https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

Exploring seismic mass-movement data with anomaly detection and dynamic time warping

Francois Kamper[1], Fabian Walter[2], Patrick Paitz[2], Matthias Meyer[1], Michele Volpi[1], and Mathieu Salzmann[1]

[1]Swiss Data Science Center, EPFL & ETH Zürich
[2]Swiss Federal Institute for Forest, Snow and Landscape Research WSL

Correspondence: Francois Kamper (francois.kamper@epfl.ch)

Abstract. Catastrophic mass movements, such as rock avalanches, glacier collapses, and destructive debris flows, are typically rare events. Their detection is consequently challenging as annotated and verified events used as training data for instrumentation and algorithm tuning are absent or limited. In this work, we explore seismic mass-movement data through the lens of anomaly detection. The idea is to screen out segments of the data that are unlikely to contain mass movements by focusing only on anomalous signals, thereby reducing the number of signals to be studied, making downstream tasks such as expert labeling and clustering of events easier. To extract anomalous signals, we design a triggering algorithm using an anomaly score computed from an isolation forest obtained from sliding windows taken from the continuous data. The extracted signals are subjected to expert labeling and/or further analyzed by dynamic time warping, a popular technique used to evaluate the dissimilarity between different types of signals. We illustrate our approach by (a) mining for seismic signals of hazardous debris-flows in Switzerland's Illgraben catchment and (b) labeling of seismic mass movement data obtained from a Greenland seismometer network.

1 Introduction

Seismic networks record ground unrest and generate large amounts of continuous data in the public domain. As global and regional earthquakes are the main focus of existing automated processing workflows by national and international seismological organizations, important natural phenomena also exciting seismic signals remain hidden in the vast amounts of available continuous data. Even though there are ongoing efforts to detect and characterize non-earthquake seismic events (Bahavar et al., 2019), a big part of the available data remains unexplored. The topic of environmental seismology focuses on these non-tectonic seismic events using the signals of rock falls, avalanches, debris flows and other mass movements to study underlying processes (Larose et al., 2015). In this context, past studies have shown the high value of seismic measurements for natural hazards science (Montagner et al., 2021).

Conventional algorithms in earthquake seismology, such as the short-term average over long-term average STA-LTA trigger (Allen, 1978), are not easily transferable to the domain of environmental seismology - especially since discrimination between

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

earthquake, noise as well as other transient waveforms and the signals of interest can become difficult. Hence, statistical learning models are needed to gain more insight into complex phenomena such as hazardous avalanches, debris flows and other mass movements (Wenner et al., 2021; Chmiel et al., 2021) and basal sliding of glaciers (Umlauft et al., 2023). In the presence of limited or no labels, unsupervised or semi-supervised methods are needed to create and refine catalogs of events, see for example Meyer et al. (2019). These type of analyses are challenging, due to high sampling rates (hundreds to thousands of Hertz) and the long-term measurements, spanning multiple years across multiple stations and networks.

From a data perspective, distinct physical seismic events (including, but not limited to, earthquakes) can be interpreted as anomalies in a background noise field. From a geophysical perspective, this background field is very complex, transient and non-stationary (Nakata et al., 2019; Fichtner et al., 2020) - so the term "noise" might be misleading for non-seismologists. Studying the properties of this seismic noise field has revolutionized passive seismology in the last decade, with applications ranging from global-scale subsurface tomography (Sager et al., 2020) to noise source location (Igel et al., 2021) and aquifer monitoring (Rodríguez Tribaldos and Ajo-Franklin, 2021). Compared to the duration of seismic events from hazardous mass movements (minutes to hours), the rate of change in the background noise field throughout such events is often negligible, taking place on diurnal to seasonal time scales. This motivates us to tackle seismic signal detection from an anomaly detection approach.

Here we explore seismic mass-movement data by combining anomaly detection with semi- and unsupervised learning, using dynamic time warping (DTW) to quantify dissimilarity between signals. The idea is based on the insight that mass-movement signals represent significant statistical anomalies in the seismic data of instruments well-placed to detect these events. From this viewpoint, we should be able to screen out large portions of the data unlikely to contain mass movement signals, thereby reducing the amount of signals to be studied. In this work we consider the isolation forest (IF) algorithm, a simple yet powerful anomaly detection method. We chose this algorithm because of (a) fast training and inference, (b) light-weight storage of models, and (c) strong empirical performance (Liu et al., 2008, 2012; Bouman et al., 2024).[1] Since vanilla anomaly detection methods cannot discriminate between different types of anomalies, the extracted signals need to be further analyzed, either by expert labeling or unsupervised/semi-supervised methods. We pursue both approaches in this work, with the latter guided by measuring dissimilarity between signals using DTW. To illustrate the value of our approach we consider refining an existing catalog of hazardous debris flows in Switzerland's Illgraben catchment, and generate a catalog from scratch for data obtained from a Greenland seismometer network.

[1] Although (a) and (b) are not strictly necessary for the applications of this paper, they could be more relevant for future work, such as extensions to the online setting.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

2 Methodology

2.1 Pre-processing

We use the Scikit-learn (version 1.4.1) (Pedregosa et al., 2011) and ObsPy (version 1.4.0) (Beyreuther et al., 2010) libraries to implement the training and signal processing procedures. The pre-processing of the raw mini-seed seismic recordings follows standard procedures in seismology. In the first step, we identify gaps in the data and discard all recordings with less than 1000 consecutive samples, as gaps in the data indicate issues on the instrumentation side. We then apply a linear de-trending and de-meaning of each recording to ensure zero-mean recordings without a drift in amplitude, followed by a zero-phase high-pass filter with a corner frequency of 0.3 Hz. Furthermore, all data are re-sampled to the same sampling rate of 100 Hz. We refer to the units of the seismic waveforms after they have been preprocessed as preprocessed counts.

2.2 Isolation forest

An IF consists of an ensemble of decision trees trained in an unsupervised manner where, in contrast to traditional decision trees and random forests, both the splitting variable and the splitting point are completely decided at random. The argument is that if we fit a decision tree to a data set $\mathcal{D} = \{\boldsymbol{x}_i : i = 1, 2, \ldots, n\}$ in this manner, anomalies in $\mathcal{D}$ tend to be isolated into singleton nodes at fairly low depths of the tree, and this property can be exploited to derive sensible anomaly scores. As in Liu et al. (2008, 2012) we refer to these random decision trees as isolation trees (iTrees).

For a test observation $\boldsymbol{x}$ and a given iTree, let us define the path length $h(\boldsymbol{x})$ as the number of edges from the root to the terminal node containing $\boldsymbol{x}$. The more anomalous $\boldsymbol{x}$, the smaller we expect $h(\boldsymbol{x})$ to be. An IF aims to estimate $\mathbb{E}_{\mathcal{D}}[h(\boldsymbol{x})]$, i.e., the expected path length for a test observation $\boldsymbol{x}$ over iTrees fitted to different datasets $\mathcal{D}$. An estimate $\widehat{\mathbb{E}}_{\mathcal{D}}[h(\boldsymbol{x})]$ is obtained by fitting iTrees to sub-samples of the data and averaging the path lengths. The test observation is flagged as anomalous if $\widehat{\mathbb{E}}_{\mathcal{D}}[h(\boldsymbol{x})]$ is sufficiently small.

For improved interpretability, the final anomaly score is calculated by normalizing and transforming the quantity $\widehat{\mathbb{E}}_{\mathcal{D}}[h(\boldsymbol{x})]$ to a value in $(0, 1)$ with higher values indicative of more anomalous observations. Normalization is achieved with division by $c_{|D|}$, a quantity representing the average number of edges from root to terminal node over all possible test observations and data sets of size $|D|$. In fact, since a test observation hitting a terminal node can be interpreted as an unsuccessful search in a binary search tree (BST), we can compute $c_{|D|} = 2 \cdot H_{|D|-1} - 2 \cdot \frac{|D|-1}{|D|}$, with $H_{|D|-1}$ the harmonic number (Liu et al., 2008, 2012). The final isolation forest anomaly score for a test observation $\boldsymbol{x}$ is given by $s(\boldsymbol{x}, \mathcal{D}) = 2^{-\frac{\widehat{\mathbb{E}}_{\mathcal{D}}[h(\boldsymbol{x})]}{c_{|D|}}} \in (0, 1)$.

2.2.1 Fitting the isolation forest

Fixed-size sliding windows have proven useful in converting time series data to a usable format for machine learning algorithms such as random forests, especially in the context of real-time monitoring (Wenner et al., 2021; Chmiel et al., 2021).

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

We follow this convention for the IF by taking sliding windows from the seismic waveforms, after they have been suitably preprocessed as discussed above. Except if indicated otherwise, by sliding windows, we mean 100 second windows taken with with 50 second overlap.

To obtain a sub-sample, we take all sliding windows corresponding to a single raw mini-seed seismic recording after it has been preprocessed. This means an ensemble of one iTree for each raw seismic mini-seed recording, which typically corresponds to a calendar day. The number of sliding windows in a sub-sample depends on the duration as well as the number and size of gaps in the corresponding recording. For a comparison of the IF anomaly score to the standard deviation of sliding windows we refer to Appendix A.

2.2.2 Isolation forest trigger

The IF trigger is activated when the IF anomaly score of a sliding window exceeds a specified onset threshold. We continue sliding windows until the IF anomaly score drops below a specified offset threshold and the flagged segment is marked from the starting point of the onset window, until the starting point of the offset window. The maximum IF anomaly score of the sliding windows taken over this period is the anomaly score associated with the entire segment, which we call the IF segment anomaly score.

The onset and offset thresholds can be either preset or calibrated to data. In the case where calibration is not possible, we recommend using an onset and offset threshold of 0.60 and 0.55, respectively, as a rule of thumb. The segments flagged by the IF trigger (IF segments) are then ranked by their corresponding IF segment anomaly scores in decreasing order.

2.3 Dynamic time warping

Suppose that we want to align two sequences $\boldsymbol{x}_1 \in \Re^{T_1}$ and $\boldsymbol{x}_2 \in \Re^{T_2}$, possibly of different lengths. We define a path $p = \{(i_k, j_k)\}_{k=1}^{K}$ such that $(i,j) \in p$ indicates that element i in $\boldsymbol{x}_1$ has been matched with element j in $\boldsymbol{x}_2$. We call a path p valid if it satisfies the following conditions:

1. $(i_1, j_1) = (1,1)$ and $(i_K, j_K) = (T_1, T_2)$.
2. $i_k \leq i_{k+1} \leq i_k + 1$ and $j_k \leq j_{k+1} \leq j_k + 1$.

These conditions ensure that (a) the first and last entry of $\boldsymbol{x}_1$ are matched with the first and last entry of $\boldsymbol{x}_2$ respectively (b) all the indices of both time series are used and (c) the path respects the flow of time in both sequences; for example if we match element 3 in $\boldsymbol{x}_1$ with element 10 in $\boldsymbol{x}_2$ then we are not allowed to match element 20 in $\boldsymbol{x}_1$ with element 2 in $\boldsymbol{x}_2$. The DTW objective is to find the valid path that minimizes the objective

$$\sum_k d(x_{1i_k}, x_{2j_k}), \tag{1}$$

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

where $d(\cdot,\cdot)$ is a chosen distance metric such as the Euclidean distance. The minimizing path determines the DTW alignment between the sequences, and the corresponding value of (1) is called the DTW distance, although it does not define a proper metric since it does not necessarily satisfy the triangle inequality.

The DTW problem can be solved using dynamic programming in $\mathcal{O}(T_1 \cdot T_2)$ time and storage complexity (Salvador and Chan, 2007). Since the signals we are studying are relatively long, using all the data at once when performing DTW is computationally prohibitive. To account for this, we consider the following two approaches for using DTW to measure dissimilarity between two signal segments:

1. **Template DTW**. Take the single sliding window over a segment with the largest IF anomaly score as a segment template. The template DTW distance between two segments is computed as the DTW distance between the corresponding templates.

2. **Segment DTW**. Take all sliding windows over a segment and compute the corresponding IF anomaly scores. To compare two segments, we first compute the DTW alignment between the two time series of anomaly scores. We match a sliding window from one segment with the sliding window of another if their corresponding anomaly scores were matched in the alignment, and then compute the DTW distance between each matched pair of sliding windows. The segment DTW distance is obtained by aggregating these distances into a single statistic using, for example, the mean or median.

The template DTW is preferable computationally, while segment DTW is able to better discriminate between segments and is preferable when a smaller number of signals are being studied. In all cases, we use the implementation of Salvador and Chan (2007) with the Euclidean distance metric, and sliding windows are normalized to zero mean and unit standard deviation before performing DTW.

