Performance of the low-latency GstLAL inspiral search towards LIGO, Virgo, and KAGRA's fourth observing run

Becca Ewing,[1,2,*] Rachael Huxford,[1,2] Divya Singh,[1,2] Leo Tsukada,[1,2] Chad Hanna,[1,2,3,4] Yun-Jing Huang,[1,2] Prathamesh Joshi,[1,2] Alvin K. Y. Li,[5] Ryan Magee,[5] Cody Messick,[6] Alex Pace,[1,2] Anarya Ray,[7] Surabhi Sachdev,[8,7] Shio Sakon,[1,2] Ron Tapia,[1,4] Shomik Adhicary,[1,2] Pratyusava Baral,[7] Amanda Baylor,[7] Kipp Cannon,[9] Sarah Caudill,[10,11] Sushant Sharma Chaudhary,[12] Michael W. Coughlin,[13] Bryce Cousins,[14,1,2] Jolien D. E. Creighton,[7] Reed Essick,[15] Heather Fong,[9,16] Richard N. George,[17] Patrick Godwin,[18,1,2] Reiko Harada,[9,16] James Kennington,[1,2] Soichiro Kuwahara,[9,16] Duncan Meacher,[7] Soichiro Morisaki,[19,7] Debnandini Mukherjee,[20,21] Wanting Niu,[1,2] Cort Posnansky,[1,2] Andrew Toivonen,[13] Takuya Tsutsui,[9] Koh Ueno,[9] Aaron Viets,[22] Leslie Wade,[23] Madeline Wade,[23] and Gaurav Waratkar[24]

[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]*Department of Astronomy and Astrophysics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*
[4]*Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA*
[5]*LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*
[6]*MIT Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[7]*Leonard E. Parker Center for Gravitation, Cosmology, and Astrophysics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201, USA*
[8]*School of Physics, Georgia Institute of Technology, Atlanta 30332, USA*
[9]*RESCEU, The University of Tokyo, Tokyo 113-0033, Japan*
[10]*Department of Physics, University of Massachusetts, Dartmouth, Massachusetts 02747, USA*
[11]*Center for Scientific Computing and Data Science Research, University of Massachusetts, Dartmouth, Massachusetts 02747, USA*
[12]*Institute of Multimessenger Astrophysics and Cosmology, Missouri University of Science and Technology, Physics Building, 1315 N. Pine St., Rolla, Missouri 65409, USA*
[13]*School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA*
[14]*Department of Physics, University of Illinois, Urbana, Illinois 61801, USA*
[15]*Canadian Institute for Theoretical Astrophysics, 60 St George St, Toronto, Ontario M5S 3H8, Canada*
[16]*Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan*
[17]*Center for Gravitational Physics, University of Texas at Austin, Austin, Texas 78712, USA*
[18]*LIGO Laboratory, California Institute of Technology, MS 100-36, Pasadena, California 91125, USA*
[19]*Institute for Cosmic Ray Research, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8582, Japan*
[20]*NASA Marshall Space Flight Center, Huntsville, Alabama 35811, USA*
[21]*Center for Space Plasma and Aeronomic Research, University of Alabama in Huntsville, Huntsville, Alabama 35899, USA*
[22]*Concordia University Wisconsin, Mequon, Wisconsin 53097, USA*
[23]*Department of Physics, Hayes Hall, Kenyon College, Gambier, Ohio 43022, USA*
[24]*Department of Physics, IIT Bombay, Powai, Mumbai 400076, India*

(Received 17 July 2023; accepted 17 January 2024; published 23 February 2024)

GstLAL is a stream-based matched-filtering search pipeline aiming at the prompt discovery of gravitational waves from compact binary coalescences such as the mergers of black holes and neutron stars. Over the past three observation runs by the LIGO, Virgo, and KAGRA Collaboration, the GstLAL search pipeline has participated in several tens of gravitational wave discoveries. The fourth observing run (O4) is set to begin in May 2023 and is expected to see the discovery of many new and interesting gravitational wave signals which will inform our understanding of astrophysics and cosmology. We describe the current configuration of the GstLAL low-latency search and show its readiness for

*rebecca.ewing@ligo.org

the upcoming observation run by presenting its performance on a mock data challenge. The mock data challenge includes 40 days of LIGO Hanford, LIGO Livingston, and Virgo strain data along with an injection campaign in order to fully characterize the performance of the search. We find an improved performance in terms of detection rate and significance estimation as compared to that observed in the O3 online analysis. The improvements are attributed to several incremental advances in the likelihood ratio ranking statistic computation and the method of background estimation.

DOI: 10.1103/PhysRevD.109.042008

I. INTRODUCTION

Since the first observing run (O1) of the LIGO Scientific, Virgo, and KAGRA Collaboration (LVK), GstLAL, a matched filtering based gravitational wave search pipeline [1], has participated in the discovery of groundbreaking gravitational wave events. GstLAL was among the search pipelines that made the first direct detection of gravitational waves from a merging binary black hole (BBH), known as GW150914 [2]. In the second observing run (O2), GstLAL was the first pipeline to observe the binary neutron star (BNS) merger known as GW170817, whose discovery kickstarted the field of multimessenger astronomy [3,4]. In the third observing run (O3), GstLAL detected $\mathcal{O}(10)$s of gravitational wave signals including the first ever neutron star-black hole binary (NSBH) mergers [5] and the very heavy BBH merger, GW190521, which resulted in a remnant object in the intermediate mass black hole (IMBH) mass region [6].

The GstLAL pipeline can be operated in one of two configurations; a low-latency or "online" mode and an "offline" mode. The online configuration of the GstLAL analysis proceeds in near real time as strain data becomes available from the interferometers [currently, LIGO Hanford (H), LIGO Livingston (L), and Virgo (V)]. The online analysis enables the prompt detection of gravitational wave events, allowing for rapid communication to the external community for electromagnetic follow-up. In order to provide the best opportunities for multimessenger astronomy, it is imperative that the low-latency analyses perform optimally. This includes reliable signal recovery, accuracy of source property estimation, and the ability of the search to keep up with real-time data and provide results as quickly as possible.

In contrast, the offline analysis proceeds on long timescales relative to the low-latency distribution of strain data. The offline analysis can benefit from a fuller understanding of the detector noise and the ability to rerank the significance of candidates against the full asynchronous background estimate collected over the entire run duration. Since the likelihood ratios and false alarm rates (FARs) of the candidates are recomputed relative to the full background, it is also possible to make adjustments to the signal model and mass model compared to what is used in the online analysis [7]. All of these factors can contribute to higher sensitivity, as quantified by the sensitive volume-time $\langle VT \rangle$, in the offline analysis.

In this paper, we will focus on the online configuration and aim to characterize the GstLAL pipeline's performance toward the fourth observing run (O4). In Sec. II we will describe the current configuration of the GstLAL online analysis. Additionally, we describe the gravitational wave low-latency test suite (`gw-lts`) software package as a tool for monitoring the performance of gravitational wave search pipelines in low latency. Then, in Sec. III we demonstrate the performance of the pipeline by presenting results from a mock data challenge (MDC). We will conclude in Sec. IV with a description of ongoing development towards O4.

II. SOFTWARE DESCRIPTION

A. GstLAL

The low-latency GstLAL inspiral workflow consists of two broad stages; a setup stage where precomputed data products are generated and stored on disk and a persistent analysis stage where strain data is filtered in near real time and candidate events are identified. We will give a brief description of the current workflow and configuration choices to be used in operating the GstLAL analysis during O4. A diagram of the low-latency workflow is shown in Fig. 1. For a more detailed description of the GstLAL analysis methods as of the end of O1 and O2 see [1,8], respectively. The GstLAL software package is described in [9].

Before the GstLAL analysis is launched, the template bank is first split into two halves in a process referred to as "checkerboarding". Each checkerboarded bank is constructed by taking alternating neighboring templates from the full bank. The checkerboarded banks are redundant as they cover the same parameter space while having unique individual templates. The full O4 template bank includes 1.8×10^6 templates, meaning that each checkerboarded bank has about 9×10^5 templates. The effectualness of the checkerboarded banks is validated in [11]. With this configuration the overall analysis can be split across two independent computing sites which improves the robustness of the analysis to upstream failures. According to the low-latency online inspiral detection (LLOID) method, the

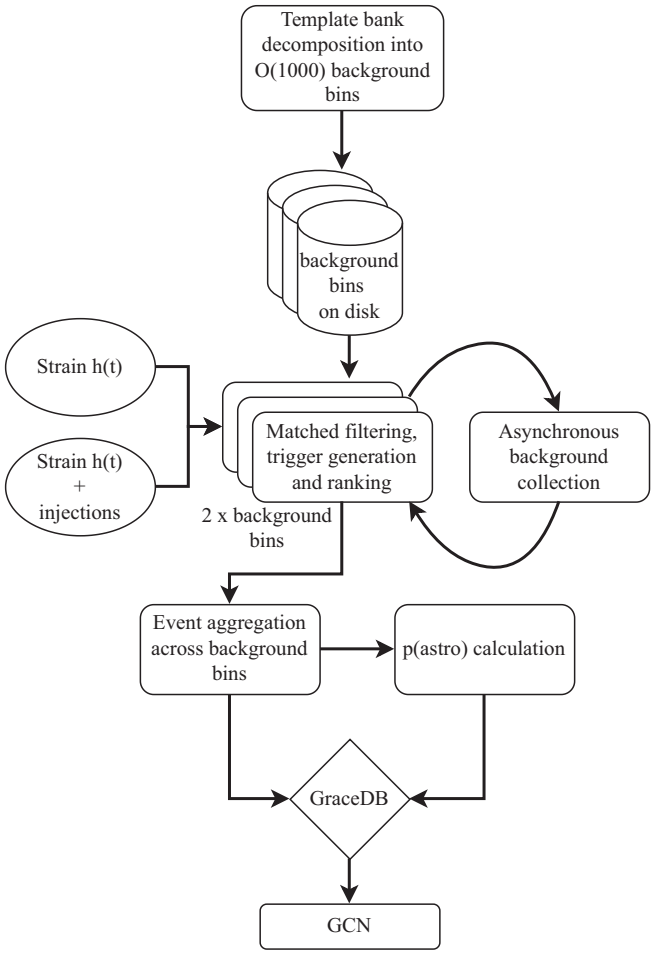

FIG. 1. The low-latency GstLAL inspiral analysis workflow. The full template bank must first be decomposed into $\mathcal{O}(1000)$ independent background bins via the LLOID method of singular value decomposition and time slicing [10]. The strain data is transferred from the interferometer sites at LIGO Livingston, LIGO Hanford, and Virgo to the computing clusters where it will be read from disk by the GstLAL pipeline. Filtering, trigger generation, and candidate ranking proceeds in parallel for each background bin independently. These filtering and ranking jobs are duplicated to process strain channels which include simulated signals injected into the data. Background statistics are collected independently in each background bin and asynchronously marginalized over the full parameter space in order to inform the FAR estimation. Candidate events are aggregated in time across all background bins, using the maximum SNR or minimum FAR as a metric for determining which candidates will be uploaded to GraceDB. Finally, candidates passing the public alert FAR threshold will be disseminated via GCN.

checkerboarded template banks must then be split into independent bins of waveforms, hereafter referred to as background bins, as shown in Fig. 1. A full derivation and motivation of the LLOID method, including the singular value decomposition (SVD) and template time slicing, is given by [10]. We first sort the template bank by the orthogonal post-Newtonian (PN) phase terms μ_1 and μ_2. These are linear combinations of the PN coefficients, ψ^0, ψ^2, and ψ^3, defined as follows [12]:

$$\mu_1 = 0.974\psi^0 + 0.209\psi^2 + 0.0840\psi^3,$$
$$\mu_2 = -0.221\psi^0 + 0.823\psi^2 + 0.524\psi^3. \quad (1)$$

Using these parameters to sort, we split the template bank into sub-banks each with ∼500 templates. Each background bin is then constructed by grouping 2 sub-banks together. When computing the decomposition, we require a 99.999% match between the reconstructed template waveforms and the initial physical waveforms. This value is chosen by balancing the need for computational efficiency with the need for accurately reconstructed waveforms. For the checkerboarded O4 template bank in [11], this produces ∼1000 background bins.

