IREU Report

Calvin Firth*
School of Physics and Astronomy, University of Minnesota

Mentor: Simone Mastrogiovanni
Sapienza Università di Roma

(Dated: August 3, 2025)

We present our results from an H_0 inference using galaxy catalogs with publicly available skymaps from the Ligo/Virgo/Kagra collaboration's O4 observing run. We find no strong constraint on H_0. We also present a web application that performs this H_0 inference on gravitational wave events in real time, and give guidance on how to reliably perform this analysis.

I. INTRODUCTION

A. Standard Sirens

Since the first detection of gravitational waves (GWs) in 2015 by the Ligo/Virgo/Kagra (LVK) collaboration [1], one of the main astrophysical applications that has been pursued is measuring cosmological parameters—particularly the Hubble constant, H_0 [2]. The Hubble constant quantifies the current-day rate at which the universe is expanding. The physical relation describing the expansion of the universe is the Hubble-Lemaitre law:

$$v = H_0 d \tag{1}$$

where v is the recessional velocity of a distant galaxy, d is the distance to that galaxy, and H_0 is the aforementioned Hubble constant [3]. The Hubble constant is most conveniently expressed in units of $\frac{km/s}{Mpc}$. Because distant galaxies will have a recessional velocity due to the expansion of the universe, their spectra will have a redshift z. For small redshifts ($z << 1$), or equivalently, small distances, the Hubble-Lemaitre law can be rewritten (to first order) in terms of the redshift as

$$cz \approx H_0 d \tag{2}$$

where c is the speed of light [3]. For larger redshifts, model-dependent higher-order corrections are required.

This suggests a recipe for measuring the Hubble constant: measure both the distances and redshifts of distant galaxies. Typically, distances are measured using the so-called "cosmic distance ladder" [4] with electromagnetic (EM) observations. What makes gravitational waves useful for measuring the Hubble constant is that the event's luminosity distance d_L can be directly estimated from its waveform [5]. This makes gravitational waves an independent measure of cosmological distances—so-called "Standard Sirens".

There are various ways to measure a GW event's redshift. In the ideal case, the event has an EM counterpart, from which you can directly and precisely measure the redshift. However, out of the > 200 confirmed GW events, only one has a confirmed EM counterpart: GW170817 [6]. In the case where the event does not have an EM counterpart ("dark sirens"), one method to infer the event's redshift information is to use galaxy catalogs.

B. Galaxy catalog method

The basic idea of the galaxy catalog method is as follows:

From the GW event signal, a probability map of the event's sky location is produced (Figure 1). Next, each

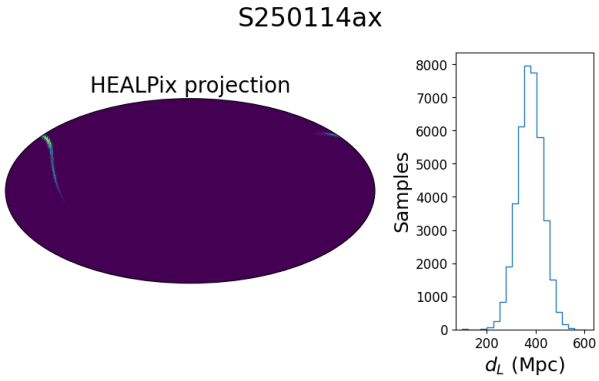

FIG. 1. An example of a sky-localization probability distribution and luminosity distance samples from the event S250114ax.

galaxy in the catalog within the sky localization area is assigned a likelihood to have hosted the GW event, essentially transforming the sky-localization probability distribution into a probability distribution over different galaxies. Each galaxy in the catalog has an associated redshift, so this distribution is equivalent to a probability distribution for the redshift of the GW event.

This redshift information is then combined with the luminosity distance estimate from the event to produce

* Communicating Author

a probability distribution $p(H_0|x_i)$ for H_0, where x_i is the information from a single gravitational wave event.

II. METHODS

For this analysis, we use the `ICAROGW` package in Python [7]. `ICAROGW` is a package designed for GW population inference. Before examining the methods in-depth, it will be useful to define Bayes' Theorem:

$$p(\theta|x) \propto p(x|\theta)p(\theta) \tag{3}$$

Bayes' Theorem applies in the case that you have some initial probability distribution and you want to update it according to some new evidence. Bayes' Theorem says that if you have a "prior" probability distribution $p(\theta)$ for some variable θ and some observation x, the correct way find the "posterior" probability distribution $p(\theta|x)$ is to multiply the prior distribution by the "likelihood function" $p(x|\theta)$.

For our purposes, `ICAROGW` takes as an input a skymap from a GW observation, which can be thought of as a posterior distribution $p(d_L, \Omega|x_i)$ for the luminosity distance d_L and the location on the sky Ω (in right ascension and declination) of the observed event. We then use `ICAROGW` to sample this distribution and create a pixelated HEALPix sky projection, as shown in Figure 1. In our case, we take 40,000 samples.

Next, we use `ICAROGW` to inspect the galaxy catalog at the pixels contained within the sky localization region. The host probability of each galaxy is weighted by its luminosity. Figure 2 shows the relative over/under-density of the weighted galaxy redshift distribution compared to a distribution that is uniform in co-moving volume. Regions with over-densities have statistical support from the galaxy catalog. This information can now be used to sample the likelihood function for H_0. A uniform distribution between 20 and 140 $\frac{km/s}{Mpc}$ is assumed as the prior $\pi(H_0)$ for H_0. We use `ICAROGW` to sample the likelihood function and compute the posterior distribution for H_0 according to Bayes' Theorem (Figure 3). For this analysis we took 25,000 samples.

