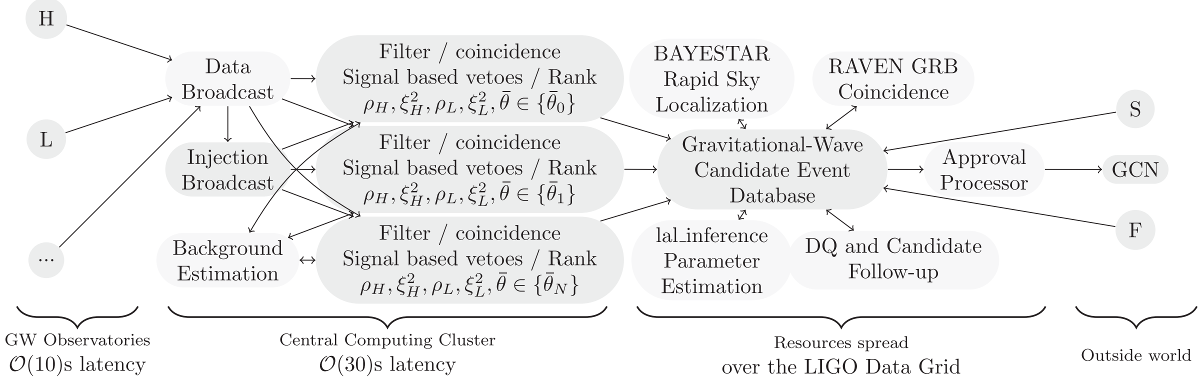
H
L
…
Data Broadcast
Injection Broadcast
Background Estimation
Filter / coincidence Signal based vetoes / Rank $\rho_H, \xi^2_H, \rho_L, \xi^2_L, \bar{\theta} \in \{\bar{\theta}_0\}$
Filter / coincidence Signal based vetoes / Rank $\rho_H, \xi^2_H, \rho_L, \xi^2_L, \bar{\theta} \in \{\bar{\theta}_1\}$
Filter / coincidence Signal based vetoes / Rank $\rho_H, \xi^2_H, \rho_L, \xi^2_L, \bar{\theta} \in \{\bar{\theta}_N\}$
BAYESTAR Rapid Sky Localization
RAVEN GRB Coincidence
Gravitational-Wave Candidate Event Database
lal_inference Parameter Estimation
DQ and Candidate Follow-up
Approval Processor
S
GCN
F
GW Observatories $\mathcal{O}(10)$s latency
Central Computing Cluster $\mathcal{O}(30)$s latency
Resources spread over the LIGO Data Grid
Outside world

FIG. 2. Diagram of the low-latency search mode of the GstLAL-based inspiral pipeline. First, data are received over a network connection from each observatory to a data broadcaster in a central computing facility. The data are then broadcast over the entire cluster with an efficient multicast protocol. The online analysis uses precomputed bank decompositions for each observatory from reference PSDs as input to jobs that combine the filtering, vetoing, coincidence, ranking, and significance estimation steps from the offline pipeline. Unlike the offline case, the online analysis workflow cannot be described as a directed acyclic graph, and in fact, data from each filtering job are exchanged bidirectionally and asynchronously to a process that constantly evaluates the global background estimates for the entire analysis. Events that are identified by any one filtering job, and subsequently pass a predetermined significance threshold, are sent to the Gravitational-Wave Candidate Event Database (GraceDB) [29] within a matter of seconds of the data being recorded at the observatories.

In Sec. IV, we explain the significance calculation for identified candidate events and the procedure for responding to significant events via alerts to our observing partners. Differences between the offline and low-latency operation modes will be highlighted when relevant.

II. MATCHED FILTERING INPUT

Matched filtering algorithms for compact binary mergers have traditionally filtered the data $d(t)$ against a set of complex template waveforms $\{h^c_i(t)\}$ in the frequency domain using the relation

$$z_i(t) = x_i(t) + iy_i(t) = 4\int_0^\infty \mathrm{d}f \frac{\tilde{h}_i^{c*}(f)\tilde{d}(f)}{S_n(f)} e^{2\pi i f t}, \qquad (1)$$

where $z_i(t)$ is the complex SNR using the ith template, $x_i(t)$ is the matched filter response to a gravitational wave signal with orbital coalescence phase ϕ_0 (the real part of the template in the time domain), $y_i(t)$ is the matched filter response to the same signal with orbital coalescence phase $\phi_0 + \pi/4$ (the imaginary part of the template in the time domain), and $S_n(f)$ is the single-sided noise PSD. The templates are normalized such that

$$1 = 4\int_0^\infty \mathrm{d}f \frac{|\tilde{h}_i^c(f)|^2}{S_n(f)}. \qquad (2)$$

Defining the SNR, $\rho(t)$, as the modulus of the complex SNR, $z(t)$, allows one to search efficiently over the unknown coalescence phase, while its expression in the frequency domain allows one to efficiently implement matched filtering using fast Fourier transform (FFT) routines.

The GstLAL-based inspiral pipeline, however, performs matched filtering in the time domain with real templates $\{h_i(t)\}$. The matched filter output is thus the real-valued $x_i(t)$ instead of the complex-valued $z_i(t)$. We can recast Eq. (1) in the time domain using the convolution theorem, which gives

$$x_i(t) = 2\int_{-\infty}^\infty \mathrm{d}f \frac{\tilde{h}_i^*(f)\tilde{d}(f)}{S_n(|f|)} e^{2\pi i f t} \qquad (3a)$$

$$= 2\int_{-\infty}^\infty \mathrm{d}\tau \hat{h}_i(\tau)\hat{d}(t+\tau), \qquad (3b)$$

where

$$\hat{d}(\tau) = \int_{-\infty}^\infty \mathrm{d}f \frac{\tilde{d}(f)}{\sqrt{S_n(|f|)}} e^{2\pi i f \tau} \qquad (4)$$

is the whitened data; the whitened template $\hat{h}_i(\tau)$ is defined similarly. As a consequence of using real templates, Eq. (3) returns the matched filter response to a single coalescence phase, while Eq. (1) returns the response to two phases. Thus, Eq. (3) must be evaluated a second time using the template corresponding to the $\pi/4$-shifted phase in order to compute the SNR. To account for using twice as many templates, the template index, i, used on real templates is related to the index used on complex templates via

$$h_{2i}(t) = \mathrm{Re}[h^c_i(t)], \qquad (5a)$$

$$h_{2i+1}(t) = \mathrm{Im}[h_i^{\mathrm{c}}(t)]. \tag{5b}$$

A different template normalization is also used, specifically

$$1 = 4\int_0^\infty \mathrm{d}f \frac{|\tilde{h}_i(f)|^2}{S_n(f)}. \tag{6}$$

In this section, we discuss the inputs to the time-domain matched filtering calculation expressed in Eq. (3). We begin by discussing the low-latency distribution of the data themselves. We then describe our method for estimating the PSD, which we use to construct the whitened data $\hat{d}(t)$ and whitened templates $\hat{h}_i(t)$. In describing our construction of the whitened data stream, we also describe the removal of loud noise transients and dealing with data dropouts in the low-latency broadcast. Finally, we describe the construction of the whitened template filters, which involves a number of computational enhancements to reduce the cost of filtering in the time domain.

A. Data acquisition

Gravitational-wave strain data acquired at the LIGO sites are digitized at a sample rate of 16384 Hz and bundled into interferometric gravitational wave detector (IGWD) frames, a custom LIGO file format described in Ref. [30], on a four-second cadence. Information about the state of the instrument and data quality is distilled from a host of auxiliary environmental and instrumental-control-system channels into a single channel, referred to as the state-vector channel. The four-second frames containing the gravitational-wave strain and state-vector channels are delivered for low-latency processing at computing clusters across the LIGO Data Grid within ∼12 s of the data being acquired.

Searches for compact binary coalescences require using hundreds or thousands of compute nodes in parallel to process all the possible template waveforms. Low-latency data must be made available to all of these nodes as soon as they arrive; thus an efficient multicast protocol is used to broadcast the data in low latency to the entire cluster. The nature of the low-latency transmission causes some small data loss within the tolerances acceptable to the pipeline, with efforts underway to reduce these losses.

B. PSD

Abstractly, we define the (one-sided) noise power spectral density $S_n(f)$ as

$$\langle \tilde{n}(f)\tilde{n}^*(f')\rangle = \frac{1}{2}S_n(f)\delta(f-f'), f > 0, \tag{7}$$

where $\langle \cdots \rangle$ denotes an ensemble average over realizations of the detector noise $n(t)$, which is assumed to be stationary and Gaussian. In practice, we cannot use Eq. (7) to calculate the PSD for a variety of reasons. To begin with, our knowledge of the detector noise comes exclusively from the observed data, which may contain signal in addition to noise. Furthermore, real data may contain brief departures from stationarity (commonly called "glitches"), which we do not want to contribute to the PSD estimate. Finally, the PSD can drift slowly over time scales shorter than the duration of a typical detector lock segment, and we want to track these changes. For low-latency applications, we also require a PSD estimate that converges quickly using only data from the past, so that we obtain an accurate estimate of the PSD as soon as possible after the data begin to flow. In this subsection, we discuss the PSD estimation algorithm and how the result is used to whiten the data and template bank. We also present the results of a study done on the convergence of an estimated PSD to its known spectrum.