As part of the background bank construction during the setup stage of the analysis, the SVD waveforms are also whitened. For the initial whitening before filtering has begun, we use a reference power spectral density (PSD) generated from several hours of O3 data. As the analysis stage proceeds we rewhiten the SVD waveforms on a weekly timescale using recent PSDs in order to account for any long term changes to the detector characteristic noise. As the analysis runs, the PSD is continuously tracked using a fast Fourier transform (FFT) length of 4 seconds. Such a short length of FFT in the whitening stage of the pipeline reduces latency at the cost of a less accurate PSD measurement which could potentially bring a loss in sensitivity while filtering.

The low-latency analysis ingests strain data, as well as data quality and interferometer state information from frame files. Each frame includes 1 second of data. The frames are distributed from the detectors via Apache Kafka, an open source event streaming platform. After streaming from the detector sites, frames are stored in shared memory partitions, where they are accessed by the GstLAL analysis. The frames are then processed in buffers 4096 bytes at a time by each filtering job in the GstLAL pipeline as shown in Fig. 1. In O4, the GstLAL pipeline will also ingest a parallel stream of strain data including simulated compact binary coalescence (CBC) signals injected into the data. These injections will be based on the inferred astrophysical distribution of sources based on the gravitational wave transient catalog (GWTC-3) [13].

It is known that the LIGO and Virgo data are not "well-behaved" and include transient and non-Gaussian noise components known as glitches. These glitches can be mistaken for astrophysical signals, especially high-mass BBH templates which are short in duration within the LIGO-Virgo frequency band. To mitigate the negative effects of non-Gaussian data, the GstLAL pipeline gates particularly glitchy whitened $h(t)$ strain data using a threshold on the amplitude of the data in units of standard deviations. In gating the strain data, we must be careful to

balance the desire to reduce false positives (i.e., mistaking a glitch for an astrophysical signal) with the desire to avoid false negatives (i.e., mistaking an astrophysical signal for a glitch). Since we know that signals from heavy mass CBC systems (for example, IMBH binaries) with high amplitudes and short durations tend to resemble glitches, we want to be conservative with gating data while filtering templates in this region of the parameter space. However, for low-mass systems we can more aggressively gate with a lower threshold since these signals typically have a smaller amplitude and much longer duration. For this reason, we choose a gate threshold which is linear in the chirp mass,

$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}}, \tag{2}$$

and calculate it for each background bin as follows [14]. We first compute a gate ratio defined as

$$R_{\text{gate}} = \frac{\sigma_{\max} - \sigma_{\min}}{\mathcal{M}_{c,\max} - \mathcal{M}_{c,\min}}, \tag{3}$$

where we choose the minimum and maximum gate thresholds, $\sigma_{\min}$ and $\sigma_{\max}$ as 15.0 and 100.0, respectively. The $\mathcal{M}_c$ minimum and maximum are taken to be $\mathcal{M}_{c,\min} = 0.8$ and $\mathcal{M}_{c,\max} = 45.0$. Then, the $h(t)$ gate threshold is calculated for each background bin as

$$\sigma_{thr} = R_{\text{gate}} \times \Delta\mathcal{M}_c + \sigma_{\min}. \tag{4}$$

Here, $\Delta\mathcal{M}_c$ is the difference between the maximum $\mathcal{M}_c$ in the given background bin and the $\mathcal{M}_{c,\min}$. This produces gate thresholds ranging from ∼15 for the smallest mass bins to ∼325 for the largest mass bins.

In the O4 configuration, we choose a filtering stride of 0.25 seconds, meaning that the matched filter output is computed in stretches of 0.25 seconds at a time. The small stride is chosen to reduce latency in the filtering stage. Triggers are defined by peaks in the signal-to-noise (SNR) time series output by the matched filtering which pass a threshold of 4.0. The GstLAL analysis has allowed for single detector candidates since O2 and will continue to do so in O4. However, when calculating the likelihood ratio of single detector candidates we apply a penalty to down rank their significance. This is a tunable parameter and the value to be used in O4 will be discussed in more detail in Sec. III B. Coincident candidates include triggers from the same template in at least two detectors. We require that the end times of coincident triggers be within 0.005 seconds of each other after accounting for the light travel time between detectors. Together, the coincidence threshold and the requirement that triggers across detectors ring up the same template provide a strong signal consistency test for candidate events.

The GstLAL pipeline uses the likelihood ratio as a ranking statistic to assign significance to gravitational wave candidates [15,16]. Recent improvements to the likelihood ratio computation towards O4 are given in [7]. These include an upgraded analytic SNR-ξ^2 signal model and a method for removing signal contamination from the background which is also described in [17]. The background noise in each detector is estimated by collecting ranking statistic data from single-detector triggers observed in coincident time. We exclude triggers from times when only one detector is operating since these triggers may be astrophysical signals. These background estimations are cumulative and "snapshotted" to disk every 4 hours. The filtering jobs which process injection strain data do not collect their own background estimations. This is because the high rate of injected signals in the data would contaminate the background and corrupt the statistics used for the FAR estimation. Instead, these injection filtering jobs use a copy of the background statistics collected by the corresponding noninjection filtering job which processes the same background bin.

While the pipeline is designed to run persistently, there is need to take the analysis down periodically. We remove each of our analyses for a short period of time on a weekly timescale with a staggered schedule so that at least one of the checkerboarded analyses is always observing. When an analysis is relaunched after this weekly downtime, we compress the background ranking statistic data by removing any values in the horizon distance history that differ fractionally from their neighbors by less than 0.003. This compression reduces the file size and memory use of the pipeline, which would otherwise grow without bounds over the duration of the observing period.

For FAR estimation, the ranking statistic data is marginalized by adding counts from the SNR-ξ^2 background distributions collected in each background bin. The histograms are marginalized over in a continuous loop, taking several hours to complete each iteration. The marginalization is cumulative in time so that as the run proceeds, we collect more and more background counts. To account for the two redundant checkerboarded analyses, we apply a FAR trials factor of 2 to each trigger.

Gravitational wave events passing a FAR threshold of one per hour will be uploaded to the gravitational wave candidate event database (GraceDB) [18]. Because the GstLAL pipeline filters the strain data in ∼1000 independent background bins, it is not only possible but highly probable that there will be multiple triggers associated with each physical gravitational wave candidate. The number of triggers per candidate could range from a few for quieter signals to several tens for louder signals. In order to reduce the number of calls made to GraceDB we aggregate these triggers in time across background bins by the maximum SNR and only upload the current best candidate. In this aggregation stage, triggers from different background bins

are grouped into candidates using a coincidence window defined by rounding $t_{\rm end} - dt$ down to the nearest half second and rounding $t_{\rm end} + dt$ up to the nearest half second. Here, $t_{\rm end}$ is the end time of the trigger and $dt = 0.2$ seconds. The first trigger received by the aggregator for a given candidate is uploaded to GraceDB immediately. Any subsequent triggers for the same candidate which are found with higher SNR are uploaded with a 4 second geometric wait time. That is, after the first upload, the second upload will not be made until 4 seconds later, the third upload until 4^2 seconds later, and so on. The aggregation stage of the pipeline is illustrated in Fig. 1.

Finally, the GstLAL pipeline calculates a probability of astrophysical origin, or $p(\text{astro})$, for each event uploaded to GraceDB. The $p(\text{astro})$ is a measure of the event's significance, and as we also compute the probability that the event originates from each CBC source class (BNS, NSBH, or BBH) it gives an indication of the likelihood that an event will have an electromagnetic counterpart. Therefore, the $p(\text{astro})$ is an important quantity to help astronomers determine when to follow up gravitational wave candidates. More information about the GstLAL pipeline's computation of $p(\text{astro})$ can be found in [19].

B. GW low-latency test suite

The `gw-lts` software is designed to provide consistency checks and real-time feedback on the reliability of science outputs of gravitational wave search pipelines. By using simulated signals injected in the strain data, we can compare the pipeline performance to what is expected.

The test suite requires a source of truth for the signals that are present in the data. For this, we rely on an injection set on disk which defines all of the injections, including all intrinsic and extrinsic parameters, and the global positioning system (GPS) times at which they appear in the strain data. Using a live estimate of the PSD and the injected signal's sky location we can compute the expected SNR. For information about recovered events, GraceDB is taken as the source of truth. The IGWN-ALERT software package is a messaging system built on Apache Kafka which sends notifications of GraceDB state changes to subscribed users. The test suite subscribes to notifications from IGWN-ALERT which are sent for any new or updated event on GraceDB. The injections are then matched with recovered alerts in low latency by finding coincidences within a small Δt which we take to be ± 1 seconds. This time window was chosen to be very small compared to a typical injection rate to avoid erroneous coincidences.

Once an injected signal is matched with a recovered event, the information is passed to an arbitrary number of independent jobs via Apache Kafka. The jobs compute metrics associated with the injection recovery such as the $\langle VT \rangle$, accuracy of source classification and sky localization, and accuracy of point estimates of the source intrinsic parameters. The `gw-lts` capabilities are described in further detail in Sec. III.

All of the metrics computed by the `gw-lts` are stored with InfluxDB, which is an open source platform for storing and querying time series data. We use the data visualization tool Grafana to display the data in real time in online dashboards. With this infrastructure, we are able to track changes in the performance of the analysis on the timescale of seconds. Additionally, from the Influx database we are able to keep an archival record of the performance metrics.

III. MOCK DATA CHALLENGE RESULTS

To demonstrate the performance of the GstLAL analysis and our readiness for O4, we participated in a MDC consisting of a forty day stretch of HLV O3 strain data taken from January 5, 2020 15:59:42 to February 14, 2020 15:59:42 UTC and replayed so as to be analyzed in a low-latency configuration. The MDC also provided a set of identical strain channels with injected BNS, NSBH, and BBH signals. Details of the injection distributions used in the MDC can be found in [20]. Injected signals were placed in the strain data at a rate of one per ∼40 seconds, leading to a total of 5×10^4 total injections throughout the MDC duration.

In this section we seek to quantify the performance of the GstLAL pipeline in its latest configuration. We will first show the recovery of known gravitational wave events in the MDC data, as well as highlight any potential retraction level events. We will then detail the results of the MDC injection campaign. Finally, we present the stability and performance of the pipeline in terms of its uptime and latency.

A. Gravitational wave events

There are nine gravitational wave events in the duration of the MDC replay data which were previously published as significant candidates in GWTC-3 [21]. These are described throughout the remainder of this section and summarized in Table I, comparing the GstLAL pipeline's recovery of the signal in O3 to that in the MDC. We recover all of the nine candidates at the 1 per hour FAR threshold for uploading to GraceDB. Of these, three were found with high significance by GstLAL in the O3 online analysis. Two were found with marginal or subthreshold significance online but with high significance offline. The remaining four candidates were found by GstLAL only in the offline analysis. The recovery of all previously published candidates shows that the pipeline is performing with at least the same capability as in O3.