3 Case studies

We present two case studies for the application of the methodology described in Sect. 2. The first study aims to refine an existing catalog of debris flows in the Illgraben torrent, Switzerland, while the second focuses on generating a catalog of events from the seismic broadband station KARAT in Greenland. Overviews of these settings and maps of the respective seismic networks are provided in Fig. 1.

3.1 Illgraben

3.1.1 Study site

Located in southern Switzerland's Canton Valais, the Illgraben is one of Europe's most active debris flow torrents. Its catchment drains an area of ca. $10\,\mathrm{km}^2$ and produces sediments at higher elevation, which are mobilized during heavy precipitation to form debris flows and sediment-laden torrential floods. Each year, several such flows with volumes of a few tens of thousands m^3

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

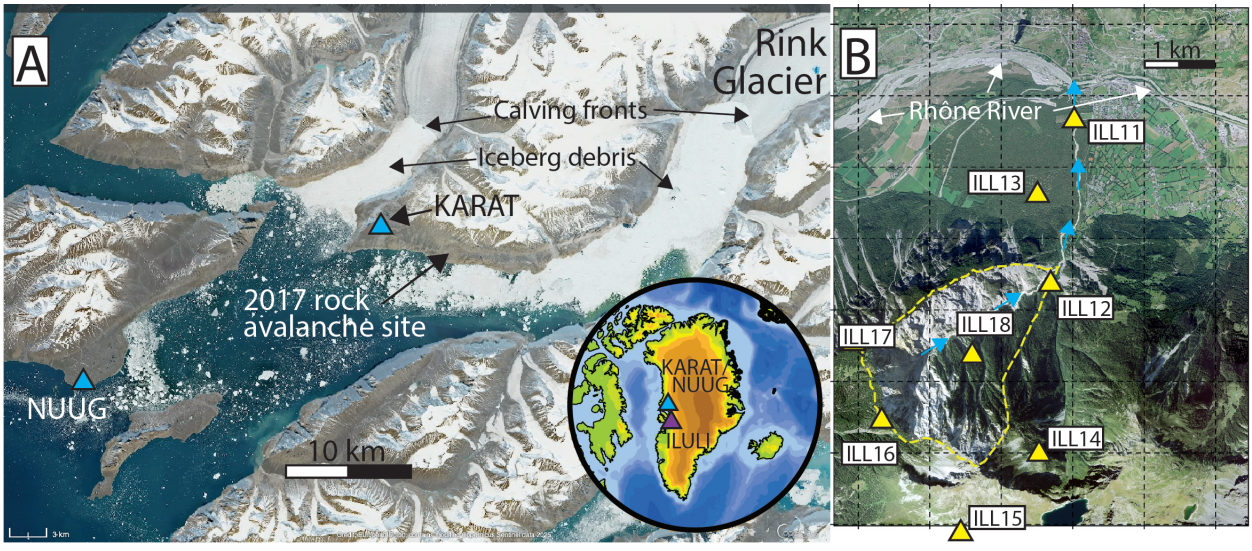

Figure 1. Study sites in Greenland (A) and Switzerland (B). (A) Karrat Fjord with seismic broadband stations (blue triangles), the location of the 2017 rock avalanche and major calving fronts indicated. White ice debris cover on the tidewater results from disintegrating icebergs. Inset shows the location of the site in Greenland. (B) Illgraben torrent with debris-flow-producing upper catchment outlined by yellow dashed lines. Blue arrows indicate flow direction. Yellow triangles represent seismometers. Sources: Copernicus (Sentinel-2 true color image) and inset from the Generic Mapping Toolbox (A), Swisstopo (B).

reach the Rhône River (Badoux et al., 2009; Hürlimann et al., 2003). Illgraben's debris flows move at several meters per second and feature the typical boulder-rich fronts, which are efficient seismic sources that can be detected on local seismic networks (Walter et al., 2017). At Illgraben, the Swiss Federal Institute for Forest, Snow and Landscape Research WSL maintains a semi-permanent seismic network that monitors debris flows and consists primarily of 1 Hz seismometers (Fig. 1). In addition, WSL's Illgraben debris flow observatory contains geophone plates, automatic cameras and depth gauges to measure flow arrival times and flow depths at various points along the torrents, especially at concrete structures stabilizing the channel ("Check Dams"; Badoux et al., 2009).

3.1.2 Debris-flow catalog

A debris flow signature can be defined to occur when the seismic waveforms of multiple stations are affected in the expected pattern as a debris flow moves down the torrent. In the case of the Illgraben seismic network we expect a debris flow to affect the upper stations ILL14-ILL18 first, and subsequently ILL12, ILL13 and ILL11 in order. The existing debris flow catalog was independently curated by cross-referencing detections made by WSL's Illgraben debris flow observatory with the seismic waveforms of the stations in the network, keeping this definition in mind. Each segment in the catalog consists of a start- and end-time, coupled with a station and confidence level. The confidence levels are defined as follows:

1. High confidence. The segment is observed during a debris-flow signature and contains a clear signal.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

2. Medium confidence. The segment is observed during a debris-flow signature and contains some signal, although somewhat suppressed. We also include here segments with a clear signal where not enough stations were active to establish if a debris-flow signature is present.

3. Low confidence. The segment is observed during a debris-flow signature; however, without the signature present in other stations it is debatable if this signal is related to a debris flow.

For the remainder of the case study we use "lower-confidence segments" to refer to both low- and medium confidence segments jointly.

3.1.3 Mining methods

We develop three different mining methods for debris flows that are not contained in the original catalog using the available labels, i.e., following a semi-supervised approach. A chosen method is applied to each station in the network in order to produce station-dependent models aimed at recommending segments which are likely debris flows. We refer to these recommended segments as detections.

A mining method consists of a triggering algorithm, scoring method, score threshold and minimum detection length. To generate a list of detections for a station we deploy the triggering algorithm over a period to generate a list of trigger segments. The scoring method assigns scores to the trigger segments and rank them in order of likelihood of being associated with a debris flow. Trigger segments that meet the score threshold and minimum detection length are kept as a list of detections and those not referenced in the original catalog are subjected to expert labeling as potential undiscovered debris flows. We consider the following mining methods:

1. **STA-LTA**. Our baseline method uses the classical STA-LTA trigger and the maximum value of the characteristic function observed over a segment as its associated score.

2. **IF**. We use the IF trigger and the IF segment anomaly score to generate and score segments.

3. **IF-DTW**. We use the IF trigger and score a segment as the mean of the template DTW distances between the segment and a subset of high-confidence segments.

For the STA-LTA and IF methods, trigger segments are ranked in decreasing order of the segment scores, and these scores can be interpreted as quantifying how severe an unknown event has affected the seismic waveforms at a station. These events can be caused by multiple sources such as debris-flows, earthquakes and anthropogenic noise. The STA-LTA or IF scoring method that views debris flows as more severe relative to other sources will achieve better performance. In the case of IF-DTW, trigger segments are ranked in decreasing order of the DTW score. Assuming that DTW can adequately capture dissimilarity between templates extracted from segments corresponding to debris-flows and other severe sources, IF-DTW will improve on the IF mining method.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

3.1.4 Calibration and evaluation of mining methods

Since it is unclear how to treat the lower-confidence segments, they do represent a challenge from a calibration and evaluation perspective. Our approach is to design the triggering algorithm for a station to capture the corresponding catalog segments (lower- and high-confidence) as well as possible, but to not allow lower-confidence segments to affect calibration of the score threshold and minimum-detection length. In this way the lower-confidence segments are explicitly encouraged to be included in the trigger segments, and to become detections if they happen to be recovered alongside high-confidence segments.

A segment in the catalog is labeled a true positive if we can find at least one detection that overlaps with it, otherwise it is labeled a false negative. A detection that does not overlap with any segment in the catalog is labeled a false positive. Denoting the number of true positives, false negatives, and false positives by TP, FN, and FP, respectively, we compute

$$\text{recall} = \frac{TP}{TP+FP}$$
$$\text{precision} = \frac{TP}{TP+FN}.$$

The recall measures the proportion of segments in the catalog found by a specified mining method, while precision measures the proportion of detections that intersect with segments in the catalog. To quantify the temporal overlap between a list of detections and catalog segments we use the intersection over union (IoU) metric, or the total time where detections and catalog segments overlap expressed relative to the total time where a detection or catalog segment is present.

A mining method is calibrated to data from a station over the training period only which we took as 2018 - 2020. Firstly we calibrate the triggering algorithm by (a) extracting all mini-seed recordings with at least one catalog segment present (both lower- or high confidence) over the training period (b) running the triggering algorithm with multiple hyper-parameter configurations over these recordings and (c) selecting the hyper-parameter configuration yielding a list of segments with the highest IoU with respect to the training catalog segments. Secondly, the calibrated triggering algorithm is deployed over the entirety of the training period to generate a list of training trigger segments. The training trigger segments are reduced by removing those that intersect with at least one lower-confidence segment, but no high-confidence segments, before being subjected to the score threshold and minimum detection length to generate detections. Finally, the score threshold and minimum detection length is selected as those values yielding the list of detections maximizing the IoU with respect to only high-confidence segments in the catalog.

For the IF trigger, we select on- and offset thresholds from $\{0.55, 0.6, 0.65, 0.7\}$ and $\{0.50, 0.55, 0.6, 0.65\}$ through a grid search, under the constraint that the onset threshold cannot be lower than the offset threshold. For the classical STA-LTA trigger, we found it difficult to choose a single grid that worked well on all stations and thus opted for a local search method instead. First, we conducted an extensive grid search on ILL11 and found that using a long-term window of 5000 seconds, a short-term window set to 10% of the long-term window, and onset and offset thresholds of 6.0 and 0.125, respectively, yielded

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

Metric	IoU (%)			Recall (%)			Precision (%)		
Station	STA-LTA	IF	IF-DTW	STA-LTA	IF	IF-DTW	STA-LTA	IF	IF-DTW
ILL11	53.06	60.32	61.02	86.96 (3)	100.0 (0)	100.0 (0)	83.33 (4)	100.00 (0)	95.83 (1)
ILL12	19.50	51.55	59.11	51.85 (13)	85.19 (4)	81.48 (5)	66.67 (7)	92.00 (2)	91.67 (2)
ILL13	25.03	64.24	74.25	50.00 (12)	75.00 (6)	91.67 (2)	75.0 (4)	100.00 (0)	100.00 (0)
ILL14	7.13	1.54	26.19	8.33 (11)	75.00 (3)	58.33 (5)	100.00 (0)	3.28 (265)	77.78 (2)
ILL15	2.18	0.71	1.67	14.29 (6)	14.29 (6)	14.29 (6)	6.25 (15)	3.85 (25)	50.00 (1)
ILL16	2.20	5.85	50.11	14.29 (12)	57.14 (6)	100.0 (0)	8.33 (22)	19.51 (33)	100.00 (0)
ILL17	10.65	7.76	42.69	5.26 (18)	68.42 (6)	89.47 (2)	100.00 (0)	10.66 (109)	89.47 (2)
ILL18	27.17	54.33	61.34	62.5 (9)	95.83 (1)	95.83 (1)	50.00 (15)	71.88 (9)	92.00 (2)

Table 1. Metrics over the training period after updating the catalog. The numbers in brackets in the recall and precision columns represent the number of false negatives and false positives, respectively. All percentages are displayed accurately up to two decimals.

a high IoU score. This configuration of hyper-parameters was used as a starting point for all stations. We then performed local neighborhood searches, with an exponential step size of 2, until no improvement in the IoU metric could be found.

To evaluate the detections produced by a mining method for a specified station we separate detections in the list that intersect with at least one lower-confidence segment, but no high-confidence segments, from the remainder. We then compute the IoU, recall and precision of the remaining detections relative to the high-confidence segments in the catalog of the corresponding station. We report the recall of low- and medium-confidence segments separately.

3.1.5 Debris flow detection: results

Tables 1 and 2 show the metrics for each mining method across all stations over the training and test periods respectively. In the recall and precision columns, the numbers in brackets indicate the number of false negatives and false positives respectively. The recall of the lower-confident segments are discussed in Appendix C.

