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

Francois Kamper[1], Fabian Walter[3], Patrick Paitz[3], Matthias Meyer[2], Michele Volpi[2], and Mathieu Salzmann[1]

[1]Swiss Data Science Center, École Polytechnique Fédérale de Lausanne, EPFL INN Building, Station 14, 1015 Lausanne, Switzerland
[2]Swiss Data Science Center, Eidgenössische Technische Hochschule Zürich, Andreasstrasse 5, 8092 Zurich, Switzerland
[3]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland

**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~~ Traditionally, global and regional earthquakes are the main focus of existing automated processing workflows by national and international seismological organizations~~,~~. Event detection and arrival time picking for earthquake source location are consequently standard tasks, which until recently have required experts to manually classify seismic transients into earthquake-related signals or other types of events. Nowadays, however, machine learning takes over these processing routines facilitated by the existence of large manually labeled data volumes available for algorithm training (**?**, and references therein). Machine learning algorithms can be trained with high-dimensional feature vectors and thus outperform conventional event detectors, like the short-term average over long-term average STA-LTA trigger (**?**), which operates on signal amplitude, only.

There exists a range of other important natural phenomena~~also exciting seismic signals~~, whose seismic signatures often

remain hidden in the vast amounts of available continuous data. ~~Even though~~ Although there are ongoing efforts to detect and characterize non-earthquake seismic events (**?**), 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 (?). In this context, past studies have shown the high value of seismic measurements for natural hazards science (?)~~that are related to moving masses on the Earth's surface (**?**). This includes catastrophic slope collapses in mountain regions (**?**), iceberg calving and resulting tsunami waves (**??**) as well as smaller events like rockfalls (**?**), avalanches (**?**) and debris flows (**?**), which nevertheless pose a threat to human lives and infrastructure. Seismometers can detect these events at kilometer distances and in case of the largest events even at hundreds of kilometers. Consequently, reliable detection is of high value for natural hazard research and monitoring.

~~Conventional algorithms in earthquake seismology, such as the short-term average over long-term average STA-LTA trigger (?), are not easily transferable to the domain of environmental seismology - especially since discrimination between earthquake, noise as well as other transient waveforms and~~ Compared to earthquakes, 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 (??) and basal sliding of glaciers (?)~~detection of mass movement signals in continuous seismic data is often more intricate: for events involving a granular mass, like avalanches and debris flows, seismic signals are generated by the chaotic superposition of particle-ground impacts (**?**, and references therein). This leads to emergent signals without identifiable seismic phases at frequencies above 1 Hz sustained over typical event durations on the minute scale (**?**). Iceberg calving signals are similar, although they also involve merging fractures and interaction with the proglacial water body (**?**). Events involving millions of cubic meters also produce seismic signals below 0.1 Hz as the bulk mass hinges over a contact point with the glacier terminus in the case of iceberg calving (**?**) or accelerates along a runout trajectory in the case of slope failure (**?**). In both cases, potentially impacted water bodies may resonate over many hours, which is often referred to as the "seiche" signal (**???**).

For the emergent character of mass movement seismograms, statistical learning models have proven useful (**????**). 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 ~~?. These type~~ **???**. These types 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 )~~ like earthquakes or mass movements can be interpreted as anomalies in a background noise field. From a geophysical perspective, this background field is very complex, transient and non-stationary (**??**) - 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 (**?**) to noise source location (**?**) and aquifer monitoring (**?**). Compared to the duration of seismic ~~events~~ signals 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~~ number of signals to be studied. In this work, we consider the isolation forest (IF) algorithm, a ~~simple~~ classical yet powerful anomaly detection method. We chose this algorithm because of ~~(a) fast training and inference, (b) light-weight storage of models, and (c)~~ its favorable computational and memory complexity, strong empirical performance (**???**) ~~.~~[1] ~~Since vanilla~~ and minimal number of hyper-parameters to tune. Since unsupervised 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~~ In this work, we pursue both approaches~~in this work, with~~ , the latter guided by measuring the 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.

## 2 Methodology

### 2.1 Pre-processing

We use the Scikit-learn (version 1.4.1) (**?**) and ObsPy (version 1.4.0) (**?**) 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~~

[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.~~

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 (**??**). We follow this convention by considering sliding windows covering 100 second periods taken with 50 second overlap. Generally, we denote the time series of preprocessed counts contained in a sliding window by bold $\boldsymbol{x} \in \Re^T$ with $T = 10000$, and refer to these as time windows for brevity. On the other hand, a waveform segment is characterized by a start- and end-time, and contains the time series of all the pre-processed counts observed over this period. We can take sliding windows over a waveform segment to generate 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 **??** we refer to these randomdecision trees~~$\{\boldsymbol{x}_i : i = 1, 2, \ldots, n\}$ of time windows which can be fed to a machine learning algorithm. If the waveform segment happens to contain time gaps, the sliding windows are taken separately over the contiguous components, and so $n$ will depend on the length of the waveform segment alongside the number and size of the gaps.

In this work, we will refer to various types of waveform segments, which is described in Table 1 for reference. Additionally, to refer to the preprocessed count corresponding to time index $t$ in a time window $\boldsymbol{x}$ we use the unbolded subscript $x_t$, while for an indexed time window $\boldsymbol{x}_i$ we use $x_{i,t}$.

### 2.2 Isolation forest

The intuition underlying the IF is that when anomalous observations are dropped down randomized decision trees they tend to follow short paths to leaf nodes. By randomized decisions trees we mean that at each node a feature to split on is randomly selected, and a splitting point taken from the corresponding observed values in a subsample of the data, also at random. Such trees, which we refer to as isolation trees (iTrees) ~~.~~ following **??** , are typically trained to a specified maximum depth. The IF itself consists of an ensemble of iTrees, trained on various subsamples of the data, which is used to compute anomaly scores of observations. Within the context of this paper, an observation corresponds to an entire time window and a splitting point to a preprocessed count.

~~For a test observation $\boldsymbol{x}$ and a given iTree, let us define the path length $h(\boldsymbol{x})$ as the~~ We illustrate this intuition in Figure 1 where we show how three time windows traverse through an iTree taken from the case study of Section 3.1. Two of these time windows were taken over a debris flow period and these require 1 and 3 splits respectively to traverse to a leaf node. The third time window does not correspond to a debris flow and requires 8 splits to reach a leaf, which in this case equals the maximum depth parameter.

#### 2.2.1 Training and evaluation

To each mini-seed seismic recording contained in a specified training period, we fit an iTree to the time windows extracted from the corresponding waveform segment so that the size of the IF ensemble corresponds to the number of ~~edges from the root to the terminal nodecontaining $\boldsymbol{x}$. The more anomalous $\boldsymbol{x}$,~~ mini-seed recordings. This was done so that the ensemble is representative

| Term | Description |
|---|---|
| waveform segment | Preprocessed counts contained between a specified start- and end-time. |
| time window | 100 second waveform segment, understood to be sliding windows taken with 50 second overlap. |
| trigger segment | Waveform segment flagged by a triggering algorithm; trigger can be replaced with IF or STA-LTA if underlying algorithm needs to be specified. |
| event segment | Waveform segment corresponding to a specific event; event can be replaced with underlying cause e.g. debris flow segment, mass movement segment, earthquake segment. |
| catalog segment | Event segment contained in a catalog. |
| region of interest (ROI) | Period inside a waveform segment containing the most anomalous preprocessed counts according to the isolation forest. Capped to a maximum of 30 minutes. |
| IF control segment | Waveform segment taken from the control station over the ROI associated with the corresponding IF segment. |

**Table 1.** Summary of terminology used for the different types of waveform segments.

of the entire training period under consideration. A specific iTree is trained on a random subsample of size $\psi = 256$ taken from the time windows extracted over the corresponding miniseed recording to a maximum depth of $\log_2(\psi) = 8$, so that the overall procedure has linear computational and memory complexity. This sub-sampling size was motivated empirically by **?**, and 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.~~ choice of the maximum depth is the Scikit-learn default. We remark that a mini-seed recording typically corresponds to a calendar day and contains approximately 1728 time windows when taken with 50s overlap. In the rare case that a recording do not contain 256 time windows, we upsample randomly with replacement so that the iTrees are always trained to subsamples of the same size.

~~For improved interpretability, the final anomaly score is calculated by normalizing and transforming the quantity $\widehat{\mathbb{E}}_{\mathcal{D}}[h(\boldsymbol{x})]$~~

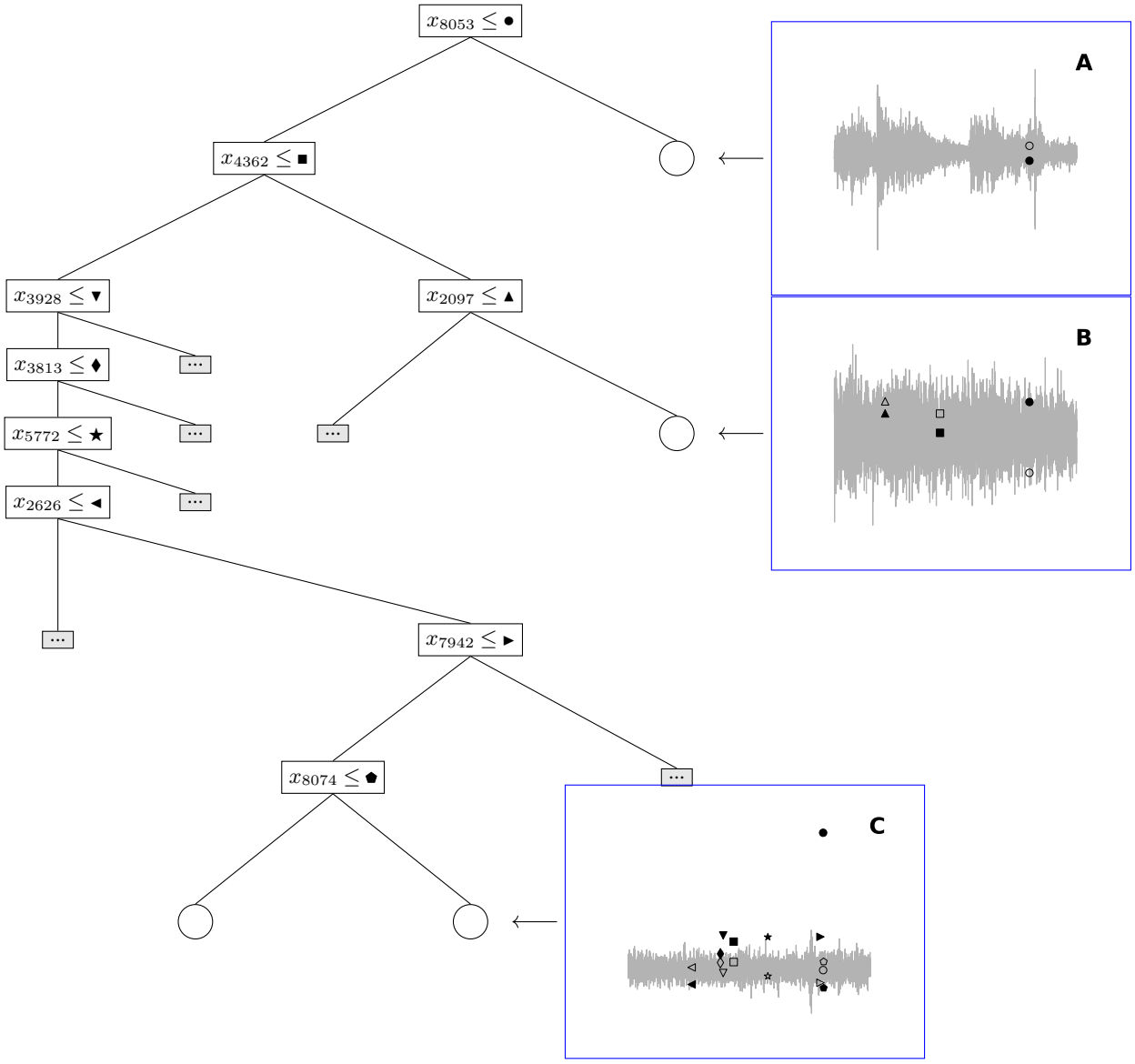


**Figure 1.** iTree trained to time windows from the waveform segment at ILL18 on 2018-05-27. The rectangular nodes show the preprocessed count represented by a filled marker, and corresponding index used as a splitting point. We show how three time windows traverse the iTree. The time windows displayed in panels A and B are taken from waveform segments corresponding to debris flows, while no debris flow signal is present in the time window of panel C. For each time window, the splitting points at the relevant time indices are shown alongside unfilled markers to indicate the corresponding preprocessed count of the time window. If the unfilled marker is above the filled one then the time window traverses to the right child of the corresponding node. We collapsed paths in the tree not relevant to the example time windows, these are represented by the small shaded rectangular nodes with ellipsis.

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 The above choices can be appreciated via an analogy between the number of steps 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 for time window $\boldsymbol{x}$ in an iTree, and the path length of 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 (??). The final isolation forest anomaly score for a test observation . For iTree $r$ in an IF we denote the former by $h_r(\boldsymbol{x})$ and refer to it as the path length for brevity. Assuming a random BST, the average path length of an unsuccessful search can be computed theoretically as $c(\psi) = 2H(\psi - 1) - \frac{2(\psi-1)}{\psi}$ with $H(j) \approx \log(j) + \gamma$ the harmonic number and $\gamma$ Euler's constant (?). Because iTrees and binary search trees (BST) have an identical typology, $c(\psi)$ serves as a reasonable reference value for the path length, although it is not necessarily true that $\mathbb{E}[h_j(\boldsymbol{x})] = c(\psi)$ for new time windows $\boldsymbol{x}$ is given by $s(\boldsymbol{x}, \mathcal{D}) = 2^{-\frac{\hat{\mathbb{E}}_{\mathcal{D}}[h(\boldsymbol{x})]}{c_{|D|}}} \in (0,1)$not used to fit the iTree. Returning to Scikit-learn's default behavior, we observe that $c(\psi) = \mathcal{O}(\log_2(\psi))$ so that by default the maximum depth grows in the order of the average path length of an unsuccessful search in a random BST.

### 2.2.2 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 (??). 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 windowswe refer to Appendix A1During evaluation, the iTrees are kept frozen and an IF anomaly score is computed for all time windows $\boldsymbol{x}$ extracted from the waveform segment covering the evaluation period. The more anomalous the time window $\boldsymbol{x}$, the higher we expect the average of the path lengths $h(\boldsymbol{x}) = \frac{1}{R}\sum_{r=1}^{R} h_r(\boldsymbol{x})$ over the iTrees to be. The *IF anomaly score* for time window $\boldsymbol{x}$ is computed in Scikit-learn as

$$s(\boldsymbol{x}) = 2^{-\frac{\tilde{h}(\boldsymbol{x})}{c(\psi)}} \tag{1}$$

where $\tilde{h}(\boldsymbol{x}) = h(\boldsymbol{x}) + \frac{1}{R}\sum_{r=1}^{R} c(n_j(\boldsymbol{x}))$ with $n_j(\boldsymbol{x})$ the number of time windows in the subsample used to construct iTree $j$ in the corresponding leaf node of $\boldsymbol{x}$. The additional term is a correction factor to account for the fact that the iTrees are not trained to their maximum granularity. We remark that $0 < s(\boldsymbol{x}) < 1$ with a higher score indicating a more anomalous time window, and when $s(\boldsymbol{x}) = 0.5$ the corrected expected path length of $\boldsymbol{x}$ is equal to the BST reference value. Given that the iTrees are kept frozen, the evaluation has linear complexity in terms of the number of time windows.

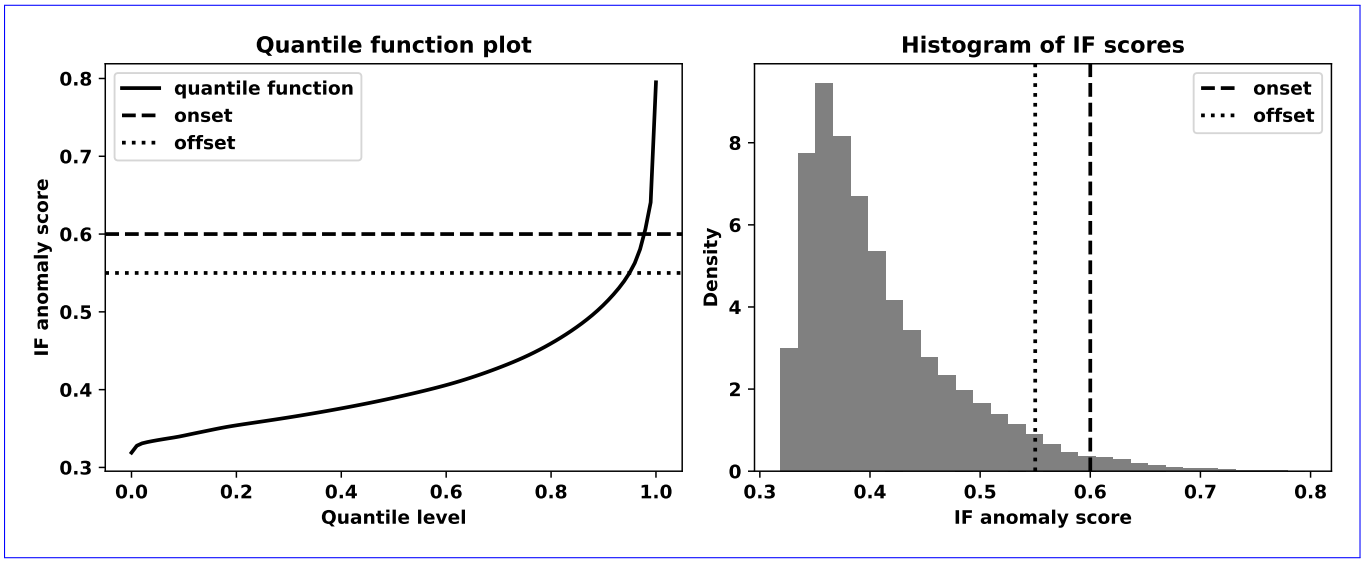


**Figure 2.** Quantile function plot and histogram of the IF anomaly scores computed for all time windows considered in the case studies.

