Systematic errors in estimation of gravitational-wave candidate significance

C. Capano and T. Dent
Max Planck Institute for Gravitational Physics, Callinstraße 38, D-30167, Hannover, Germany

C. Hanna
The Pennsylvania State University, University Park, PA 16802, USA

Y.-M. Hu*
Max Planck Institute for Gravitational Physics, Callinstraße 38, D-30167, Hannover, Germany
School of Physics and Astronomy, Kelvin Building, University of Glasgow, Glasgow, G12 8QQ, UK and
Tsinghua University, Beijing, 100084, China

M. Hendry and C. Messenger
School of Physics and Astronomy, Kelvin Building, University of Glasgow, Glasgow, G12 8QQ, UK

J. Veitch
School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
(Dated: January 7, 2016)
(LIGO-P1500247-v1)

We investigate a critical issue in determining the statistical significance of candidate transient gravitational-wave events in a ground-based interferometer network. Given the presence of non-Gaussian noise artefacts in real data, the noise background must be estimated empirically from the data itself. However, the data also potentially contains signals, thus the background estimate may be overstated due to contributions from signals. It has been proposed to mitigate possible bias by removing single-detector data samples that pass a multi-detector consistency test (and thus form coincident events) from the background estimates. We conduct a high-statistics Mock Data Challenge to evaluate the effects of removing such samples, modelling a range of scenarios with plausible detector noise distributions and with a range of plausible foreground astrophysical signal rates; thus, we are able to calculate the exact false alarm probabilities of candidate events in the chosen noise distributions. We consider two different modes of selecting the samples used for background estimation: one where coincident samples are removed, and one where all samples are retained and used. Three slightly different algorithms for calculating the false alarm probability of candidate events are each deployed in these two modes. The three algorithms show good consistency with each other; however, discrepancies arise between the results obtained under the 'coincidence removal' and 'all samples' modes, for false alarm probabilities below a certain value. In most scenarios the median of the false alarm probability (FAP) estimator under the 'all samples' mode is consistent with the exact FAP. On the other hand the 'coincidence removal' mode is found to be unbiased for the mean of the estimated FAP over realisations. While the numerical values at which discrepancies become apparent are specific to the details of our numerical experiment, we believe that the qualitative differences in the behaviour of the median and mean of the FAP estimator have more general validity. On the basis of our study we suggest that the FAP of candidates for the first detection of gravitational waves should be estimated *without* removing single-detector samples that form coincidences.

I. INTRODUCTION

The global network of advanced gravitational wave (GW) detectors is poised to make its first direct detection [1–4]. The coalescence of binary systems containing neutron stars and/or black holes is the most likely source of transient gravitational waves [5] and the detection of such a compact binary coalescence (CBC) event would open the new window of GW astronomy [6]. The observation of CBC events would not only allow us to test General Relativity, but also help to give hint on the equation-of-state of neutron-stars (NSs). The study of populations of CBC events would help to deepen our understanding of stellar evolution for binary massive stars, especially the very late stages [7–9].

To support possible detections, it will be necessary to determining the confidence that the candidate signals are associated with astrophysical sources of GWs [10] rather than spurious noise events. Thus, a FAP estimate is produced in order to classify candidate events. Claims for the detection of previously undetected or unknown physical phenomena have been held up to high levels of scrutiny, e.g. the Higgs boson [11] and B-modes in the polarization of the cosmic microwave background (CMB) [12]. The same will be true for the direct detection of GWs: a high level of statistical confidence will be required as well as a thorough understanding of the accuracy and potential systematic bias of procedures to determine the FAP of a candidate event.

Existing CBC search pipelines (*e.g.* [13–15]) assess the FAP of a detection candidate by estimating its probability under the null hypothesis. The null hypothesis is that candidate events are caused by non-gravitational-wave processes acting

* yiming.hu@aei.mpg.de

on and within the interferometers. In order to determine this probability, we assign a test statistic value to every possible result of a given observation or experiment performed on the data: larger values indicate a higher deviation from expectations under the null hypothesis. We compute the FAP as the probability of obtaining a value of the test statistic equal to, or larger than, the one actually obtained in the experiment. The smaller this probability, the more significant is the candidate.

In general the detector output data streams are the sum of terms due to non-astrophysical processes, known as background noise, and of astrophysical GW signals, labelled foreground. If we were able to account for all contributions to noise using predictive and reliable physical models, we would then obtain an accurate calculation of the FAP of any observation. However, for real GW detectors, in addition to terms from known and well-modelled noise processes, the data contains large numbers of non-Gaussian noise transients ("glitches", see for example [16, 17]) whose sources are either unknown or not accurately modelled, and which have potentially large effects on searches for transient GWs. Even given all the information available from environmental and auxiliary monitor channels at the detectors, many such noise transients cannot be predicted with sufficient accuracy to account for their effects on search algorithms. Thus, for transient searches in real detector noise it is necessary to determine the background noise distributions empirically, i.e. directly from the strain data.

Using the data to empirically estimate the background has notable potential drawbacks. It is not possible to operate GW detectors so as to 'turn off' the astrophysical foreground and simply measure the background; if the detector is operational then it is always subject to both background and foreground sources. In addition, our knowledge of the background noise distribution is limited by the finite amount of data available. This limitation applies especially in the interesting region of low event probability under the noise hypothesis, corresponding to especially rare noise events.

CBC signals are expected to be well modelled by the predictions of Einstein's general relativity [18]. The detector data are cross-correlated (or matched filtered) [19] against a bank of template CBC waveforms, resulting in an signal-to-noise ratio (SNR) time series for each template [20]. If this time series crosses a predetermined threshold, the peak value and the time of the peak are recorded as a *trigger*. Since GWs propagate at the speed of light (see e.g. [6]), the arrival times of signals will differ between detectors by $\simeq$40 ms or less, i.e. Earth's light crossing time. Differences in arrival times are governed by the direction of the source on the sky in relation to the geographical detector locations. We are thus able to immediately eliminate the great majority of background noise by rejecting any triggers which are not coincident in two or more detectors within a predefined coincidence time window given by the maximum light travel time plus trigger timing errors due to noise. Only triggers coincident in multiple detectors with consistent physical parameters such as binary component masses are considered as candidate detections. In order to make a confident detection claim one needs to estimate the rarity, i.e., the *FAP* of an event, and claim detection only when the probability of an equally loud event being caused by noise is below a chosen threshold.

The standard approach used in GW data analysis for estimating the statistical properties of the background is via analysis of time-shifted data, known as "time slides" [21, 22]. This method exploits the coincidence requirement of foreground events by time-shifting the data from one detector relative to another. Such a shift, if larger than the coincidence window, would prevent a zero-lag signal remaining untouched in a time-shifted analysis. Therefore, from a single time-shifted analysis ("time slide") the output coincident events should represent one realisation of the background distribution of coincident events, given the sets of single-detector triggers, assuming that the background distributions are not correlated between detectors. By performing many time slide analyses with different relative time shifts we may accumulate instances of background coincidences and thus estimate their rate and distribution.

The time-slides approach has been an invaluable tool in the analysis and classification of candidate GW events in the initial detector era. Initial LIGO made no detections, however note that in 2010, the LSC (LIGO Scientific Collaboration) and Virgo collaborations (LVC) performed a blind injection challenge [10] to check the robustness and confidence of the pipeline. A signal was injected in "hardware" (by actuating the test masses) in the global network of interferometers and analysed by the collaboration knowing only that there was the possibility of such an artificial event. The blind injection was recovered by the templated binary inspiral search with a high significance [26]; however, the blind injection exercise highlighted potential issues with the use of time-shifted analysis to estimate backgrounds in the presence of astrophysical (or simulated) signals.

Simply time-shifting detector outputs with respect to each other does not eliminate the possibility of coincident events resulting from foreground (signal) triggers from one detector passing the coincidence test with random noise triggers in another detector. Thus, the ensemble of samples generated by time-shifted analysis may be "contaminated" by the presence of foreground events in single-detector data, compared to the ensemble that would result from the noise background alone. The distribution of events caused by astrophysical signals is generally quite different from that of noise: it is expected to have a longer tail towards high values of SNR (or other event ranking test-statistic used in search algorithms). Thus, depending on the rate of signals and on the realisation of the signal process obtained in any given experiment, such contamination could considerably bias the estimated background. If the estimated background rate is raised by the presence of signals, the FAP of coincident search events (in non-time-shifted or "zero-lag" data) may be overestimated, implying a conservative bias in the estimated FAP. The expected number of false detection claims will not increase due to the presence of signals in time-slide analyses, however some signals may fail to be detected due to an elevated background estimate. [1]

[1] The reader may ask what a "false detection claim" means if signals are

Besides the standard time-shift approach, other background estimation techniques have been developed [15, 23, 24]. All are variants on one key concept: events that are not coincident within the physically allowed time window cannot be from the same foreground source. All therefore use non-coincident triggers as proxies for the distribution of background events due to detector noise. Differences between methods occur in implementation, and fall into two categories. In the standard scheme, many time slides are applied to a set of triggers from a multi-detector observation and all resultant coincidences are retained. We label this the 'all samples' mode of operation. Concern about the potential conservative bias of including foreground GW events of astrophysical original in the time-shifted distribution motivates a modification to this procedure: one can instead choose to first identify all coincidences between different detectors in zero-lag and exclude them from the subsequent time-slide analysis. We label this the 'coincidence removal' mode of operation.

In this paper we describe the results of an mock data challenge (MDC) in which participants applied three different background estimation algorithms, each applied to the MDC data in the two modes of operation described above: the 'coincidence removal' and 'all samples' modes. Since some aspects of the MDC data generation were not fully realistic, the algorithms were simplified compared to their use on real data; two were based on methods currently in use in the LVC, while the third introduces a new approach. The MDC consisted of simulated realisations of single-detector triggers – maxima of matched filter SNR ρ above a fixed threshold value chosen as $\rho_0 = 5.5$. The trigger SNR values were generated according to analytically modelled background and foreground distributions unknown to the MDC participants. The background distributions spanned a range of complexity including a realistic Initial detector distribution [25]. The foreground triggers were chosen to model one of the most likely first sources, binary neutron stars, with an expected detection rate from zero to the maximum expected for the first Advanced detector science runs [5]. Participants were asked to apply their algorithms to these datasets and report, in both modes, the estimated FAP of the loudest event found in each realisation. The MDC used analytic formulae to generate the background distributions; although we will not, in reality, have access to such formulae, they allow us to compute an "exact" FAP semi-analytically. We use these exact values, alongside other figures of merit, as a benchmark for comparing the results obtained in the 'coincidence removal' and 'all samples' modes, both within each background estimation algorithm and between the different algorithms.

In the following section of this paper we provide details of the MDC, describing the background and foreground distributions, the data generating procedure, and the "exact" FAP calculation. In Sec. III we then describe the different methods of background estimation used in this MDC. We report the results of the challenge in Sec. IV, comparing and contrasting the results obtained from each of the algorithms in each of their 2 modes of operation. Finally in Sec. V we provide a summary and present our conclusions.

present. This refers to cases where the search result contains both foreground events, and background events with comparable or higher ranking statistic values. In particular, if the loudest event in the search is due to background a detection claim would be misleading.

II. THE MOCK DATA CHALLENGE

Our MDC is designed to resolve the question of whether to remove ('coincidence removal') or not to remove ('all samples') coincident zero-lag events from the data used to estimate FAP. Simulated single detector triggers are generated from known background and foreground distributions: challenge participants are then given lists of trigger arrival times and single detector SNRs. The ranking statistic used for each coincident trigger is the combined multi-detector SNR. The calculation of FAP could be carried out with any appropriate ranking statistic. When applied to real data, different pipelines may define their ranking statistics in slightly different ways; in order to make a fair comparison in the challenge we use the multi-detector SNRs as a simplified ranking statistic. The cumulative distribution functions (CDFs) of the SNR in each detector are described analytically and hidden from the challenge participants. With this analytic form, the challenge designers are able to compute the exact FAP at the ranking statistic value of the loudest coincident event in each realisation.

The challenge consisted of 14 independent experiments, each with a different foreground and background distribution for which 10^5 observational realisations are generated (see section A). This number is large enough to reach the interesting statistics region while computationally feasible. Each realisation contained, on average, $\sim 10^4$ single-detector triggers in each of two detectors. A realisation should be considered analogous to a single GW observing run. Participants were asked to estimate the FAP of the loudest coincident event in each of the 10^5 realisations for each experiment. The 14 experiments cover a variety of background distributions; the foregrounds differ in that the astrophysical rate ranges between zero, i.e., no events, and a relatively high value corresponding to ~ 3 detections per realisation. Analysis of the differences between the estimated and exact FAP values enable us to quantify the effect of removing coincident zero-lag triggers in the background estimate, as opposed to retaining all samples. It also allows us to directly compare implementations between different algorithms in each mode of operation. Finally, it allows us to quantify the accuracy and limiting precision of our FAP estimates and thus their uncertainties.

We divide our simulations into three groups according to astrophysical rate, and independently into three groups according to background distribution complexity ("simple", "realistic" and "extreme"). To this nine combinations we have appended an additional four simulations, three of which have exactly the same background distributions as three of the original nine but contain no signals. The final simulation contains a background distribution with a deliberately extended tail such that the generation of particularly loud background triggers is possible. The primary properties of each experi-

	Background property			
Foreground rate	simple	realistic	extreme	ext. tail
zero	1,3	12	14	-
low	-	10	2	7
medium	9,13	8	6	-
high	5	11	4	-

TABLE I: The classification of each experiment in the MDC in terms of background complexity and astrophysical foreground rate. See main text for definitions.

ment are given in Table I and details are listed in Tables II, III and IV. . A *low* foreground rate corresponds to <0.01 expected coincident signal triggers per realisation, a *medium* rate corresponds to 0.01–1 coincidences per realisation, and *high* rate corresponds to >1 per realisation. We do not consider foreground rates above ∼3 coincidences per realisation since we are motivated by FAP estimation for the first advanced era GW detections.

A. Modelling the detector noise backgrounds

The CDF of the background single-detector SNR triggers is modelled as the exponential of a piecewise polynomial function in the SNR ρ via

$$C(\rho) = \begin{cases} 1 - \exp\left(\sum_{i=0}^{6} a_i \left(\rho - \rho_{\rm th}\right)^i\right), & \text{for } \rho \leq \rho_{\rm sp} \\ 1 - C_{\rm sp} \exp\left(b\left(\rho - \rho_{\rm sp}\right)\right), & \text{for } \rho > \rho_{\rm sp}, \end{cases} \quad (1)$$

where the trigger generation threshold is set as $\rho_{\rm th} = 5.5$. The polynomial coefficients a_i must satisfy the constraint that the CDF remains monotonic in ρ; additionally, a_0 is determined by the constraint that the CDF should range between 0 and 1. We define the CDF differently in the regions below and above a switching-point $\rho_{\rm sp}$ value in order to satisfy the constraints on the CDF model, such that the CDF and its derivative with respect to ρ are continuous at the switching point. Hence, a choice of $C_{\rm sp}$ determines the values of $\rho_{\rm sp}$ and b. Details of the background distribution parameters chosen for each simulation can be found in Appendix A; here we describe the broader properties of the chosen distributions.

In cases with a "simple" background, the coefficients of our model (Eq. 1) are all zero with the exception of a_0 and $\rho_{\rm sp} = \infty$. The CDF then follows the simple form $C = 1 - \exp(-a_0(\rho - \rho_{\rm th}))$ for the single-detector SNR. A "realistic" background is modelled by basing our analytic CDF model on distributions of existing GW trigger data [26]. The "extreme" backgrounds attempt to model distributions containing multiple independent populations of detector noise artefacts resulting in CDFs that exhibit large variations in their gradients as a function of SNR. We give examples of each type of background distribution in Fig. 1. The single experiment described as containing an "extended tail" is similar to the extreme cases in the sense that its gradient varies substantially as a function of SNR. However, this variation occurs at much smaller values of $1 - C$, thus it is rare that realisations have events generated from the "tail". This rarity and shallowness of the tail are designed to mimic the behaviour of an astrophysical foreground (with the exception of being coincident between detectors).

The trigger time of a background event is a random variable generated from a uniform distribution spanning the the length of an observation realisation. The number of such triggers within a realisation is drawn from a Poisson distribution with parameter λ_j, the expected number of triggers in the j'th detector. The two detectors are treated independently and for events to be labelled coincident, their trigger times must satisfy

$$|t_1 - t_2| \leq \delta t, \quad (2)$$

where t_1 and t_2 are the times associated with a trigger from the first and second detectors respectively and δt is our allowed coincidence window. We can therefore estimate the expected number of coincident events n within a single realisation as

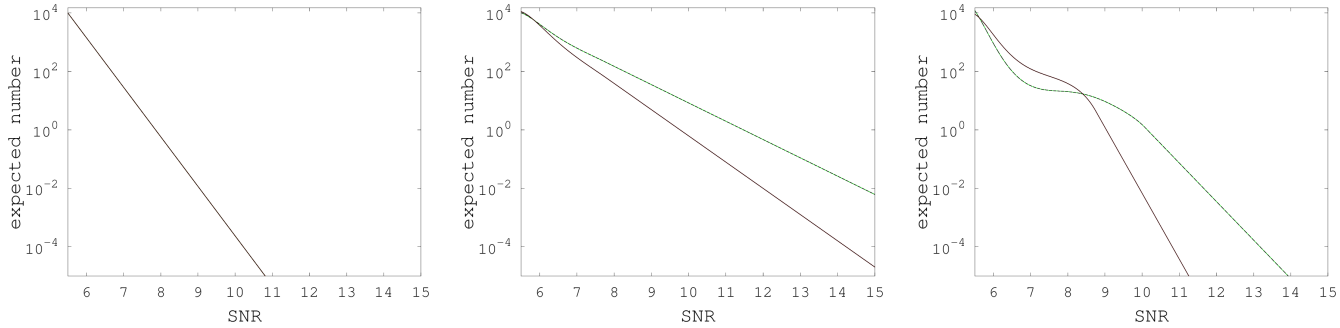
$$n = \frac{2\lambda_1 \lambda_2 \delta t}{T}, \quad (3)$$

where T is the total time of the observation and we have assumed that $\lambda_j \delta t / T \ll 1$ in both detectors. In order to generate a large enough number of background triggers to adequately model a GW observation, we use $\lambda_1 \sim \lambda_2 \sim 10^4$. This choice is also motivated by the expected dependence of the uncertainty in estimation of FAP on the numbers of triggers, and the computational cost to challenge participants. We set the coincidence window $\delta t = 50\,\text{ms}$ to broadly model realistic values and in order to obtain a desired ∼10 coincidences per realisation the total observation time is set to $T = 10^6$ s.

