An Unsupervised Machine-Learning Approach to Understanding Seismicity at an Alpine Glacier

Journal Article

Author(s):
Sawi, Theresa; Holtzman, Ben; Walter, Fabian Thomas [ID]; Paisley, John

JGR Earth Surface

RESEARCH ARTICLE

Key Points:

- We use unsupervised machine-learning and clustering to reveal time-varying features in seismic records from Gornergletscher, Swiss Alps
- We find clusters of seismic noise that correspond to varying hydraulic conditions at the glacier
- We propose an updated timeline for a glacial lake drainage and describe seismic signatures of subglacial meltwater flow

Correspondence to:
T. Sawi,
tsawi@ldeo.columbia.edu

Author Contributions:
Conceptualization: T. Sawi, Ben Holtzman
Data curation: Fabian Walter, John Paisley
Formal analysis: T. Sawi
Methodology: John Paisley
Software: John Paisley
Supervision: Ben Holtzman
Writing – original draft: T. Sawi
Writing – review & editing: Ben Holtzman, Fabian Walter, John Paisley

An Unsupervised Machine-Learning Approach to Understanding Seismicity at an Alpine Glacier

T. Sawi[1] , **Ben Holtzman[1]** , **Fabian Walter[2]**, and **John Paisley[3]**

[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA, [2]ETH Zürich Versuchsanstalt für Wasserbau, Hydrologie und Glaziologie VAW, Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft WSL, Birmensdorf, Switzerland, [3]Department of Electrical Engineering, Columbia University, New York, NY, USA

Abstract It is critical to understand the dynamic conditions of Earth's cryosphere, yet the subglacial and englacial environments that control many aspects of ice behavior are inherently difficult to observe. The study of seismicity in glaciers and ice sheets has provided valuable insights about the cryosphere for decades, more recently aided by tools from machine learning. Here, we present an unsupervised machine-learning approach to discovering and interpreting cryoseismic patterns using 5 weeks of seismic data recorded at Gornergletscher, Switzerland. Our algorithm utilizes non-negative matrix factorization and hidden Markov modeling to reduce spectrograms into characteristic, low-dimensional "fingerprints," which we reduce further using principal component analysis, then cluster using k-means clustering. We investigate the timing, locations, and statistical properties of the clusters in relation to temperature, GPS and lake-level measurements, and find that signals associated with lake flooding tend to occupy one cluster, whereas signals associated with afternoon and evening melt-water flow reside in others. We suggest that the one cluster contains signals that include the true initiation of the flood's englacial and subglacial drainage components. This work demonstrates an unsupervised machine-learning approach to exploring both continuous and event-based glacial seismic data.

Plain Language Summary Earth's climate change poses extreme risks for the stability of glaciers and ice sheets. One such glacier, Gornergletscher in the Swiss Alps, has been monitored for decades, along with a lake on the glacier which from the 1950s until recently filled and drained every summer, releasing huge amounts of water under, through, and on top of the glacier. To study this glacier, which is hundreds of meters thick, and its lake, scientists employ "cryoseismology," the study of seismic excitations in the ice. In our study, we use unsupervised machine-learning, a type of artificial intelligence that identifies patterns in large data sets, to organize 5 weeks of cryoseismic records into clusters composed of similar types of signals. Although the machine-learning algorithm has no knowledge of glacial hydrology nor its evolution over time, the output places the signals associated with the beginning of lake flooding in one cluster, and signals associated with afternoon and evening melt-water flow in others. We also find that "icequakes" are clustered based on extremely subtle, similar, vibrations. From these findings, we provide an improved chronology for the glacial lake flood, and a better understanding of the cryoseismic signals caused by water flowing underneath and through glaciers.

1. Introduction

Climate change is disrupting cryospheric systems across the globe (IPCC, 2022), highlighting the importance of monitoring and understanding the conditions of glaciers and ice sheets. Even with powerful tools for monitoring, it remains difficult to observe the interiors and ice-bed interfaces of glaciers and ice sheets. Fortunately, decades of research in cryoseismology have resulted in powerful tools for monitoring surficial, englacial and subglacial processes. Cryoseismic sources are diverse, including impulsive "icequakes" caused by ice flow and deformation (see reviews by Aster and Winberry [2017] and Podolskiy and Walter [2016]) and high-frequency (0.5–20 Hz) tremor caused by turbulent water flow and bedload transport in sub- and englacial water channels (e.g., Bartholomaus et al., 2015; Eibl et al., 2020; Gimbert et al., 2016; Lindner et al., 2020; Röösli et al., 2014; Vore et al., 2019). "Noise"—that is, continuous seismic signals that are mostly devoid of impulsive events—can include information about time-varying seismic sources and acoustic properties of the medium (Lindner, 2020), and have been linked to sliding at the ice-bed interface (Podolskiy et al., 2021; Umlauft et al., 2021) and varying environmental conditions including wind, temperature and precipitation (Chaput et al., 2018).

The large volume and variety of cryoseismic data (not to mention its noisy environment) can make the characterization, especially the detection, location, and classification of events, a challenging process when using typical seismic analytic techniques (e.g., template matching, cross-correlation, auto-correlation). Given a certain data distribution, however, larger data sets, tend to be better-suited for machine-learning models, which, as with all statistical models, are typically improved with larger representative samples. Machine learning has recently gained enormous interest within the seismological community (e.g., reviews by Kong et al. [2019] and Bergen et al. [2019]), with most studies aimed at improving analyst-led tasks such as waveform phase detection, association, and event classification (e.g., Mousavi et al., 2020; Park et al., 2020; Ross et al., 2018, 2019; Zhu & Beroza, 2019). The machine-learning methods in such studies are typically supervised, meaning the models are optimized to match manually labeled data (e.g., Murphy, 2022). Supervised machine learning has shown remarkable success at classifying seismic signals at volcanoes (e.g., Ren et al., 2020, and reviews by Malfante et al. [2018] and Carniel & Guzmán [2020]), where the combination of tectonic-like and fluid-mediated behaviors are at times analogous to temperate glacial settings.

Typically, data from field studies will likely emerge unlabeled, underscoring the need for the development of unsupervised machine-learning (UML). UML infers patterns in data sets without first learning on labeled data, either by clustering data into groups based on similarity or by reducing data to lower-dimensional, statistically relevant "features" (Murphy, 2022). Since no labels are predicted, UML results can be difficult to interpret physically, and since there is no defined target, not all clusters or features may be of scientific interest. Despite these challenges, numerous seismic studies have produced insight through unsupervised feature-extraction, clustering, or a combination of the two (e.g., Chamarczuk et al., 2019; Sick et al., 2015; Steinmann et al., 2021; Trugman & Shearer, 2017; Yoon et al., 2015), including in numerous glaciated, volcanic and/or geothermal settings (e.g., Holtzman et al., 2018; Jenkins et al., 2021; Lamb et al., 2020; Ren et al., 2020; Seydoux et al., 2020). Seydoux et al. (2020), for example, uses a Gaussian mixture model to cluster features that were automatically extracted using a deep scattering network (a type of convolutional neural network) from continuous seismic data at Greenland, and are able to identify precursory seismic activity leading to a massive landslide. Another example of UML applied to cryoseismic settings is in Jenkins et al. (2021), where the authors compare two types of clustering (Gaussian mixture model and deep-embedded clustering) on 2 years of continuous seismic data from the Ross Ice Shelf (Antarctica), and find that numerous clusters correspond to oceanographic and atmospheric forcing.

Here, we employ an UML algorithm that extracts subtle, characteristic, time-varying features from spectrograms and transforms them into low-dimensional representations called fingerprints, an approach introduced by Holtzman et al. [2018]. We apply an algorithm, now called Spectral Unsupervised Feature Extraction, or "SpecUFEx," to 5 weeks of glacial icequakes and continuous glacial seismic noise from Gornergletscher, Switzerland. Our aim is not to distinguish obviously different types of signals, for example, between noise and icequakes or between basal icequakes and surface icequakes, but rather to discover subtle differences in a well-constrained data sets that may have eluded visual or other detection. We find strong evidence that different clusters of fingerprints correspond to different modes of glacial hydrologic behavior, including the englacial and subglacial components of a glacial lake outburst flood and diurnal meltwater flow through systems of subglacial conduits. Groups of highly similar icequakes are clustered together as well. Semi-automated, unsupervised machine-learning approaches such as ours will provide a practical path for understanding the behavior of glaciers and ice sheets, particularly as data streams—and concerns over the state of glaciers and ice sheets—grow rapidly.

2. Background

Gornergletscher is the second largest glacier in the Swiss Alps, spanning approximately 57 km^2 at an average elevation of 2,500 m (Figure 1). Gornergletscher converges with Grenzgletscher, flowing generally westward at a speed of ~4–10 cm/day (Garcia et al., 2019). Until 2007, an ice-marginal lake, Gornersee, formed every spring at the confluence of the two glaciers, and by midsummer drained rapidly in a glacial outburst flood, releasing up to 6 million m^3 s^{-1} of water in a combination of subglacial, englacial, and supraglacial flow (Huss et al., 2007; Werder et al., 2009, 2010).

Seismicity at Gornergletscher is primarily controlled by meltwater runoff. Figure 2 shows seismicity rates peaking in the afternoon as meltwater percolates through crevasses to the ice-bed interface, elevating the glacier to near flotation levels and allowing for reduced basal drag and increased ice flow (Roux et al., 2010; Walter, 2009; Walter et al., 2008). Walter (2009) detected hundreds to thousands of icequakes daily in June and July 2007, with
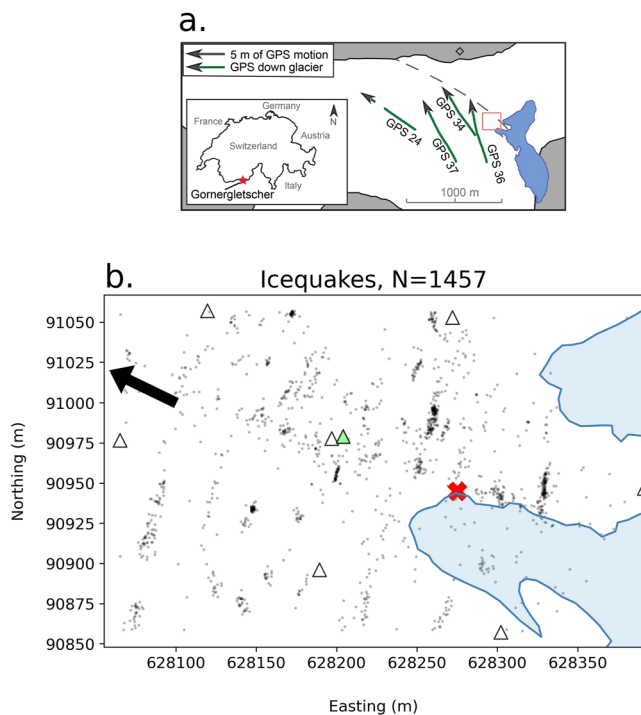
Figure 1. Map of study region and icequakes. (a) From Garcia et al. (2019). Two glaciers, Gornergletscher (north) and Grenzgletscher (south), converge at Gornersee (blue polygon) and flow roughly north-westward, as indicated by GPS displacement (green lines and arrows). Precipitation and temperature data come from an autonomous weather station (black diamond). The red box shows the approximate limits of the seismic array footprint, and the inset map shows Gornergletscher on the national scale. (b) Seismic array (black triangles) and icequake locations (black dots) from Walter (2009). Station J8 (green triangle) is used in this study. Blue polygon shows the lake at its maximum extent. Black arrow shows the flow direction of the glacier. Red X marks the approximate location of the moulin.

the vast majority (estimated 99%) caused by tensile openings of crevasses in the top 20 m of the ice. A few tens of basal icequakes were also detected daily, typically in the early morning hours when subglacial water pressure was at its lowest. At Gornergletscher, icequakes frequently occur alongside continuous background signals generated by sources such as ice-proximal water flow (Lindner, 2020), wind- or rain-induced vibrations in the ice, or superimposed stick-slip events near the glacier bed (Umlauft et al., 2021), highlighting some of the complexities that can arise when analyzing glacial cryoseismic data.