### 2.2.2 Isolation forest trigger

~~The IF~~ Our objective is to find waveform segments in the seismic data containing counts that exhibit anomalous behavior. To this end we propose to use a trigger that operates on the IF anomaly scores of time windows, which we call the *IF trigger*. This 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~~ , and we mark the waveform segment starting from the start time of the onset window ~~, until the starting point~~ until the start time of the offset window as anomalous. We refer to waveform segments flagged by the IF trigger as *IF segments*. ~~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~~annotations if available. In the case where calibration is not possible ~~, we recommend using an onset and offset threshold of 0.60 and~~ one needs to resort to rule-of-thumb values. While such recommendations are inherently heuristic in nature, we argue that such specifications should meet the following requirements:

1. An IF anomaly score of 0.5 means that the average path length of the corresponding time window over the ensemble is equal to the BST reference value. Since approximately 11.42% of all time windows considered in the case studies had an IF anomaly score of at least 0.5, this value should serve as a lower bound for both thresholds.

2. In the Case study of Section 3.2 there are several waveform segments corresponding to mass movements containing time windows with IF anomaly scores reaching values close to 0.7 (see Table B2) suggesting that this value should serve as an upper bound both thresholds.

3. In cases where the IF anomaly score spikes briefly, having an offset threshold greater than the onset can lead to spuriously long IF segments, since we need to wait for the anomaly score to spike again above the offset threshold for the trigger to deactivate. To avoid such cases we require the onset threshold to be at least as large as the offset.

Our rule of thumb suggestion is to set the offset- and onset thresholds equal to 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.~~ and 0.60 respectively. To further contextualize these choices, we show a quantile-function and histogram plot of all the IF anomaly scores computed in the case studies in Figure 2. Around 5.05% and 2.24% of time time windows had an IF anomaly score of 0.55 and 0.60 respectively.

To quantify the degree of anomalous behavior contained in a IF segment, we use the maximum IF anomaly score associated with the corresponding time windows and call this the *IF segment score*. This score can be used to propose a ranking of the IF segments for exploration purposes. We define the *region of interest* (ROI) of a waveform segment as the 30 minute sub segment containing the most anomalous preprocessed counts; for segments shorter than 30 minutes, the ROI is the entire segment. In the case of waveform segments longer then 30 minutes, the ROI is extracted by identifying the time window with the highest IF anomaly score and iteratively expanding by adding the time window in the direction of the larger IF anomaly score, until the 30 minute cap is reached.

Figure 3 shows scatter plots of the IF anomaly scores against the log standard deviation of time windows observed at stations ILL11 and ILL14 during 2018 in the Illgraben seismic network considered in case study of Sect. 3.1. The scatter plots form a hook-like pattern with time windows related to debris flows ranked highly in terms of the IF anomaly score for fixed log standard deviation values of the seismogram amplitude. The figure suggests that the IF segment score can rank IF segments related to debris flows (or mass movements in general) highly in stations like ILL11, but this will not always be the case, as illustrated by the ILL14 scatter plot. For stations like the latter, additional tools are needed to improve the exploration procedure in both the semi-supervised and unsupervised setting. At the same time, Figure 3 shows that time segments with higher seismic amplitudes (higher standard deviation) tend to have higher anomaly scores. However, highly anomalous debris flow segments are not associated with the largest standard deviation (in particular, for ILL14). This shows that anomaly is based on waveform information beyond the seismogram amplitudes.

### 2.3 Dynamic time warping

~~Suppose that~~

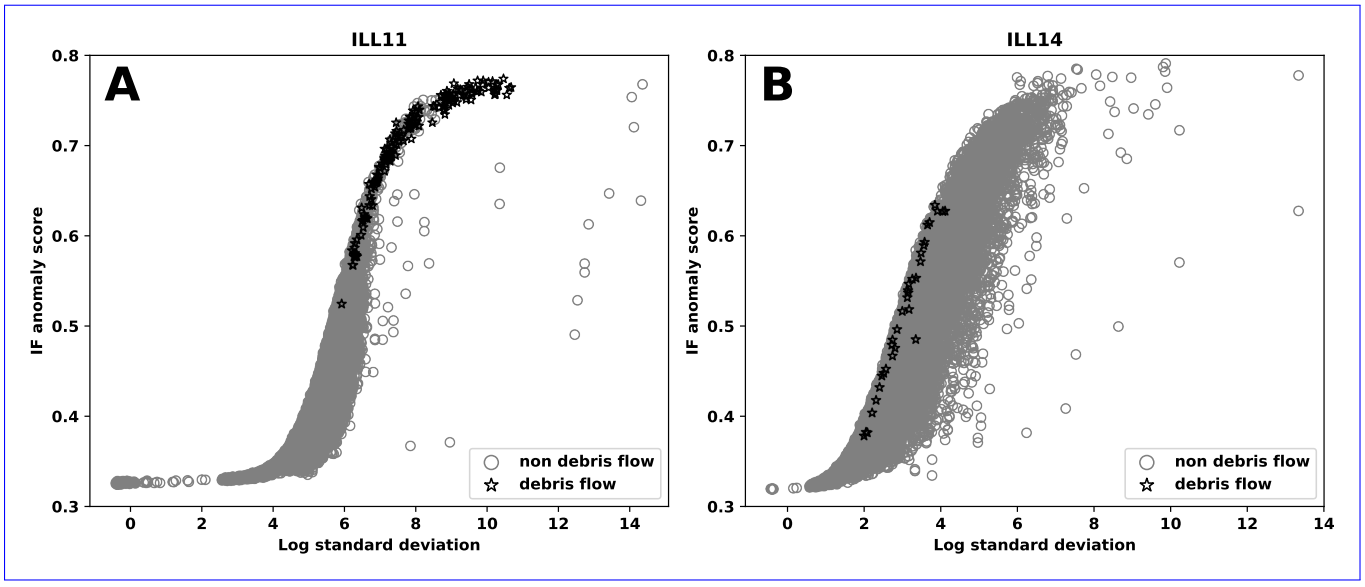


**Figure 3.** Scatter plots of IF anomaly score against the log standard deviation of time windows observed at station ILL11 (panel A) and ILL14 (panel B) during 2018. Similar plots for the remaining stations are shown in Figure A1.

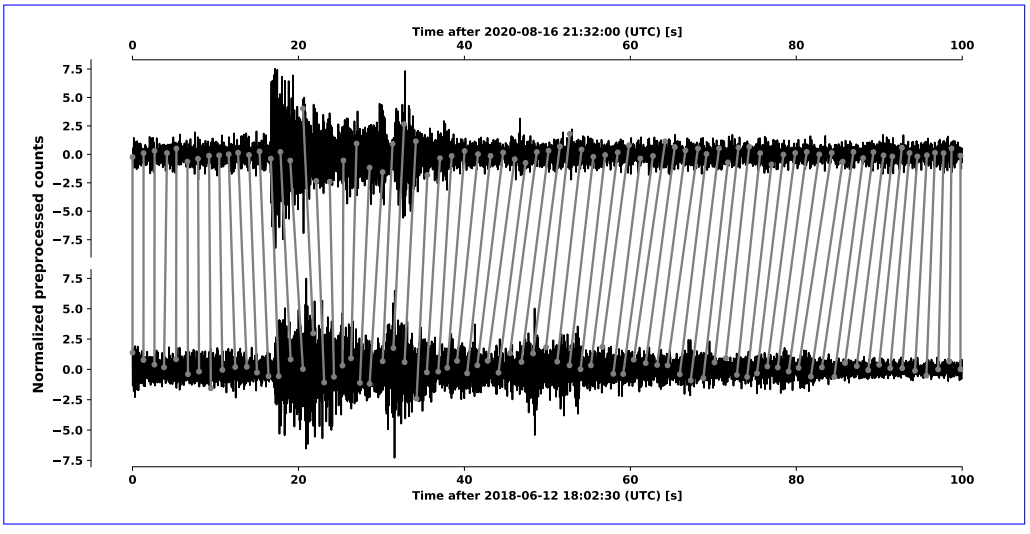


**Figure 4.** Illustration of DTW between two time windows taken from debris flow segments at station ILL18 of the Illgraben seismic network. To avoid unnecessary clutter, not all matches between the time windows are indicated.

To improve exploration of the IF segments we consider measuring dissimilarity between waveform segments using dynamic time warping (DTW). In DTW we align two sequences by matching entries between them such that the overall distance between matched entries are minimized, subject to constraints on how matches can be made. We illustrate this procedure in Figure 4 showing two time windows taken from debris flow segments from the Illgraben seismic network at station ILL18. Both time windows are normalized to zero mean and unit standard deviation with matched normalized preprocessed counts indicated by gray line segments. Note that the first and last entry of the top time window are matched with the first and last entry of the second respectively, and none of the gray line segments cross. These illustrate the constraints on the way in which the entries of the sequences are allowed to be matched.

More formally, 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~~ The second condition ensures that 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 $\boldsymbol{x}_2$ ~~.~~(this is why gray line segments in Figure 4 do not cross). We remark that a valid path can contain repeated values with the interpretation of allowing local stretching/compression to align the sequences. The DTW objective is to find the valid path that minimizes the objective

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

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 (2) 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 (**?**). ~~Since the signals we are studying are relatively long, using all~~ This does pose a computational bottleneck when comparing IF segments since such segments can span hours and therefore contain hundreds of thousands of preprocessed counts. In this work we address this bottleneck using a pre-alignment of time windows contained in segments as depicted in Figure 5. The pre-alignment is done by extracting the time windows for each segment, computing the corresponding time series of IF anomaly scores and aligning them using DTW. Two time windows are matched if their corresponding anomaly scores were matched in the pre-alignment, and we compute the DTW distance between each pair of matched time windows. These distances are then aggregated into a single value using the median. We refer to this procedure as segment DTW and the corresponding median as the segment DTW

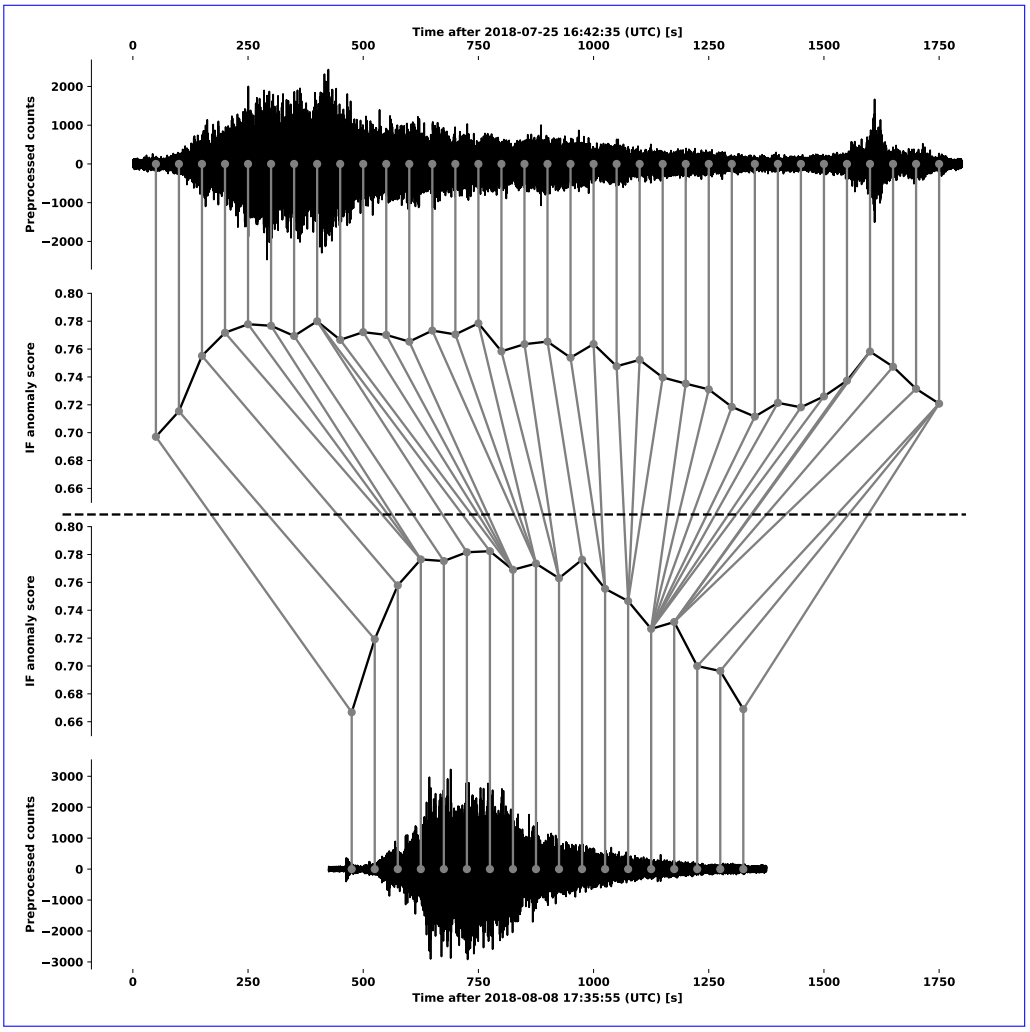


**Figure 5.** Illustration of DTW between two waveform segments corresponding to debris flows taken from station ILL18 in the Illgraben seismic network. The gray circles indicate the midpoints of time windows while the gray lines track the matching of time windows through the alignment of the IF anomaly scores.

distance.

We remark that in all cases time windows are normalized to zero mean and unit standard deviation before DTW is performed, and waveform segments are confined to their corresponding ROIs before application of segment DTW. The DTW alignment between time series of IF anomaly scores is done exactly, while the alignment between time windows is done approximately using the method of **?** with a radius of 1.

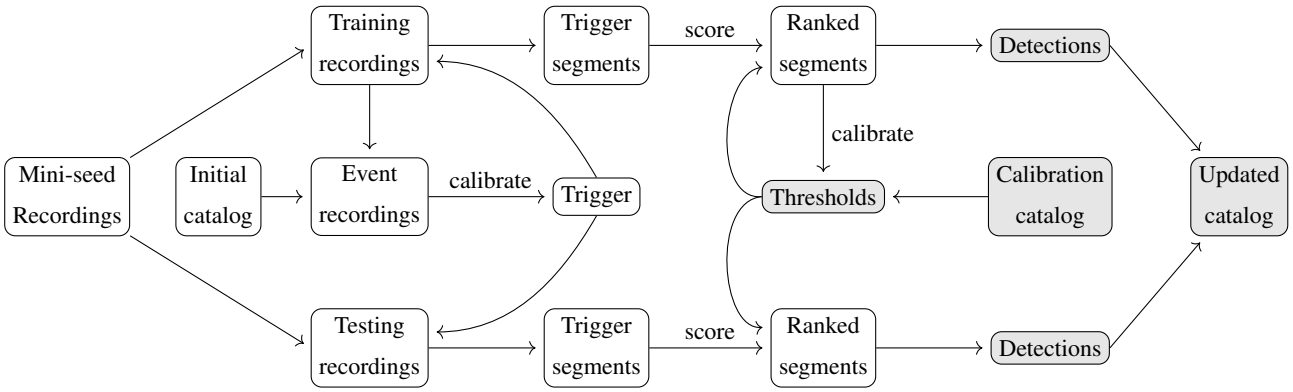


**Figure 6.** Exploration workflow in the semi-supervised case. The components that change following an update of the event catalog are indicated by shaded nodes.