Note, however, that the MDC does not model some aspects of a real search for CBC signals, most notably the need for many thousands of different template waveforms to optimise the sensitivity to different possible source parameters. The multiplicity of templates has various implications for estimating FAP. The numbers of single-detector background triggers will increase, but the probability that any given pair of triggers will form a coincident noise event will drop since random noise triggers are unlikely to have consistent physical source parameters across different detectors. The complexity and computational load of modelling sets of single-detector triggers would be considerably increased, since the times and SNRs of triggers will be nontrivially correlated between templates with similar waveforms.

B. Modeling an astrophysical foreground

In the majority of experiments (10 of the 14) an astrophysical foreground was simulated. We model the astrophysical signal distribution as originating from the inspiral of equal mass $1.4 - 1.4 M_\odot$ binary neutron stars having a uniform distribution in volume and in time. For each source the binary orientation is selected from an isotropic distribution (uniform in the cosine of the inclination angle ι), the polarisation angle ψ is uniform on the range $[0, 2\pi)$ and the sky position is distributed isotropically on the sphere parametrised by



(a) A "simple" example background (experiment 3).

(b) A "realistic" example background (experiment 12).

(c) An "extreme" example background (experiment 14).

FIG. 1: Examples of different background distributions used in the MDC. In each of the three examples we show the complementary CDF ($1 - C$) versus the single-detector SNR for each detector. There were no foreground distributions present in these experiments.

right ascension α (with range $[0, 2\pi)$) and declination δ (range $[-\pi/2, \pi/2)$). Given a set of these parameters we can compute the optimal single-detector SNR $\rho_{\rm opt}$ as

$$\rho^2_{\rm opt} = 4 \int_{f_{\rm min}}^{f_{\rm ISCO}} \frac{|\tilde{h}(f)|^2}{S_{\rm n}(f)} df \tag{4}$$

where the lower and upper integration limits are selected as 10 Hz and as the innermost stable circular orbit frequency = 1570 Hz for our choice of total system mass. The detector noise spectral density $S_{\rm n}(f)$ corresponds to the advanced LIGO design [2], and the frequency domain signal in the stationary phase approximation is given by

$$\tilde{h}(f) = \frac{Q(\{\theta\})\mathcal{M}^{5/6}}{d} \sqrt{\frac{5}{24}} \pi^{-2/3} f^{-7/6} e^{i\Psi(f)}. \tag{5}$$

Here the function $Q(\{\theta\})$, where $\{\theta\} = (\alpha, \delta, \psi, \cos\iota)$, describes the antenna response of the detector; d is the distance to the source, and $\mathcal{M}$ is the "chirp mass" of the system given by $\mathcal{M} = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$. Since we consider that such signals, if present in the data, are recovered with exactly matching templates, the phase term $\Psi(f)$ does not influence the optimal SNR of Eq. 4. Hence the square of the observed (or matched filter) SNR ρ is drawn from a non-central χ^2 distribution with 2 degrees of freedom and non-centrality parameter equal to $\rho^2_{\rm opt}$.

We generate foreground events within a sphere of radius 1350 Mpc such that an optimally oriented event at the boundary has <0.3% probability of producing a trigger with SNR > $\rho_{\rm th} = 5.5$. Each event is given a random location (uniform in volume) and orientation from which we calculate the corresponding optimal SNR and relative detector arrival times. The matched filter SNR is modelled as a draw from the non-central chi-squared distribution. For each detector, if the matched filter SNR is larger than $\rho_{\rm th}$, independently of the other detector, it is recorded as a single detector trigger. The arrival time in the first detector (chosen as the LIGO Hanford interferometer) is randomly selected uniformly within the observation time and the corresponding time in the second detector (the LIGO Livingston interferometer) is set by the arrival time difference defined by the source sky position. We do not model statistical uncertainty in the arrival time measurements, hence when a foreground event produces a trigger in both detectors the trigger times will necessarily lie within the time window and will generate a coincident event.

C. The definition of false alarm probability (FAP) for the MDC

In order to define the FAP for any given realisation of an experiment we require a ranking statistic which is a function of the coincident triggers within a realisation. In this MDC the chosen ranking statistic was the combined SNR of coincident events, defined as

$$\rho^2 = \rho_1^2 + \rho_2^2, \tag{6}$$

where $\rho_{1,2}$ are the SNRs of the single-detector triggers forming the coincident event. Challenge participants were required to estimate the FAP of the "loudest" coincident event within each realisation, i.e. the event having the highest ρ value, independent of its unknown origin (background or foreground). The FAP of an outcome defined by a loudest event ρ^* is the probability of obtaining at least one background event having $\rho \geq \rho^*$ within a single realisation. Given that single-detector background distributions fall off with increasing $\rho_{1,2}$, the louder a coincident event is, the less likely it is for a comparable or larger ρ value to be generated by noise, and the smaller the FAP.

With access to the analytical description of the backgrounds from both detectors we may compute the single trial FAP $\mathcal{F}_1$

as

$$1-\mathcal{F}_1(\rho) = \int_{\rho_{\rm th}}^{\sqrt{\rho^2-\rho_{\rm th}^2}} d\rho_1 \int_{\rho_{\rm th}}^{\sqrt{\rho^2-\rho_1^2}} d\rho_2\, p_1(\rho_1)p_2(\rho_2),$$
$$= \int_{\rho_{\rm th}}^{\sqrt{\rho^2-\rho_{\rm th}^2}} d\rho_1 p_1(\rho_1)\, C_2\left(\sqrt{\rho^2-\rho_1^2}\right), \qquad (7)$$

where $p_1(\rho_1)$ and $p_2(\rho_2)$ are the probability distribution functions (PDFs) of the background distributions (obtained by differentiating the corresponding CDFs with respect to ρ_j), and $C_2(\rho_2)$ is the CDF for the second detector.

To account for the fact that we are interested in the "loudest" coincident event within each realisation we must perform an additional marginalisation over the unknown number of such coincident events. To do this we model the actual number of coincidences as drawn from a Poisson distribution with known mean n (values for the different MDC experiments are given in Table IV). The FAP of the "loudest" event is modelled as the probability of obtaining one or more coincident events with a combined SNR $\geq \rho$ and is given by

$$\mathcal{F}(\rho) = \sum_{j=0}^{\infty}\left(1-(1-\mathcal{F}_1(\rho))^j\right)\frac{n^j e^{-n}}{j!}. \qquad (8)$$

Challenge participants only had access to the trigger $\rho_{1,2}$ values and trigger times in each realisation and were not given the distributions from which they were drawn. Estimates of the loudest coincident event FAP $\mathcal{F}$ from all participants will be compared to the "exact" values computed according to Eq. 8.

D. The expected error on estimated false alarm probability (FAP)

Inferring the FAP, as defined above, from a finite sample of data will have associated uncertainty, i.e., the computed values will be approximate. Methods to estimate the FAP at a given combined SNR value $\mathcal{F}(\rho)$ involve counting the number of noise events $N(\rho)$ above that value:

$$N(\rho) = \iint_{\rho\geq\rho_{\rm th}} n_1(\rho_1)\, n_2(\rho_2)\, d\rho_1 d\rho_2$$
$$= \Lambda_1\Lambda_2 - \iint_{\rho<\rho_{\rm th}} n_1(\rho_1)\, n_2(\rho_2)\, d\rho_1 d\rho_2, \qquad (9)$$

where $n_i(\rho_i)$ is the number density of background triggers from detector i and Λ_i is the total number of background triggers from detector i. The region of integration is bounded by a threshold on the coincident SNR statistic of Eq. 6, though in general one may choose other functional forms for $\rho(\rho_1,\rho_2,\dots)$.

It is possible to compute (either analytically or numerically) the error on $N(\rho)$ given any functional form for ρ. However, we seek a simple "rule of thumb" as a general approximation. We replace the region $\rho< \rho_{\rm th}$ with an equivalent hyper-cuboid with lengths ρ_i^*, such that for an event to be counted towards the FAP it must have a SNR greater than ρ_i^* in either detectors. In this case, the number of louder triggers as a function of ρ can be approximated by

$$N(\rho) \approx \Lambda_1\Lambda_2 - \int_0^{\rho_1^*} d\rho_1 \int_0^{\rho_2^*} d\rho_2\, n_1(\rho_1)n_2(\rho_2)$$
$$\approx \Lambda_1\Lambda_2 - N_1'(\rho_1^*)N_2'(\rho_2^*), \qquad (10)$$

where

$$N_i'(\rho_i^*) \equiv \int_0^{\rho_i^*} n_i(\rho_i)\, d\rho_i, \qquad (11)$$

is the cumulative number of triggers from detector i. We then define the inferred FAP as

$$\mathcal{F}(\rho) \approx \frac{N(\rho)}{\Lambda_1\Lambda_2}$$
$$\approx 1 - \frac{N_1'(\rho_1^*)N_2'(\rho_2^*)}{\Lambda_1\Lambda_2} \qquad (12)$$

We wish to characterise the error in $\mathcal{F}(\rho)$ given the error in the number of triggers counted above ρ_i in each detector. We expect that this error will increase when fewer triggers are available to estimate the single detector counts. Transforming Eq. 12 to use the counts above a threshold ρ_i^* in each detector via $N_i(\rho_i) \equiv \Lambda_i - N_i'(\rho_i)$, we have

$$\mathcal{F}(\rho) \approx 1 - \frac{(\Lambda_1 - N_1(\rho_1^*))(\Lambda_2 - N_2(\rho_2^*))}{\Lambda_1\Lambda_2}. \qquad (13)$$

Assuming a negligible error on the total count of triggers in each detector Λ_i, we can then write

$$\sigma^2_{\mathcal{F}(\rho)} \approx \sum_i \left(\frac{\partial\mathcal{F}(\rho)}{\partial N_i(\rho_i^*)}\right)^2 \sigma^2_{N_i(\rho_i^*)}. \qquad (14)$$

Taking the distribution of counts $N_i(\rho)$ to be Poisson, we have standard errors $\sigma^2_{N_i(\rho_i^*)} = N_i(\rho_i^*)$; for the two-detector case we then find

$$\sigma^2_{\mathcal{F}(\rho)} \approx \frac{(\Lambda_2 - N_2(\rho_2^*))^2 N_1(\rho_1^*) + (\Lambda_1 - N_1(\rho_1^*))^2 N_2(\rho_2^*)}{\Lambda_1^2\Lambda_2^2}, \qquad (15)$$

hence the fractional error is

$$\frac{\sigma_{\mathcal{F}(\rho)}}{\mathcal{F}(\rho)} \approx \frac{\sqrt{\frac{(\Lambda_2-N_2(\rho_2^*))^2}{N_1(\rho_1^*)N_2{}^2(\rho_2^*)} + \frac{(\Lambda_1-N_1(\rho_1^*))^2}{N_2(\rho_2^*)N_1{}^2(\rho_1^*)}}}{\frac{\Lambda_2}{N_2(\rho_2^*)} + \frac{\Lambda_1}{N_1(\rho_1^*)} - 1}. \qquad (16)$$

In the limit of low FAPs, $N_1'(\rho_1^*) \ll \Lambda_1$ and $N_2'(\rho_1^*) \ll \Lambda_2$, our expression simplifies to

$$\frac{\sigma_{\mathcal{F}(\rho)}}{\mathcal{F}(\rho)} \approx \frac{\sqrt{\left(\frac{\Lambda_2}{N_2(\rho_2^*)}\right)^2 \frac{1}{N_1(\rho_1^*)} + \left(\frac{\Lambda_1}{N_1(\rho_1^*)}\right)^2 \frac{1}{N_2(\rho_2^*)}}}{\frac{\Lambda_2}{N_2(\rho_2^*)} + \frac{\Lambda_1}{N_1(\rho_1^*)}}. \qquad (17)$$

Now we consider two limiting cases. First, when the distribution of counts is similar in each detector such that $\Lambda_1 \approx \Lambda_2$ and $N_1(\rho_1^*) \approx N_2(\rho_2^*)$, we have

$$\frac{\sigma_{\mathcal{F}(\rho)}}{\mathcal{F}(\rho)} \approx \sqrt{\frac{1}{2N_1(\rho_1^*)}}. \tag{18}$$

Second, when we are probing much further into the "tail" of the distribution of one detector, e.g., $\Lambda_1/N_1(\rho_1^*) \gg \Lambda_2/N_2(\rho_2^*)$, we have

$$\frac{\sigma_{\mathcal{F}(\rho)}}{\mathcal{F}(\rho)} \approx \sqrt{\frac{1}{N_1(\rho_1^*)}}. \tag{19}$$

In both cases the fractional error is related to the inverse of the single-detector counts, not the combined counts $N(\rho)$ as one might naively expect. A similar contribution to the uncertainty in false alarm rate estimation, due to Poisson counting errors for single-detector events, was found in [27]. We note that a number of approximations were made to derive our "rule of thumb", though we show the level of agreement between this estimate and the results of the MDC analysis in Section IV.

III. BACKGROUND ESTIMATION ALGORITHMS

A. Standard offline analysis: false alarm probability (FAP) via inverse false alarm rate (IFAR)

We now describe the time slide method implemented in the all-sky LIGO-Virgo search pipeline for CBC [21, 22, 26, 28, 29] and indicate how the method has been adapted for the simplified high-statistics study presented in this paper.

Each coincident event obtained in the search is characterized by its estimated coalescence time and binary component masses, and in addition by the values of SNR ρ and the signal-based chi-squared test χ^2 [30] in each detector, which together are intended to achieve separation of signals from non-Gaussian noise transients. The event ranking statistic used, ρ_c, is the quadrature sum of re-weighted SNRs $\hat{\rho}_i(\rho_i, \chi_i^2)$ [21, 26] over participating detectors i.[2] Exactly the same coincidence test is performed in the time-shifted analyses as in the search, resulting in a set of values $\{\rho_{c,b}\}$ from time-shifted events, considered as background samples.[3]

With the search performed over a duration T of two- or more-detector coincident data, and time-shifted analyses covering a total duration $T_b \equiv sT$, defining a background multiplier s, the *estimated false alarm rate (FAR)* of a candidate event having ranking statistic ρ_c^* is calculated as the observed rate of louder background events over the time-shifted analyses:

$$\mathrm{FAR}(\rho_c^*) \equiv \frac{\sum_{\{\rho_{c,b}\}} \Theta(\rho_{c,b} - \rho_c^*)}{T_b} \equiv \frac{n_b(\rho_c^*)}{T_b}, \tag{20}$$

where $\Theta(x) = 1$ if $x > 0$ and 0 otherwise. n_b is the number of events louder than ρ_c^*. Note that the FAR may equal zero for a high enough threshold ρ_c^*.

The test statistic used to determine FAP is inverse FAR (IFAR), i.e. 1/FAR; thus a false alarm corresponds to obtaining a given value of n_b/T_b or lower under the null hypothesis.

Consider ranking the ρ_c value of a single search event relative to a total number N_b of time-shifted background event values. Under the null hypothesis every ranking position is equally probable, thus the probability of obtaining a count n_b or smaller of background events is $P_0(n_b \text{ or less}|1) = (1 + n_b)/N_b$. Since n_b decreases monotonically with increasing ρ_c, if a search event has a ρ_c value equal to or greater than a given threshold ρ_c^*, the number of louder background events $n_b(\rho_c)$ must be equal to or less than $n_b(\rho_c^*)$. Thus we may also write[4]

$$P_0(\rho_c \geq \rho_c^*|1) \leq \frac{1 + n_b(\rho_c^*)}{N_b}. \tag{21}$$

Then, if there are k such search events due to noise, the probability of at least one being a false alarm above the threshold ρ_c (implying an estimated IFAR as large as T_b/n_b) is

$$P_0(1 \text{ or more } \geq \rho_c^*|k) = 1 - \left(1 - P_0(\rho_c \geq \rho_c^*|1)\right)^k. \tag{22}$$

The implementation is simplified by considering the largest possible number of time-shifted analyses, such that a pair of single-detector triggers coincident in one analysis cannot be coincident for any other time shift. This implies that the relative time shifts are multiples of $2\delta t$, and the maximum number of time-shifted analyses is $s = T/(2\delta t) - 1$. The resulting time-shifted coincidences are then simply all possible combinations of the single-detector triggers, minus those coincident in the search ("zero-lag"), since every trigger in detector 1 will be coincident with every trigger in detector 2 either in zero-lag or for some time shift. Identifying ρ_c^* with the loudest coincident search event value $\rho_{c,\max}$ we have

$$1 + n_b(\rho_{c,\max}) = 1 + \sum_i \sum_j \Theta(\rho_{1,i}^2 + \rho_{2,j}^2 - \rho_{c,\max}^2), \tag{23}$$

where the sums run over all single-detector triggers $\{\rho_{1i}\}$, $\{\rho_{2j}\}$, $i = 1 \ldots \Lambda_1$, $j = 1 \ldots \Lambda_2$.

