LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY
- LIGO -
CALIFORNIA INSTITUTE OF TECHNOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LIGO-T2400305–v3 2024/10/27

O3 Data Release Request: Second-Trend Data from Seismometers, Wind Speed Monitors, and Accelerometers

Jonathan Richardson, Evangelos Papalexakis, Rutuja Gurav, Pooyan Goodarzi

California Institute of Technology
LIGO Project, MS 18-34
Pasadena, CA 91125
Phone (626) 395-2129
Fax (626) 304-9834
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Project, Room NW22-295
Cambridge, MA 02139
Phone (617) 253-4824
Fax (617) 253-7014
E-mail: info@ligo.mit.edu

LIGO Hanford Observatory
Route 10, Mile Marker 2
Richland, WA 99352
Phone (509) 372-8106
Fax (509) 372-8137
E-mail: info@ligo.caltech.edu

LIGO Livingston Observatory
19100 LIGO Lane
Livingston, LA 70754
Phone (225) 686-3100
Fax (225) 686-7189
E-mail: info@ligo.caltech.edu

http://www.ligo.caltech.edu/

1 Motivation and Science Case

Machine Learning (ML) and novel data analysis techniques are revolutionizing the way data is interpreted and utilized in today's scientific research. These technologies enable researchers to process vast amounts of data with unprecedented speed and accuracy, uncovering patterns that were previously indiscernible. Large-scale scientific instruments, such as the LIGO detectors, generate vast amounts of data. Certain data channels are directly utilized in scientific investigations (e.g., strain data for astrophysical analyses), while others are employed for monitoring, diagnosing, and controlling the instrument and its environment. Within LIGO, hundreds of thousands of auxiliary data channels are recorded continuously.

Our team at UC Riverside has developed a machine learning clustering tool designed to continuously identify the environmental "state" of each LIGO detector based on data obtained from auxiliary environmental channels. It was developed in coordination with site-based commissioners who, through experience, have developed the expertise to heuristically interpret many of these data streams (e.g., identify the "high microseism" or "high anthropogenic noise" states). Our model formalizes and extends these interpretations, by distilling information from a large number of heterogeneous sensors into a single state word that can be continuously recorded over time. This process is conceptually illustrated in Figure 1.

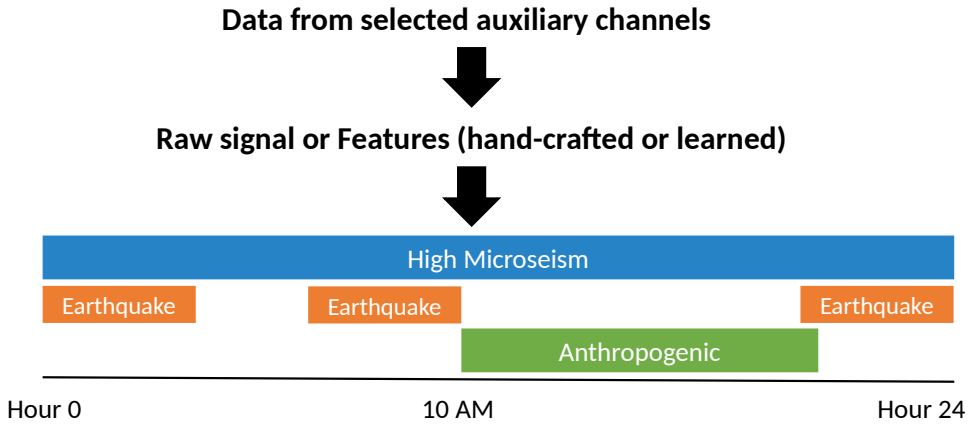

Figure 1: Conceptual overview of the ML approach. This unsupervised clustering model distills information from numerous heterogeneous sensors into a single environmental "state" which is continuously updated and recorded.

This model has the potential to become a powerful diagnostic commissioning tool for the LIGO detectors. The environmental "states" identified by the model can be correlated with events of interest in the detector's in-loop channels, such as periods of increased glitches or controls instabilities. As one example, Figure 2 shows the output of our clustering model when run on a subset of auxiliary channels sensitive to seismic motion during O3. The top three panels show the raw time series from a seismic-motion sensor with three different bandpass filters applied, whose passbands correspond to the characteristic frequencies of earthquakes, microseisms, and anthropogenic sources. The bottommost panel shows the resulting seismic states identified by the clustering model.

Not only does this *unsupervised* model recover strikingly similar states to those that commissioners heuristic identify, there is a strong correlation between the identified states and the rate of glitch occurrences in the main strain channel (the states are identified independently of the strain channel, using only auxiliary environmental sensor data). The rate of glitch occurrences is shown by the black trace in Figure 2. Although this example is limited to identifying well-understood seismic states, the clustering model can generalize to the identification of environmental states more broadly, which could include other, non-seismic sensors such as anemometers, microphones, and electronics noise monitors. Thus there is potential to discover previously unknown associations with environmental conditions, as well.