### 2.4 Semi-supervised workflow

In the semi-supervised setting we assume access to an initial catalog of event segments that can be used to guide the exploration procedure in order to obtain a more complete catalog (Figure 6). We first split all available mini-seed recordings into a training- and testing period, the former used exclusively for calibrating the procedure. From the training mini-seed recordings we extract only those containing at least one initial catalog segment, and these recordings are used to calibrate a specified triggering algorithm. The calibrated trigger is applied to both the training- and testing recordings to flag segments in the data. These trigger segments are scored and subsequently ranked using a specified scoring procedure for 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 : **Template DTW**. Take the single sliding window over a segment with~~ training- and testing periods separately. The calibrated trigger and scored segments remain frozen in subsequent updates of the catalog. A score threshold and minimum segment length are calibrated by comparing the trigger segments with segments contained in a specified calibration catalog over the training period. Trigger segments meeting the score threshold- and minimum segment length requirements become detections, separately for the training and testing period. The detections are then subjected to expert labeling and used to produce an updated catalog. The calibration catalog is set to be the initial catalog during the first round of updates, and replaced with the updated catalog during subsequent ones. We also allow any of the initial catalog, calibration catalog and trigger segments to be pruned before any comparison to allow for the removal of waveform segments connected to an event with a specified degree of uncertainty.

For calibration purposes we compare a list of waveform segments to segments contained in a specified catalog using the intersection over union (IoU) metric. The IoU metric is defined to be the total time where the waveform segments in the list and catalog segments overlap expressed relative to the total time where either is present. In addition, we define a segment in the catalog to be a true positive if we can find at least one waveform segment in the list that overlaps with it, otherwise it is labeled

a false negative. We define a true positive this way to avoid multiple counts of an event in the case where multiple waveform segments in the list overlap with it. If a waveform segment in the list does not overlap with any catalog segment, it is labeled a false positive. Using these definitions, we define

$$\text{recall} = \frac{TP}{TP+FP}$$

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

The recall therefore measures the proportion of catalog segments contained in a list of waveform segments, while precision measures the proportion of waveform segments in the list that intersect with catalog segments.

We consider three different versions of the semi-supervised workflow which we call STA-LTA, IF and IF-DTW. For the first we use the classical STA-LTA trigger where waveform segments are scored with the maximum value of the characteristic function over 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. **Segment DTW**. Take all sliding windows over a segment and compute the corresponding IF anomaly scores. To compare two segments , we first compute the~~ corresponding period while for the second we use the IF trigger and score segments with the IF segment score. In the case of the IF-DTW we again use the IF trigger but with an alternative scoring method using segment DTW. For this scoring method, we perform segment DTW between all pairs of initial catalog segments contained in the training period and use the pairwise segment DTW distances to construct a dendrogram under complete linkage. We then remove those initial catalog segments that do not form a sub cluster with other catalog segments before merging with the dendrogram (singleton merges), since such segments are considered unusual w.r.t. the majority of catalog segments according to the segment DTW distance. This procedure is illustrated in Figure A2. An IF segment is then scored with the average segment 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~~ the segment and the remaining initial catalog segments. If the IF segment happens to overlap with one of the catalog segments, the corresponding 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 ? with the Euclidean distance metric, and sliding windows are normalized to zero mean and unit standard deviation before performing DTW.~~ excluded when computing the score.

### 2.5 Unsupervised workflow

In the unsupervised case we run the IF trigger using the rule of thumb thresholds and construct a clustering guided by the segment DTW distance. We found that performing pairwise segment DTW between a large number of IF segments to form a dendrogram can be computationally intractable due to the quadratic number of comparisons. In the case where the number of

IF segments exceeds 200, we instead opted for an approach following **?**. The idea is to compute the segment DTW distances between an IF segment and a set of reference segments and use the corresponding distances as features to characterize the IF segments. Beyond the computational advantages, this approach yields a proper metric in the space of segment DTW distances and also allows additional features to be incorporated.

To find suitable reference segments, we first perform segment DTW between the leading 200 IF segments according to the IF segment score, construct a dendrogram using complete linkage, and cluster the segments using the dynamic tree cut (**?**) algorithm with a deep split of 3 and minimum cluster size of 1. Inside each cluster we select the leading IF segment to use as a reference. We also included the IF segment score, length of the ROI and a feature describing the anomalous behavior of waveform segments at a control station over the ROI associated with the IF segments. The control station should be sufficiently far from the target station so that local events (in particular, mass movements) at the latter do not effect the former at the same time, and sufficiently close so that regional/global events (in particular, earthquakes) 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 an earthquake rather than a mass movement. We define the *IF control segment* as the waveform segment taken from the control station over the ROI associated with the corresponding IF segment. Then we fit a new IF to the control station, and take the corresponding maximum IF anomaly score of the IF control segment as the *control IF segment score*.

To cluster the IF segments we performed min-max scaling of all the features followed by hierarchical clustering under Ward linkage. The dynamic tree cut algorithm with a deep split of 3 and minimum cluster size of 10 was used to determine the clusters.

## 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. 7.

### 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 $\mathrm{m}^3$ reach the Rhône River (**??**). Illgraben's debris flows move at several meters per second and feature the typical boulder-rich

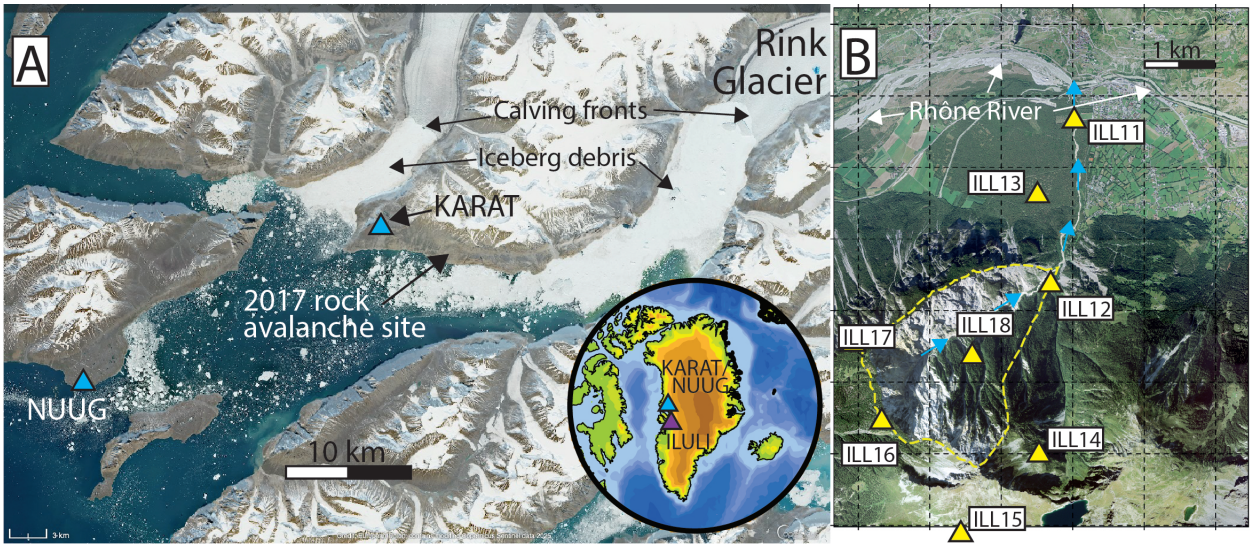


**Figure 7.** Study sites in Greenland (A) and Switzerland (B). (A) Karrat Fjord with seismic broadband stations (blue triangles), the location of the 2017 rock avalanche and major calving fronts indicated. White ice debris cover on the tidewater results from disintegrating icebergs. Inset shows the 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~~ using the Generic Mapping Toolbox and modified from **?** (A), Swisstopo (B).

fronts, which are efficient seismic sources that can be detected on local seismic networks (**?**). 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. 7). In addition, WSL's ~~Illgraben~~ debris flow observatory at Illgraben 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"; **?**). We consider data from the Illgraben seismic network for the period covering 2018-04-10 to 2022-08-28. Summary statistics are given in Table A1 of Appendix A3.

### 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~~ This is illustrated in Fig. 8, where we show waveform segments and the corresponding IF anomaly scores on 2021-07-06 when a debris flow ~~to affect the~~ was active in the torrent. We see that the debris flow first effects the upper stations ILL14-ILL18 ~~first,~~ and subsequently ILL12, ILL13 and ILL11 in order. ~~The existing debris flow catalog~~

An existing catalog of debris flow segments, each coupled with a specific station, was independently curated by cross-

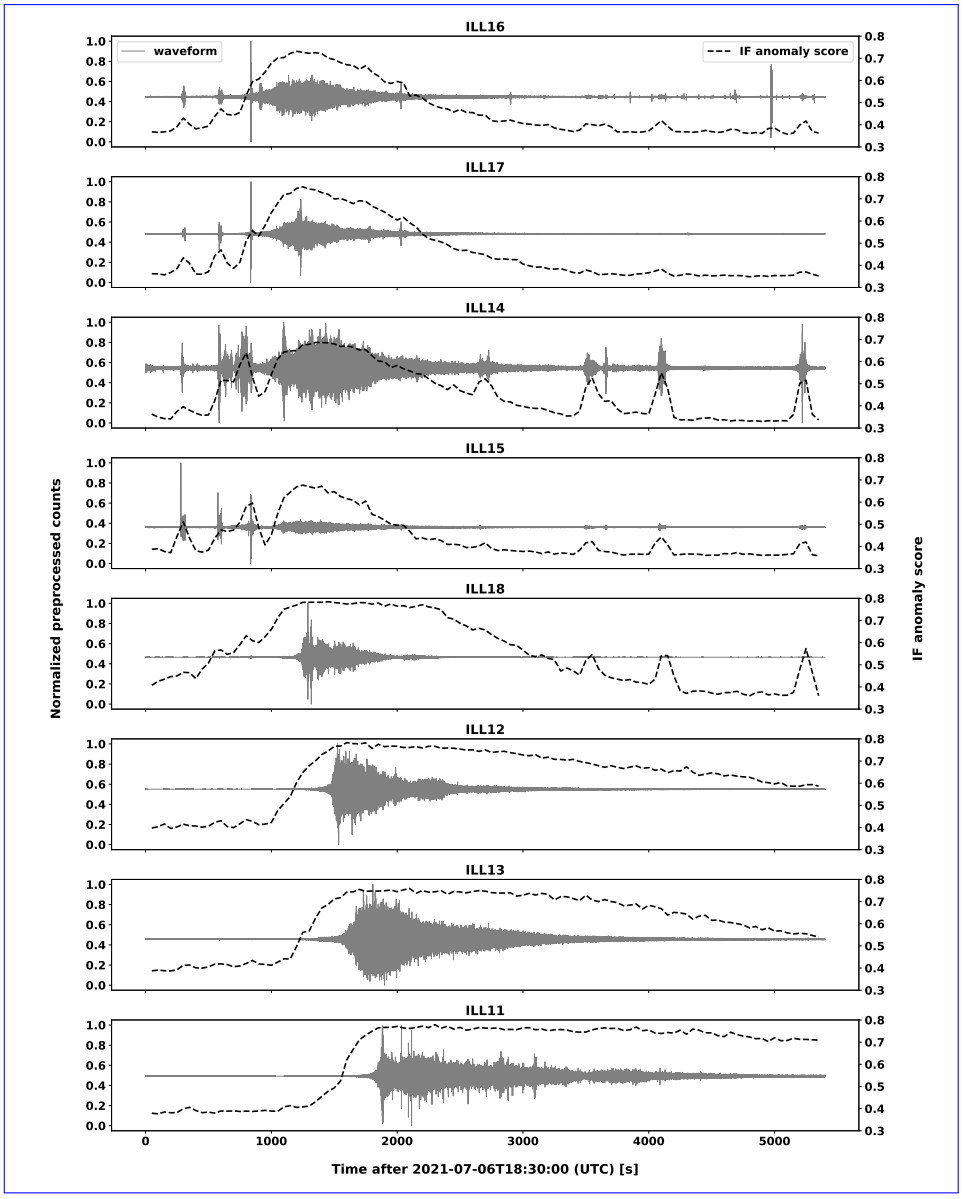


**Figure 8.** Time series plots of waveform segments and IF anomaly scores for stations in the Illgraben network from 2021-07-06 18:30:00 (UTC) to 2021-07-06 20:00:00 (UTC). Preprocessed counts were normalized to the range $[0, 1]$ for each station. The IF anomaly tends to increase significantly before the visible onset of the debris flow particularly at stations ILL11 - ILL13 and ILL18. While this has not been systematically evaluated, we do not consider this early increase in the IF score to be attributable to the acausal filtering described in Sect. **??** since such preprocessing tends to suppress the amplitude of waveforms over debris flow events.

referencing detections made by WSL's Illgraben debris flow observatory with the seismic waveforms of the stations in the network, keeping ~~this~~ the above definition in mind. ~~Each~~ Since a debris flow signature does not always manifest as clearly as in Fig. 8, each debris flow segment in the catalog ~~consists of a start- and end-time, coupled with a station and confidence level. The confidence levels are~~ is associated with a confidence level, which is defined as follows:

1. High confidence. The segment is observed during a debris-flow signature and contains a clear signal.
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 "~~The existing catalog will be referred to as the WSL catalog with corresponding summary statistics given in Table A2 of Appendix A3. In the remainder of this section we refer to lower- and medium confidence segments in a catalog as lower-confidence segments. A trigger segment overlapping with a lower confidence segment in a catalog, but with no high-confidence segment, is called a lower-confidence ~~segments" to refer to both low- and medium confidence segments jointly.~~ trigger segment.

### 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~~

### 3.1.3 Calibration

To each station in the Illgraben seismic network we apply the 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.

### 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 catalogsegments (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 segmentin 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 segmentin strategy of Fig. 6 using data from 2018-2020 for training. The calibration of 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 trainingperiod 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 for a station is done by using all corresponding segments in the WSL catalog

as an initial catalog. The 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.~~ then extracted, scored and kept frozen. 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 of the IoU metric, 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 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~~ can be found. The selected hyper parameters for both triggers are reported in Table A3.

~~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~~ For calibration of the thresholds we used the catalog produced in the preprint version of this paper (**?**). This catalog was produced following three major updates of the WSL catalog using the workflow of Fig. 6, but where a simpler form of DTW was used to score the segments. We provide more detail in Sect. A2, but note that after the second update in the formulation of this catalog the lower confidence class was expanded to include other types of mass movements such as rockfalls, landslides and slope failures. For a chosen station we prune away 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~~ catalog and trigger segments according to the preprint catalog before computing the IoU metric. 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~~in the catalog of the corresponding station. We report the recall of low- and medium-confidence segmentsseparately.~~ . The selected score thresholds and minimum detection lengths for all the methods are reported in Table A4.

#### 3.1.4 ~~Debris flow detection: results~~Evaluation

~~Tables A5 and A6 show the metrics for each mining method across all stations over the training and test periods respectively. In the recall and precision columns,~~ Following the calibration procedure the detections made were used to update the preprint catalog. To keep the number of detections to investigate manageable we excluded detections from stations ILL14, ILL15, ILL16 and ILL17 from the IF and STA-LTA methods. These stations are located above Illgraben's upper catchment near substantial noise sources associated with a water reservoir, a skiing resort, cow grazing and human activity around various buildings. For IF-DTW, all detections were used to update the catalog. As in 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 **??**. These metricsare computed following three updates of the original catalog made~~ calibration of the thresholds, for a given station, we prune away

Metric

ILL1

Metrics over the training period after updating the catalog. The numbers in brackets in the recall and precision columns represent the number of false ne

IL

ILL1813.0852.8259.3733.33 (10)100.00

**Table 2.** Metrics Average metrics over different station groups during the testing test period after updating the catalogfrom 2021-2022. The numbers in brackets in the recall and precision columns represent the number of false negatives and false positives, respectively. All percentages metrics are displayed accurately up as percentages rounded to two decimalsdecimal places. The symbol "-" means average metric of the best performing method is indicated in bold which in all cases is IF-DTW. We remark that when STA-LTA fails to make a detection over the corresponding metric could not test period at a specified station, the precision cannot be computed because no detections were made over and in such cases we simply allocated a zero value to the testing periodmetric.

lower-confidence catalog and detections, this time according to the evaluation catalog. The corresponding filtered detections and catalog segments over a given time period are compared and used to calculate IoU, recall and precision metrics. We provide detailed tables of these values over the training period.[1] 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 and test periods in Table A5 and Table A6, respectively, where the test period was taken to be 2021-2022. The respective number of low, medium and high confidence segments increased from 15,24,240 in the WSL catalog to 197,44,257 in the final updated catalog.

Table 2 provides average IoU, recall and precision metrics over the test period for three different station groups. The first station group containing stations ILL11, ILL12, ILL13 and ILL18 for the STA-LTA and IF methods , while for IF-DTW we also include represents stations where the performance of all the methods is better relative to the second group, which contains stations ILL14, ILL15, ILL16 and ILL17. After an update, the score threshold and minimum detection length of the 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

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

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 periodThe third group corresponds to all the stations. With detection IoU, Recall and Precision of up to 68 %, 100 % and 98 %, respectively, we see that IF-DTW outperforms IF. IF, in turn, outperforms the STA-LTA method in terms of the averages for all metrics and in all groups. While the performance of all methods are worse for the ILL14, ILL15, ILL16 and ILL17 station groups, the degradation is more severe for the STA-LTA and IF methods.