A timeline of reported events for the 2007 lake glacial outburst flood is as follows. First, lake drainage began on 4 July, as Gornersee breached its shore and flowed over the ice to a lake-proximal moulin (Werder et al., 2009). From 7 to 9 July, water began to drain subglacially, and from 9 to 15 July the majority of the lake drained through a crevasse exposed at mid-basin height, with the drainage rate varying as new hydro-fractures were initiated (Werder et al., 2009). The drainage concluded when the lake was mostly empty and the lake-pressure transducer subaerially exposed (Werder et al., 2010). Although sediment-transport data were not collected during the 2007 study period, a borehole study at the northern slope of Gornergletscher (about one km downstream of our study region) determined that subglacial conduits there contained likely little sediment (Iken et al., 1996). The annual outburst floods typically accelerate the flow of Gornergletscher (Sugiyama et al., 2007, 2008, 2010), triggering icequake activity in the top 20 m of the ice (Garcia et al., 2019; Riesen et al., 2010; Roux et al., 2010). Accordingly, a burst of shallow icequakes marked the onset of the 2007 lake drainage (although no such increase in basal seismicity was observed (Walter, 2009).

3. Data

Our seismic data are from an eight-instrument array of three-component Geospace GS-11D geophones operating in shallow boreholes at Gornergletscher between 14 June and 22 July 2007. For this proof-of-concept study, we focus on data from station J8, which demonstrates good data quality and lies nearest the center of the array (1). Future studies could analyze spectrograms from multiple stations by either performing this analysis on each station separately and comparing results, or combining the spectrogram data sets from multiple stations prior to applying SpecUFEx, and clustering multiple stations at once. We use only the vertical-component data, which are less affected by tilt noise. The instrument has an 8 Hz natural frequency, and since data are flat to velocity in the frequency ranges of interest, no instrument corrections are done. Station J8 sampled continuously at 1,000 sps and was occasionally redrilled (every 1–2 weeks) to account for the effect of differential melting of the ice surface beneath the instrument. It is unlikely that the redrilling systematically changed the coupling, given that the geophones never fully re-froze to the borehole, which was typically filled with water.

We perform twin UML analyses on two separate data sets: one containing 2-s-long icequake waveforms from the catalog created by Walter (2009), and one containing sixty-second-long records of background noise compiled for this study. The two experiments help explore how features extracted from icequakes arise from systematic effects of background noise. Additionally, exploring the isolated noise allows for a more nuanced view of seismic tremor behavior and gives insights into sub- and englacial dynamic and hydrologic behavior.

3.1. Icequake Selection and Locations

For the icequake analysis, we use the catalog of Walter (2009), which includes over 100,000 icequakes in June and July 2007. The icequakes are high frequency (~50 Hz), low moment magnitude ($M < 1$), short-duration (~1 s), and events have prominent Rayleigh wave arrivals. The icequake waveforms, compiled by Walter (2009),
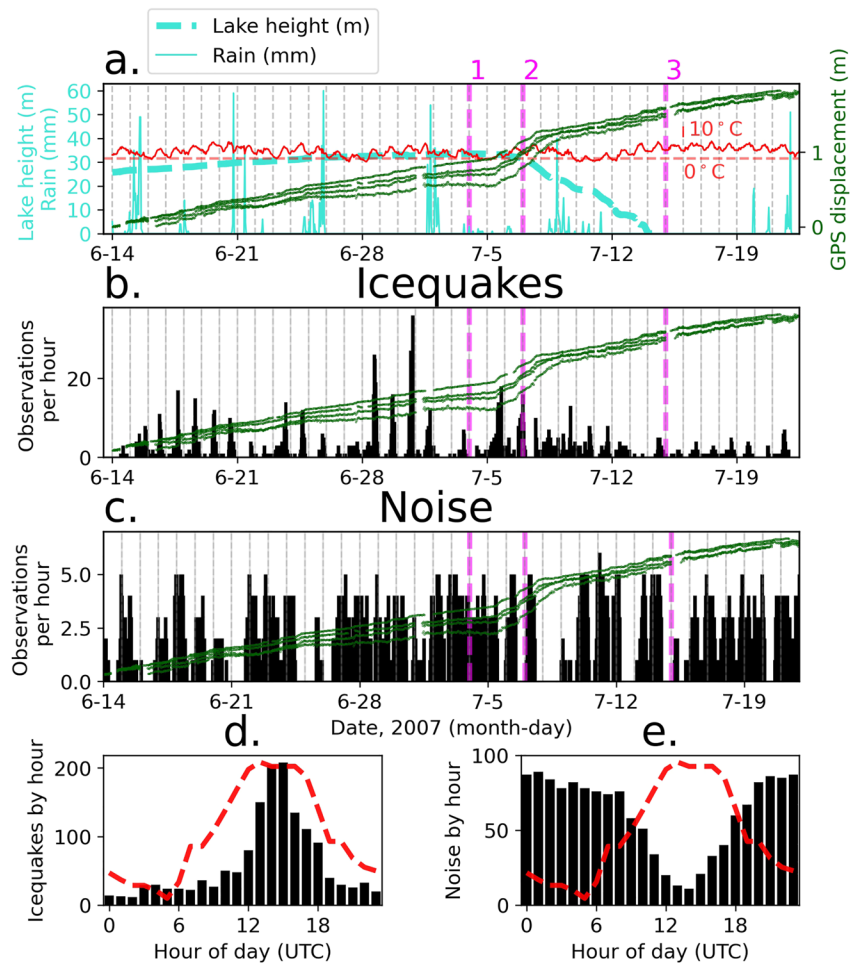
Figure 2. Timing of icequake and noise samples with ancillary geophysical data. (a) GPS displacement (green lines), lake level (thick dashed blue line), precipitation (thin solid blue line), and ambient temperature (red line) from Garcia et al. (2019). Fuchsia dashed lines show the timing of lake-drainage events: 1. is the onset of supraglacial lake drainage, 2. is the onset of the sub- and englacial drainage, and 3. is when drainage stops. Icequakes (b) and noise samples (c.) binned by hour through the study season. Counts of icequakes (d) and noise samples (e) by hour of the day (UTC). The red dashed line shows the mean hourly deviation from daily mean temperature throughout the season, with scale shown in panel (a).

occur within two-second-long windows that are defined to include 0.5 s of waveform before the Rayleigh wave arrival and 1.5 s after it. A subset ($N = $ ~24,000) of the detected events were located by Roux et al. (2010) using a grid-search inversion algorithm (W. H. K. Lee & Stewart, 1981). Mean location errors for events located within the network are assumed to be no more than 10 m (Roux et al., 2010). We limit our study to icequakes located within the seismic-array footprint and in the top 50 m of ice where locations are most accurate ($N = 2,806$). In previous experiments, we have analyzed icequakes outside of that boundary, in which cases we found that systematic location errors caused by poor azimuthal coverage obfuscated interpretation of results. About 6% of the events ($n = 157$) are unable to be loaded due to a digitizer error.

We verified that there was only one icequake per window by using a short term average/long term average (STA/LTA) filter. STA/LTA filters are calculated by taking the ratio of two moving averages of the waveform signal; one short window and one long window. An impulsive signal will cause the ratio to exceed a prescribed threshold and trigger the STA/LTA filter. For the icequakes, we use a short-term window of 0.05 s, a long-term window of 0.4 s, and a threshold of 7.5. We only retain records where the STA/LTA threshold is triggered exactly once in a 0.25 s period, resulting in a data set of 1,457 icequakes (Figure 1). We tuned the STA/LTA parameters on a subset of icequakes ($n \sim 100$), and, through visual inspection, found them to be effective for filtering out seismic records with multiple, or no, icequake arrivals.

Table 1
Parameters Used to Generate Icequake and Noise Spectrograms

	Icequakes	Noise
STFT window length (s)	0.08	1.1
STFT window overlap	25%	25%
Number of STFT windows	33	72
Minimum frequency (Hz)	15	1
Maximum frequency (Hz)	80	80

3.2. Noise Selection

The second experiment is performed on 1,472 min-long segments of seismic noise. We define noise as continuous seismic records free of impulsive icequakes. We detect icequakes by using an STA/LTA filter with a 0.03 s short-term window, a 0.5 s long-term window, and a ratio threshold of 10. Similar to the icequakes, we tuned the STA/LTA parameters using a subset of noise segments ($n \sim 100$) to adequately filter out seismic records with icequake arrivals. We extract 1-min-long waveforms every 10 min during the same 5-week time period as the icequakes and check for prominent icequakes in the record using the STA/LTA filter. We choose to sample every 10 min to ease the computational load, since calculating spectrograms for every minute of the 2-month period would be excessively time-consuming. If the filter exceeds the STA/LTA threshold, we advance 5 s and extract another minute-long window on which to apply the filter. Once an icequake-free window is found, we advance 10 minutes and repeat the process. Figure 2e shows the distribution of selected noise windows binned by hour. Because icequakes occur frequently during the afternoon hours (at times >5 icequakes per minute) many noise windows are discarded during those hours, leading to under-sampling of noise in the afternoon.

3.3. Ancillary Data

We compare seismic data to GPS, lake level, temperature and precipitation measurements made nearby to understand how environmental factors may influence seismicity, as shown in Figure 2a. Four GPS stations were located on the ice between 100 m and 3 km down-glacier from the lake, providing position estimates every 2 min with horizontal and vertical accuracy of ±1.4 and ±3 mm, respectively (Sugiyama et al., 2007). We use GPS displacement estimates from Garcia et al. (2019), referenced to the GPS station location at the beginning of our study. We detrend the displacement estimates using the pre-drainage glacier velocities reported by Garcia et al. (2019) for each GPS station (3.8, 4.7, 6.8 and 6.8 cm/day for stations 24, 34, 36, and 37, respectively). We then take an hourly running average for each station. This approach allows us to visualize changes in glacier displacement (green lines, Figure 2) associated with diurnal variations, varying temperature, precipitation, and the glacial outburst flood. Lake levels were measured by a pressure transducer sampling once per minute (thick blue dashed line, Figure 2a; adapted from Werder et al. [2009]). Relative ambient temperature (red solid line) and precipitation levels (thin solid blue line) are from an automated weather station located on the northern margin of Gornergletscher (black diamond, Figure 1), 2.5 km north of, and 600 m above, the seismic array (Garcia et al., 2019).

4. Methods

We define a workflow in which we use short time Fourier transforms (STFT) to transform waveforms into spectrograms, employ SpecUFEx to reduce the spectrograms into fingerprints using non-negative factorization (NMF) and hidden Markov models (HMM), apply principal component analysis (PCA) to the fingerprints, cluster the PCA-fingerprints using k-means clustering evaluated with silhouette scores, and then highlight a representative subset of events from each cluster for more in-depth exploration. We then investigate the timing, locations, and statistical properties of events in the clusters in relation to the ancillary geophysical measurements (i.e., temperature, GPS, and lake-level) described above.

4.1. Spectrogram Generation

We calculate spectrograms from digital counts that are proportional to ground-velocity measurements using a short-time Fourier transform (STFT) and the parameters listed in Table 1. Frequency bounds were chosen to capture the majority of the spectral energy (Figures 3a and 3b). The minimum frequency, f_{min}, is near the theoretical limit defined by the reciprocal of an STFT window length chosen to allow for at least 30 time samples per seismic record, a value we have found empirically to be the minimum number of time samples needed for the HMM. The HMM learns on time-varying patterns between concurrent STFT windows, such that too few windows will yield too little temporal information. We choose f_{max} in order to explore possible signals of interest in higher frequencies while avoiding instrument noise that occurs at and above 90 Hz. Following Holtzman
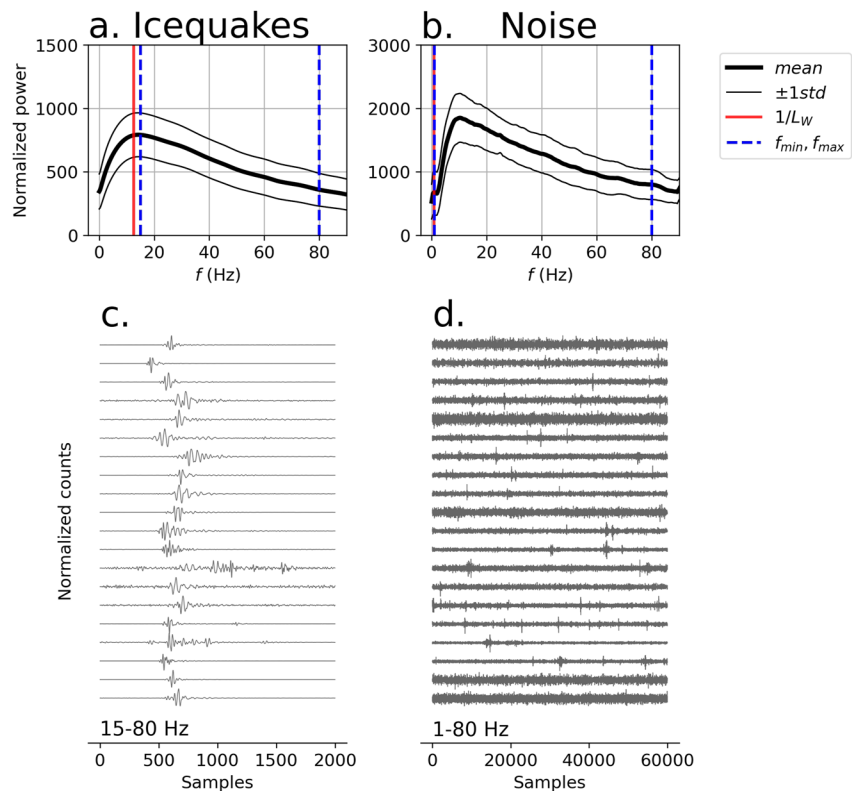
Figure 3. Spectrogram parameter choices and randomly selected waveforms for icequakes (left column) and noise (right column). Top panels show the mean and standard deviations of spectra for icequakes (a) and noise (b) calculated by summing the STFT windows across all frequencies over time after normalization and converting to decibels (described in Section 4.1, "Spectrogram Generation"). Blue dashed lines show the frequency bounds, and the red line shows the minimum frequency resolved by the STFT window length. Lower panels show 20 randomly selected waveforms for (c) icequakes and (d) noise. Each waveform is normalized by its maximum amplitude after the bandpass filter, listed below the traces, is applied.