It is important to note two things at this time: The first being the fact that in doing this analysis, assumptions are made about the astrophysical population of binary mergers, which significantly effect the resulting posterior distribution. Thus, the redshift information is not coming solely from the galaxy catalog, but is also informed by these assumptions. The second is that selection bias must be accounted for in this type of analysis. Due to the nature of GW detectors, not all GW events are equally likely to be detected, so there is a bias in which events are actually observed. To account for this selection bias, `ICAROGW` uses "injections", which are basically simulated GW waveforms, to determine which waveforms are detectable by current detectors [7].

The output product of our analysis is the individual event posterior $p(H_0|x_i)$. To combine multiple individual event posteriors to obtain a cumulative posterior $p(H_0|\{x\})$, we simply take product of the individual event posteriors and then normalize the result: $p(H_0|\{x\}) \propto \prod_i p(H_0|x_i)$.

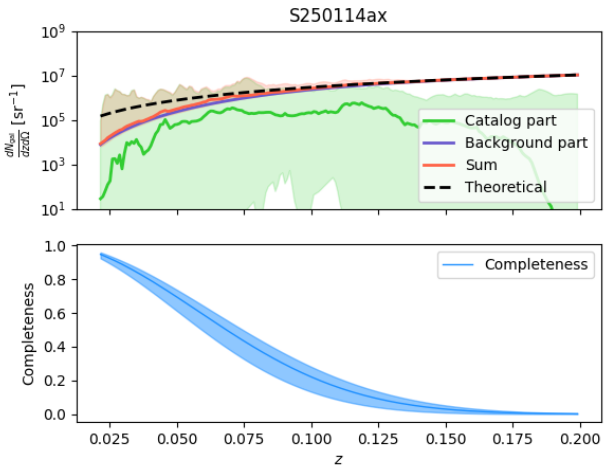

FIG. 2. Top: The number density of galaxies in the sky-localization area of S250114 as a function of redshift. The green curve is the number density in the catalog, while the blue is the number density expected to be missing from the galaxy catalog due to its finite sensitivity (see section II A). The blue curve is added to the green curve to correct for the missing galaxies, and the result is the red line. This curve is then compared to the theoretical number density of galaxies (black dashed line) to obtain statistical information about over and under-densities.
Bottom: The assumed completeness of the galaxy catalog as a function of redshift. This is the fraction of galaxies that are actually expected to be present in your galaxy catalog.

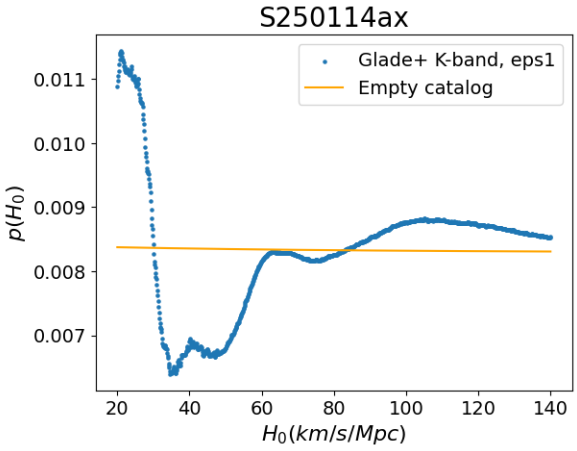

FIG. 3. An example H_0 posterior distribution from S250114ax. The orange line is the posterior produced when using an empty galaxy catalog. In that case, the posterior is informed only by the population assumptions, so it is plotted in order to ascertain which features in the posterior are actually coming from the galaxy catalog.

A. Glade+ Catalog

For this analysis, we use the GLADE+ galaxy catalog. GLADE+ is an all-sky catalog containing about 22 million galaxies compiled from various catalogs and surveys [8]. Each galaxy in the catalog has a measured redshift, and peculiar velocity corrections are implemented for galaxies with redshift $z < 0.05$. For this analysis, we exclusively use the infrared K-band of the catalog.

An intricacy previously unmentioned is the fact that the galaxy catalog is *incomplete*. In our analysis, we assume that the incompleteness is due to the fact that the surveys are limited in sensitivity, and thus can only detect sources of a certain brightness. To properly account for incompleteness in this analysis, one must impose a completeness correction. The first step is to define an "apparent magnitude threshold" representing the limiting sensitivity of the catalog for each pixel (Figure 4).

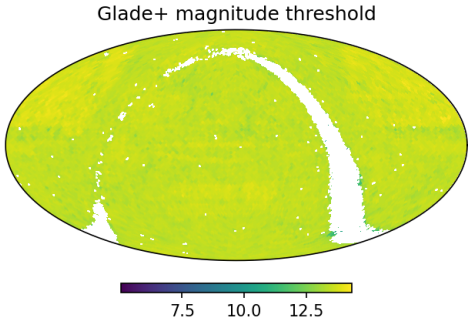

FIG. 4. A skymap showing the apparent magnitude threshold used for each pixel for the Glade+ galaxy catalog.

Per [7], ICAROGW defines the apparent magnitude threshold as the median apparent magnitude of the galaxies within the pixel. Galaxies fainter than the apparent magnitude threshold are then removed from the catalog. Using this apparent magnitude threshold, the completeness correction is applied by calculating the expected density of galaxies missing due to the threshold using a Schechter luminosity function.