1. Estimation and whitening

We use a median and a running geometric mean to meet these requirements for each analyzed segment of data. The median estimate operates on medium time scales and is robust against shorter time-scale fluctuations in the noise, while the running geometric mean tracks longer time-scale changes in the PSD, averaging the PSD estimates with the most recent estimates weighted more strongly. The time scales of the median and geometric mean are set, respectively, by the tunable parameters n_{med} and n_{avg}.

The PSD calculation begins by partitioning the strain time series into blocks of length N points with each block overlapping the previous by $N/2 + Z$ points, where N and Z are even-valued integers. Each block of data, denoted $d_j[k]$, is windowed and Fourier transformed,

$$\tilde{d}_j[\ell] = \sqrt{\frac{N}{\sum_{k=0}^{N-1} w[k]^2}}\Delta t \sum_{k=0}^{N-1} d_j[k]w[k]e^{-2\pi i\ell k/N}, \tag{8}$$

where $k \in [0, N-1]$ is the time index, $\ell \in [0, \frac{N}{2}]$ is the frequency index, Δt is the time sample step, and

$$w[k] = \begin{cases} 0, & 0 \le k < Z \\ \sin^2\frac{\pi(k-Z)}{N-2Z}, & Z \le k < N-Z \\ 0, & N-Z \le k \le N-1 \end{cases} \tag{9}$$

is a zero-padded Hann window function. The mean and Nyquist terms of Eq. (8), $\tilde{d}[0]$ and $\tilde{d}[N/2]$, are set to zero, and the zero-padded Hann window is defined such that the sequence of overlapping window functions sum to unity everywhere. The squared magnitude of Eq. (8) is proportional to the instantaneous PSD and has a frequency resolution of $\Delta f = \frac{1}{N\Delta t}$.

The median of the most recent n_{med} instances of the instantaneous PSD, $S_j^{\mathrm{med}}[\ell]$, is determined for each frequency bin ℓ. Mathematically,

$$S_j^{\mathrm{med}}[\ell] = \mathrm{median}\{2\Delta f|\tilde{d}_k[\ell]|^2\}_{k=j-n_{\mathrm{med}}}^{k=j}. \quad (10)$$

The median is relatively insensitive to short time-scale fluctuations, which must occur over a time scale of $\frac{1}{2}n_{\mathrm{med}}(N/2 - Z)\Delta t$ in order to affect the median.

The median is used to estimate the geometric mean of the last n_{med} samples for each frequency bin. Assuming the noise is a stationary, Gaussian process allows us to assume that the measured frequency bins of the estimated PSD are χ^2-distributed random variables. The geometric mean of a χ^2-distributed random variable is equal to the median divided by a proportionality constant β. The logarithm of the running geometric mean of median estimated PSDs, $\log S_j[\ell]$, is computed from one part $\log S_j^{\mathrm{med}}[\ell]/\beta$ and $(n_{\mathrm{avg}} - 1)$ parts $\log S_{j-1}[\ell]$. Mathematically,

$$S_j[\ell] = \exp\left[\frac{n_{\mathrm{avg}} - 1}{n_{\mathrm{avg}}}\log S_{j-1}[\ell] + \frac{1}{n_{\mathrm{avg}}}\log\frac{S_j^{\mathrm{med}}[\ell]}{\beta}\right]. \quad (11)$$

Changes to the PSD must occur over a time scale of at least $n_{\mathrm{avg}}(N/2 - Z)\Delta t$ to be fully accounted for by Eq. (11).

To whiten the data and the templates, the arithmetic mean is estimated from the geometric mean. The arithmetic mean of a χ^2-distributed random variable is equal to the geometric mean multiplied by $\exp(\gamma)$, where γ is Euler's constant. If the noise assumptions are violated, then the true arithmetic mean of the spectrum will differ from the measured spectrum by some unknown factor. This estimated arithmetic mean is referred to as $S_n(f)$ in the continuum limit [see, e.g., Eq. (7)].

The low-latency operating mode must whiten the data and update the running geometric mean of the PSD at the same time. The whitening process is done after the running geometric mean has been updated and is performed by dividing each frequency bin of Eq. (8) by the square root of the corresponding frequency bin in the estimated arithmetic mean of the PSD. Mathematically,

$$\tilde{\tilde{d}}_j[\ell] = \frac{\tilde{d}_j[\ell]}{\sqrt{S_j[\ell]\exp(\gamma)}}, \quad (12)$$

$$\hat{d}_j[k] = 2\Delta t\sqrt{\sum_{m=0}^{N-1} w[m]^2\Delta f}\sum_{\ell=0}^{N/2}\tilde{\tilde{d}}_j[\ell]e^{2\pi i\ell k/N}. \quad (13)$$

The extra terms in the inverse Fourier transform are necessary for unity variance.

The low-latency analysis typically uses $N = f_s(8\ \mathrm{s})$ and $Z = f_s(2\ \mathrm{s})$, where $f_s = 1/\Delta t$ is the sampling frequency, resulting in 1/4 Hz frequency resolution. This introduces four seconds of latency into the analysis. Unlike the low-latency case, the offline analysis begins with a known list of data segments. The PSD of each segment is estimated using $N = f_s(32\ \mathrm{s})$ and $Z = 0$; the final result of the running average is written to disk as a "reference PSD." The median of the reference PSDs is used to whiten the template bank before matched filtering. However, the data segments are whitened in a procedure similar to the low-latency analysis, using $N = f_s(32\ \mathrm{s})$ and $Z = f_s(8\ \mathrm{s})$. At the time of writing, the typical values used are $n_{\mathrm{med}} = 7$ and $n_{\mathrm{avg}} = 64$ for both modes of operation. The only procedural difference between the offline and low-latency whitening steps is that the offline analysis seeds the running average with the segment's reference PSD.

2. Convergence

For low-latency applications, we require a PSD estimate that converges quickly using only data from the past, so that we obtain an accurate estimate of the PSD as soon as possible after the data begin to flow. To quantify the convergence, we create noise with a known power spectrum and compute a quantity that is proportional to the expected SNR for a given PSD in the absence of noise (commonly referred to as the "optimal SNR"). In the stationary phase approximation, binary waveforms in the frequency domain, $h(f)$, are proportional to $f^{-7/6}$ [31], and thus

$$\rho \propto \int_{f_1}^{f_2} \mathrm{d}f\frac{f^{-7/3}}{S_n(f)}. \quad (14)$$

We choose $f_1 = 10$ Hz and $f_2 = 2048$ Hz. Specifically, we compare the quantity computed using the *measured*

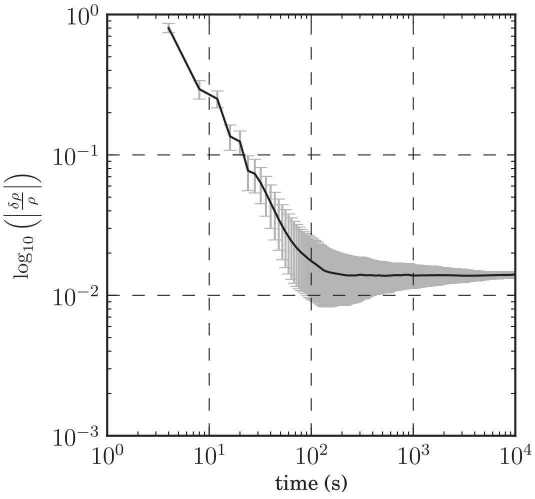

FIG. 3. PSD convergence properties. Estimating the PSD is a critical part of ensuring that events are detected and assigned the appropriate significance. This figure illustrates the convergence properties of the PSD estimation in terms of the impact on SNR. Within 20 s the PSD will have an $\mathcal{O}(10\%)$ impact on SNR, and within 200 s the impact drops to $\mathcal{O}(1\%)$, where it remains.

spectrum $S_n(f)$, which we denote simply as ρ, to the SNR computed from the *known* spectrum $\hat{S}_n(f)$, which we denote as $\hat{\rho}$. Figure 3 shows the fractional change of ρ with respect to $\hat{\rho}$ as a function of time,

$$\frac{\delta\rho}{\rho}(t) = \frac{\rho(t) - \hat{\rho}}{\rho(t)}. \tag{15}$$

We find that convergence happens quickly relative to the length of the data. The approximation of the true PSD does not affect the measured SNR after tens of seconds.

C. Data conditioning

Matched filtering is optimal under the condition that the noise, $n(t)$, is both stationary and Gaussian. Although nonstationarity on long time scales can be handled by tracking the PSD, short noise transients, commonly referred to as glitches, can cause high-SNR matched filter outputs that mimic signal detections. Glitches are handled by either removing them from the data or using signal consistency checks to vet the matched filter output. Section III C provides more details on the latter.