Figure 2 shows the count of observed candidates versus inverse false alarm rate (IFAR). The expected background counts are calculated using an estimated livetime which is equal to the wall clock time from the first to the last candidate. Additionally, we apply a trials factor of 2 to the

TABLE I. Gravitational wave candidates from January 5, 2020 15:59:42 to February 14, 2020 15:59:42 UTC as recovered by the GstLAL pipeline during the O3 online analysis and during the MDC. The instruments provided are those which participated in the event, that is, contributed a trigger with SNR > 4.0. Here, the SNR is the recovered network SNR, FAR is the false alarm rate in inverse years, and $p(\text{astro})$ is the probability of astrophysical origin. The two low-significance candidates identified with FAR above the public alert threshold in the O3 online analysis are indicated with FAR > 1.2 per year. We note that this FAR threshold is after a trials factor corresponding to the number of operating pipelines has been applied. The last four candidates in the table were not recovered by GstLAL in the O3 online analysis.

	O3 Online				MDC			
Name	Inst	SNR	FAR (yrs^{-1})	$p(\text{astro})$	Inst	SNR	FAR (yrs^{-1})	$p(\text{astro})$
GW200112_155838	L1	18.79	4.05×10^{-4}	>0.99	L1	18.46	1.01×10^{-7}	>0.99
GW200115_042309	H1L1	11.42	6.61×10^{-4}	>0.99	H1L1	11.48	2.55×10^{-4}	>0.99
GW200128_022011	···	···	>1.2	···	H1L1	9.98	1.44×10^{-4}	>0.99
GW200129_065458	H1L1V1	26.61	2.11×10^{-24}	>0.99	H1L1V1	26.30	1.78×10^{-17}	>0.99
GW200202_154313	···	···	>1.2	···	H1L1	11.09	1.69×10^{-2}	>0.99
GW200208_130117	···	···	···	···	H1L1	10.56	4.92×10^{-5}	>0.99
GW200208_222617	···	···	···	···	H1L1	8.00	2.02×10^{3}	0.48
GW200209_085452	···	···	···	···	H1L1	9.96	1.20	>0.99
GW200210_092254	···	···	···	···	H1L1	9.28	3.64×10^{3}	0.27

FARs since we only include candidates from one of the checkerboarded analyses. Figure 2 shows the nine known gravitational wave events recovered in the MDC. The candidate IFAR statistics agree with the expected counts from noise at lower IFAR and diverge due to the presence of signals at higher IFAR.

GW200112155838 was a BBH candidate observed in LIGO Livingston data as a single detector candidate with chirp mass $35.37M_\odot$ in O3 and $33.37M_\odot$ in the MDC. In the MDC, we recover the event with a comparable SNR as that observed in O3, however in the MDC the FAR is significantly lower.

GW200115042309 was the first confident NSBH detection found in O3. The event was found as a coincident trigger in LIGO Hanford and LIGO Livingston data. The SNR recovery, FAR estimation, and chirp mass estimation are all equivalent in the MDC to what was observed in O3.

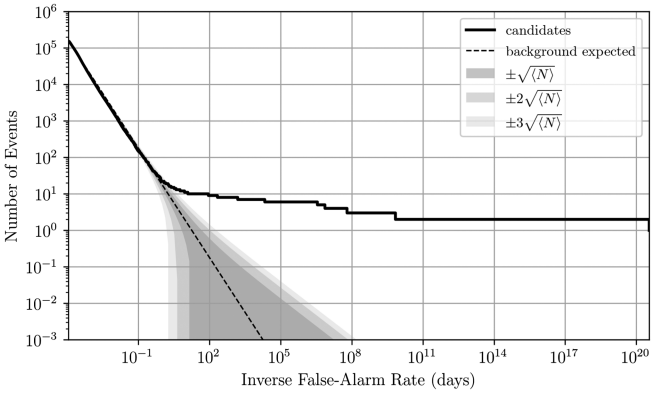

FIG. 2. Count of observed candidates vs IFAR in days. The dashed line is an estimation of the expected number of background counts, assuming a FAR threshold of 1×10^4 per day. The one, two, and three σ error regions are indicated by the shaded regions.

GW200129065458 was a BBH and the loudest gravitational wave signal in the duration of the MDC with O3 SNR = 26.61. The event was recovered well below the public alert FAR threshold in both O3 and the MDC.

GW200128022011 and **GW200202154313** are both BBH candidates found by GstLAL in the O3 online analysis with low significance. Both candidates were found with FAR above the O3 public alert threshold of 1.2 per year, where a trials factor corresponding to the number of operating pipelines has been applied. Later, during the offline analysis they were recovered as significant candidates and included in GWTC-3 [21]. In the MDC we recover both candidates with significantly lower FARs, both well below the public alert threshold. Therefore, if similar events occur during O4, we can expect to recover them as significant public alerts.

GW200208130117 was not recovered by GstLAL in the O3 online analysis, however it was found in the offline analysis by GstLAL as a highly significant candidate [21]. As recovered in the MDC, this event is a BBH candidate with chirp mass $34.50M_\odot$ and a much lower FAR than what was found in either the O3 online or offline analyses.

GW200208222617 was only recovered by GstLAL as a subthreshold candidate in the O3 offline analysis, and its inclusion as a significant candidate in GWTC-3 was due to its recovery by other CBC pipelines [21]. The GstLAL pipeline did not recover this event in O3 online. In the MDC the event was recovered with low significance at SNR = 8.00 and FAR = 2.02×10^3 per year.

GW200209085452 and **GW200210092254** were not recovered by GstLAL in the O3 online analysis, however they were found in the offline analysis by GstLAL. GW200209085452 was recovered in the MDC as a

significant candidate, although with a FAR of 1.20 per year it is less significant than in the O3 offline analysis. This event is a BBH candidate with chirp mass $= 39.45M_{\odot}$ found in LIGO Hanford and LIGO Livingston data. GW200210_092254 was found as a subthreshold candidate in the MDC with FAR $= 3.64 \times 10^3$ per year. If astrophysical, the event would be an NSBH candidate with chirp mass $= 7.89M_{\odot}$. GW200210_092254 was considered as a highly significant candidate in GWTC-3 due to its recovery by other pipelines, however the GstLAL trigger for this candidate had a $p(\text{astro})$ below the threshold of 0.50 and on its own would be considered marginal. We note that the candidate was recovered with a higher FAR and a lower $p(\text{astro})$ in the MDC than in the O3 offline analysis. Still, the recovery of these candidates in the MDC, even at subthreshold significance which is the case for GW200210_092254, demonstrates an improvement over the O3 online sensitivity.

The improved performance of the GstLAL pipeline in the MDC as compared to the O3 online analysis can be attributed to a number of incremental improvements made to the likelihood ratio ranking statistic and background estimation. [7] describes an improved signal model and [17] introduces a new method for a time-dependent background wherein contamination is reduced by removing signals counts from the background SNR $- \xi^2$ histograms. Each of these changes have introduced a small improvement to the $\langle VT \rangle$ which, when combined, leads to a noticeable increase in sensitivity and corresponding number of detected events.

B. Retractions

In O3, there were 23 public gravitational wave candidates which were subsequently determined to be terrestrial in origin and thus retracted. Of these, GstLAL contributed to 15. In O4, we hope to significantly reduce the number of retractions produced by the GstLAL pipeline. Four of the 23 retractions took place during the stretch of data covered by the MDC. These are S200106au, S200106av, S200108v, and S200116ah [22–24].

GstLAL did not upload triggers for S200106au and S200106av during O3. In the MDC, these events would have occurred at a time before the pipeline had collected sufficient background to begin ranking candidates, and thus we did not upload triggers for these events. The retractions S200108v and S200116ah were GstLAL-only candidates in O3, both being found as single detector candidates in LIGO Livingston. Again, the time corresponding to S200108v would have been early enough in the MDC cycle that the pipeline was not ranking or uploading triggers yet, so we cannot make any comparison to our performance in O3 for this retraction. Finally, we did not produce any trigger below the 1 per hour FAR threshold for uploading to GraceDB corresponding to S200116ah in the MDC, despite the analysis being fully burned in and operating in a stable state. This means that a similar noise event in O4 may not lead to a spurious candidate and subsequent retraction. However, there is insufficient data within the MDC to infer any changes in performance in terms of the rate of retractions.

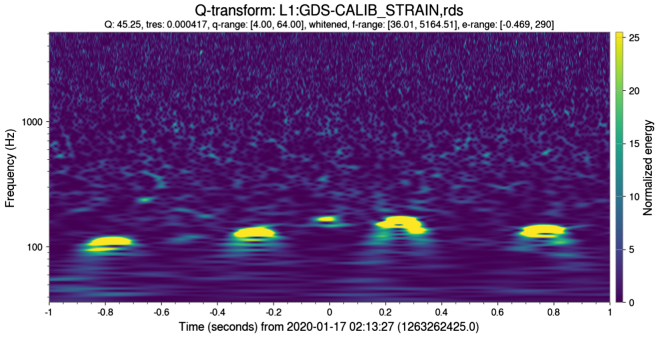

FIG. 3. Spectrogram of L1 $h(t)$ data for ± 1 second around the time of the retraction level candidate recovered in the MDC. This candidate is expected to be terrestrial in origin due to the clear presence of glitches in this data.

In addition to the retracted candidates uploaded by GstLAL in O3, for the purpose of the MDC we define a "retraction level candidate" as any gravitational wave candidate uploaded with a FAR less than one per year which is not in the list of previously published candidates discussed earlier in this section. Over the duration of the MDC, we find one such retraction level candidate. This was a single-detector candidate found in LIGO Livingston with an SNR of 14.5 and a FAR of 1.67 per year which was low enough to be counted as significant. The LIGO Livingston data around the event time shows the clear presence of scattering glitches, as shown in Fig. 3. Further evidence of terrestrial origin for this candidate is that no coincident triggers were recovered in LIGO Hanford or Virgo despite both of these detectors operating normally at the time. Candidates recovered in only a single detector are more susceptible to uncertainty since they lack the strong signal consistency test of coincidence in multiple detectors. For this reason, there has been a penalty applied to the ranking statistic of single detector candidates which down weights their significance. In the O3 offline analysis and in the MDC we used a singles penalty of 12 in log likelihood ratio, however in order to reduce the number of similar retraction level events in O4, we plan to increase the singles penalty to 13. With this penalty applied, the retraction event in the MDC would be down weighted and expected to be recovered with a FAR greater than two per year.

C. Recovered injections

There were 5×10^4 simulated signals injected into the five week duration of the MDC strain data. Of these, many had component masses and spins outside the region of parameter space covered by our template bank. In addition to the injection parameters, the expected recovery of each

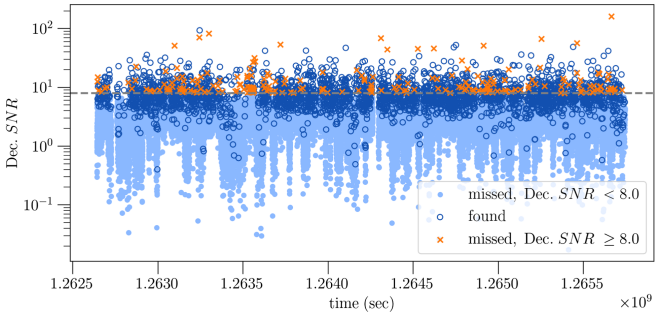

FIG. 4. Time-series of injected decisive SNR for injections with component masses and spins within the O4 template bank. Dark blue circles indicate injections that were recovered below a 2 per day FAR threshold. Orange crosses and light blue markers indicate injections not recovered below this FAR threshold, where orange crosses are injections with decisive SNR > 8.0. Times on the horizontal axis are GPS times shifted to the original O3 epoch.

injection is dependent on the set of interferometers producing science quality data at the time of the injection. During times when no interferometers are operating we of course do not expect to recover any injections. We define the decisive SNR as the SNR in the second most sensitive interferometer during times when multiple interferometers were observing, and the only available SNR otherwise. The decisive SNR is a more informative measure of the loudness of an injection than the network SNR since it wraps in information about the set of operating interferometers. While all injections have network SNR ≥ 4.0, we find that many injections have decisive SNR < 4.0.