et al. (2018), we normalize each STFT window by its median, convert it to decibels with a reference value of 1, and set any negative values to zero in order to comply with the non-negativity constraint of NMF (described below). Empirically in this and a previous study (Holtzman et al., 2018), we find that these normalization steps improve SpecUFEx and clustering performance in terms of identifying more discrete clusters of fingerprints.

4.2. SpecUFEx: Spectral Unsupervised Feature Extraction

SpecUFEx (Holtzman et al., 2018) applies two layers of unsupervised feature extraction to the spectrograms: nonnegative matrix factorization (NMF) and hidden Markov modeling (HMM). The workflow is illustrated in Figure 4. NMF reduces the spectrograms into a basis dictionary of spectral patterns that are characteristic of the entire data set, and, for each spectrogram, an activation coefficient matrix (ACM) that expresses how the dictionary patterns vary over time (e.g., D. D. Lee & Seung, 1999). Examples of spectral patterns are the columns in the NMF dictionary (Figure 4a). The NMF dictionary is shared for all spectrograms, but the ACMs are unique for each spectrogram. The dominant spectral patterns of the spectrogram are retained in the ACM even as the frequency resolution is drastically reduced (Figures 4a and 4b).

The ACMs are then used as input for a HMM (Baum & Petrie, 1966). The HMM assumes that the temporal evolution of the ACMs' spectral patterns depends on hidden states defined by frequency patterns that tend to co-occur in time. Like (Holtzman et al., 2018), we choose 15 hidden states for the HMM. The hidden states are given in the HMM emissions matrix (EM), which, similar to the NMF dictionary, is shared for all spectrograms. The state transition matrices (STMs, Figure 4c) show how hidden states vary in time for each ACM. The final product, a fingerprint (Figure 4c, right), counts how many times each state transitions to another (or itself) from one time step to the next. The fingerprints only retain the number of transitions, not their order, making them
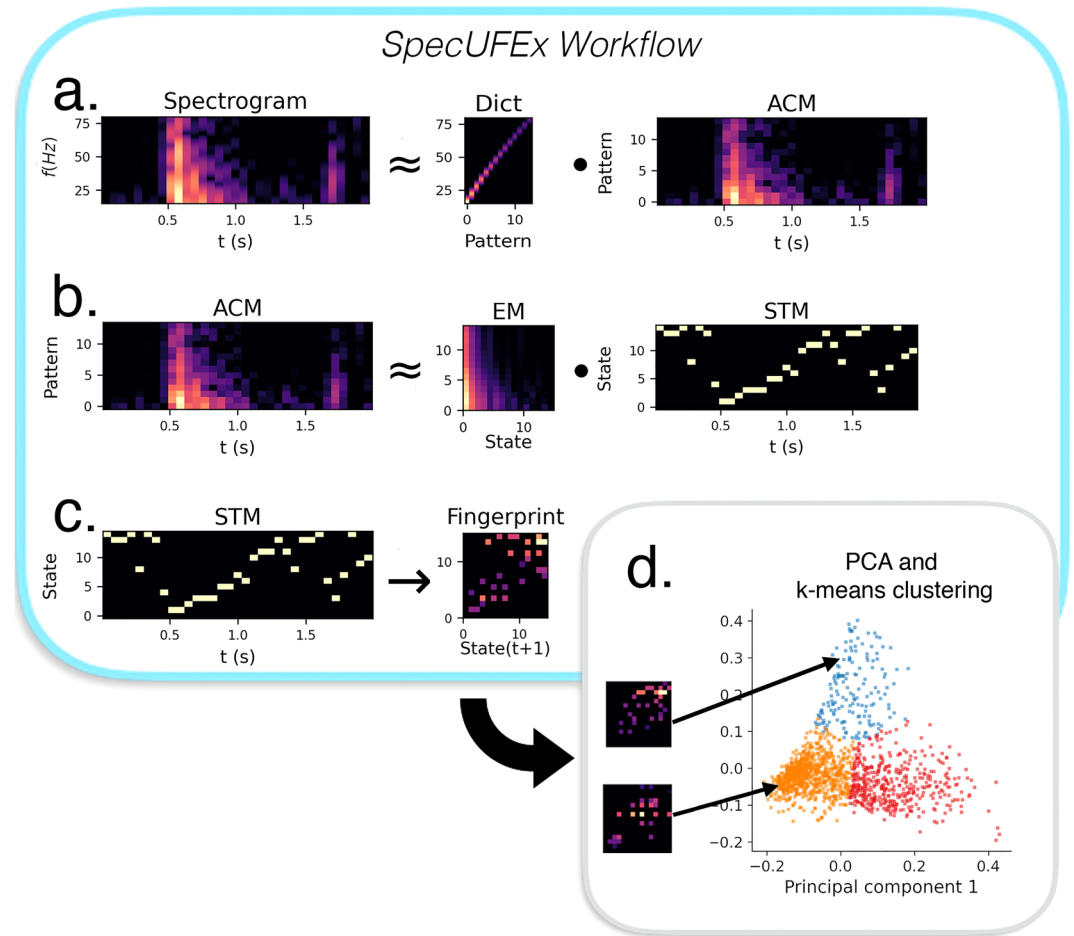
Figure 4. SpecUFEx workflow. SpecUFEx (Holtzman et al., 2018) applies two layers of unsupervised feature extraction to the spectrograms: (a) non-negative matrix factorization and (b) hidden Markov modeling. (c) Fingerprints are made by counting state-transitions from the HMM state transition matrix (STM). (d) Following principal component analysis (PCA), *k*-means clustering is applied to the fingerprints.

time-invariant with respect to the spectrogram. As a result, full temporal sequences cannot be reconstructed from the fingerprints.

4.3. Principal Component Analysis

The fingerprints are 255-dimensional (15×15 features) and generally sparse, so we reduce their dimensionality further prior to clustering using PCA on the raw fingerprint values (e.g., Karl Pearson, 1901; Lever et al., 2017); Figure 4d). This additional dimensionality reduction allows for clearer visualization of the clustering results and other trends in the data. We choose the number of principal components so that at least 86% of the variance of the original fingerprint data set is preserved.

4.4. Clustering and Validation

We cluster the PCA-reduced fingerprints (Figure 4d) with the k-means clustering algorithm of Pedregosa et al. (2011), which separates data points, x_i, into J clusters based on their proximity to centroids μ_j by optimizing the function:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\|x_i - \mu_j\|^2 \right).$$
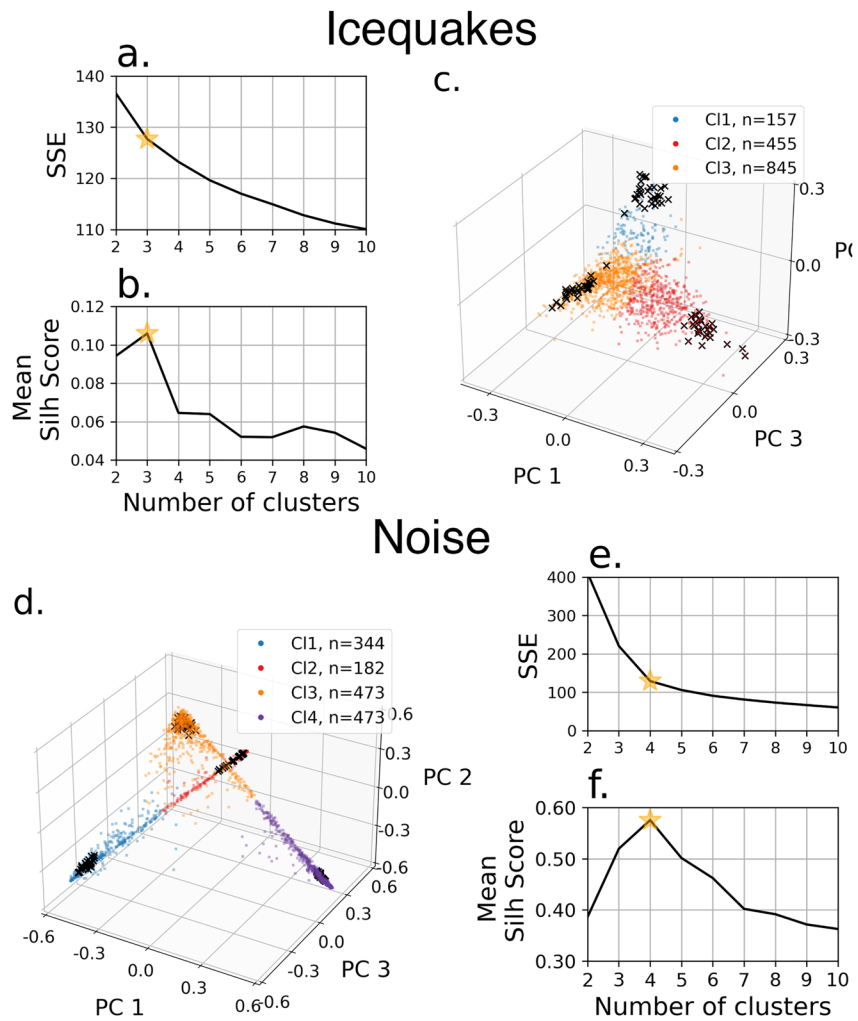
(1)

Figure 5. PCA projections of fingerprints from the (c) icequakes and (d) noise samples colored by *k*-means cluster assignments. Black x's are 30 "representative" fingerprints for each cluster, defined as those with the highest silhouette ("Silh") scores in their cluster. Note that representative fingerprints can overlap in PCA-space. Line plots show cluster validation metrics for two to 10 clusters of icequakes (a and b) and (e and f) noise. The sum of squared errors (SSE, panels [a and e]) is the distance of each point to its cluster centroid. Panels (b and f) show the mean silhouette score of all clusters.

We choose *J*, the number of clusters, based on the clusters' mean silhouette scores (SiS; Rousseeuw, 1987). SiS quantify how well-clustered a given fingerprint is by assigning it a score from −1 to 1 with positive values indicating the point is on average closer to other points in its cluster than other points in the next closest neighboring cluster. The silhouette score, $SiS(x)$ is defined as:

$$SiS(x) = \frac{b - a}{\max(a, b)}, \tag{2}$$

where *a* is the mean Euclidean distance of a fingerprint, *x*, to every other fingerprint in its cluster, while *b* is the mean Euclidean distance of *x* to every fingerprint in the closest neighboring cluster. Negative silhouette scores imply that a fingerprint is erroneously clustered, meaning that it is on average closer to points in a different cluster than its own cluster. We also evaluate the clustering results using the sum of squared errors (SSE) between the data points and their assigned cluster centroids. By plotting SSE for an increasing number of clusters, we can find at which point adding more clusters leads to smaller improvements in the SSE, leading to a break in the slope (see Figure 5e).