So far we have worked with specific values for the number of search events due to noise k and time-shifted background events N_b, however these are not known in advance

[2] In real data the search may be divided into event bins determined by the component masses and participating interferometers [22, 28]; however the present study does not attempt to simulate these complications.

[3] In real data an additional time clustering step is performed on the search and on each time-shifted analysis in order to reduce the number of strongly-correlated coincident events separated by short intervals ($\lesssim 1$ s) resulting from the multiplicity of filter templates. In this study, however, single-detector events are already uncorrelated by construction thus such clustering is not performed.

[4] Note that a statistic value slightly *below* ρ_c^* may also map to the same number of louder background events $n_b(\rho_c^*)$, thus the condition $\rho_c \geq \rho_c^*$ is more restrictive than $n_b \leq n_b(\rho_c^*)$.

and should be treated as stochastic (see also [**?**]). We assume that N_b is large enough that we can neglect its fluctuations, but we model k as a Poisson process with mean rate $\mu = \langle N_b \rangle / s \simeq N_b/s$. (In fact we know that $k + N_b = \Lambda_1\Lambda_2$, the product of single-detector trigger counts, thus we assume that $\Lambda_{1,2}$ are large Poisson-distributed numbers and $s \gg 1$.) We then marginalize over k using the Poisson prior:

$$p(k|\mu) = \frac{\mu^k e^{-\mu}}{k!}.$$

After marginalization the dependence on μ vanishes to obtain

$$\mathcal{F}(\rho_c^*) = p(\rho_{c,\max}) \simeq 1 - \exp\left(-\frac{1 + n_b(\rho_{c,\max})}{s}\right). \tag{24}$$

Thus, false alarms louder than $\rho_{c,\max}$ arising by random coincidence from our sets of single-detector triggers are approximated by a Poisson process with expected number $(2\delta t/T)(1 + n_b(\rho_{c,\max}))$. For this MDC, the values of the coincidence window and analysis time chosen imply $s \simeq 10^7$, giving a limit $p \gtrsim 10^{-7}$ to any FAP estimate. We have verified that the p-value of Eq. (24) is distributed uniformly on $(0, 1]$ for MDC data sets containing uncorrelated noise triggers.

So far we have considered the case where all single-detector triggers are kept in constructing the background values. To implement the case of removing zero-lag coincident triggers, we simply exclude these from the sums over pairs of triggers on the RHS of Eq. (23).

B. All possible coincidences (APC) approach

The all possible coincidences (APC) approach is described in detail in [24]. Here we provide a brief synopsis.

To estimate the FAP of zero-lag triggers, we first find the probability of getting a trigger from the background distribution with combined SNR $\geq \rho$ in a single draw. When not removing zero-lag triggers from the background estimate, this is:

$$\overline{\mathcal{F}}(\rho) = P_0(\rho|1) = \frac{n_b(\rho)}{\Lambda_1\Lambda_2 - k}. \tag{25}$$

Both background and zero-lag triggers are constructed by finding every possible combination of triggers in detector 1 and detector 2. Background triggers are then any coincidence such that $\Delta t = |t_1 - t_2| > \delta t$, while zero-lag triggers are those with $\Delta t \leq \delta t$. These can be found by adding the matrices $Z = X + Y$, where $X_{ij} = \rho^2_{1,i}\, \forall\, j$ and $Y_{ij} = \rho^2_{2,j}\, \forall i$. The elements of Z are thus the ρ^2 of all possible combination of triggers.

When removing zero-lag triggers from the background, the single detector triggers that form the zero-lags are removed from the X and Y matrices prior to finding Z. This changes the denominator in Eq. (25) to $(\Lambda_1 - k)(\Lambda_2 - k)$. However, if $\Lambda_1, \Lambda_2 \gg k$, then the denominator is approximately $\Lambda_1\Lambda_2$ in either case; we use this approximation in the following.

Since Eq. (25) is a measured quantity, it has some uncertainty $\delta\overline{\mathcal{F}}$. This is given by:

$$\left(\frac{\delta\overline{\mathcal{F}}}{\overline{\mathcal{F}}}\right)^2\Bigg|_{\rho=\sqrt{\rho_1^2+\rho_2^2}} = \sum_{i=1,2}\left(\frac{\delta F_i(\rho_i)}{F_i(\rho_i)}\right)^2, \tag{26}$$

where $F_i(\rho_i)$ is the estimated survival function in the ith detector, given by:

$$F_i(\rho_i) = \frac{n_i(\rho_i)}{\Lambda_i}. \tag{27}$$

Here, $n_i(\rho_i)$ is the number of triggers in the ith detector with SNR $\geq \rho_i$. We estimate δF_i by finding the range of F_i for which n_i varies by no more than one standard deviation. Using the Binomial distribution this is (similar to equation 22):

$$\overset{\max}{\min} F_i = \frac{\Lambda_i(2n_i + 1) \pm \sqrt{4\Lambda_i n_i(\Lambda_i - n_i) + \Lambda_i^2}}{2\Lambda_i(\Lambda_i + 1)}. \tag{28}$$

The error is thus:

$$\pm\, \delta F_i = \mp F_i \pm \overset{\max}{\min} F_i. \tag{29}$$

This error estimate can be asymmetric about F_i; to propagate to $\delta\mathcal{F}$, we use $+(-)\delta F_1$ and $+(-)\delta F_2$ to find $+(-)\delta\mathcal{F}$.

Equation (25) estimates the probability of getting a trigger with combined SNR ρ in a *single* draw from the background distribution. If we do k draws, the probability of getting one or more events from the background with combined SNR $\geq \rho$ is:

$$\mathcal{F}(\rho) = 1 - (1 - \overline{\mathcal{F}}(\rho))^k, \tag{30}$$

with error:

$$\pm\, \delta\mathcal{F}(\rho) = k(1 - \overline{\mathcal{F}})^{k-1}(\pm\delta\overline{\mathcal{F}}). \tag{31}$$

Thus, if we have two detectors with Λ_1 and Λ_2 triggers, k of which form zero-lag, or correlated, coincidences, then we can estimate the probability (and the uncertainty in our estimate of the probability) that each trigger was drawn from the same distribution as background, or uncorrelated, coincidences using Eqs. (25)–(31). The smaller this probability is for a zero-lag coincidence, the less likely it is that that coincidence was caused from uncorrelated sources. Since gravitational waves are expected to be the only correlated source across detectors, we use this probability as an estimate for the FAP.

As this study is concerned with just the loudest zero-lag events, it is useful to evaluate the smallest FAP that can be estimated using this method, and its uncertainty. From Eq. (25), the smallest single-draw $\overline{\mathcal{F}}$ that can be estimated is $(\Lambda_1\Lambda_2)^{-1}$. By definition, this occurs at the largest combined background SNR, $\rho^\dagger$. If the combined SNR of the loudest zero-lag event is not $> \rho^\dagger$, then $\rho^\dagger$ must be formed from the largest SNRs in each detector, so that $n_i = 1$. Assuming $\Lambda_1, \Lambda_2 \gg 1$, then from Eqs. (30) and (26) we find:

$$\min\overline{\mathcal{F}} \pm \delta\overline{\mathcal{F}} \underset{N_{1,2}\gg 1}{\approx} \frac{k}{\Lambda_1\Lambda_2}\left[1 \pm \left\{\begin{array}{c} 2.3 \\ 0.87 \end{array}\right\}\right]. \tag{32}$$

If the combined SNR of the loudest zero-lag is $> \rho^\dagger$, then we cannot measure its FAP. In this case, we use Eq. (32) to place an upper limit on $\overline{\mathcal{F}}$.

Determining the $\overline{\mathcal{F}}$ for every zero-lag trigger can require storing and counting a large number of background triggers. To save computational time and storage requirements, we reduce the number of background triggers that have $\mathcal{F} >$ some fiducial $\mathcal{F}_0$ by a factor of $\mathcal{F}/\mathcal{F}_0$ for each order of magnitude increase in $\mathcal{F}$. We then apply a weight of $\mathcal{F}/\mathcal{F}_0$ to the remaining background triggers when finding $\mathcal{F}$ for the zero-lag. For this study, $\mathcal{F}_0$ was chosen to be 10^{-5}. Thus, between $\mathcal{F} = 10^{-4}$ and 10^{-5}, 1 out of every 10 background triggers was kept, with a weight of 10 applied to the remaining. Likewise, between $\mathcal{F} = 10^{-3}$ and 10^{-2}, 1 out of every 100 background triggers was kept, with a weight of 100 applied to the remaining; etc. This substantially reduces the number of background triggers that need to be counted and stored; e.g., for $\lambda_1\lambda_2 = 10^8$, only $\sim$ 5000 background triggers are needed, a saving of about 5 orders of magnitude. The trade-off is our accuracy in measuring the FAP is degraded for triggers with $\mathcal{F} > \mathcal{F}_0$. This is assumed to be acceptable in a real analysis, since triggers with larger $\mathcal{F}$ are, by definition, less significant.[5]

C. The gstlal approach

The method to estimate the FAP of coincident events based on the likelihood ratio ranking statistic described in [15] was modified for this test to use a single parameter, ρ_c. The FAP for a single coincident event can be found as

$$P_0(\rho_c) = \int_{\Sigma_{\rho_c}} \prod_i p(\rho_i)\, d\rho_i, \tag{33}$$

where $p(\rho_i)\, d\rho_i$ are the probability densities of getting an event in detector i with SNR ρ_i, and Σ_{ρ_c} is a surface of constant coincident SNR. The distributions $p(\rho_i)\, d\rho_i$ are measured by histogramming the single detector SNR values either with or without the coincident events included. To get the cumulative distribution for a single event we have

$$P_0(\rho_c > \rho_c^*|1) = 1 - \int_0^{\rho_c^*} P_0(\rho_c)\, d\rho_c. \tag{34}$$

The multiple event FAP is found in the same way as (22).

Notice that for this MDC, an artificial lower boundary of 10^{-6} is set, as the participant decided any estimation below it is subject to excessive uncertainty and thus not reliable.

IV. RESULTS

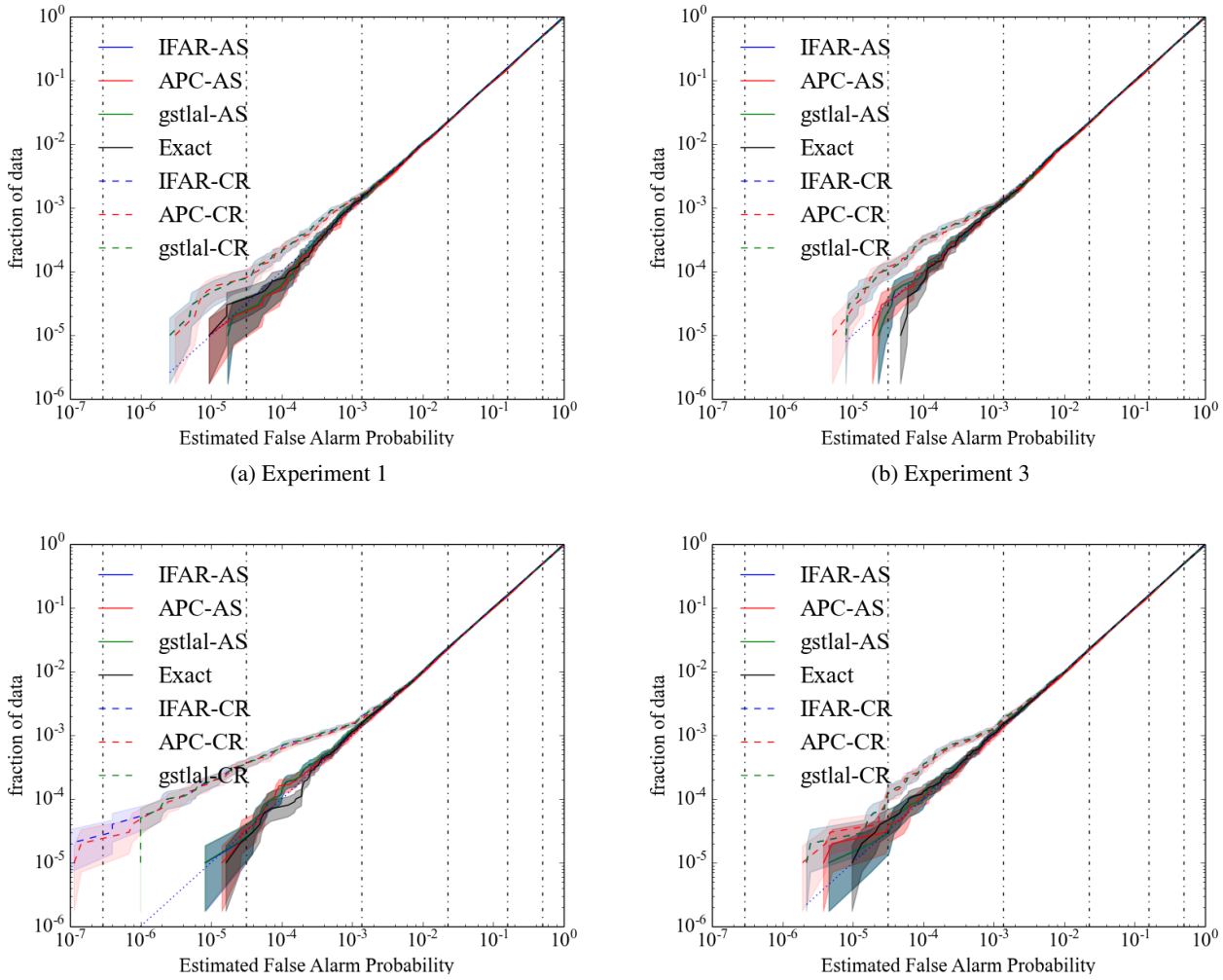
To achieve our aims of comparing the estimation from the 'coincidence removal' and 'all samples' modes, we will examine multiple properties of the submitted challenge results. We first examine the self-consistency of each set of results for each simulation in Sec. IV A. For experiments in the absence of signals, the fraction of realisations with an estimated FAP smaller than a certain threshold should be identical to the value of that threshold; we denote this property as self-consistency. In Sec. IV B we then investigate the accuracy of the FAP estimates by direct comparison with the exact calculated values for each realisation in each simulation. In Sec. IV C we select certain range of data and compare the median and mean of estimate with the exact value for both modes. In Sec. IV D we then construct Receiver Operating Characteristic (ROC) plots for each experiment as a way to compare estimates via their detection efficiency at fixed false positive rates. Finally in Sec. IV E we address the general issue of FAP estimate precision and attempt to extrapolate our findings to predict the likely uncertainties rephrased on significance estimates for GW detection. The challenge was attempted by 3 different teams using a similar but independently implemented algorithms (see Sec. III). Each algorithm was operated in 2 modes, one in which zero-lag triggers were included in the background estimate and the other in which they were removed. For each realisation of each experiment this gives us 6 FAP estimates to compare. In the main text we include only plots from selected simulations that highlight the main features of the comparisons; all other plots can be found in Appendix B.

A. Self consistency tests: p-p plots

In Fig. 2 we show the relationship between the estimated FAP values and their cumulative frequency of occurrence. When the zero-lag coincidences are drawn from the background distribution from which the FAP values are derived then we expect the curves to trace the diagonal. The figure shows results for the 4 experiments (1, 3, 12 and 14) for which there were only background triggers. As we probe lower FAP values (i.e., rarer events) we begin to see counting noise due to the finite number of rare events. However, we see a marked difference between the 'coincidence removal' and 'all samples' modes and no discernible differences between algorithms. In all cases the 'all samples' mode stays consistent with the diagonal within the expected fluctuations due to the finite number of samples. The 'coincidence removal' results, however, always systematically overproduces very small numerical values of FAP, with deviation from the expected behaviour for all values below $\sim 10^{-3}$.

Experiments 1 and 3 were both designed to have simple background distributions: the logarithms of their CDF tails are linear in SNR with each detector having the same distribution, each experiment having a different slope. Experiment 14 was designed to have an extreme background distribution with multiple CDF features. The behaviour of the p–p plots in these 3 cases is very similar with the 'coincidence removal' mode deviating (by $\sim$1–2 standard deviations from the diagonal) for FAPs $< 10^{-3}$. At exact FAP values of 10^{-4} the 'coincidence removal' mode tends to assign $\sim$3 times as many

[5] In retrospect, this background degradation was not really necessary for this study, since we were only interested in the FAP of the loudest zero-lag event in each realisation. However, we wished to keep the analysis method as similar as possible to what would be done in a real analysis.