The GstLAL-based inspiral pipeline removes glitches from the data in two ways. In some cases, the matched filter outputs of glitches have considerably higher amplitudes than any expected output from a compact binary signal and can thus be safely removed from the data through a process called gating. Once the data have been whitened, they have unit variance. If a momentary excursion greater than some number of standard deviations, σ, is observed in the whitened data, then the gating process zeros the excursion in the whitened data with a 0.25 s padding on each side. An example of this is shown in Fig. 4. When gating the strain data, care must be taken to choose a threshold that will not discard real gravitational wave signals. The threshold is chosen by testing with simulated gravitational wave signals.

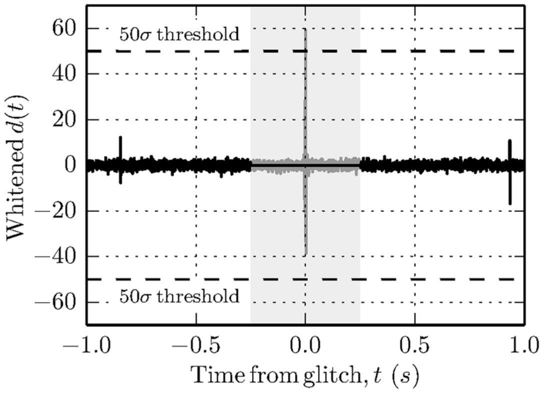

FIG. 4. Data conditioning. In this two second block of LIGO S6 data, three noise transients (glitches) are visible. The glitch at time zero surpassed the threshold of 50 standard deviations (σ), triggering the gate to veto a ±0.25 s window around the glitch by replacing the data with zeros (black). The gray trace shows what the data looked like prior to gating.

The choice of 0.25 s padding is conservative for LIGO PSDs, where the whitening filter in the time domain can be approximated as a narrow sinc function. An initial LIGO PSD and the time domain representation of its corresponding whitening filter, estimated from data taken during the initial LIGO's sixth science run [32,33] (referred to as S6), are shown in Fig. 5. The ~0.98 whitening filter's square magnitude is contained within ±10 ms of the filter's peak; thus we expect no significant spectral leakage when gating glitches with 0.25 s padding.

In many cases, auxiliary information is available through environmental and instrumental monitors that can ascertain times of clear coupling between local transient noise sources [34,35]; in cases where data quality is known to be poor, vetoes are applied after the strain data are whitened. Since whitened data are, by definition, uncorrelated between adjacent samples for stationary Gaussian processes, vetoes are applied by simply replacing the whitened data during vetoed times with zeros.

Large noise transients occasionally remain in the absence of clear instrumental or environmental coupling.

In some cases, the matched filter outputs of these glitches have considerably higher amplitude than any expected output from a compact binary signal and can thus be safely removed from the data through a process called gating. Once the data have been whitened, it has unit variance. If a momentary excursion greater than some number of standard deviations, σ, is observed in the whitened data, then the gating process zeros the excursion in the whitened data with a 0.25 s padding on each side. An example of this is shown in Fig. 4. When gating the strain data, care must be

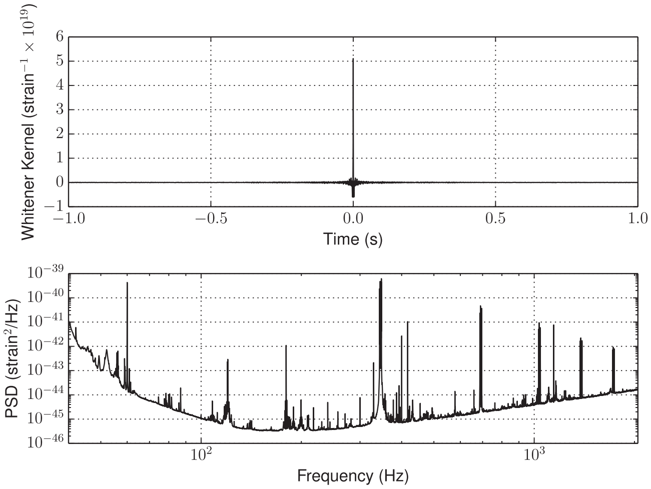

FIG. 5. Top: Time domain representation of the whitening filter computed from a PSD estimated in an analysis of one week of S6 data. The ~0.98 of the filter's square magnitude is contained enclosed within ±10 ms of the peak. Bottom: The PSD used to compute the whitening filter.

taken to choose a threshold that will not discard real gravitational wave signals. The threshold is chosen by testing with simulated gravitational wave signals.

D. Template bank decomposition

In order to detect any compact binaries within a region of the mass parameter space, we filter the data against a bank of template signals. As the true binary parameter space is continuous, actual signals may not exactly match any one template from the bank; such signals incur a loss of SNR. The parameters of the templates in the bank are chosen to minimize this loss of SNR using as few templates as possible [36–38]. Techniques for efficiently covering the binary parameter space with templates have been extensively developed [39–45]. We assume here that such a template bank has already been constructed, and we describe how the bank is decomposed to more efficiently filter the data.

The standard methods for template bank construction naturally lead to banks of highly redundant templates. In the frequency domain, filtering directly with the physical templates has the advantage of admitting computationally efficient searches over the unknown signal coalescence phase and time; this advantage is lost in the time domain. The GstLAL-based inspiral pipeline therefore does not directly filter the data against the physical template waveforms themselves. Rather, it employs the Low Latency Online Inspiral Detection (LLOID) method [24] (see also Sec. III A), which combines singular value decomposition (SVD) [46–48] with near-critical sampling to construct a reduced set of orthonormal filters with far fewer samples.

In order to prepare the templates for the LLOID decomposition, the template bank is first split into partially overlapping "split banks" of templates with similar time-frequency evolution based on the template parameters, as depicted in Fig. 6. Templates corresponding to binary black hole systems with circular orbits and component spins parallel to the orbital angular moment can be characterized by the component masses m_i and the dimensionless spin parameters $\chi_i = \vec{S}_i \cdot \hat{L}/m_i^2$ for $i = 1, 2$, where $\vec{S}_i$ are the spin vectors and $\hat{L}$ is the orbital angular momentum unit vector. Circularized binary neutron star templates with aligned spins can also be characterized by m_i and χ_i, however, not as accurately due to neutron-star specific effects such as tidal disruption. The templates for these systems are binned in a two-dimensional space, first by an effective spin parameter $\chi_{\rm eff}$,

$$\chi_{\rm eff} \equiv \frac{m_1\chi_1 + m_2\chi_2}{m_1 + m_2}, \qquad (16)$$

and then by chirp mass $\mathcal{M}$,

$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}. \qquad (17)$$

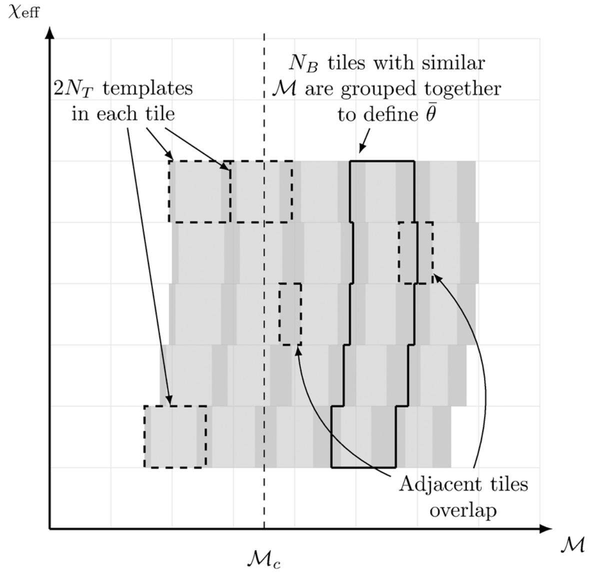
$\chi_{\rm eff}$
$2N_T$ templates in each tile
N_B tiles with similar $\mathcal{M}$ are grouped together to define $\bar{\theta}$
Adjacent tiles overlap
$\mathcal{M}_c$
$\mathcal{M}$

FIG. 6. An illustration of how the physical parameter space is tiled into regions in which the LLOID decomposition is done. The physical parameter space is projected onto the $\mathcal{M}$, χ plane. Tiles of equal template number, $2N_T$, are constructed and overlapped in the $\mathcal{M}$ direction by $\mathcal{O}(10\%)$. Above a specified chirp mass, $\mathcal{M}_c$, waveforms that use the full inspiral-merger-ringdown description are used. Below $\mathcal{M}_c$, waveforms that model only the inspiral phase are used. Tiles of similar chirp mass are then grouped together to define a one-dimensional family of similar parameters, $\bar{\theta}$, used in the evaluation of the likelihood-ratio ranking statistic (Sec. III E).

$2N_T$ real templates are placed in each split bank, where N_T is typically $\mathcal{O}(100)$. The factor of 2 is a result of using two orthogonal real-valued templates in place of one complex-valued template (Sec. II). The input templates in adjacent $\mathcal{M}$ bins are overlapped in order to mitigate boundary effects from the SVD. Overlapping regions are clipped after reconstruction such that the output has no redundant template waveforms.[1] The waveforms are then whitened using reference PSDs, as described in Sec. II B, and each split bank is decomposed via the LLOID method, described below and in Fig. 7.