III. LVK O4 EVENT ANALYSIS

We used the aforementioned techniques to analyze all confirmed GW events from the LVK collaboration's fourth operating run (O4) thus far. Out of the 193 events we used, ICAROGW was only able to complete the analysis for 161 events. We combined the posteriors using different criteria for which events are included. In particular, we used a "sky-localization cutoff", meaning we only include events whose 90% confidence sky-localization areas are below some threshold. On the page below we show the combined H_0 posterior for different sky-localization cutoffs (Figure 5).

A. S250207bg

The event with the most significant H_0 out of those we analyzed is S250207bg. It is a very well-localized and nearby event, and thus is in a region of the galaxy catalog with high completeness and notable over and underdensities. Below we show plots containing information about this event, including its H_0 posterior.

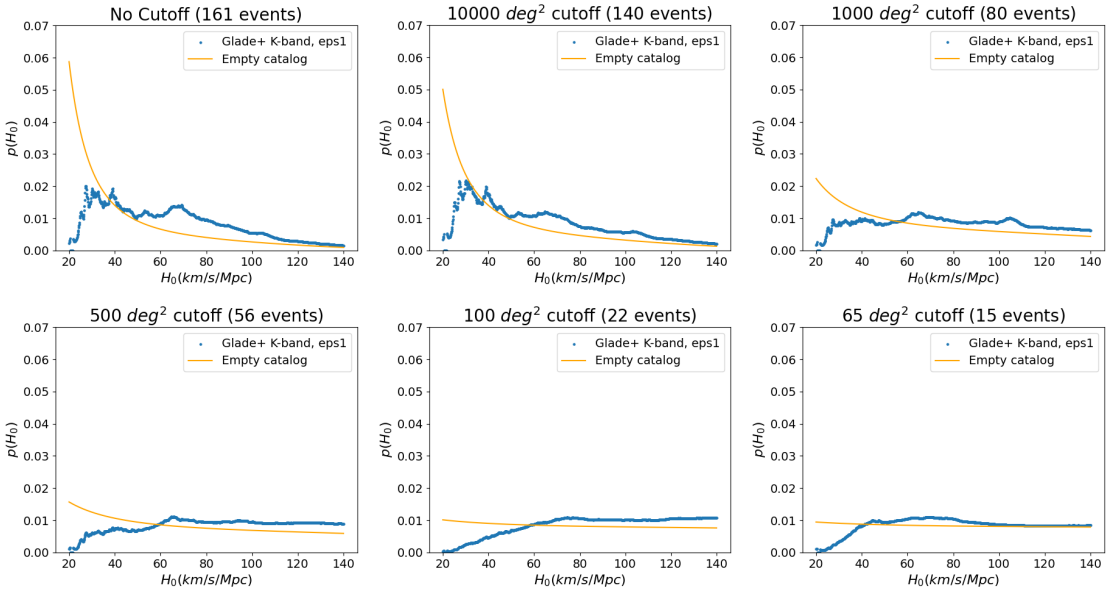

FIG. 5. The combined posteriors with different sky localization cutoffs. The blue curves are the combined posteriors when using the Glade+ catalog for each event, while the orange curves are the combined posteriors when using an empty catalog for each event.

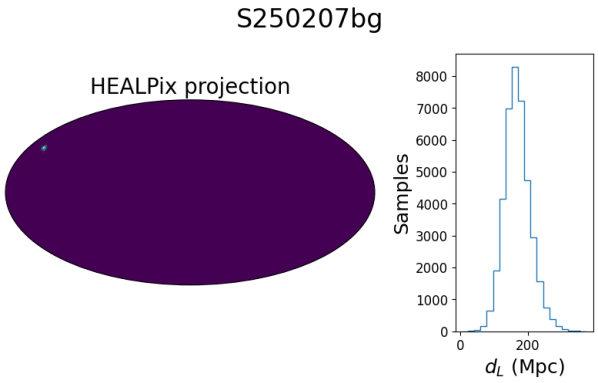

FIG. 6. The sky-localization map and luminosity distance samples for S250207bg. S250207bg has a 90% sky-localization area of 19.2° and a d_L of 180 ± 38 Mpc.

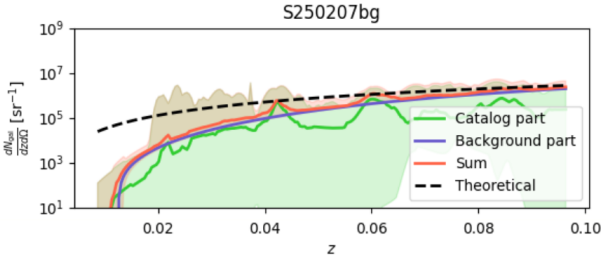

FIG. 7. The number density of galaxies in the S250207bg localization region as a function of redshift. Note the significant density peaks around $z = 0.02, 0.04$, and 0.06.

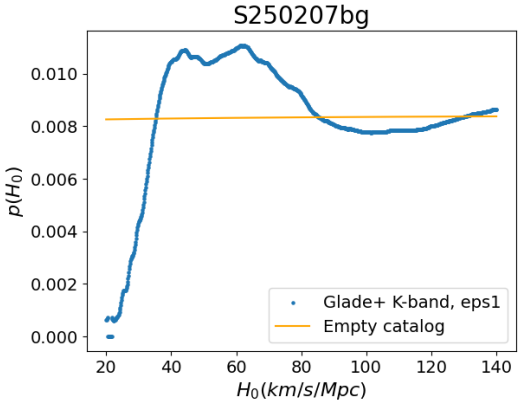

FIG. 8. The H_0 posterior for S250207bg. Note that this event's posterior provides significant support for values of H_0 in the range $35 - 80 \frac{km/s}{Mpc}$ when using the Glade+ catalog.