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 ??A3) to manage sensitivity towards amplitude, in order to avoid flagging an excessive number of false positive segments (see Fig. A1). However, these long window sizes leadsto 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 A1. The results further show that detecting debris flows from stations ILL11, ILL12, ILL13 and inviting false negatives. We illustrate this in Fig. 9 where two debris flows segments associated with 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 on 2021-07-31 are shown. Due to the long window sizes, the characteristic function of the flows are suppressed by preceding increased amplitude in the seismic waveforms. We include more examples in Fig. A2 and Fig. A3.

The sensitivity of the STA-LTA trigger to amplitude and its proneness to event masking means that it is difficult to find a configuration of hyper parameters where both the number of false positives and false negatives are small. On the other hand, the IF anomaly score has a natural robustness to amplitude due to the threshold splits along the time axis used to construct iTrees (see Fig. 1). This is why the IF provides superior metrics in terms of the IoU, recall and precision with an increase of 19.82%, 45.50% and 17.71% on average for all stations over those produced by STA-LTA. We remark that since the IF anomaly score is computed from the path lengths in iTrees, which are built in a randomized manner, there is nothing explicitly guiding the score 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. different types of events. This is in contrast to IF-DTW, where the score reflects the dissimilarity between the IF segment and debris flow segments according to the segment DTW distance. At stations ILL11 - ILL13 and ILL18 the average improvement for the IoU, recall and precision metrics offered by IF-DTW over IF is 10.17%, 11.65% and 1.15% respectively, and this increases to 33.59%, 23.24% and 58.46% for stations ILL14 - ILL17. The reason for the difference in the scale of the improvement is because the IF anomaly score happens to rank debris-flow time windows highly at stations ILL11-ILL13 and ILL18 (see Fig. 3 and Fig. A1).

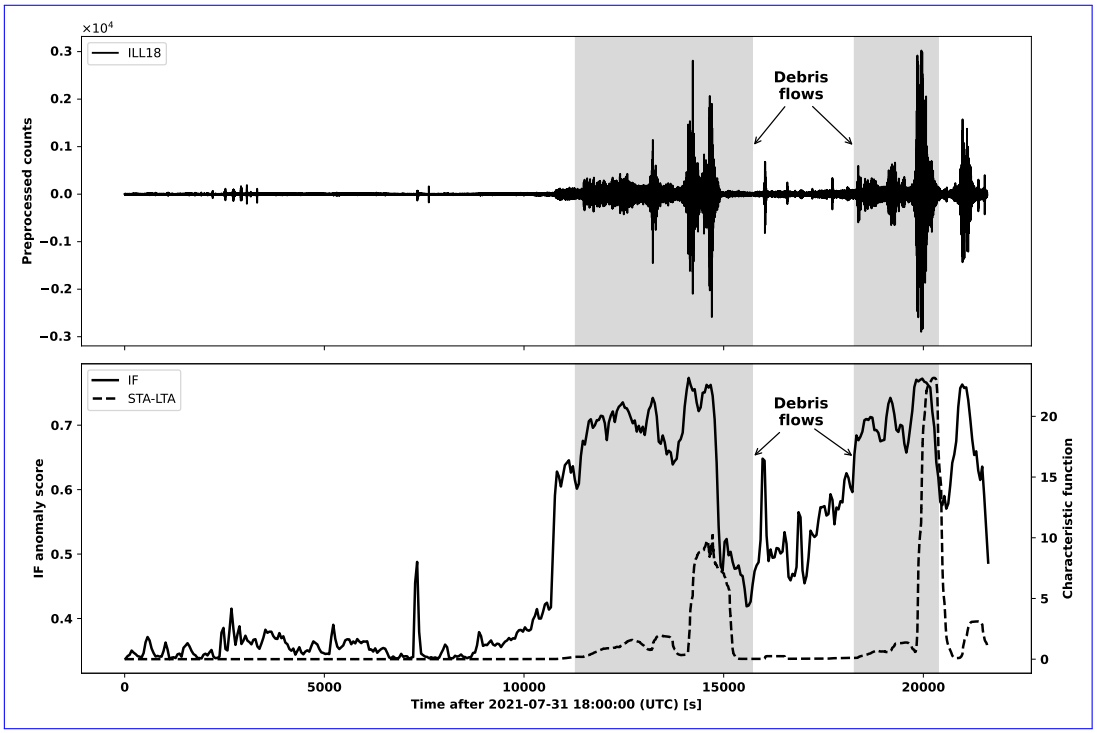


**Figure 9.** Illustration of the behavior of the STA-LTA characteristic function relative to the IF anomaly score at ILL18 on 2021-07-31. Debris flows are represented by the shaded regions.

## 3.2 Greenland

### 3.2.1 Study site

Our Greenlandic site locates on the western coast at the Karrat Fjord (Fig. 7). 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 (**?**). The rock avalanche and precursory slip events left clear seismic signatures on the nearby broadband station NUUG, installed in the village Nuugaatsiaq (**??**). 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 rock 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~~

While our primary focus is the generation of a catalog of events for the KARAT station, we illustrate the unsupervised exploration procedure of Sect. 2.5 by applying it to waveforms obtained from NUUG over the period 2017-01-01 to 2017-06-28, using waveforms from ILULI over the period 2017-01-01 to 2017-12-31 as the control station (see Table B1 for summary statistics). The IF trigger flagged 194 segments in 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: 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. 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. 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~~ corresponding waveforms. The highest ranking IF segment according to the IF segment score corresponds to the rock avalanche discussed in the preceding paragraph, with a corresponding value of 0.7373 and control IF segment score of 0.7171 (both accurate to four decimals). Time series of the preprocessed counts and IF anomaly scores contained in the rock avalanche segment are shown in Fig. 10. The same figure contains a boxplot of the heights at which individual IF segments merge with the remainder inside a dendrogram constructed from the pairwise segment DTW distances between the 194 IF segments under complete linkage. ~~We note that the height at which segments merge into a cluster quantifies the diversity of the segments with larger heightcorresponding 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. 10. 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~~The larger the merge height, the more dissimilar the corresponding IF segment is with respect to the remaining segments. The rock avalanche segment achieved the largest height with a value of 8618.74. We emphasize that its enormous volume made this event a rare example of a mass movement that is strong enough to show up as a highly anomalous seismic signal at both NUUG and the control station ILULI.

### 3.2.2 KARAT ~~results~~clustering

We ~~analyze data~~ apply the exploration procedure of Sect. 2.5 to waveforms 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 **??** and the results are illustrated in Fig. **??** 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~~

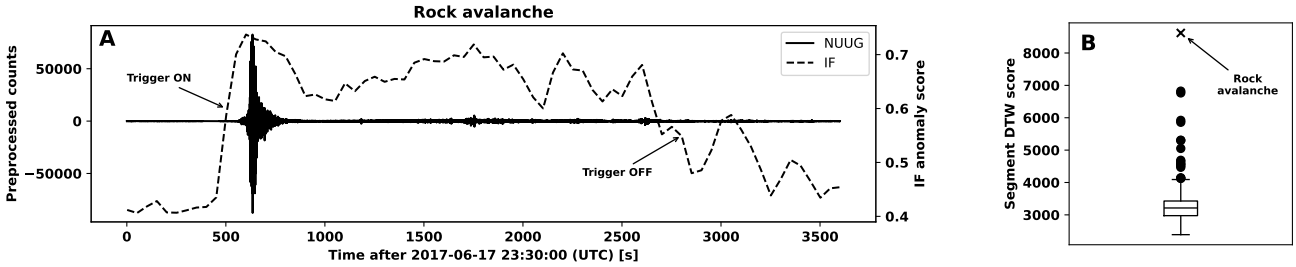


**Figure 10.** ~~Seismic~~ Panel A: Rock avalanche waveform segment observed at the NUUG station overlaid with the IF anomaly scores~~close to the 2017 Rock Avalanche (panel A). The segment represented by the onset- and offset triggers indicated on the plot represents the most anomalous segment flagged in 2017.~~ Panel B~~shows a box-plot summarizing an agglomerative clustering~~ : Boxplot of the ~~highest 50 anomalous~~ heights at which individual IF segments ~~flagged by~~ merge with the ~~IF trigger, based on~~ remainder inside a dendrogram constructed from the pairwise segment DTW ~~scoring method discussed in Sect. 2.3,~~ distances 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.~~

~~**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~~ for the period 2022-05-30 to 2023-10-20 using waveforms from ILULI over the period 2022-01-01 to 2023-12-31 as the control station (see Table B1 for summary statistics). The IF trigger flagged a total of 605 segments in the KARAT waveforms of which we could relate 96 to mass movements, most of which are likely iceberg calving events. To determine if an IF segment is related to a mass movement the seismic waveforms and corresponding spectograms around the segment flagged by the IF trigger are investigated by a domain scientist and a mass movement label is recommended based on well-known characteristics of calving seismograms (see Fig. 14). Once such a label is recommended, additional verification is performed using satellite images if available. A clearly missing piece of a glacier terminus confirms a calving event (Fig. 15). In some cases, the capsizing of a large tabular iceberg may produce a similar seismic signature (Fig. 17). An equivalent procedure is used to confirm whether an IF segment is related to a teleseismic or regional earthquake, with additional verification from the United States Geological Survey and the Geological Survey of Denmark and Greenland (**??**) earthquake catalogs. Fig. B2 shows for each $k \in 1, 2, ..., 605$ the proportion of the $k$ leading IF segments that are related to mass movements. The graph spikes fairly quickly with a maximum value of 45.83% at $k = 48$ showing that the IF segment score tends to rank mass movements highly.

The unsupervised workflow split the 605 IF segments into 13 clusters as summarized in Table 3; the corresponding Ward linkage dendrogram and clustering of the IF segments are shown in Figure B1. Cluster 13 is exclusively populated by highly ranked IF segments flagged before 2022-08-15 when the instrument was streaming sporadically and with high amplitudes (see Fig. B3). Since these likely correspond to issues on the instrumentation side, these segments are labeled as instrument-relatednoise.

| | | Membership proportion | | | | |
|---|---|---|---|---|---|---|
| Cluster | Size | EQ | MM | Noise | High-pass screened | Figure |
| 1 | 107 | 0.00 | 0.00 | 0.00 | 100.00 | - |
| 2 | 67 | 0.00 | 0.00 | 0.00 | 100.00 | - |
| 3 | 66 | 1.52 | 0.00 | 1.52 | 96.97 | - |
| 4 | 62 | 0.00 | 0.00 | 1.61 | 98.39 | - |
| 5 | 58 | 0.00 | 0.00 | 0.00 | 100.00 | - |
| 6 | 52 | 86.54 | 5.77 | 7.69 | 0.00 | - |
| 7 | 45 | 0.00 | 0.00 | 0.00 | 100.00 | - |
| **8** | 44 | **0.00** | **86.36** | **13.64** | **0.00** | Fig. 11 |
| **9** | 27 | **0.00** | **100.00** | **0.00** | **0.00** | Fig. 12 |
| **10** | 24 | **8.33** | **29.17** | **8.33** | **54.17** | Fig. 13 |
| 11 | 20 | 0.00 | 5.00 | 95.00 | 0.00 | - |
| **12** | 20 | **0.00** | **100.00** | **0.00** | **0.00** | Fig. 14 & 16 |
| 13 | 13 | 0.00 | 0.00 | 100.00 | 0.00 | - |

**Table 3.** ~~Dendrogram~~ Summary of clusters from exploration procedure of Sect. 2.5 applied to the ~~top 50 anomalies detected at~~ waveforms from the KARAT station. ~~We use IN, HEL, AN, CAL, REQ, TEQ and ID~~ Shown are the membership percentages for ~~instrument noise~~the 4 categories. The tags EQ, ~~helicopter, anthropogenic noise, calving events, regional~~ MM stands for earthquakes ~~, teleseismic earthquakes~~ and ~~iceberg disintegration~~ mass movements respectively. The last column contains references to representative figures containing waveform-spectrogram plots for representative IF segments from chosen clusters.

The ~~other two segments in this cluster are caused by helicopters arriving /~~noise class was expanded to include IF segments corresponding to anthropogenic events such as installation and service work near the station, helicopters arriving and departing from the station ~~. Our analysis show that 21 out of the top 50 IF segments corresponds to calving eventsshowcasing 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 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 (?). Finally, energy partitioning between different > 1~~alongside electronic glitches/spikes. The majority of the IF segments in cluster 11 correspond to these events, with one IF segment possibly related to a mass movement, but with a degree of uncertainty. Around 29.90% of the IF segments in cluster 1 were extracted from 2022-09-25 and 2022-09-26, two days with 198 and 192 gaps in the corresponding recordings, so that spectrograms could not be computed.

These segments contained unusually enhanced low-frequency ($< 0.5$ ~~Hz frequency bands may change in response to altered calving front geometries (?). 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 ($\approx 0.70$) indicative of a regional earthquake. This works almost perfectly, but for one major calving event that reached the ILULI station. Fig. B3 suggests both higher amplitudes and IF anomaly scores~~ Hz) content. This was confirmed by extracting the raw waveform over the IF segment and preprocessing it by increasing the corner frequency of the zero-phase high-pass filter to 0.5 Hz and reapplying the IF trigger with no retraining. If none of the remaining IF segments in the cluster correspond to a mass movement the cluster is inspected further, otherwise the entirety of the original IF segments in the cluster is inspected. In the remainder of the paper we refer to this process as a high-pass screening. Similarly, 95.16% of the IF segments in cluster 4 were extracted from 2022-09-25. Applying the high-pass screening process reduced the cluster to a single IF segment corresponding to noise. Similarly, we found that clusters 2, 3, 5, 7, 10 contain IF segments flagged due to increased low-frequency content, particularly during the months of September-January ~~.~~ (see Fig. B3 ). 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.~~ process the remainder of the clusters, we first apply the high-pass screening procedure to decide if a cluster should be inspected.

The results of the high-pass screening procedure left clusters 6, 8, 9, 10 and 12 to be inspected. Clusters 9 and 12 consisted of IF segments, which all resembled mass movements signatures. We give representative waveform-spectrogram plots of these clusters in Fig. 12, 14 and 16. Mass-movement related IF segments make up 86.36% of cluster 8 which also contains a few noise related events; a representative example is given in Fig. 11. Cluster 10 consists of a number of exceedingly long IF segments with an average of 10.77 hours. The the high-pass screening procedure left 14 IF segments, 10 of which are related to mass movements (some of the original IF segments are split into multiples by the high-pass screening). Included in the remaining IF segments is the Dixon fjord rock avalanche and tsunami (?) which is illustrated in Fig. 13. One of the remaining segments is related to a regional earthquake, another to a teleseismic earthquake and the remaining two to noise. The majority of the IF segments in cluster 6 correspond to earthquakes (86.54%) alongside a few mass-movement- and noise related segments. One of these mass-movement events corresponds to a major calving event that occurred on 2023-07-27 and reached the ILULI station.

## 4 Summary, Conclusion and Outlook

We have showcased the ability of the IF trigger to flag mass movements in seismic waveforms to the degree that the method ~~should~~ can be considered as an alternative to conventional algorithms when mining seismic data for such events. ~~In particular, we applied IF and STA-LTA triggers~~ Applied to continuous seismic records from a debris flow catchment, ~~which~~ our IF and STA-LTA triggers had been subjected to minimal pre-proccessing, and showed that the IF trigger can improve over the classical

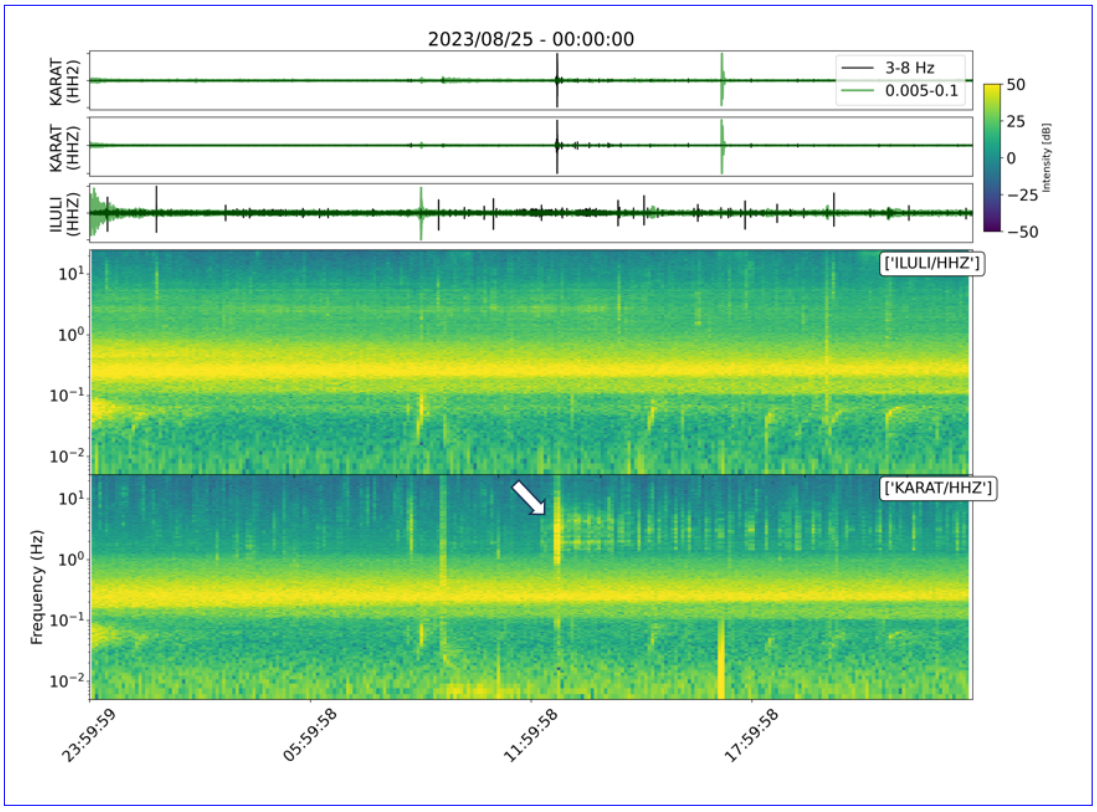


**Figure 11.** Seismic waveforms and corresponding spectrogram around the segment in cluster 8 flagged by the IF trigger on 2023-08-25 . The IF trigger flags a typical calving seismogram.

