Supplementary Material for

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

**S. Saltiel1, N. Groebner2, T. Sawi3, C. McCarthy3**

1Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

1. Strabo Analytics, Inc, New York, NY, USA.
2. Lamont-Doherty Earth Observatory, The Earth Institute, Columbia University of New York, NY, USA

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

**Contents of this file**

Text S1 to S5, Video 1

# Introduction

Supplementary materials 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 results from the entire suite of machine learning classification algorithms. Text S3 includes a comparison of predictions using the spectrum vs log spectrum. Text S4 includes a prediction experiment if entire experiments are held for testing. Video 1 includes an example rock bed experiment with audible stress-drops corresponding to AE waveforms. 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 and others, [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 and others., [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 right side 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 and sustain 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 performed experiments with both stable and changing temperature to explore the effect on stick-slip instability (see section 3), stress-drops, and resulting AEs (future work).

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 and others, [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 A picture containing kitchen appliance, miller

Description automatically generatedepoxied 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, reaching the bottom 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 experimental work of ice bed slip has shown that stability decreases with slip velocity [Zoet and others, 2013; Saltiel and others, 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 (see Figure 2 of the main text).

# Text S2: Results from Suite of ML Classification Algorithms

We systematically tested a suite of ML classification algorithms on both the full recorded waveforms and a short segment containing the first arrival (the product of processing steps described in Section 4 of the main text). We tested three types of input features: the waveforms themselves (normalized amplitude), spectra created from the waveforms, and the base-10 logarithm of those spectra. The waveform lengths, input features, and ML algorithms are listed in Table S1. Each combination of waveform length, input feature, and algorithm was tested with one or more initial test-train splits to produce a total of 42 separate numerical experiments. Figures S3 and S4 show the distributions of prediction accuracies for each of these combinations of algorithms and catalogs.

|  |  |  |
| --- | --- | --- |
| **Waveform length** | **Input feature type** | **Algorithm and source** |
| * Full recorded waveform, 120 μs * First arrival, 15 μs segment starting 5 μs from the trigger | * Waveform normalized amplitude * Spectrum (Welch’s method, numpy package) * Log10 of the spectrum | * Naïve Bayes, scikit-learn * XGBoost, xgboost library * Support vector machine (SVM), scikit-learn * Random forest, scikit-learn * Fully-connected (FC) neural network (multilayer perceptron), scikit-learn * K-nearest neighbors, scikit-learn |

**Table S1:** List of the parameters for the machine learning experiments. All combinations of parameters were tested, and several were tested using multiple test-train splits. A total of 42 experiments were run.

**A chart of a graph

Description automatically generated**

**Figure S3:** Whisker plot showing the distribution of prediction accuracies using the processed waveform catalog for each *algorithm* tested. Each distribution contains the accuracies of multiple runs produced by varying the amount of the waveform used and the input features, discussed above. Highest median accuracy is obtained by support vector machines; second highest by random forests. The number of runs for each algorithm is listed on the chart.

A diagram of a diagram

Description automatically generated

**Figure S4:** Whisker plot showing the distribution of prediction accuracies for each *input data type*. Spectra have the lowest median accuracy and the largest variation in accuracy because the low frequency power dominates the spectral power and thus contains little information (see section S3 below). The base-10 logarithm of the spectrum retains the high frequency information and demonstrates the highest median and absolute accuracy.

# Text S3: Predictions using Spectrum vs Log Spectrum

We first undertook our analysis using spectra, 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 S5a), and thus the predictions are relatively poor (Figure S4). By taking the log of the spectrum we retain the useful higher frequency information (Figure S5b), and predictions are more accurate.

# a)

# A graph of a person with a blue line Description automatically generated

# b)

**A graph with lines and numbers

Description automatically generated**

**Figure S5: a)** Spectrum from every till (teal) and rock (red) event, and the feature importance used to make Random Forest Classifier model predictions (black). 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 S4: 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 **Text

Description automatically generated with medium confidenceChart, bar chart

Description automatically generated**as the testing data (Figure S6a).

**Figure S6: 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 S7). Our data shows a sharp cut off at these values, so it is unlikely to remain a perfect classifier with more data, but it does suggest how predictions might be made given the overlapping event populations.

Chart, scatter chart

Description automatically generated**Figure S7:** 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 C0270, would be very difficult to predict correctly. C0270 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.

**Video 1:** Example rock bed experiment with audible stress-drops that correspond to waveforms recorded by AE sensor frozen into ice block. Instances of multiple arrivals can be seen (likely emitted from the two sliding interfaces) but a single waveform is always isolated by using the processing steps described in section 4 of the main text.