IV. INVESTIGATION OF SYSTEMATIC EFFECTS

Because many individual event posteriors need to be combined to produce precise measurements of H_0 using this method, even small systematic effects can accumulate into large errors in the resulting posterior. Thus, it is especially important to identify and mitigate systematic effects in the analysis. In this section, we discuss our studies of various systematic effects, motivated by our analysis of LVK O4 events.

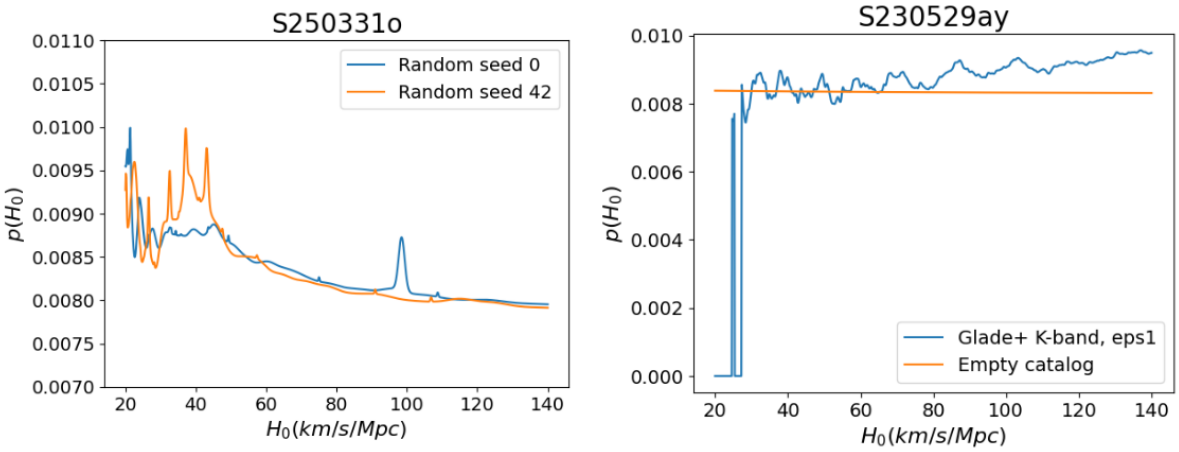

FIG. 9. Left: An example of an event with a random fluctuation in its H_0 posterior. Note the peak at $H_0 \approx 100 \frac{km/s}{Mpc}$ in the blue curve that isn't present in the orange curve. The only difference between the two analyses is the random seed used. Right: An example of a posterior with fluctuations to 0 at low values of H_0.

A. Random Fluctuations

One concerning effect observed during our analysis of O4 events was that changing the random seed used for the analysis could lead to the appearance and disappearance of structures within an event's H_0 posterior. Additionally, some events had random fluctuations that would set the posterior distribution $p(H_0|x_i)$ to 0 at certain values of H_0 (Figure 9). This is especially problematic when combining posteriors, as only one event having a value of 0 would cause the entire combined posterior to be 0 at that point when taking the product of the individual posteriors.

1. *NeffPE*

We suspected that some of the random 0 values in the posteriors were due to a parameter called "NeffPE". NeffPE is a parameter that `ICAROGW` takes as input when sampling the likelihood function $p(x_i|H_0)$. NeffPE is meant to be numerical stability threshold that represents the minimum number of effective parameter estimation samples. If the stability of the likelihood function at a particular value of H_0 is below the NeffPE threshold, `ICAROGW` returns 0 as the value of the likelihood function there, meaning that that value of H_0 will be 0 in the posterior. To test this, we set NeffPE to -1 when doing the analysis, meaning that the numerical stability of the likelihood would never fall below the NeffPE threshold, as it cannot be negative. Thus, if the 0 values are being caused by this threshold, they should not appear when NeffPE is -1. This is exactly what we saw. When setting NeffPE to -1, there were no anomalous 0 values in individual event posteriors, as an example shows in Figure 10.

However, it is not advisable to set NeffPE to -1. For this analysis we set NeffPE to 10, a value commonly used in this type of analysis.

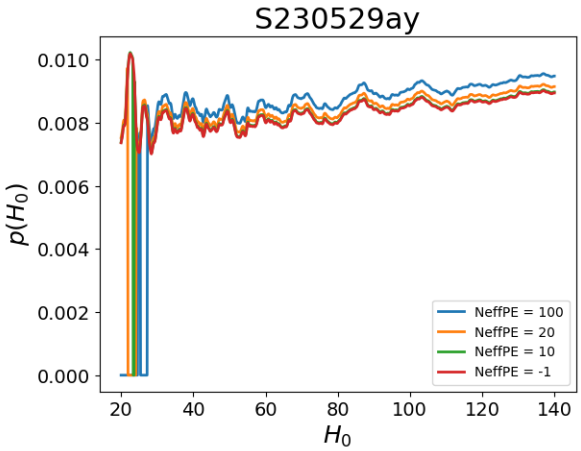

FIG. 10. The posteriors produced for the same event with different values of NeffPE. Note that as NeffPE decreases, so does the incidence of zero values in the posterior, and when NeffPE is set to -1, there are no zeros in the posterior.