4.5. Statistical Measures

We employ two statistical measures to understand the basic properties of the waveforms prior to their transformation into spectrograms. These properties are used solely for exploring results post-SpecUFEx, and are not used as features for UML. The first is the sum of the absolute amplitude of a waveform, a variation of the "real-time seismic amplitude measurement" (RSAM) introduced by Endo and Murray (1991) and frequently utilized in volcano observatories as a proxy for volcanic activity. RSAM is defined as:

$$\text{RSAM}(x) = \sum_{t=0}^{T} |x(t)|, \tag{3}$$

where $x(t)$ are digital counts (proportional to seismic ground velocity) at time-step t after being filtered through a four-pole Butterworth bandpass filter using the minimum and maximum frequencies given in Table 1.

We also calculate the spectral centroid, $SC(x)$, of each waveform's spectrum. The spectral centroid is an audio information retrieval metric meant to quantify the pitch of a waveform (Klapuri & Davy, 2007). We apply a fast Fourier transform to the waveform and treat the resultant spectrum as a distribution of N evenly separated frequency bins from which the centroid, the weighted mean, is calculated

$$SC(x) = \frac{\sum_{n=1}^{N} f(n)s(n)}{\sum_{n=1}^{N} s(n)}, \tag{4}$$

where $f(n)$ is the center frequency of bin n, and $s(n)$ is the mean value of all frequencies in the bin.

5. Results

5.1. PCA and Clustering

Figure 5 shows the first three principal components of the fingerprints after applying PCA to both the icequake and noise data, along with the results of our clustering evaluation procedure. Based on silhouette scores, we choose three clusters for the icequake fingerprints and four clusters for the noise fingerprints. When a data point has a negative silhouette scores, it indicates that it, the data point, would have had a higher silhouette score in a different cluster, and therefor is erroneously clustered. Only 100 of 1,457 icequake fingerprints (<7%) and 1 of the 1,472 noise fingerprints (<0.1%) have negative silhouette scores, meaning that very few of the fingerprints are erroneously clustered. Figures 5c and 5d shows the fingerprints colored by their cluster assignment, projected onto their first three principal subspaces. The colors are assigned randomly, and have no connection across icequake and noise experiments. In this figure, ∼28% and ∼86% of the variance of the original fingerprint data set is preserved for icequakes and noise, respectively. Rather than clustering in distinct, isolated groups, the PCA plots demonstrate that our data form a continuous manifold with 3 or 4 branches (one for each cluster) extending outwards.

We select a subset of fingerprints from each cluster to highlight for analysis, chosen as the 30 events with the highest SiS in each cluster. We refer to these as the "representative" events from each cluster, marked with black crosses in Figures 5c and 5d. The 30 representative events in each cluster are located near the ends of the branches of the PCA manifold, therefore emphasizing variations between clusters more clearly than events near the center of the manifold, which can be close to events assigned to other clusters. By highlighting the representative events in our analysis, we can visualize minute variations between clusters that can be overlooked when examining the cluster population as a whole.

5.2. Icequake Locations

Figure 6a shows the icequake locations colored by cluster assignment with the 30 representative icequakes from each cluster outlined in black. Representative icequakes in the red (2nd) cluster tend to occur along a linear pattern of icequakes that trace out a large (>40m), north-south striking, lake-marginal crevasse. Icequakes in the representative orange cluster generally appear along crevasses as well, and both the orange and blue representative icequakes are dispersed about the study region, covering a wide range of azimuths and distances. Unlike the red and orange clusters, the icequakes in the blue cluster are not generally located along crevasses. Figure 6b
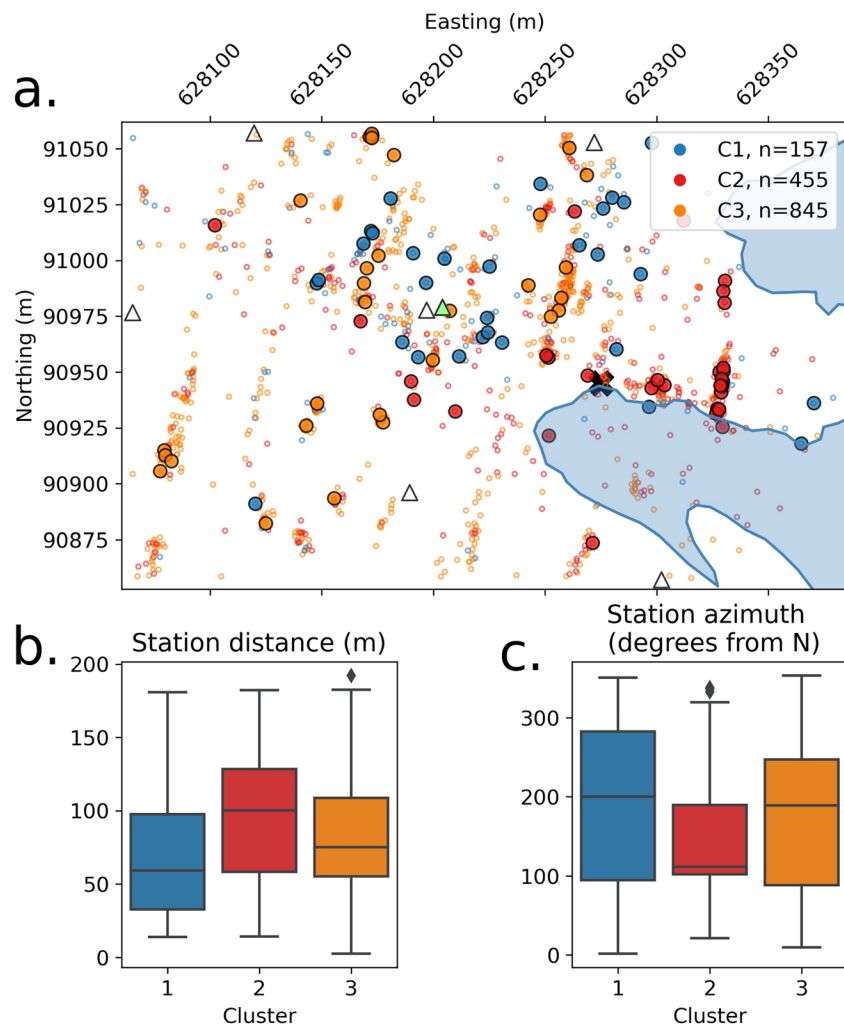
Figure 6. (a) Icequake locations colored by cluster assignments. The 30 representative icequakes from each cluster are outlined in black, and the moulin is represented by a black X. Symbols are otherwise identical to Figure 1. Box-and-whisker plots for station-event (b) epicentral distances and (c) azimuths. Boxes enclose the 25th to 75th percentiles of the location measurements, and the horizontal line in the box shows the 50th percentile. The whiskers extend to 1.5 times the range between the 25th and 50th percentiles, and individual points beyond the whiskers are outliers.

shows box-and-whisker plots for the epicentral icequake-station distances for all events in each cluster. Boxes enclose the 25th to 75th percentiles of the location measurements, and the horizontal line in the box shows the 50th percentile. The whiskers extend to 1.5 times the range between the 25th and 50th percentiles, and individual points beyond the whiskers are outliers. The red (2nd) cluster's representative events tend to be farther from the station than the blue and orange clusters', events in the blue cluster are overall closer. Figure 6c describes the station-event azimuths for all events in each cluster. The blue (1st) and orange (3rd) clusters are more dispersed about the study region, whereas the red (2nd) cluster has a narrower distribution, with the majority of events occurring between 100 and 190° from north.

5.3. Seasonal Cluster Behavior

Figure 7 shows bar-plots of the icequake and noise clusters throughout the June-July season, where each row belongs to a different cluster. The fuchsia lines indicate notable lake-drainage events: (1) the onset of supraglacial lake drainage, (2) the onset of sub- and englacial drainage, and (3) the end of drainage, as reported by Werder et al. (2009). For the icequakes, the blue (1st) cluster events occur primarily on or after the morning of 5 July, with the majority of events occurring between 8 and 11 July. The red (2nd) and orange (3rd) clusters appear
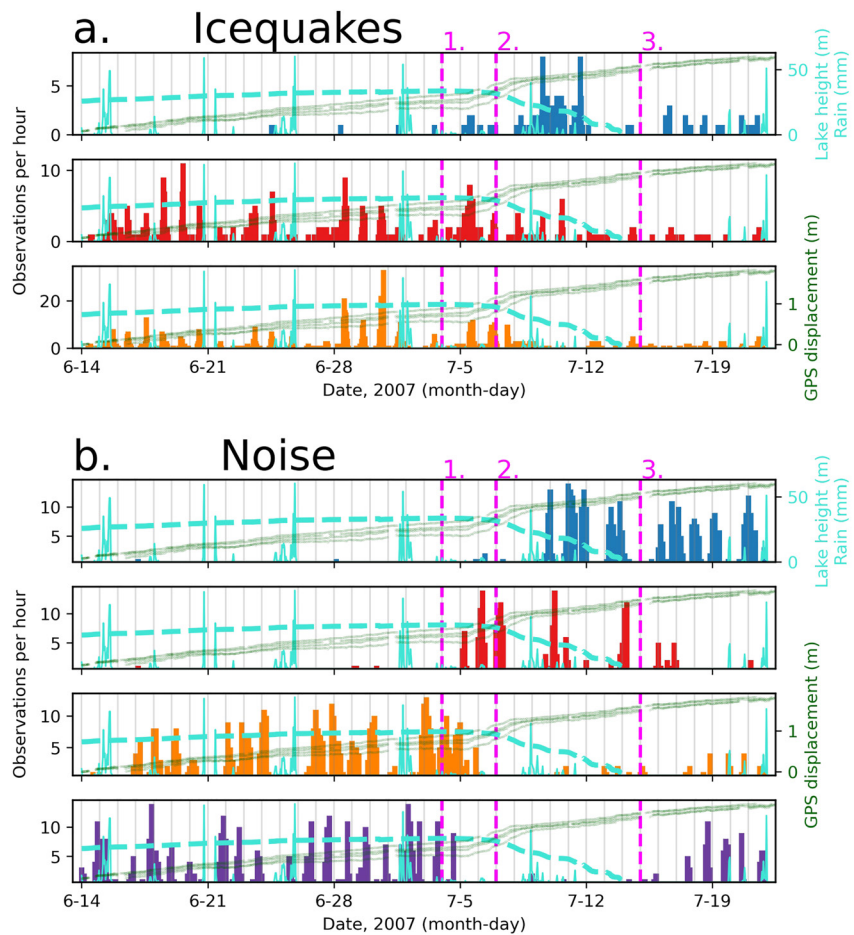
Figure 7. Comparison of cluster timing to geophysical data for icequakes (top) and noise samples (bottom), both binned in 3-hr windows. Each row corresponds to a cluster, with colors as in Figure 5. Green traces and thick, blue dashed lines show GPS displacement and lake height, respectively. Thin blue line is precipitation. Fuchsia dashed lines mark important glacial flood events: (1) supraglacial lake drainage, (2) sub- and englacial drainage, and (3) end of drainage.

continuously throughout the season, although they decrease following the onset of the sub- and englacial lake drainage (7 July, 2nd fuchsia line).
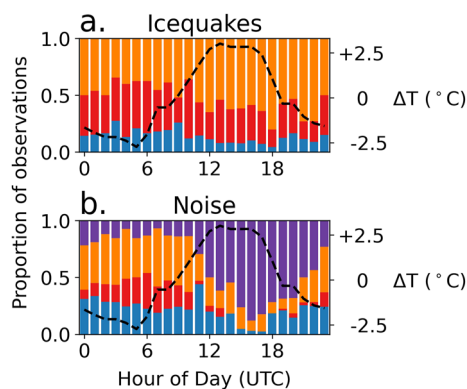
For the noise data, the blue (1st) clusters mostly take place on or after 9 July, 2 days after lake drainage initiates. The red (2nd) cluster noise records occur primarily between 5 July (2 days before drainage begins) and 17 July (2 days after drainage ends). The orange (3rd), and purple (4th) clusters include noise records occurring up until 5 July. The purple cluster pauses on July 4 and resumes on 17 July. Similarly, the orange cluster almost completely pauses on 5 July before resuming activity on the 17th.

