Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning

Léonard Seydoux, Randall Balestriero, Piero Poli, Maarten De Hoop, Michel Campillo, Richard Baraniuk

Accelerating seismicity before the 2017 Nuugaatsiaq landslide revealed with unsupervised deep learning

L. Seydoux[1], R. Balestriero[2], P. Poli[1], M. de Hoop[3], M. Campillo[1], and R. Baraniuk[2]

[1]Institut des sciences de la Terre, Université Grenoble-Alpes, UMR CNRS 5375, France

[2]Electrical & Computational Engineering, Rice University, Houston, TX, 77005, USA

[3]Computational & Applied Mathematics, Rice University, Houston, TX, 77005, USA

May 19, 2020

Abstract

The continuously growing amount of seismic data collected worldwide is outpacing our abilities for analysis, since to date, such datasets have been analyzed in a human-expert-intensive, supervised fashion. Moreover, analyses that are conducted can be strongly biased by the standard models employed by seismologists. In response to both of these challenges, we develop a new unsupervised machine learning framework for detecting and clustering seismic signals in continuous seismic records. Our approach combines a deep scattering network and a Gaussian mixture model to cluster seismic signal segments and detect novel structures. To illustrate the power of the framework, we analyze seismic data acquired during the June 2017 Nuugaatsiaq, Greenland landslide. We demonstrate the blind detection and recovery of the repeating precursory seismicity that was recorded before the main landslide rupture, which suggests that our approach could lead to more informative forecasting of the seismic activity in seismogenic areas.

Introduction

Current analysis tools for seismic data lack the capacity to investigate the massive volumes of data collected worldwide in a timely fashion, likely leaving crucial information undiscovered. The current reliance on human-expert analysis of seismic records is not only unscalable, but it can also impart a strong bias that favors the observation of already known signals [1]. As a case in point, consider the detection and characterization of non-volcanic tremors, which were first observed in the southwestern Japan subduction zone two decades ago [2]. The complex signals generated by such tremors are very hard to detect in some regions due to their weak amplitude. Robustly detecting these new classes of seismic signals in a model-free fashion would have a major impact in seismology (e.g., for the purpose of forecasting earthquakes), since we would better understand the physical processes of seismogenic zones (subduction, faults, etc.).

Recently, techniques from machine learning have opened up new avenues for rapidly exploring large seismic datasets with minimum a priori knowledge. Machine learning algorithms are data-driven tools that approximate non-linear relationships between observations and labels (supervised learning) or that reveal patterns from unlabeled data (unsupervised learning). Supervised algorithms rely on the quality of the predefined labels, often obtained via classical algorithms [3, 4] or even manually [5, 6, 7, 8]. Inherently, supervised strategies are used to learn how to detect or classify specific classes of already-known signals and, therefore, cannot be used for discovering new classes of seismic signals. Seismology can significantly benefit from the development of unsupervised approaches because the data is mostly unlabeled. Unsupervised tools are likely the best candidates to explore seismic data without the need for any explicit signal model and hence, discover new classes of seismic signals. For this reason, unsupervised methods are more relevant for seismology, where the data is largely unlabeled and new classes of seismic signals should be sought. While supervised strategies are often easier to implement thanks to the evaluation of a prediction error, unsupervised strategies mostly rely on implicit models that are challenging to design. Unsupervised-learning based studies have mostly been applied to data from volcano monitoring systems, where a large variety of seismo-volcanic signals is usually observed [9, 10, 11, 12]. Some unsupervised methods have also been recently applied to induced seismicity [13, 14], global seismicity [15], and local-vs-distance earthquakes [16]. In both cases (supervised or unsupervised), the keystone to success lies in the data representation namely, we need to define an appropriate set of relevant waveform features for solving the task of interest. The features can be manually defined [17, 7, 18] or learned with appropriates techniques such as artificial neural networks [5, 3], the latter belonging to the field of deep learning.

In this paper, we develop a new unsupervised deep-learning method for clustering signals in continuous multichannel seismic time-series. Our strategy combines a deep scattering network [19, 20] for automatic feature extraction and a Gaussian mixture model for clustering. Deep scattering networks belong to the family of deep convolutional neural networks, where the convolutional filters are restricted to wavelets and with modulus activations [19]. The restriction to wavelets filters allows the deep scattering networks to have explicit and physics-related properties that greatly simplifies the architecture design in contrast with classical deep convolutional neural network (frequency band, time scales of interest, amplitudes). Scattering networks have shown to perform high-quality classification of audio signals [21, 20, 22] and electrocardiograms [23]. A deep scattering network decomposes the signal's structure through a tree of wavelet convolutions, modulus operations and average-pooling, providing a stable representation at multiple time and frequency scales [20]. The resulting representation is particularly suitable for discriminating complex seismic signals that may differ in nature (source and propagation effects) with several order of different durations, amplitudes and frequency contents. After decomposing the time series with the deep scattering network, we exploit the representation in a two-dimensional feature space that results from a dimension reduction for visualization and hence, interpretation purposes. The two-dimensional features are finally fed to a Gaussian mixture model for clustering the different time segments.

The design of the wavelet filters have been conducted in many studies, and in each case led to data-adapted filter banks based on intuition on the underlying physics [24, 25, 26] (e.g. music classification, speech processing, bioacoustics, etc.). In order to follow the same idea of optimal wavelet design in a fully explorative way, we propose to learn the mother wavelet of each filter bank with respect to the clustering loss. By imposing a reconstruction constraint to the different layers of the deep scattering network, we guarantee to fully fit the data distribution together with improving the clustering quality. Our approach therefore preserves the structure of a deep scattering network while learning a representation relevant for clustering. It is an unsupervised representation learning method located in between the time-frequency analysis widely used in seismology and the deep convolutional neural networks. While classical convolutional networks usually require a large amount of data for learning numerous coefficients, our strategy can still work with small datasets thanks to the restriction to wavelet filters. In addition, the architecture of the deep scattering network is dictated by physical intuitions (frequency and time scales of interest). This is in contrast to the tedious task of designing deep convolutional neural networks, which today is typically pursued empirically.

Results

Seismic records of the 2017 Nuugaatsiaq landslide. We apply our strategy for blindly clustering and detecting the low-amplitude precursory seismicity [27] to the June 2017 landslide that occurred near Nuugaatsiaq, Greenland at 23:39 UTC. The volume of the rockfall was estimated between 35 to 51 million cubic meters by differential digital elevation models, forming thus a massive landslide [28]. This landslide triggered tsunami waves that impacted the small town of Nuugaatsiaq and caused four reported injuries [28].

The continuous seismic wavefield was recorded by a three-component broadband seismic station (NUUG) located 30 km away from the landslide epicenter

(Fig. 1A). We select the daylong three-component seismograms from 2017-06-17 00:00 to 2017-06-17 23:38 in order to remove the signal due to the mainshock and focus on the content of the seismic wavefield recorded before. A detailed inspection of the east component records revealed that a small event was occurring repetitively before the landslide, starting approximately 9 hours before the rupture and accelerating over time [27, 29]. The accelerating behavior of this seismicity suggests that an unstable initiation was at work before the massive landslide. This signal is not directly visible in raw seismic records; it is of very weak amplitude, has a smooth envelope, and most of its energy located in between 2 and 8 Hz (see a zoom onto the last 4 hours of data in Fig. 1B) and was first highlighted with three-component template matching [27]. While some of these events may be visible in the seismograms filtered between 2 and 9 Hz at times close to the landslide, a large part are hidden in the background noise [27]. The structure of such signal makes it hard to detect via traditional detection algorithms such as STA/LTA (the ratio between Short-Term Average and the Long-Term Average of the seismic signal [30]), because they are sensitive to brutal signal changes with decent signal-to-noise ratios [15]. Besides, STA/LTA only delivers an information about the presence of a signal in the continious trace without any clue of the similarity with other signals, which is our primary goal here.