2. *Number of samples*

We also investigated the effect of the number of samples taken from the skymap and the likelihood function on these random fluctuations. To do this, we selected 10 random events from the O4 observing run, and compared the random fluctuations in their posteriors for multiple different random seeds while varying the number of samples. We found that the number of samples taken from the skymap did not have a significant effect on the H_0 posterior, so we kept that value constant at 40,000 sam-

ples throughout this test. For each of the 10 events, we tested values of 5,000; 10,000; 15,000; 20,000; 25,000; and 30,000 for the number of likelihood samples taken (a parameter called "Nparallel" in `ICAROGW`). For each value of Nparallel, we used 5 unique random seeds to perform the analysis, so there were 5 posteriors per event for each value of Nparallel. We then compared the deviations between the 5 posteriors as a function of number of Nparallel. For each event and value of Nparallel, we calculated the standard deviation of the 5 posteriors at each value of H_0, and then took the average of those standard deviations (see Figure 11 for clarification). This "average

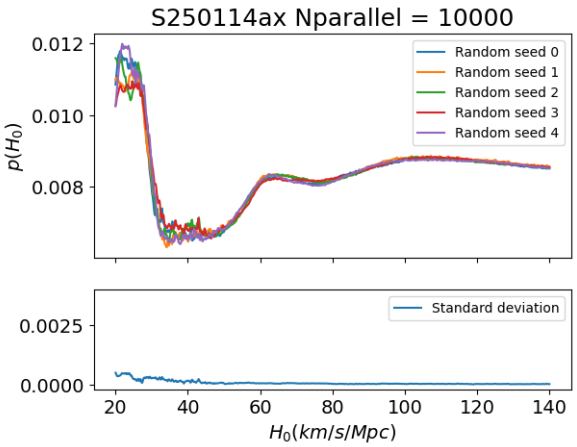

FIG. 11. An example of the random fluctuation tests done with Nparallel. The top panel shows the posteriors produced by five distinct random seeds for a particular event and a particular value of Nparallel. The bottom panel shows the standard deviation between the values of the five posteriors at each value of H_0. The "average standard deviation" is simply the average of this curve in the bottom panel.

standard deviation" gives a measure of the magnitude of the random fluctuations. Figure 12 shows this average standard deviation plotted against the number of likelihood samples for all 10 events.

To illustrate the effect of Nparallel, Figure 13 shows the posteriors of two events at an Nparallel of 5,000 and 30,000. These particular events were chosen because one (S250207bg) is very well localized on the sky (90% confidence area around 20 square degrees) while the other (S230529ay) is poorly localized (90% confidence area $> 20,000$ square degrees).

B. Sky localization/distance cutoffs

To test the effects of these random fluctuations on the combined posterior of many events, we repeated the analysis of the same 161 O4 events analyzed in section III using two different random seeds. We then compare the combined posteriors produced by the two different random seeds. As in section III, we impose different sky-localization cutoffs and observe the effects (Figure 14).

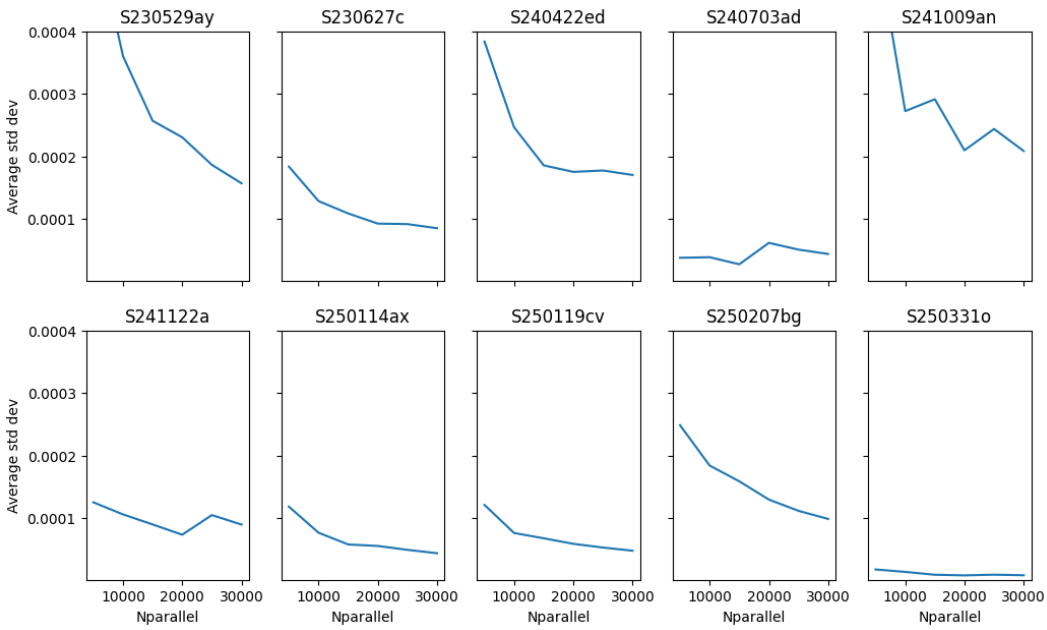

FIG. 12. The average standard deviation (as defined in Figure 11) as a function of Nparallel for all 10 events tested. Note that the average standard deviation almost exclusively decreases as Nparallel is increased.