STA-LTA trigger up to ~~4~~ 2.75 times in terms of the average IoU metric. The performance of ~~the~~ both the IF and 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 (**??**). It was the goal of this study to mine for mass movements without such prior knowledge, and our results show that in this ~~regard~~ sense the IF trigger is better suited than the STA-LTA trigger.

The potential of using DTW to measure dissimilarity between waveform segments for the purpose of mass-movement identification was illustrated in both a semi-supervised and unsupervised setting. In the case of the former, an improvement of 8.16 times in terms of the average IoU metric was observed over a pure IF exploration method for debris flows in the Illgraben catchment at selected stations.

Since reasonable mass movement detectors can be obtained at some stations just by thresholding the IF anomaly score, this

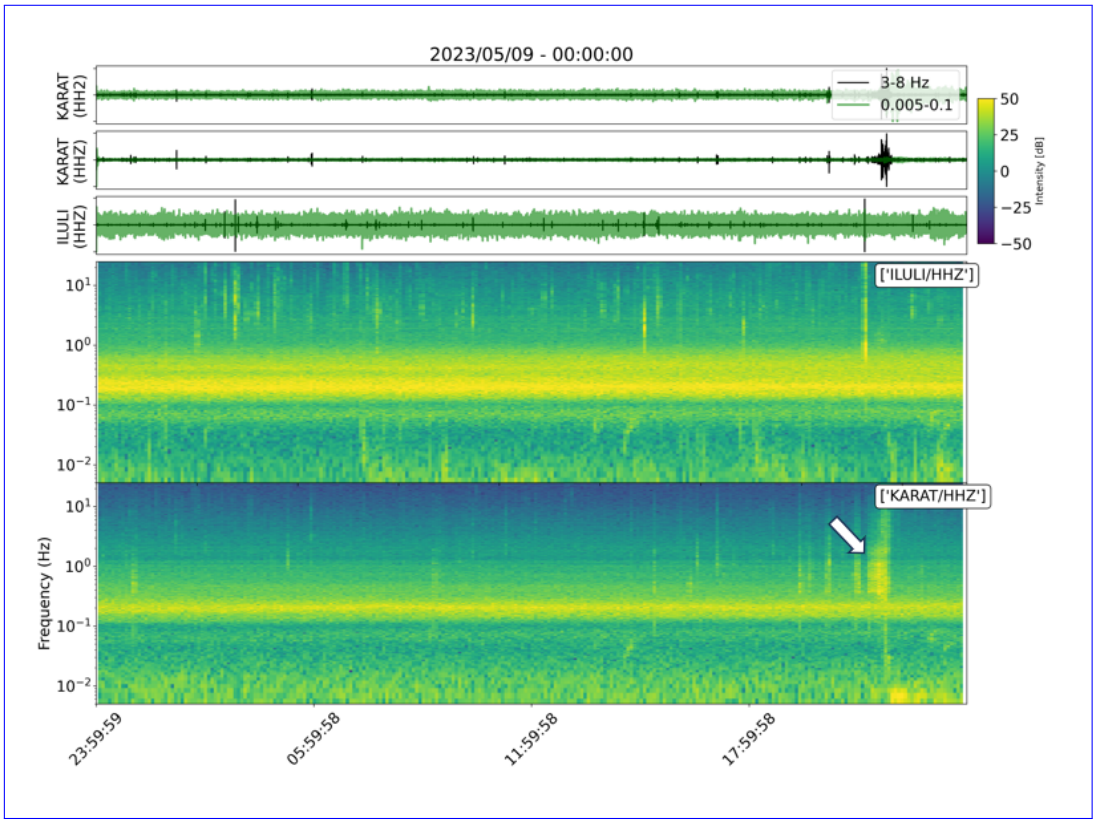


**Figure 12.** ~~Log median absolute value of the daily preprocessed~~ Seismic waveforms ~~(A)~~ and corresponding spectrogram around the segment in cluster 9 flagged by the IF ~~anomaly score (B) observed at KARAT~~trigger on 2023-05-09 . The IF trigger flags a typical calving seismogram.

score could serve as a useful feature when building more sophisticated classifiers in addition to those, for example, used in **??**. 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~~affecting 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 (???) or more broadly~~The isolation forest is a popular anomaly detection algorithm and has inspired many subsequent developments (**??**). Notable extensions include the extended isolation forest (**?**), deep isolation forest (**?**) and an IF variant that can identify anomalous subsequences in stationary time series data (**?**). Although identification of anomalous subsequences considered in the latter work aligns with our use of the isolation forest, the assumption of stationarity

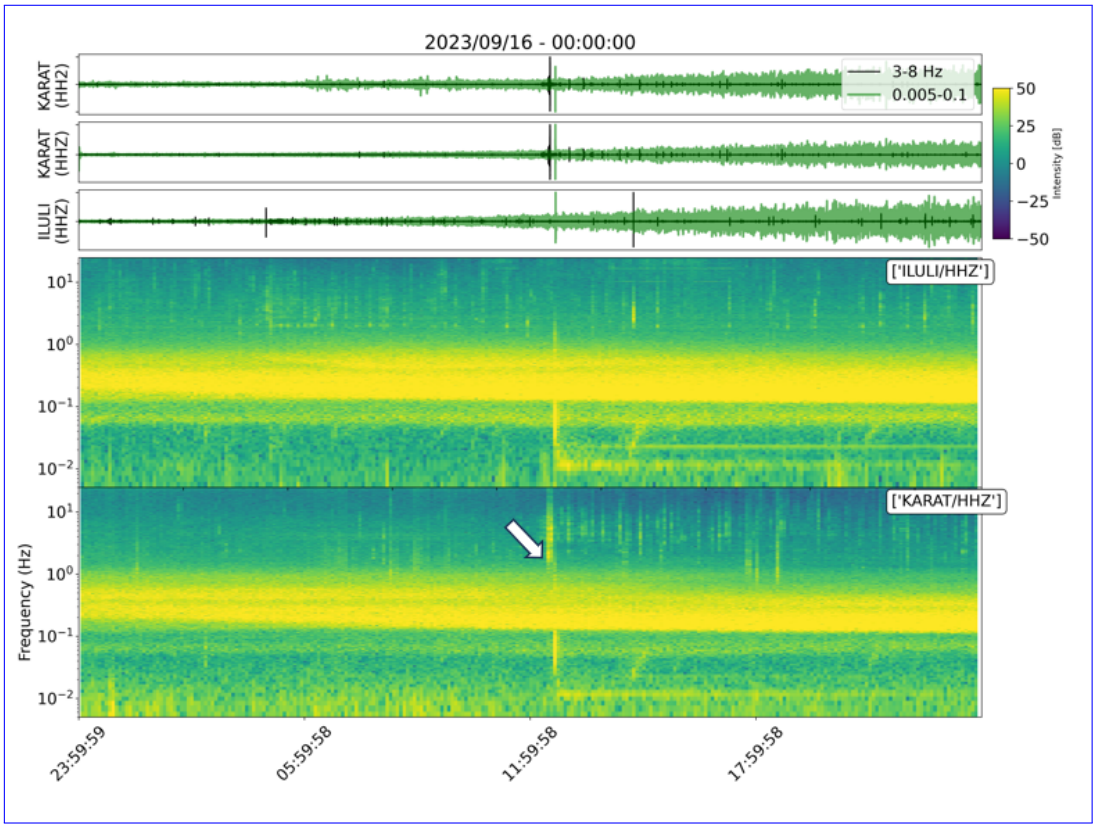


**Figure 13.** Seismic waveforms and corresponding spectrogram around the segment flagged in cluster 10 by the IF trigger on 2023-09-16 around the time when the Dixon fjord rock avalanche and tsunami occurred.

is too restrictive when considering seismic data. The extended isolation forest (EIF) has been shown to outperform the standard isolation forest in many applications (**?**) and can serve as a ready-made replacement for IF in our applications. In the deep isolation forest data are fed through various randomly initialized multi-layer perceptrons (MLPs) and subsequently processed by classical IFs. By exchanging the MLP with other neural network architectures, dynamics related to how mass movements propagate through space and time can be encoded into the anomaly score which could improve corresponding detection. More broadly, alternative anomaly detection methods in the time series ~~context (??). Another avenue for future research is to extend the IF and IF-DTW mining methods of Sect. 3.1 to be online~~ domain can be explored as reasonable alternatives to isolation-based methods (**????**). Features describing the anomalous behavior of waveform segments have been used as part of feature sets for supervised learning in seismological studies (**???**). We remark that such feature sets and others used in a similar context (**??**) can be processed through the isolation forest to produce an anomaly score. In this work, since our objective is exploration, we did not commit to a specific feature set and chose to work directly with the waveforms.

The use of DTW for waveform searching over cross correlation was suggested by **?** and the DTW distance matrix used as a basis

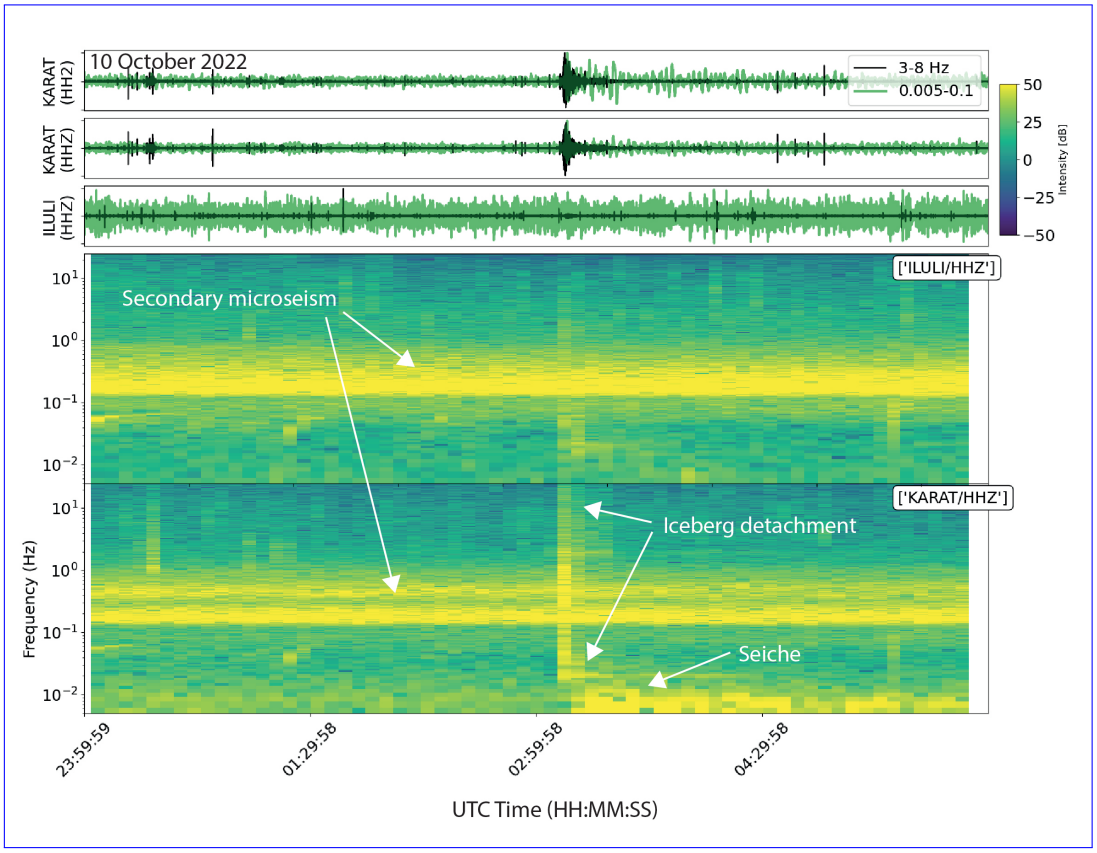


**Figure 14.** Seismic waveforms and corresponding spectrogram around the segment flagged in cluster 12 by the IF trigger on 2022-10-10 which according to satellite images constitutes a calving event (Fig. 15). One horizontal component (HH2) and the vertical component (HHZ) are shown for KARAT and the vertical component is shown for ILULI. The spectrograms show the continuous energy of the secondary microseism generated by standing waves in ocean basins (**??**). The IF trigger flags a typical calving seismogram with broadband signals representing the iceberg detachment (**?**) and a low-frequency (<0.01 Hz) signal generated by calving-induced water oscillations within the fjord ("seiche"; **?**).

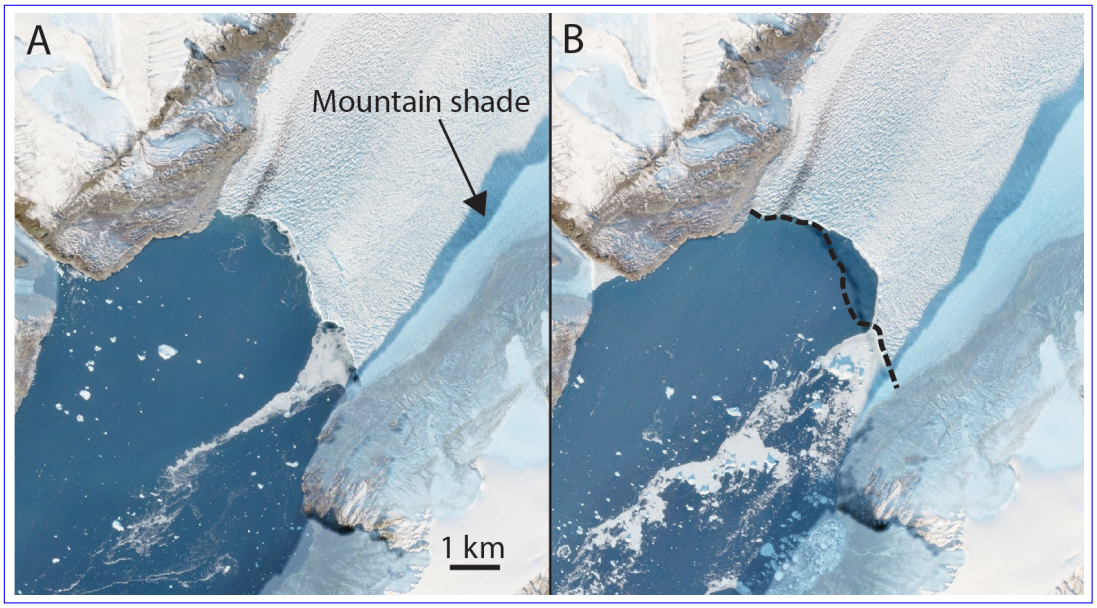


**Figure 15.** Satellite image pair of Rink Glacier calving front (Fig. 7) on 2022-10-07 (A) and 2022-10-12 (B) before and after the calving event 2022-10-10, respectively. The black dashed line represents the before-calving terminus and the missing area indicates a calving volume of about 0.5 $km^3$ assuming a terminus thickness of 500 – 600 m (**?**). Source: Sources: Copernicus (Sentinel-2 true color image).

for k-means clustering of waveform segments by **?**. In the case of the latter, 90 waveform segments were manually extracted over a 10 hour period and provided to the clustering algorithm. Since the extracted waveform segments spanned periods of 2 - 7 seconds, the computation of the pairwise DTW distances is computationally feasible. We did explore performing exact DTW between aggregated values of time windows, including time series of the IF anomaly scores and principal component (PCA) projections, but this degraded the performance of exploration procedures. Alternatively, DTW can be applied in a multivariate context to time series features of time windows as discussed in the preceding paragraph or inspiration can be draw from application of DTW in the audio domain (**?**). More ambitiously, self-supervised neural network approaches (**??**) or contrastive learning in the presence if weakly labeled data (**?**) can be used to learn features of waveform segments either to apply DTW to, or use directly in a clustering or semi-supervised procedure. Approximate differentiable DTW distance functions (**?**) can be incorporated into neural network architectures to learn features of waveform segments. Highly optimized applications of DTW (**???**) should be considered in future work, either to accelerate the use of DTW in this work or to apply it in a different manner.