Fig. 4 shows the time series of decisive SNR for all injections throughout the MDC. For the purpose of this paper, we focus on injections whose parameters fall inside our bank, that is injections with component masses between $1.0M_\odot$ and $200M_\odot$, with total masses $m_1 + m_2 < 400.0M_\odot$ and mass ratios $q = m_1/m_2 < 20$. For objects with mass $< 3.0M_\odot$ the template bank restricts spins perpendicular to the orbital plane $|s_{i,z}| < 0.05$ and for objects with mass $> 3.0M_\odot$ allows $|s_{i,z}| < 0.99$. We use the effective precession spin, χ_p, defined in [25] as

$$\chi_p = \frac{\max(a_1 \cdot s_1, a_2 \cdot s_2)}{a_1 \cdot m_1^2} \tag{5}$$

to quantify the in-plane spin of injections. Here, $a_1 = 2 + 3/2q$, $a_2 = 2 + 3q/2$, $s_i = \sqrt{s_{i,x}^2 + s_{i,y}^2}$ and the mass ratio q assumes $m_1 \geq m_2$. Since the template bank does not include any in-plane spins, we focus on injections with $\chi_p < 1 \times 10^{-3}$. However, we find that all injections with one component mass $< 3.0M_\odot$ have spins outside the range of the bank, therefore we relax the spin conditions on these components. The mass and spin restrictions that we use are summarized in Table II. Finally, to account for the fact that not all interferometers were providing science quality data at all times, we highlight injections with an estimated decisive SNR ≥ 8.0. These cuts leave a total of 1457 injections during the five week MDC. Of these, there are 597 BBH, 482 BNS, and 378 NSBH injections.

TABLE II. Restrictions on the masses (m_1, m_2, M, and q), spins perpendicular to the orbital plane ($s_{1,z}$ and $s_{2,z}$), and spins parallel to the orbital plane (χ_p) of injections according to the O4 template bank boundaries. The "–" in the first two rows indicates that we make no restrictions on the spins for injections with $m_i < 3.0M_\odot$. This relaxation is done because the template bank restricts NS spins $|s_{i,z}| < 0.05$, which would effectively remove most BNS and NSBH injections from consideration.

m_1	m_2	M	q	$\lvert s_{1,z}\rvert$	$\lvert s_{2,z}\rvert$	$\lvert\chi_p\rvert$
1,3	1,3	<6	<0.33	...	...	...
3,200	1,3	<203	<20	...	...	...
3,200	3,200	<400	<20	<0.99	<0.99	<0.001

The injected SNRs are not known in advance of the MDC, but we estimate them using `gw-lts`. We calculate the injected strain time series using the injection end time, sky position, and other source intrinsic parameters assuming an `IMRPhenomPv2_NRTidalv2` waveform [26]. We use a running estimate of the detector PSDs and estimate the SNR with a lower (upper) frequency cutoff of 10.0 (1600.0) Hz. Figure 5 shows the recovered and estimated injected SNR for each detector. If the template bank did not have a sufficient minimal match, we may expect to see systematically lower recovered SNRs than the expected values. However, we find that the recovered SNR generally aligns with the expected SNR. We note that the figure shows a wider spread in the recovered SNRs for LIGO Hanford than LIGO Livingston. This is likely due to the greater sensitivity in LIGO Livingston such that

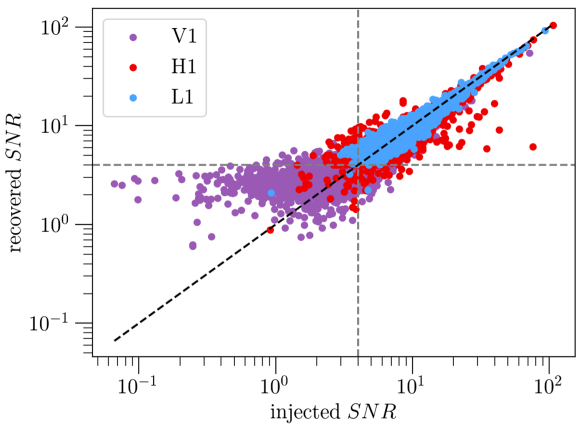

FIG. 5. Recovered and estimated injected SNR in each interferometer: LIGO Hanford (H), LIGO Livingston (L), and Virgo (V). The black dashed line shows the diagonal and the gray dashed lines show injected and recovered SNRs $= 4.0$, the threshold which defines a trigger.

TABLE III. Injection efficiencies as defined in Eq. (6) computed using four FAR thresholds to count "found" injections: one per hour (the GraceDB upload threshold), two per day (the public alert threshold), one per month, and two per year. Source categories are defined in Table II and "ALL" combines injections from the three source categories.

FAR	BNS	NSBH	BBH	ALL
2.78×10^{-4} Hz	0.95	0.77	0.87	0.87
2.31×10^{-5} Hz	0.95	0.71	0.84	0.84
3.85×10^{-7} Hz	0.89	0.65	0.77	0.78
3.16×10^{-8} Hz	0.86	0.62	0.71	0.74

injections with low-recovered SNRs in LIGO Livingston may be expected to have had even lower SNRs in LIGO Hanford and thus not be recovered at the FAR threshold for inclusion in the figure.

An injection is considered "found" if it is recovered by the pipeline with a FAR passing some predetermined threshold, and "missed" otherwise. We will quote most results in the following sections with respect to a 2 per day FAR threshold. At the time of writing, this is the threshold expected to be used in O4 for sending public alerts [27]. However, the FARs of CBC signals will be subject to a trials factor corresponding to the number of operating pipelines, so the effective alert threshold will be lower. We define the injection recovery efficiency as

$$\frac{\text{found injections}}{(\text{found injections} + \text{missed injections})}. \tag{6}$$

At the 2 per day FAR threshold the efficiency was $= 0.84$ for all injections in the template bank. The recovered injection efficiencies for each source class are shown in Table III at four typical FAR thresholds. As is expected, the efficiencies are better at more conservative FAR thresholds. The analysis has the highest efficiency for injections consistent with BNS sources, and the lowest efficiency for NSBH sources. The injections in the MDC included precession effects while the O4 GstLAL template bank was constructed assuming only spins aligned with the orbital angular momentum. Therefore, the relatively lower NSBH recovery efficiency is expected as the precession will have a more significant effect on the gravitational waveform for binaries with more extreme mass ratios [28]. In future work, we may seek to quantify the efficacy of the template bank in the precessing parameter space.

We would expect the pipeline to recover all injections above some decisive SNR or network SNR threshold. However, Fig. 4 shows that there are several very high SNR missed injections throughout the duration of the MDC. We find that most of the missed injections with decisive SNR > 20.0 are high-mass BBH injections and a few are high-mass ratio NSBH injections. This results in a decrease of the BBH recovery efficiency as the injections increase in SNR, which is contradictory to our expectations. These injections are missed due to falling outside of the SNR-ξ^2 signal region used in the likelihood ratio calculation. The signal region is an analytic model which depends on the allowed mismatch[1] between recovered SNR time series and the template waveform as part of the autocorrelation ξ^2 test. If the allowed mismatch range is too strict, it will result in a narrow signal model which can exclude real signals. This effect is exaggerated at high SNR where we expect larger mismatches due to the discreteness of the template bank as well as waveform systematics. In the MDC, we used a mismatch range of 0.1–10%. The optimal mismatch range in the signal model is an open area of study. See [7] for a more detailed discussion.

1. Injection parameter recovery

In this section we will quantify the accuracy of point estimates of the source intrinsic parameters made by the GstLAL pipeline. These estimates simply come from the template parameters of the trigger which rang up the maximum SNR across background bins. An understanding of the parameter accuracy obtained by search pipelines can be useful to full parameter estimation efforts. For example, the Bayesian inference library Bilby [29,30] relies on the choice of prior probability distributions for intrinsic parameters. When parameters are well determined by the searches, Bilby can use narrow distributions around those values, otherwise more broad prior distributions must be used. We present parameter accuracy results for the chirp mass $\mathcal{M}_c$, effective inspiral spin χ_{eff}, mass ratio q, and the coalescence end time, t_{end}. The χ_{eff} is a mass-weighted combination of the component spins parallel to the orbital angular momentum $\hat{L}$, defined as

$$\chi_{\text{eff}} = \frac{(m_1\vec{s_1} + m_2\vec{s_2}) \cdot \hat{L}}{m_1 + m_2}, \tag{7}$$

where we take $\hat{L}$ to be in the z-direction. Both the injected and recovered masses quoted in this paper are in the detector frame. The error on a recovered parameter, λ is defined as

$$\text{error} = \frac{\text{recovered } \lambda - \text{injected } \lambda}{\text{injected } \lambda}, \tag{8}$$

for all parameters except for the end time, where we simply take the error as the difference between the recovered and injected end times in milliseconds.

[1]The "mismatch" can also mean the fractional loss in SNR due to differences between the template parameters and the true waveform. However, here we refer to the mismatch as defined in [7] which is an unnormalized quantity, therefore retaining a dependence on the SNR.

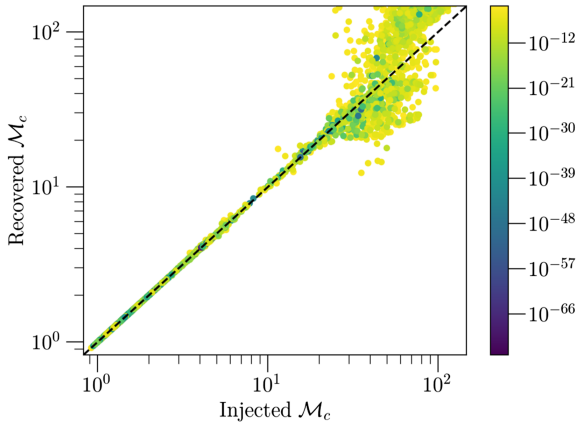

FIG. 6. Injected $\mathcal{M}_c$ for injections found with FAR $< 2.31 \times 10^{-5}$ Hz is shown on the horizontal axis. The vertical axis shows the recovered $\mathcal{M}_c$. The color bar is FAR.

It is well known that the chirp mass, $\mathcal{M}_c$, is one of the best measured parameters in gravitational wave detections, however the recovered accuracy is highly dependent upon the mass of the system. Injections with small $\mathcal{M}_c <$ 10–20$M_\odot$ are recovered with very accurate $\mathcal{M}_c$, but above this level, the accuracy starts to fall off, as shown in Fig. 6. For BNS injections, we find a mean $\mathcal{M}_c$ error 2.06×10^{-4} with a standard deviation of 8.33×10^{-4}. Similarly, the $\mathcal{M}_c$ in the NSBH region is recovered very well with mean -2.14×10^{-4} and standard deviation 6.26×10^{-3}. The BBH region has a higher $\mathcal{M}_c$ error over all, and additionally a much larger spread in the error with mean 1.54×10^{-1} and standard deviation 4.53×10^{-1}. Histograms of the recovered $\mathcal{M}_c$ error for each source class are shown in Fig. 7. The figure shows a skew in the recovered $\mathcal{M}_c$ error for BBH injections toward more positive values, i.e., we are more likely to overestimate the $\mathcal{M}_c$ in this region than to underestimate it. This skew is likely due to differences in the injected waveforms compared to the waveforms used to generate the template bank, however we leave a more in depth investigation of this issue to future work.

Figure 8 and Fig. 9 are scatter plots of the injected and recovered $\chi_{\rm eff}$ and q, respectively. These plots show that there is very little correlation between the injected and recovered values of these parameters. The mean and standard deviation on the recovered error for these parameters are given in Table IV.