5.4. Diurnal Variability of Clusters

Because the hourly sampling distributions of the icequakes and noise are strongly biased (see Figures 2d and 2e), we examine in Figure 8 the proportion of samples from each cluster for each hour of the day throughout the season. The mean difference from the daily mean temperature (black dashed line) is provided to indicate when peak meltwater production approximately occurs. For the icequakes, the hourly distributions are largely uniform throughout the day (Figure 8a). The noise cluster hourly distributions,
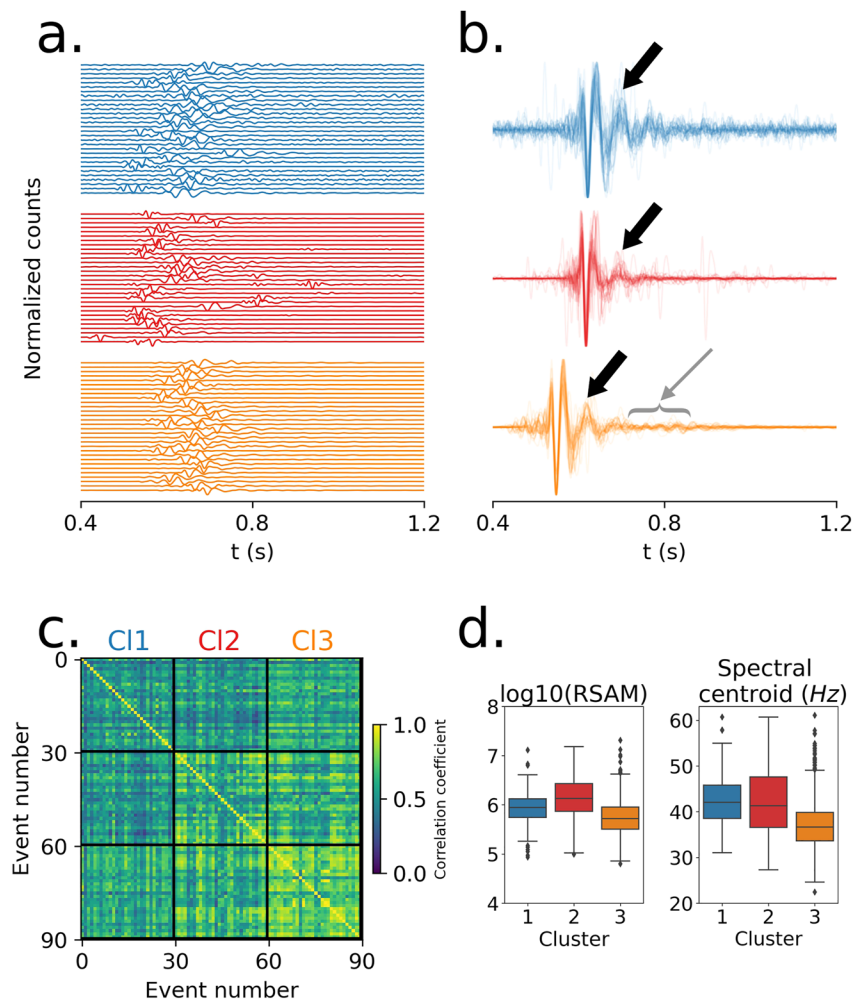

Figure 8. Proportion of (a) icequakes and (b) noise observations that occur during each hour of the day. Local time is UTC + 2. Black dashed line shows the seasonal mean hourly deviation from daily mean temperature.

Figure 9. Waveform characteristics of icequake clustering results. (a) Top 30 representative waveforms filtered from 15 to 80 Hz and normalized by maximum amplitude, zoomed in to the main impulse in the two-second waveform. (b) Same as (a), but aligned to maximize cross-correlation coefficients and stacked. (c) Cross-correlation matrix on entire two-second waveforms for the 20 representative icequakes in each cluster. (d) Box-and-whisker plots showing log(RSAM) and spectral centroids for all icequakes by cluster. RSAM (real-time seismic amplitude measurement) is the time-summed, non-normalized, filtered waveform amplitude, and the spectral centroid is the weighted mean of frequencies in the waveform spectrum.

however, differ greatly from the overall sampling distribution, evident by the colored bands of variable thickness in Figure 8b. The purple (4th) cluster occurs more frequently than the other clusters in the afternoon and early evening hours (12:00 to 21:00 UTC), with the majority of events occurring around 16:00 UTC. The majority of the orange (3rd) cluster's records occur in the late evening and early morning hours (23:00 to 10:00 UTC). Most of the red (2nd) cluster's records occur in the early morning hours, between approximately 2:00 and 10:00 UTC. The blue noise cluster's distributions become increasingly depleted in the afternoon hours (15:00 to 18:00), but are otherwise constant through time.

5.5. Icequake Cluster Waveform and Spectral Characteristics

Figure 9a shows waveforms for the 30 representative icequakes in each cluster. For visualization of waveform similarity (Figure 9b), we normalize the waveforms by their maximum amplitude, shift them by the time lag that maximizes their cross-correlation coefficient, then plot them on top of each other within a 0.8-s window. This approach allows us to see the subtle features that characterize the waveforms in each cluster. For example, the third cycle in each stack (indicated by the thick black arrows in Figure 9b) has a distinct shape for each cluster.

This arrival, which we refer to as the primary coda, is relatively high-amplitude in the blue (1st) cluster compared to the other two clusters. Another feature is the high degree of similarity in the extended codas for the orange cluster (gray bracketed region with thin gray arrow, Figure 9b), a feature that is lacking in the other two clusters. Additionally, the relative amplitudes of the first two peaks (the first peak being the Rayleigh wave arrival) vary between clusters. For the blue and orange clusters, the second peak tends to be higher than the first, whereas for the red cluster, the first peak appears generally higher than the second peak. Another notable feature is that the first icequake cluster (blue) has a visibly lower signal-to-noise ratio than the other two clusters due to the presence of a constant background noise.

Figure 9c shows the full-waveform cross-correlation (CC) matrix for the 90 icequakes (30 representative icequakes each for three clusters) sorted by cluster. Although the cross-correlation coefficients (CC) are generally high amongst all icequakes (CC ~0.8), the third cluster has particularly high CC. The CC for the first cluster are relatively low, despite icequakes being clustered together based on fingerprint similarity. Figure 9d shows box-and-whisker plots characterizing the RSAM and spectral centroid values for all icequakes events by cluster. The boxes enclose the 25th to 75th percentiles of the statistics, and the horizontal line in the box shows the 50th percentile. The whiskers extend to 1.5 times the range between the 25th and 50th percentiles, and individual points beyond the whiskers are outliers. The red (2nd) cluster has the highest median RSAM, followed by the blue (1st) cluster, then orange (3rd). Icequakes in the orange cluster have a substantially lower spectral centroids than the other two icequake clusters. The blue and red icequake clusters have median spectral centroids at ~42 Hz, whereas the orange cluster median is at ~36 Hz.

5.6. Noise Cluster Waveform and Spectral Characteristics

Identical to our procedure with the icequakes, we attempted to align the representative noise waveforms in the time domain to highlight coherent details. Cross-correlation, however, resulted in coefficients generally close to 0, indicating that the noise waveforms are not similar in the time domain at the minute-long time scale. Patterns do emerge, however, in the spectral domain. We calculate the clusters' spectra by summing the STFT windows (i.e., the columns of the spectrograms) over time, thus preserving the filtering and median-normalization steps applied prior to SpecUFEx (see Section 4.1, "Spectrogram Generation"). Figure 10a shows the 30 representative spectra for each cluster, emphasizing differences in amplitude and overall shapes of the spectra between clusters. Spectra in the red (2nd) cluster, for example, retain more energy above 20 Hz compared to those in the other three clusters. The orange (3rd) and purple (4th) clusters have similar shapes, but between 10 and 30 Hz, the purple cluster has more energy than the orange cluster. Above 30 Hz, the blue (1st) cluster has higher energy than the purple and orange clusters, but not the red cluster.

Figures 10b and 10c show the RSAM and spectral centroid values, respectively, colored by cluster for each noise sample throughout the study period. Yellow vertical lines mark the approximate hottest time of day, and the numbered vertical dashed lines are the same as above: (a) is the beginning of supraglacial drainage, (b) is when sub- and englacial drainage begins, and (c) is the end of drainage. Prior to drainage, these two statistics oscillate daily, with higher RSAM values and lower spectral centroids later in the day, and lower RSAM values and higher spectral centroids in the early morning hours. The orange (3rd) and purple (4th) clusters capture this diurnal pattern, with orange clusters containing the early morning noise and purple clusters containing the afternoon noise. This diurnal oscillating pattern is interrupted on 5 July, when the red (2nd) cluster appears amid a sudden increase in RSAM values and decrease in spectral centroid values. Another, weaker oscillating pattern is set up between the blue (1st) and red clusters, before returning on 15 July to a mode similar to the original oscillating pattern between the orange and purple clusters, except the blue cluster is now partly taking the place of the orange cluster.

Figure 11 plots spectral centroid against RSAM colored by (a) cluster, (b) day of year and (c) time of day. In the early season, the daily anti-correlated nature of the spectral centroid and RSAM values is particularly apparent, with higher spectral-centroid values generally occurring in the earlier times of day (00:00-12:00 UTC). Black arrows in Figure 11a describe a theoretical progression of events: (1) a stable, diurnal pattern of noise alternating between orange and purple clusters prior to drainage, (2) the transition to the red cluster, and to sub- and englacial component of drainage (3) the return to a post-drainage modified diurnal pattern of purple and orange clusters. The dashed line is a proposed boundary in log(RSAM)-spectral-centroid space between a steady state of glacier
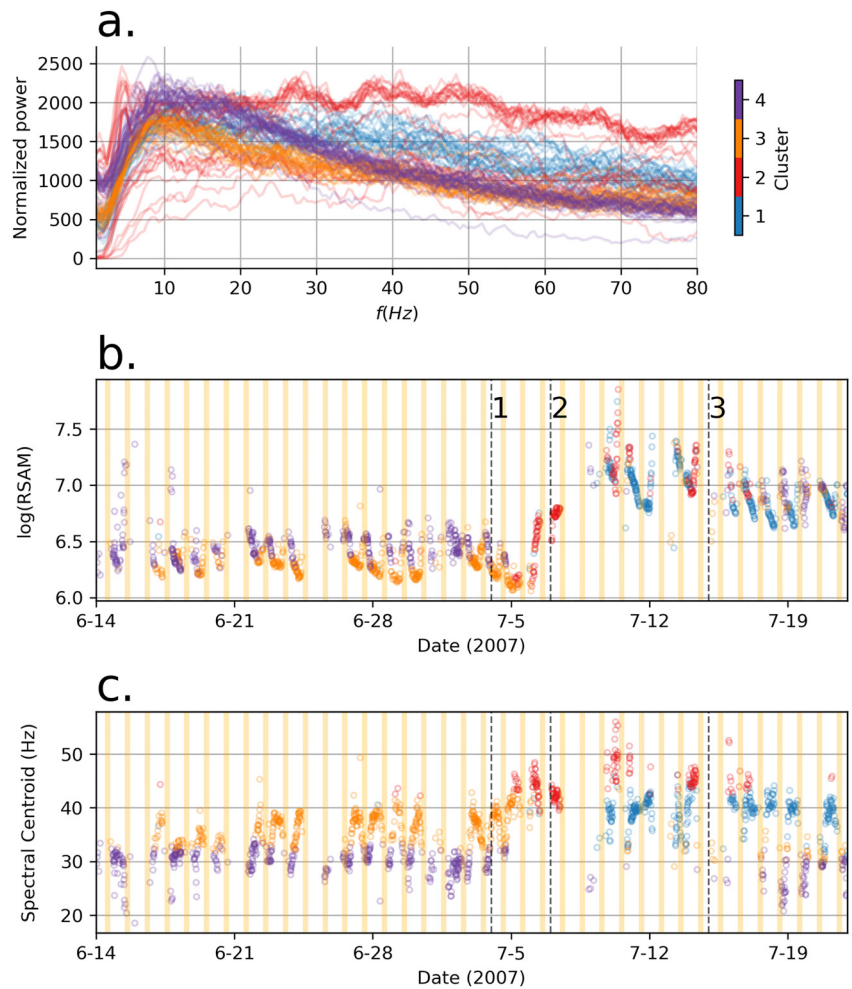
Figure 10. Spectral and waveform characteristics of noise clustering results. (a) Power spectral density (PSD) of the 30 representative noise samples, calculated by summing spectrogram amplitudes across all frequencies over time and normalizing by the median. Lower panels show the (b) log10(RSAM) and (c) spectral centroids (described in text) of all noise samples colored by cluster, with yellow vertical bars indicating afternoon hours (∼16:00 local time). Numbered, vertical dashed lines are the same as in Figures 2 and 7.

hydraulics driven by meltwater input oscillating on a daily timescale, and a dynamic state controlled by the catastrophic flooding event.