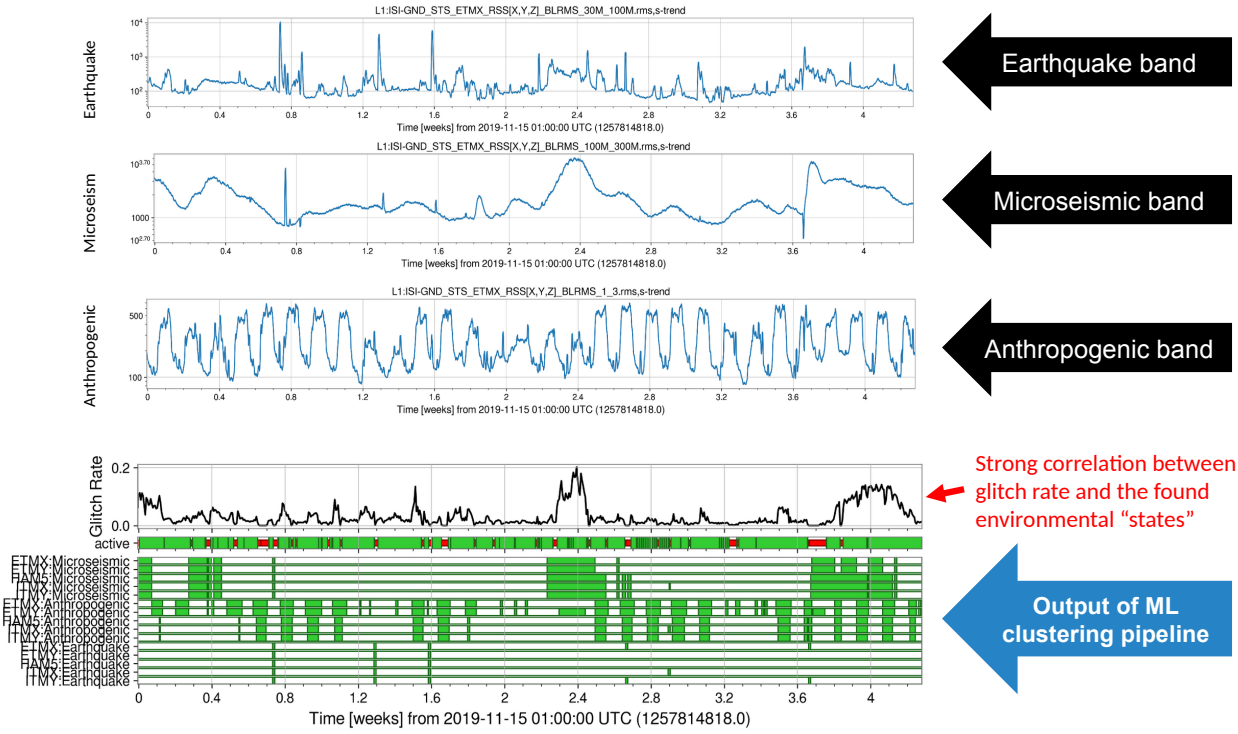

Figure 2: Example output of the ML clustering model. Using band-limited RMS (BLRMS) signals from seismic sensors (blue traces), the model identifies similar environmental states to those heuristically known to commissioners. Moreover, the identified states (green bars) have physical significance for interferometer performance, as evidenced by their clear association with periods of elevated glitch activity in the main strain channel (black trace).

Beyond the LVK Collaboration, we aim to present this work in the "Applied Data Science" tracks of machine learning conferences, where the most impactful research results typically include an accompanying public data release. In our case, releasing the underlying LIGO auxiliary channel data will not only enable independent verification of our model's performance, but will also engage the broader ML community by presenting a well-defined challenge to develop better models applicable to this class of problems. The release of additional O3 auxiliary channel data will thus contribute a valuable dataset, in its own right, towards the machine learning community's efforts of building multivariate time series analysis algorithms. In the following section, we specify the precise scope of our data release request.

2 Dataset Specifications

We are requesting that the RMS **second-trends** for a larger number of environmental sensor signals collected during O3 be released. The specific scope our request is as follows:

- **Interferometers:** H1 and L1.
- **Time intervals:** All ANALYSIS_READY segments during O3a and O3b.
- **Time resolution:** Second-trends.
- **Channels:** 402 in total. Table 1 contains links to machine-readable channel lists.
- **Data Volume:** 61 GB.
- **Timeline for release:** Prior to December 2024 (the date of a targeted ML conference).

GWOSC has already downloaded the requested O3 auxiliary data from NDS, and Our team at UC Riverside has prepared a documentation for the dataset.

Channel Type	IFO	Chs/IFO	Machine-Readable Channel List
ISI-GND_STS channels	H1	15	H1/isi_gnd_sts_channels.txt
	L1	15	L1/isi_gnd_sts_channels.txt
ISI-GND_STS-BLRMS channels	H1	105	H1/isi_gnd_sts_blrms_channels.txt
	L1	102	L1/isi_gnd_sts_blrms_channels.txt
PEM-ACC channels	H1	38	H1/pem_acc_channels.txt
	L1	61	L1/pem_acc_channels.txt
PEM-LOWFMIC channels	H1	3	H1/pem_lowfmic_channels.txt
	L1	3	L1/pem_lowfmic_channels.txt
PEM-SEIS channels	H1	9	H1/pem_seis_channels.txt
	L1	9	L1/pem_seis_channels.txt
PEM-TEMP channels	H1	8	H1/pem_temp_channels.txt
	L1	9	L1/pem_temp_channels.txt
PEM-WIND channels	H1	12	H1/pem_wind_channels.txt
	L1	13	L1/pem_wind_channels.txt

Table 1: Requested channels. Each line links to a machine-readable list of channel names.