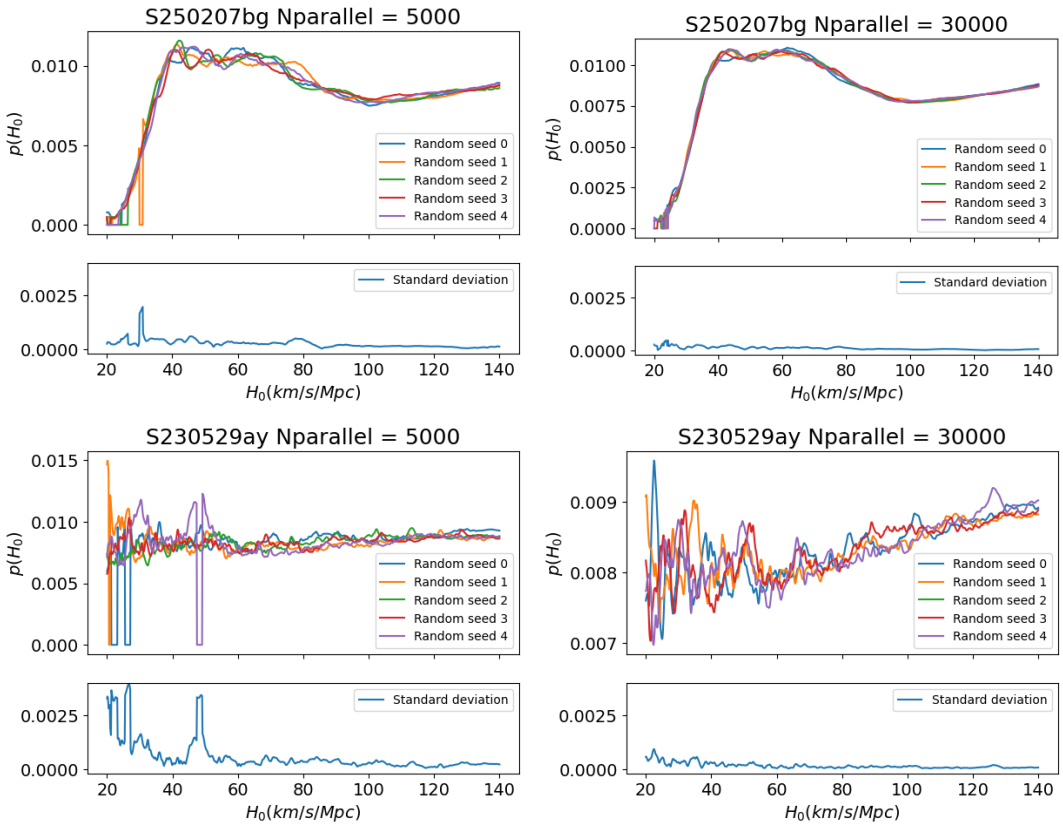

FIG. 13. The random fluctuation tests at an Nparallel of 5,000 for two distinct events (S250207bg and S230529ay). The top two figures correspond to S250207bg, while the bottom two refer to S230529ay. S250207bg is very well localized ($\approx$ 20 square degrees 90% sky-localization area) while S230529ay is poorly localized ($>$ 20,000 square degrees 90% area). For Nparallel = 5,000 both events have 0 values in their posteriors, and relatively large random fluctuations, especially S230529ay. For Nparallel = 30,000 the anomalous 0 values are eliminated for both events. S250207bg has very small random fluctuations, while S230529ay has slightly larger fluctuations, but still significantly smaller than the Nparallel = 5,000 case.

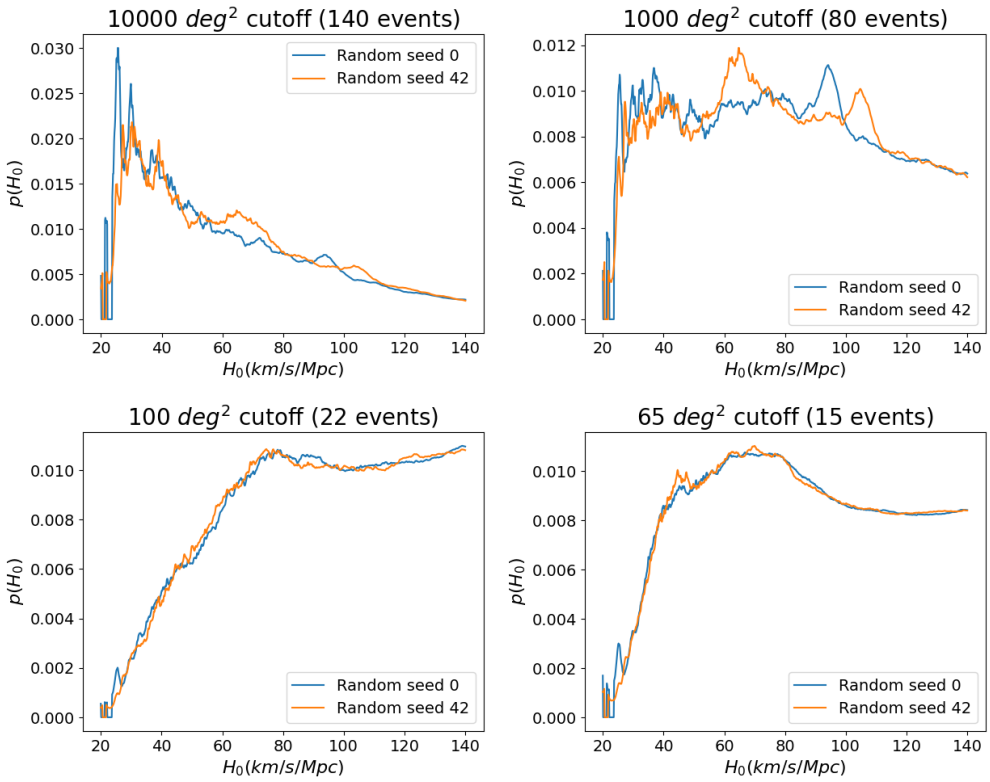

FIG. 14. A comparison of the combined posteriors produced by different random seeds for different sky-localization cutoffs. The magnitude of the fluctuations generally decrease as the cutoffs are tightened and fewer events are included.