5.7. Representative Spectrograms and Fingerprints

Figure 12 shows the five most representative spectrograms from each cluster for (a) icequakes and (b) noise along with their corresponding fingerprints. For visualization, each spectrogram and fingerprint is normalized to the highest value in each matrix, respectively. In Figure 12a, we see a continuous background noise in the blue icequake cluster (Cl1) that is not apparent in the red and orange clusters (Cl2 and Cl3, respectively). Close inspection of Cl2 and Cl3 reveal differences, for example, the extended decay of Cl3 waveforms, and the higher-frequency energy in the impulses of Cl2. The impulses in Cl1 look relatively weaker and of shorter duration than in the other two clusters.

The spectrograms of the noise clusters in Figure 12b are generally of constant amplitude and frequency content, but at times punctuated by thin bright vertical lines, presumably very small icequakes. These small icequakes appear randomly through time and do not appear to drive clustering results. Clusters 1 (blue cluster) and 2 (red cluster) have a more broadband signal than the other two clusters, and Cluster 2 may even exhibit some harmonics, although the timing of these events does not coincide with previous reports of gliding tremor and
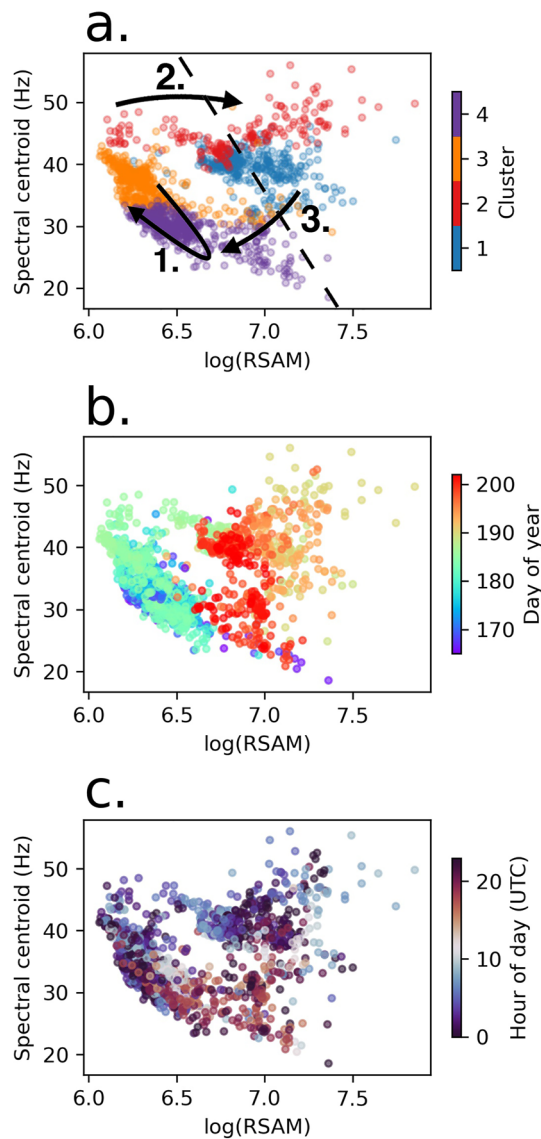
Figure 11. Hysteresis plots showing evolution of noise waveform and spectral characteristics colored by (a) cluster, (b) day of year, and (c) hour of day. Horizontal axes show log10(RSAM), a measure of digital counts per second, and vertical axes show spectral centroid, defined as the amplitude-normalized mean of the waveform spectrum (described in text). In panel (a), black arrows describe (1) diurnal pattern of noise between orange and purple clusters prior to drainage, (2) the transition to the red cluster, and to the sub-glacial component of drainage (3) the return to a modified diurnal pattern and the purple and orange clusters. The dashed line is a proposed boundary between a steady state of glacier hydraulics driven by diurnal melt flow, and a dynamic state controlled by the flooding event.

other harmonic signals at Gornergletscher (Heeszel et al., 2014; Umlauft et al., 2021). The noise spectrograms in our data set hardly vary in terms of amplitude and frequency on the minute-long time scale, and therefore are represented in the HMM state transition matrix by one state repeating in time. In the fingerprints, this is represented by one bright element highlighted on the main diagonal, since the diagonal elements in the fingerprints represent one state transitioning to itself. In the noise fingerprints shown in Figure 12b, clusters 1, 2, 3, and 4 have elements 4, 3, 2, 1 highlighted on the main diagonal, respectively.

6. Discussion

6.1. Cluster-Based Determination of the Onset of Subglacial Drainage

Werder et al. (2009) identified the onset of subglacial drainage as the time when the lake level begins to decrease from its highest level, indicated by the second vertical dashed line in Figures 2, 10b, and 10c. Sub- and englacial drainage, however, likely initiated earlier, since lake level reflects a balance between water input to the lake and water export by drainage, and variable meltwater input might delay by minutes to hours the visible drop in lake level from increased subglacial drainage. Given that about 99% of noise windows in the red (2nd) noise cluster begin on or after 5 July—2 days before the reported subglacial flooding– we suggest that the red noise cluster contains records that include the true initiation of the flood's englacial and subglacial drainage components. The red noise cluster coincides with the onset, though not the end, of the glacial surge inferred from GPS data (Figure 7), so we cannot definitively rule out that the source of the spectral features controlling the clustering could alternatively be due to till deformation, slip at the ice-bed interface, or minute fractures occurring in the ice caused by increased ice flow and deformation. Yet, since the red noise cluster continues beyond the end of the glacier acceleration, we suggest that its noise is more likely the result of hydraulic excitations driven by turbulent water flow.

Figure 10a shows how the normalized amplitude of noise in the red (2nd) noise cluster is elevated in the higher frequencies (>25 Hz) compared to the other clusters, and Figure 10b shows how its RSAM values, that is, the energy in its waveforms, are higher as well. This combination of high-amplitude, broadband signal is consistent with studies relating velocity of turbulent flow to spectral power (Gimbert et al., 2014; Nanni et al., 2021; Tsai & Rice, 2010). Additionally, Lindner (2020) uses plane-wave beam-forming to reveal how the source of the seismic noise shifts suddenly to the lake front on 8 July as lake drainage increases. Given the timing and the broad-spectra, high-energy noise samples in the red cluster, we infer that seismicity in that cluster is generated by turbulent water flow as the englacial and subglacial components of the flooding initiates.

6.2. Diurnal Cluster Trends and the Reorganization of Subglacial Water Conduits

Hourly cluster trends are of interest because of the known diurnally varying conditions at Gornergletscher, including ambient temperature, meltwater production, and glacial velocity; all of which directly or indirectly influence glacial seismicity. The noise clusters discovered here exhibit strong diurnal variations that are different than the distribution of noise samples as a whole (Figure 8b). Noise samples in the purple (4th) cluster, for example, peak at 16:00 UTC, lagging about 3 hr after the hottest time of day, whereas noise records in the orange cluster preferentially occur during the cooler
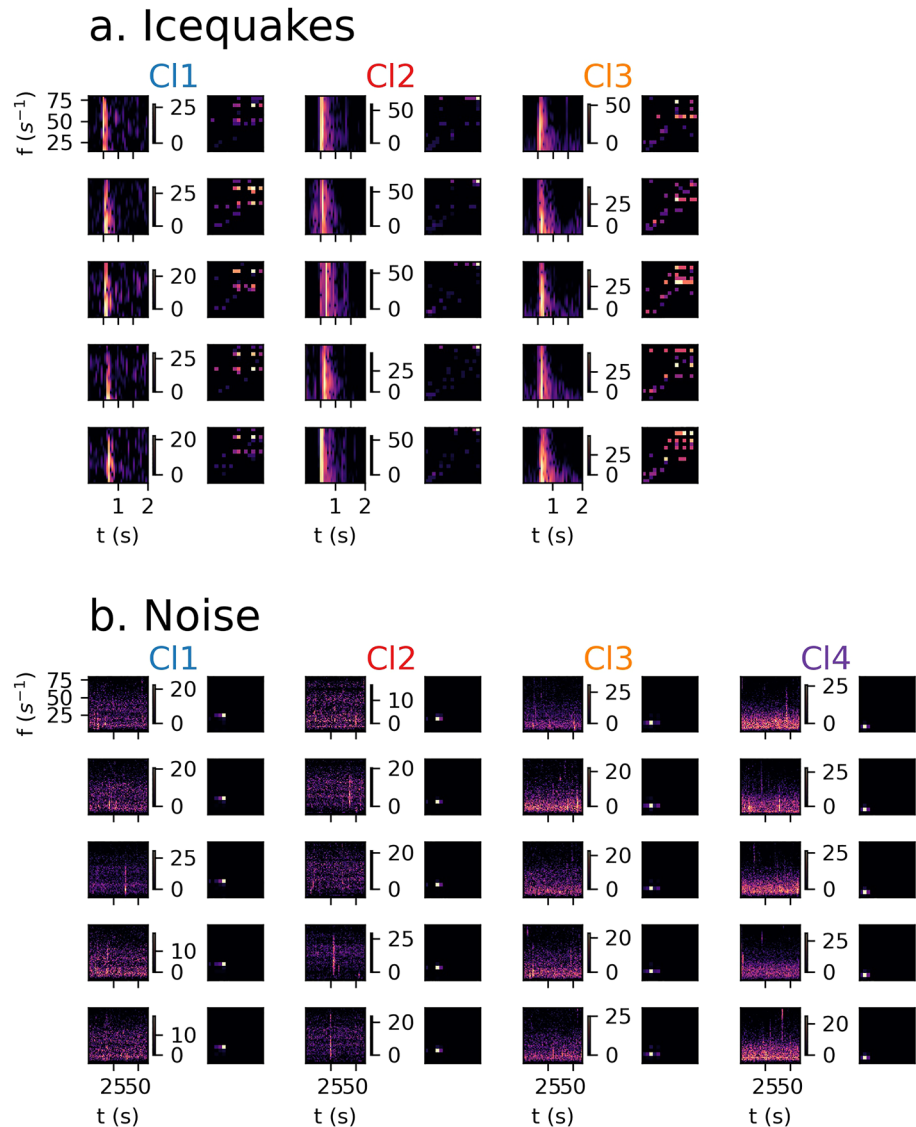
a. Icequakes

b. Noise

Figure 12. The five spectrograms with the highest silhouette scores from each cluster for (a) icequakes and (b) noise. Color bars show power spectral density normalized by its median and converted to decibels.

times of the day. The temporal trends of the orange and purple clusters appear to coincide with diurnal variations of sources of meltwater input, and so we suggest that these two noise clusters contain seismic signals of two types of meltwater flow.

Beyond simply a temporal correlation, Figure 10b shows that waveforms from the purple (4th) noise cluster have significantly higher amplitudes than those from the orange (3rd) cluster, consistent with numerous studies reporting observations of higher energy seismic signals arising from increased volumes of subglacial water flow (e.g., Bartholomaus et al., 2015; Eibl et al., 2020; Vore et al., 2019). Additionally, noise samples in the orange (3rd) cluster have higher spectral centroids than those in the purple cluster (Figure 10c), consistent with a physical model describing seismic resonance of fluids flowing through narrower or impeded conduits (e.g., Röösli et al., 2014), or with a tremor source located more closely to the station (since the attenuation of higher frequencies increases with greater station-event distance). Thus, the purple cluster may contain noise windows with the seismic signature of unimpeded, high-volume afternoon water-flow through sub- and englacial channels. Conversely, the orange cluster could have noise windows with impeded water-flow through a distributed, poorly connected subglacial network, perhaps constricted by basal ice creeping during the nighttime hours. This

negative correlation between spectral centroids and RSAM measurements in the pre- and post-drainage periods is indicated by the black arrow labeled "1" in Figure 11a.

The subglacial conduit networks would likely be deformed or destroyed during the glacial lake outburst flood as distributed drainage modes switch to channelized flow under large influxes of water, as indicated by modeling of this particular drainage event by Werder et al. (2013) and illustrated with the arrow labeled "2" in Figure 11a. Post-drainage, however, the channels could reform, eventually returning to a similar style of subglacial hydrology signified by the return of orange and purple clusters, indicated in Figure 11a by the arrow labeled "3". The blue (1st) noise cluster also has a stable, diurnal mode post-drainage, yet this cluster does not appear much prior to drainage.