While the focus of this work was exploring existing seismic waveforms, the methods considered could be extended to the online setting so that they can be used for ~~debris flow~~ mass movement detection in real time. In fact, assuming appropriate pre-processing, the IF ~~method~~ version of the workflow depicted in Fig. 6 is already online since a detection can be labeled as a ~~debris flow~~ mass movement the moment the IF anomaly score hits the score threshold, subject to the minimum detection

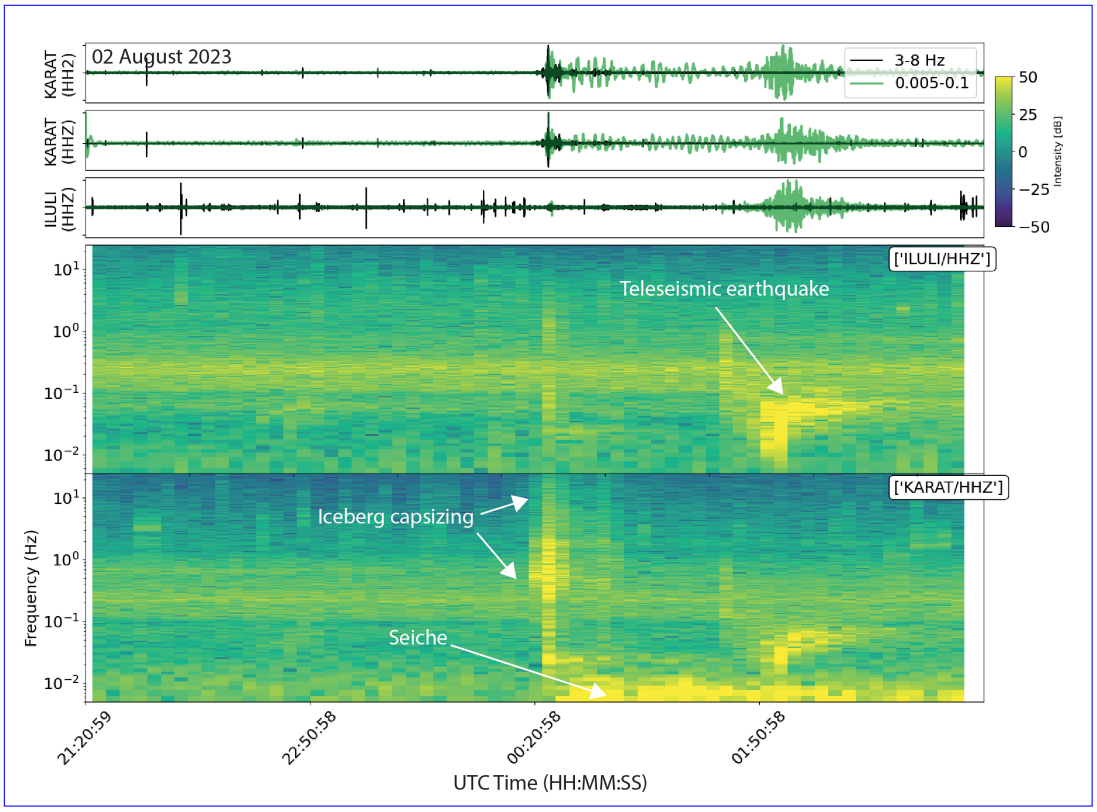


**Figure 16.** Seismic waveforms and corresponding spectrograms around the segment flagged in cluster 12 by the IF trigger on 2023-08-03, which according to satellite images constitutes the capsizing of a tabular iceberg (Fig. 17). The tabular iceberg was within 500 m of the calving front and thus likely contacted the calving front as it capsized. This generated a broadband signal similar to iceberg detachment (Fig. 14). 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) (**?**)).

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 (?). 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 (??) which has been applied within the context of seismology (?). Contrastive learning and DTW hybrids are also a possibility.~~ Generalization of the IF to larger seismometer networks should be relatively easy given the computational and memory efficiency of the method. The DTW methods can be extended to the online setting by streaming

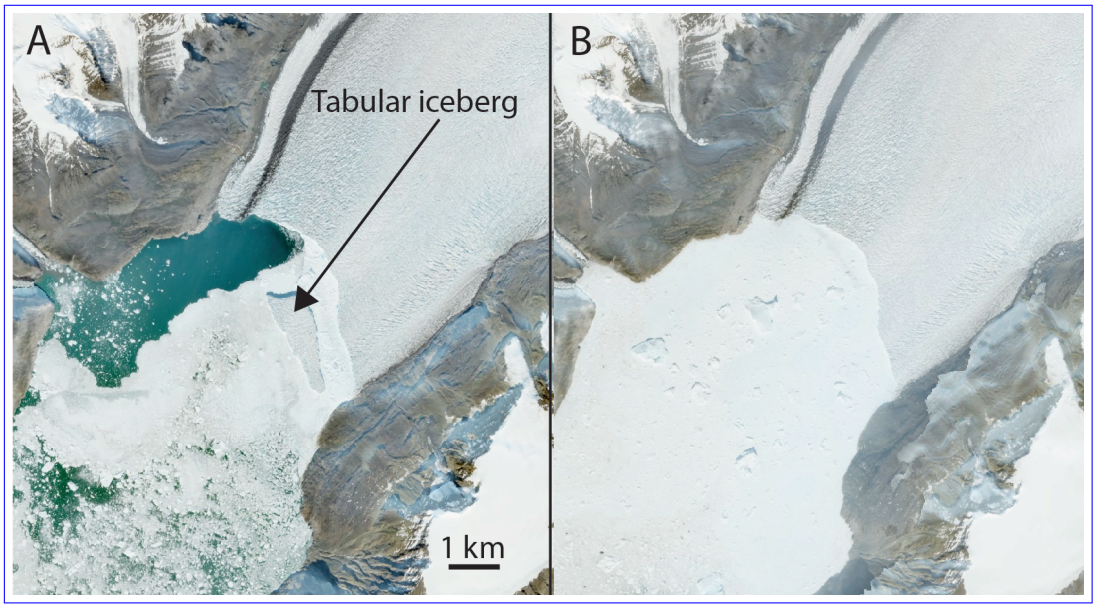


**Figure 17.** Satellite image pair of Rink Glacier calving front (Fig. 7) on 2023-08-01 (A) and 2023-08-03 (B) before and after the IF trigger segment on 2023-08-02, respectively. Assuming a full-thickness iceberg with a depth of $500 - 600$ m (**?**), the iceberg had a volume of about $0.5\,\mathrm{km}^3$ and may have contacted the terminus during capsize. Source: Copernicus (Sentinel-2 true color image).

the corresponding distances as soon as the IF trigger activates, although overcoming the computational challenges in such a step is critical.

. The seismic data used for the Illgraben case study are available at **?**. The source code is available at https://github.com/FKamper/seismic-isolation-forest.

. 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.

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

. 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.

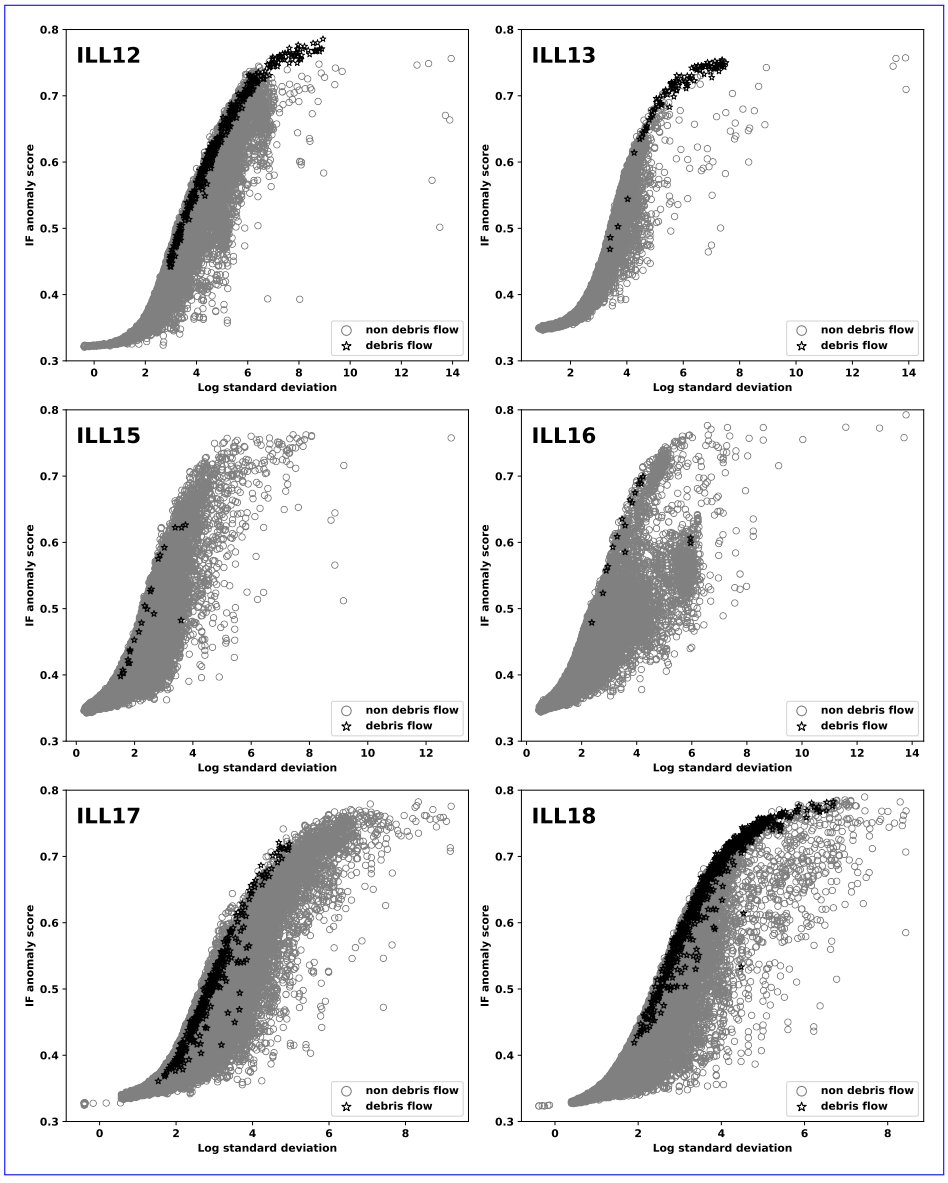


**Figure A1.** Scatter plots of IF anomaly score against the log standard deviation of time windows observed at stations ILL12, ILL13, ILL15, ILL16, ILL17 and ILL18 during 2018.

## Appendix A: ~~Isolation forest anomaly score~~ Illgraben supplementary information

### A1 Isolation forest anomaly scores

~~Figure ?? and ??~~

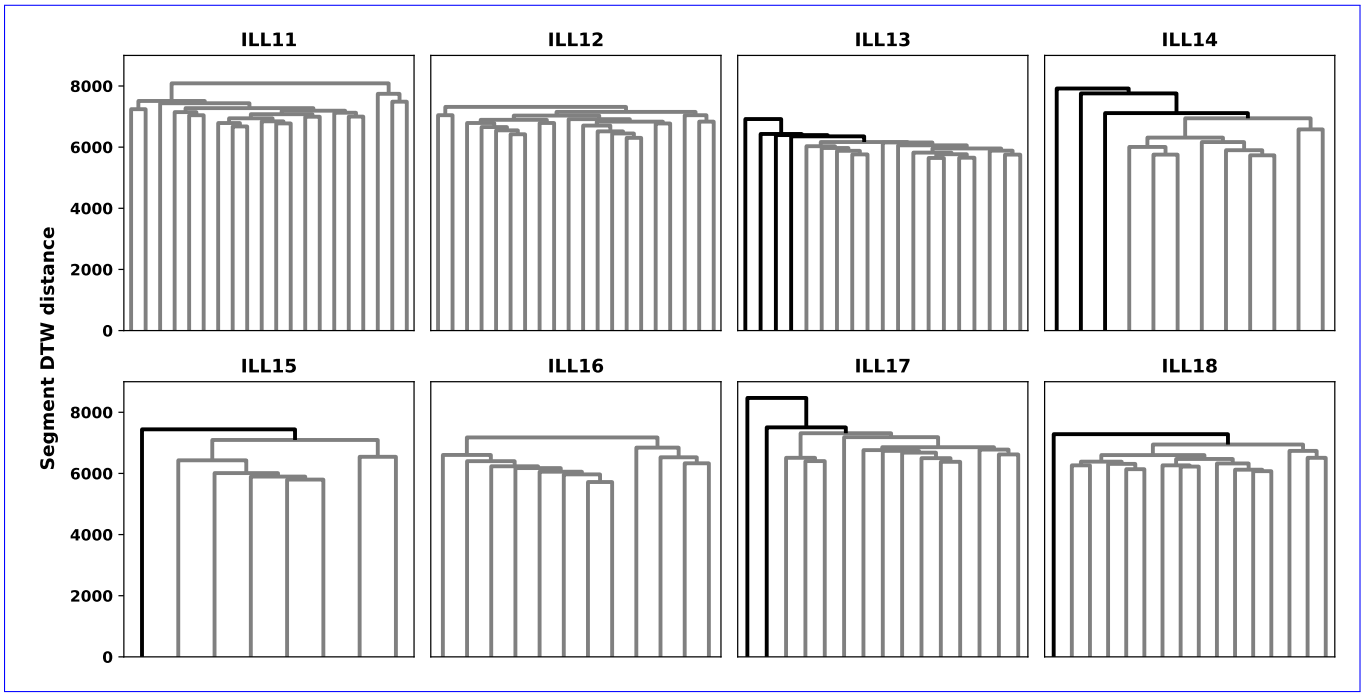


**Figure A2.** ~~Plots of~~ Dendrograms constructed for high-confidence training segments in the ~~IF anomaly score vs the log standard deviation of sliding windows taken from 2018~~ WSL catalog for ~~ILL11-ILL14~~each station in the Illgraben seismic network.~~The plots are restricted to show only those sliding windows with an anomaly score exceeding 0.5.~~

Figure A1 shows the IF anomaly score plotted against log standard deviation for ~~each station~~ stations in the Illgraben seismic network ~~for~~ (excluding ILL11 and ILL14) in 2018. ~~Sliding~~ Time windows overlapping with ~~catalog segments (all confidence levels)~~ debris flow segments with high confidence are indicated by star marks ~~. The relationship forms a non-linear wave-like pattern , with~~ and other confidence levels were excluded. The hook-like pattern observed in Figure 3 persists. Debris-flow time windows are highly ranked by the IF anomaly ~~scores of debris flow segments at stations ILL11,~~ score at stations 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~~.

## A2 Dendrograms

Figure A2 shows complete linkage dendrograms for each station corresponding to the segment DTW distances between high-confidence segments in the WSL catalog over the training period. Singleton merges are defined as those segments that do not merge with a sub cluster. These segments are removed before the mean DTW segment distances of IF trigger segments are computed.

~~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.~~

| Station | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|
| ILL11 | 130<br>1,114,064,574 | 155<br>1,330,498,078 | 140<br>1,202,056,672 | 148<br>1,267,142,728 | 95<br>771,963,166 |
| ILL12 | 123<br>1,050,286,704 | 172<br>1,466,552,357 | 144<br>1,221,974,157 | 72<br>598,765,256 | 111<br>953,626,064 |
| ILL13 | 122<br>1,045,012,833 | 163<br>1,382,823,864 | 160<br>2,399,407,988 | 149<br>1,270,807,077 | 111<br>952,407,056 |
| ILL14 | 124<br>1,047,005,509 | 163<br>1,480,274,085 | 143<br>1,269,611,815 | 173<br>1,478,024,717 | 112<br>946,655,226 |
| ILL15 | 86<br>733,166,663 | 106<br>887,423,288 | 161<br>2,462,657,197 | 108<br>982,738,337 | 75<br>631,977,746 |
| ILL16 | 86<br>733,993,359 | 158<br>1,347,432,481 | 164<br>2,593,025,871 | 105<br>944,259,026 | 74<br>632,515,769 |
| ILL17 | 121<br>1,036,724,359 | 162<br>1,367,039,472 | 159<br>2,351,660,008 | 141<br>1,203,039,698 | 112<br>950,456,007 |
| ILL18 | 181<br>1,531,176,693 | 197<br>1,687,134,959 | 239<br>2,060,845,390 | 203<br>1,732,218,104 | 168<br>1,253,195,329 |

**Table A1.** Number of mini-seed recordings (top of each cell) and counts (bottom of each cell) for each station by year in the Illgraben seismic network, before any pre-processing is performed.