(a) Experiment 1

(b) Experiment 3

(c) Experiment 12

(d) Experiment 14

FIG. 2: Plots of estimated FAP value versus the fraction of events with that value or below (known as a p-p plot). If the estimate is self-consistent we would expect the value to be representative of its frequency of occurrence; the diagonal line indicates a perfectly self-consistent FAP estimate. We show results for the four experiments where the triggers were generated from background only. The solid lines are the results obtained for our three algorithms in 'all samples' mode while the dashed lines are for the 'coincidence removal' mode of operation. Shaded regions include the uncertainty due to Poisson counting noise. Vertical dashed lines indicate the FAP associated with integer multiples of Gaussian standard deviations, i.e. the equivalent of $n\sigma$ confidence.

realisations with an estimated FAP at or below this value. By contrast the 'all samples' mode remain consistent throughout within the 1–σ counting uncertainties. For experiment 12, intended to have a 'realistic' background distribution, deviation from the diagonal occurs at approximately the same point (FAP $\sim 10^{-3}$) for 'coincidence removal'; here, for an estimated FAP of 10^{-4}, there are $\sim$7 times the number of estimated values at or below this level. The discrepancy in this case and the previous 3 experiments cannot be accounted for by counting uncertainty over experiments.

The deviations seen for the 'coincidence removal' case do not have direct implications for point estimates of FAP in specific realisations; they also do not indicate a bias in those estimates in the sense of a systematic *mean* deviation of the estimated FAP away from the exact value. The result does however indicate that for rare events in a background-only dataset, using a 'coincidence removal' mode gives a greater than $\mathcal{F}$ chance of obtaining an event of estimated FAP $\mathcal{F}$. This result is also expected to hold for experiments where the majority of realisations do not contain foreground triggers, i.e. those with 'low' signal rates.

We may understand the onset of systematic discrepancies between the two modes as follows. The change in estimated significance due to removal of coincident triggers will only have a major effect – comparable to the estimated value itself – when much of the estimated background (louder than the loudest zero-lag event) is generated by the triggers that are being removed. This is likely only to be the case when the

loudest event itself contains one of the loudest *single-detector* triggers. Thus, the probability of a substantial shift in the estimate due to removal is approximately that of the loudest trigger in a single detector forming a random coincidence; for the parameters of this MDC this probability is $2\lambda_1\lambda_2\delta_t/T \simeq 10^{-3}$.

B. Direct comparison with exact false alarm probability (FAP)

In this section, we show the direct comparison of estimated FAP values with the exact FAP. In a parameter estimation problem we may consider both the accuracy and precision of the estimates as figures of merit: ideally the spread of estimated values compared to the exact value should be small and the estimated values should concentrate around the exact value. The estimated values could be influenced by a number of factors including random fluctuations in the statistics of triggers, structures like hidden tails could bias the estimates, and there may be contamination from a population of foreground triggers. Where possible we attempt to understand the performance of the algorithms in terms of these factors, and to quantify their influences.

Although the comparison shows obvious difference between the 'all samples' and 'coincidence removal' modes, readers are reminded that the quantitative result does not necessarily reflect the behaviour in an actual GW search. In Figs. 3–7 the estimation shows a large scatter in estimated FAP values below 10^{-3}; this value is, though subject to the design of the MDC, which does not attempt to model all aspects of a CBC search on real data.

1. Low foreground rate

To compare estimates of FAP we display them for each realisation, plotting $\log \mathcal{F}$ since the occurrence of very low values is crucial for detection claims (either true or false). In these figures, a perfect estimation would lie on the line $y = x$; if an algorithm provides an underestimate by assigning a smaller FAP, it will fall below the diagonal line; an overestimate would lie above the diagonal line.

For the experiments with no signal triggers or low foreground rate, the triggers are at most slightly contaminated by foreground signals, so the estimation of the FAP should be correspondingly unaffected by their presence. Where there are no foreground triggers present, even the extreme backgrounds, e.g. experiment 14, shown in Fig. 3, don't appear to adversely affect the estimation and the spread is relatively small around the diagonal line. However, in the 'coincidence removal' mode, for all algorithms there is a tendency to *underestimate* $\mathcal{F}$ for small values. This is not conservative, i.e. it is over-optimistic, in the sense that underestimating $\mathcal{F}$ claims that the experiment outcome is rarer than they are in reality (in the absence of signal).

In Fig. 3 we also see that gstlal estimation is very close to that of the IFAR approach. For other experiments (see App. B 2), their results do show small discrepancies most notably in their different lower limits for estimated FAP values.

For all APC results, the estimation method used was designed such that only for rare events (those with low FAP values) were the results computed to the highest accuracy (hence the large spread in FAP estimation seen in Fig. 3). This is motivated by the fact that the astrophysical events that we are ultimately interested in will necessarily have small FAP values, and by the computational load of the challenge itself.

2. Medium foreground rate

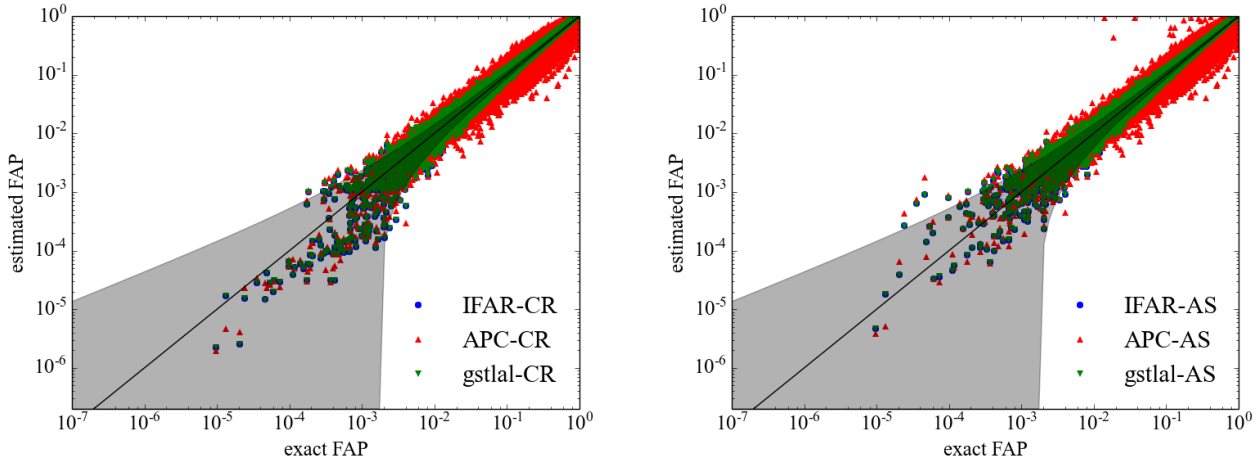
The experiments with "medium" foreground rate have an average of ∼half the realisations containing a foreground coincidence. Foreground triggers are drawn from a long-tailed astrophysical distribution in SNR and are likely to be loud if present. In such realisations, any bias in the estimation of FAP $\mathcal{F}$ due to this "contamination" would be in the direction of overestimation. This kind of bias is considered to be conservative since it would underestimate the rarity of possible astrophysical events under the null hypothesis.

We use results from experiments 9 and 6, shown in Figs. 4 and 5 resp., as examples of a medium foreground rate. We see greater variance in the FAP estimates in the low FAP region, in comparison to experiments with zero or low foreground rates. This increased variance is clearly caused by the presence of foreground events since nearly all points below 10^{-5} on the x-axis are due to signals. We again see general agreement between algorithms (with the exception of APC at high $\mathcal{F}$ values) however there is now evidence of discrepancies at the lowest values of estimated FAP. This is mostly due to the different choices of lowest estimatable value between algorithms, and is independent of the MDC dataset. There is now also stronger indication that 'coincidence removal' shifts estimates to lower values compared to the 'all samples' mode.[6] Further investigation of this systematic difference between modes will be made in Sec. IV C.

The experiment shown in Fig. 5 has a background distribution with a shallow 'platform' (seen in Fig. 6), classed as an 'extreme' background. Here we see similar behaviour to experiment 9, but with even greater variation in estimated FAP values, spanning ∼4 orders of magnitude for all exact FAP values below $\sim 10^{-3}$.

The platform feature ranging in SNR between approximately $8 - 11$ contains on average less than 0.1 triggers per realisation. Therefore in many cases the background at high SNR is not well represented and could fool our algorithms towards underestimation of $\mathcal{F}$, while in other cases the contamination due to foreground triggers could mimic the background and lead to an overestimation of $\mathcal{F}$.

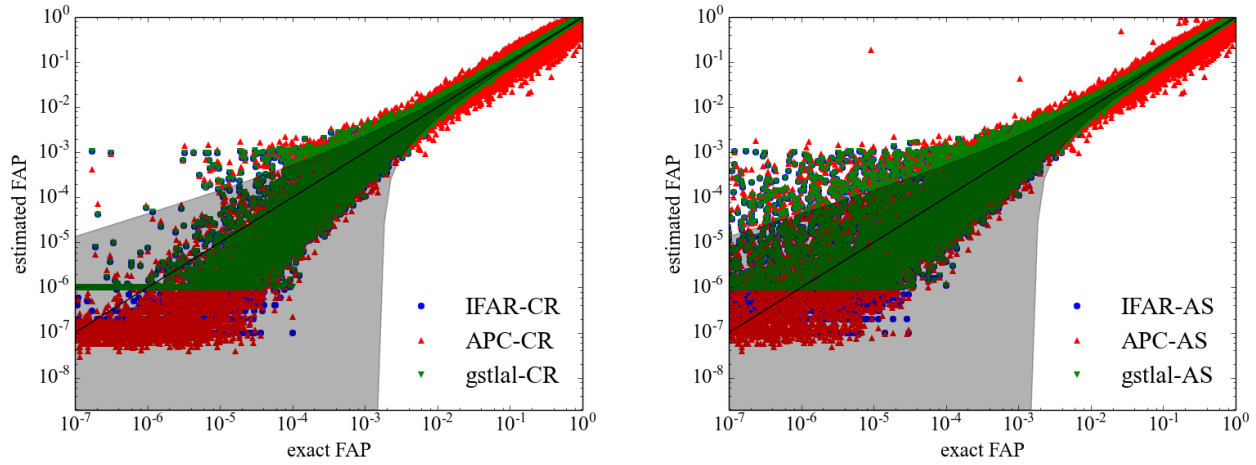
[6] Note that since these plots have logarithmic scales, a very small difference in $\mathcal{F}$ may have a large apparent effect at low values.



(a) Direct comparison with 'coincident removal'

(b) Direct comparison with 'all samples'

FIG. 3: Direct comparisons of FAP estimates with the exact FAP for experiment 14 (containing no signal events). In both plots the majority of blue points are masked by the green points since these methods provide closely matching results. The estimates are concentrated on the diagonal in both 'coincidence removal' and 'all samples' cases. However, in the 'coincidence removal' mode, all algorithms place the majority of points under the diagonal for exact FAP values ($<10^{-3}$), indicating a non-conservative estimate.



(a) Direct comparison with 'coincidence removal'

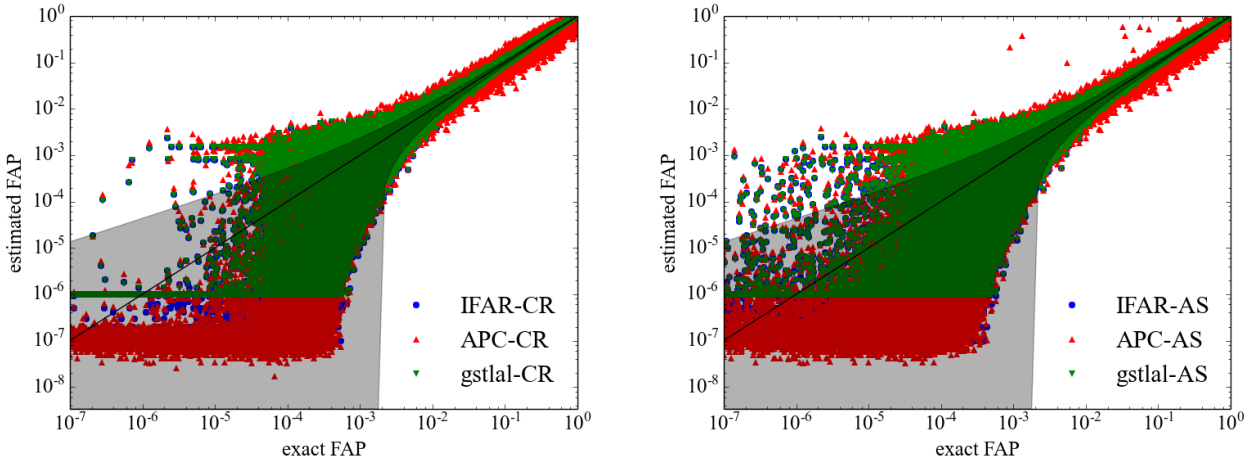
(b) Direct comparison with 'all samples'

FIG. 4: Direct comparisons of FAP estimates with the exact FAP for experiment 9. The medium level foreground rate in this case leads to a number of realisations containing signals, resulting in a larger vertical spread. Different algorithms fix different lower boundary values for the estimated FAP, visible in the sharp lower edge in estimated $\mathcal{F}$ of gstlal results. The shaded region represents the expected uncertainty from Eq. 19.

3. *High foreground rate*

Experiments with "high" foreground rate have, on average, $>$ 1 foreground event per realisation. Results from experiment 11 are shown in Fig. 7 where, as is generally the case, there is good agreement between algorithms but clear variation between 'coincidence removal' and 'all samples' modes. Compared with experiments with lower foreground rates, the presence of many contaminating foreground signals shifts estimates to higher values. For 'coincidence removal' mode, this shift reaches the point where, in this experiment, the bulk of the distribution now appears consistent with the exact FAP with relatively narrow spread.[7] For the 'all samples' mode the contamination due to signals is greater and so is the corresponding shift to higher values of FAP (i.e. to the conservative side); in fact, underestimates of $\mathcal{F}$ are extremely rare. For all algorithms and for both 'coincidence removal' and 'all samples', as the foreground rate increases, a horizontal feature

[7] We again remind the reader again that the comparison is presented on logarithmic axes.



(a) Direct comparison with 'coincidence removal'

(b) Direct comparison with 'all samples'

FIG. 5: Direct comparisons of FAP for experiment 6. The presence of a platform feature in the tail of the distribution causes the spread in estimates values to be wider than for experiment 9. The shaded region represents the expected uncertainty from Eq. 19.

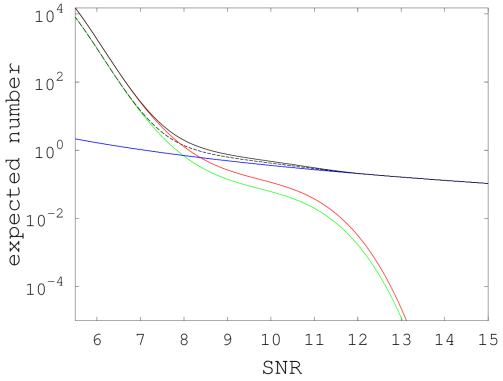


FIG. 6: Reverse CDF of trigger SNRs for experiment 6. The red and green curves represent the two individual detectors, while the blue curve represents the astronomical signals. The black lines represent the combined distribution of both background and foreground triggers.

appears in these comparison plots, which we discuss in the following section.

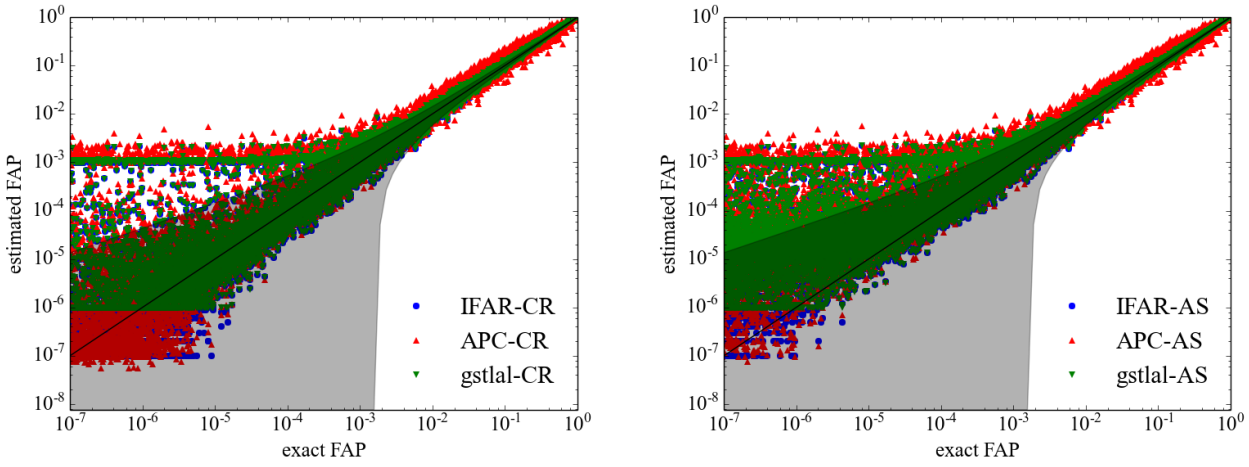
4. *Horizontal bar feature*

In all high foreground rate scenarios, horizontal features appear at $\sim 10^{-3}$ in estimated FAP, which are also marginally visible in medium rate experiments. The process of FAP estimation for the loudest coincident event is based on collecting the fraction of all possible unphysical coincidences which are louder. The estimation will be strongly biased when there exists a foreground trigger in one detector that is extremely loud and either not found in coincidence in zero-lag, or coincident with a trigger with very low SNR. In such cases it is highly likely that when performing background estimation it would result in background coincidences which are louder than the loudest zero-lag event (the details of this process are specific to each algorithm). Assuming a method that makes use of all possible unphysical trigger combinations between detectors, this corresponds to $\sim 10^4$ louder background coincidences out of a possible $\sim 10^8$ in total. Considering an expected ~ 10 zero-lag coincidences this gives an estimated FAP of these asymmetric events as $\sim 10^{-3}$.