The diurnally oscillating pattern could conversely be related to glacial sliding, which is also modulated by daily meltwater flow. In Figure 10a, however, we see how the majority of the spectral energy in the purple and orange clusters occurs in the 5–40 Hz band; a signal which has in previous studies been linked to sub- and englacial hydrology due to a combination of turbulent water flow and sediment saltation during transport (Bartholomaus et al., 2015; Eibl et al., 2020; Gimbert et al., 2016; Tsai & Rice, 2010; Vore et al., 2019). Additionally, a beam-forming analysis discovered diurnally alternating sources of meltwater input within continuous seismic data in the weeks leading up to the lake drainage, indicating that the dominant seismic noise originates from the west during the daytime and the northeast in the late evening/morning hours (Lindner, 2020). Future studies, particularly those focused on physical modeling, might aim to better understand which aspects of the sources and paths are contributing to the subtle differences between the spectra in Figure 10a and how they relate to the hydrology and dynamics at Gornergletscher.

6.3. Icequake Clusters and Hydrology

To understand the clusters' connection to glacial behavior, we compare the timing of the observations in each cluster to ancillary geophysical data, including ambient temperature, on-ice GPS displacement, precipitation measurements, and lake-level data from an immersible pressure transducer. The acceleration of ice flow shown by the GPS data and the reported onset of the outburst flood (4 July; earliest dashed fuchsia lines in Figure 7) both approximately coincide with the blue icequake cluster, which also continues after the ice-flow acceleration has ceased. Given the blue cluster's timing with the glacial lake outburst flood and background noise present in its records, we assume that this cluster is related to the lake drainage. In addition to the background noise, the primary coda (black arrow, 9b) of the blue (1st) cluster has a slightly different shape than the other two clusters, indicating that a systematic variation in source and/or path may exist during the time of increased lake drainage. There is no apparent season-wide or otherwise hydraulically induced trend within the red (2nd) and orange (3rd) icequake clusters. Instead, given their locations (Figure 6a), these clusters appear related to the opening of crevasses.

6.4. Identifying Similar Icequake Waveforms

The 30 representative icequake waveforms in the red (2nd) and orange (3rd) clusters are similar in terms of waveform cross-correlation (CC) (see Figure 9c), indicating that in this case, similar fingerprints correspond to similar waveforms. Most traditional methods of finding similar earthquakes rely on pair-wise similarity metrics, most commonly CC coefficients. CC coefficients, however, are sensitive to parameters such as station-event azimuth, wave-propagation direction, and low signal-to-noise ratio, and can therefore systematically reject what are in reality, similar events (e.g., Gao et al., 2021). Based on the range of azimuths of the orange representative icequakes in Figure 6c, it seems that SpecUFEx is insensitive to variable azimuths in this case. CC analysis can also fail when excessive random noise is superimposed on one or both waveforms in the pair. Such noise, however, does not appear to hinder clustering of the blue (1st) cluster, although it does lower its CC coefficients as expected (Figure 9c).

In addition to a lower signal-to-noise ratio, the blue cluster's icequakes also have larger primary codas (thick black arrow, Figure 9b) relative to the first two phase arrivals, compared to the orange and red clusters. The varying shapes of the primary codas could be reflections from, for example, the ice surface, ice-bed interface, or crevasses, indicating changing material properties of one or more of these interfaces. Further analysis must be

undertaken to explore what physical aspects of the icequakes' rupture characteristics, paths, or a combination of the two are leading to the waveforms observed in the icequake clustering results.

SpecUFEx could be a useful tool for detecting repeating sequences of tectonic earthquakes. Repeating earthquake sequences, or "REQS", are quasi-periodically occurring earthquakes located on a single rupture patch that generate near-identical waveforms, and are frequently used to infer tectonic loading rates and gain insight on variable source and path characteristics (see review by Uchida and Bürgmann [2019]). Methods for detecting REQS typically employ CC coefficients, and therefore suffer from the shortcomings described above. Comparing similar SpecUFEx fingerprints, as opposed to waveforms, could help complement existing REQS catalogs and to discover new repeating earthquake sequences as well. Although not a tectonic example, icequakes in the red cluster have highly correlated waveforms that occur over a period of weeks in a tight, geographic cluster around a lake marginal crevasse, and so could be a possible candidate for a "repeating icequake sequence".

6.5. Comparing Representative Events of Icequakes and Noise

Figure 12 shows the five spectrograms and fingerprints with the highest silhouette scores for each cluster for (a) icequakes and (b) noise samples. Visualizing the fingerprints for the five representative events allows us to qualitatively compare the differences between events and noise. We see that icequake fingerprints are more complicated than the noise fingerprints, which is to be expected since the icequakes, with their multiple phase arrivals, vary more through time than the noise samples, which, within a 60-s window, are essentially invariant. Fingerprints show the state transitions from the HMM's state-sequence matrix (Figure 4c), and since the noise state-sequence matrices do not vary much through time, they are represented in a fingerprint by one (or two) states on the main diagonal.

An important result of the icequake clustering is how the blue (1st) cluster is permeated with a noise from ∼25–60 Hz associated with the onset of lake drainage (see Figure 12a). This indicates that in this case, clustering is based in part on the background noise as well as the signal itself. Icequakes in other temperate glacial settings may also occur in the presence of hydraulic noise, so careful attention must be paid when interpreting results as to whether they are due to the signal of interest, or some other systematic "noise." By analyzing the noise alongside the signal, our approach allows us to ascertain the impacts of noise on icequake clustering results, which, in this case, are substantial.

6.6. Clustering Choices

We choose to employ k-means clustering on the principal components of the fingerprints as an initial step toward unsupervised classification of glacial seismic data, although there are many other, more sophisticated clustering and dimensionality reduction methods we could have employed. One simplification of k-means is that it defines hard boundaries between clusters, meaning that one point can only be assigned to one cluster even if it is on the boundary between two or more clusters. A number of points in our PCA manifolds lie at the boundaries of clusters (Figure 5), however, in many cases we ignore these points by focusing on the "representative" events that are closer to the ends of the clusters. A clustering algorithm such as Gaussian mixture model (e.g., Murphy, 2022) or fuzzy clustering (Dunn, 1973) could possibly better characterize gradual changes between clusters by allowing the events at the boundaries of the clusters to be mixtures of multiple clusters. The disadvantage of this technique is that certain subsequent interpretive steps that we take, such as comparing cluster evolution through time, becomes more difficult when data are not assigned to a single cluster.

In this study, we perform PCA prior to *k*-means clustering to aid in visualization and interpretation of results. We did, however, also try clustering the raw fingerprints, and found similar results as clustering on their principal components. This was to be expected, given that we analyze enough principal components to so that at least 86% of the variance of the raw fingerprint data set is preserved. Other than PCA, we experimented with one additional dimensionality reduction technique, t-distributed stochastic neighbor embedding, or t-SNE (Hinton & Roweis, 2002), prior to k-means clustering. When comparing clustering results between PCA and t-SNE, we found that 81% of icequake records were clustered similarly between the two methods, and 98% of noise records were clustered similarly. These three test cases illustrate that our clustering choices appear largely independent of choice of dimensionality reduction.

6.7. Applicability to Other Glacial (or Non-Glacial) Settings

Although in this work we focus on a particularly well-studied glacier, Gornergletscher, here we discuss the applicability of transferring this method to other glaciers, ice sheets, or other seismically active settings. UML is an empirical process, and interpretation of results is highly dependent on the availability (and reliability) of ancillary data. Our analysis benefits from a wealth of geophysical data from Gornergletscher (e.g., temperature, lake level, locations), but we acknowledge that applicability to other settings may be limited by the data available there. In addition to ancillary data, all studies using SpecUFEx benefit greatly from a priori knowledge of the timescales of interest at the system, in part because the input spectrograms must be of the same lengths. At Gornergletscher, for example, we know that surface icequakes have a duration of about 1 s (e.g., Roux et al., 2010; Walter, 2009), so 2-s long record windows are appropriate for analyzing that signal. A longer signal (5 s) may be dominated not by the icequake, but rather by background noise, which, as we demonstrate in this study, could affect clustering results. Other glacial (and non-glacial) settings may have more complicated behavior, for example, icequakes of varying duration, or short-duration tremor episodes. Trial and error may be necessary if prior knowledge of the timescales of interest is not available.

The data processing steps taken prior to SpecUFEx are crucial, and depend on the quality of data, the geologic setting, and the research goals. For example, one of the aims of this study is to compare subtle differences between noise samples, therefore, for the noise experiment we tune our STA/LTA filters to remove seismic records with icequakes. At other settings, however, one could choose to not separate the icequakes from the noise beforehand, and instead use SpecUFEx to help characterize the differences between those two classes. Other, rarer events such as basal and intermediate-depth events have also been reported at this glacier (e.g., Walter, 2009; Walter et al., 2010) but were purposely excluded from this work. Waveforms from these events can appear markedly different from their surficial counterparts, both in terms of frequency content and duration, and they typically represent only a small fraction (<1%) of all events. Since we are interested in more subtle differences between icequakes, we opt to omit these rare events. Future studies, however, could approach using UML to distinguish these rare events from the icequake population as a whole.

7. Conclusion

Here, we demonstrate the utility of unsupervised machine learning for exploring and characterizing large cryoseismic data sets that contain a diverse variety of seismic signals. In the case of the Gornergletscher Summer 2007 season, these signals include icequakes from surficial crevasses, seismic noise from diurnal meltwater flow, and a glacial outburst flood. By reducing the spectrograms of these various signals into low-dimensionality fingerprints, we are able to cluster icequakes with similar waveforms, and noise with similar spectral and waveform characteristics. We discover similarities in icequake waveforms and noise spectra that are at times highly subtle and would elude simpler seismic analysis techniques. The icequakes that are "representative" of their cluster, that is, ones with the highest silhouette scores, exhibit strong intra-cluster similarities and systematic differences between clusters, for example, each cluster has a characteristic shape of a late-phase arrival that could be interpreted as being caused by different sources (e.g., crevasse opening or not). For the noise analysis, one noise cluster with particularly high-energy spectra appears 2 days before surficial signals of lake drainage occur, and thus could be considered a seismic precursor to the onset of lake drainage, or could mark the true start of lake drainage itself. Other noise clusters exhibit a diurnal pattern (i.e., one cluster occurs primarily in the morning hours, and another in the evening). These diurnal clusters are interrupted (or perhaps overwhelmed) by the lake drainage, but return after its conclusion, and thus could indicate when a stable mode of diurnal meltwater flow has returned as the most prominent state. Given that glaciers exhibit a wide range of complex seismic signals and that the volume of cryoseismic data is increasing exponentially, glacial seismic data exploration may be improved through methods that incorporate unsupervised machine learning. This work demonstrates the value of spectral unsupervised feature extraction and the SpecUFEx workflow as a seismic monitoring and discovery tool, particularly in systems where a diverse range of fluid-driven processes can occur.

Data Availability Statement

The glacial seismic data set is available from the Swiss Seismological Service (https://doi.org/10.12686/sed/networks/4d) and can be accessed at http://eida.ethz.ch/. SpecUFEx is implemented in MATLAB, and can be downloaded from github.com/tsawi/SawiEtAl_2022.

