[Geophysical Research Letters]

Supporting Information for

Characterization of Seismicity from Different Glacial Bed Types: Machine Learning Classification of Laboratory Stick-Slip Acoustic Emissions

S. Saltiel[1], N. Groebner[2], T. Sawi[3], C. McCarthy[3], B.K. Holtzman[3]

[1] Nevada Seismological Laboratory, University of Nevada, Reno, NV, USA

[2] Strabo Analytics, Inc, New York, NY, USA.

[3] Lamont-Doherty Earth Observatory, The Earth Institute, Columbia University of New York, NY, USA

Corresponding author: Seth Saltiel (ssaltiel@unr.edu)

Contents of this file

Text S1 to S5

Additional Supporting Information (Files uploaded separately)

Captions for Movie S1

Introduction

Supporting information of additional details on experimental methods and materials, as well as data processing. Text S1 includes details of ice, rock, and till sources and preparation procedures; apparatus design; and experimental protocols. Text S2 includes data cleaning and normalization processing steps. Dataset S1 is a movie of the experiment and data stream in real-time, including audible stick-slips that are simultaneous with AEs and mechanical stress-drops being recorded. The datasets generated for this study are available on figshare.com at doi: 10.6084/m9.figshare.21257730, and Jupyter notebook for processing data is available at https://github.com/StraboAI/IcesAEs.

Text S1: Experimental Details

For this study we only used bulk ice samples, frozen slowly from deionized water in a slightly oversized die, and subsequently cut down to 50 x 50 x 100 mm with a microtome housed in a cold room (~ - 12 °C). The bulk freezing process results in large, non-uniform grain size compared to 'standard ice,' created using a narrow range of seed ice grain sizes [Cole 1979]. Saltiel et al., [2021] showed an insignificant frictional difference between the two types, so we employed bulk ice in this study. The simplified freezing process is much less time intensive and allows the ultrasonic transducers to be frozen directly into the ice sample (Figure S1), minimizing travel distance from the ice-bed interface and contact surfaces which can greatly diminish recorded acoustic amplitudes. The sliding surfaces were roughened with a no. 100 grit sandpaper using the same procedure as McCarthy et al., [2017], who determined a roughness average (Ra) of 7 ± 1 μm using a profilometer (Mitutoyo SF-210).

Figure S1: Bulk ice with an ultrasonic transducer (AE sensor) frozen into it. The bulk freezing process allows the suspension of the sensor in the deionized water during slow freezing. The sensor is oriented to face the sides of the block, where the ice-bed interface, source of AEs, will be when loaded into the apparatus.

As in Saltiel et al., [2021], we control temperature with Peltier thermoelectric coolers in front and behind the ice block, as well as circulation of chiller fluid through the side blocks where both temperature and flow rate of chiller fluid were actively controlled to reach the desired temperature. Resistance Temperature Detectors (RTDs) ported directly behind the till or rock monitor the temperature as close to the sliding interfaces as possible. Unlike in Saltiel et al., [2021], we preformed experiments with both stable and changing temperature to explore the effect on stick-slip instability, stress-drops, and resulting AEs.

Actively chilled aluminum side blocks were employed with either frozen till or rock attached to their ice-facing sides (Figures 1a, S2). All till experiments used a sample collected from the Matanuska glacier in south-central Alaska and were prepared using the same procedure described in Saltiel et al., [2021]. For rock beds, we employed Barre granite quarried from Barre Township, Vermont, that was cut into two 10 x 50 x 50 mm

slabs. A hole was drilled into the back side of the rock with the size and orientation of the side blocks' RTD port, to embed the RTD and measure the temperature directly behind the ice-rock interface. These slabs were then epoxied onto the aluminum side blocks and roughened using no. 100 grit sandpaper.