To test whether the decrease in the magnitude of fluctuations as the cutoffs were tightened was solely due to including fewer events, or if the small sky localization also contributed, we compared the posteriors of the 16 worst localized events (Figure 15).

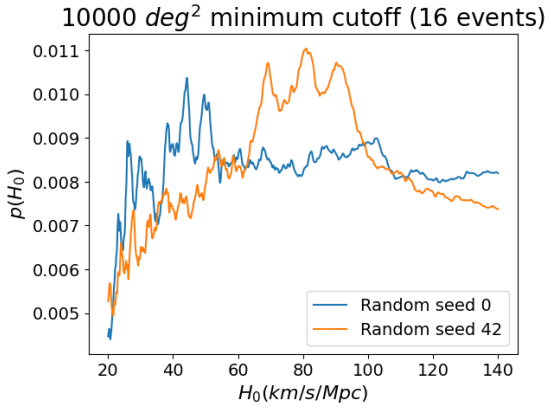

FIG. 15. The combined posterior produced by two different random seeds including the 16 worst localized events (10,000 square degree *minimum* cutoff) events. The two posteriors vary significantly from one another.

Additionally, we performed a test where we combined the posteriors of the 16 most distant events (Figure 16).

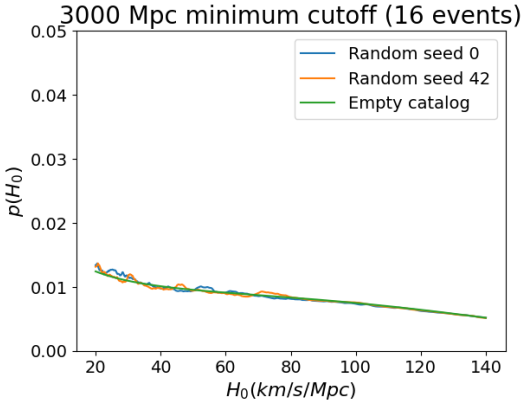

FIG. 16. The combined posterior produced by two different random seeds including the 16 most distant ($> 3,000$ Mpc) events, as well as that produced by using an empty galaxy catalog. All three curves are very similar.

V. LIVE ANALYSIS APPLICATION

We also briefly present a prototype version of a web application to perform galaxy catalog H_0 inference in real-time as gravitational wave events are observed (Link). The application consists of two parts—a background analysis program and a user-interface. The background program runs continuously and searches for public GW event notices. When a notice is received, the program performs the analysis and writes the results to an external database. The UI is produced using Streamlit, a Python package used for creating web applications. When a user accesses this Streamlit application, it reads the analysis results from the aforementioned database and displays various figures, including the H_0 posterior. The application automatically refreshes and checks for updates from the background program.

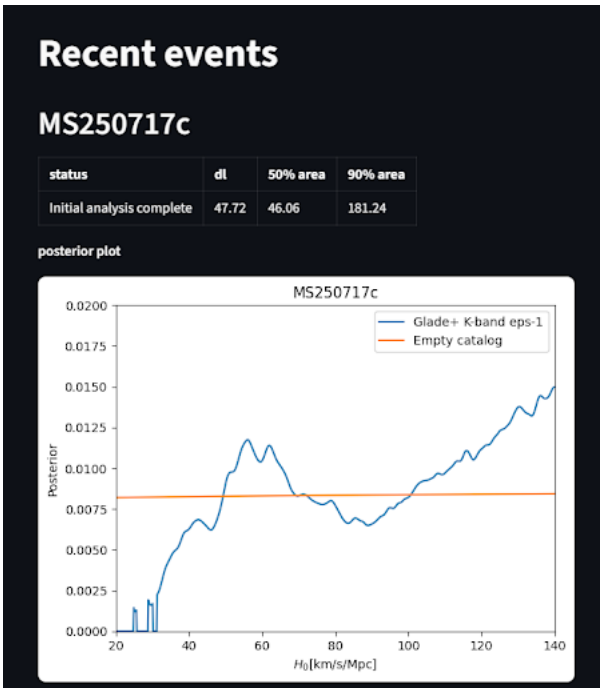

FIG. 17. An image of the user interface of the live analysis app. This page displays the analysis results of the most recent events. The event displayed in this image is an event from a mock alert from LVK.

VI. DISCUSSION

A. O4 Event Analysis

The posteriors of most O4 events were dominated by population assumptions. This can be seen in the fact that the combined posteriors (Figure 5) do not differ significantly from those inferred using an empty galaxy catalog. However, there is some support for values near the "H_0 tension" area around $70\frac{km/s}{Mpc}$. This becomes particularly evident as the sky-localization threshold is decreased. For the 65 deg^2 cutoff posterior, we find that most of the structure is coming from the aforementioned event S250207bg (section III A).