A histogram of the difference between the injected and recovered injection end times is given in Fig. 10. Table IV shows the mean $t_{\rm end}$ difference across all source classes and detectors is 6.23 milliseconds with a standard deviation of 30.22 milliseconds. The 90th percentile on $|t_{\rm end}|$ is 25.6 milliseconds and we recover every injection with a recovered $t_{\rm end}$ less than a second away from the injected value. Figure 10 shows that for BNS and BBH injections, most of the recovered end times fall within ±50 milliseconds of the true injected end time. For BNS injections, the mean $t_{\rm end}$

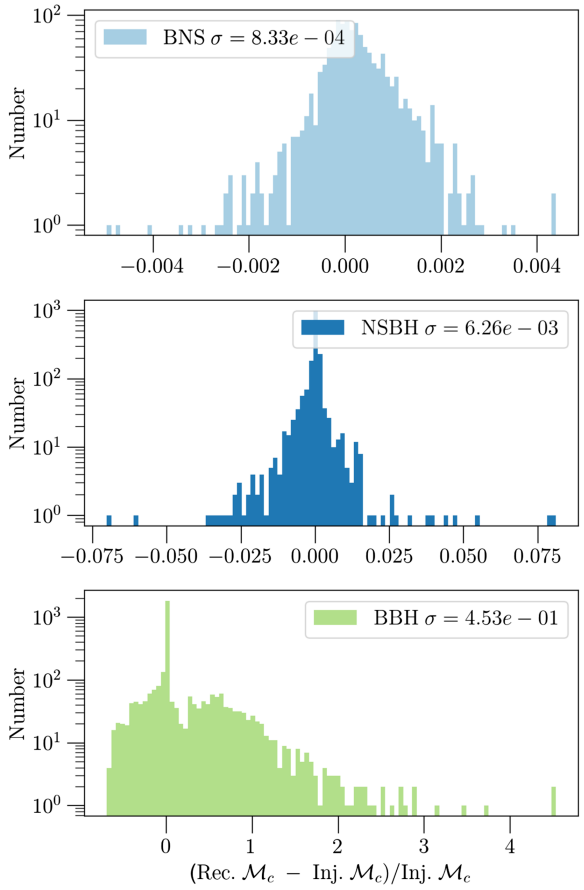

FIG. 7. Recovered $\mathcal{M}_c$ error for injections found with FAR $< 2.31 \times 10^{-5}$ Hz. The top panel shows injections in the BNS range of the parameter space, the middle panel shows NSBH detections, and the bottom panel shows BBH injections. The σ value in each panel indicates the standard deviation on the recovered $\mathcal{M}_c$ error.

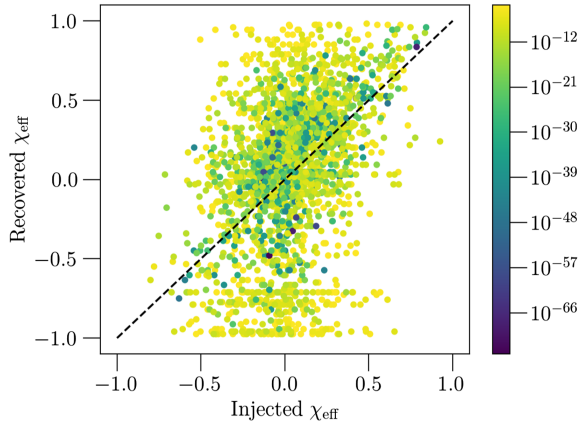

FIG. 8. Injected $\chi_{\rm eff}$ for injections found with FAR $< 2.31 \times 10^{-5}$ Hz is shown on the horizontal axis. The vertical axis shows the recovered $\chi_{\rm eff}$. The color bar is FAR.

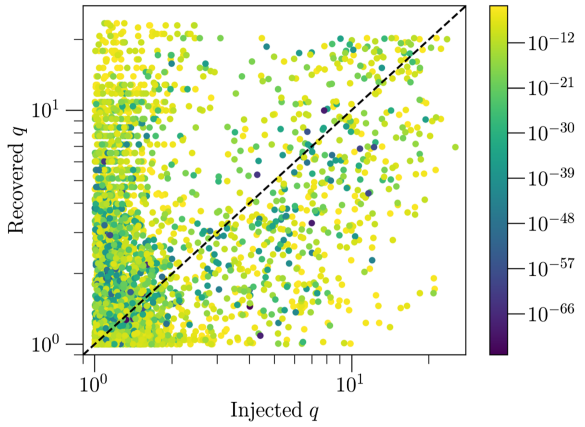

FIG. 9. Injected mass ratio, $q = m_1/m_2$, for injections found with FAR $< 2.31 \times 10^{-5}$ Hz is shown on the horizontal axis. The vertical axis shows the recovered mass ratio. The color bar is FAR.

difference is −0.90 milliseconds with a standard deviation 18.0. And for BBH injections, the mean $t_{\rm end}$ difference is 6.03 milliseconds with a standard deviation 11.3. However, for NSBH templates, there is a wider distribution, skewing towards higher positive values of the end time difference, with mean 18.7 milliseconds and standard deviation 59.3. This is expected to be due to waveform systematics.

2. Search sensitivity

The comoving volume is defined as [31,32]

$$V_c = 4\pi D_H \int_0^z \mathrm{d}z \frac{(1+z)D_A^2}{E(z)}, \tag{9}$$

with z as the redshift. D_H is the Hubble distance, D_A is angular distance, and $E(z)$ is the Hubble parameter. Multiplying this quantity by an observation time gives the surveyed spacetime volume. We compute the total injected volume-time, $\langle VT\rangle_{\rm inj}$, using the max redshift to which injections were distributed and the time range over which injections were placed. For the MDC this time range is 3.456×10^6 seconds. With this quantity we can then estimate the online $\langle VT\rangle$ in the MDC as

$$\langle VT\rangle = N_f \times \langle VT\rangle_{\rm inj}. \tag{10}$$

Here, N_f is the fraction of "found" injections out of the total number of injections in the data. We independently compute the $\langle VT\rangle$ for each source population. The max redshifts of the injection distributions are $z = 0.15$, 0.25, and 1.9 for BNS, NSBH, and BBH respectively. Table V gives the injected $\langle VT\rangle$ for each source class.

TABLE IV. Mean, $\bar{X}$, standard deviation, σ, and the 50th, 75th, and 90th percentiles on the recovered parameter error. The error is defined as in Eq. (8) for all parameters except for the end time, where we simply take the difference in milliseconds between the recovered and injected values as the error. Results are computed only including injections which were recovered below a FAR threshold of two per day. The percentiles are computed for the absolute value of each distribution.

	$\bar{X}$	σ	P_{50}	P_{75}	P_{90}
$\mathcal{M}_c$	0.15	0.45	0.007	0.33	0.73
$\chi_{\rm eff}$	5.77	252	1.34	3.71	10.8
q	1.39	2.86	0.45	1.67	4.97
$t_{\rm end}$	6.23	30.22	3.8	9.78	25.6

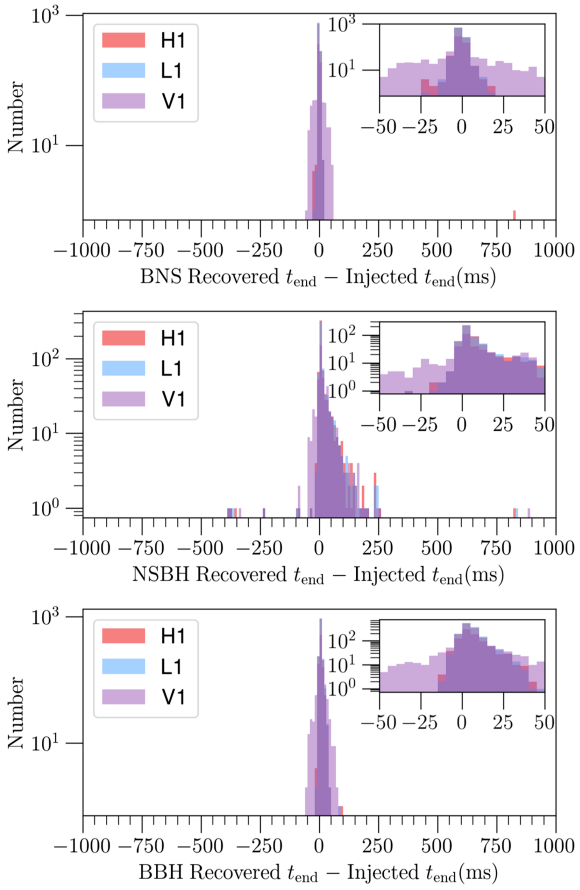

FIG. 10. Recovered end time accuracy in milliseconds of injections recovered with FAR $< 2.31 \times 10^{-5}$ Hz. Results are shown for each interferometer: LIGO Hanford (red), LIGO Livingston (blue), and Virgo (purple). BNS injections are shown in the upper panel, NSBH in the center, and BBH in the lower panel.

TABLE V. Values of the $\langle VT\rangle$ in cubic gigaparsec-years measured at the end of the MDC using a FAR threshold of 2.31×10^{-5} Hz compared to the injected $\langle VT\rangle$ in each source class.

$\langle VT\rangle$ Gpc^3 yrs	BNS	NSBH	BBH
$\langle VT\rangle_{\rm inj}$	1.08×10^{-1}	4.34×10^{-1}	29.1
$\langle VT\rangle$	3.49×10^{-4}	8.08×10^{-4}	1.23×10^{-1}

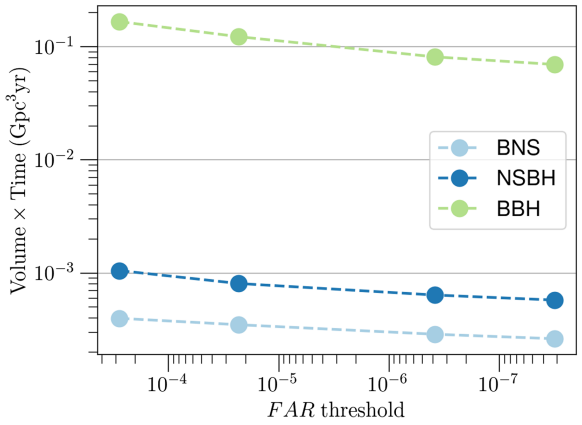

FIG. 11. $\langle VT \rangle$ in each source class at the end of the MDC at four different FAR thresholds; 2 per year, 1 per month, 2 per day, and 1 per hour. BNS $\langle VT \rangle$ is shown in light blue, NSBH $\langle VT \rangle$ in dark blue, and BBH $\langle VT \rangle$ in green.

Figure 11 shows the final $\langle VT \rangle$ in the MDC for each source class using four different FAR thresholds to determine whether injections count as found. In Fig. 12, we show the cumulative $\langle VT \rangle$ over the duration of the MDC using different thresholds to count found injections. For each source class, we find the highest $\langle VT \rangle$ by using a threshold of network SNR $= 10.0$ and the lowest $\langle VT \rangle$ with a threshold of decisive SNR $= 8.0$.