The template matching strategy consists in a search for similar events in a time series with evaluating a similarity function (cross-correlation) between a pre-defined example of the event (template) and the continuous records. This method is sensitive to the analyzed frequency band, the duration and the quality of the template (often manually defined), making the template matching strategy a severely supervised strategy, yet powerful [31]. Revealing this kind of seismicity with an unsupervised template-matching based strategy could be done with performing the cross-correlation of all time segments (autocorrelation), testing every time segments as potential template event [32]. Considering that several durations, frequency bands, etc. should be tested, this approach is nearly impossible to perform onto large datasets for computational limitations [15].

In the present study, we propose to highlight this precursory event in a blind way over a daylong, raw seismic record. Our goal is to show that even if the precursory signal was not visible after a detailed manual inspection of the seismograms and late times, it could have been correctly detected by our approach. The reader should bear in mind that clustering is an exploratory task [33]; we do not aim at overperforming techniques like template matching, but to provide a first, preliminary and statistical result that could simplify further detailed analyses like template selection for template matching detection.

Feature extraction from a learnable deep scattering network. A diagram of the proposed clustering algorithm is shown in Fig 2. The theoretical definitions are presented in the supplementary materials. Our model first builds a deep scattering network that consists in a tree of wavelet convolutions and modulus operations (Eq. S5). At each layer, we define a bank of wavelet filters with constant quality factor from dilations and stretching of a mother wavelet. This is done according to a geometric progression in the time domain in order to cover a frequency range of interest (the *scale* defined in Eq. S2). The input seismic signal is initially convolved with a first bank of wavelets at different scales, which modulus leads to a first-order scalogram (`conv1`), a time and frequency representation of one-dimensional signals widely used in seismology [34]. In order to speed up computations, we low-pass filter the coefficients in `conv1`, and perform a temporal downsampling (`pool1`) with an average-pooling operation [35]. The coefficients of `pool1` are then convolved with a second wavelet bank, forming the second-order convolution layer (`conv2`). These succession of operations can be seen as a two-layer demodulation, where the input signal's envelope is extracted at the first layer (`conv1`) for several carrier frequencies, and where the frequency content of each envelope is decomposed again at the second layer (`conv2`) [20].

We define a deep scattering network as the sequence of convolution-modulus operations performed at higher orders, allowing to scatter the signal structure through the tree of time and frequency analyses. We finally obtain a locally invariant signal representation by applying an average-pooling operation to the all-order pooling layers [19, 21, 20]. This pooling operation is adapted for concatenation, with an equal number of time samples at each layer (Fig. 2). The scattering coefficients are invariant to local time translation, small signal deformations and signal overlapping. They incorporate multiple time scales (at different layers) and frequencies scales (different wavelets). The tree of operations performed in a scattering network forms a deep convolutional neural networks, where the convolutional filters are restricted to wavelets, and where the activation function is the modulus operator [19]. Scattering networks are located in between (1) classical time and frequency analysis routinely applied in seismology that is often limited to a typical time scale, and (2) deep convolutional neural networks where the unconstrained filters are often hard to interpret, and where the network architecture is often challenging to define. In contrast, deep scattering networks can be designed in a straightforward way, thanks to the analytic framework defined in [19].

From one layer to another, we increase the frequency range of the filter banks in order to consider at the same time small-duration details of the waveform, and larger-duration histories (see Table 1, case D for the selected architecture in the present study). The number of wavelets per octaves and number of octaves defines the frequency resolution and bandwidth of each layer, and the depth (total number of layers) of the scattering network controls the final temporal resolution of the analysis. Following the recommendations cross-validated onto audio signal classification [20], we use a large number of filters at the first layer, and we gradually increase the number of octaves while reducing the number of wavelets per octave from the first to the last layer (Table 1, case D). That way, the representation is highly redundant at the layer `conv1` and gets sparser at the higher-order layers `conv2` and `conv3`, where fewer filters are used at each frequency to decompose the signal. This has the main effect of improving the contrast between signals of different nature [20]. We finally choose the network depth based on the range of time scales of interest. In the present study, we aim at investigating mostly impulsive earthquake-like signals that may last between several seconds to less that one minute. A deeper scattering network could be of interest in order to analyze the properties of longer-duration signals such as seismic tremors [36] or background seismic noise. Finally, with our choice of pooling factors, we obtain a temporal resolution of 35 seconds for each scattering coefficient.

Clustering seismic signals. The scattering coefficients are built in order to be linearly separable [23] so that the need for a high-dimensional scattering representation is greatly reduced. In fact, it is even possible to enforce the learning to favor wavelets that not only solve the task but also provide a lower-dimensional representation of the signal. We do so by reducing the dimension of the scattering coefficients with projection onto the first two principal components (Eq. S10). This also improves the data representation in two dimensions and eases the interpretation. More flexibility could be also added to the procedure by using the latent representation of an autoencoder instead of principal component analysis, because autoencoders can lower the dimension of any datasets with non-linear projections. However, such dimension reduction must be thoroughly investigated because it adds a higher-level complexity to the overall procedure (autoencoder learning rate, architecture, etc.), and will define the goal of future studies.

The two-dimensional scattering coefficients are used to cluster the seismic data. We use a Gaussian mixture model [37] for clustering, where the idea is to find the set of K normal distributions of mean μ_k and covariance Σ_k (where $k = 1 \dots K$) that best describe the overall data (illustrated in Fig. 2 inset, and described in Eq. S11). A categorical variable is also inferred in order to allocate each data sample into each cluster in this procedure, which is the final result of our algorithm. Gaussian mixture model clustering can be seen as a probabilistic and more flexible version of the K-means clustering algorithm, where each covariance can be anisotropic, the clusters can be unbalanced in term of internal variance, and where the decision boundary is soft [37].

Initialized with Gabor wavelets [38], we learn the parameters governing the shape of the wavelets with respect to the clustering loss (Eq. S8) with the *Adam* stochastic gradient descent [39] detailed in the supplementary material (Eq. S14). The clustering loss is defined as the negative log-likelihood of the data to be fully described by the set of normal distributions. We define the wavelets onto specific knots, and interpolate them with Hermite cubic splines onto the same time basis of the seismic data for applying the convolution (see the dedicated section in the material and methods). We ensure that the mother wavelet at each layer satisfies the mathematical definition of a wavelet filter in order to keep all the powerful properties of a deep scattering network [23]. We finally add a constraint on the network in order to prevent the learning procedure to dropout some signals that make the clustering task hard (e.g. outlier signals). This is done by imposing a reconstruction loss from one layer to its parent signal, noticing that a signal should be reconstructed from the sum of the convolutions of itself with a bank of wavelet filters (Eq. S13).