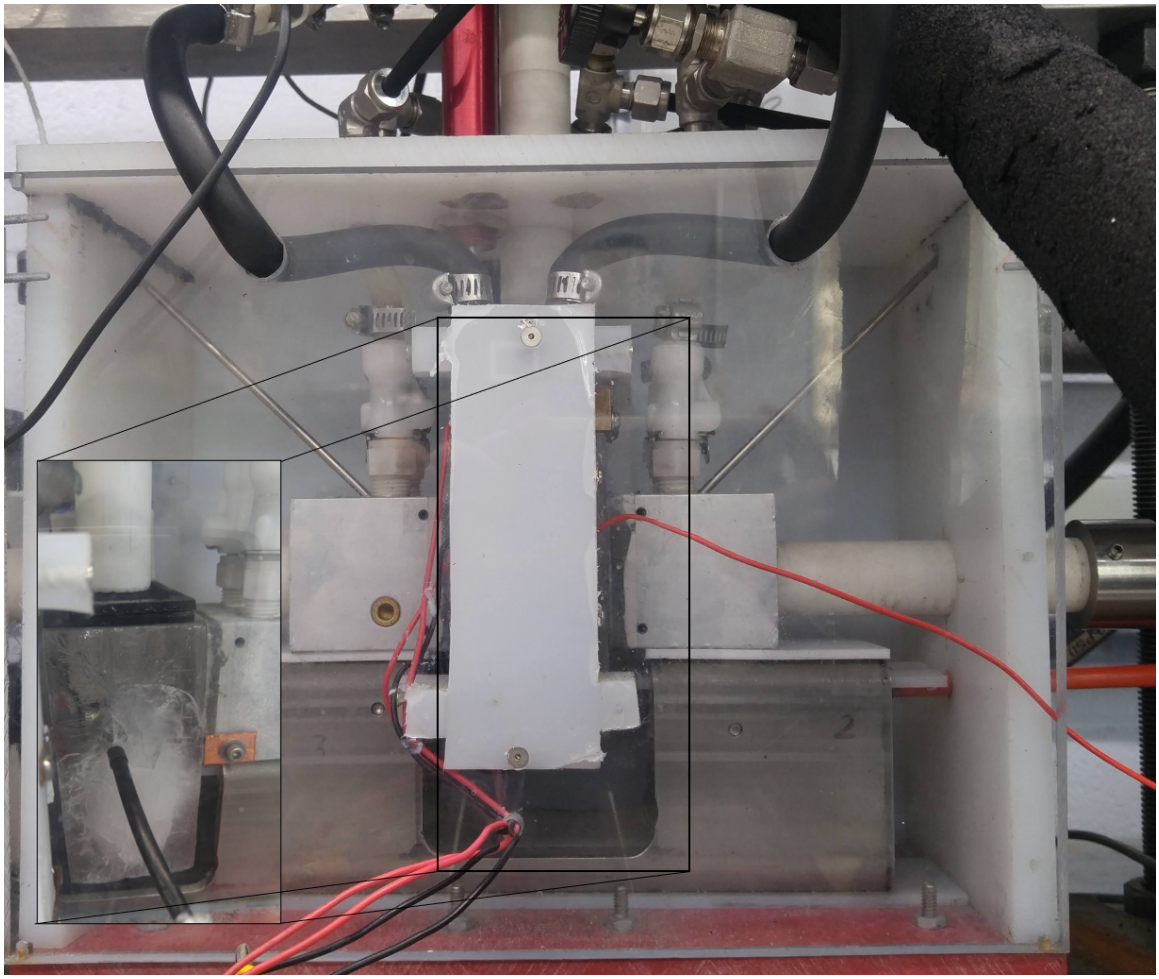


Figure S2: Photo of apparatus fully loaded. Since the peltier coolers cover the ice block, a photo without the cover is inset in the bottom left corner showing the central ice block at the end of an experiment, at the end of its full displacement.

All experiments were undertaken at ~50 kPa of normal stress and a load point velocity of 100 μm/s (just over 3 km/yr) for the entire displacement of 40 mm. This relatively high load point velocity was chosen because previous work has shown that stability decreases with slip velocity [Zoet et al., 2013, Saltiel et al., 2021]. Since the load point Linear Variable Differential Transformer (LVDT) only has 20 mm of stroke, the load point was stopped halfway through each experiment and then LVDT was reset to complete the rest of the experimental displacement. In this way, every experiment included a hold of about 60 seconds during which the shear stress relaxed and then reloaded, usually resulting in the largest stress-drop and AE of each experiment.