3. Sky localization and source classification

For efficient electromagnetic follow-up, it is vital that accurate sky localizations and source classifications are provided to the public in low latency. The sky localization information informs where electromagnetic observers should search on the sky to find coincident events. The accuracy and precision of sky localization information can have a direct impact on the time it takes to identify a counterpart, especially for narrow field of view telescopes. Sky localizations are produced in low latency for all events on GraceDB using Bayestar [33,34]. The sky localization calculation depends on the SNR time series around the coalescence time of the event. These are uploaded to GraceDB by the search pipelines as part of the event metadata. Since the potential for bright electromagnetic counterparts is highly dependent on the nature of the binary source, it is also important to provide accurate source classification so that observers may make informed decisions about when to follow up gravitational wave events. The probability that a gravitational wave candidate is astrophysical in origin is the $p(\text{astro})$, computed by GstLAL using the multicomponent FGMC formalism [19,35,36]. We use a population model with a Salpeter distribution for the source component masses m_1, m_2 given by [37]

$$p(m_1, m_2) \propto \frac{m_1^{-2.35}}{m_1 - m_{\text{min}}}, \tag{11}$$

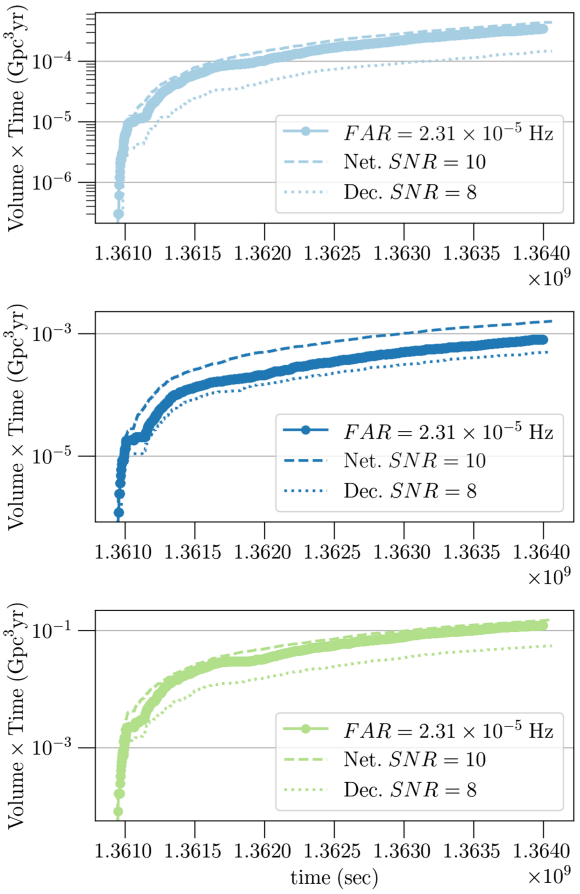

FIG. 12. The plot shows the cumulative $\langle VT \rangle$ time-series over the duration of the MDC for each source class; BNS in light blue (top panel), NSBH in dark blue (middle panel), and BBH in green (bottom panel). The $\langle VT \rangle$ is calculated using three different thresholds for counting "found" injections: FAR $< 2.31 \times 10^{-5}$ Hz (dot markers), network SNR > 10.0 (dashed line), and decisive SNR > 8.0 (dotted line).

with $m_{\text{min}} = 0.8 M_\odot$ and a uniform distribution in component spins, $s_{1,z}$, $s_{2,z}$. The $p(\text{astro})$ is further divided into the probability that a gravitational wave candidate originates from a BNS, NSBH, or BBH as

$$\begin{aligned} p(\text{astro}) &= 1 - p(\text{Terrestrial}) \\ &= p(\text{BNS}) + p(\text{NSBH}) + p(\text{BBH}). \end{aligned} \tag{12}$$

For this purpose, we use a cutoff of $3.0 M_\odot$ as the maximum neutron star mass to define the BNS, NSBH, and BBH regions. In this section, we will briefly summarize the accuracy of the Bayestar skymaps as well as the FGMC $p(\text{astro})$ for injections recovered by GstLAL during the MDC. Detailed information about recent developments to the FGMC $p(\text{astro})$ calculation and a comparison between offline and online $p(\text{astro})$ results for GstLAL events is given in [19].

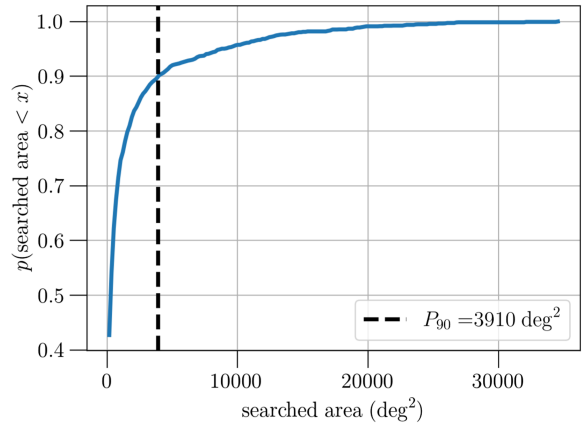

FIG. 13. Cumulative distribution function of sky-map searched area (deg^2) for injections recovered with FAR less than two per day.

To quantify the sky localization performance we will consider the skymap searched area and searched probability. The searched area is given in square degrees and represents the area within the credible region containing the injection's true source location. This gives an indication of the accuracy of the skymaps. The searched area can be interpreted as the sky area an electromagnetic observer would have to tile before reaching the true injection sky location assuming they start at the highest probability sky position given by the skymap. Therefore, a smaller searched area is desirable as it means an observer could find the injection with minimal telescope pointings. The searched probability is similarly the probability within the credible region containing the injection's true sky location. We aim to find 90% of injections within the 90% credible region of the true source location. If the searched probability P-P plot lies off the diagonal, we can make the interpretation that there is an inconsistency in the sky localizations, i.e., they may be accurate but lack precision, or vice versa.

Figure 13 shows the cumulative distribution function of the searched area for all injections recovered by GstLAL with FAR less than two per day. We find that the 90th percentile of searched area is 3910 deg^2. This is about 19% of one hemisphere of the sky. Since we know that injections recovered in coincidence by two or three interferometers will have more accurate sky localizations, we give statistics on the searched areas by coincidence type in Table VI. Fig. 14 shows the cumulative distribution function of the searched probabilities. We find that the 50th, 75th, and 90th percentiles on the searched probability are 0.53, 0.79, and 0.93, respectively. The CDF lies very near to the diagonal, indicating that Bayestar is producing well-calibrated sky localizations.

We will briefly discuss the performance of the $p(\mathrm{astro})$ in the MDC using the Sankey diagram in Fig. 15. The diagram is read from left to right. The width of each source band on the left corresponds to N_{source}, the number of recovered events in each source class: BNS, NSBH, BBH, and terrestrial. The terrestrial events are any candidates uploaded to GraceDB which do not coincide in time with an injection. The width of each band on the right is the sum of recovered $p(\mathrm{source})$. This diagram gives an indication of the relative misclassification between sources.

TABLE VI. 50th, 75th, and 90th percentiles on the searched area of injections of each coincidence type recovered with FAR less than two per day. Values are given in deg^2.

	P_{50}	P_{75}	P_{90}
ALL	271	1080	3910
1 IFO	3150	10,400	18,400
2 IFO	301	893	2470
3 IFO	31.9	140	357

For BNS injections, we find that $p(\mathrm{BNS})$ accounts for 90.3% of the recovered probabilities, while $p(\mathrm{NSBH})$ accounts for 9.7%. The recovered BBH and terrestrial probabilities of BNS injections are negligible. This indicates that BNS signals were most commonly mistaken as NSBH. For NSBH injections, we find $p(\mathrm{NSBH})$ is 64.1% of the recovered $p(\mathrm{astro})$, $p(\mathrm{BBH})$ makes up 33.8%, and $p(\mathrm{BNS})$ makes up just 2.10%. There is a significant amount of misclassification between BNS and NSBH signals. However, the misclassification is asymmetrical with very few NSBH injections being assigned high $p(\mathrm{BNS})$ whereas many more BNS injections are assigned a high $p(\mathrm{NSBH})$. The large proportion of NSBH signals assigned high $p(\mathrm{BBH})$ is even more concerning as these represent potentially electromagnetically bright signals which might not be followed up by astronomers due to the apparent high probability of originating from a BBH merger. This misclassification is an ongoing area of study and corrections to the $p(\mathrm{astro})$ calculation which mitigate

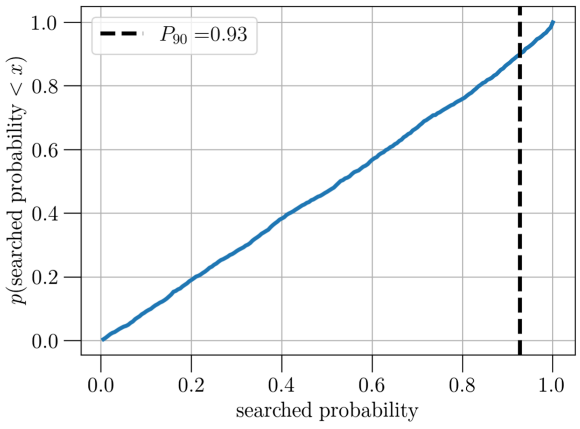

FIG. 14. Cumulative distribution function of skymap searched probability for injections recovered with FAR less than two per day.

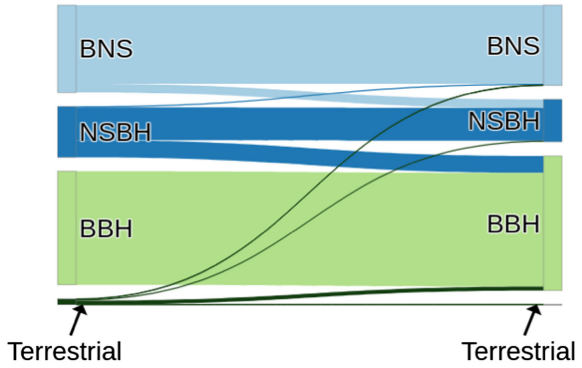

FIG. 15. Sankey diagram showing $p(\text{astro})$ classification of events uploaded to GraceDB during the MDC with FAR less than two per day. Events with an end time within ± 1 second of an injection are classified as either BNS, NSBH, or BBH using a neutron star mass boundary of $3.0 M_{\odot}$. Events that do not correspond in time with an injection are all classified as terrestrial.

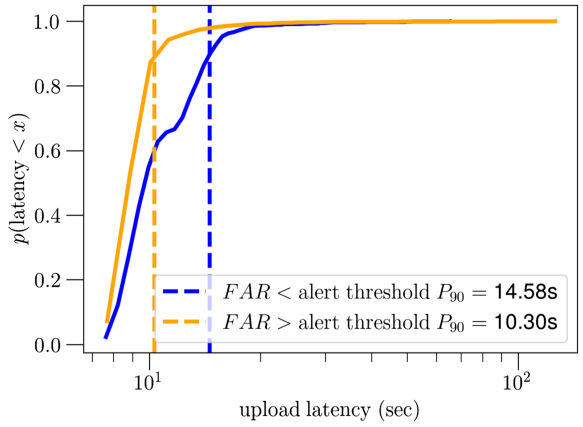

FIG. 17. Cumulative distribution of event upload latencies. The orange (blue) curve shows the distribution of latencies for the first event uploaded with FAR higher (lower) than the public alert threshold. The dashed lines show the location of the 90th percentile for each distribution.

this effect are discussed in [19]. The $p(\text{astro})$ calculation performs very well for BBH classification, with 100% recovered $p(\text{BBH})$.

4. Latency performance

Figure 16 shows a histogram of the upload latency for all recovered injections on GraceDB. The upload latency is defined as the difference between the time the event appears on GraceDB and the event coalescence time. The distribution is bimodal as a result of the built-in 4 second geometric wait time between uploads, discussed in Sec. II A. The first peak in the upload latency distribution is at 8–9 seconds and the second peak is at 13–14 seconds. This shows that the GstLAL pipeline is able to keep up with filtering data in real time and regularly produce gravitational wave candidates within $\mathcal{O}(10)$ seconds of the coalescence time.