These metrics are computed following three updates of the original catalog made over the training period.[2] The updates are performed by including those false positive detections which actually correspond to debris flows as newly discovered debris flows to the catalog, with assigned confidence levels and possibly modified start- and/or end-times, based on expert labeling. If a new debris flow is discovered from a given station, segments from other stations forming part of the debris-flow signature is included in the catalog as well. In addition, existing entries in the catalog can be modified, again based on expert labeling, either in terms of confidence level or of start- and/or end-times. To keep the number of detections to investigate manageable, we only investigate detections from stations ILL11, ILL12, ILL13 and ILL18 for the STA-LTA and IF methods, while for IF-DTW we also include stations ILL14, ILL16 and ILL17. After an update, the score threshold and minimum detection length of the

[2]These include smaller updates following, for example, experimentation with the hyper-parameter grids.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

Metric	IoU (%)			Recall (%)			Precision (%)		
Station	STA-LTA	IF	IF-DTW	STA-LTA	IF	IF-DTW	STA-LTA	IF	IF-DTW
ILL11	41.61	71.74	71.74	87.50 (2)	100.00 (0)	100.0 (0)	87.50 (2)	100.00 (0)	100.00 (0)
ILL12	16.18	52.31	71.70	46.15 (7)	76.92 (3)	100.0 (0)	100.00 (0)	90.91 (1)	92.86 (1)
ILL13	19.76	50.18	66.65	64.71 (6)	76.47 (4)	100.0 (0)	73.33 (4)	100.00 (0)	100.00 (0)
ILL14	0.00	4.37	34.33	0.00 (12)	91.67 (1)	50.00 (6)	0.00 (2)	6.29 (164)	85.71 (1)
ILL15	0.00	0.99	-	0.00 (7)	14.29 (6)	0.00 (7)	0.00 (6)	4.00 (24)	- (0)
ILL16	0.00	4.08	56.14	0.00 (9)	55.56 (4)	100.00 (0)	0.00 (21)	23.81 (16)	90.00 (1)
ILL17	-	9.85	39.24	0.00 (12)	83.33 (2)	91.67 (1)	- (0)	11.9 (74)	61.11 (7)
ILL18	13.08	52.82	59.37	33.33 (10)	100.00 (0)	100.00 (0)	31.25 (11)	88.24 (2)	100.00 (0)

Table 2. Metrics over the testing period after updating the catalog. The numbers in brackets in the recall and precision columns represent the number of false negatives and false positives, respectively. All percentages are displayed accurately up to two decimals. The symbol "-" means that the corresponding metric could not be computed because no detections were made over the testing period.

mining methods are re-calibrated, and deployed again over the training period. Following two rounds of updates we notice that the upper stations frequently flag segments related to catchment activity as being similar to debris-flows. Such activity includes events such as rockfalls, landslides, and slope failures. Since we are exploring the data, and because this type of activity could related to debris flows, these detections were included as low-confidence debris flow segments in the catalog. After making these changes, we perform one more round of recalibration and update of the catalog over the training period, before deploying the mining methods and updating the catalog over the testing period.

We see that the IF mining method generally outperforms its STA-LTA counterpart with the comparison particularly striking at stations ILL12, ILL13 and ILL18. We found that the STA-LTA method tends to prefer exceedingly long window sizes (see Table B2) to manage sensitivity towards amplitude (see Fig. D1). However, these long window sizes leads to event masking, where a first event will suppress the characteristic function over a neighboring subsequent event. In the case of debris flows, this can lead to false negatives, particularly at more active stations. We provide concrete examples in Appendix D. The results further show that detecting debris flows from stations ILL11, ILL12, ILL13 and ILL18 is relatively easy, because good quality detectors can be obtained here by simply thresholding the IF segment anomaly score. Detection at ILL16 and ILL17 is more difficult and template DTW is needed to discriminate between debris flows and events arising from other sources such as those of an anthropogenic nature. A similar remark applies to ILL14, although the improvement is not as striking. Detection at ILL15 remains difficult.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

3.2 Greenland

3.2.1 Study site

Our Greenlandic site locates on the western coast at the Karrat Fjord (Fig. 1). In this fjord system a $35 - 58$ million m^3 rock avalanche occurred on 17 June 2017 generating a tsunami wave that destroyed parts of the nearby village Nuugaatsiaq and claimed 4 fatalities (Svennevig et al., 2020). The rock avalanche and precursory slip events left clear seismic signatures on the nearby broadband station NUUG, installed in the village Nuugaatsiaq (Poli, 2017; Seydoux et al., 2020). To investigate the detectability of the 17 June 2017 rock avalanche and comparable signals, we focus on station NUUG as well as KARAT, a broadband seismometer that was installed in summer 2022 about 6 km west of the avalanche epicenter. Finally, we also use the broadband station ILULI, which has been operating since 2009 in the village of Ilulissat, approximately 280 km south of Karrat Fjord.

3.2.2 Exploration procedure

We consider generating a catalog from scratch for a specified target seismic station. We first fit the IF to the seismic data of the station and deploy the IF trigger with rule-of-thumb onset and offset thresholds. The top 50 IF segments are then subjected to expert labeling. To aid in this task:

1. We fit an IF to a control station in order to obtain the control IF. The control station should be sufficiently far from the target station so that local events at the latter do not effect the former at the same time, and sufficiently close so that regional/global events effect both stations at around the same time. The argument is that if we observe two high anomaly scores at both stations, this is likely caused by a regional/global event, which in most cases is an earthquake.

2. To compute the control anomaly score for a IF segment of the target station we first limit the segment to 30 minutes. This is achieved by identifying the sliding window with the highest IF anomaly score and iteratively expanding by adding the sliding window in the direction of the larger score. We then compute the maximum IF anomaly score of sliding windows taken from the corresponding segment in the control seismic data using the control IF.

3. We perform DTW between pairs of the top 50 IF segments using the segment DTW approach described in Sect. 2, with the segment length limited as before. Distances are aggregated into a single statistic using the median. Finally, an agglomerative clustering of the top 50 IF segments is performed using the computed pairwise segment DTW distances under complete linkage. We note that the height at which segments merge into a cluster quantifies the diversity of the segments with larger height corresponding to more diversity.

Before considering KARAT, we illustrate the proposed methodology by applying it to seismic data from the NUUG station in the Greenland seismic network in 2017, and summarize the results in Fig. 2. The most anomalous segment of the seismic waveforms according to the IF trigger corresponds to the infamous rock avalanche of 2017 and the uniqueness of this destructive mass movement is further emphasized by the agglomerative clustering.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

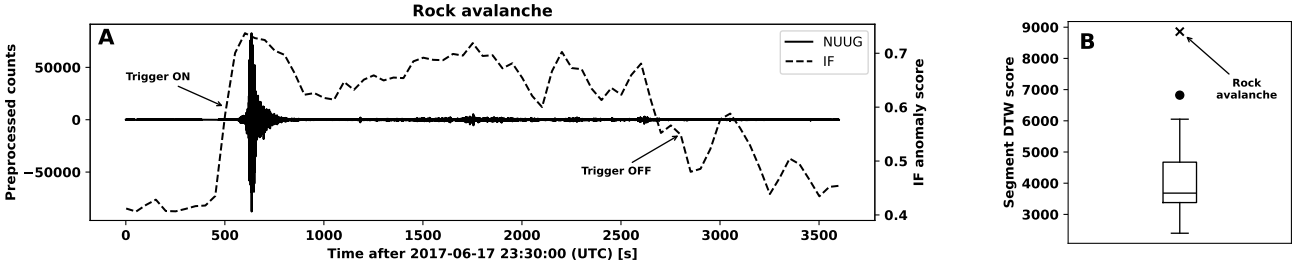

Figure 2. Seismic waveform observed at the NUUG station overlaid with the IF anomaly scores close to the 2017 Rock Avalanche (panel A). The segment represented by the onset- and offset triggers indicated on the plot represents the most anomalous segment flagged in 2017. Panel B shows a box-plot summarizing an agglomerative clustering of the highest 50 anomalous segments flagged by the IF trigger, based on the segment DTW scoring method discussed in Sect. 2.3, under complete linkage. For each of these anomalous segments, we compute the DTW score at which it merges with an existing cluster of segments. The box-plot was constructed from these scores.

3.2.3 KARAT results

We analyze data obtained from the KARAT station in the Greenland seismic network during 2022 and 2023, and use the nearby ILULI station as a control. Details of the expert labeling procedure is given in Appendix E and the results are illustrated in Fig. 3 in the form of a dendrogram constructed from the agglomerative clustering. The first, second, and third columns of the segment labels correspond to the source of the event, rank and control anomaly score respectively. The dendrogram is split into 4 clusters which we describe in increasing order of diversity:

- **Cluster A**. Consists entirely of teleseismic earthquakes.
- **Cluster B**. Consists predominantly of calving events alongside a regional earthquake, iceberg disintegration and a segment we were not able to label.
- **Cluster C**. Consists predominantly of calving events alongside a regional earthquake and some noise signals. Cluster is more diverse compared to cluster B.
- **Cluster D**. Mostly populated by segments flagged before 2022-08-15 when the instrument was streaming sporadically and with high amplitudes (see Fig. 4). Since these likely correspond to issues on the instrumentation side, these segments are labeled as instrument noise. The other two segments in this cluster are caused by helicopters arriving/departing from the station.

Our analysis show that 21 out of the top 50 IF segments corresponds to calving events showcasing the ability of the IF trigger to flag mass movements. The dendrogram shows that segment DTW is able to discriminate well between teleseismic earthquakes, calving events and instrument noise.

The segment DTW does detect diversity in the signals generated by calving events, as indicated by the splitting of these

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

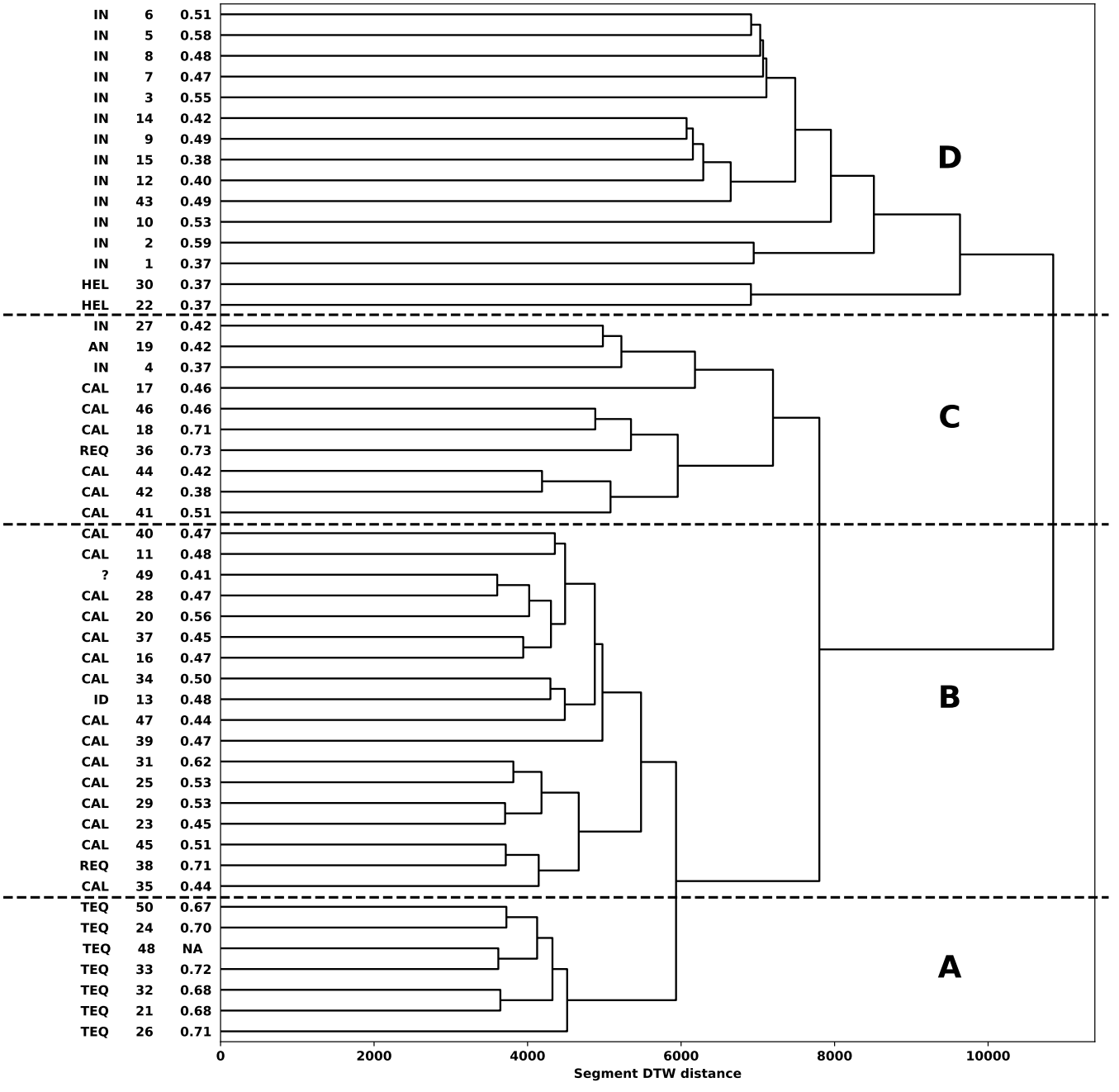

Figure 3. Dendrogram of the top 50 anomalies detected at KARAT. We use IN, HEL, AN, CAL, REQ, TEQ and ID for instrument noise, helicopter, anthropogenic noise, calving events, regional earthquakes, teleseismic earthquakes and iceberg disintegration respectively.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