In experiment 11 (Fig. 7), there are $\sim$ 650 such realisations. For $\sim$ 500 of them, the cause is a single astrophysical signal appearing as an extremely loud trigger in one detector, while for the other detector the combination of antenna pattern and non-central χ^2 random fluctuations results in a sub-threshold SNR and is hence not recorded as a trigger. The remaining $\sim$ 150 events also have very loud SNRs in one detector, but in these cases the counterpart in the second detector appears only as a relatively weak trigger. When foreground events appear with asymmetrical SNRs between the two detectors, removing coincident triggers from the background estimate could mitigate overestimation occurring in such cases; while for the $\sim$500 realisations that contain a single loud foreground trigger which does not form a coincidence, overestimation will occur regardless of the method used.

5. *Uncertainty estimate*

Throughout figure 3 - 7, a shaded region was plotted which represents the uncertainty predicted from Eq. 19. In the derivation of Eq. 19 several simplifying assumptions were used, thus some discrepancy between the theoretical expectation and the actual spread is not surprising. However, this expression does capture the order of magnitude of the uncertainty, so as a "rule of thumb" it serves as a guide in the estimation of uncertainty for the FAP.



(a) Direct comparison with 'coincidence removal'

(b) Direct comparison with 'all samples'

FIG. 7: Direct comparisons of FAP for experiment 11. The high foreground rate in this case causes general shifts towards larger estimates for the FAP. The horizontal bar feature visible in both plots is also most prominent in this high rate case. The shaded region represents the expected uncertainty from Eq. 19.

C. Box plots

In this section we characterise the estimated FAP values in more detail by conditioning them on the value of exact FAP. We take decade-width slices in exact FAP and for each summarise the distributions of estimated values. For a given experiment, algorithm, and mode, we isolate those results corresponding to the chosen FAP decade and take the ratio of estimated to exact FAP value. We then calculate and plot a central box, median point, mean value, whiskers and individual outliers.

The central box contains the central half of all estimated values, covering from 25% to 75% of the total sorted values, i.e. between the first and the third quartile, also known as the inter-quartile range (IQR). The box is divided by a vertical line identifying the 50% (median) quartile. When the distribution is relatively concentrated, and the most extreme samples lie within 1.5× the IQR, then the whisker ends at the most extreme samples. Otherwise the whisker is drawn to be 1.5× the IQR, and outliers beyond the whisker are drawn as individual points. We also indicate on these plots the mean value of ratio for each distribution.

Since we are more interested in the region of low FAP values, where detection of foreground signals will likely occur, we take bins corresponding to values of exact FAP between $(10^{-5}$–$10^{-4})$, $(10^{-4}$–$10^{-3})$, and $(10^{-3}$–$10^{-2})$. For each bin, we draw box plots on the ratio between the estimated and exact FAP value, using a logarithmic x-axis scale. The vertical purple line corresponds to where the log of the ratio is zero, meaning that the estimation and exact FAP are identical; left hand side means the estimated FAP value is smaller than the actual value, which translates to an underestimation of FAP. In all plots the vertical axis gives the experiment index ranging from the lowest foreground rates to the highest. For each index there are 3 coloured boxes associated with each algorithm. Figures are divided into 2 plots, one for the 'coincidence removal' mode and the other for 'all samples'.

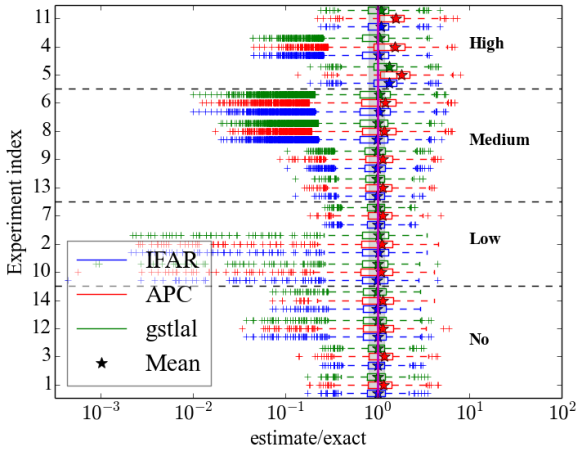
1. *False alarm probability range* 10^{-3}–10^{-2}

In Fig. 8 we see relatively tight and symmetric distributions for all algorithms when considering the IQR with strong agreement specifically between the gstlal and IFAR algorithms. We remind the reader that the APC algorithm was not optimised at high FAP values and hence shows very slightly broader distributions. We note that the extrema of the FAP ratios in most experiments range symmetrically over $\lesssim \pm 1$ order of magnitude. However, for some experiments, most notably 12, 10, 2, 8, 6 and 4 there are large deviations in the extrema towards underestimates of FAP. Such deviations would be classed as non-conservative, i.e. events are estimated as more rare than indicated by the exact calculation. This effect is somewhat reduced for the 'all samples' mode, most evidently for experiments 2 and 10.

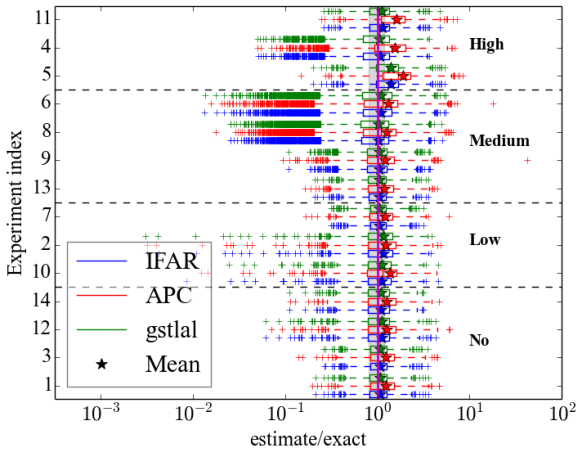
The points identified with star symbols in Fig. 8 show the means of the distribution of ratios. In general, the means for the 'coincidence removal' mode are slightly more consistent with the expected vertical line than for the 'all samples' mode. This trend will be amplified in subsequent sections as we investigate lower values of exact FAP. For this $(10^{-3}$–$10^{-2})$ region we note that, for reasons discussed earlier, the means computed from APC tend to overestimate the expected value.

2. *False alarm probability range* 10^{-4}–10^{-3}

As we move to lower FAP ranges, shown in Fig. 9, we start to see the effects of having lower numbers of results. By definition, for experiments with no foreground we expect to see a factor of $\approx$10 fewer results in the decade $(10^{-4}$–$10^{-3})$ than

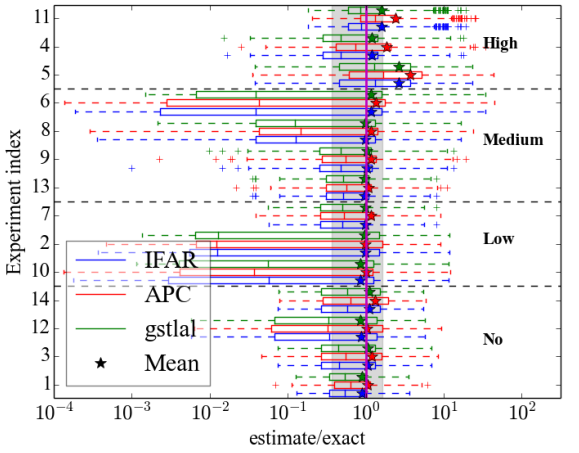


(a) Box plots based on ‘coincidence removal’.

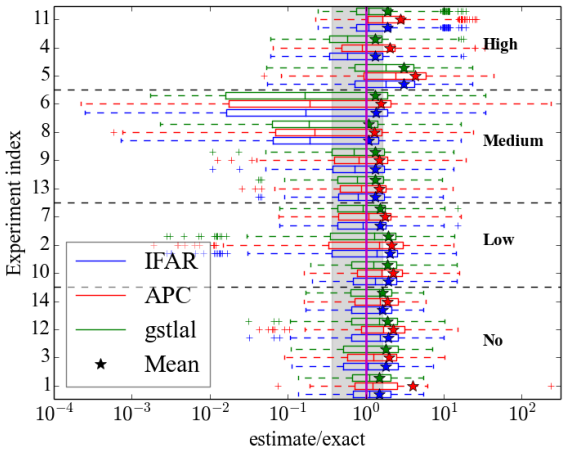


(b) Box plots based on ‘all samples’

FIG. 8: Box plots of the ratio of estimated to exact FAP value, for exact FAPs between 10^{-3} and 10^{-2}. The shaded region represents the expected uncertainty from Eq. 19.



(a) Box plots based on ‘coincidence removal’



(b) Box plots based on ‘all samples’

FIG. 9: Box plots of the ratio of estimated to exact FAP value, for exact FAPs between 10^{-4} and 10^{-3}. The shaded region represents the expected uncertainty from Eq. 19.

in the decade (10^{-3}–10^{-2}), implying larger statistical fluctuations due to the reduced number of samples. We also see intrinsically broader distributions, as the estimation methods themselves are constrained by the infrequency of loud, low-FAP events. As seen in previous figures of merit, results differ only slightly between algorithms with the largest differences coming from the issue of inclusion or removal of coincident triggers.

Overall, we see ranges in the extrema spanning ±1 order of magnitude for both ‘coincidence removal’ and ‘all samples’ modes. However, for experiments 10, 2, 6, and 8 the lower extrema extend to ∼ 4 order of magnitude below the expected ratio for the ‘coincidence removal’ mode. This behaviour is mitigated for the ‘all samples’ mode: note that for experiment 10 the extrema are reduced to be consistent with the majority of other experiments. In general it is clear that the IQRs for the ‘coincidence removal’ mode are broader in logarithmic space than for ‘all samples’. This increase in width is always to lower values of the ratio, implying underestimation of the FAP. This trend is also exemplified by the locations of the median values: for the ‘coincidence removal’ mode, low foreground rates yield medians skewed to lower values by factors of ∼2–200. For the 3 high foreground rate experiments the IQRs and corresponding medians appear consistent with the exact FAP. For the ‘all samples’ mode the IQRs and medians are relatively symmetric about the exact FAP and the IQRs are in all cases narrower than for the ‘coincidence removal’ mode.

In this FAP range the difference in distribution means between the ‘coincidence removal’ and ‘all samples’ modes becomes more obvious. The removal mode results consistently return mean estimates well within factors of 2 for all no, low and medium foreground rates. For high foreground rates they consistently overestimate the means by up to a factor of ∼ 3. For the ‘all samples’ mode there is a clear overestimate of the ratio (implying a conservative overestimate of the FAP) for all experiments irrespective of foreground rate or background

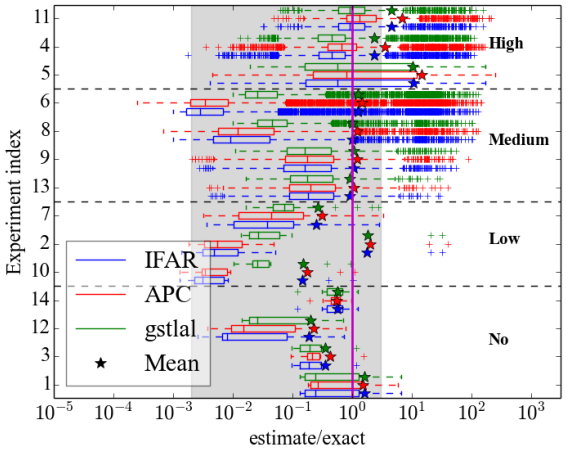
complexity. This overestimate is in general a factor of ∼2. Note though that the estimates from both modes for the three high foreground rate experiments are very similar in their distributions and means.

3. *False alarm probability range* 10^{-5}–10^{-4}

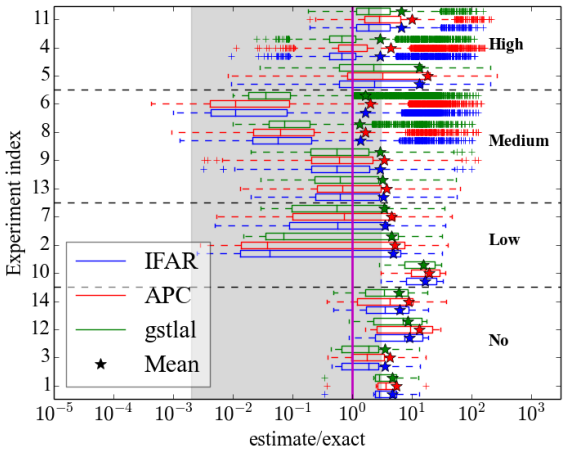
In this FAP range the uncertainties and variation in the results are strongly influenced by the low number of events present at such low FAP values. Among all the experiments with no foreground rate suffer the most. Nonetheless, in Fig. 10 we see similarities between algorithms and striking differences between 'coincidence removal' and 'all samples' modes. Firstly, in all cases the variation in extrema is comparable, in this case spanning ∼ ±3 orders of magnitude. The IQRs are broadly scattered and in many cases do not intersect with the exact FAP. This is not indicative of poor estimation but indicative of highly non-symmetric distributions.

For no and low foreground rates there is a marked difference between results from 'coincidence removal' and 'all samples' modes. For 'coincidence removal' all distributions are skewed to low values which is also a characteristic for the medium foreground rate experiments. For example, in experiment 12 there are no estimates in any of the realisations in this range that overestimate the FAP. Removal methods in general in this range of very low FAP for low and medium foreground rates provide IQRs of width ∼1 order of magnitude with medians shifted by between ∼1–2 orders of magnitude below the exact values. For 'all samples' mode, all no and low foreground experiments (with the exception of experiment 2) provide conservative (over)estimates of the FAP ratio with IQRs and extrema ranges spanning <1 and ∼1 order of magnitude respectively. We see that for experiment 10 there are no 'all samples' estimates in any realisation that underestimate the FAP. With 'all samples' there is then a marked change as we move to medium level foreground rates and the distributions become relatively symmetric in log-space with all medians lower than, but mostly within a factor of 2 of, the exact FAP. Experiments 6 and 8 both have medium level foreground rates and give rise to results that are very similar between 'coincidence removal' and 'all samples' results and that exhibit highly skewed distributions to low values with long distribution tails extending to high values. This trend of similarity is then continued for high foreground rates where there is little distinction between either algorithm of 'coincidence removal' mode. In these cases however, the distributions appear relatively symmetric in log-space with reasonably well constrained IQRs.

Considering the means of the distributions, we see similar behaviour but with more variation than in the previous FAP ranges. Starting with 'all samples' mode there is consistent conservative bias in the mean of the ratio of estimated to exact FAP. For low/no and high foreground rates this bias is ∼1 order of magnitude which reduces to a factor of ∼3 overestimate for medium level foregrounds. For the 'coincidence removal' mode, no and low foreground rates return distribution means that are scattered symmetrically about the exact FAP with a variation of ∼1 order of magnitude. For all medium



(a) Box plots based on 'coincidence removal'



(b) Box plots based on 'all samples'

FIG. 10: Box plots for the ratio between estimated and exact FAP values for FAPs between 10^{-5} and 10^{-4}. The shaded region represents the expected uncertainty from Eq. 19.

level foregrounds including experiments with low, medium and high background complexity, the mean estimates are very tightly localised around the exact FAPs with variations of 10's of percent. For high foreground rates the means obtained from both 'coincidence removal' and 'all samples' modes are all consistently overestimates of the exact FAP by up to ∼1 order of magnitude.

By looking at the bulk distribution in the box plots, it seems that 'coincidence removal' will generally underestimate the FAP, while 'all samples' is systematically unbiased over all 14 experiments. However, note that if we only look at the mean values, then 'all samples' modes almost always overestimate the FAP, while 'coincidence removal' is generally consistent with the exact FAP. This indirectly proves that 'coincidence removal' mode is a mean-unbiased estimator. For a significant event, the exact FAP is very close to zero, thus any difference due to underestimation may be very small in linear space though not necessarily small in log space, while

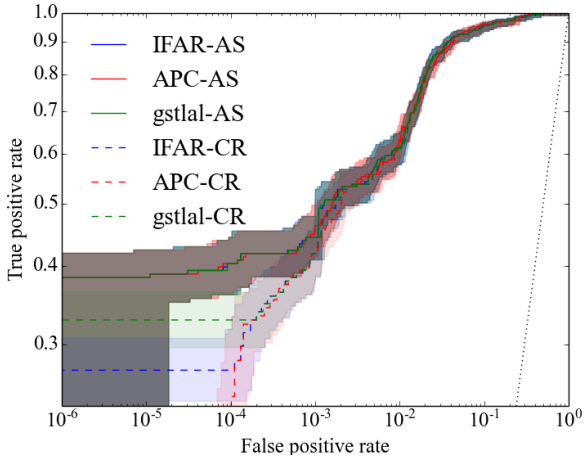


FIG. 11: ROC plot for experiment 2. The error bars in both horizontal and vertical directions are calculated with a binomial likelihood under a 68% credible interval. Solid lines correspond to 'all samples', while dashed lines corresponds to 'coincidence removal'. The dotted line represents the expected performance of random guess, no rational analysis would perform worse than it.

overestimation could bias the value with a large relative deviation. When the FAP is very small, the estimation uncertainty (variance) is large relative to the exact value; then, since estimated values are bounded below by zero, in order to achieve a mean-unbiased estimate a large majority of estimated values are necessarily below the exact value, i.e. underestimates. In other words, the distribution is highly skewed.