Each split bank is divided into various time slices after prepending the templates with zeros such that every template has the same number of sample points; this allows us to efficiently sample different regions of our waveforms with the appropriate Nyquist frequency instead of oversampling the low-frequency regions of the waveform with the sampling frequency required for the high-frequency regions. The SVD is then performed on each time slice of each split bank and truncated such that we retain only the

[1] Split banks that contain the lowest and highest $\mathcal{M}$ templates in a given χ bin are padded with duplicate templates from within the split bank in order to keep the clipping uniform between split banks.

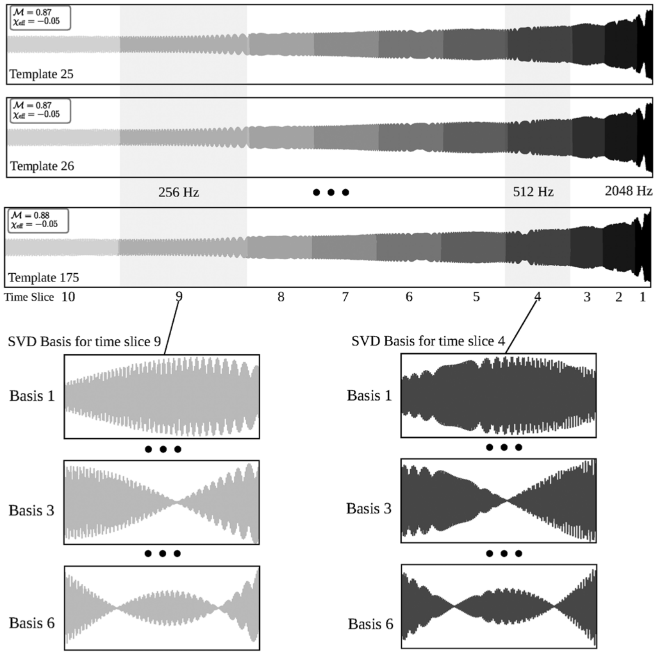

FIG. 7. An example of the LLOID decomposition [24]. In this example, $N_T = 195$ binary inspiral waveforms (390 including the two possible phases) with a chirp mass between 0.87 and 0.88 are first "whitened" by dividing them by a realistic noise amplitude spectral density from advanced LIGO (aLIGO). The line features in the spectrum are responsible for the amplitude modulation of the waveforms. The waveforms, which are prepended with zeros when necessary so that all of the templates in a given decomposition have the same number of sample points, were decomposed into 30 time slices at sample rates ranging between 128 Hz and 2048 Hz (only the last 10 slices are shown). A basis filter set from the waveforms in each time slice was constructed using the SVD [46]. Only 6–10 basis waveforms per slice were needed to reconstruct both phases of the 195 input waveforms to an accuracy of better than 99.9%.

most important basis waveforms returned by the SVD algorithm, as measured by the match between the original templates and the reconstructed waveforms [24].

In addition to being used for the LLOID decomposition, split banks are binned by the lowest chirp mass in each split bank to construct bins of similar templates. These are referred to as $\bar{\theta}$ bins and define a binning of the likelihood-ratio detection statistic defined in Sec. III E.

III. EVENT IDENTIFICATION STAGE

Borrowing the language commonly used in particle experiments, the GstLAL-based inspiral pipeline identifies "triggers" from individual interferometer data streams. Triggers which arrive in coincidence are elevated to the "event" classification and ranked by the likelihood-ratio ranking statistic. In this section, we discuss how a list of triggers is generated by the matched-filtering algorithm and how coincidences are identified and ranked as events.

A. Matched filtering and the LLOID method

As discussed in Sec. II D, groups of templates are partitioned into time slices as part of the LLOID decomposition [24]. Specifically, any split-bank $\mathbf{H}$ can be written as a collection of matrices $\mathbf{H}^s$,

$$\mathbf{H} = \{\mathbf{H}^s\}, \tag{18}$$

where each $\mathbf{H}^s$ contains time slice s of all $2N_T$ templates in the split bank,

$$\mathbf{H}^s = \{\hat{h}_i^s(t) : i \in [0, 2N_T - 1]\}. \tag{19}$$

The index s is chosen to be the largest at the start of the template waveforms, decreasing until $s = 0$ for the last slice (as seen in Fig. 7). Each slice of the split bank, $\mathbf{H}^s$, is decomposed via the SVD to provide basis functions u. These basis functions can be used to reconstruct any $\hat{h}_i$ to a predetermined tolerance, i.e.,

$$\hat{h}_i^s(t) \approx \sum_{\nu=0}^{N-1} v_{i\nu}^s \sigma_\nu^s u_\nu^s(t), \tag{20}$$

where $\mathbf{u}^s = \{u_\nu^s(t)\}$ is a matrix composed of N basis vectors, $\mathbf{v}^s = \{v_{i\nu}^s\}$ is a reconstruction matrix, and $\vec{\sigma} = \{\sigma_\nu^s\}$ is a vector of singular values whose magnitudes are directly proportional to how important a corresponding basis vector is to the reconstruction process [46]. Now instead of evaluating Eq. (3b) $2N_T$ times for each slice of $2N_T$ templates, we can evaluate

$$U_\nu^s(t) = 2\int_{-\infty}^{\infty} \mathrm{d}\tau u_\nu^s(t) \hat{d}^s(t+\tau) \tag{21}$$

$N < 2N_T$ times for each slice, where $\hat{d}^s(t)$ is sampled at the same rate as $u_\nu^s(t)$. The matched filter output time series is calculated for each time slice $\mathbf{H}^s$, then upsampled via sinc interpolation and added to the output of other time slices (in order of decreasing s) to obtain the output of Eq. (3b). The matched filter output accumulated up through slice s is defined recursively for each template in a given split bank as

$$x_i^s(t) = \overbrace{(H^\uparrow x_i^{s+1})(t)}^{\text{Previous} x_i} + \underbrace{\sum_{\nu=0}^{N-1} v_{i\nu}^s \sigma_\nu^s U_\nu^s(t)}_{\text{Current} x_i}, \tag{22}$$

where $H^\uparrow$ acts on a time series sampled at f_{s+1} and upsamples it to f_s. Recall that the GstLAL-based inspiral pipeline uses two real-valued templates in place of one complex-valued template (Sec. II), and thus the computed SNR is the quadrature sum of matched filter outputs from waveforms which differ only in coalescence phase by $\pi/4$,

$$\rho_j(t) = \sqrt{x_{2j}(t)^2 + x_{2j+1}(t)^2}, \qquad j \in [0, N_T - 1]. \tag{23}$$

Note that there are half as many SNRs as there are templates, which is a result of using real-valued templates in place of complex-valued templates (Sec. II). Evaluating Eq. (22) saves a factor of 10^4 in computational cost over a direct time domain convolution of the template waveforms for typical aLIGO search parameters [24].

B. Triggers

The raw SNR time series, typically sampled at 2 KHz, is discretized into triggers before being stored. The discretization is done by maximizing the SNR over time in one-second windows and recording the peak if it crosses a predetermined threshold. With the typical threshold of $\mathrm{SNR} = 4$, it is probable to have at least one trigger in every one-second interval for every template for every detector data set analyzed. Although the number of triggers can easily be hundreds of thousands per second (due to modern template banks containing hundreds of thousands of templates [17]), storing them is a marked improvement over storing the raw SNR time series, which is over 3 orders of magnitude more voluminous. However, we do not discard the raw SNR time series information immediately, because it is needed for the next stage of the pipeline (Sec. III C). For each trigger, we record the parameters of the template, the trigger time, the SNR, and the coalescence phase. The trigger time is computed via subsample interpolation to nanosecond precision; while low SNR triggers suffer from poor timing resolution, high SNR triggers can be resolved to better resolution than that of the sample rate [9,49,50]. Triggers are identified in parallel across each template in a given $\bar{\theta}$ bin.

C. Signal-based vetoes

Detector data often contain glitches that are not removed during the data conditioning stage (Sec. II C). Therefore, ranking triggers solely by SNR is not sufficient to separate noise from transient signals. Fortunately, we can exploit consistency checks to improve our ability to discriminate spurious glitches from true gravitational-wave events. Requiring multiple-detector coincidence (Sec. III D) is one powerful check, but here we discuss a separate check on waveform consistency for a single detector's matched-filter output. This waveform consistency check determines how similar the SNR time series of the data is to the SNR time series expected from a real signal.