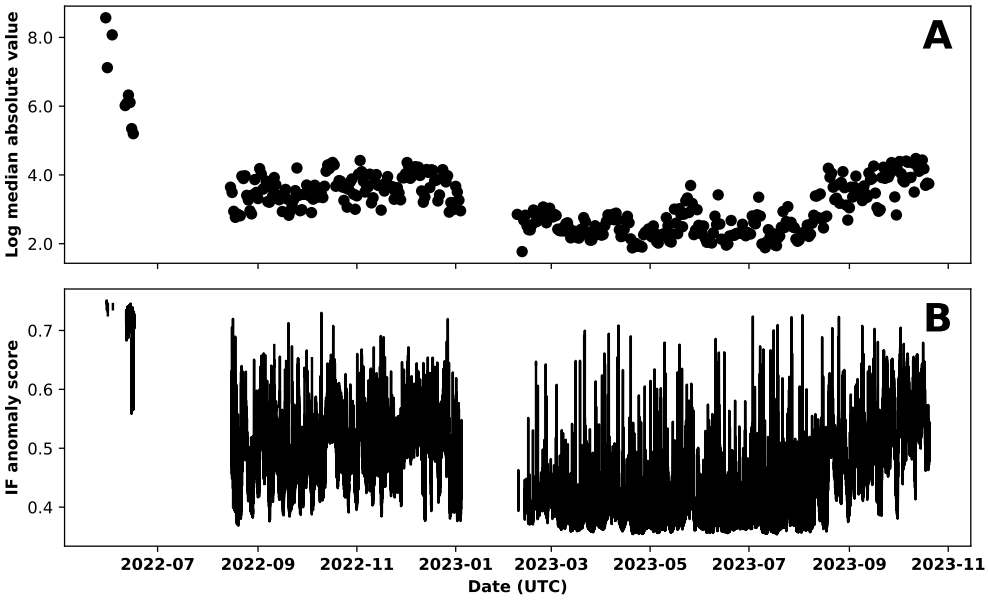

Figure 4. Log median absolute value of the daily preprocessed waveforms (A) and IF anomaly score (B) observed at KARAT.

events into two clusters. Such signals can be diverse due to several reasons. The location of the calving front with respect to the recording station likely plays an important role. High-frequency signals decay are subject to most attenuation, and thus tend to be missing for large source-station distances. Moreover, for relatively small calving events, ground tilt of calving-induced fjord water oscillations ("seiches") can only be detected in the vicinity of the respective fjord (Amundson et al., 2012). Finally, energy partitioning between different > 1 Hz frequency bands may change in response to altered calving front geometries (Walter et al., 2010).

The dendrogram suggests that the segment DTW distance struggles to discriminate between regional earthquake and calving event signals, although it is not clear if enough signals of the former is available to establish a cluster. However, discrimination between these two event sources can be improved by considering the control anomaly scores with high values (≈ 0.70) indicative of a regional earthquake. This works almost perfectly, but for one major calving event that reached the ILULI station.

Fig. 4 suggests both higher amplitudes and IF anomaly scores during the months of September-January. Wind noise, ocean swell, snow cover and other meteorological conditions may explain this observation. To remove the effect of these phenomena on the anomaly scores one can consider training seasonal IF models.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

4 Discussion and further research

We have showcased the ability of the IF trigger to flag mass movements in seismic waveforms to the degree that the method should be considered as an alternative to conventional algorithms when mining seismic data for such events. In particular, we applied IF and STA-LTA triggers to continuous seismic records from a debris flow catchment, which had been subjected to minimal pre-proccessing, and showed that the IF trigger can improve over the classical STA-LTA trigger up to 4 times in terms of the IoU metric. The performance of the STA-LTA trigger could likely be improved by further data processing like band-pass filtering to focus on the most relevant frequencies. However, this requires prior knowledge as source-station distances affect peak frequencies of debris flow seismograms and background noise may pollute certain frequency bands rendering them less suitable for seismic monitoring (Walter et al., 2017; Lai et al., 2018). It was the goal of this study to mine for mass movements without such prior knowledge, and our results show that in this regard the IF trigger is better suited than the STA-LTA trigger.

Since reasonable mass movement detectors can be obtained at some stations just by thresholding the IF anomaly score, this score could serve as a useful feature when building more sophisticated classifiers in addition to those, for example, used in Chmiel et al. (2021); Zhou et al. (2025). Furthermore, running the IF trigger over a network of seismic stations can provide insights into how the network responds to mass-movements and other sources of events. Such insights could include (a) difficulty of detecting mass-movements from different stations (b) identification of other sources significantly effecting stations and (c) examples of how these sources manifest in the seismic waveforms. Within this context, we have shown the ability of DTW-based dissimilarity scores to discriminate between signals arising from various event sources, and to quantify diversity of signals associated with specific sources.

There is a rich literature surrounding anomaly detection that could provide reasonable alternatives to the IF. For example, we could consider extensions of the IF (Hariri et al., 2019; Staerman et al., 2019; Xu et al., 2023) or more broadly anomaly detection methods in the time series context (Blázquez-García et al., 2021; Schmidl et al., 2022). Another avenue for future research is to extend the IF and IF-DTW mining methods of Sect. 3.1 to be online so that they can be used for debris flow detection in real time. In fact, assuming appropriate pre-processing, the IF method is already online since a detection can be labeled as a debris flow the moment the IF anomaly score hits the score threshold, subject to the minimum detection length requirement. The IF-DTW strategy can be made online by streaming the DTW distances of sliding windows relative to the templates the moment the IF trigger activates. Care should be taken in terms of the computational cost associated with this approach. A more lightweight alternative is to only update the DTW distances if a new sliding window is the most anomalous window observed since the trigger activated. An alternative to using the mean of the DTW distances for scoring segments is to use the distances as features for a machine learning model, possibly including templates from other events as well (Wu et al., 2018). Finally, it is not clear if DTW is the most appropriate method for measuring dissimilarity between signals. A promising alternative is to use contrastive approaches (Franceschi et al., 2019; Yue et al., 2022) which has been applied within the context of seismology (Meyer et al., 2021). Contrastive learning and DTW hybrids are also a possibility.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

Code and data availability. The seismic data used for the Illgraben case study are available at Swiss Seismological Service (SED) at ETH Zurich (2012). The source code is available at https://github.com/FKamper/seismic-isolation-forest.

Author contributions. FK, FW and PP conceptualized the study and was responsible for data curation. FK, FW and PP developed analysis methodology. FK and PP developed the software. FK performed the formal analysis. FW, MV, MM and MS provided supervision. MV, MM and MS provided validation. FK, FW, PP, MM, MV and MS wrote the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This project has been partly supported by the SDSC collaborative grant "DATSSFLOW" C21-03. Although all content were developed by the authors, GPT-4-turbo was used for code-related queries and GitHub Copilot (version 1.350.0) for doc-string generation and code completion. Any suggestion made by AI tools were reviewed by the authors.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

References

Allen, R. V.: Automatic earthquake recognition and timing from single traces, Bull. Seismol. Soc. Am., 68, 1521–1532, https://doi.org/10.1785/BSSA0680051521, 1978.

Amundson, J. M., Clinton, J. F., Fahnestock, M., Truffer, M., Lüthi, M. P., and Motyka, R. J.: Observing calving-generated ocean waves with coastal broadband seismometers, Jakobshavn Isbræ, Greenland, Ann. Glaciol., 53, 79–84, https://doi.org/10.3189/2012/AoG60A200, 2012.

Badoux, A., Graf, C., Rhyner, J., Kuntner, R., and McArdell, B. W.: A debris-flow alarm system for the Alpine Illgraben catchment: design and performance, Nat. Hazards, 49, 517–539, https://doi.org/10.1007/s11069-008-9303-x, 2009.

Bahavar, M., Allstadt, K. E., Van Fossen, M., Malone, S. D., and Trabant, C.: Exotic seismic events catalog (ESEC) data product, Seismol. Res. Lett., 90, 1355–1363, https://doi.org/https://doi.org/10.1785/0220180402, 2019.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J.: ObsPy: A Python toolbox for seismology, Seismol. Res. Lett., 81, 530–533, https://doi.org/10.1785/gssrl.81.3.530, 2010.

Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. A.: A review on outlier/anomaly detection in time series data, ACM Comput. Surv., 54, 1–33, https://doi.org/10.1145/3444690, 2021.

Bouman, R., Bukhsh, Z., and Heskes, T.: Unsupervised anomaly detection algorithms on real-world data: how many do we need?, J. Mach. Learn. Res., 25, 1–34, https://www.jmlr.org/papers/v25/23-0570.html, 2024.

Chmiel, M., Walter, F., Wenner, M., Zhang, Z., McArdell, B. W., and Hibert, C.: Machine learning improves debris flow warning, Geophys. Res. Lett., 48, e2020GL090 874, https://doi.org/10.1029/2020GL090874, 2021.

Fichtner, A., Bowden, D., and Ermert, L.: Optimal processing for seismic noise correlations, Geophys. J. Int., 223, 1548–1564, https://doi.org/10.1093/gji/ggaa390, 2020.

Franceschi, J.-Y., Dieuleveut, A., and Jaggi, M.: Unsupervised scalable representation learning for multivariate time series, in: Adv. Neural Inf. Process. Syst., edited by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., vol. 32, https://proceedings.neurips.cc/paper_files/paper/2019/file/53c6de78244e9f528eb3e1cda69699bb-Paper.pdf, 2019.

Geological Survey of Denmark and Greenland: Registered earthquakes in Greenland, https://www.geus.dk/natur-og-klima/jordskaelv-og-seismologi/registrerede-jordskaelv-i-groenland, data list generated automatically: 2025-07-13 14:17 UTC, 2025.

Hariri, S., Kind, M. C., and Brunner, R. J.: Extended isolation forest, IEEE Trans. Knowl. Data Eng, 33, 1479–1489, https://doi.org/10.1109/TKDE.2019.2947676, 2019.

Hürlimann, M., Rickenmann, D., and Graf, C.: Field and monitoring data of debris-flow events in the Swiss Alps, Can. Geotech. J., 40, 161–175, https://doi.org/10.1139/t02-087, 2003.

Igel, J. K., Ermert, L. A., and Fichtner, A.: Rapid finite-frequency microseismic noise source inversion at regional to global scales, Geophys. J. Int., 227, 169–183, https://doi.org/10.1093/gji/ggab210, 2021.

Lai, V. H., Tsai, V. C., Lamb, M. P., Ulizio, T. P., and Beer, A. R.: The seismic signature of debris flows: Flow mechanics and early warning at Montecito, California, Geophysical Research Letters, 45, 5528–5535, 2018.

Larose, E., Carrière, S., Voisin, C., Bottelin, P., Baillet, L., Guéguen, P., Walter, F., Jongmans, D., Guillier, B., Garambois, S., et al.: Environmental seismology: What can we learn on earth surface processes with ambient noise?, J. Appl. Geophys., 116, 62–74, https://doi.org/10.1016/j.jappgeo.2015.02.001, 2015.

Liu, F. T., Ting, K. M., and Zhou, Z.-H.: Isolation forest, in: IEEE Data Mining, pp. 413–422, https://doi.org/10.1109/ICDM.2008.17, 2008.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

Liu, F. T., Ting, K. M., and Zhou, Z.-H.: Isolation-based anomaly detection, ACM Trans. Knowl. Discov. Data, 6, 1–39, https://doi.org/10.1145/2133360.2133363, 2012.

Longuet-Higgins, M. S.: A theory of the origin of microseisms, Philos. Trans. R. Soc. A, 243, 1–35, https://doi.org/https://doi.org/10.1098/rsta.1950.0012, 1950.

McNamara, D. E. and Buland, R. P.: Ambient noise levels in the continental United States, Bull. Seismol. Soc. Am., 94, 1517–1527, https://doi.org/10.1785/012003001, 2004.

Medrzycka, D., Benn, D. I., Box, J. E., Copland, L., and Balog, J.: Calving behavior at Rink Isbræ, West Greenland, from time-lapse photos, Arct. Antarct. Alp. Res., 48, 263–277, https://doi.org/10.1657/AAAR0015-059, 2016.