### A3 Summary statistics

We give the following summary statistics for the Illgraben seismic network:

- Table A1 contains statistics related to sample sizes.
- Table A2 contains counts of the number of event segments of different confidence levels for each station and year pair, in the WSL and updated catalogs.

## Appendix B: ~~Mining methods hyper-parameters~~

~~Table ?? contain the hyper parameters selected by the calibration procedure for the IF and~~

### A1 Hyper parameters

Table A3 contains the chosen hyper parameters for the IF and STA-LTA trigger while Table A4 contain the calibrated thresholds for IF, IF-DTW ~~mining methods~~ and STA-LTA methods.

| Station | 2018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|
| ILL11 | 0 / 0 / 4 | 0 / 0 / 9 | 0 / 0 / 7 | 1 / 2 / 11 | 0 / 0 / 2 |
| | 1 / 1 / 3 | 6 / 1 / 11 | 2 / 0 / 9 | 2 / 2 / 14 | 2 / 0 / 2 |
| ILL12 | 0 / 0 / 4 | 0 / 0 / 9 | 0 / 0 / 7 | 0 / 1 / 6 | 0 / 0 / 4 |
| | 0 / 0 / 5 | 1 / 2 / 12 | 2 / 1 / 10 | 1 / 1 / 10 | 2 / 1 / 4 |
| ILL13 | 0 / 0 / 3 | 0 / 0 / 9 | 0 / 0 / 7 | 1 / 2 / 11 | 0 / 0 / 4 |
| | 1 / 1 / 2 | 1 / 1 / 12 | 0 / 0 / 10 | 1 / 2 / 14 | 0 / 0 / 4 |
| ILL14 | 0 / 1 / 1 | 1 / 2 / 6 | 2 / 0 / 5 | 0 / 0 / 7 | 0 / 0 / 4 |
| | 0 / 1 / 1 | 1 / 3 / 7 | 5 / 1 / 4 | 36 / 0 / 8 | 29 / 0 / 4 |
| ILL15 | 0 / 1 / 1 | 0 / 2 / 2 | 2 / 0 / 5 | 0 / 0 / 4 | 0 / 0 / 3 |
| | 0 / 1 / 1 | 1 / 2 / 2 | 3 / 0 / 4 | 6 / 0 / 6 | 0 / 0 / 3 |
| ILL16 | 0 / 1 / 1 | 1 / 1 / 6 | 2 / 0 / 5 | 0 / 2 / 5 | 0 / 0 / 3 |
| | 0 / 1 / 1 | 6 / 2 / 7 | 5 / 0 / 6 | 3 / 3 / 7 | 0 / 0 / 3 |
| ILL17 | 0 / 0 / 4 | 2 / 1 / 6 | 1 / 1 / 5 | 0 / 1 / 8 | 0 / 0 / 4 |
| | 2 / 0 / 4 | 4 / 4 / 9 | 9 / 2 / 6 | 15 / 2 / 9 | 17 / 0 / 4 |
| ILL18 | 0 / 0 / 3 | 0 / 2 / 7 | 1 / 2 / 6 | 0 / 1 / 9 | 0 / 0 / 2 |
| | 2 / 0 / 6 | 1 / 3 / 9 | 6 / 3 / 9 | 11 / 3 / 13 | 13 / 0 / 2 |

**Table A2.** Counts of the number of event segments in the initial (above in each cell) and updated (below in each cell) subdivided by station and year. An entry of $a/b/c$ refers to the number of counts of events of low-, medium- and high confidence.

| Station | IF | | STA-LTA | | | |
|---|---|---|---|---|---|---|
| | Onset Threshold | Offset Threshold | Onset Threshold | Offset Threshold | Short-term Window | Long-term Window |
| ILL11 | 0.65 | 0.65 | 6.0 | 0.1250 | 500 | 5000 |
| ILL12 | 0.70 | 0.65 | 12.0 | 0.0625 | 500 | 10000 |
| ILL13 | 0.65 | 0.65 | 12.0 | 0.0625 | 250 | 5000 |
| ILL14 | 0.55 | 0.50 | 3.0 | 0.5000 | 2000 | 10000 |
| ILL15 | 0.55 | 0.50 | 3.0 | 2.0000 | 1000 | 20000 |
| ILL16 | 0.60 | 0.50 | 12.0 | 0.5000 | 250 | 20000 |
| ILL17 | 0.55 | 0.50 | 3.0 | 0.5000 | 2000 | 40000 |
| ILL18 | 0.65 | 0.65 | 24.0 | 0.5000 | 500 | 40000 |

**Table A3.** Hyper parameters selected for the IF- and classical STA-LTA trigger. Window sizes are given in seconds.

## A2 Preprint catalog

| Station | IF | | IF-DTW | | STA-LTA | |
|---|---|---|---|---|---|---|
| | Score threshold | Minimum detection length | Score threshold | Minimum detection length | Score threshold | Minimum detection length |
| ILL11 | 0.6895 | 600.0 | 8481.4217 | 600.0 | 8.0929 | 1942.78 |
| ILL12 | 0.7614 | 1650.0 | 7148.4049 | 1650.0 | 19.9125 | 1992.51 |
| ILL13 | 0.7466 | 400.0 | 6940.8372 | 700.0 | 17.2562 | 1325.11 |
| ILL14 | 0.6070 | 850.0 | 6716.4948 | 250.0 | 3.8547 | 16457.40 |
| ILL15 | 0.6875 | 1450.0 | 6607.6618 | 300.0 | 6.8055 | 7149.11 |
| ILL16 | 0.6782 | 700.0 | 6650.1578 | 650.0 | 32.4546 | 4609.80 |
| ILL17 | 0.6088 | 1650.0 | 7168.1900 | 450.0 | 13.6962 | 21343.48 |
| ILL18 | 0.7472 | 800.0 | 7233.1762 | 1000.0 | 41.1311 | 2366.17 |

**Table A4.** Score thresholds and minimum detection lengths calibrated for the IF, IF-DTW and STA-LTA semi-supervised strategies. The score thresholds are given accurate to 4 decimals and minimum detection lengths are given in seconds.

The preprint catalog was generated by three major updates of the WSL catalog using the methodology of Fig. 6, and a few smaller updates due to, for example, small experiments. As in the update described in Section 3.1.4 the detections made by the IF and STA-LTA methods at stations ILL14, ILL15, ~~while Table ?? contains those selected for the~~ ILL16 and ILL17 were nor considered in these updates. Detections from at station ILL15 from IF-DTW was excluded as well because we could not obtain meaningful results in the preprint. 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, one more update of the catalog was performed.

## A3 Metrics

Tables A5 and A6 contain comprehensive metrics for each station in the Illgraben network over the training and testing periods respectively. Table A7 and A1 show the recall achieved by the IF, IF-DTW and STA-LTA method of the lower-confidence catalog 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.

| Metric | IoU (%) | | | Recall (%) | | | Precision (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Station | STA-LTA | IF | IF-DTW | STA-LTA | IF | IF-DTW | STA-LTA | IF | IF-DTW |
| ILL11 | 53.06 | 60.95 | 60.99 | 86.96 (3) | 100.0 (0) | 100.0 (0) | 83.33 (4) | 95.83 (1) | 95.83 (1) |
| ILL12 | 19.5 | 53.78 | 62.65 | 51.85 (13) | 81.48 (5) | 88.89 (3) | 66.67 (7) | 91.67 (2) | 96.0 (1) |
| ILL13 | 25.03 | 64.24 | 75.04 | 50.0 (12) | 75.0 (6) | 95.83 (1) | 75.0 (4) | 100.0 (0) | 100.0 (0) |
| ILL14 | 7.13 | 1.54 | 32.81 | 8.33 (11) | 75.0 (3) | 83.33 (2) | 100.0 (0) | 3.3 (264) | 90.91 (1) |
| ILL15 | 2.18 | 0.72 | 6.52 | 14.29 (6) | 14.29 (6) | 42.86 (4) | 6.25 (15) | 3.85 (25) | 27.27 (8) |
| ILL16 | 2.20 | 6.04 | 49.33 | 14.29 (12) | 57.14 (6) | 100.00 (0) | 8.33 (22) | 20.00 (32) | 100.00 (0) |
| ILL17 | 10.65 | 7.80 | 49.45 | 5.26 (18) | 68.42 (6) | 94.74 (1) | 100.00 (0) | 10.74 (108) | 100.00 (0) |
| ILL18 | 27.17 | 54.15 | 62.77 | 62.50 (9) | 95.83 (1) | 100.00 (0) | 50.00 (15) | 71.88 (9) | 96.00 (1) |

**Table A5.** 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.

| Metric | IoU (%) | | | Recall (%) | | | Precision (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Station | STA-LTA | IF | IF-DTW | STA-LTA | IF | IF-DTW | STA-LTA | IF | IF-DTW |
| ILL11 | 41.61 | 71.70 | 71.70 | 87.50 (2) | 100.00 (0) | 100.00 (0) | 87.50 (2) | 100.00 (0) | 100.00 (0) |
| ILL12 | 16.18 | 53.29 | 71.40 | 46.15 (7) | 76.92 (3) | 100.00 (0) | 100.00 (0) | 100.00 (0) | 92.86 (1) |
| ILL13 | 19.76 | 50.18 | 66.65 | 64.71 (6) | 76.47 (4) | 100.00 (0) | 73.33 (4) | 100.00 (0) | 100.00 (0) |
| ILL14 | 0.00 | 3.85 | 31.87 | 0.00 (11) | 90.91 (1) | 72.73 (3) | 0.00 (2) | 5.78 (163) | 53.33 (7) |
| ILL15 | 0.00 | 0.97 | 28.67 | 0.00 (8) | 12.5 (7) | 62.5 (3) | 0.00 (6) | 4.00 (24) | 71.43 (2) |
| ILL16 | 0.00 | 4.08 | 52.35 | 0.00 (9) | 55.56 (4) | 100.00 (0) | 0.00 (21) | 23.81 (16) | 100.00 (0) |
| ILL17 | 0.00 | 9.88 | 40.21 | 0.00 (12) | 83.33 (2) | 100.0 (0) | - (0) | 11.9 (74) | 54.55 (10) |
| ILL18 | 13.08 | 55.28 | 61.37 | 33.33 (10) | 100.00 (0) | 100.00(0) | 31.25 (11) | 88.24 (2) | 100.00 (0) |

**Table A6.** 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.

## Appendix B: ~~STA-LTA examples~~

### A1 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 ~~??~~A3). We illustrate this in Fig. A1, 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-

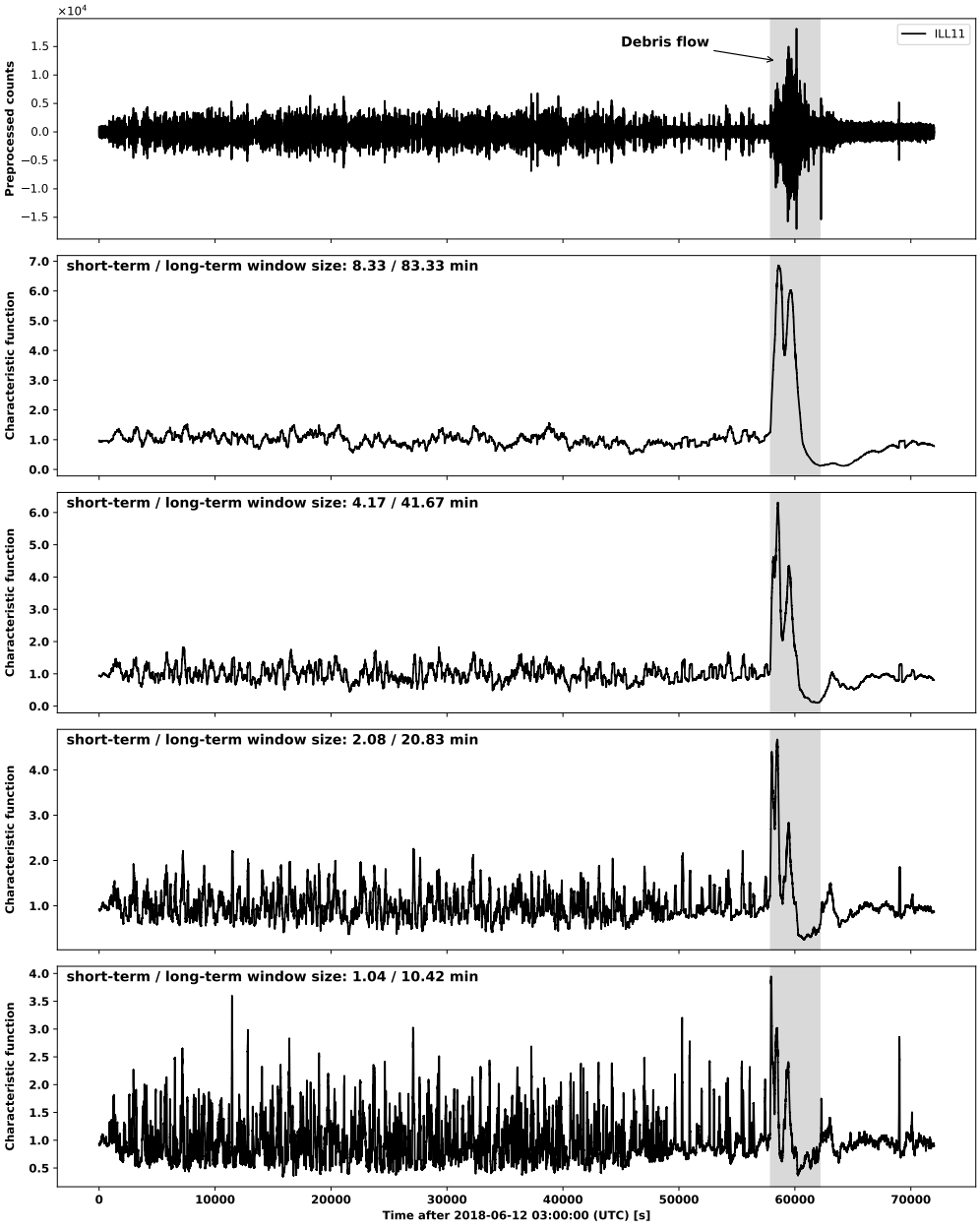


**Figure A1.** Illustration of the effect of the window sizes on the characteristic function of the STA-LTA trigger.

| | 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 | 77.78 | 88.89 | 50.00 | 50.00 | 50.00 |
| ILL12 | 3 | 3 | 0.00 | 0.00 | 33.33 | 0.00 | 0.00 | 66.67 |
| ILL13 | 2 | 2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| ILL14 | 6 | 5 | 16.67 | 16.67 | 16.67 | 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 | 27.27 | 0.00 | 33.33 | 0.00 |
| ILL17 | 15 | 6 | 6.67 | 20.00 | 40.00 | 16.67 | 16.67 | 50.00 |
| ILL18 | 9 | 6 | 0.00 | 22.22 | 44.44 | 50.00 | 50.00 | 66.67 |
| Overall | 59 | 30 | 5.69 | 18.22 | 31.33 | 14.58 | 27.92 | 29.17 |

**Table A7.** 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.

LTA trigger with the short- and long-term windows given in Table ??A3. 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. A2, A3 and 9 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.

~~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.~~

| Station | IF TriggerNumber of events | | IF Low-confidence recall (% | |
|---|---|---|---|---|
| | Low Confidence | Medium Confidence | STA-LTA | IF |
| ILL11 | 0.654 | 0.652 | 0.730.00 | 3.3375.00 |
| ILL12 | 0.703 | 0.652 | 0.760.00 | 14.1733.33 |
| ILL13 | 0.651 | 0.652 | 0.750.00 | 6.670.00 |
| ILL14 | 0.5565 | 0.500 | 0.610.00 | 14.176.15 |
| ILL16ILL15 | 0.606 | 0.500 | 0.680.00 | 11.670.00 |
| ILL18ILL16 | 0.653 | 0.653 | 0.750.00 | 13.330.00 |
| ILL11ILL17 | 6.0032 | 0.122 | 8.330.00 | 83.3325.00 |
| ILL13ILL18 | 12.0024 | 0.063 | 4.1720.83 | 83.3350.00 |
| ILL15heightOverall | 3.00138 | 2.0014 | 16.672.60 | 333.3323.69 |

ILL173.000.5033.33666.6713.70355.72 ILL1824.000.508.33666.6741.1339.44 height

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: Recall of lower-confidence segments

Table A7 and A1 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.

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.

**Table A1.** 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 B: Greenland labelingsupplementary information

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. 14). Once a label is recommended additional verification are performed if possible. For example We include the following supplementary information for the case study of Sect. 3.2:

1. **Calving events**. Satellite images such as those depicted in Fig. 15Table B1 shows the number of mini-seed recordings and counts for NUUG, KARAT and ILULI.