Under the assumption that the signal in the data exactly matches the matched-filter template up to a constant, it is possible to predict the local matched-filter SNR by computing the template autocorrelation function and scaling it to the known SNR. However, the known SNR is a result of the matched-filter response from two identical but out of phase templates; thus instead of scaling the autocorrelation function to the SNR, a complex SNR series is constructed from the two matched-filter outputs,

$$z_j(t) = x_{2j}(t) + i x_{2j+1}(t), \qquad j \in [0, N_T - 1]. \tag{24}$$

These are compared to the complex autocorrelation function,

$$R_j(t) = \int_{-\infty}^{\infty} \mathrm{d}f \, \frac{|\tilde{h}_{2j}(f)|^2 + |\tilde{h}_{2j+1}(f)|^2}{S_n(|f|)} e^{2\pi i f t}, \tag{25}$$

where $t = 0$ is chosen to be the peak time, t_p. By convention, each *real* template is normalized such that its autocorrelation is $\frac{1}{2}$ at the peak time, and thus $R_j(0) = 1$. We compute a signal consistency test value, ξ^2, as a function of time given the complex SNR time series $z_j(t)$, a trigger's peak complex SNR $z_j(0)$, and the autocorrelation function time series $R_j(t)$ as

$$\xi_j^2(t) = |z_j(t) - z_j(0) R_j(t)|^2. \tag{26}$$

If the gravitational wave strain data contain only noise [i.e., $\tilde{d}(f) = \tilde{n}(f)$], then (see Appendix A for derivation)

$$\langle \xi_j^2(t) \rangle = 2 - 2|R_j(t)|^2. \tag{27}$$

In practice, a value of ξ^2 is computed for each trigger by integrating $\xi^2(t)$ in a window of time around the trigger and normalizing it using Eq. (27). The integral takes the form

$$\xi_j^2 = \frac{\int_{-\delta t}^{\delta t} \mathrm{d}t |z_j(t) - z_j(0) R_j(t)|^2}{\int_{-\delta t}^{\delta t} \mathrm{d}t (2 - 2|R_j(t)|^2)}, \tag{28}$$

where δt is a tunable parameter that defines the size of the window around the peak time over which to perform the integration. Typically, δt is calculated in terms of an odd-valued autocorrelation length (ACL), specified as a number of samples such that $\delta t = (\mathrm{ACL} - 1)\Delta t_s/2$, where $\Delta t_s = f_s^{-1}$ is the sampling time step. A suitable value for ACL was found to be 351 samples when filtering is conducted at a 2048 Hz sample rate, resulting in $\delta t \sim 85.4$ ms; this value was found by using Monte Carlo simulations in real data.

Figure 8 plots the SNR and scaled autocorrelation for a template that recovered a simulated signal in initial LIGO data. Subtracting the measured SNR time series from the predicted series shown in this figure is what is done in (28) on a trigger-by-trigger basis.

We note that ξ^2 differs from the traditional time-frequency χ^2 test in [26], and it is not, in fact, a χ^2-distributed number in Gaussian noise. However, the statistics of the ξ^2 test are recorded for both noise and simulated signals and can therefore be used in the likelihood-ratio test described in Sec. III E.

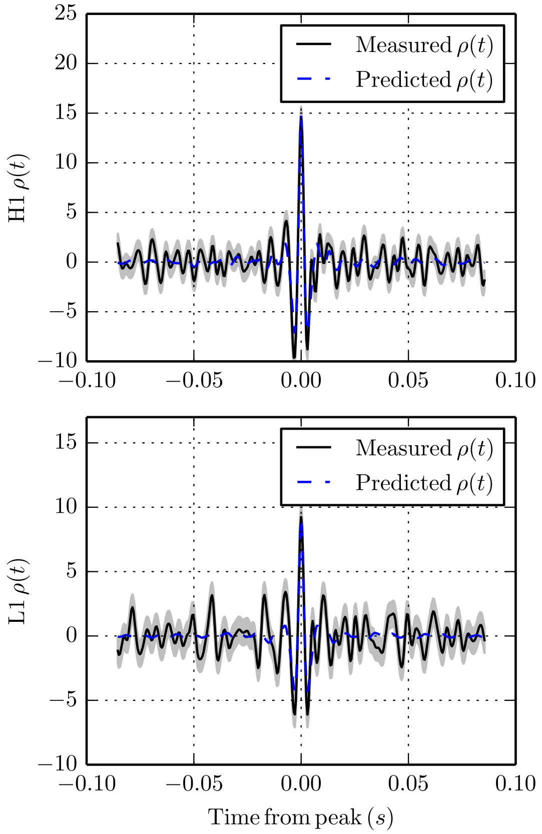

FIG. 8. Ingredients in the autocorrelation-based least-squares test as described in (26). The two panels show the SNR time series near a simulated signal in initial LIGO data (black lines) along with the predicted SNR computed from the template autocorrelation. Subtracting these two time series and integrating their squared magnitude provides a signal consistency test, ξ^2, at the time of a given trigger that can be used to reject nonstationary noise transients.

D. Coincidence

Demanding that two triggers are found in temporal coincidence between the LIGO sites is a powerful technique to suppress the background of the search. For a single detector trigger, we define the time of an event to coincide with the peak of its SNR time series. Given a trigger in one detector, we check for corresponding triggers in the other detector within an appropriate time window, which takes into account the maximum gravitational-wave travel time between detectors and statistical fluctuations in the measured event time due to detector noise. For the two LIGO detectors, the time window is typically ±15 ms. We further require that the mass and spin template parameters are the same for the two triggers. This exact match requirement potentially results in a small loss of SNR for real signals, since the loudest trigger in each detector will in general not have the exact same template parameters due to independent noise in the detectors. However, taking into account such fluctuations requires detailed knowledge of the metric on the signal manifold [51], which may not be easily available. Furthermore, the exact match restriction suppresses the noise and drastically simplifies the pipeline.

E. Event ranking

Each trigger from each detector has independently computed ρ, ξ^2, and t_p values. After coincidences are formed, it is necessary to rank the coincident events from least likely to be a signal to most likely to be a signal and to assign a significance to each. The GstLAL-based inspiral pipeline uses the likelihood-ratio statistic described in [27] to rank coincident events by their SNR, ξ^2, the instantaneous sensitivity of each detector (expressed as the horizon distance, $\{D_{\mathrm{H1}}, D_{\mathrm{L1}}\}$), and the detectors involved in the coincidence (expressed as the set $\{\mathrm{H1}, \mathrm{L1}\}$). For the case where only the aLIGO observatories H1 and L1 are participating, the likelihood ratio of an event found in coincidence is defined as

$$\begin{aligned}&\mathcal{L}(\{D_{\mathrm{H1}},D_{\mathrm{L1}}\},\{\mathrm{H1},\mathrm{L1}\},\rho_{\mathrm{H1}},\xi^2_{\mathrm{H1}},\rho_{\mathrm{L1}},\xi^2_{\mathrm{L1}},\bar{\theta})\\&=\mathcal{L}(\{D_{\mathrm{H1}},D_{\mathrm{L1}}\},\{\mathrm{H1},\mathrm{L1}\},\rho_{\mathrm{H1}},\xi^2_{\mathrm{H1}},\rho_{\mathrm{L1}},\xi^2_{\mathrm{L1}}|\bar{\theta})\mathcal{L}(\bar{\theta})\\&=\frac{P(\{D_{\mathrm{H1}},D_{\mathrm{L1}}\},\{\mathrm{H1},\mathrm{L1}\},\rho_{\mathrm{H1}},\xi^2_{\mathrm{H1}},\rho_{\mathrm{L1}},\xi^2_{\mathrm{L1}}|\bar{\theta},\mathrm{signal})}{P(\{D_{\mathrm{H1}},D_{\mathrm{L1}}\},\{\mathrm{H1},\mathrm{L1}\},\rho_{\mathrm{H1}},\xi^2_{\mathrm{H1}},\rho_{\mathrm{L1}},\xi^2_{\mathrm{L1}}|\bar{\theta},\mathrm{noise})}\mathcal{L}(\bar{\theta}),\end{aligned} \tag{29}$$

where $\bar{\theta}$ is a label corresponding to the template bank bin being matched-filtered (Sec. II D). The numerator and denominator are factored into products of several terms in [27], assuming that the noise distributions for each interferometer are independent of each other. The computation of each term in the factored numerator and denominator is discussed in detail in [27]; in this paper, we will give only a short summary of the denominator. The denominator is factored such that

$$\begin{aligned}&P(\{D_{\mathrm{H1}},D_{\mathrm{L1}}\},\{\mathrm{H1},\mathrm{L1}\},\rho_{\mathrm{H1}},\xi^2_{\mathrm{H1}},\rho_{\mathrm{L1}},\xi^2_{\mathrm{L1}}|\bar{\theta},\mathrm{noise})\\&\quad\propto\prod_{\mathrm{inst}\in\{\mathrm{H1},\mathrm{L1}\}}P(\rho_{\mathrm{inst}},\xi^2_{\mathrm{inst}}|\bar{\theta},\mathrm{noise}).\end{aligned} \tag{30}$$

The detection statistics ρ and ξ^2 from noncoincident triggers are used to populate histograms for each detector, which are then normalized and smoothed by a Gaussian smoothing kernel to approximate $P(\rho_{\mathrm{inst}},\xi^2_{\mathrm{inst}}|\bar{\theta},\mathrm{noise})$. Running in the low-latency operation mode requires a burn-in period until the analysis collects enough noncoincident triggers to construct an accurate estimate of the (ρ,ξ^2) PDFs. Neither operation mode tracks time dependence of these PDFs; instead the PDFs are constructed from cumulative histograms. Future work may add time dependence.