Meyer, M., Weber, S., Beutel, J., and Thiele, L.: Systematic identification of external influences in multi-year microseismic recordings using convolutional neural networks, Earth Surf. Dyn., 7, 171–190, https://doi.org/10.5194/esurf-7-171-2019, 2019.

Meyer, M., Wenner, M., Hibert, C., Walter, F., and Thiele, L.: Using system context information to complement weakly labeled data, in: Workshop on weakly supervised learning, co-located with Int. Conf. Learn. Represent. (ICLR) 2021, https://arxiv.org/abs/2107.10236, 2021.

Montagner, J.-P., Mangeney, A., and Stutzmann, E.: Seismology and environment, in: Encyclopedia of Solid Earth Geophysics, Encyclopedia of Earth Sciences Series, edited by Gupta, H. K., Springer, Cham, https://doi.org/10.1007/978-3-030-58631-7_258, 2021.

Nakata, N., Gualtieri, L., and Fichtner, A.: Seismic ambient noise, Cambridge University Press, Cambridge, United Kingdom, ISBN 9781107195423, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825–2830, https://www.jmlr.org/papers/v12/pedregosa11a.html, 2011.

Poli, P.: Creep and slip: Seismic precursors to the Nuugaatsiaq landslide (Greenland), Geophys. Res. Lett., 44, 8832–8836, https://doi.org/10.1002/2017GL075039, 2017.

Rodríguez Tribaldos, V. and Ajo-Franklin, J. B.: Aquifer monitoring using ambient seismic noise recorded with distributed acoustic sensing (DAS) deployed on dark fiber, J. Geophys. Res. Solid Earth, 126, e2020JB021 004, https://doi.org/10.1029/2020JB021004, 2021.

Sager, K., Boehm, C., Ermert, L., Krischer, L., and Fichtner, A.: Global-scale full-waveform ambient noise inversion, J. Geophys. Res. Solid Earth, 125, e2019JB018 644, https://doi.org/10.1029/2019JB018644, 2020.

Salvador, S. and Chan, P.: Toward accurate dynamic time warping in linear time and space, Intell. Data Anal., 11, 561–580, https://cs.fit.edu/~pkc/papers/tdm04.pdf, 2007.

Schmidl, S., Wenig, P., and Papenbrock, T.: Anomaly detection in time series: a comprehensive evaluation, Proc. VLDB Endow., 15, 1779–1797, https://doi.org/10.14778/3538598.3538602, 2022.

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. d., Campillo, M., and Baraniuk, R.: Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning, Nat. Commun., 11, 3972, https://doi.org/10.1038/s41467-020-17841-x, 2020.

Staerman, G., Mozharovskyi, P., Clémençon, S., and d'Alché Buc, F.: Functional isolation forest, in: Asian Conf. Mach. Learn., pp. 332–347, PMLR, https://proceedings.mlr.press/v101/staerman19a.html, 2019.

Svennevig, K., Dahl-Jensen, T., Keiding, M., Merryman Boncori, J. P., Larsen, T. B., Salehi, S., Munck Solgaard, A., and Voss, P. H.: Evolution of events before and after the 17 June 2017 rock avalanche at Karrat Fjord, West Greenland–a multidisciplinary approach to detecting and locating unstable rock slopes in a remote Arctic area, Earth Surf. Dyn., 8, 1021–1038, https://doi.org/10.5194/esurf-8-1021-2020, 2020.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

Swiss Seismological Service (SED) at ETH Zurich: 9s - Temporary deployments in Switzerland associated with landslides, https://doi.org/10.12686/SED/NETWORKS/XP, 2012.

Umlauft, J., Johnson, C. W., Roux, P., Trugman, D. T., Lecointre, A., Walpersdorf, A., Nanni, U., Gimbert, F., Rouet-Leduc, B., Hulbert, C., et al.: Mapping glacier basal sliding applying machine learning, J. Geophys. Res. Earth Surface, 128, e2023JF007 280, https://doi.org/10.1029/2023JF007280, 2023.

U.S. Geological Survey: Earthquake catalog (1568 to 2018) for the USGS national seismic hazard model and nuclear regulatory commission, https://doi.org/10.5066/P95SNP2J, accessed 2025-06-27, 2023.

Walter, F., O'Neel, S., McNamara, D., Pfeffer, W., Bassis, J. N., and Fricker, H. A.: Iceberg calving during transition from grounded to floating ice: Columbia Glacier, Alaska, Geophys. Res. Lett., 37, https://doi.org/10.1029/2010GL043201, 2010.

Walter, F., Amundson, J. M., O'Neel, S., Truffer, M., Fahnestock, M., and Fricker, H. A.: Analysis of low-frequency seismic signals generated during a multiple-iceberg calving event at Jakobshavn Isbræ, Greenland, J. Geophys. Res. Earth Surf., 117, https://doi.org/10.1029/2011JF002132, 2012.

Walter, F., Burtin, A., McArdell, B. W., Hovius, N., Weder, B., and Turowski, J. M.: Testing seismic amplitude source location for fast debris-flow detection at Illgraben, Switzerland, Nat. Hazards Earth Syst. Sci., 17, 939–955, https://doi.org/10.5194/nhess-17-939-2017, 2017.

Wenner, M., Hibert, C., van Herwijnen, A., Meier, L., and Walter, F.: Near-real-time automated classification of seismic signals of slope failures with continuous random forests, Nat. Hazards Earth Syst. Sci., 21, 339–361, https://doi.org/10.5194/nhess-21-339-2021, 2021.

Wu, L., Yen, I. E.-H., Yi, J., Xu, F., Lei, Q., and Witbrock, M.: Random warping series: A random features method for time-series embedding, in: Int. Conf. Artif. Intell. Stat., pp. 793–802, PMLR, https://proceedings.mlr.press/v84/wu18b/wu18b.pdf, 2018.

Xu, H., Pang, G., Wang, Y., and Wang, Y.: Deep isolation forest for anomaly detection, IEEE Trans. Knowl. Data Eng., 35, 12 591–12 604, https://doi.org/10.1109/TKDE.2023.3270293, 2023.

Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., and Xu, B.: Ts2vec: Towards universal representation of time series, in: AAAI Conf. Artif. Intell., vol. 25, pp. 8980–8987, https://doi.org/10.1609/aaai.v36i8.20881, 2022.

Zhou, Q., Tang, H., Hibert, C., Chmiel, M., Walter, F., Dietze, M., and Turowski, J. M.: Enhancing debris flow warning via machine learning feature reduction and model selection, J. Geophys. Res. Earth Surf., 130, e2024JF008 094, https://doi.org/10.1029/2024JF008094, 2025.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

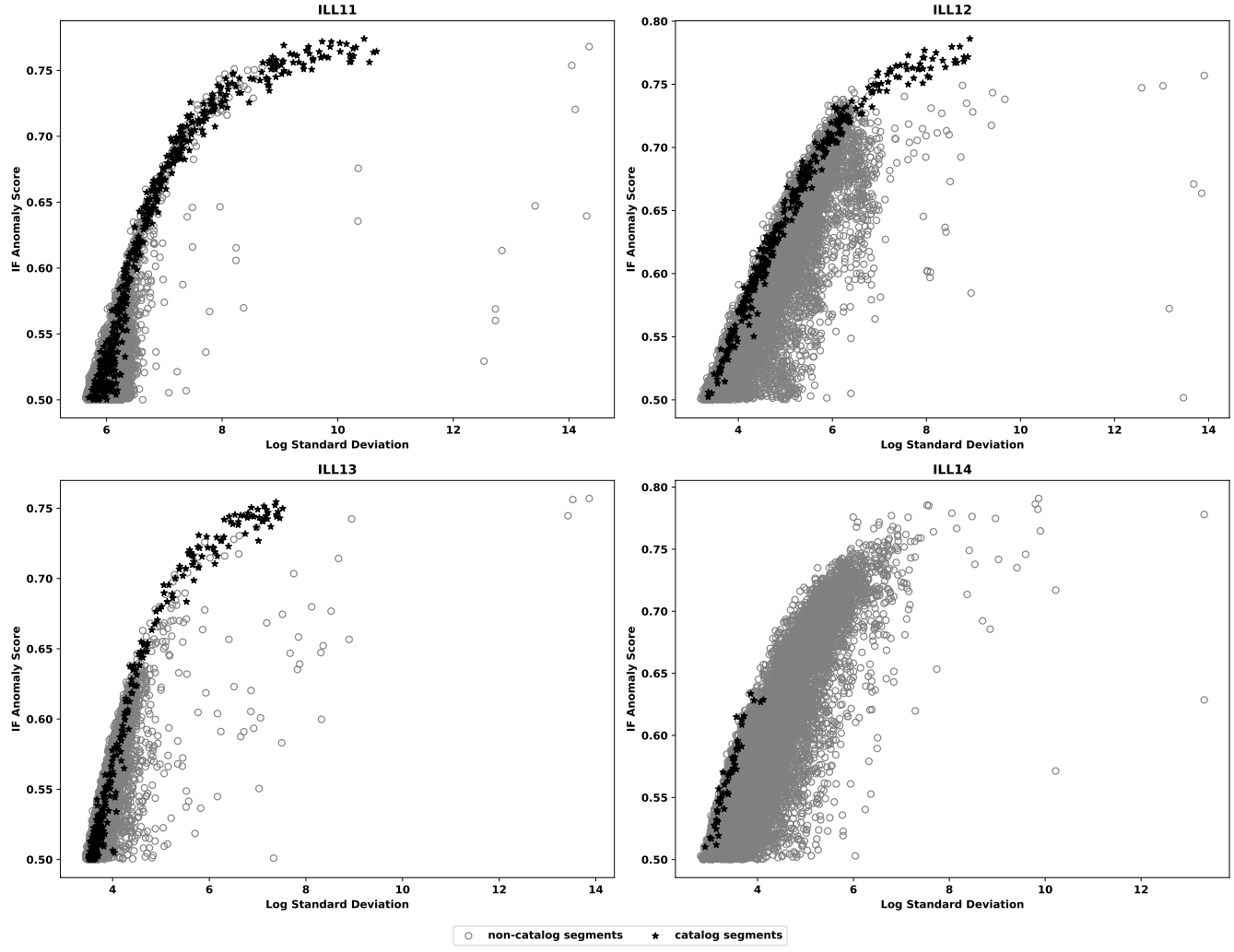

Figure A1. Plots of the IF anomaly score vs the log standard deviation of sliding windows taken from 2018 for ILL11-ILL14. The plots are restricted to show only those sliding windows with an anomaly score exceeding 0.5.

Appendix A: Isolation forest anomaly score

Figure A1 and A2 shows the IF anomaly score plotted against log standard deviation for each station in the Illgraben seismic network for 2018. Sliding windows overlapping with catalog segments (all confidence levels) are indicated by star marks. The relationship forms a non-linear wave-like pattern, with the IF anomaly scores of debris flow segments at stations ILL11, ILL12, ILL13 and ILL18 highly ranked. Interestingly, even though the IF anomaly score of debris-flow segments at the other stations do not rank as highly, they are fairly highly rank in the log standard deviation bands in which they appear.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