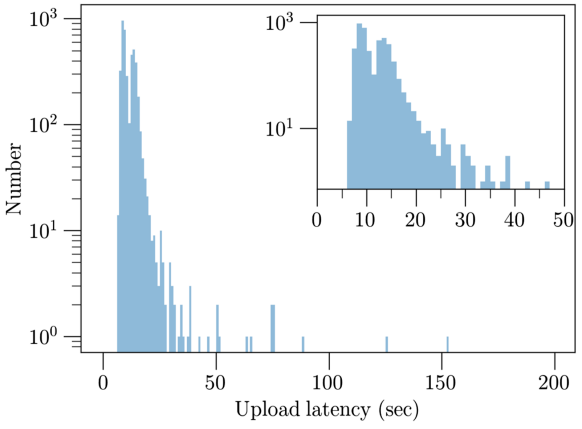

FIG. 16. Upload latency, defined as the difference between the GPS time of upload to GraceDB and the event coalescence time, in seconds.

The 4 second wait time in the event aggregation is implemented in order to reduce the total number of uploads by waiting a long enough time to collect many triggers across background bins before making an upload. Even though there is no built-in wait time for the first event upload, subsequent uploads are necessarily delayed by this method. If an event with FAR below the alert threshold comes after the first upload it can be delayed by at least 4 seconds. Figure 17 shows the cumulative distribution of upload latency for the first events above and below the alert threshold for each superevent in the MDC. We find that the low FAR event uploads are significantly delayed by the event aggregation process. After identifying this issue in the MDC we plan to make changes to the aggregation scheme in order to reduce the latency of alert quality uploads before the start of O4.

IV. CONCLUSION

GstLAL is a matched-filtering based gravitational wave search pipeline, which is operated in a low-latency configuration in order to identify signals within seconds of their arrival. We have introduced the `gw-lts` software as a useful auxiliary tool for characterizing the performance of such an analysis in real time. We have presented the performance of the GstLAL pipeline on a mock data challenge consisting of 40 days HLV data from O3 along with an injection campaign of simulated BNS, NSBH, and BBH signals.

Within the MDC data we recover nine previously published gravitational wave candidates at the one per hour FAR threshold. As only five of these were identified by the GstLAL pipeline in low latency in O3, we have demonstrated an improvement in the pipeline's signal recovery. We attribute this improvement to several incremental updates to the likelihood ratio computation [7] and

the new method of removing signals from the background as introduced in [17]. During the MDC, we find only one candidate which, if uploaded during O4, would likely be retracted. The candidate is identified in only a single detector and we expect that increasing the penalty applied to single detector candidates in the likelihood ratio would reduce our recovery of such spurious signals in the future. We have detailed the results of the injection campaign including efficiency of signal recovery across the parameter space, accuracy of estimated parameters, search sensitivity, sky localization and source classification accuracy, and typical latencies.

The configuration and performance of the GstLAL pipeline as described in Sec. III is a close approximation to what will be used in the fourth observing run of the LVK Collaboration. However, since the conclusion of the MDC used in this paper, work has been ongoing and several areas for possible improvements have been identified. These changes in configuration and the corresponding improvements in performance are given in the Appendix.

ACKNOWLEDGMENTS

The authors are grateful for computational resources provided by the LIGO Laboratory and supported by National Science Foundation Grants No. PHY-0757058 and No. PHY-0823459. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the National Science Foundation (NSF) and operates under cooperative Agreement No. PHY-1764464. The authors gratefully acknowledge the Italian Istituto Nazionale di Fisica Nucleare (INFN), the French Centre National de la Recherche Scientifique (CNRS) and the Netherlands Organization for Scientific Research, for the construction and operation of the Virgo detector and the creation and support of the EGO consortium. The authors are grateful for computational resources provided by the Pennsylvania State University's Institute for Computational and Data Sciences (ICDS) and the University of Wisconsin Milwaukee Nemo and support by Grants No. NSF PHY-2011865, No. NSF OAC-2103662, No. NSF PHY-1626190, No. NSF PHY-1700765, and No. NSF PHY-2207728. This paper carries LIGO Document No. LIGO-P2300124. M. W. C. acknowledges support from the National Science Foundation with Grant No. PHY-2010970 and OAC- 2117997. R. E. is supported by the Natural Sciences & Engineering Research Council of Canada (NSERC). *Software:* Confluent Kafka [38], Grafana [39], GraceDB [18], GstLAL [40], GWCELERY [41], GW-LTS [42], IGWN-ALERT [43], InfluxDB [44], LIGO-SCALD [45].

APPENDIX: GstLAL PERFORMANCE WITH UPDATED O4 CONFIGURATION

In this section, we demonstrate the performance of the GstLAL pipeline using an updated configuration. The improved performance in this updated run are attributed to the following areas of development, which proceeded after the conclusion of the MDC presented in Sec. III.

The analytic SNR-ξ^2 signal model used in the likelihood ratio ranking statistic has now been tuned to use a more optimal allowed mismatch region, which is wider for the high SNR region of parameter space. This change is expected to improve the recovery efficiency for very loud signals.

The presence of non-Gaussian noise transients, known as glitches, in the strain data has long been a problem for gravitational wave searches. Integrated data quality (iDQ) is a machine-learning based algorithm used to assign probabilities of the presence of a glitch in a segment of strain data [46]. In the O3 offline analysis, iDQ was incorporated into the GstLAL ranking statistic as a means of reducing the significance of candidates found during particularly glitchy stretches of data. Although it is no longer used in the ranking statistic, it is now possible to use iDQ state information as a gate on the strain data, so that segments of data with a high glitch probability will be excluded from the filtering. This should mitigate the negative effects of non-Gaussian data and is expected to result in fewer retraction level candidates and an improved $\langle VT \rangle$. Although the iDQ gate was used in the MDC analysis presented here, the feature is still under development and is not planned for use in the O4 production configuration.

As mentioned in Sec. III C 3, the $p(\text{astro})$ performance in the BNS and NSBH region of the parameter space was sub-optimal in the MDC, with a significant amount of misclassification between the two source types, as well as NSBH misclassification as BBH. It is imperative that sources including a neutron star are not falsely classified as BBHs since this may discourage astronomers from following up these potentially electromagnetically bright signals. Since the conclusion of the MDC analysis presented in this paper, work has been ongoing toward improving the $p(\text{astro})$ source classification. This effort is discussed in more detail in [19].

Finally, as discussed in Sec. III C 4, we have made improvements to the event aggregation method to ensure that events with FAR below the public alert threshold will be uploaded with as little latency as possible. By adjusting the geometric cadence factor in the aggregator and removing the wait time for these low FAR events we expect a reduction in the latency of these uploads by up to several seconds.

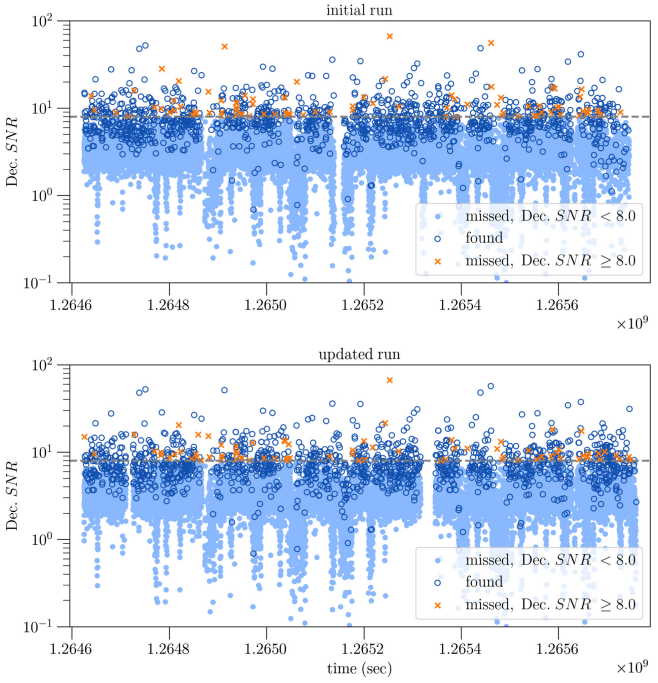

FIG. 18. Time series of injected decisive SNR for injections with component masses and spins within the O4 template bank in the initial run, March 15, 2023 to March 28, 2023 (upper panel) and updated run April 24, 2023 to May 7, 2023 (lower panel). Dark blue circles indicate injections that were recovered below a 2 per day FAR threshold. Orange crosses and light blue points indicate injections not recovered below this FAR threshold, where orange crosses are injections with decisive SNR > 8.0. Times on the horizontal axis are GPS times shifted to the original O3 epoch.

We reanalyze the final two weeks of the MDC data using the updated pipeline configuration in order to demonstrate the effect on the pipeline performance. The "initial run" presented in Sec. III was analyzed from March 15, 2023 to March 28, 2023 and the "updated run" was analyzed from April 24, 2023 to May 7, 2023. These runs correspond to identical stretches of O3 replay data with the same injections present in each.

Figure 18 shows the decisive SNR time-series for all injections within the two week span. There are several high SNR injections missed by the initial analysis (upper panel) but found in the updated analysis (lower panel). This

TABLE VII. Injection efficiencies as defined in Eq. (6) computed using a FAR threshold of 2.31×10^{-5} Hz to count "found" injections. The "initial run" is the same as presented in Sec. III and the "updated run" uses the configuration improvements described in this section. Source categories are defined in Table II and "ALL" combines injections from the three source categories.

	BNS	NSBH	BBH	ALL
Initial run	0.95	0.71	0.84	0.84
Updated run	0.93	0.72	0.90	0.86

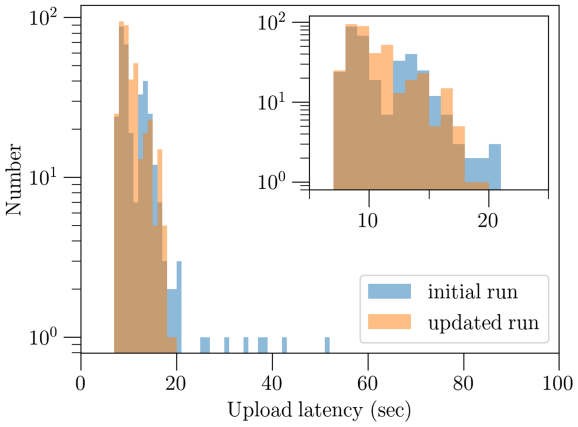

FIG. 19. Histogram of event upload latencies for the initial run (orange) and updated run (blue).

indicates an improvement as a result of the widened mismatch region of the SNR-ξ^2 signal model. Table VII compares the injection efficiencies between the two runs. For the BNS and NSBH regions the efficiency is comparable, while there is a 6% improvement in the BBH efficiency from 0.84 in the initial run to 0.90 in the updated run.

Figure 19 and Fig. 20 show the improvements in event upload latency between the two runs. The histogram in Fig. 19 shows that upload latencies for the updated run are shifted slightly lower overall with respect to the initial run, while the lower edge remains the same. This is expected as the improvements focused on lowering the latency of subsequent uploads by reducing the geometric cadence factor from 4 seconds to 2 seconds. Figure 20 shows a

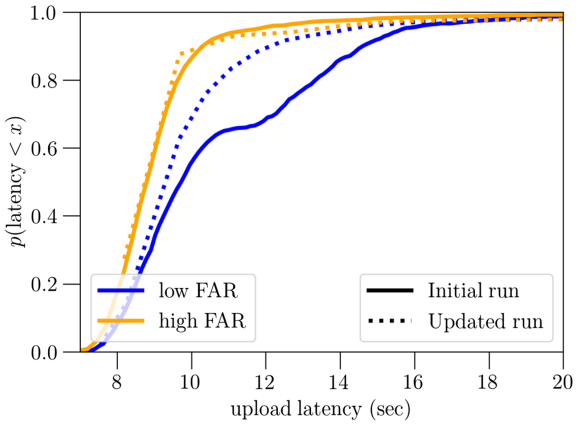

FIG. 20. Cumulative distribution of event upload latencies for the initial run (solid lines) and updated run (dashed lines). The orange (blue) curves show the distribution of latencies for the first event uploaded with FAR higher (lower) than the public alert threshold. The 90th percentile on the upload latency of high FAR candidates in the initial (updated) run is 10.30 (9.86) seconds. The 90th percentile on the upload latency of low FAR candidates in the initial (updated) run is 14.58 (12.04) seconds.