Rather than collecting noncoincident (ρ,ξ^2) statistics from individual templates, we group linearly dependent templates together to avoid the computational cost and complexity of tracking each template separately. Furthermore, it has been observed that groups of linearly dependent templates produce similar PDFs, and thus coarse graining the parameter space allows one to approximate these PDFs for collections of templates. Therefore, the $\bar{\theta}$ in the likelihood ratio is a label that identifies a specific template bank bin. Exactly how templates are grouped together into background bins is left as a tuning decision for the user, but typically $\mathcal{O}(1000)$ templates from each detector are grouped together.

Examples of the ρ and ξ^2 distributions, estimated from an analysis of one week of S6 data, are shown in Fig. 9. The analysis considered data recorded between September 14, 2010, 23:58:48 UTC and September 21, 2010, 23:58:48 UTC. These boundaries were chosen to include the blind injection performed on September 16, 2010, at 06:42:23 UTC, often referred to as the "Big Dog." The warm colormap corresponds to the natural logarithm of the estimated noise probability density function. The cool colormap corresponds to a PDF generated by adding the coincident triggers to the single detector triggers before smoothing and normalizing. The cool-colormap distribution was then masked to show only regions which deviate from the background estimate. The location of the Big Dog parameters is marked with a black X.

Examples of two other likelihood-ratio components from the Big Dog analysis are shown in Fig. 10. The top plot shows the joint SNR PDF, which is used in the numerator

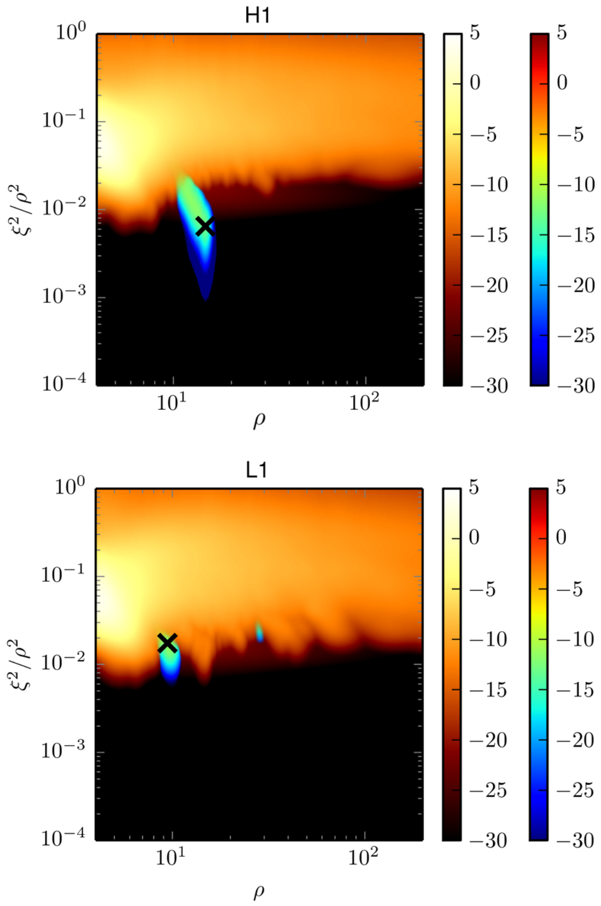

FIG. 9. PDFs used in the likelihood ratio calculation, generated by histogramming, then smoothing, and normalizing the triggers. The plots shown are from an analysis of S6 data beginning at September 14, 2010, at 23:58:48 UTC and ending at September 21, 2010, at 23:58:48 UTC, which includes the blind injection known as the Big Dog. The warm colormap corresponds to the natural logarithm of marginalized probability density function estimated from noncoincident triggers only; the cool-colormap region was computed by adding coincident triggers to the histograms before smoothing and normalizing. Regions of the cool-colormap model consistent with the warm-colormap model were then masked. The location of the Big Dog in the (ρ, ξ^2) plane is marked with a black X.

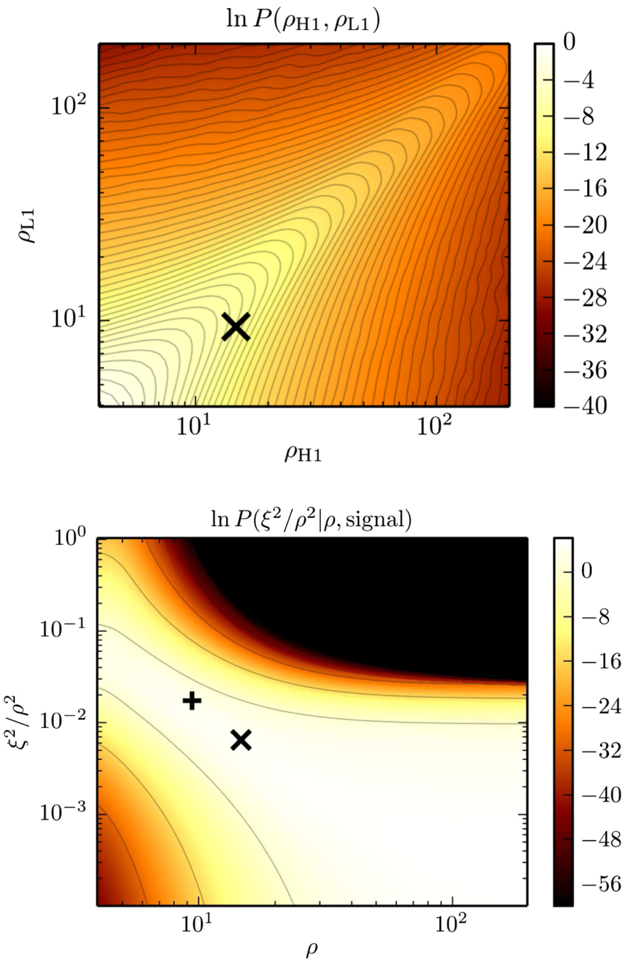

FIG. 10. Instances of two of the distributions included in the calculation of the likelihood-ratio numerator, generated from an analysis of S6 data beginning at September 14, 2010, at 23:58:48 UTC and ending at September 21, 2010, at 23:58:48 UTC. Top: The joint SNR PDF used to enforce amplitude consistency across observatories. The location of the measured Big Dog parameters is marked with a black X. Bottom: The (ρ, ξ^2) signal distribution used in the numerator of the likelihood ratio. The locations of the measured Big Dog parameters are marked with a black X for Hanford and a black + for Livingston.

of the likelihood ratio to enforce amplitude consistency [27]; the bottom plot shows the signal hypothesis model of the (ρ, ξ^2) plane. The semianalytic models used to generate these plots are described in [27].

F. Event clustering

Signals can produce several high-likelihood events at the same time in different templates; we wish to ensure that we consider only the most likely event associated with a signal. In the offline analysis, we use a clustering algorithm that picks out the maximum likelihood-ratio event globally across the input template bank within a ±4 s window. The online analysis does not cluster events globally to reduce latency. Instead, the online analysis keeps the maximum likelihood-ratio event in each $\bar{\theta}$ bin within a ±1 s window.

IV. EVENT PROCESSING AND SENSITIVITY ESTIMATION

The result of the pipeline components described in Sec. III is a list of events ranked from most to least likely to be a gravitational-wave signal. In this section, we discuss how the significance is estimated, what the procedure is in the case that a sufficiently significant candidate is identified, and how simulated waveforms are used to characterize the sensitivity of the analysis to gravitational waves.

A. Event significance estimation

Most coincident events are noise, and thus the p-value, the probability that noise would produce an event with a ranking statistic at least as large as the one under consideration, is the standard tool used to identify candidate gravitational-wave events. The p-value has conventionally been evaluated by performing *time slides*, where a set of time shifts that are much larger than the gravitational-wave travel time (tens of milliseconds) between gravitational-wave detectors is introduced into one or more data sets and the coincidence and event-ranking procedure is repeated in the same way as it is done without the time shifts [52]. Instead of performing time slides, the GstLAL-based inspiral pipeline uses triggers not found in coincidence to compute a kernel density estimate of the probability density of noiselike events in each background bin, $P(\ln\mathcal{L}|\bar{\theta}, \text{noise})$ [27,28]. The background bins are then marginalized over to obtain $P(\ln\mathcal{L}|\text{noise})$ and the complementary cumulative distribution,

$$C(\ln\mathcal{L}^*|\text{noise}) = \int_{\ln\mathcal{L}^*}^{\infty} \mathrm{d}\ln\mathcal{L} P(\ln\mathcal{L}|\text{noise}). \quad (31)$$

The p-value we seek describes the probability that a population of M independent coincident noiselike events contains at least one event with a log likelihood ratio greater than or equal to some threshold $\ln\mathcal{L}^*$. This can be written as the complement of the binomial distribution [28],

$$\begin{aligned} P(\ln\mathcal{L} \geq \ln\mathcal{L}^*|\text{noise}_1, \ldots, \text{noise}_M) \\ &= 1 - \binom{M}{0}(1 - e^{-C(\ln\mathcal{L}^*|\text{noise})})^0 (e^{-C(\ln\mathcal{L}^*|\text{noise})})^M \\ &= 1 - e^{-MC(\ln\mathcal{L}^*|\text{noise})}, \end{aligned} \quad (32)$$

where $e^{-C(\ln\mathcal{L}^*|\text{noise})}$ is the probability that a Poisson process with mean rate $C(\ln\mathcal{L}^*|\text{noise})$ will yield an event with the log likelihood ratio less than $\ln\mathcal{L}^*$. The binomial coefficient and the term that follows are both clearly unity and were only explicitly written for pedagogical reasons.