B. Systematic Effects

We find that random fluctuations to zero in H_0 posteriors cannot be completely eliminated unless NeffPE is set to -1, as shown in Figure 10. However, with a low value of NeffPE (10), increasing Nparallel has the effect of mitigating random zeros. Additionally, increasing Nparallel nearly universally decreases the amplitude of random fluctuations (Figure 12). In most of the analyzed events, there are significant improvements in increasing Nparallel up to around 20,000; and then there are diminishing returns as Nparallel is increased further. Increasing Nparallel lengthens the computation time, so one must optimize the value of Nparallel considering both random fluctuations and computation time. We find 25,000 to be a good balance.

The magnitude of the random fluctuations are not consistent between all events, as shown in Figure 13. The much better localized event S250207bg has very small fluctuations, while the poorly localized event S230529ay has larger fluctuations. We also observe this trend when combining multiple posteriors. When including only the 15 best localized events, there are very small random fluctuations between different random seeds, but when including only the 16 worst localized events, the random fluctuations are much larger (Figure 14). Thus, it is advisable to include more well localized events than poorly localized ones when combining posteriors. Furthermore, we find that events at large distances are less prone to random fluctuations and provide very little H_0 information. This is because the galaxy catalog is largely incomplete at large distances, so the population assumptions dominate the H_0 posterior.

C. Live Analysis Application

The app is not yet complete, however, the main features (analyzing events in real-time and displaying the results) are functional. The user interface is currently quite minimal, and will likely be expanded before full deployment.

The current database being used to store the analysis results only allows for 500,000 read/write commands per month, which would likely be too little for this app if it is running continuously. Prior to full deployment, a more permanent solution must be found.

Thus far, the background analysis program has been run on a typical laptop. This demonstrates its low computational needs, but would best be deployed on a cluster, as it is intended to run indefinitely. The front-end

Streamlit page is run on a cloud-based server. The complete code is contained in a Git repository:

VII. CONCLUSIONS

To conclude, we presented a web application that performs real-time galaxy catalog H_0 inference with galaxy catalogs. Additionally, we present our studies done to ensure that this low-latency analysis is robust and consistent.

From our studies, we can give suggestions on procedures to take to ensure the reliability of this type of analysis:

1. If possible, use a lower value of NeffPE to avoid potential fluctuations to zero in the H_0 posterior. We find 10 to be a suitable value. However, we do not recommend reducing NeffPE any further, as maintaining numerical stability is important.

2. To mitigate random fluctuations in the H_0 posterior, use a large value for Nparallel (the number of samples from the likelihood function). Accordingly, one must ensure to take enough skymap samples to accommodate the number of likelihood samples. We find that there are diminishing returns to increasing Nparallel beginning around a value of 20,000-30,000. For this reason, we use a value of 25,000. However, a higher Nparallel is always preferable for mitigating random fluctuations, but increases the computational requirements.

3. Consider instituting a sky-localization cutoff when combining multiple posteriors, as events with larger sky-localization areas are more prone to random fluctuations. Additionally, keep in mind that the posteriors of events at large distances have very little contribution from the galaxy catalog and are dominated by population assumptions. However, we caution against removing these events when combining posteriors, as this may introduce a selection bias that is not currently accounted for.

ACKNOWLEDGMENTS

This research was conducted through the University of Florida International Research Experience for Undergraduates in Gravitational Physics program, NSF grant 2348913.

Thank you to all involved in the IREU program for making the experience enjoyable and successful.

Thank you to Dr. Simone Mastrogiovanni and group members for graciously hosting me.

[1] LIGO Scientific Collaboration and Virgo Collaboration, Observation of Gravitational Waves from a Binary Black Hole Merger, Physical Review Letters **116**, 061102 (2016).
[2] T. L. S. Collaboration, the Virgo Collaboration, and t. K. Collaboration, Constraints on the Cosmic Expansion History from GWTC–3, The Astrophysical Journal **949**, 76 (2023).
[3] E. Harrison, The Redshift-Distance and Velocity-Distance Laws, The Astrophysical Journal **403**, 28 (1993).
[4] A. G. Riess, S. Casertano, W. Yuan, L. M. Macri, and D. Scolnic, Large Magellanic Cloud Cepheid Standards Provide a 1% Foundation for the Determination of the Hubble Constant and Stronger Evidence for Physics beyond ΛCDM, The Astrophysical Journal **876**, 85 (2019).
[5] B. F. Schutz, Determining the Hubble constant from gravitational wave observations, Nature **323**, 310 (1986).
[6] LIGO Scientific Collaboration and Virgo Collaboration, GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, Physical Review Letters **119**, 161101 (2017).
[7] S. Mastrogiovanni, G. Pierra, S. Perriès, D. Laghi, G. C. Santoro, A. Ghosh, R. Gray, C. Karathanasis, and K. Leyde, ICAROGW: A python package for inference of astrophysical population properties of noisy, heterogeneous, and incomplete observations, Astronomy & Astrophysics **682**, A167 (2024).
[8] G. Dálya, R. Díaz, F. R. Bouchet, Z. Frei, J. Jasche, G. Lavaux, R. Macas, S. Mukherjee, M. Pálfi, R. S. de Souza, B. D. Wandelt, M. Bilicki, and P. Raffai, GLADE+ : An extended galaxy catalogue for multimessenger searches with advanced gravitational-wave detectors, Monthly Notices of the Royal Astronomical Society **514**, 1403 (2022).