D. ROC analyses

The FAP value is a measure of the rarity of observed events, but in this section we treat the estimated FAP as a test statistic. This allows us to use ROC plots to compare the ability to distinguish signal from noise for each method. In practice this involves taking one of our experiments containing 10^5 realisations and, as a function of a variable threshold on our test statistic (the FAP), computing the following. The false positive rate (FPR) is the fraction of loudest events due to background that had estimated FAP values below the threshold. The true positive rate (TPR) is computed as the fraction of loudest events due to the foreground that had estimated FAPs below the threshold. For each choice of threshold a point can be placed in the FPR-TPR plane creating an ROC curve for a given test-statistic. Better performing statistics have a higher TPR at a given FPR. A perfect method would recover 100% of the signals while incurring no false alarms, corresponding to a ROC curve that passes through the upper left corner. A random classifier assigning uniformly distributed random numbers to the FAP would lead to identical FPR and TPR, yielding a diagonal ROC curve.

Error regions for the FPR and TPR are computed using a binomial likelihood function. In general, as can be seen in our ROC plots, as the foreground rate goes up, the more events are available to compute the TPR, reducing the vertical uncertainties. Conversely, the more background events are available, the smaller the horizontal uncertainties.

In the following subsections we focus on the experiments where there are clear discrepancies, leaving cases where there is agreement between methods to Appendix B 3. We stress that ROC curves allow us to assess the ability of a test-statistic to distinguish between realisations where the loudest event is foreground vs. background; however they make no direct statement on the accuracy or precision of FAP estimation.

1. *Low foreground rate*

There are 3 experiments, 2, 7 and 10, that have low foreground rates. The ROC curve for experiment 2 in Fig. 11 exhibits a number of interesting features. Firstly, there is general agreement between algorithms; deviations are only visible between 'coincidence removal' and 'all samples' modes of operation. At a FPR of $\sim 10^{-3}$ and below, the 'all samples' mode appear to achieve higher TPRs, when accounting for their respective uncertainties, by $\sim$ 10%. This indicates that in this low-rate case, where $\approx$ 1 in 1000 loudest events were actual foreground signals, the 'all samples' mode is more efficient at detecting signals at low FPRs. We can identify all experiments that show such deviations, and all have tail features or obvious asymmetry between the two detectors' background distributions, combined with a low to medium foreground rate.

2. *Medium foreground rate*

Experiments 6, 8, 9 and 13 contain medium foreground rates and collectively show two types of behaviour. Experiments 9 and 13 show general agreement between algorithms and 'coincidence removal' and 'all samples' modes (see Figs. 37 and 40). Experiments 6 and 8 show similar deviations to those seen in the p–p plots in Section IV A. This similarity is not surprising since the vertical axes of the p-p plots are identical to horizontal axes of the ROC plots, with the distinction that they are computed on background-only and background-foreground experiments respectively.

Here we focus on Experiment 8 shown in Fig. 12 which contained realistic but slightly different background distributions in each detector. As seen in the low-foreground example there is good agreement between algorithms but differences between 'coincidence removal' and 'all samples' modes. In this case, due to the increased number of foreground events, this difference is more clearly visible and the disscrepancies are inconsistent with the estimated uncertainties. So for medium foreground rates we conclude that as a detection statistic, the 'all samples' mode obtains higher TPRs at fixed FPR for low values of FAP. We remind the reader that detection claims will be made at low FAP values, although the absolute values appearing in our MDC may not be representative of those obtained by algorithms in a realistic analysis.

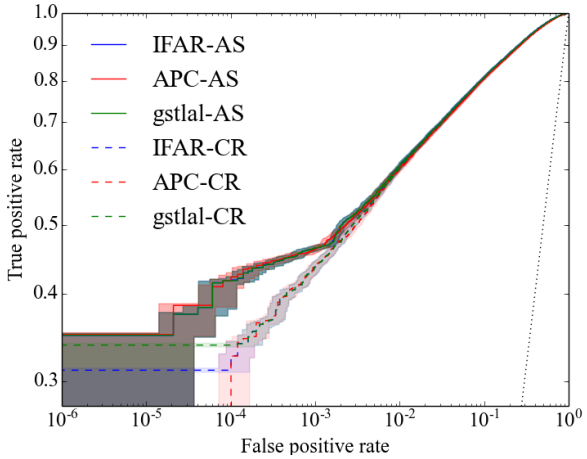


FIG. 12: ROC plot for experiment 8. The error bars in both horizontal and vertical direction are calculated with a binomial likelihood under a 68% credible interval. Solid lines correspond to 'all samples', while dashed lines corresponds to 'coincidence removal'. The dotted line represents the expected performance of random guess, no rational analysis would perform worse than it.

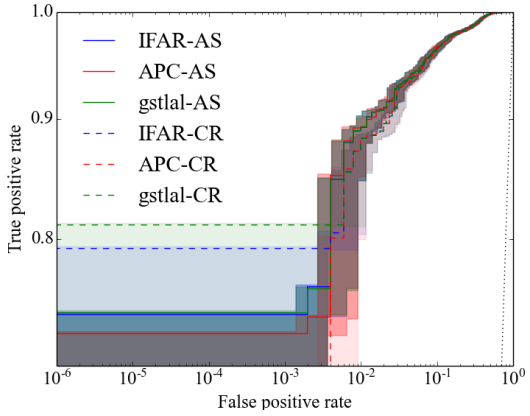


FIG. 13: ROC plot for experiment 4. The error bars in both horizontal and vertical direction are calculated with a binomial likelihood under a 68% credible interval. Solid lines correspond to 'all samples', while dashed lines correspond to 'coincidence removal'. The dotted line represents the expected performance of random guess, no rational analysis would perform worse than it.

3. *High foreground rate*

The high rate experiments 4, 5 and 11 all show similar behaviour. Here we show Fig. 13 as an example where we see general agreement (within our estimated uncertainties) between algorithms and between the 'coincidence removal' and 'all samples' modes. The high rates used result in >90% of realisations containing a loudest coincident event from the foreground. The 3 experiments were examples of all 3 levels of background complexity respectively and the results indicate that in terms of detection efficiency, all algorithms and approaches perform equally well at any fixed FPR. This high rate scenario is most likely to be relevant in the epochs following the first direct detection.

E. Uncertainty in estimation

From the results presented in the previous sections we can conclude that the relative uncertainty in the estimation of FAP increases as FAP decreases. As shown in Figs. 3, 4, 5, and 7, with the exception of APC results, which are designed to be accurate only at low FAP values, both other estimation algorithms show larger spread as the FAP value goes lower. Specific features in the background distributions would vary the actual spread, but the trend is consistent. When the value of the exact FAP is as small as 10^{-4}, the relative uncertainty can exceed 100%; since the estimated FAP can't be negative, the errors in estimated FAP are not symmetrically distributed.

Any claims of GW detection will necessarily be made at low values of FAP and most likely at low or medium level foreground rate. Using Fig. 10 and focusing on the low and medium foreground rate experiments 10, 2, 7, 13, and 9 it is clear from both 'coincidence removal' and 'all samples' modes that a single loudest event measurement of FAP will be drawn from a very broad distribution. For 'all samples' mode, in all but experiment 10 for medium or low foregrounds, the IQR looks symmetric mostly consistent with the true value within ±1 order of magnitude. For the equivalent 'coincidence removal' mode results, the IQR is much smaller than its counterpart in the 'all sample' mode in log-space, and hence more precise. The extrema between approaches are comparable but the bulk of the distribution is more concentrated in the 'coincidence removal' case. Note however that the median and IQR for the 'coincidence removal' mode are both uniformly below the exact value, in some cases by orders of magnitude.

Our results show that the uncertainty in the FAP value can be predicted by the "rule of thumb" Eq. 19, derived in Sec. II D. The scope of the MDC described in this work was designed to probe multiple background types and foreground levels with large numbers of realisations. The number of triggers simulated for each realisation does not match the number that would be present in a realistic GW search; nor does the coincidence test correspond to a real search in which many thousands of distinct, but nontrivially correlated templates are used. Hence, the FAP value at which the uncertainty in our estimates approaches 100% can be very different in reality than the 10^{-3} value seen in our simulations. We expect that Eq. 19 will serve as a guide in estimating uncertainties due to different realisations of noise in the FAP values produced within realistic GW searches.

V. DISCUSSION

We have designed and generated an MDC to evaluate the performance of methods to determine the significance of transient GW candidate events, simulating wide ranges of qualita-

tively different background distributions and of possible signal rates. We compared the results obtained via three different algorithms, each operating in two different modes: estimating the distribution and rate of background events by considering either 'coincidence removal': excluding single-detector triggers found to be coincident in the search from the background estimate; or 'all samples': including all single-detector triggers in the background estimate. These background estimates were then used to assign a false alarm probability (FAP) to the loudest coincident event in each realisation of each experiment in the MDC. Our methods for comparison of these results include self-consistency checks via the use of *p*-*p* plots, for those experiments not containing any foreground signals; direct comparison of estimates with the exact value of FAP resulting from the generating background distributions; a boxplot comparison of the result distributions from each experiment; and finally an ROC analysis to identify detection efficiency at fixed FPR.

Based on these comparison analyses we find the following key conclusions:

a. The results of all experiments indicate a good consistency among all three algorithms; disagreements only occur between the modes of 'coincidence removal' and 'all samples' for low values of false alarm probability (FAP).

b. For all except high foreground rates, the 'coincidence removal' mode is more likely to underestimate the FAP than overestimate, though producing an unbiased mean value; while the 'all samples' mode tends to overestimate, especially for smaller FAPs.

c. For high foreground rates, both the 'coincidence removal' and 'all samples' modes overestimate FAP as indicated by the mean over realisations, while the 'coincidence removal' mode has a larger fraction of underestimated FAP.

d. We only observe extreme underestimation of FAP from either complex or realistic background distributions. When the foreground rate is not high, or the background distributions have no tail or asymmetry between detectors there is evidence that the 'coincidence removal' mode can underestimate the FAP.

e. Due to different detector responses and random noise fluctuations, an astrophysical event may induce a very loud trigger in one detector and a sub-threshold trigger in the other. This would lead to a systematic overestimation of FAP for all algorithms and modes, as shown as Fig. 7.

f. The evaluation of FAP is found to be entirely self-consistent only for the 'all samples' mode. In this MDC, the 'coincidence removal' mode would claim a fraction of 10^{-4} realisations containing only noise events to have FAP 10^{-5}, hence the estimated FAP for this mode does not have a frequentist interpretation at low values. Such a deviation, however, is expected to occur at far lower FAP values for a real analysis of GW data.

g. In general, FAP estimates computed using 'all samples' were found to be more effective at distinguishing foreground events from background at fixed FPR. This was most notable in experiment 8 which contained a medium level foreground and a realistic background.

h. For all but high foreground rates, coincidence removal methods have the merit of appearing to be unbiased estimators concerning the mean of FAP estimates. However, the distributions of these estimates are highly asymmetric, especially for low FAP values. Single realisations from 'coincidence removal' mode are consequently highly likely to have underestimated values of FAP. By contrast, estimates from the 'all samples' mode show a conservative bias (i.e. towards overestimation) concerning the mean over realisations; but for low FAP events, these estimates more likely to lie around the exact value or above it.

i. The relative uncertainty in the estimation is larger when the FAP is smaller. The relative uncertainty reaches 100% when the FAP is about 10^{-4}, for the experiment parameters chosen in this MDC. This value depends on the expected number of coincident events and the number of single detector triggers.

At the time of writing this we eagerly await results from the Advanced detector era. While we are aiming to make several detections over the lifespan of the upgraded Advanced facilities, we should bear in mind that the first detection(s) may be relatively quiet and belong to a distribution with a low astrophysical event rate. In this case, we recommend a sacrifice in possible accuracy of FAP estimation in favour of conservatism. Considering also the self-consistency and relative efficacy of the methods in distinguishing signal realisations from noise, we recommend the use of the 'all samples' mode of any of our 3 algorithms, anticipating that false alarms will likely be (conservatively) overestimated rather than underestimated for the first GW detections.

ACKNOWLEDGMENTS

The authors have benefited from discussions with many LIGO Scientific Collaboration (LSC) and Virgo collaboration members and ex-members including Drew Keppel and Matthew West. We acknowledge the use of LIGO Data Grid computing facilities for the generation and analysis of the MDC data. C. M. is supported by a Glasgow University Lord Kelvin Adam Smith Fellowship and the Science and Technology Research Council (STFC) grant No. ST/ L000946/1. J. V. is supported by the STFC grant No. ST/K005014/1.

REFERENCES

[1] G. M. Harry and the LIGO Scientific Collaboration, Classical and Quantum Gravity **27**, 084006 (2010).

[2] J. Aasi *et al.* (LIGO Scientific), Class.Quant.Grav. **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].

[3] F. Acernese *et al.* (VIRGO), Class.Quant.Grav. **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].

[4] J. Aasi *et al.* (LIGO Scientific Collaboration, Virgo Collaboration), (2013), arXiv:1304.0670 [gr-qc].

[5] J. Abadie *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Class. Quantum Grav. **27**, 173001 (2010).

[6] B. S. Sathyaprakash and B. F. Schutz, Living Reviews in Relativity **12**, 2 (2009), arXiv:0903.0338 [gr-qc].

[7] M. Dominik, E. Berti, R. O'Shaughnessy, I. Mandel, K. Belczynski, C. Fryer, D. E. Holz, T. Bulik, and F. Pannarale, Astrophys. J. **806**, 263 (2015).

[8] M. Morscher, B. Pattabiraman, C. Rodriguez, F. A. Rasio, and S. Umbreit, Astrophys. J. **800**, 9 (2015).

[9] F. Antonini, S. Chatterjee, C. L. Rodriguez, M. Morscher, B. Pattabiraman, V. Kalogera, and F. A. Rasio, ArXiv e-prints (2015).

[10] J. Abadie *et al.*, *Evidence for the Direct Detection of Gravitational Waves from a Black Hole Binary Coalescence*, Tech. Rep. LIGO-P1000146 (LIGO Scientific Collaboration and Virgo Collaboration, 2011).

[11] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A. A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins, and et al., Physics Letters B **716**, 1 (2012), arXiv:1207.7214 [hep-ex].

[12] P. A. R. Ade *et al.* ((BICEP2 Collaboration)), Phys. Rev. Lett. **112**, 241101 (2014).

[13] S. Babak, R. Biswas, P. R. Brady, D. A. Brown, K. Cannon, C. D. Capano, J. H. Clayton, T. Cokelaer, J. D. E. Creighton, T. Dent, A. Dietz, S. Fairhurst, N. Fotopoulos, G. González, C. Hanna, I. W. Harry, G. Jones, D. Keppel, D. J. A. McKechan, L. Pekowsky, S. Privitera, C. Robinson, A. C. Rodriguez, B. S. Sathyaprakash, A. S. Sengupta, M. Vallisneri, R. Vaulin, and A. J. Weinstein, Phys. Rev. D **87**, 024033 (2013), arXiv:1208.3491 [gr-qc].

[14] "Gstreamer plugins for the LSC Algorithm Library," .

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

[16] J. Aasi *et al.*, Classical and Quantum Gravity **32**, 115012 (2015), arXiv:1410.7764.

[17] E. Chassande-Mottin, in *AIP Conference Proceedings*, Vol. 1535 (2013) pp. 252–259, 1210.7173.

[18] C. Cutler and E. Flanagan, Phys. Rev. D **49**, 2658 (1994).

[19] L. S. Finn, Physical Review D **46**, 5236 (1992).

[20] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, Phys.Rev. **D85**, 122006 (2012), arXiv:gr-qc/0509116 [gr-qc].

[21] S. Babak, R. Biswas, P. Brady, D. Brown, K. Cannon, *et al.*, Phys.Rev. **D87**, 024033 (2013).

[22] B. Abbott *et al.* (LIGO Scientific Collaboration), Phys. Rev. D **80**, 047101 (2009).

[23] W. M. Farr, J. R. Gair, I. Mandel, and C. Cutler, Phys. Rev. D **91**, 023005 (2015), arXiv:1302.5341 [astro-ph.IM].

[24] C. Capano, "in preparation".

[25] M. Briggs *et al.* (LIGO Scientific Collaboration), (2012), arXiv:1205.2216 [astro-ph.HE].

[26] J. Abadie *et al.* (LIGO Collaboration, Virgo Collaboration), Phys.Rev. **D85**, 082002 (2012), arXiv:1111.7314 [gr-qc].

[27] M. Was, M.-A. Bizouard, V. Brisson, F. Cavalier, M. Davier, *et al.*, Class.Quant.Grav. **27**, 015005 (2010), arXiv:0906.2120 [gr-qc].

[28] D. Keppel, *Signatures and dynamics of compact binary coalescences and a search in LIGO's S5 data*, Ph.D. thesis, Caltech, Pasadena, CA (2009).

[29] C. Capano, *Searching for Gravitational Waves from Compact Binary Coalescence Using LIGO and Virgo Data*, Ph.D. thesis, Syracuse University (2011).

[30] B. Allen, Phys.Rev. **D71**, 062001 (2005), arXiv:gr-qc/0405045 [gr-qc].

[31] T. Dent *et al.*, (2012), in preparation.