When calculating an event's significance during an experiment of undetermined length, such as the low-latency processing of data during a science run, it is convenient to express the significance in terms of how often the noise is expected to yield an event with a log likelihood ratio $\geq \ln\mathcal{L}^*$. This is referred to as the false-alarm rate (FAR) [28]; for an experiment of length T, we define this as

$$\text{FAR} = \frac{C(\ln\mathcal{L}^*|\text{noise})}{T}. \quad (33)$$

The time used in the calculation of the FAR is the total elapsed observing time regardless of instrument state for the low-latency configuration and the total time where at least two detectors are operating for the offline analysis operation. The offline configuration definition is historically what has been used; however, the online definition leads to intuitive false alarm rates for sharing low-latency events with external observing partners.

The procedure to estimate the background distribution described thus far does not account for the clustering described in Sec. III F. Events with low $\ln\mathcal{L}$ are more common than those with high $\ln\mathcal{L}$, and thus the clustering process removes low $\ln\mathcal{L}$ events preferentially. The normalization of the background model is determined by the observed events above a log-likelihood ratio threshold chosen to be safely out of the region affected by clustering. This is acceptable because low $\ln\mathcal{L}$ events are, by construction, the least likely to contain a signal. Consequently, we only consider events well above this threshold as viable candidates. Work is currently underway to create a background model that accounts for clustering.

A plot of the significance results from the Big Dog run discussed in Sec. III E is shown in Fig. 11. The Big Dog was found with a p-value of 5.4×10^{-9} (5.7σ), which corresponded to a FAR of 1.1×10^{-14} Hz (1 per $\sim 2.7 \times 10^6$ yr); Table I lists the recovered parameters of the Big Dog.

B. Generating alerts

When operating in a low-latency analysis configuration, one of the primary goals of the GstLAL-based inspiral pipeline is to identify candidate events and upload them to

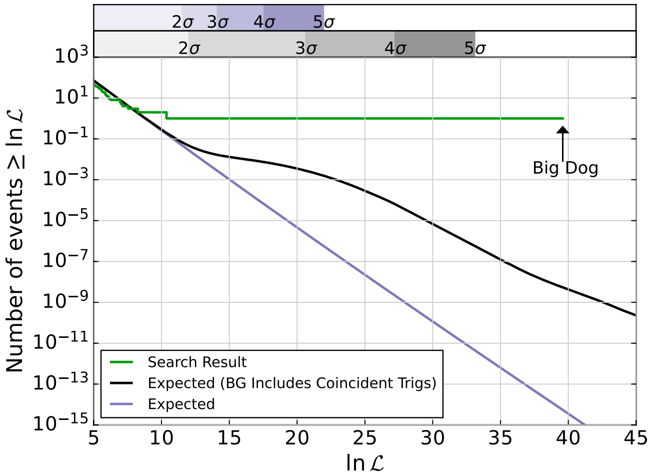

FIG. 11. Number of observed events as a function of the log likelihood ratio in an analysis of S6 data beginning on September 14, 2010, at 23:58:48 UTC and ending at September 21, 2010, at 23:58:48 UTC. The Big Dog injection, found with a false alarm probability of 5.4×10^{-9} (5.7σ), is marked on the observed distribution (green). The black line represents the predicted number of events when observed events are included in the background model, while the blue line is the predicted number when the observed events are not included in the background model.

the Gravitational-wave Candidate Event Database (GraceDB) [29] as quickly as possible in order to issue alerts to observing partners [12].

Events that pass a given FAR threshold are identified within ∼1 min of the gravitational-wave signals arriving at Earth. The basic parameters of the event are transmitted to GraceDB, including the time of the event, the SNR and ξ^2 values for the triggers in each detector, and the parameters of the best-fit template (for example, mass and spin values, the significance estimate, etc.). Furthermore, the instantaneous estimate of the PSD is uploaded to GraceDB, as well as the histogram data used in computing the p-value.

TABLE I. The result of the analysis of S6 data beginning on September 14, 2010, at 23:58:48 UTC and ending on September 21, 2010, at 23:58:48 UTC. Only the parameters found for the recovered Big Dog injection are shown. The Big Dog was the most significant event found in this analysis period.

p-value	5.4×10^{-9} (5.7σ)
FAR (Hz)	1.1×10^{-14}
$\log \mathcal{L}$	39.6
ρ_{H}	14.7
ξ^2_{H}	1.4
ρ_{L}	9.4
ξ^2_{L}	1.5
$\mathcal{M}$ ($\mathrm{M}_\odot$)	4.7

An event upload automatically initiates several automated and human follow-up activities to aid rapid communication with observing partners [53]. First, a rapid sky localization routine known as BAYESTAR [9,50,54] uses the event information and the PSD to estimate the event's sky position within minutes. At the same time, deeper parameter estimation analysis begins in order to provide updated position reconstruction, as well as the full posterior probability distributions of the binary parameters [55], on a time scale that ranges from hours to days.

In addition to parameter estimation, data-quality information is also mined to provide rapid feedback to analysts. Time-frequency spectrograms are automatically generated to indicate the stationarity of noise near an event [56]. Furthermore, low-latency mining of LIGO's auxiliary channels provide additional information about the state of the detector and environment when an alert is first generated [16,57,58].

The suitability of the low-latency pipeline for generating data for external alerts has been studied extensively in [9,54].

C. Software injections

Simulated gravitational waveforms known as "software injections" are used to assess the pipeline response to real gravitational-wave signals. The LIGO strain data are duplicated and simulated compact binary waveforms are digitally added to the duplicated data streams. In low latency, the new data with software injections added are broadcast to the LIGO Data Grid in parallel to the normal data set so that a simultaneous run can measure the instantaneous sensitivity of the low-latency analysis to compact binary sources. In the offline mode, strain data are read from the disk, software injections are added, and the new data are written back to disk before the offline inspiral pipeline processes the data.

Injections are considered "found" if a coincident event with the correct template parameters is found with a FAR ≤ 30 d at the time of the injection, and "missed" otherwise. The volume of space the pipeline is sensitive to, V, is approximated as a sphere and computed via

$$V = 4\pi \int_0^\infty \mathrm{d}r \epsilon(r) r^2, \tag{34}$$

where $\epsilon(r)$ is an efficiency parameter given by the ratio of found to total injections modeled to be a distance r away. We define our estimated range, the average furthest distance a signal can originate from and still be detected, as

$$R = \left(\frac{3V}{4\pi}\right)^{1/3}. \tag{35}$$

It is important to note the range depends on parameters of the compact binary system. For example, the range for a

1.4–1.4M$_\odot$ binary neutron star system will be different from that of a 10–10M$_\odot$ binary black hole system; thus different injection sets must be used to determine the pipeline's sensitivity to different regions of the compact binary parameter space. Equation (35) is compared to the analytically computed SenseMon range, an estimate of pipeline sensitivity calculated from the PSD [59]. Comparing the sensitivity estimated from the PSD to the sensitivity estimated from injections provides additional confidence in the sensitivity estimates.

Typically, injections are added at a much higher rate than the expected gravitational wave signal rate. However, their cadence is chosen such that they do not bias the PSD estimate described in Sec. II B. In practice, injections are typically added about once per minute so that it is possible to evaluate the average response to certain signal types over the entire experiment duration.

V. CONCLUSION

The GstLAL-based inspiral pipeline is a stream-based pipeline that allows for time-domain compact binary searches capable of identifying and uploading candidate gravitational-wave signals within seconds. This provides rapid feedback to the gravitational-wave detector control rooms and enables prompt event alerts for electromagnetic follow-up by observing partners. The analysis techniques were designed for second- and third-generation gravitational-wave detectors and have been demonstrated to be applicable even to the computationally challenging case of the future Einstein Telescope [60].

GstLAL and all related software is available for public use and licensed under the GPL [14].

ACKNOWLEDGMENTS

The authors wish to thank B. Sathyaprakash, the LIGO Scientific Collaboration, and the Compact Binary Coalescence working group for many useful discussions. We gratefully acknowledge the support of the Eberly Research Funds of Penn State and the National Science Foundation through Grants No. PHY-0757058, No. PHY-0923409, No. PHY-1104371, No. PHY-1454389, No. PHY-1607585, and No. PHY-1307429. K. C. was supported in part by JSPS KAKENHI Grant No. 16H06714. This document has LIGO Document No. P1600009.