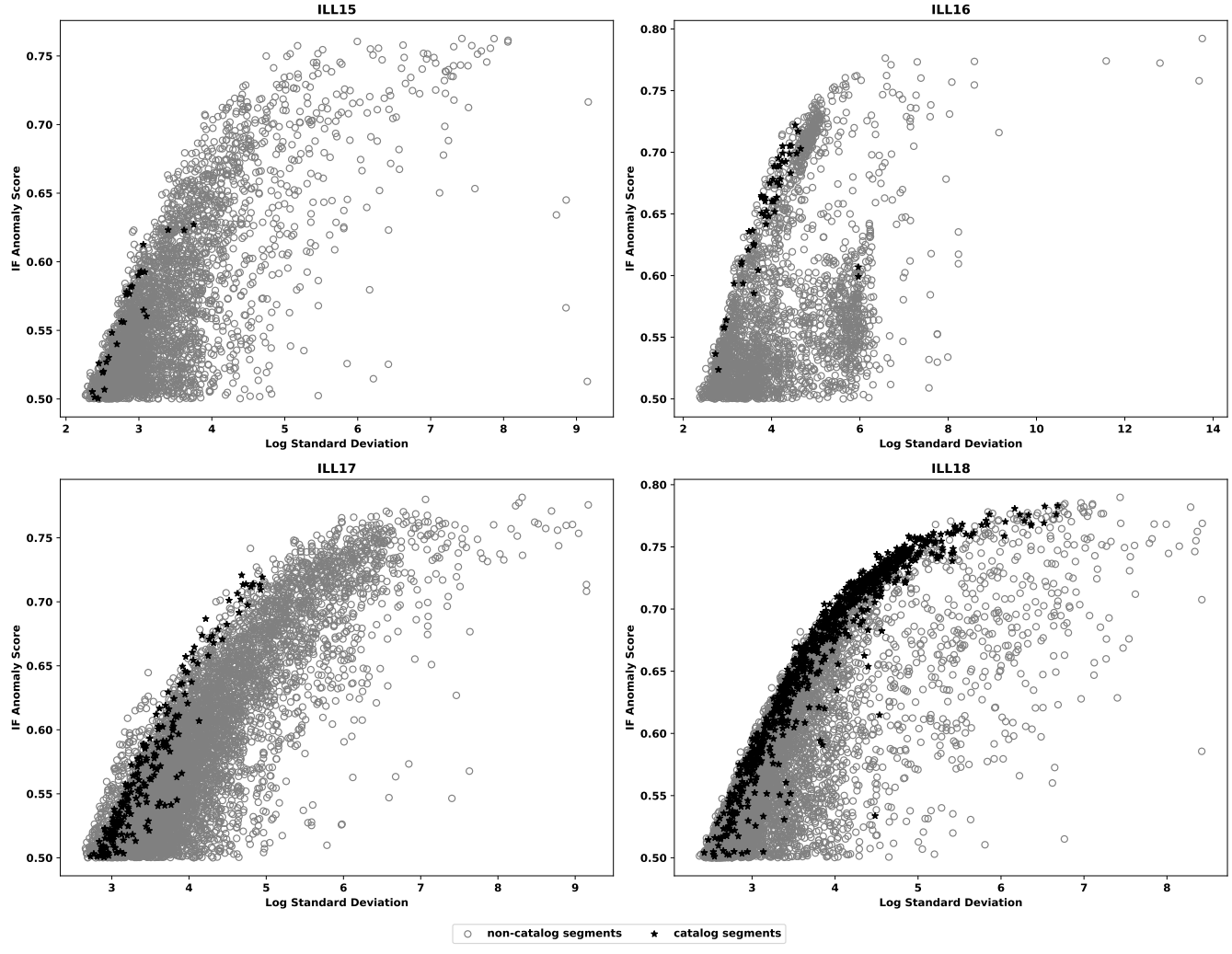

Figure A2. Plots of the IF anomaly score vs the log standard deviation of sliding windows taken from 2018 for ILL15-ILL18. The plots are restricted to show only those sliding windows with an anomaly score exceeding 0.5.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

	IF Trigger		IF		IF-DTW	
Station	Onset Threshold	Offset Threshold	Score Threshold	Minimum Detection Length	Score Threshold	Minimum Detection Length
ILL11	0.65	0.65	0.73	3.33	8313.60	10.00
ILL12	0.70	0.65	0.76	14.17	7223.35	34.17
ILL13	0.65	0.65	0.75	6.67	6792.46	11.67
ILL14	0.55	0.50	0.61	14.17	6774.87	2.50
ILL15	0.55	0.50	0.69	24.17	6645.90	3.33
ILL16	0.60	0.50	0.68	11.67	6714.87	10.83
ILL17	0.55	0.50	0.61	27.50	7278.21	15.00
ILL18	0.65	0.65	0.75	13.33	7496.32	22.50

Table B1. Hyper parameters selected for the IF and IF-DTW mining methods. The minimum detection lengths are given in minutes. All values displayed are accurate up to two decimals.

Station	Onset Threshold	Offset Threshold	Short-term Window Size	Long-term Window Size	Score Threshold	Minimum Detection Length
ILL11	6.00	0.12	8.33	83.33	8.09	32.38
ILL12	12.00	0.06	8.33	166.67	19.91	33.21
ILL13	12.00	0.06	4.17	83.33	17.26	22.09
ILL14	3.00	0.50	33.33	166.67	3.85	274.29
ILL15	3.00	2.00	16.67	333.33	6.81	119.15
ILL16	12.00	0.50	4.17	333.33	32.45	76.83
ILL17	3.00	0.50	33.33	666.67	13.70	355.72
ILL18	24.00	0.50	8.33	666.67	41.13	39.44

Table B2. Hyper parameters selected for STA-LTA mining method. The minimum detection lengths, short- and long-term windows are given in minutes. All values displayed are accurate up to two decimals.

Appendix B: Mining methods hyper-parameters

Table B1 contain the hyper parameters selected by the calibration procedure for the IF and IF-DTW mining methods, while Table B2 contains those selected for the STA-LTA method.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

	Number of events		Low-confidence recall (%)			Medium-confidence recall (%)		
Station	Low Confidence	Medium Confidence	STA-LTA	IF	IF-DTW	STA-LTA	IF	IF-DTW
ILL11	9	2	22.22	66.67	100.00	50.00	0.00	50.00
ILL12	2	3	0.00	0.00	0.00	0.00	33.33	66.67
ILL13	2	2	0.00	0.00	0.00	0.00	0.00	0.00
ILL14	5	5	20.00	20.00	40.00	0.00	40.00	0.00
ILL15	4	3	0.00	0.00	0.00	0.00	33.33	0.00
ILL16	11	3	0.00	9.09	45.45	0.00	33.33	0.00
ILL17	8	6	0.00	37.50	62.50	16.67	16.67	33.33
ILL18	9	6	0.00	22.22	44.44	50.00	50.00	50.00
Overall	50	30	5.28	19.43	36.55	14.58	25.83	25.00

Table C1. Number of events for each confidence class and recall of lower-confidence segments for the different mining strategies over the training period. All values displayed are accurate up to two decimals.

	Number of events		Low-confidence recall (%)			Medium-confidence recall (%)		
Station	Low Confidence	Medium Confidence	STA-LTA	IF	IF-DTW	STA-LTA	IF	IF-DTW
ILL11	4	2	0.00	0.00	75.00	0.00	50.00	50.00
ILL12	3	1	0.00	33.33	33.33	0.00	0.00	100.00
ILL13	1	2	0.00	0.00	0.00	0.00	0.00	50.00
ILL14	47	0	0.00	8.51	72.34	0.00	0.00	0.00
ILL15	4	0	0.00	0.00	0.00	0.00	0.00	0.00
ILL16	3	3	0.00	0.00	33.33	0.00	33.33	66.67
ILL17	24	2	0.00	33.33	45.83	0.00	100.00	100.00
ILL18	24	3	20.83	50.00	20.83	33.33	100.00	100.00
Overall	110	13	2.60	15.65	35.08	4.17	35.42	58.33

Table C2. Number of events for each confidence class and recall of lower-confidence segments for the different mining strategies over the testing period. All values displayed are accurate up to two decimals.

Appendix C: Recall of lower-confidence segments

Table C1 and C2 show the recall achieved by the various mining strategies for the lower-confidence segments over the training and test period respectively. There are more lower- than medium-confidence debris flow segments partly due to the inclusion of catchment and other activity in the lower-confidence class. Overall, the IF-DTW strategy exhibit the highest recall, followed by IF and then STA-LTA.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

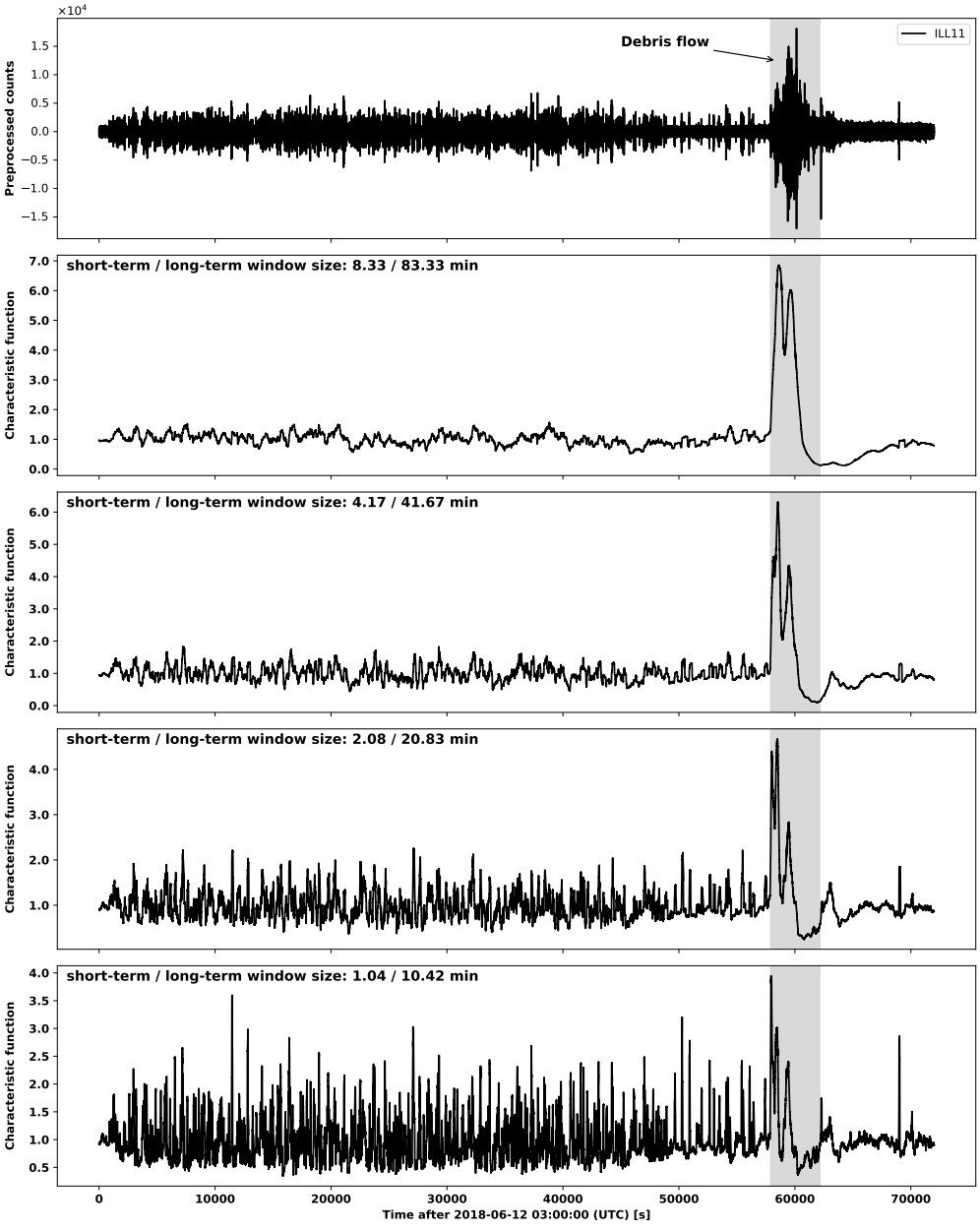

Figure D1. Illustration of the effect of the window sizes on the characteristic function of the STA-LTA trigger.

Appendix D: STA-LTA examples

STA-LTA triggers are known to be sensitive to changes in the amplitude of seismic waveforms. To better capture debris flows, the STA-LTA trigger accommodates for this by taking exceedingly long window lengths, sometimes spanning hours (see Table

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

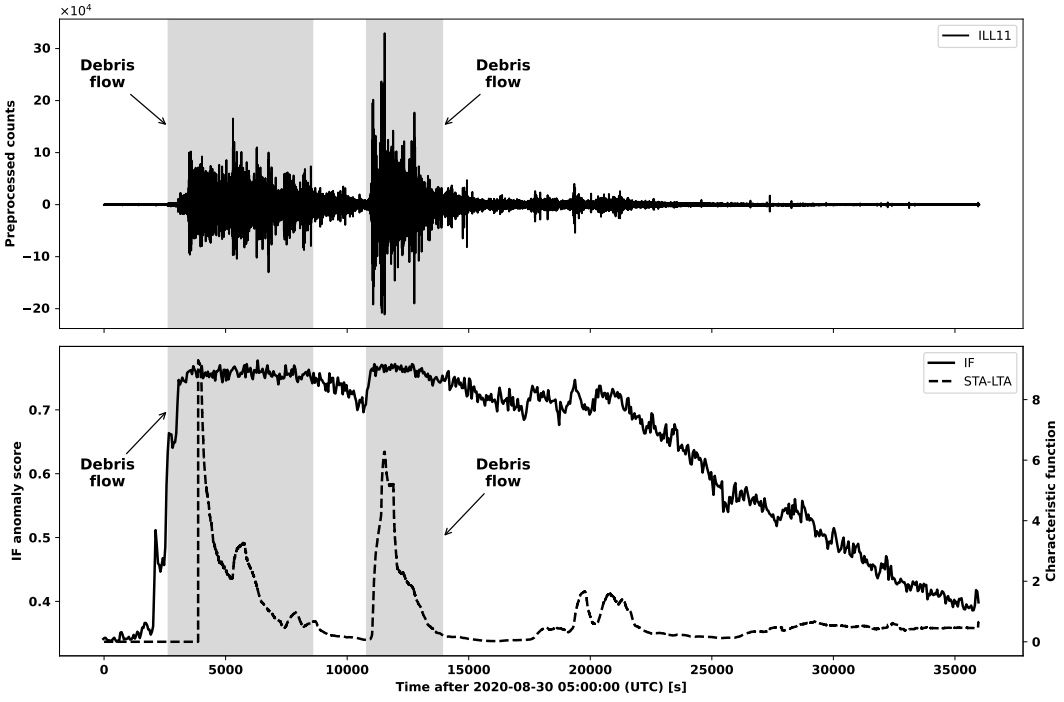

Figure D2. Illustration of the behavior of the STA-LTA characteristic function relative to the IF anomaly score at ILL11 on 2020-08-30. Debris flows are represented by the shaded regions.