The number of clusters is also inferred by our procedure. We initialize the Gaussian mixture clustering algorithm with a (large) number $K = 10$ clusters at the first epoch, and let all of these components be used by the Expectation-Minimization strategy [37]. This is shown at the first epoch in the latent space in Fig. 3A, where the Gaussian component mean and covariance are shown in color with the corresponding population cardinality on the right-hand side. As the learning evolves, we expect the representation to change the coordinates of the two-dimensional scattering coefficients in the latent space (black dots), leading to Gaussian components that do not contribute anymore to fit the data distribution, and therefore to be automatically disregarded in the next iteration. We can therefore infer a number of clusters from a maximal value. At the first epoch (Fig. 3A), we observe that the seismic data samples are scattered in the latent space, and that the Gaussian mixture model used all of the 10 components to explain the data.

The clustering loss decreases with the learning epochs (Fig. 3C). We declare the clustering to be optimal when the loss stagnates (reached after approximately 7,000 epochs). The learning is done with batch-processing, a technique that allows for faster computation by randomly selecting smaller subsets of the full dataset. This also avoids falling into local minima, as we can observe around epoch 3,500, and guarantees to reaching a stable minimum that does not evolve anymore after epoch 7,000 (Fig. 3C). After 10,000 training epochs, as expected, we observe that the scattering coefficients have been concentrated around the clusters centroids obtained with the Gaussian mixture model (Fig. 3B). The set of useful components have been reduced to 4, a consequence of a better learned representation due to the learned wavelets at the last epoch (Fig. 3D). The cluster colors range from colder to warmer colors depending on the population size.

The clustering loss improves by a factor of approximately 4.5 between the first and the last epoch (Fig. 3C). At the same time, we observe that the reconstruction loss is more than 15 times smaller than at the first training epoch (Table 1). This indicates that the basis of wavelets filter banks used in the deep scattering network is powerful to accurately represent the seismic data with ensuring a good-quality clustering at the same time.

Analysis of clusters. An analysis of the temporal evolution of the clusters is presented in Fig. 4. The within-cluster cumulative detections obtained after training (epoch 10,000) are presented in Fig. 4A for clusters 1 and 2, and in Fig. 4B for clusters 2 and 3. The two most populated clusters 1 and 2 (Fig. 4A) gather more than 90% of the overall data (observed on the histograms in Fig. 3B). They both show a linear detection rate over the day with no particular concentration in time and, therefore, relate to the background seismic noise. Clusters 3 and 4 (Fig. 4B) show different non-linear trends that include 10% of the remaining data.

The temporal evolution of cluster 4 is presented in Fig. 4B. The time segments that belong to cluster 4 are extracted and aligned to a reference time segment (at the top) with local cross-correlation for better readability (see further details about the strategy in the supplementary materials). We see that these time segments contain seismic events localized in time with relatively high signal-to-noise ratio and sharp envelope. These events do not show a strong similarity in time, but they strongly differ from the event belonging to other clusters, explaining why they have been gathered in the same cluster. The detection rate is sparse in time, indicating that cluster 4 is mostly related to a random background seismicity or other signals which interest is beyound the scope of the present manuscript.

The temporal evolution of cluster 3 shows three behaviors. First, we observe a nearly-constant detection rate from the beginning of the day to approximately 07:00. Second, the detection rate lowers between 07:00 and 13:00 where only 4% of the within-cluster detections are observed. An accelerating seismicity is finally observed from 13:00 up to the landslide time (23:39 UTC). The time segments belonging to cluster 3 are reported on Fig. 4D in gray colorscale, and aligned with local cross-correlation with a reference (top) time segment. The correlation coefficients obtained for the time lag that maximizes the alignment are indicated in orange color in Fig. 4E. As with the template matching strategy, we clearly observe the increasing correlation coefficient with the increasing event index [27], indicating that the signal-to-noise ratio increases towards the landslide rupture. This suggests that the repeating event may still exist earlier in the data even before 15:00, but that the detection threshold of the template matching method is limited by the signal-to-noise ratio [27]. In contrast, we observe that the probability of these 171 events remains high in our approach,with 97% of the precursory events previously found [27] recovered.

A interesting observation is the change of behavior in the detection rate of this cluster at nearly 07:00 (Fig. 4B). The events that happened before 07:00 have all a relatively high probability to belong to cluster 3, refuting the hypothesis that noise samples have randomly been misclassified by our strategy (Fig. 4E). The temporal similarity of all these events in Fig. 4D is particularly visible for later events (high index) because the signal-to-noise ratio of these events increases towards the landslide [27]. The two trends may be whether related to similar signals generated at same position (same propagation) with a different source, or by two types of alike-looking events that differ in nature, but that may have been gathered in the same cluster because they strongly differ from the other clusters. This last hypothesis can be tested with using hierarchical clustering [40]. Our clustering procedure highlighted those 171 similar events in a totally unsupervised way, without the need of defining any template from the seismic data. The stack of the 171 waveforms is shown in black solid line in Fig. 4D, indicating that the template of these events is defined in a blind way thanks to our procedure. In addition, these events have very similar properties (duration, seismic phases, envelope) in comparison with the template defined in [27].

Discussion and conclusions

We have developed a new strategy for clustering and detecting seismic events in continuous seismic data. Our approach extends a deterministic deep scattering network by learning the wavelet filter-banks and applying a Gaussian mixture model. While scattering networks correspond to a special deep convolutional neural network with fixed wavelet filter-banks, we allow it to fit the data distribution by learnability of the different mother wavelets; yet we preserve the structure of the deep scattering network allowing interpretability and theoretical guarantees. We combine the powerful representation of the learnable scattering network with Gaussian mixture clustering by learning the shape of the wavelet filters according to the clustering loss. This allows to learn a representation of multichannel seismic signals that maximizes the quality of clustering, leading to an unsupervised way of exploring possibly large datasets. We also impose a reconstruction loss as each layer of the deep scattering network, following the ideas of convolutional autoencoders, thus preventing to learn trivial solutions such as zero-valued filters.

Our strategy is capable of blindly recovering the small-amplitude precursory signal reported in [27, 29]. This indicates that waveform templates can be recovered from our method without the need of any manual inspection of the seismic data prior to the clustering process, and tedious selection of waveform template in order to perform high-quality detection. Such unsupervised strategy is of strong interest in the exploration of seismic datasets, where the structure of seismic signals can be complex (low-frequency earthquakes, non-volcanic tremors, distant vs. local earthquakes, etc.), and where some class of unknown signals is likely to be disregarded by a human expert.

In the proposed workflow, only a few parameters need be chosen, namely the number of octaves and wavelets per octave at each layer $J^{(\ell)}$ and $Q^{(\ell)}$, the number of knots $\mathcal{K}$ the pooling factors and the network depth M. This choice of parameters is extremely constrained by the underlying physics. The number of octaves at each layer controls the lowest analyzed frequency at each layer, and therefore, the largest time scale. The pooling factor and number of layers M should be chosen according to the analyzed time scale at each layer, and the final maximal time scale of interest for the user. We discuss our choice of parameters with testing several parameter sets summarized in Table 1 and with corresponding results summarized in Fig. S5 for the cumulative detection curves, within-cluster population sizes and learned mother wavelets. All the results obtained with different parameters show extremely similar cluster shapes in the time domain, and the precursory signal accelerating shape is always recovered. We see that a low number of 3 or 4 clusters are found in almost all cases, with a highly similar detection rates over the day. Furthermore, we observe that the shape of the learned wavelets remain highly similar between the different data-driven tests, and in particular, the third-order wavelet is highly similar with all the tested parameters (Fig. 5G). This result makes sense because the coefficients that output from the last convolutional layer *conv3* are over-represented in comparison with the other ones. We also observe that the procedure still works with only a few amount of data (Fig. 5A–C), a very strong advantage compared with classical deep convolutional neural networks that often require a large amount of data to be successfully applied.