APPENDIX: EXPECTATION VALUE OF SIGNAL CONSISTENCY TEST VALUE IN NOISE

Expanding Eq. (26) and taking the ensemble average, we find

$$\begin{aligned}\langle \xi_j^2(t)\rangle &= \langle |z_j(t) - z_j(0)R_j(t)|^2\rangle,\\ &= \langle |z_j(t)|^2\rangle - 2\mathrm{Re}[\langle z_j^*(t)z_j(0)\rangle R_j(t)]\\ &\quad + \langle |z_j(0)|^2\rangle |R_j(t)|^2. \qquad \text{(A1)}\end{aligned}$$

Starting with Eq. (3a),

$$\begin{aligned}\langle |z_j(t)|^2\rangle &= \left\langle \left| 2\int_{-\infty}^{\infty} \mathrm{d}f \frac{\tilde{n}(f)(\tilde{h}_{2j}^*(f) + i\tilde{h}_{2j+1}^*(f))}{S_n(f)} e^{2\pi i t f}\right|^2\right\rangle,\\ &= 4\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{d}f_1 \mathrm{d}f_2 \Bigg(\langle \tilde{n}(f_1)\tilde{n}(f_2)\rangle e^{2\pi i t(f_1 - f_2)}\\ &\quad \times \frac{(\tilde{h}_{2j}^*(f_1) + i\tilde{h}_{2j+1}^*(f))(\tilde{h}_{2j}(f_2) - i\tilde{h}_{2j+1}(f_2))}{S_n(|f_1|)S_n(|f_2|)}\Bigg),\\ &= 2\int_{-\infty}^{\infty} \mathrm{d}f \frac{|\tilde{h}_{2j}(f)|^2 + |\tilde{h}_{2j+1}(f)|^2}{S_n(|f|)},\\ \langle |z_j(t)|^2\rangle &= \langle |z_j(0)|^2\rangle = 2, \qquad \text{(A2)}\end{aligned}$$

where Eq. (7) was used in the last step. Computing $\langle z_j^*(t)z_j(0)\rangle$ follows the same steps, except the $z_j(0)$ term does have a complex exponential to cancel the complex exponential accompanying $z_j^*(t)$; thus

$$\langle z_j^*(t)z_j(0)\rangle = 2R_j^*(t), \qquad \text{(A3)}$$

$$\langle \xi_j^2(t)\rangle = 2 - 2|R_j(t)|^2. \qquad \text{(A4)}$$

[1] B. Abbott *et al.*, Phys. Rev. Lett. **116**, 061102 (2016).
[2] B. Abbott *et al.*, Phys. Rev. Lett. **116**, 241103 (2016).
[3] J. Aasi *et al.*, Classical Quantum Gravity **32**, 074001 (2015).
[4] F. Acernese *et al.*, Classical Quantum Gravity **32**, 024001 (2015).
[5] Y. Aso, Y. Michimura, K. Somiya, M. Ando, O. Miyakawa, T. Sekiguchi, D. Tatsumi, and H. Yamamoto, Phys. Rev. D **88**, 043007 (2013).
[6] B. Iyer *et al.*, LIGO-India, Proposal of the Consortium for Indian Initiative in Gravita tional-wave Observations (IndIGO), Report No. LIGO-DCC-M1100296, 2011.
[7] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Astrophys. J. Lett. **833**, L1 (2016).
[8] J. Abadie *et al.*, Classical Quantum Gravity **27**, 173001 (2010).
[9] L. P. Singer *et al.*, Astrophys. J. **795**, 105 (2014).
[10] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), Astrophys. J. **826**, L13 (2016).

[11] S. Adrián-Martínez *et al.* (ANTARES, IceCube, LIGO Scientific, Virgo Collaborations), Phys. Rev. D **93**, 122010 (2016).

[12] Identification and follow up of electromagnetic counterparts of gravitational wave candidate events, http://www.ligo.org/scientists/GWEMalerts.php, accessed 01-15-2016.

[13] B. P. Abbott *et al.* (LIGO Scientific, Virgo Collaborations), Classical Quantum Gravity **33**, 134001 (2016).

[14] Gstlal, https://www.lsc-group.phys.uwm.edu/daswg/projects/gstlal.html, accessed 07-01-2015.

[15] Gstreamer, https://gstreamer.freedesktop.org.

[16] Lalsuite, https://www.lsc-group.phys.uwm.edu/daswg/projects/lalsuite.html, accessed 07-01-2015.

[17] B. Abbott *et al.*, Phys. Rev. D **93**, 122003 (2016).

[18] J. Abadie *et al.*, Astron. Astrophys. **541**, A155 (2012).

[19] S. Klimenko *et al.*, Classical Quantum Gravity **25**, 114029 (2008).

[20] R. Lynch, S. Vitale, R. Essick, E. Katsavounidis, and F. Robinet, arXiv:1511.05955.

[21] B. P. Abbott *et al.* (Virgo, LIGO Scientific Collaborations), Phys. Rev. D **93**, 122004 (2016).

[22] D. Buskulic *et al.* (LIGO Scientific, Virgo Collaborations), Classical Quantum Gravity **27**, 194013 (2010).

[23] S. Babak *et al.*, Phys. Rev. D **87**, 024033 (2013).

[24] K. Cannon *et al.*, Astrophys. J. **748**, 136 (2012).

[25] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. Creighton, Phys. Rev. D **85**, 122006 (2012).

[26] B. Allen, Phys. Rev. D **71**, 062001 (2005).

[27] K. Cannon, C. Hanna, and J. Peoples, arXiv:1504.04632.

[28] K. Cannon, C. Hanna, and D. Keppel, Phys. Rev. D **88**, 024025 (2013).

[29] Gravitational wave candidate event database, https://www.lsc-group.phys.uwm.edu/daswg/projects/gracedb.html, accessed 01-15-2016.

[30] S. Anderson *et al.*, Report No. LIGO-DCC-T970130, 2009.

[31] S. Droz, D. J. Knapp, E. Poisson, and B. J. Owen, Phys. Rev. D **59**, 124016 (1999).

[32] J. Abadie *et al.*, Phys. Rev. D **85**, 082002 (2012).

[33] J. Abadie *et al.*, Astrophys. J. **760**, 12 (2012).

[34] J. Slutsky *et al.*, Classical Quantum Gravity **27**, 165023 (2010).

[35] N. Christensen *et al.* (LIGO Scientific, Virgo Collaborations), Classical Quantum Gravity **27**, 194010 (2010).

[36] B. J. Owen, Phys. Rev. D **53**, 6749 (1996).

[37] B. J. Owen and B. Sathyaprakash, Phys. Rev. D **60**, 022002 (1999).

[38] T. A. Apostolatos, Phys. Rev. D **52**, 605 (1995).

[39] T. Cokelaer, Phys. Rev. D **76**, 102004 (2007).

[40] B. Abbott *et al.*, Phys. Rev. D **78**, 042002 (2008).

[41] I. W. Harry, A. H. Nitz, D. A. Brown, A. P. Lundgren, E. Ochsner, and D. Keppel, Phys. Rev. D **89**, 024010 (2014).

[42] S. Babak, Classical Quantum Gravity **25**, 195011 (2008).

[43] I. W. Harry, B. Allen, and B. Sathyaprakash, Phys. Rev. D **80**, 104014 (2009).

[44] G. M. Manca and M. Vallisneri, Phys. Rev. D **81**, 024004 (2010).

[45] S. Privitera, S. R. Mohapatra, P. Ajith, K. Cannon, N. Fotopoulos, M. A. Frei, C. Hanna, A. J. Weinstein, and J. T. Whelan, Phys. Rev. D **89**, 024003 (2014).

[46] K. Cannon, A. Chapman, C. Hanna, D. Keppel, A. C. Searle, and A. J. Weinstein, Phys. Rev. D **82**, 044025 (2010).

[47] K. Cannon, C. Hanna, and D. Keppel, Phys. Rev. D **84**, 084003 (2011).

[48] K. Cannon, C. Hanna, and D. Keppel, Phys. Rev. D **85**, 081504 (2012).

[49] S. Fairhurst, New J. Phys. **11**, 123006 (2009).

[50] L. P. Singer and L. R. Price, Phys. Rev. D **93**, 024013 (2016).

[51] C. Robinson, B. Sathyaprakash, and A. S. Sengupta, Phys. Rev. D **78**, 062002 (2008).

[52] C. D. Capano, Ph.D. thesis, Syracuse University, 2011.

[53] https://www.lsc-group.phys.uwm.edu/daswg/projects/lvalert.html, accessed 01-15-2016.

[54] C. P. Berry *et al.*, Astrophys. J. **804**, 114 (2015).

[55] J. Veitch *et al.*, Phys. Rev. D **91**, 042003 (2015).

[56] Ligo data viewer, https://www.lsc-group.phys.uwm.edu/daswg/projects/ligodv.html, accessed 01-15-2016.

[57] R. Biswas *et al.*, Phys. Rev. D **88**, 062003 (2013).

[58] R. Essick, L. Blackburn, and E. Katsavounidis, Classical Quantum Gravity **30**, 155010 (2013).

[59] J. Abadie *et al.*, arXiv:1003.2481.

[60] D. Meacher, K. Cannon, C. Hanna, T. Regimbau, and B. Sathyaprakash, Phys. Rev. D **93**, 024018 (2016).