Expt	Background parameters						
	a_0	a_1	a_2	a_3	a_4	a_5	a_6
1	-10.0000	-5	0	0	0	0	0
	-10.0000	-5	0	0	0	0	0
2	-7.2240	-0.32	0.53	-0.73	0.12	0.067	-0.018
	-7.2240	-0.32	0.53	-0.73	0.12	0.067	-0.018
3	-7.8000	-3.9	0	0	0	0	0
	-7.8000	-3.9	0	0	0	0	0
4	-6.2800	-0.3	0.5	-0.7	0.1	0.07	-0.02
	-4.8800	-1	0	-0.6	0	-0.04	-0.05
5	-8.4000	-4.2	0	0	0	0	0
	-8.0000	-4	0	0	0	0	0
6	-8.0704	-3	0.8	0.01	-0.05	0.007	-0.0004
	-8.0704	-3	0.8	0.01	-0.05	0.007	-0.0004
7	-9.1072	-4	0.7	0.09	-0.05	0.005	-0.0002
	-9.6200	-5.55	-0.37	0	0	0	0
8	-3.5040	-1.4	0	-0.16	0	-0.034	-0.026
	-4.6400	-2	0	-0.2	0	-0.03	-0.03
9	-7.0000	-3.5	0	0	0	0	0
	-6.6000	-3.3	0	0	0	0	0
10	-2.4800	-1	0	-0.1	0	-0.03	-0.02
	-5.8400	-3	0	-0.1	0	-0.03	-0.03
11	-4.0800	-1	0	-0.3	0	-0.05	-0.03
	-8.3200	-4	0	-0.1	0	-0.035	-0.025
12	-3.5040	-1.4	0	-0.16	0	-0.034	-0.026
	-4.6400	-2	0	-0.2	0	-0.03	-0.03
13	-7.8000	-3.9	0	0	0	0	0
	-7.8000	-3.9	0	0	0	0	0
14	-6.2800	-0.3	0.5	-0.7	0.1	0.07	-0.02
	-4.8800	-1	0	-0.6	0	-0.04	-0.05

TABLE II: Parameters of our background model distributions.

Expt	IFO	$C_{\rm SP}$	b	$\rho_{\rm SP}$
1	1	1e-10	-5.0000	10.1052
1	2	1e-10	-5.0000	10.1052
2	1	1e-4	-1.2690	9.9447
2	2	1e-4	-1.2690	9.9447
3	1	1e-10	-3.9000	11.4041
3	2	1e-10	-3.9000	11.4041
4	1	1e-4	-3.0350	10.0791
4	2	5e-4	-5.1906	8.7474
5	1	1e-10	-4.2000	10.9823
5	2	1e-10	-4.0000	11.2565
6	1	1e-9	-6.8668	13.0625
6	2	1e-9	-6.8668	13.0625
7	1	1e-9	-4.3611	12.9193
7	2	1e-9	-6.8728	9.2876
8	1	2e-2	-1.4415	7.7886
8	2	5e-3	-2.0660	7.8256
9	1	1e-9	-3.5000	11.4209
9	2	1e-9	-3.3000	11.7798
10	1	5e-2	-1.0889	8.0018
10	2	1e-3	-3.0410	7.8544
11	1	3e-4	-7.1603	9.1251
11	2	1e-5	-4.2931	8.2823
12	1	2e-2	-1.4415	7.7886
12	2	5e-3	-2.0660	7.8256
13	1	1e-9	-3.9000	10.8137
13	2	1e-9	-3.9000	10.8137
14	1	1e-4	-3.0350	10.0791
14	2	5e-4	-5.1906	8.7474

TABLE III: Parameters of background distributions (tail).

Appendix A: Parameters of simulated trigger distributions

In this section, we list the parameters used to define the background distributions. Recall that we adopt a form of SNR distribution for the background triggers given by Eq. 1 using input polynomial coefficients a_i as listed in Table II for all 14 experiments.

For the tails of the CDFs, the form is changed to a simpler representation as defined in Eq. 1 in order to make sure that the background distribution is well behaved as the SNR rises. The corresponding parameters b, $C_{\rm SP}$, and $\rho_{\rm SP}$ are listed in Table III. Notice that here the actual control parameter is $C_{\rm SP}$, while b and $\rho_{\rm SP}$ are derived values which could be subject to round-off error.

The rate of both background triggers and foreground triggers are controlled by parameters listed in Table IV. Here n is the predicted average coincidence number in one realisation, while the measured n is the actual value concluded from the data, their consistency reflect our confidence in the generation of the mock data. The *AstroRate (all)* is the expected rate for astrophysical foreground events in each realisation, but as only a fraction of them have large enough SNR to be detectable, thus the *AstroRate (loud)* represents such detectable event rate.

Expt	λ_1	λ_2	n	measured n	AstroRate (loud)	AstroRate (all)
1	10500	10500	11.025	11.0098	0	0
2	11500	11500	13.225	13.2453	0.001	0.0022
3	9900	9900	9.801	9.7874	0	0
4	12000	9000	10.8	10.8023	2.96	6.41
5	9800	10100	9.898	9.8857	2.74	5.94
6	8000	15000	12	12.0195	0.548	1.19
7	10300	9900	10.197	10.2206	0.0011	0.0024
8	10100	11100	11.211	11.2202	0.438	0.95
9	9700	10600	10.282	10.2785	0.11	0.24
10	12000	10800	12.96	12.9552	0.0001	0.0003
11	9800	10700	10.486	10.4938	3.07	6.65
12	10100	11100	11.211	11.2047	0	0
13	9900	9900	9.801	9.7814	0.022	0.048
14	12000	9000	10.8	10.7857	0	0

TABLE IV: Parameters for rates of both background and foreground triggers?

Appendix B: Additional results

1. CDF distribution of SNR

In this section, we show the reverse CDF distribution of the triggers' SNR. For one experiment, two detectors could have different background distributions, but they share the same astronomical foreground distribution. In figure 14 to 23, two detectors' background SNR distribution is demonstrated. Background for two detectors are represented by red and green line, while the foreground distribution is shown as the blue line, and combined distribution is the black line.

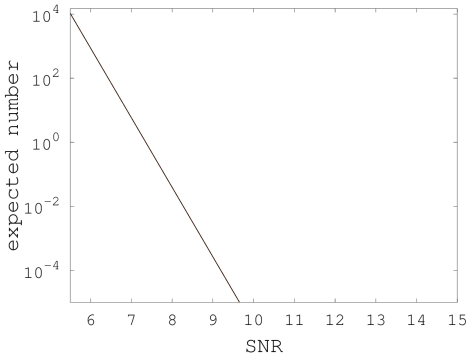


FIG. 14: Reverse CDF distribution of the triggers' SNR for experiment 1. The red and green curves represent the two individual detectors, while the blue curve represents the astronomical signals. The black lines represent the combined distribution of both background and foreground triggers.

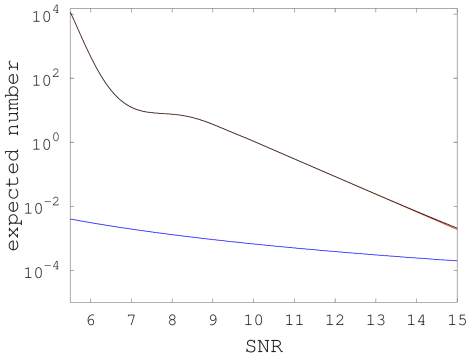


FIG. 15: Reverse CDF of the triggers' SNR for experiment 2: colours assigned as in Fig. 14.

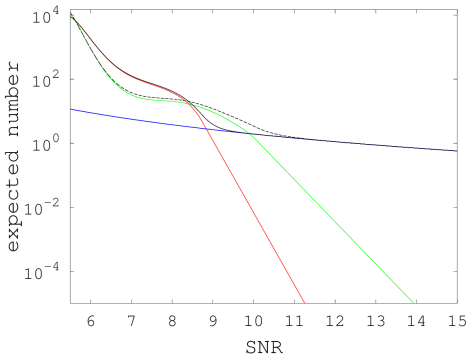


FIG. 16: Reverse CDF of the triggers' SNR for experiment 4: colours assigned as in Fig. 14.

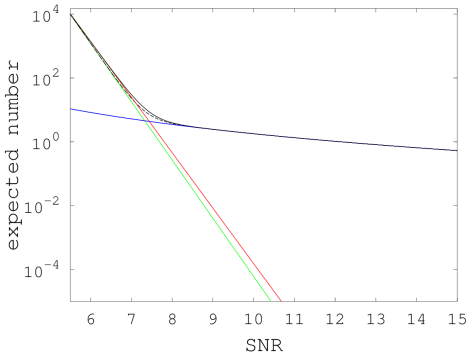


FIG. 17: Reverse CDF of the triggers' SNR for experiment 5: colours assigned as in Fig. 14.

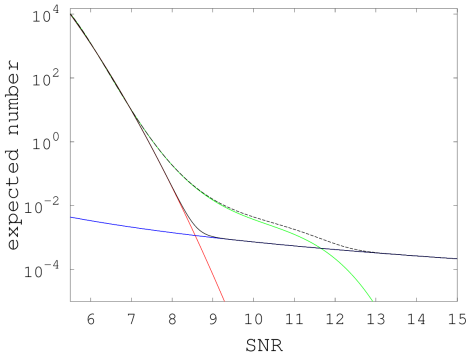


FIG. 18: Reverse CDF of the triggers' SNR for experiment 7: colours assigned as in Fig. 14.

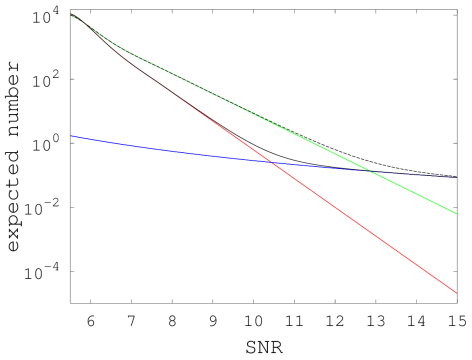


FIG. 19: Reverse CDF of the triggers' SNR for experiment 8: colours assigned as in Fig. 14.

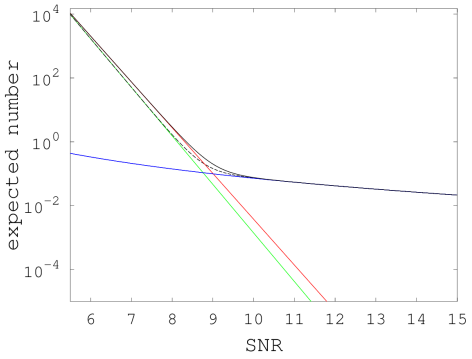


FIG. 20: Reverse CDF of the triggers' SNR for experiment 9: colours assigned as in Fig. 14.

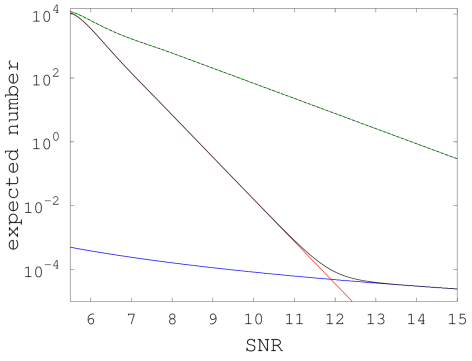


FIG. 21: Reverse CDF of the triggers' SNR for experiment 10: colours assigned as in Fig. 14.

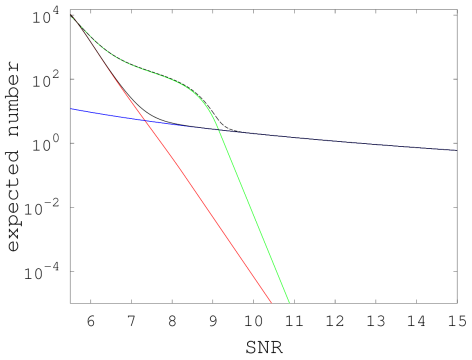


FIG. 22: Reverse CDF of the triggers' SNR for experiment 11: colours assigned as in Fig. 14.

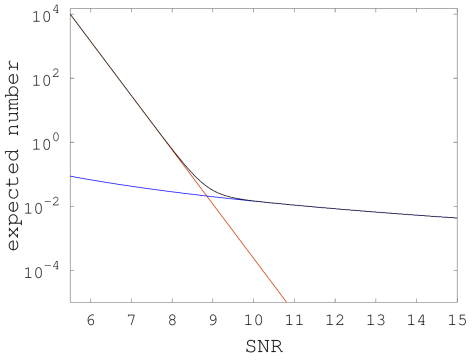
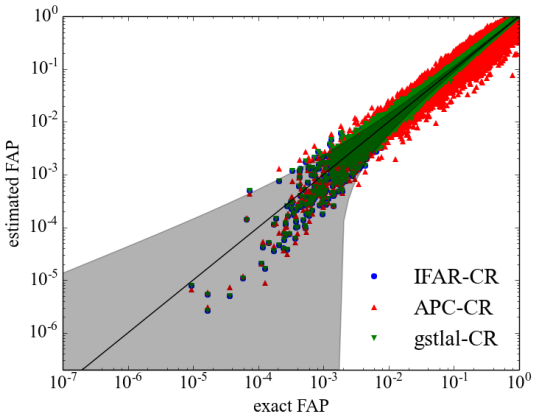


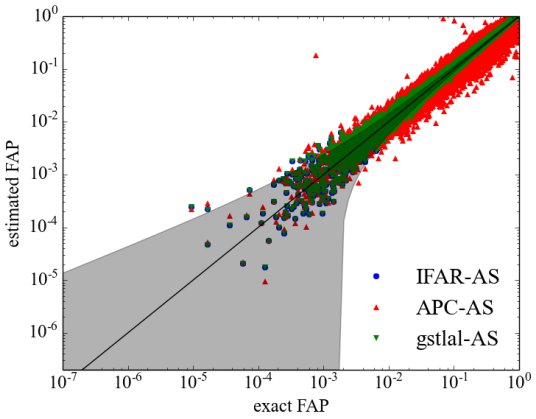
FIG. 23: Reverse CDF of the triggers' SNR for experiment 13: colours assigned as in Fig. 14.

2. Direct comparison

In Figs. 24, 25, 26, 27, 28, 29, 30, 31, 32, and 33 we present plots of the direct comparison between actual FAP and the FAP estimations from all methods.

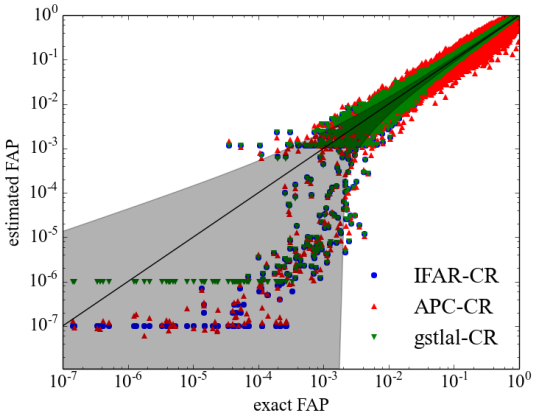


(a) direct comparison with ‘coincidence removal’

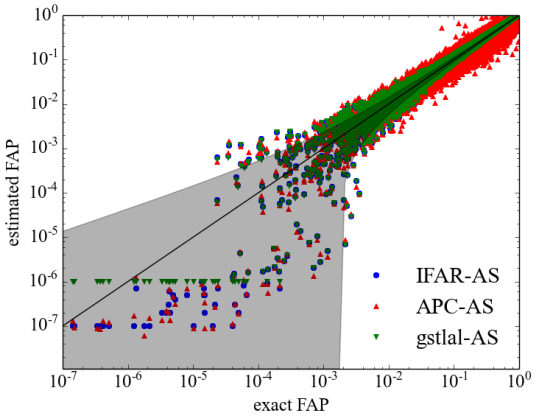


(b) direct comparison with ‘all samples’

FIG. 24: Direct comparisons on experiment 1.

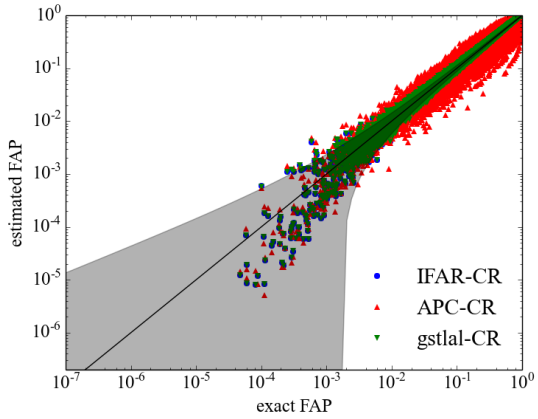


(a) direct comparison with ‘coincidence removal’

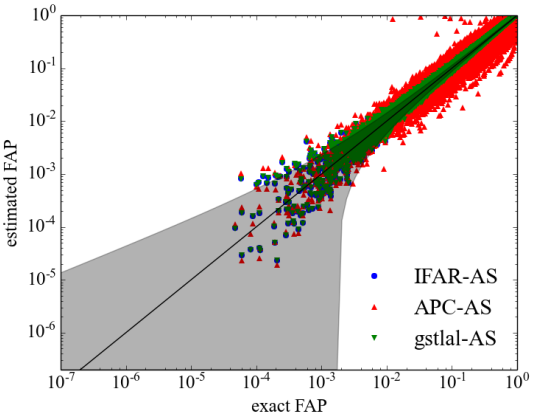


(b) direct comparison with ‘all samples’

FIG. 25: Direct comparisons on experiment 2.