Besides being adapted to small amount of data, our strategy can also work with large amount of data, as scalability is garanteed by batch processing, and using only small-complexity operators (convolution and pooling). Indeed, batch processing allows to control the amount of data seen by the scattering network and GMM at a single iteration, each epoch being defined when the whole dataset have been analyzed by the algorithm. There is no limitation to the total amount of data being analyzed because only the selected segments at each iteration are fed to the network. At longer time scales, the number of clusters needed to fit the seismic data must change, however, with an expectation that the imbalance between clusters should increase. We illustrate this point another experiment performed on the continuous seismogram recorded at the same station over 17 days, including the date of the landslide (from 2017-06-01 to 2017-06-18). With this larger amount of data, the clustering procedure still converges and exhibit 9 new clusters. The hourly within-clusters detections of these new clusters are presented in Fig. 5. Among the different clusters found by our strategy, we observe that more than 93% of the data is identified in slowly evolving clusters, most likely related to fluctuations of the ambient seismic noise (Fig. 5, clusters A to E). The most populated clusters (A and B) occupy more than 61% of the time, and are most likely related to diffuse wavefield without any particular dominating source. Interestingly, we observe two other clusters with large population with a strong localisation in time (clusters C and D in Fig. 5). A detailed analysis of the ocean-radiated microseismic energy [44, 45] allowed us to identify the location and dominating frequency of the sources reponsible for these clusters to be identified (illustrated in Fig. S2 and S3 in the supplementary material). The source time function of the best-matching microseismic sources have been reported on clusters C and D in Fig. 5.

Compared with these long-duration clusters, the clustering procedure also reports very sparse clusters where less than 7% of the seismic data is present. Because of clustering instabilities caused by the large class imbalance of the seismic data, we decided to perform a second-order clustering on the low-populated clusters. This strategy follows the idea of hirearchical clustering [40], where the firstly identified clusters are analyzed several consecutive times in order to discover within-cluster families. For the sake of brevity, we do not intend to perform a deep-hirarchical clustering in the present manuscript, but to illustrate the potential strength of such strategy in seismology, where the data is essentially class-imbalanced. We perform a new clustering from the data obtained in the merged low-populated clusters (F to I in Fig. 5). This additional clustering procedure detected two clusters presented in Fig. 6A. These two clusters have different temporal cumulated detections and exhibits different population sizes. A zoom of the cumulated within-cluster detections is presented in Fig. 6B, and show a high similarity with clusters 3 and 4 previously obtained in Fig. 3 from the daylong seismogram. This result clearly proves that the accelerating precursor is captured by our strategy even when the data is highly imbalanced. If the scattering network provide highly relevant features, clustering seismic data with simple clustering algorithms can be a hard task that can be solved with hierachical clustering, as illustrated in the present study. This problem can also be better tackled by other clustering algorithms such as spectral clustering [41] which has the additional ability to detect outliers. Clustering the outlier signals may then be an alternative to GMM in that case. Another possibility would be to use the local similarity search with hashing functions [15] in order to improve our detection database onto large amount of seismic data.

The structure of the scattering network shares some similarities with the FAST algorithm (for Fingerprint And Similarity Search [15]) from a architectural point of view. FAST uses a suite of deterministic operations in order to extract waveforms features and feed it to a hashing system in order to perform a similarity search. The features are extracted from the calculation of spectrogram, Haar wavelet transforms and thresholding operations. While being similar, the FAST algorithm involves a number of paramaters that are not connected to the underlying physics. For instance, the thresholding operation has to be manually inspected [15], as well as the size of the analyzing window. In comparison, our alrotihm's parameters are based on physical intuition, and does not imply any signal windowing (only the resolution of the final result can be controlled). FAST is not a machine learning strategy because no learning is involved; in contrast, we do learn the representation of the seismic data that best solves the task of clustering. While FAST needs a large amount of data to be run in an optimal way [15], our algorithm still works with a few number of samples.

This work shows that learning a representation of seismic data in order to cluster seismic events in continuous waveforms is a challenging task that can be tackled with deep learnable scattering networks. The blind detection of the seismic precursors to the 2017 Landslide of Nuugaatsiaq with a deep learnable scattering network is a strong evidence that weak seismic events of complex shape can be detected with a minimum amount of prior knowledge. Discovering new classes of seismic signals in continuous data can, therefore, be better addressed with such strategy, and could lead to a better forecasting of the seismic activity in seismogenic areas.

Aknowldgements

L.S., P.P. and M.C. acknowledge support from the European Research Council under the European Union Horizon 2020 research and innovation program (grant agreement no. 742335, F-IMAGE). M.V.d.H. gratefully acknowledges support from the Simons Foundation under the MATH + X program and from NSF under grant DMS-1815143. R.B. and R.G.B. were supported by NSF grants IIS-17-30574 and IIS-18-38177, AFOSR grant FA9550-18-1-0478, ONR grant N00014-18-12571, and a DOD Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047. L.S. thanks Romain Cosentino for very helpful discussions and comments. The facilities of IRIS Data Services, and specifically the IRIS Data Management Center, were used for access to waveforms and related metadata used in this study. IRIS Data Services are funded through the Seismological Facilities for the Advancement of Geoscience and EarthScope (SAGE) Project funded by the NSF under Cooperative Agreement EAR-1261681. The authors declare that they have no competing financial interests. Correspondence and requests for materials should be addressed to L.S. (email: leonard.seydoux@univ-grenoble-alpes.fr).

Figures and tables

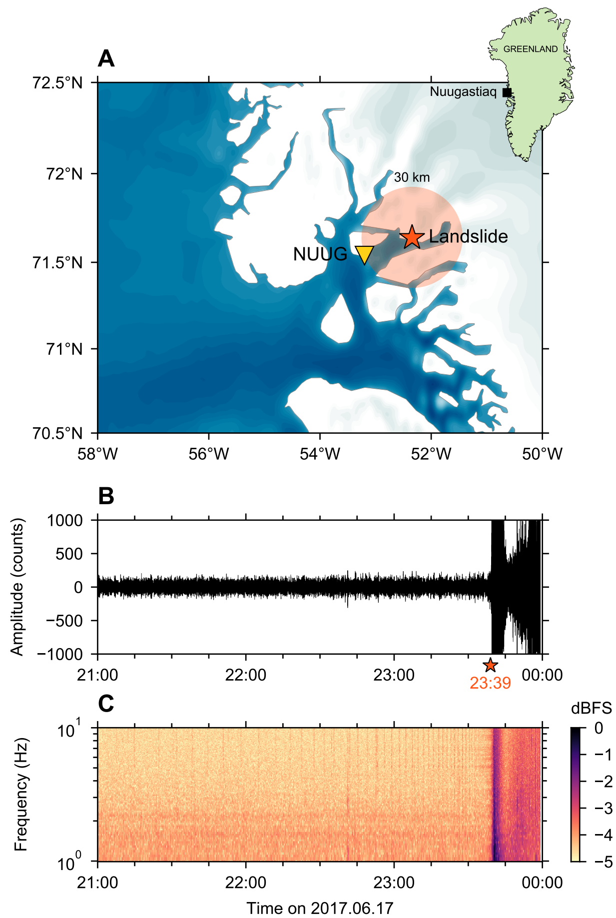

Figure 1: **Geological context and seismic data.** **A** Location of the landslide (red star) and the seismic station NUUG (yellow triangle). The seismic station is located in the vicinity of the small town of Nuugaatsiaq, Greenland (top-right inset). **B** Raw record of the seismic wavefield collected between 21:00 UTC and 00:00 UTC on 2017-06-17. The seismic waves generated by the landslide main rupture are visible after 23:39 UTC. **C** Fourier spectrogram of the signal from B obtained over 35-second long windows.