References

Aster, R. C., & Winberry, J. P. (2017). Glacial seismology. *Reports on Progress in Physics*, *80*(12), 126801. https://doi.org/10.1088/1361-6633/aa8473

Bartholomaus, T. C., Amundson, J. M., Walter, J. I., O'Neel, S., West, M. E., & Larsen, C. F. (2015). Subglacial discharge at tidewater glaciers revealed by seismic tremor. *Geophysical Research Letters*, *42*(15), 6391–6398. https://doi.org/10.1002/2015GL064590

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, *37*(6), 1554–1563. https://doi.org/10.1214/aoms/1177699147

Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*, 6433. https://doi.org/10.1126/science.aau0323

Carniel, R., & Guzmán, S. R. (2020). Machine learning in volcanology: A review. *Updates in Volcanology*. https://doi.org/10.5772/intechopen.94217

Chamarczuk, M., Nishitsuji, Y., Malinowski, M., & Draganov, D. (2019). Unsupervised learning used in automatic detection and classification of ambient-noise recordings from a large-N array. *Seismological Research Letters*, *91*, 1–389. https://doi.org/10.1785/0220190063

Chaput, J., Aster, R. C., McGrath, D., Baker, M., Anthony, R. E., Gerstoft, P., et al. (2018). Near-surface environmentally forced changes in the Ross Ice Shelf observed with ambient seismic noise. *Geophysical Research Letters*, *45*(11), 187–211. https://doi.org/10.1029/2018GL079665

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57. https://doi.org/10.1080/01969727308546046

Eibl, E. P. S., Bean, C. J., Einarsson, B., Pàlsson, F., & Vogfjörd, K. S. (2020). Seismic ground vibrations give advanced early-warning of subglacial floods. *Nature Communications*, *11*, 2504. https://doi.org/10.1038/s41467-020-15744-5

Endo, E. T., & Murray, T. (1991). Real-time seismic amplitude measurement (RSAM): A volcano monitoring and prediction tool. *Bulletin of Volcanology*, *53*(7), 533–545. https://doi.org/10.1007/BF00298154

Gao, D., Kao, H., & Wang, B. (2021). Misconception of waveform similarity in the identification of repeating earthquakes. *Geophysical Research Letters*, *48*, 13. https://doi.org/10.1029/2021GL092815

Garcia, L., Luttrell, K., Kilb, D., & Walter, F. (2019). Joint geodetic and seismic analysis of surface crevassing near a seasonal glacier-dammed lake at Gornergletscher, Switzerland. *Annals of Glaciology*, *60*(79), 1–13. https://doi.org/10.1017/aog.2018.32

Gimbert, F., Tsai, V. C., Amundson, J. M., Bartholomaus, T. C., & Walter, J. I. (2016). Subseasonal changes observed in subglacial channel pressure, size, and sediment transport. *Geophysical Research Letters*, *43*(8), 3786–3794. https://doi.org/10.1002/2016GL068337

Gimbert, F., Tsai, V. C., & Lamb, M. P. (2014). A physical model for seismic noise generation by turbulent flow in rivers. *Journal of Geophysical Research: Earth Surface*, *119*(10), 2209–2238. https://doi.org/10.1002/2014JF003201

Heeszel, D. S., Walter, F., & Kilb, D. L. (2014). Humming glaciers. *Geology*, *42*(12), 1099–1102. https://doi.org/10.1130/G35994.1

Hinton, G. E., & Roweis, S. (2002). Stochastic neighbor embedding. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15). MIT Press. Retrieved from https://proceedings.neurips.cc/paper/2002/file/6150ccc6069bea-6b5716254057a194ef-Paper.pdf

Holtzman, B. K., Paté, A., Paisley, J., Waldhauser, F., & Repetto, D. (2018). Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field. *Science Advances*, *4*(5), eaao2929. https://doi.org/10.1126/sciadv.aao2929

Huss, M., Bauder, A., Werder, M., Funk, M., & Hock, R. (2007). Glacier-dammed lake outburst events of Gornersee, Switzerland. *Journal of Glaciology*, *53*(181), 189–200. https://doi.org/10.3189/172756507782202784

Iken, A., Fabri, K., & Funk, M. (1996). Water storage and subglacial drainage conditions inferred from borehole measurements on Gornergletscher, Valais, Switzerland. *Journal of Glaciology*, *42*(141), 233–248. https://doi.org/10.3189/S0022143000004093

IPCC. (2022). *The Ocean and Cryosphere in a Changing Climate: Special Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/9781009157964

Jenkins, W. F., Gerstoft, P., Bianco, M. J., & Bromirski, P. D. (2021). Unsupervised deep clustering of seismic data: Monitoring the Ross ice shelf, Antarctica. *Journal of Geophysical Research: Solid Earth*, *126*(9), e2021JB021716. https://doi.org/10.1029/2021JB021716

Karl Pearson, F. R. S. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Klapuri, A., & Davy, M. E. (2007). *Signal processing methods for music transcription, chapter 5*. Science & Business Media: Springer.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2019). Machine learning in seismology: Turning data into insights. *Seismological Research Letters*, *90*(1), 3–14. https://doi.org/10.1785/0220180259

Lamb, O., Lees, J., Marin, L. F., Lazo, J., Rivera, A., Shore, M., & Lee, S. (2020). Investigating potential icequakes at Llaima volcano, Chile. *Volcanica*, *3*(1), 29–42. https://doi.org/10.30909/vol.03.01.2942

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Lee, W. H. K., & Stewart, S. W. (1981). *Principles and application of microearthquake networks. Advances in Geophysics* (Vol. 2). Academic Press.

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of significance: Principal component analysis. *Nature Methods*, *14*(7), 641–642. https://doi.org/10.1038/nmeth.4346

Lindner, F. (2020). *Englacial structure and the subglacial drainage system: Seismological studies* (Doctoral dissertation). ETH Zurich. https://doi.org/10.3929/ETHZ-B-000417199

Lindner, F., Walter, F., Laske, G., & Gimbert, F. (2020). Glaciohydraulic seismic tremors on an Alpine glacier. *The Cryosphere*, *14*(1), 287–308. https://doi.org/10.5194/tc-14-287-2020

Malfante, M., Dalla Mura, M., Métaxian, J. P., Mars, J. I., Macedo, O., & Inza, A. (2018). Machine learning for volcano-seismic signals: Challenges and perspectives. *IEEE Signal Processing Magazine*, *35*(2), 20–30. https://doi.org/10.1109/MSP.2017.2779166

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*, 3952. https://doi.org/10.1038/s41467-020-17591-w

Murphy, K. P. (2022). *Probabilistic machine learning*. MIT Press.

Nanni, U., Gimbert, F., Roux, P., & Lecointre, A. (2021). Observing the subglacial hydrology network and its dynamics with a dense seismic array. *Proceedings of the National Academy of Sciences*, *118*(28), e2023757118. https://doi.org/10.1073/pnas.2023757118

Park, Y., Mousavi, S. M., Zhu, W., Ellsworth, W. L., & Beroza, G. C. (2020). Machine-learning-based analysis of the Guy-Greenbrier, Arkansas earthquakes: A tale of two sequences. *Geophysical Research Letters*, *47*(6), e2020GL087032. https://doi.org/10.1029/2020GL087032

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Podolskiy, E. A., Murai, Y., Kanna, N., & Sugiyama, S. (2021). Ocean-bottom and surface seismometers reveal continuous glacial tremor and slip. *Nature Communications*, *12*(1), 3929. https://doi.org/10.1038/s41467-021-24142-4

Podolskiy, E. A., & Walter, F. (2016). Cryoseismology. *Reviews of Geophysics*, *54*(4), 708–758. https://doi.org/10.1002/2016RG000526

Ren, C. X., Peltier, A., Ferrazzini, V., Rouet-Leduc, B., Johnson, P. A., & Brenguier, F. (2020). Machine learning reveals the seismic signature of eruptive behavior at Piton de la Fournaise volcano. *Geophysical Research Letters*, *47*(3), 1–11. https://doi.org/10.1029/2019GL085523

Riesen, P., Sugiyama, S., & Funk, M. (2010). The influence of the presence and drainage of an ice-marginal lake on the flow of Gornergletscher, Switzerland. *Journal of Glaciology*, *56*(196), 278–286. https://doi.org/10.3189/002214310791968575

Röösli, C., Walter, F., Husen, S., Andrews, L. C., Lüthi, M. P., Catania, G. A., & Kissling, E. (2014). Sustained seismic tremors and icequakes detected in the ablation zone of the Greenland ice sheet. *Journal of Glaciology*, *60*(221), 563–575. https://doi.org/10.3189/2014joG13j210

Ross, Z. E., Meier, M. A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, *108*(5), 2894–2901. https://doi.org/10.1785/0120180080

Ross, Z. E., Yue, Y., Meier, M. A., Hauksson, E., & Heaton, T. H. (2019). Phaselink: A deep learning approach to seismic phase association. *Journal of Geophysical Research: Solid Earth*, *124*(1), 856–869. https://doi.org/10.1029/2018JB016674

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Roux, P. F., Walter, F., Riesen, P., Sugiyama, S., & Funk, M. (2010). Observation of surface seismic activity changes of an Alpine glacier during a glacier-dammed lake outburst. *Journal of Geophysical Research*, *115*(3), 1–13. https://doi.org/10.1029/2009JF001521

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. D., Campillo, M., & Baraniuk, R. (2020). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature Communications*, *11*(1), 3972. https://doi.org/10.1038/s41467-020-17841-x

Sick, B., Guggenmos, M., & Joswig, M. (2015). Chances and limits of single-station seismic event clustering by unsupervised pattern recognition. *Geophysical Journal International*, *201*(3), 1801–1813. https://doi.org/10.1093/gji/ggv126

Steinmann, R., Seydoux, L., Beauce, E., & Campillo, M. (2021). Hierarchical exploration of continuous seismograms with unsupervised learning.

Sugiyama, S., Bauder, A., Huss, M., Riesen, P., & Funk, M. (2008). Triggering and drainage mechanisms of the 2004 glacier-dammed lake outburst in Gornergletscher, Switzerland. *Geophys. Res*, *113*(F4), F04019. https://doi.org/10.1029/2007JF000920

Sugiyama, S., Bauder, A., Riesen, P., & Funk, M. (2010). Surface ice motion deviating toward the margins during speed-up events at Gornergletscher, Switzerland. *Journal of Geophysical Research*, *115*(3), F03010. https://doi.org/10.1029/2009JF001509

Sugiyama, S., Bauder, A., Weiss, P., & Funk, M. (2007). Reversal of ice motion during the outburst of a glacier-dammed lake on Gornergletscher, Switzerland. *Journal of Glaciology*, *53*(181), 172–180. https://doi.org/10.3189/172756507782202847

Trugman, D. T., & Shearer, P. M. (2017). GrowClust: A hierarchical clustering algorithm for relative earthquake relocation, with application to the Spanish Springs and Sheldon, Nevada, earthquake sequences. *Seismological Research Letters*, *88*(2), 379–391. https://doi.org/10.1785/0220160188

Tsai, V. C., & Rice, J. R. (2010). *A model for turbulent hydraulic fracture and application to crack propagation at glacier beds* (Vol. 115, No. F3). American Geophysical Union (AGU). https://doi.org/10.1029/2009jf001474

Uchida, N., & Bürgmann, R. (2019). Repeating earthquakes. *Annual Review of Earth and Planetary Sciences*, *47*(1), 305–332. https://doi.org/10.1146/annurev-earth-053018-060119

Umlauft, J., Lindner, F., Roux, P., Mikesell, T. D., Haney, M. M., Korn, M., & Walter, F. T. (2021). Stick-slip tremor beneath an alpine glacier. *Geophysical Research Letters*, *48*(2), e2020GL090528. https://doi.org/10.1029/2020GL090528

Vore, M. E., Bartholomaus, T. C., Winberry, J. P., Walter, J. I., & Amundson, J. M. (2019). Seismic tremor reveals spatial organization and temporal changes of subglacial water system. *Journal of Geophysical Research: Earth Surface*, *124*(2), 427–446. https://doi.org/10.1029/2018JF004819

Walter, F. (2009). *Seismic activity on Gornergletscher during Gornersee outburst floods* (Dissertation). (Vol. 31, No. 1–2, pp. 14–27). ETH. https://doi.org/10.1007/BF03322148

Walter, F., Deichmann, N., & Funk, M. (2008). Basal icequakes during changing subglacial water pressures beneath Gornergletscher, Switzerland. *Journal of Glaciology*, *54*(186), 511–521. https://doi.org/10.3189/002214308785837110

Walter, F., Dreger, D. S., Clinton, J. F., Deichmann, N., & Funk, M. (2010). Evidence for near-horizontal tensile faulting at the base of Gornergletscher, a Swiss alpine glacier. *Bulletin of the Seismological Society of America*, *100*(2), 458–472. https://doi.org/10.1785/0120090083

Werder, M. A., Hewitt, I. J., Schoof, C. G., & Flowers, G. E. (2013). Modeling channelized and distributed subglacial drainage in two dimensions. *Journal of Geophysical Research: Earth Surface*, *118*(4), 2140–2158. https://doi.org/10.1002/jgrf.20146

Werder, M. A., Loye, A., & Funk, M. (2009). Dye tracing a jökulhlaup: I. Subglacial water transit speed and water-storage mechanism. *Journal of Glaciology*, *55*(193), 889–898. https://doi.org/10.3189/002214309790152447

Werder, M. A., Schuler, T. V., & Funk, M. (2010). Short term variations of tracer transit speed on alpine glaciers. *The Cryosphere*, *4*(3), 381–396. https://doi.org/10.5194/tc-4-381-2010

Yoon, C. E., O'Reilly, O., Bergen, K. J., & Beroza, G. C. (2015). Earthquake detection through computationally efficient similarity search. *Science Advances*, *1*(11), e1501057. https://doi.org/10.1126/sciadv.1501057

Zhu, W., & Beroza, G. C. (2019). PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, *216*(1), 261–273. https://doi.org/10.1093/gji/ggy423