Text S2: Data Cleaning, Trimming, and Normalization

We implemented a data cleaning, trimming, and normalization approach based on that implemented by Nolte and Pyrak-Nolte [2022]. First, waveforms were trimmed to a total of 1200 samples, including 400 samples before the trigger point, giving a total window of 15 microseconds. Waveforms were then normalized by the sum of the squared amplitudes of the first 400 samples after the trigger, multiplied by a cosine taper. Zero and large amplitude waveforms were removed, defined as having a sum of the first 400 normalized samples greater than 15. This threshold was found to give the best catalog of non-noise events without removing too many. 325 events were then removed that a have high amplitude low frequency noise component. Finally, the waveforms were realigned to the first maximum peak after the trigger, which refined alignment by a few samples in most cases. From this catalog of normalized, filtered, and aligned 1200-sample waveforms, we used a trial-and-error approach to determine how much of the pre- and post-trigger waveforms to use for training the models and found a total length of 150 samples, with 45 before the trigger, was optimal. This subsample of the waveforms emphasizes the first arrivals of each AE, which are more dependent on source effects, while ignoring the coda, which depend more on path effects. Although, as we will show in the next section, the original, unprocessed catalog was able to produce as high prediction accuracies, the processed waveforms were clearer to interpret, the main point of this study.

Text S3: Results from Suite of ML Classification Algorithms

We systematically tested of a suite of ML classification algorithms, the original, full catalog and that created by the trimming and cleaning processing steps described above, using both waveforms and spectra. Figures S3 – S6 show the distributions of prediction accuracies for each of these combinations of algorithms and catalogs.

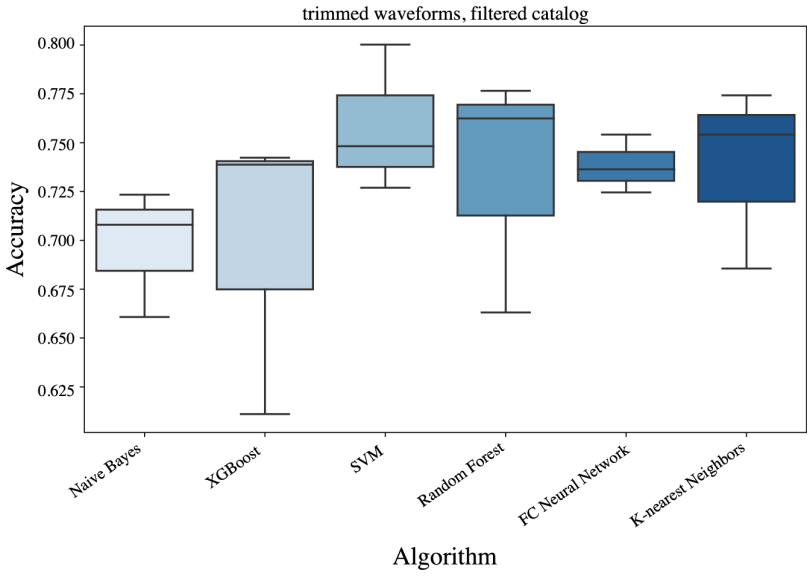


Figure S3: Whisker plot showing the distribution of prediction accuracies using the processed waveform catalog for each algorithm tested. Random Forest Classifier shows the highest mean accuracy of the all the algorithms which give a distribution, and, most importantly, provides feature importance for interpretation, so we focus on those results.

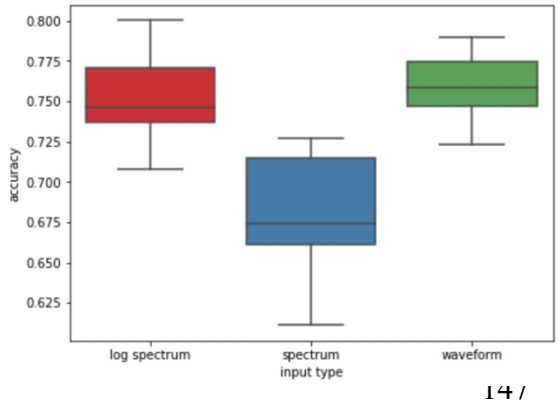


Figure S4: Whisker plot showing the distribution of prediction accuracies for each input data type tested, using the processed catalog. Waveforms show the tightest distribution and highest mean. Spectra are not very accurate, because the low frequency power dominates the spectral power and thus contains little information (see S4 below). But $\log_{10}$ of the spectrum retain the high frequency information and accuracy can be as high as the predictions using waveforms.