2. Table B2 shows the leading 10 IF segments among those that are related to mass movements.

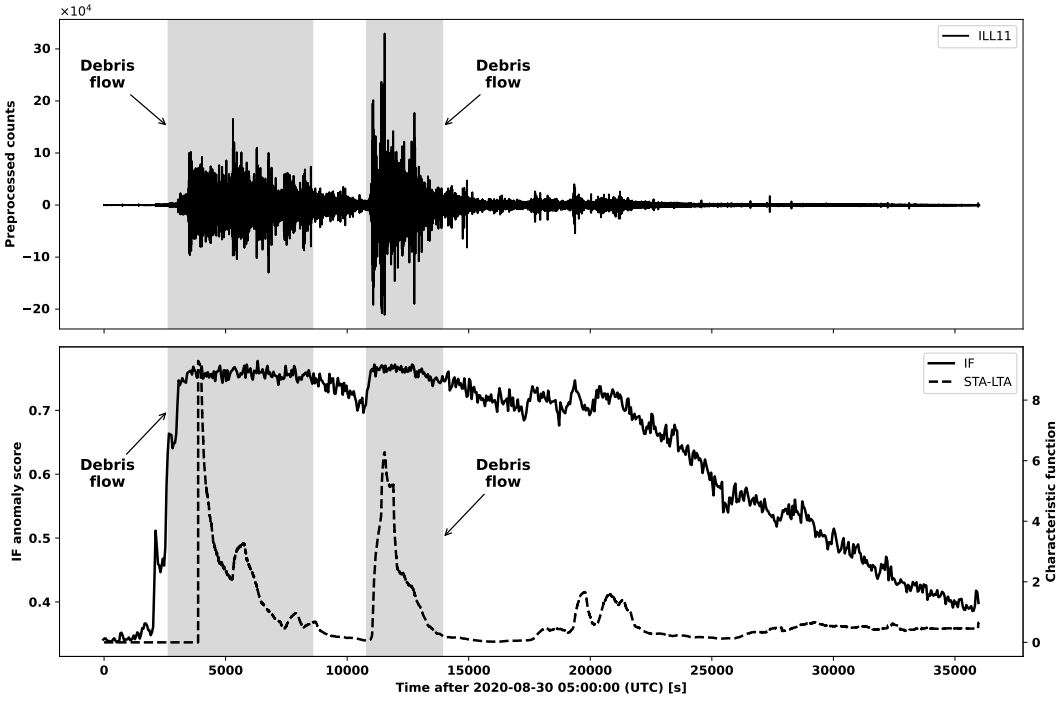


**Figure A2.** Illustration of the behavior of the STA-LTA characteristic function relative to the IF anomaly score at ILL11 on 2020-08-30. Debris flows are represented by the shaded regions.

3. Fig. B1 shows the Ward-linkage dendrogram of the IF segments extracted from the KARAT station.

4. ~~**Teleseismic earthquakes**. USGS earthquake catalog (**?**), see Table **??**~~Fig. B2 plots the cumulative fraction of mass-movements contained among the leading $k$ IF segments for $k \in \{1, 2, ..., 605\}$.

5. ~~**Regional earthquakes**. GEUS earthquake catalog (**?**), see Table **??**.~~ Fig. B3 contain time series plots of the amplitude of seismic waveforms and the IF anomaly score for the KARAT station.

~~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 7. Figures 14, 15, 16 and 17 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.~~

~~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 (**??**). Moreover, the IF trigger flags a typical calving seismogram with broadband signals representing the~~

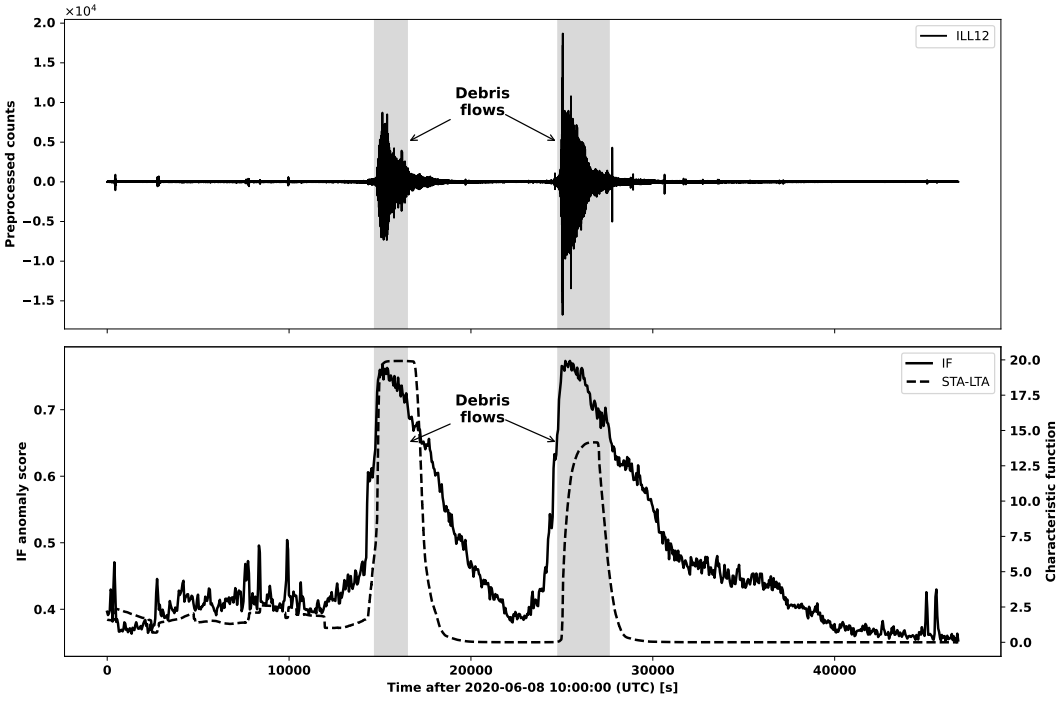


**Figure A3.** Illustration of the behavior of the STA-LTA characteristic function relative to the IF anomaly score ~~at ILL12 on 2020-06-08~~. Debris flows are represented by the shaded regions.

~~iceberg detachment (?) and a low-frequency (<0.01 Hz) signal generated by calving-induced water oscillations within the fjord ("seiche"; ?).~~

~~Satellite image pair of Rink Glacier calving front (Fig. 7) 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 $km^3$ assuming a terminus thickness of 500 – 600 m (?). Source: Sources: Copernicus (Sentinel-2 true color image).~~

~~Similar to Fig. 14, except that it shows the IF detection of a seismic event, which according to satellite images constitutes the capsizing of a tabular iceberg (Fig. 17). 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. 14). 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) (?)).~~

~~Satellite image pair of Rink Glacier calving front (Fig. 7) 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 (?), the iceberg had a volume of about 0.5 $km^3$ and may have contacted the terminus during capsize. Source: Copernicus (Sentinel-2 true color image).~~

| Station | Rank<br>Number of recordings | Number of counts |
|---|---|---|
| 21NUUG: 2017 | 2022-09-19T18:14:59.410000Z179 | 2022-09-19T18:18:19.410000ZM 7.6 - 35 km SSW of Aguililla,Mo |
| 24ILULI: 2017 | 2023-09-08T22:19:10.000000Z365 | 2023-09-08T22:28:20.000000Z3,153,620,341 |
| KARAT: 2022 - 2023 | M 6.8 405 | 6,620,667,831 |
| ILULI: 2022 - Al Haouz,Morocco 2023 | 726 | 6,264,509,194 |
| 26height | | |

**Table B1.** Number of mini-seed recordings and counts contained in the data used in the case study of Sect. 3.2, before any pre-processing is performed. The relevant years over which the statistics are extracted are contained in the Station column following the colon.

| Time window start | 2022-09-19T18:35:16.365000ZIF anomaly score |
|---|---|
| 2022-10-10T03:10:45.350000Z | M 7.6 - 35 km SSW of Aguililla,Mexico 320.726770 |
| 2023-08-03T00:25:42.32000050.000000Z | 2023-07-16T07:04:52.320000Z0.723054 |
| 332023-03-21T16:57:30.0000002023-08-25T12:41:40.000000Z | 2023-03-21T17:05:50.000000Z0.720424 |
| 482023-10-16T09:47:30.0000002023-07-27T07:00:00.000000Z | 2023-10-16T15:35:50.000000Z0.720379 |
| 502023-05-19T03:15:42.3200002023-07-03T09:34:52.320000Z | 2023-05-19T03:23:12.320000Z0.719906 |
| Teleseismic earthquakes in cluster A of Fig. ??. Rank 2022-12-27T03:01:40.000000Z | Start 0.715773 |
| 36 2023-04-11T11:37:22.320000Z | 2023-03-21T06:57:30.00000000.704491 |
| 2022-10-17T09:29:55.350000Z | 2023-03-21T07:0.703665 |
| 2023-10-02T15:01:40.00000049.415000Z | M: 3.9, Latitude: 69.088°N, Longitude: 53.429°W, 38 0.70109 |
| 2023-08-21T09:53:20.000000Z | 2023-04-19T01:38:12.320000Z 0.700739 |

**Table B2.** Regional earthquakes in clusters B and C Top ranking mass movements detected at KARAT, Greenland. Shown are the starting times of Figthe most anomalous time window according to the IF for each event and the corresponding IF anomaly score. ??CAL and ID stands for calving and iceberg disintegration events.

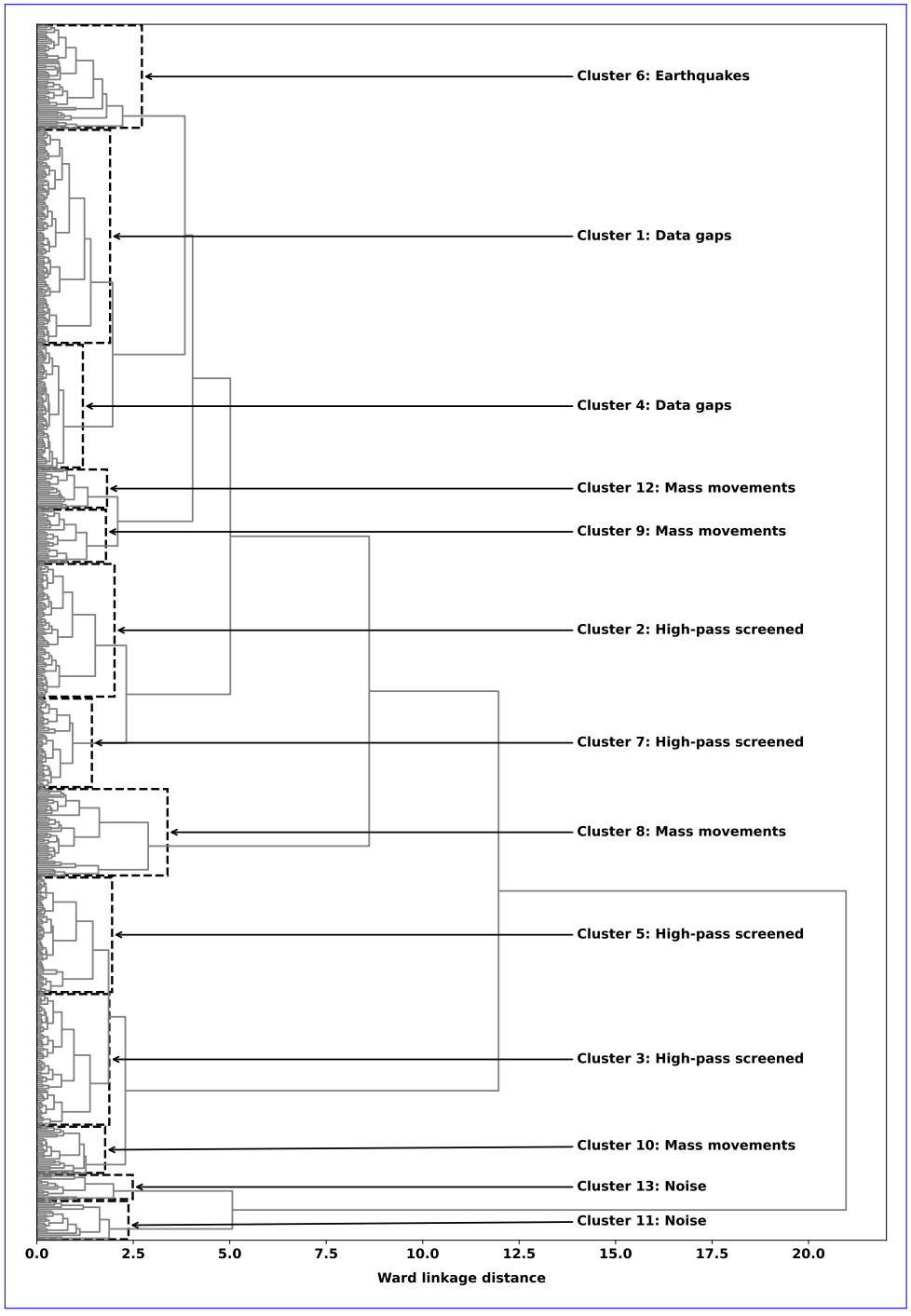


**Figure B1.** Ward linkage dendrogram of the 605 IF segments flagged at KARAT. Clusters are indicated by bounding boxes. High-pass screened means that after applying the high-pass screening rule described in Sect. 3.2.2 none of the remaining IF segments could be related to mass movements and was not explored.

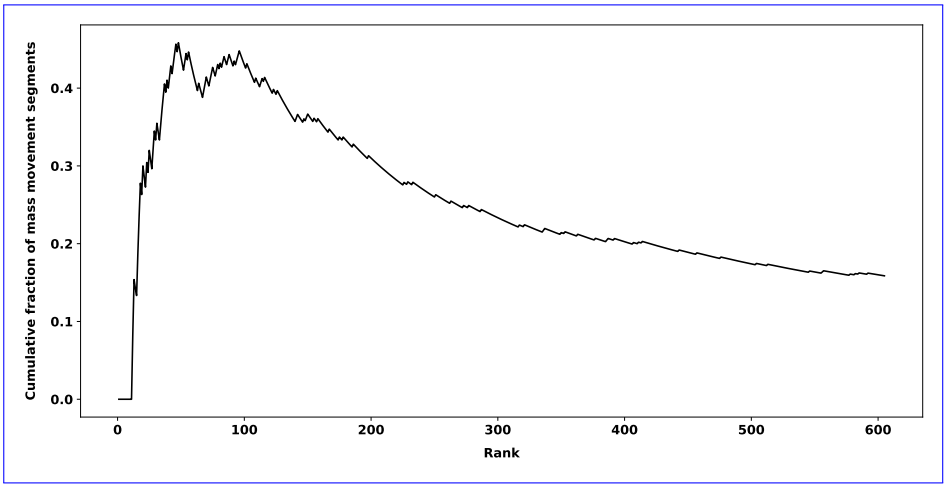


**Figure B2.** Cumulative fraction of mass-movements contained in the leading $k \in \{1, 2, \ldots, 605\}$ IF segments according to the IF anomaly score. The index $k$ is represented by the rank label on the x axis.

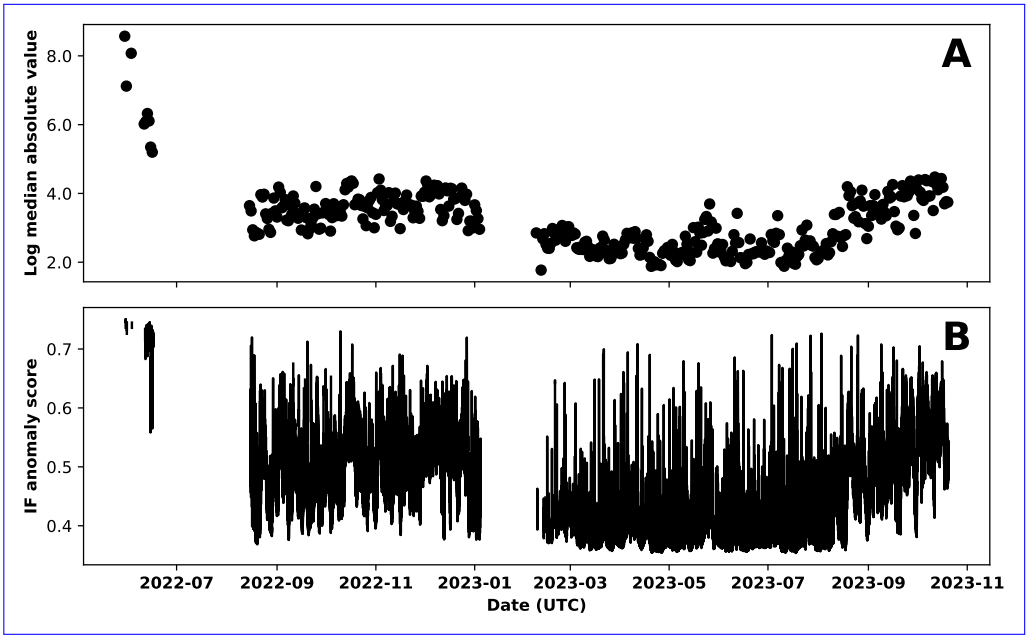


**Figure B3.** Log median absolute value of the daily preprocessed waveforms (A) and IF anomaly score (B) observed at KARAT.