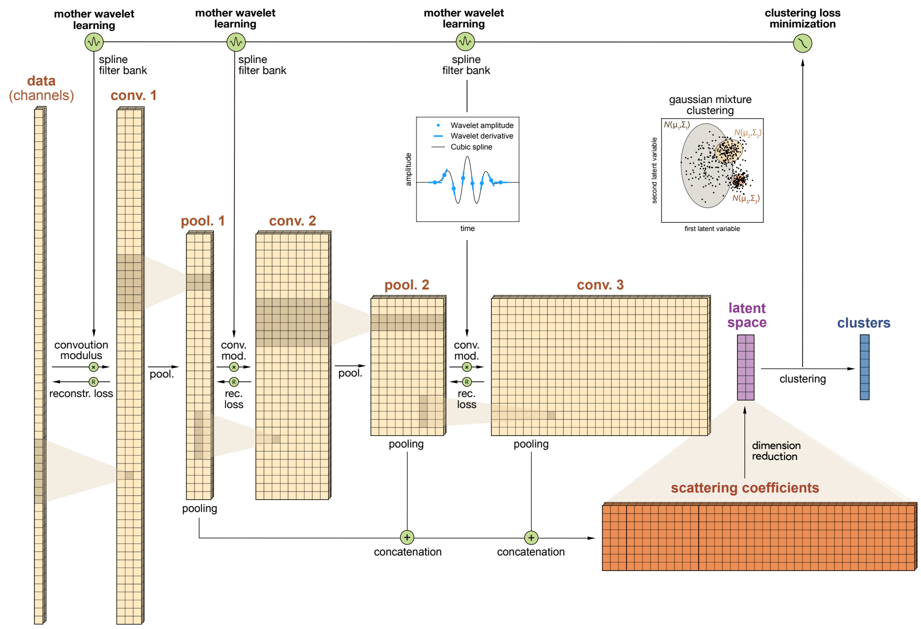

Figure 2: **Deep learnable scattering network with Gaussian mixture model clustering.** The network consists in a tree of convolution and modulus operations successively applied to the multichannel time series (layers conv 1 – 3). A reconstruction loss in calculated at each layer in order to constrain the network not to cancel out any part of the signal (Eq. S13). From one layer to another, the convolution layers are downsampled with an average pooling operation (pool 1 – 2), except for the last layer which can be directly used to compute the deep scattering coefficients. This allows to analyze large time scales of the signal structure with the increasing depth of the deep scattering network at reasonable computational cost. The scattering coefficients are finally obtained from the equal pooling and concatenation of the pool layers, forming a stable high-dimensional and multiple time and frequency scale representation of input multichannel time series. We finally apply a dimension reduction to the set of scattering coefficients obtained at each channel in order to form the low-dimensional latent space (here two-dimensional as defined in Eq. S10). We use a Gaussian mixture model in order to cluster the data in the latent space (Eq.S11). The negative log-likelihood of the clustering is used to optimize the mother wavelet at each layer (inset) with *Adam* [39] stochastic gradient descent (Eq.S14). The filter bank of each layer ℓ is then obtained by interpolating the mother wavelet in the temporal domain $\psi_0^{(\ell)}(t)$ with Hermite cubic splines (Eq. S9), and dilating it over the total number of filters $J^{(\ell)}Q^{(\ell)}$ (see Eq. S2).

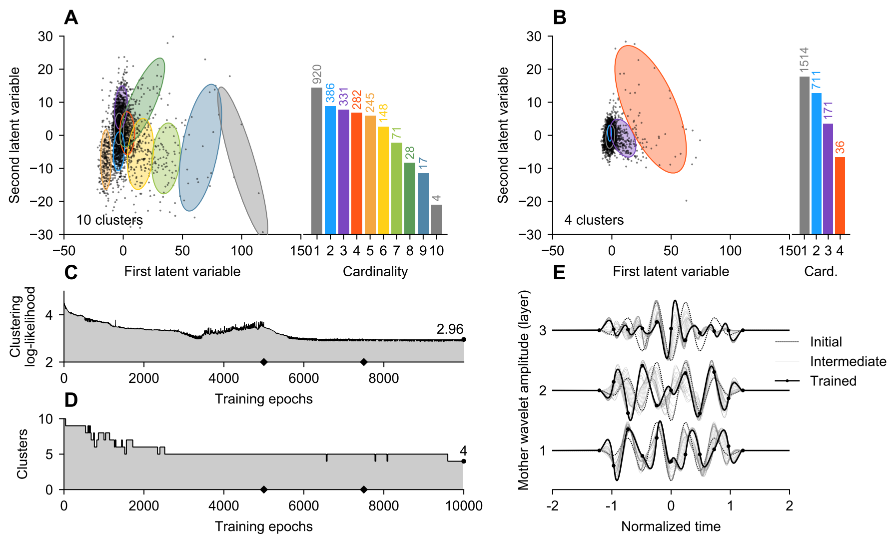

Figure 3: **Learning results.** Scattering coefficients in the latent space at initialization (**A**) and after learning (**B**). The covariance of each component of the Gaussian mixture model is represented by a colored ellipse centered at each component mean. All of the 10 components are used at initial stage with a steadily decaying number of elements per clusters, while only 4 are used at final stage with unbalanced population size. The clustering negative log-likelihood (**C**, top) decreases with the learning epochs indicating that the clustering quality is improved by the learned representation. We also observe that the reconstruction loss fluctuates and remains as low as possible (**C**, bottom). The number of cluster with respect to the increasing training epoch is shown in (**D**). Finally, the initial, intermediate and final wavelets at each layer (**E**) are shown in the time domain interpolated from 11 knots.

	Data		Scattering network				Learning		
Title	Start	End	$J^{(\ell)}$	$Q^{(\ell)}$	$\mathcal{K}$	Pool.	Clusters	Loss (clus.)	Loss (rec.)
A	15:00	23:30	3, 6, 6	8, 2, 1	7	2^{10}	$10 \to 4$	3.79	4.20
B	15:00	23:30	3, 6, 6	8, 2, 1	11	2^{10}	$10 \to 3$	3.42	5.40
C	15:00	23:30	3, 6, 6	8, 2, 1	15	2^{10}	$10 \to 3$	3.17	5.49
⋆ **D**	**00:30**	**23:30**	**4, 6, 6**	**8, 4, 3**	**11**	$\mathbf{2^{10}}$	$\mathbf{10 \to 4}$	**2.96**	**3.06**
E	00:30	23:30	3, 6, 6	8, 2, 1	11	2^{9}	$10 \to 6$	3.67	1.76
F	00:30	23:30	3, 6, 6	8, 2, 1	11	2^{11}	$10 \to 4$	3.11	3.06

Table 1: **Set of different tested parameters** (with corresponding cumulative detection curves shown in Fig. 5). The results presented in Figs. 3 and 4 are obtained with the set of parameters D (black star and bold typeface), with the lowest clustering loss.