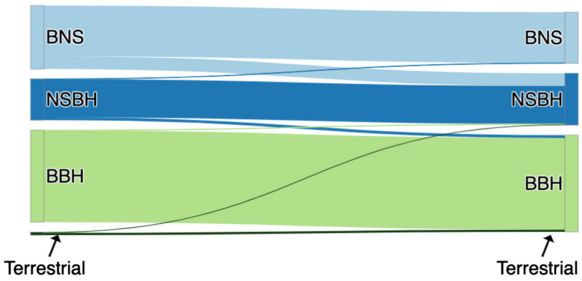

FIG. 21. Sankey diagram showing $p(\text{astro})$ classification of events uploaded to GraceDB during the updated run from April 24, 2023 to May 7, 2023. Events with an end time within ± 1 second of an injection are classified as either BNS, NSBH, or BBH using a neutron star mass boundary of $3.0M_{\odot}$. Events that do not correspond in time with an injection are all classified as terrestrial.

similar improvement in upload latency for events below the public alert threshold. We find a 2.54 second reduction in the 90th percentile on upload latency for events with FAR below the public alert threshold, from 14.58 seconds in the initial run to 12.04 seconds in the updated run. This indicates a significant improvement in the latency of alert-quality candidates.

Figure 21 is a Sankey diagram showing the $p(\text{astro})$ classification performance in the updated run. The initial results showed a significant amount of misclassification between BNS and NSBH as well as between NSBH and BBH injections. Here we find 20.2% of recovered BNS being classified as NSBH and only 6.83% of recovered NSBH being classified as BBH. While the updated run shows more confusion between BNS and NSBH, we see that the confusion between NSBH and BBH is significantly reduced. This is considered as an overall improvement, since the classification more accurately indicates whether a source may be electromagnetically bright, which is more likely the case for BNS and NSBH mergers as opposed to BBH. The reason for this improvement is as follows.

Throughout the $p(\text{astro})$ calculation, misclassification of sources are assumed to be the result of Gaussian noise fluctuations causing a gravitational wave signal to match better with templates that are further away in parameter space than the one that would recover it in the absence of noise [19,47]. In the initial run, this probability of template migration was estimated while making several simplistic assumptions some of which were later found to be inaccurate. In particular, the geometry of the template bank was modeled as a four-dimensional unit sphere and the observed SNR assumed to be a single Gaussian distributed real number [47]. In reality however, the geometry of the bank is sufficiently more complex than that of a spherical surface. Furthermore, due to the two-phase matched filtering process, the observed SNR is actually a complex number whose real and imaginary parts can be thought of as independent Gaussians. Marginalizing over the unknown signal phase leads to the absolute value of the SNR having a distribution that is broader than the Gaussian of [47].

Both of these ambiguities can cause an underestimation of misclassification by the model of Ref. [47], leading to the $p(\text{astro})$ calculation confidently classifying, e.g., an NSBH as a BBH even at low SNR. A detailed remodeling of the misclassification with realistic noise distributions and accurate representations of template bank geometry is beyond the scope of this paper and is part of an ongoing investigation.

Hence, to avoid strongly misclassifying potential electromagnetically bright sources, we instead construct a pessimistic model of noise induced template migration by assuming that the noise degrees of freedom are entirely contained in the signal manifold [19]. Implementing this model in the updated run leads to a significant improvement in the classification of simulated sources that have the potential to be electromagnetically bright. Further improvements in classification are expected to result upon changing the population model in Eq. (11) with the true distribution of source parameters [19].

[1] C. Messick *et al.*, Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data, Phys. Rev. D **95**, 042001 (2017).
[2] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. **116**, 061102 (2016).
[3] LIGO Scientific and Virgo Collaborations, GCN **21505** (2017), https://gcn.gsfc.nasa.gov/other/G298048.gcn3.
[4] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW170817: Observation of gravitational waves from a binary neutron star inspiral, Phys. Rev. Lett. **119**, 161101 (2017).
[5] R. Abbott *et al.* (LIGO Scientific, KAGRA, and Virgo Collaborations), Observation of gravitational waves from two neutron star–black hole coalescences, Astrophys. J. Lett. **915**, L5 (2021).
[6] R. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), GW190521: A binary black hole merger with a total mass of $150M_{\odot}$, Phys. Rev. Lett. **125**, 101102 (2020).
[7] L. Tsukada *et al.*, Improved ranking statistics of the GstLAL inspiral search for compact binary coalescences, Phys. Rev. D **108**, 043004 (2023).

[8] S. Sachdev *et al.*, The GstLAL search analysis methods for compact binary mergers in Advanced LIGO's second and Advanced Virgo's first observing runs, arXiv:1901.08580.

[9] K. Cannon *et al.*, GstLAL: A software framework for gravitational wave discovery, arXiv:2010.05082.

[10] K. Cannon *et al.*, Toward early-warning detection of gravitational waves from compact binary coalescence, Astrophys. J. **748**, 136 (2012).

[11] S. Sakon *et al.*, Template bank for compact binary mergers in the fourth observing run of Advanced LIGO, Advanced Virgo, and KAGRA, arXiv:2211.16674.

[12] S. Morisaki and V. Raymond, Rapid parameter estimation of gravitational waves from binary neutron star coalescence using focused reduced order quadrature, Phys. Rev. D **102**, 104020 (2020).

[13] R. Abbott *et al.* (KAGRA, Virgo, and LIGO Scientific Collaborations), Population of merging compact binaries inferred using gravitational waves through GWTC-3, Phys. Rev. X **13**, 011048 (2023).

[14] K.-L. Lo, S. Sachdev, K. Blackburn, and A. Weinstein, Searching for Gravitational Waves from the Coalescence of High Mass Black Hole Binaries (2016), https://dcc.ligo.org/LIGO-T1600261.

[15] K. Cannon, C. Hanna, and D. Keppel, Method to estimate the significance of coincident gravitational-wave observations from compact binary coalescence, Phys. Rev. D **88**, 024025 (2013).

[16] K. Cannon, C. Hanna, and J. Peoples, Likelihood-ratio ranking statistic for compact binary coalescence candidates with rate estimation, arXiv:1504.04632.

[17] P. Joshi, L. Tsukada, and C. Hanna, Method for removing signal contamination during significance estimation of a GstLAL analysis, Phys. Rev. D **108**, 084032 (2023).

[18] B. Moe, P. Brady, B. Stephens, E. Katsavounidis, R. Williams, and F. Zhang, GraceDB: A Gravitational Wave Candidate Event Database (2014), https://dcc.ligo.org/DocDB/0113/T1400365/005/.

[19] A. Ray *et al.*, When to point your telescopes: Gravitational wave trigger classification for real-time multi-messenger followup observations, arXiv:2306.07190.

[20] S. S. Chaudhary *et al.*, Low-latency alert products and their performance in anticipation of the fourth LIGO-Virgo-KAGRA observing run, arXiv:2308.04545.

[21] R. Abbott *et al.* (LIGO Scientific, Virgo, and KAGRA Collaborations), GWTC-3: Compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run, Phys. Rev. X **13**, 041039 (2023).

[22] LIGO Scientific and Virgo Collaborations, GCN **26641** (2020), https://gcn.gsfc.nasa.gov/other/S200106au.gcn3.

[23] LIGO Scientific and Virgo Collaborations, GCN **26665** (2020), https://gcn.gsfc.nasa.gov/other/S200108v.gcn3.

[24] LIGO Scientific and Virgo Collaborations, GCN **26785** (2020), https://gcn.gsfc.nasa.gov/other/S200116ah.gcn3.

[25] P. Schmidt, F. Ohme, and M. Hannam, Towards models of gravitational waveforms from generic binaries II: Modelling precession effects with a single effective precession parameter, Phys. Rev. D **91**, 024043 (2015).

[26] T. Dietrich, A. Samajdar, S. Khan, N. K. Johnson-McDaniel, R. Dudi, and W. Tichy, Improving the NRTidal model for binary neutron star systems, Phys. Rev. D **100**, 044003 (2019).

[27] LIGO Scientific, Virgo, and KAGRA Collaborations, LIGO/Virgo/KAGRA Public Alerts User Guide, https://emfollow.docs.ligo.org/userguide/ (2023).

[28] S. Fairhurst, R. Green, C. Hoy, M. Hannam, and A. Muir, Two-harmonic approximation for gravitational waveforms from precessing binaries, Phys. Rev. D **102**, 024055 (2020).

[29] R. J. E. Smith, G. Ashton, A. Vajpeyi, and C. Talbot, Massively parallel Bayesian inference for transient gravitational-wave astronomy, Mon. Not. R. Astron. Soc. **498**, 4492 (2020).

[30] G. Ashton *et al.*, BILBY: A user-friendly Bayesian inference library for gravitational-wave astronomy, Astrophys. J. Suppl. Ser. **241**, 27 (2019).

[31] D. W. Hogg, Distance measures in cosmology, arXiv:astro-ph/9905116.

[32] H.-Y. Chen, D. E. Holz, J. Miller, M. Evans, S. Vitale, and J. Creighton, Distance measures in gravitational-wave astrophysics and cosmology, Classical Quantum Gravity **38**, 055010 (2021).

[33] L. P. Singer and L. R. Price, Rapid Bayesian position reconstruction for gravitational-wave transients, arXiv:1508.03634.

[34] L. P. Singer *et al.*, Going the distance: Mapping host galaxies of LIGO and Virgo sources in three dimensions using local cosmography and targeted follow-up, Astrophys. J. Lett. **829**, L15 (2016).

[35] W. M. Farr, J. R. Gair, I. Mandel, and C. Cutler, Counting and confusion: Bayesian rate estimation with multiple populations, Phys. Rev. D **91**, 023005 (2015).

[36] S. J. Kapadia *et al.*, A self-consistent method to estimate the rate of compact binary coalescences with a Poisson mixture model, Classical Quantum Gravity **37**, 045007 (2020).

[37] E. E. Salpeter, The luminosity function and stellar evolution, Astrophys. J. **121**, 161 (1955).

[38] https://developer.confluent.io/get-started/python/.

[39] https://grafana.com/docs/.

[40] GstLAL, https://git.ligo.org/lscsoft/gstlal (2023).

[41] gwcelery, https://git.ligo.org/emfollow/gwcelery (2023).

[42] B. Ewing, gw-lts, https://git.ligo.org/rebecca.ewing/gw-lts (2023).

[43] A. Pace, igwn-alert, https://igwn-alert.readthedocs.io/ (2023).

[44] https://docs.influxdata.com/influxdb/v2.7/.

[45] P. Godwin, ligo-scald, https://git.ligo.org/gstlal-visualisation/ligo-scald (2023).

[46] R. Essick, P. Godwin, C. Hanna, L. Blackburn, and E. Katsavounidis, iDQ: Statistical inference of non-Gaussian noise with auxiliary degrees of freedom in gravitational-wave detectors, Mach. Learn. **2**, 015004 (2020).

[47] H. K. Y. Fong, From simulations to signals: Analyzing gravitational waves from compact binary coalescences, Ph.D. thesis, Toronto U., 2018.