B2). We illustrate this in Fig. D1, where we study the behavior of the STA-LTA trigger in relation to the seismic waveform observed at ILL11 on 2018-06-12, which contains a debris flow. In all plots, the debris flow is represented by the shaded region. The top graph shows the preprocessed waveform, and the second graph shows the characteristic function of the STA-LTA trigger with the short- and long-term windows given in Table B2. In the remaining plots the window sizes of the STA-LTA trigger are successively divided by two as we proceed towards the bottom. As the window sizes become smaller, it becomes harder to see where the debris flow manifests in the characteristic function.

Having longer window sizes is not without consequence. One particular issue arises when there is increased amplitude (for whatever reason) in the seismic waveform within the long-term or short-term window before a debris flow occurs. Here, the averaging suppresses the characteristic function over the debris-flow period relative to the case if the increase in amplitude did not occur. Managing the trade-off between this phenomenon and the sensitivity towards amplitude can be difficult, particularly in more active stations. We give three examples in Figs. D2, D3 and D4 where two debris-flows occur relatively close in time. The characteristic function over the period associated with the second debris flow is suppressed by the increased amplitude in the seismic waveform over the period associated with the first, leading to false negatives. The IF anomaly score does not suffer from this issue.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

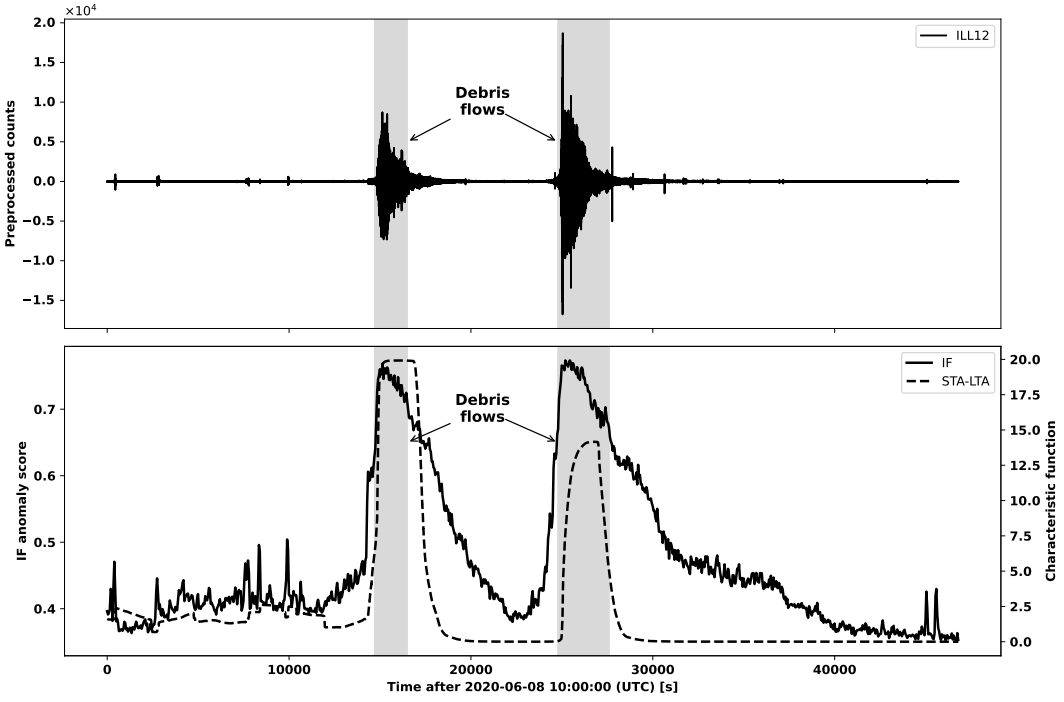

Figure D3. Illustration of the behavior of the STA-LTA characteristic function relative to the IF anomaly score at ILL12 on 2020-06-08. Debris flows are represented by the shaded regions.

Appendix E: Greenland labeling

We provide some insight into how events were labeled in the Greenland data. First the seismic waveforms and corresponding spectograms around the segment flagged by the IF trigger is investigated by a domain scientist and a label is recommended based on well-known characteristics of calving seismograms (see Fig. E1). Once a label is recommended additional verification are performed if possible. For example:

1. **Calving events**. Satellite images such as those depicted in Fig. E2.
2. **Teleseismic earthquakes**. USGS earthquake catalog (U.S. Geological Survey, 2023), see Table E1.
3. **Regional earthquakes**. GEUS earthquake catalog (Geological Survey of Denmark and Greenland, 2025), see Table E2.

Satellite image availability is contingent upon cloud-free conditions and thus often does not allow for a ground-truth check. For this study we focused on Rink Glacier, the most active calving front near station KARAT 1. Figures E1, E2, E3 and E4 contain examples of a well-constrained calving event and an iceberg capsizing event. The regional and teleseismic earthquake catalogs are considered reliable ground-truth sources.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

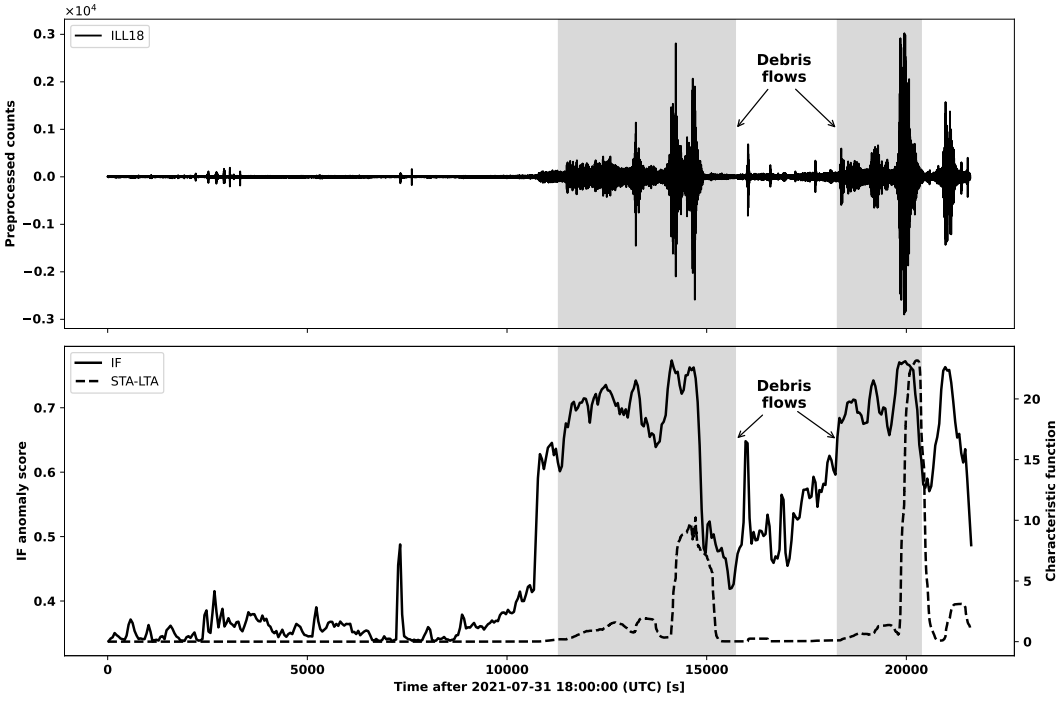

Figure D4. Illustration of the behavior of the STA-LTA characteristic function relative to the IF anomaly score at ILL18 on 2021-07-31. Debris flows are represented by the shaded regions.

Rank	Start	Stop	Remarks
21	2022-09-19T18:14:59.410000Z	2022-09-19T18:18:19.410000Z	M 7.6 - 35 km SSW of Aguililla, Mexico
24	2023-09-08T22:19:10.000000Z	2023-09-08T22:28:20.000000Z	M 6.8 - Al Haouz, Morocco
26	2022-09-19T18:35:16.365000Z	2022-09-19T18:41:06.365000Z	M 7.6 - 35 km SSW of Aguililla, Mexico
32	2023-07-16T06:55:42.320000Z	2023-07-16T07:04:52.320000Z	M 7.2 - 106 km S of Sand Point, Alaska
33	2023-03-21T16:57:30.000000Z	2023-03-21T17:05:50.000000Z	M 6.5 - 40 km SSE of Jurm, Afghanistan
48	2023-10-16T09:47:30.000000Z	2023-10-16T15:35:50.000000Z	M 6.4 - 78 km NNW of Adak, Alaska
50	2023-05-19T03:15:42.320000Z	2023-05-19T03:23:12.320000Z	M 7.7 - southeast of the Loyalty Islands

Table E1. Teleseismic earthquakes in cluster A of Fig. 3.

Rank	Start	Stop	Remarks
36	2023-03-21T06:57:30.000000Z	2023-03-21T07:01:40.000000Z	M: 3.9. Latitude: 69.088°N. Longitude: 53.429°W.
38	2023-04-19T01:34:02.320000Z	2023-04-19T01:38:12.320000Z	M: 4.1. Latitude: 65.913°N. Longitude: 37.537°W.

Table E2. Regional earthquakes in clusters B and C of Fig. 3.

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

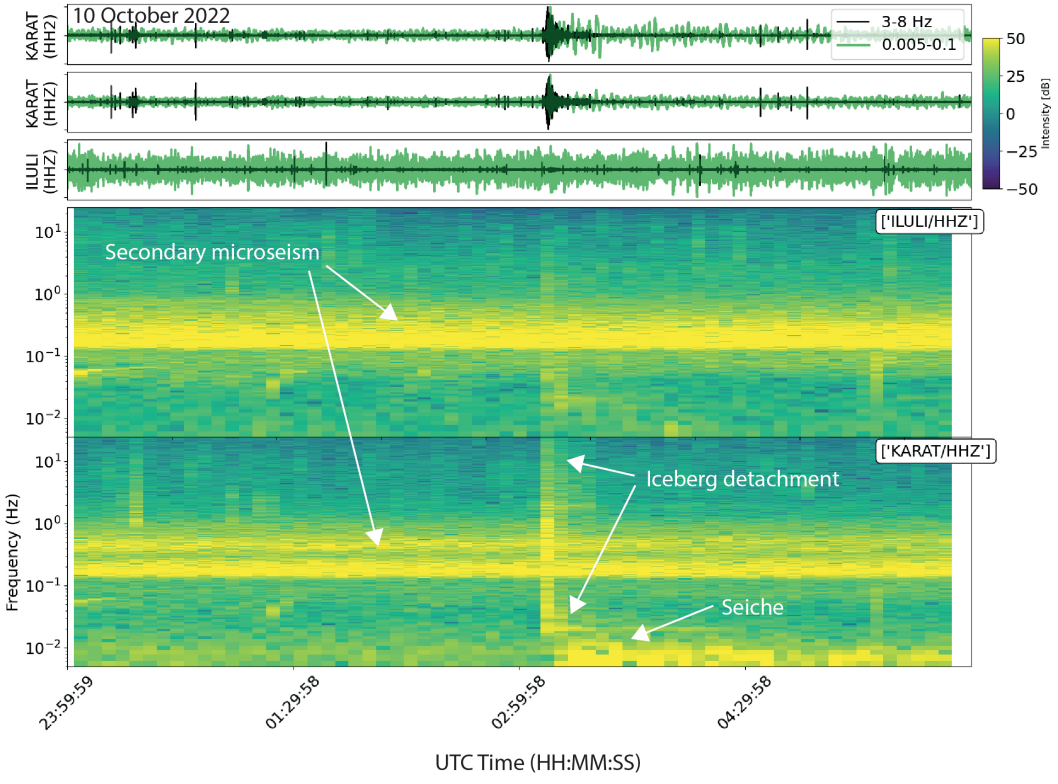

Figure E1. Seismic waveforms and corresponding spectograms around the segment flagged by the IF trigger on 2022-10-10. One horizontal component (HH2) and the vertical component (HHZ) are shown for KARAT and the vertical component is shown for ILULI. The spectrograms show the continuous energy of the secondary microseism generated by standing waves in ocean basins (Longuet-Higgins, 1950; McNamara and Buland, 2004). Moreover, the IF trigger flags a typical calving seismogram with broadband signals representing the iceberg detachment (Walter et al., 2012) and a low-frequency (<0.01 Hz) signal generated by calving-induced water oscillations within the fjord ("seiche"; Amundson et al., 2012).