Supplementary materials

Deep scattering network

A complex wavelet $\psi \in \mathcal{L}$ is a filter localized in frequency with zero average, center frequency ω_0 and bandwidth $\delta\omega$. We define the functional space $\mathcal{L}$ of any complex wavelet ψ as

$$\mathcal{L} = \left\{ \psi \in L^2_c(\mathbb{C}), \int \psi(t) dt = 0 \right\}, \tag{S1}$$

where $L^2_c(\mathbb{C})$ represents the space of square integrable functions with compact time support c on $\mathbb{C}$. At each layer, the mother wavelet $\psi_0 \in \mathcal{L}$ is used to derive a number of JQ wavelets of the filter bank ψ_j with dilating the mother wavelet by means of scaling factors $\lambda_j \in \mathbb{R}$ such as

$$\psi_j(t) = \lambda_j \psi_0(t\lambda_j), \quad \forall j = 0 \ldots JQ - 1\,. \tag{S2}$$

where the mother wavelet is centered at the highest possible frequency (Nyquist frequency). The scaling factor $\lambda_j = 2^{-j/Q}$ is defined as powers of 2 in order to divide the frequency axis in portions of octaves depending on the desired number of wavelets per octaves Q and total number of octaves J which controls the frequency axis limits and resolution at each layer. The scales are designed to cover the whole frequency axis, from the Nyquist angular frequency $\omega_0 = \pi$ down to a smallest frequency $\omega_{QJ-1} = \omega_0 \lambda_J$ defined by the user.

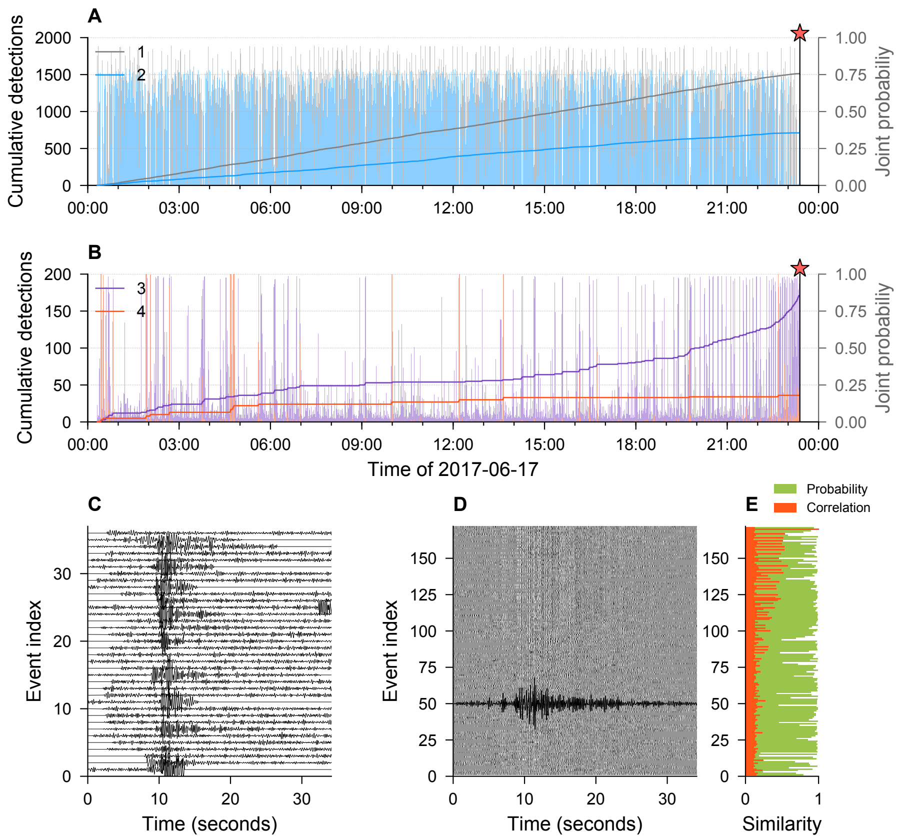

Figure 4: **Analysis of clusters in the time domain.** Within-cluster cumulative number of detection of events in clusters 1 and 2 (**A**) and clusters 3 and 4 (**B**) at epoch 10,000. The relative probability for each time window to belong to each cluster is represented with lighter bars. The waveforms extracted within the last two clusters (purple and red) are extracted and aligned with respect to a reference waveform within the cluster, for cluster 4 (**C**) and cluster 3 (**D**). The seismic data have been bandpass-filtered between 2 and 8 Hz for better visualization of the different seismic events. (**E**) similarity measurement in the time domain (correlation) and in the latent space (probability) for the precursory signal.

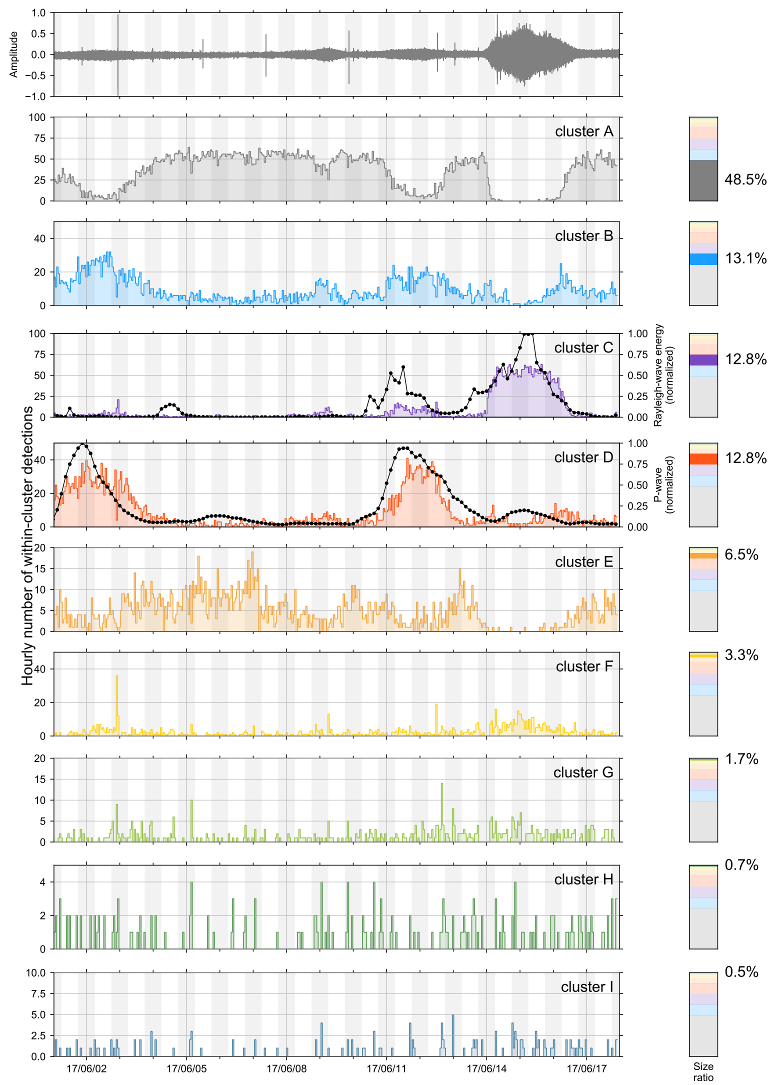

Figure 5: **Clustering results obtained long-duration seismic data.** The broadband seismogram recorded by the station NUUG (Fig. 1) from 2017-06-01 to 2017-06-18 is presented in the top plot. The hourly within-cluster detection rate is presented for each of the 9 clusters (A to I). The right-hand side insets indicate the relative population size of each clusters. The best-correlating microseismic energy have been reported on top of clusters C and D, respectively automatically identified from offshore the city of Nuugastiaq, and in the middle of the North Atlantic (see Fig. S2 and S3 for more details).