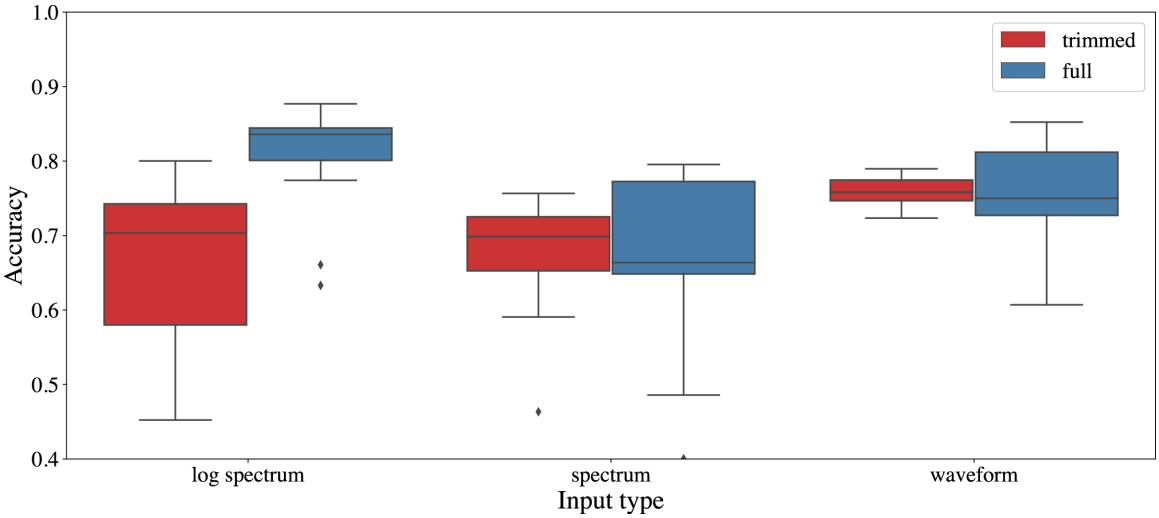
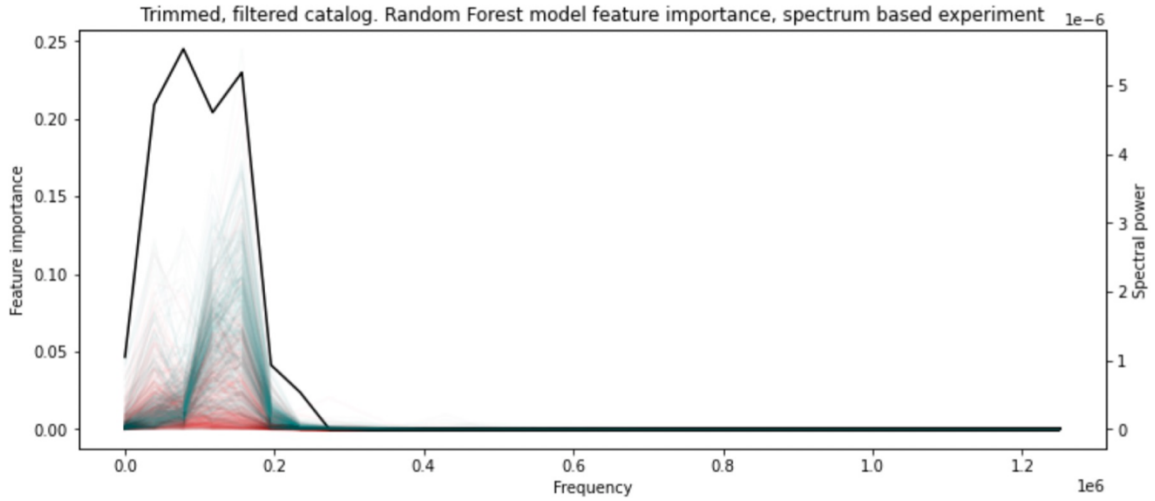


Figure S5: Whisker plot showing the distribution of prediction accuracies for the original, 'full', catalog of events vs. the processed, 'trimmed', catalog, using the processing steps described in text S2. Although the full catalog is able to give as good, or sometimes better predictions accuracies, which is not surprising since it contains more information, we focus our analysis on the processed, 'trimmed', catalog since the results are easier to interpret, the main focus of this study.