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

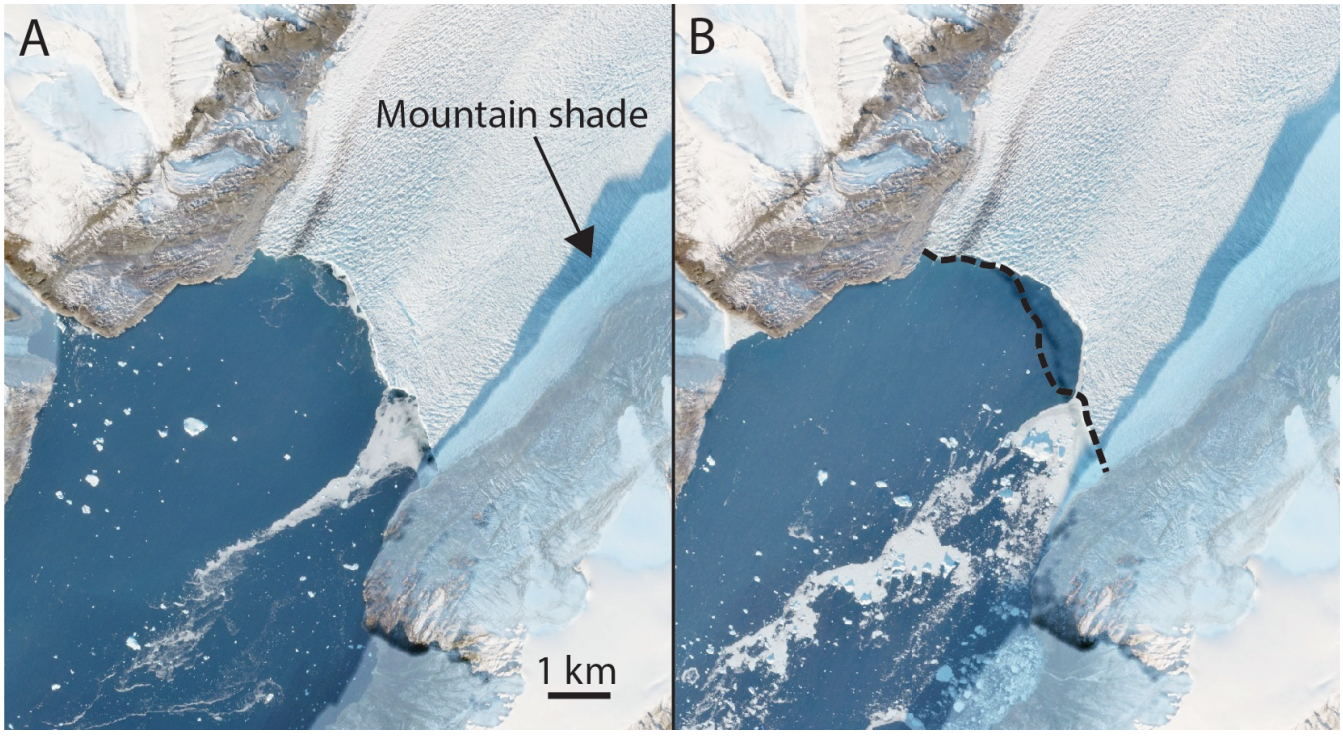

Figure E2. Satellite image pair of Rink Glacier calving front (Fig. 1) on 2022-10-07 (A) and 2022-10-12 (B) before and after the calving event 2022-10-10, respectively. The black dashed line represents the before-calving terminus and the missing area indicates a calving volume of about $0.5\,\mathrm{km}^3$ assuming a terminus thickness of $500-600\,\mathrm{m}$ (Medrzycka et al., 2016). Source: Sources: Copernicus (Sentinel-2 true color image).

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

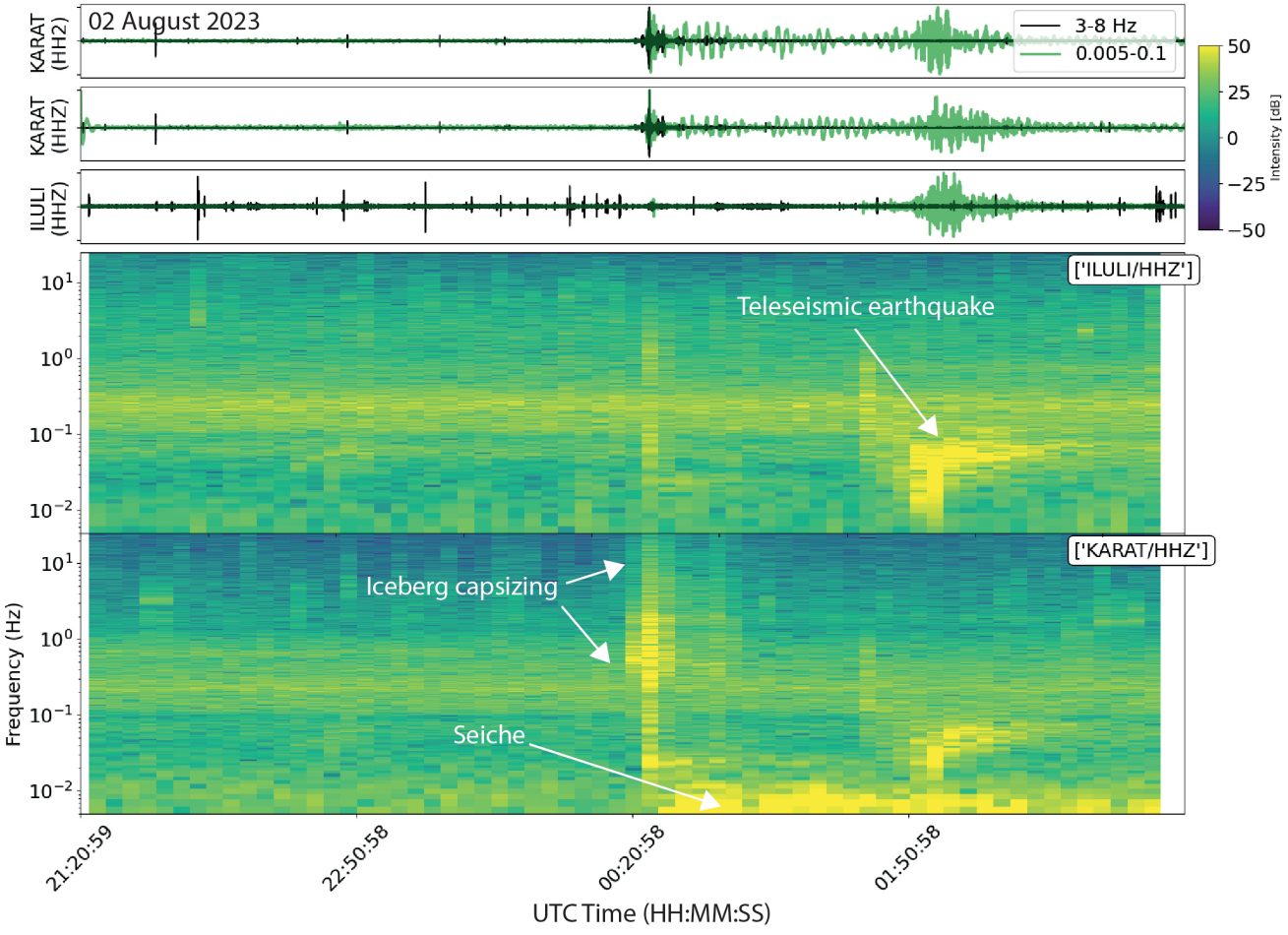

Figure E3. Similar to Fig. E1, except that it shows the IF detection of a seismic event, which according to satellite images constitutes the capsizing of a tabular iceberg (Fig. E4). The tabular iceberg was within 500 m of the calving front and thus likely contacted the calving front as it capsized. This generated a broadband signal similar to iceberg detachment (Fig. E1). Shortly after the capsizing, both KARAT and ILULI recorded a teleseismic earthquake (M5.9, 266 km South of Burica, Panama, UTC time: 2023-08-03 01:25:21, location 5.640 °N 82.606 °W) (U.S. Geological Survey, 2023)).

https://doi.org/10.5194/egusphere-2025-3864
Preprint. Discussion started: 9 September 2025

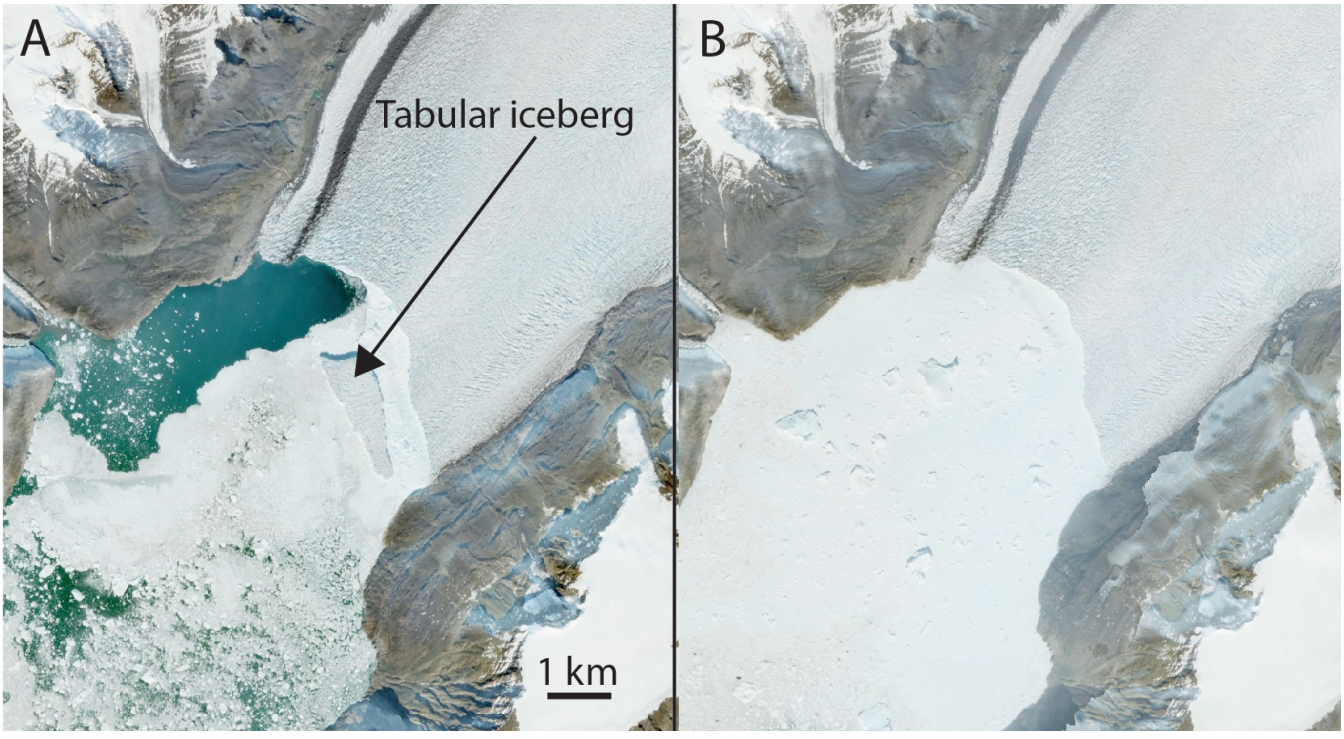

Figure E4. Satellite image pair of Rink Glacier calving front (Fig. 1) on 2023-08-01 (A) and 2023-08-03 (B) before and after the IF trigger segment on 2023-08-02, respectively. Assuming a full-thickness iceberg with a depth of $500 - 600$ m (Medrzycka et al., 2016), the iceberg had a volume of about $0.5\,\mathrm{km}^3$ and may have contacted the terminus during capsize. Source: Copernicus (Sentinel-2 true color image).