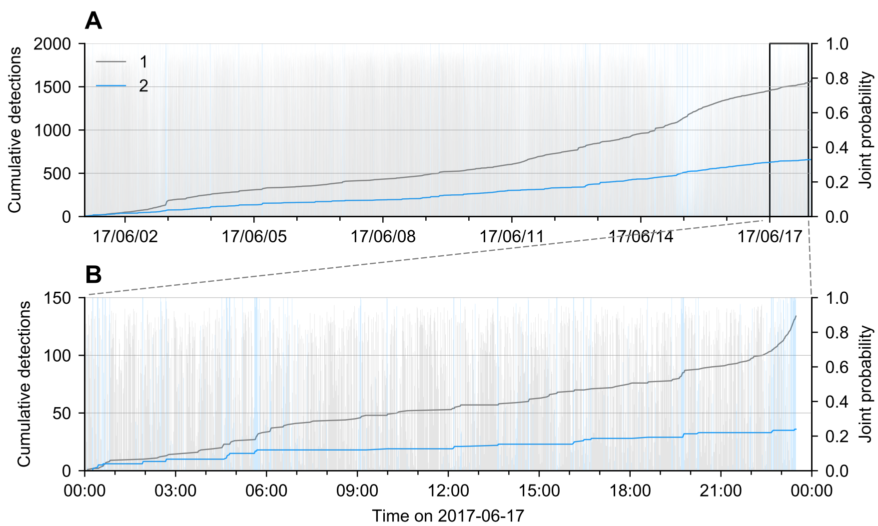

Figure 6: **Hierarchical clustering of long-duration seismic data.** (**A**) Within-cluster cummulative detection oversed for second-order clustering of former clusters F to I presented in Fig. 5 from 2017-06-01 to 2017-06-18 . (**B**) Zoom on the day 2017-06-17 from the detections presetend in A. Similarly to Fig. 3, the relative probability for each time window to belong to each cluster is represented with lighter bars.

We define the first convolution layer of the scattering network (conv1 in Fig. 2) as the convolution of any signal $x(t) \in \mathbb{R}^C$ (where C denotes the number of channels) with the set of $J^{(1)}Q^{(1)}$ wavelet filters $\psi_j^{(1)}(t) \in \mathcal{L}$ as

$$U_j^{(1)}(t) = \left| x * \psi_j^{(1)} \right| (t) \in \mathbb{R}^{C \times J^{(1)} \times Q^{(1)}}, \tag{S3}$$

where $*$ represents the convolution operation. The first layer of the scattering network defines a scalogram, a time-frequency representation of the signal $x(t)$ according to the shape of the moher wavelet $\psi_0^{(1)}$ widely used in the analysis of one-dimensional signals including seismology.

The first-order scattering coefficients $S_j^{(1)}(t)$ are obtained after applying an average-pooling operation $\phi(t)$ over time to the first-order scalogram $U_j^{(1)}(t)$

$$S_j^{(1)}(t) = \left(U_j^{(1)} * \phi_1 \right)(t) = \left(|x * \psi_{j_1}| * \phi_1 \right)(t). \tag{S4}$$

The average-pooling operation is equivalent to a low-pass filtering followed by a downsampling operation [35]. It ensures the scattering coefficients to be locally stable with respect to time, providing a representation stable to local deformations and translations [21]. This property is essential in the analysis of complex signals such as seismic signals that can often be perturbed by scattering or present a complex source time function.

The small details information that has been removed by the pooling operation with Eq. S4 could be of importance to properly cluster different seismic signals. It is recovered by cascading the convolution, modulus and pooling operations on higher-order convolutions performed on the first convolution layer (thus defining the high-order convolution layers shown in Fig. 2):

$$S_j^{(\ell)}(t) = U_j^{(\ell)}(t) * \phi_j^{(\ell)}(t), \tag{S5}$$

where $U^{(0)}(t) = x(t)$ is the (possibly multichannel) input signal (Fig. 2). The scattering coefficients are obtained at each layers from the successive convolution of the input signal with different filters banks $\psi^{(\ell)}(t)$. In addition, we apply an average pooling operation to the output of the convolution-modulus operators in order to downsample the successive convolutions without aliasing. This allow for observing larger and larger time scales in the structure of the input signal at reasonnable computational cost.

We define the relevant features $\mathbf{S}(t)$ of the continuous seismic signal to be the concatenation of all-orders scattering coefficients obtained at each time t as

$$\mathbf{S}(t) = \{S^{(\ell)}\}_{\ell=1\ldots M} \in \mathbb{R}^F, \tag{S6}$$

with M standing for the depth of the scattering network, and $F = J^{(1)}Q^{(1)}(1 + \ldots (1 + J^{(M)}Q^{(M)}))$ is the total number of scattering coefficients (or features). When dealing with multiple-channel data, we also concatenate the scattering coefficients obtained at all channels. The feature space therefore is a high-dimensional representation that encodes multiple time-scales properties of the signal over a time interval $[t, t+\delta t]$. The time resolution δt of this representation then depends on the size of the pooling operations. The choice of the scattering network depth thus should be chosen so that the final resolution of analysis is larger that maximal duration of the analyzed signals.

Seismic signals can have several orders of different magnitude, even for signals lying in the same class. In order to make our analysis independent from the amplitude, we normalize the scattering coefficient by the amplitude of their "parent". The scattering coefficients of order m are normalized by the amplitude of the coefficients $m-1$ down to $m=2$. For the first layer (which has no parent), the scattering coefficients are normalized by the coefficients of the absolute value of the signal [42].

Adaptive Hermite cubic splines

Instead of learning all the coefficients of the mother wavelet $\psi_0^{(\ell)}$ at each layer in the frequency domain, as one would do in a convolutional neural network, we restrict the learning to the amplitude and the derivative on a specific set of $\mathcal{K}$ knots $\{t_k \in c\}_{k=1\ldots\mathcal{K}}$ laying in the compact temporal support c (see Eq. S1). The mother wavelet $\psi_0^{(\ell)}$ can then be approximated with Hermite cubic splines [23], a third-order polynomial defined on the interval defined by two consecutive knots $\tau_k = [t_k, t_{k+1}]$. The four equality constraints

$$\begin{cases} \psi_0^{(\ell)}(t_k) = \gamma_k \\ \psi_0^{(\ell)}(t_{k+1}) = \gamma_{k+1} \\ \dot{\psi}_0^{(\ell)}(t_k) = \theta_k \\ \dot{\psi}_0^{(\ell)}(t_{k+1}) = \theta_{k+1} \end{cases}, \tag{S7}$$

uniquely determine the Hermite cubic spline solution piecewise on the consecutive time segments τ_k, given by

$$\psi_{0,\Gamma,\Theta}^{(\ell)}(t) = \sum_{k=1}^{\mathcal{K}-1} \gamma_k f_1\left(x_k(t)\right) + \gamma_{k+1} f_2\left(x_k(t)\right) + \theta_k f_3\left(x_k(t)\right) + \theta_{k+1} f_4\left(x_k(t)\right) \mathbf{1}_{\tau_k}, \tag{S8}$$

where $\Gamma = \{\gamma_k\}_{k=1\dots\mathcal{K}-1}$ and $\Theta = \{\theta_k\}_{k=1\dots\mathcal{K}-1}$ respectively are the set of value and derivative of the wavelets on the knots, where $x(t) = \frac{t-t_k}{t_{k+1}-t_k}$ is the normalized time on the interval τ_k, and where the Hermite cubic functions $f_i(t)$ are defined as

$$\begin{cases} f_1(t) = 2t^3 - 3t^2 + 1, \\ f_2(t) = -2t^3 + 3t^2, \\ f_3(t) = t^3 - 2t^2 + t, \\ f_4(t) = t^3 - 2t^2. \end{cases} \tag{S9}$$