Text S4: Predictions using Spectrum vs Log Spectrum

We first undertook our analysis using spectrum, to test the predictive power of spectral information. But since the low frequency power dominates, using straight spectral power greatly diminishes the amount of data available (Figure S6a), and thus the predictions are relatively poor (Figure S4). By taking the log of the spectrum the higher frequency information is useful (Figure S6b) and predictions are more accurate.

a)



b)

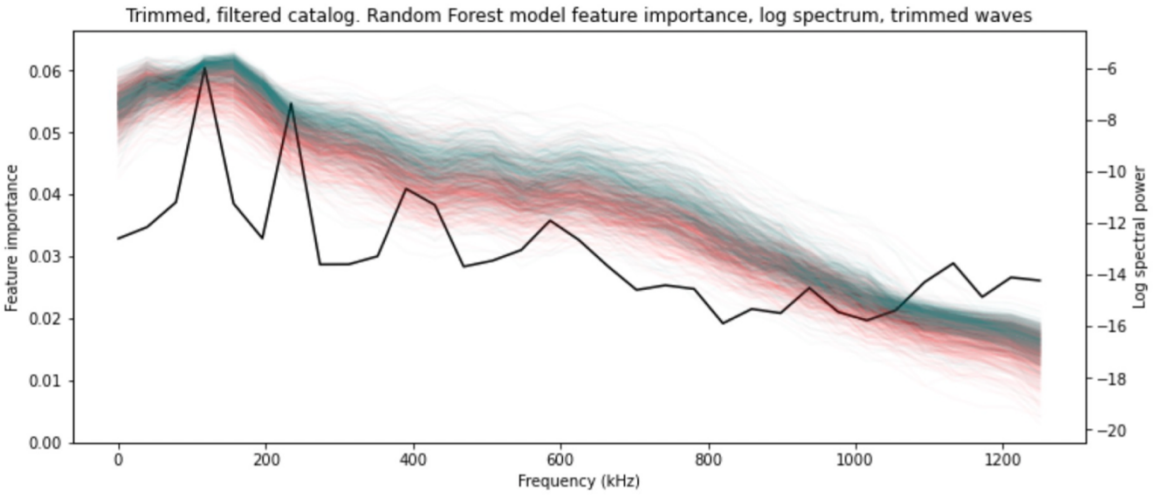
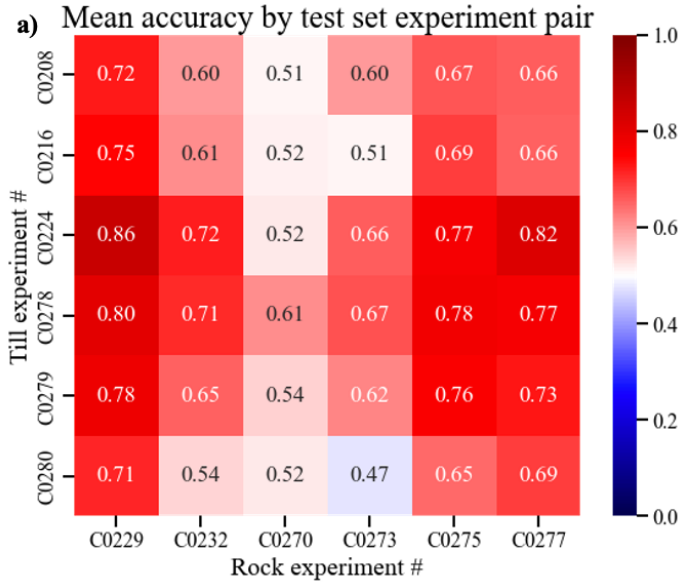


Figure S6: a) Spectrum from every till (teal) and rock (red) event, and the feature importance used to make Random Forest Classifier model predictions. Most spectral power is below 200 kHz, **b)** by taking the log spectrum, the higher frequency information is useable and prediction accuracy is improved.

Text S5: Testing Experimental Differences

To ensure that the prediction is not based on some aspect of the waveform specific to the ice sample or other uncontrolled aspect of the experiment and not the bed type which we are testing for, we also tested each experiment independently, not allowing the algorithm to train on data from the same experiment as the testing. We divide the data into training and test sets based on experiment, i.e., for a given model training run the waveforms from 5 till and 5 rock experiments are used for the training set, and the remaining 1 till and 1 rock experiment are used for testing. By separating training and test sets by experiment, any experiment-dependent features of the waveforms would be irrelevant for classification. The prediction accuracy is summarized by a 6 till by 6 rock experiments matrix, giving the accuracy for 36 models with each combination used as the testing data (Figure S7).

a)



b)