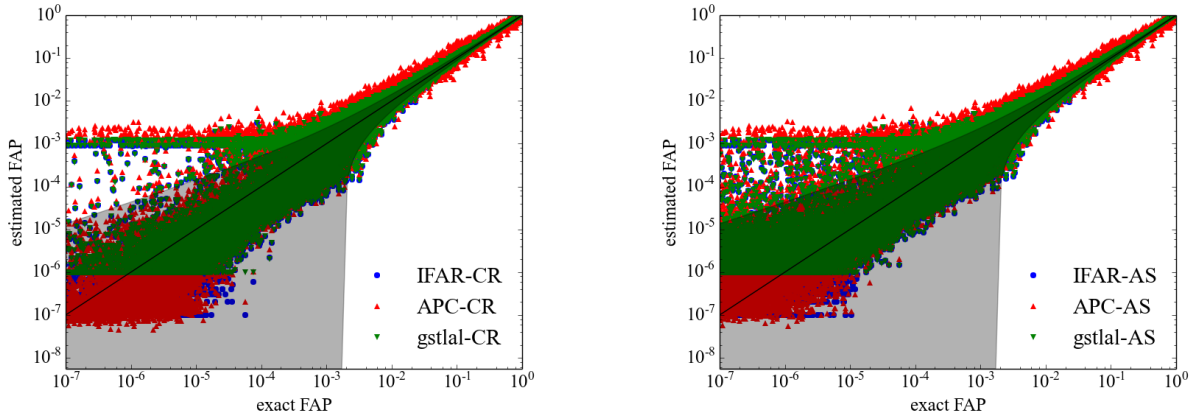


(a) direct comparison with ‘coincidence removal’



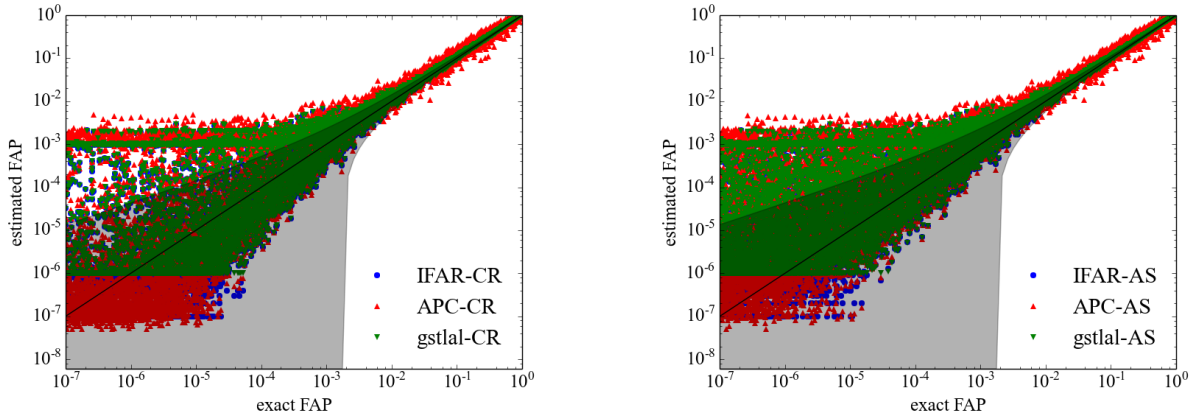
(b) direct comparison with ‘all samples’

FIG. 26: Direct comparisons on experiment 3.



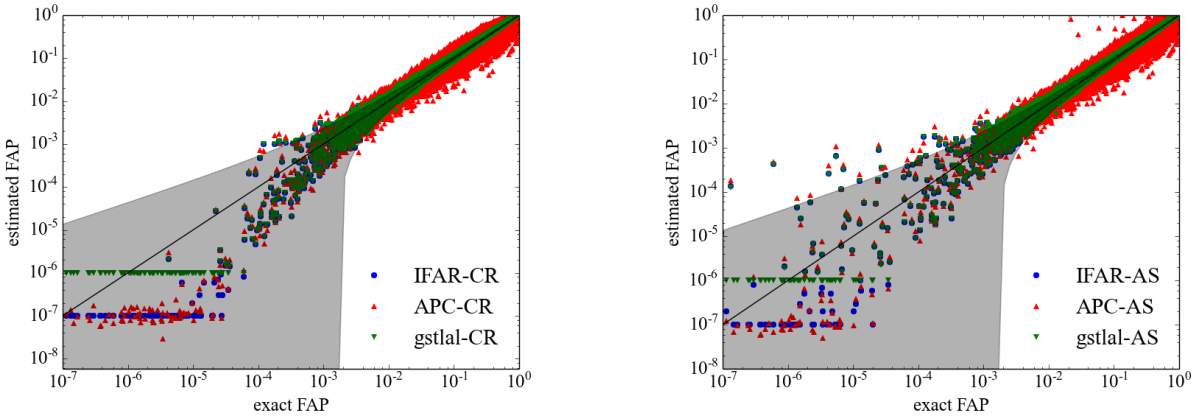
(a) direct comparison with 'coincidence removal' (b) direct comparison with 'all samples'

FIG. 27: Direct comparisons on experiment 4.



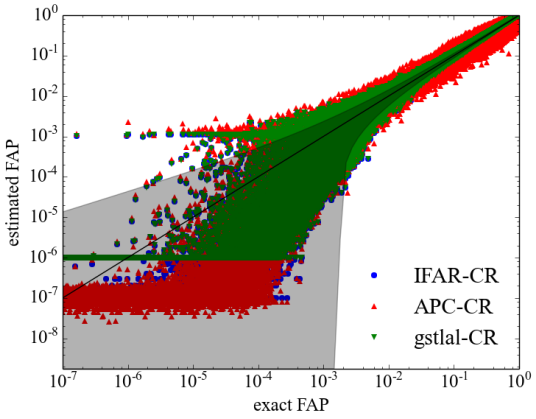
(a) direct comparison with removal (b) direct comparison with 'all samples'

FIG. 28: Direct comparisons on experiment 5.

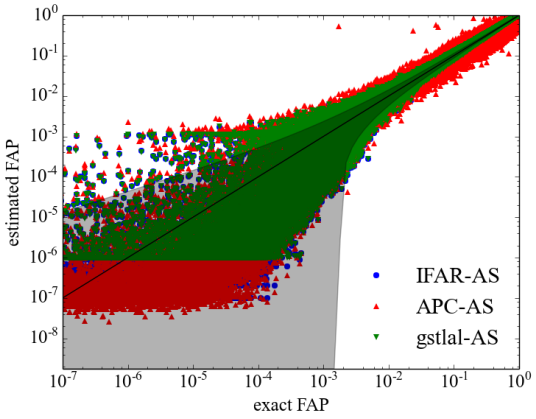


(a) direct comparison with 'coincidence removal' (b) direct comparison with 'all samples'

FIG. 29: Direct comparisons on experiment 7.

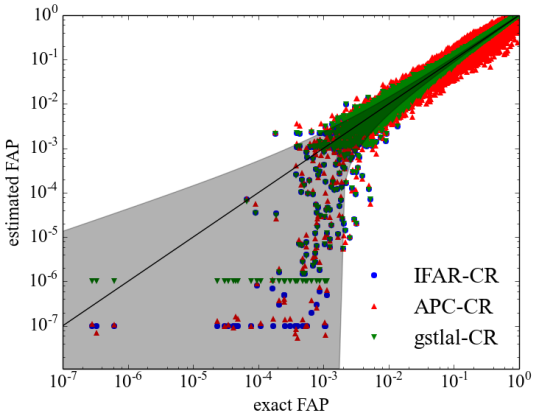


(a) direct comparison with ‘coincidence removal’

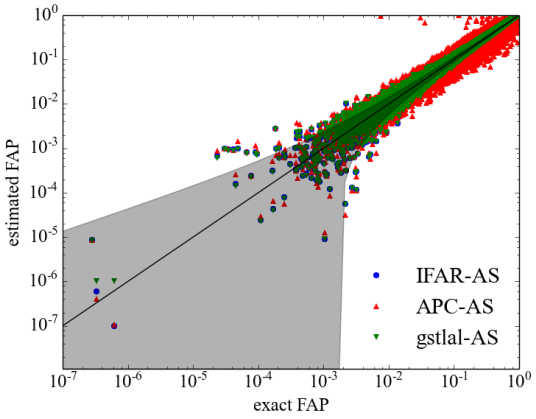


(b) direct comparison with ‘all samples’

FIG. 30: Direct comparisons on experiment 8.

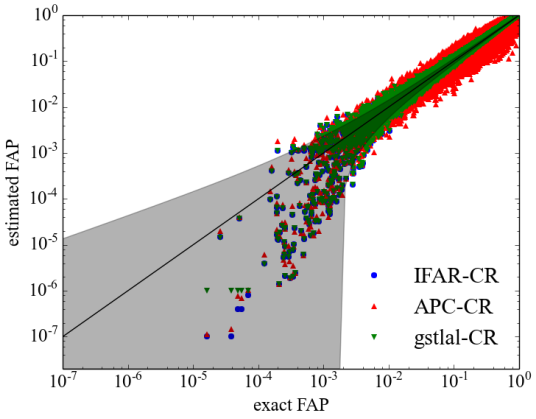


(a) direct comparison with ‘coincidence removal’

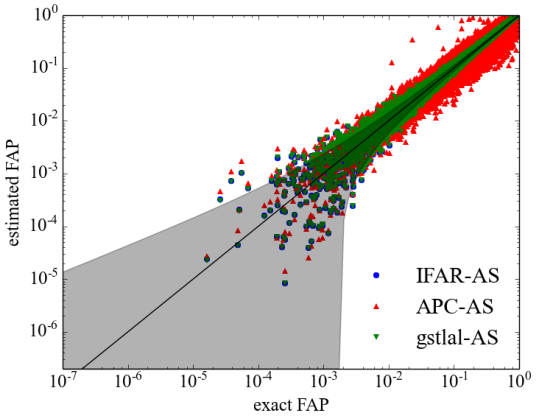


(b) direct comparison with ‘all samples’

FIG. 31: Direct comparisons on experiment 10.

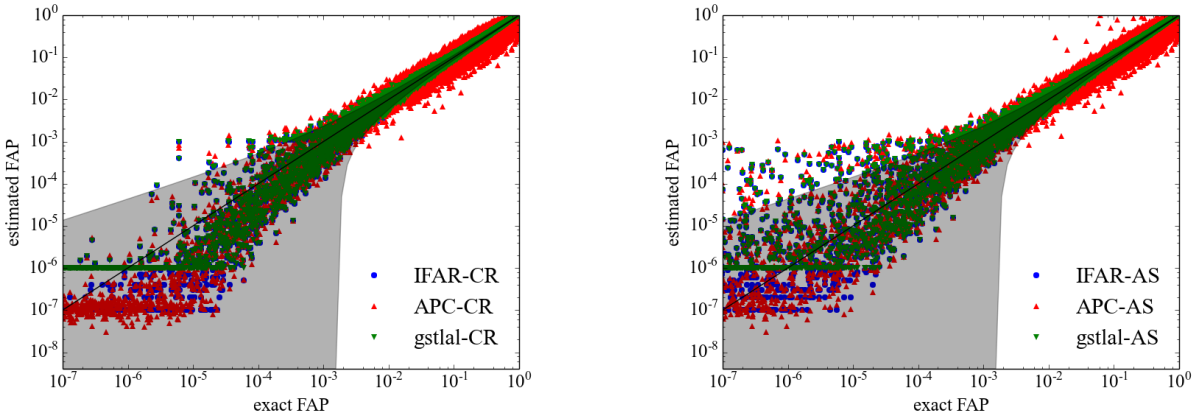


(a) direct comparison with ‘coincidence removal’



(b) direct comparison with ‘all samples’

FIG. 32: Direct comparisons on experiment 12.



(a) direct comparison with ‘coincidence removal’

(b) direct comparison with ‘all samples’

FIG. 33: Direct comparisons on experiment 13.

3. ROC plots

In this section we include all remaining ROC plots. Note that it is only possible to create a ROC plot when there are foreground events in the data, so we only show 7 ROC plots, Figs. 34, 35, 36, 37, 38, 39, and 40, complementing those already shown in Sec. IV D.

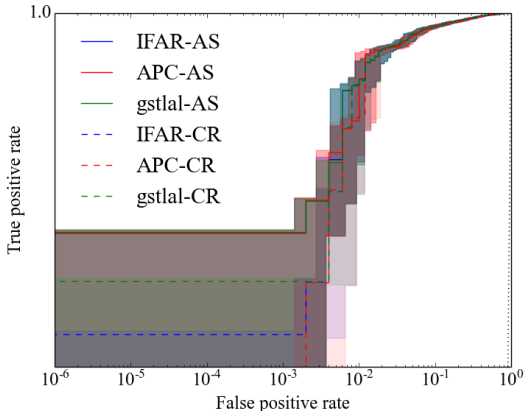


FIG. 34: ROC plot for experiment 5.

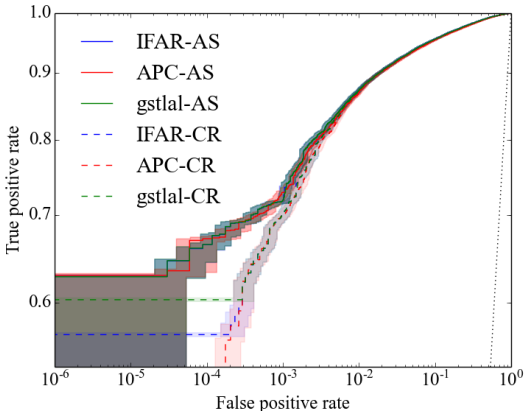


FIG. 35: ROC plot for experiment 6.

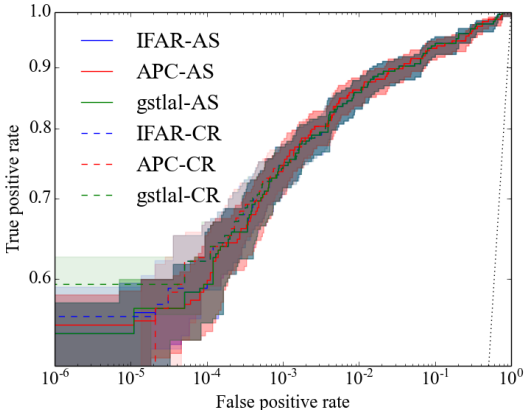


FIG. 36: ROC plot for experiment 7.

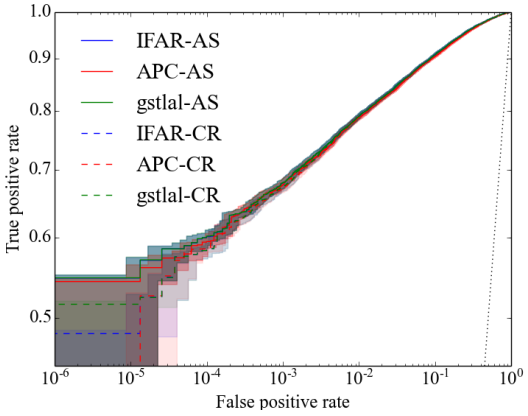


FIG. 37: ROC plot for experiment 9.

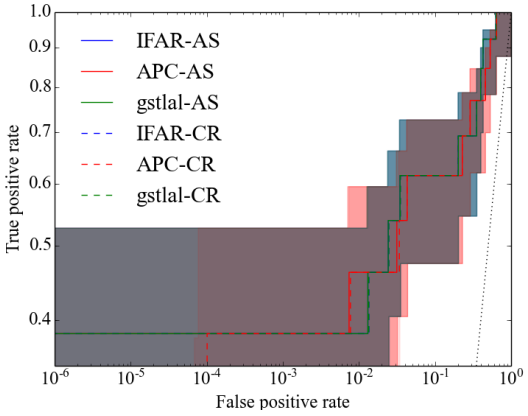


FIG. 38: ROC plot for experiment 10.

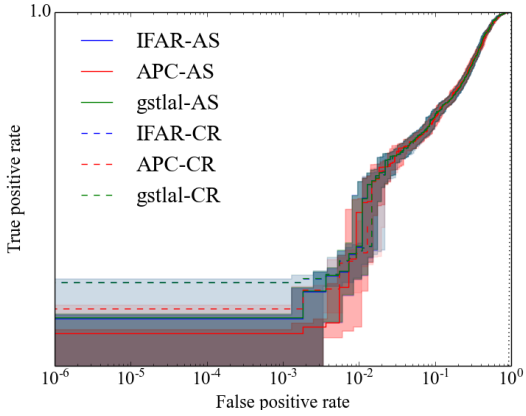


FIG. 39: ROC plot for experiment 11.

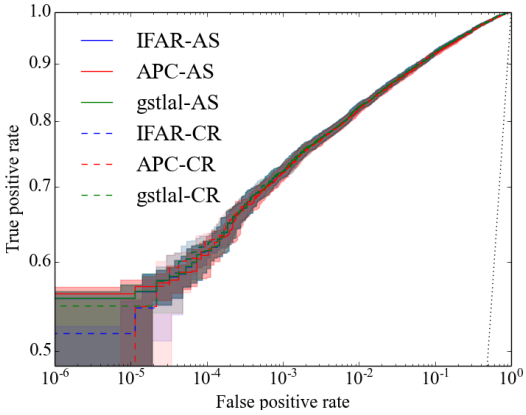


FIG. 40: ROC plot for experiment 13.