Exp. (C0#)	Bed type	Temp (~°C)	# of events	Accu-racy
208	Till	-3	94	67%
216	Till	-3	465	71%
224	Till	-2.5	245	86%
278	Till	-3.4	421	82%
279	Till	-4	181	74%
280	Till	-4	141	63%
229	Rock	-6 => -4	319	82%
232	Rock	-4 => -3	184	51%
270	Rock	-3	296	38%
273	Rock	-3.8	95	42%
275	Rock	-4.3 => -3.9	150	71%
277	Rock	-3.9	226	68%

Figure S7: a) Mean prediction accuracy given different sets of rock and till experiments used as testing dataset. In each case, the other experiments were used as training data, producing a model for each combination of testing experiments (6 till and 6 rock experiments make for 36 different train and test datasets, and models). Although some experimental variation is expected, relatively consistent results across testing datasets (either randomly selected from all experiments or from an individual one) shows that the overall predictability is not experiment dependent. **b)** Table on right provides the temperature range, number of events, and accuracy for each individual experiment.

This prediction accuracy calculates how often the model could correctly classify individual waveforms as coming from till or rock beds, but we envision a tool whereby a collection

of seismic events recorded from a given location would be analyzed to determine the probability it came from a till- or rock-bedded section of a glacier. So, the more relevant accuracy is if a single experiment can be accurately predicted to be till or rock, and how many events would be needed to make such a prediction accurate. Since its clear from Figure 5 that there are overlapping 'till-like' rock events and visa-versa, the direct prediction does not have to be used for the overall population prediction. For example, we find that all the experiments can be correctly predicted if 37.5% 'rock-like' events, or 62.5% 'till-like' events, is used as the cut-off for overall prediction (Figure S8). Our data shows a sharp cut off at these values, so it likely would not remain a perfect classifier with more experiments, but it does suggest how predictions might be made given the overlapping event populations.

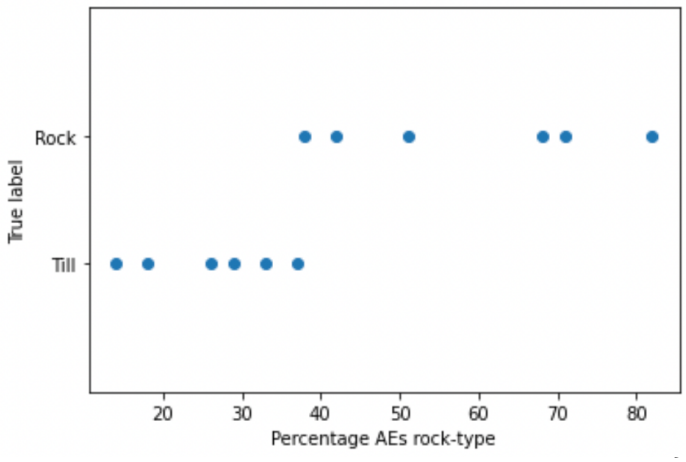


Figure S8: Each experiments percentage of events predicted as rock, which we label as 'rock-like' events. The till and rock experiments perfectly separate if more than 37.5% of the events are predicted as rock.

Since there are rock experiments with more 'till-like' events than 'rock-like' events, it is possible that the model is 'defaulting' to till since there are slightly more till than rock events overall. We do not believe this is the case, given the significant overlap in the characteristics of rock and till events (Figure 5). While the rock stress-drops have a tighter distribution (Figure 4c), these stress drops do not follow a simple relationship with recurrence interval, as would be expected with a single healing rate and as seen with the till experiments (Figure 4d). Although there is not enough data to fully constrain, Figure 4d suggests that some rock experiments sit on the till healing relation (stress-drops of about 25 kPa per second of recurrence interval), while others have lower healing rates. This may explain the imbalance in prediction accuracy, why there are more 'till-like' rock AEs than 'rock-like' till AEs. Some experiments near the cut-off, such as 270, would be very difficult to predict correctly. 270 is one of the rock experiments with a high healing rate (~22 kPa/s), which might contribute to its having more 'till-like', misclassified events.

Movie S1: Movie of experiment and AE recording in real-time. Audible stick-slips and mechanical stress-drop data (not shown) both simultaneously occur with the recorded AEs. Some events appear to have two arrivals, probably one from each ice interface, since they have different path lengths they arrive at the sensor at slightly different times even if they occur at the same time. In these cases, the processing steps from text S2 remove the later arrival.