We finally ensure that the Hermite spline solution lays in the wavelets functional space $\mathcal{L}$ defined in Eq. S1 by additionnaly imposing

- the compactness of the support: $\gamma_1 = \theta_1 = \theta_K = \gamma_K = 0$,
- the null average: $\gamma_k = -\sum_{n \neq k} \gamma_n$,
- that the coefficients are bounded: $\max_t \ \gamma_t < \infty$.

The parameters γ_k and θ_k solely control the shape of the mother wavelet and are the only parameters that we learn in our strategy. Notice that thanks to the above constraints, for any value of those parameters, the obtained wavelet is guaranteed to belong into the functional space of wavelets $\mathcal{L}$ defined in Eq. S1 with compact support. By simple approximation argument, Hermite cubic splines can approximate arbitrary functions with a quadratically decreasing error with respect to the increasing number of knots $\mathcal{K}$. Once the mother filter has been interpolated, the entire filter-bank is derived according to Eq. S2.

Clustering in a low-dimensional space

We decompose the scattering coefficients $\mathbf{S}$ onto its two first principal components by means of singular value decomposition $\mathbf{S} = \mathbf{UDV}^{\dagger}$, where $\mathbf{U} \in \mathbb{R}^{F \times F}$ and $\mathbf{V} \in \mathbb{R}^{T \times T}$ are respectively the feature- and time-dependant singular matrices gathering the singular vectors column-wise, $\mathbf{D}$ are the singular values, and where T is the total number of time samples in the scattering representation. We define the latent space $\mathbf{L} \in \mathbb{R}^{2 \times T}$ as the projection of the scattering coefficients onto the first two feature-dependent singular vectors. Noting $\mathbf{U} = \{\mathbf{u}_i\}_{i \in [1 \ldots F]}$ and $\mathbf{V} = \{\mathbf{v}_j\}_{j \in [1 \ldots T]}$ where $\mathbf{u}_i$ and $\mathbf{v}_j$ are respectively the singular vectors, the latent space is defined as

$$\mathbb{R}^{2 \times T} \ni \mathbf{L} = \sum_{i=1}^{2} \mathbf{S}\mathbf{u}_i \tag{S10}$$

To tackle clustering tasks, it is common to resort to centroidal-based clustering. In such strategy, the observations are compared to cluster prototypes and associated to the clusters with prototype the closest to the observation. The most-famous centroidal clustering algorithm is probably the K-means algorithm. Its extension, the Gaussian mixture model extends it by allowing non uniform prior over the clustering (unbalanced in the clusters) and by allowing to adapt the metric used to compare an observation to a prototype by means of a covariance matrix. To do so, Gaussian mixture model resorts to a generative modeling of the data. When using a Gaussian mixture model, the data are assumed to be generated according to a mixture of K independant normal (Gaussian) processes $\mathcal{N}(\mu_k, \Sigma_k)$ as in

$$x \sim \prod_{k=1}^{K} \mathcal{N}(\mu_k, \Sigma_k)\mathbf{1}_{\{t=k\}} \tag{S11}$$

where t is a Categorical variable governed by $t \sim \mathcal{C}at(\pi)$. As such, the parameters of the model are $\{\mu_k, \Sigma_k, k = 1 \ldots K\} \cup \{\pi\}$. The graphical model is given by $p(x,t) = p(x|t)p(t)$ and the parameters are learned by maximum likelihood with the expectation-maximization technique, where for each input x, the missing variable (unobserved) t is inferred using expectation with respect to the posterior distribution as $E_{p(t|x)}(p(x|t)p(t))$. Once this latent variable estimation has been done, the parameters are optimized with their maximum likelihood estimator. This two step process is then repeated until convergence which is guaranteed [43].

Learning the wavelets with gradient descent

The clustering quality is measured in term of negative log-likelihood $\mathcal{T}$ with respect to the Gaussian mixture model formulation (here calculated with the expectation-minimization method). The negative log-likelihood is used to learn and adapt the Gaussian mixture model parameters (via their maximum likelihood estimates) in order to fit the model to the data. We aim at adapting our learnable scattering filter-banks in accordance to the clustering task to increase the clustering quality. The negative log-likelihood will thus be used to adapt the filter-bank parameters.

This formulation alone contains a trivial optimum at which the filter-banks disregard any non stationary event leading to a trivial single cluster and the absence of representation of any other event. This would be the simplest clustering task and would minimize the negative log-likelihood. As such it is necessary to force the filter-banks to not just learn a representation more suited for Gaussian mixture model clustering but also not to disregard information from the input signal. This can be done naturally by enforcing the representation of each scattering to contain enough information to reconstruct the layer input signal. Thus, the parameters of the filters are learned to jointly minimize the negative log-likelihood and a loss of reconstruction.

Reconstruction loss

The reconstruction $\hat{x}(t)$ of any input signal $x(t)$ can be formally written in the single-layer case as

$$\hat{x}(t) = \sum_{i=1}^{JQ} \frac{1}{C(\lambda_i)} \sum_{t'} \psi_i(t - t') \left|(x * \psi_i)(t')\right| \tag{S12}$$

where $C(\lambda_i)$ is a renormalization constant at scale λ_i, and $*$ stands for convolution. While some analytical constant can be derived from the analytical form of the wavelet filter, we instead propose a learnable coefficient obtained by incorporating a batch-normalization operator. The model thus considers $\hat{x} = (\text{BatchNorm} \circ \text{Deconv} \circ |\cdot| \circ \text{BatchNorm} \circ \text{Conv})(x)$. From this, the reconstruction loss is simply given by the expression

$$\mathcal{L}(x) = \|x - \hat{x}\|_2^2. \tag{S13}$$

We use this reconstruction loss for each of the scattering layers.

Stochastic gradient descent

With all the losses defined above we are able to leverage some flavor of gradient descent [39] in order to learn the filter parameters. Resorting to gradient descent is here required as analytical optimum is not available for the wavelet parameters as we do not face a convex optimization problem. During training, we thus iterate over our dataset by means of mini-batches (a small collection of examples seen simultaneously) and compute the gradients of the loss function with respect to each of the wavelet parameters as

$$G(\theta) = \frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \left(\frac{\partial \mathcal{T}}{\partial \theta}(x_n) + \sum_{i=1}^{\ell} \frac{\partial \mathcal{L}^{(i)}}{\partial \theta} \left(x_n^{(i)} \right) \right), \tag{S14}$$

with $\mathcal{B}$ being the collection of indices in the current batch and θ being one of the wavelet parameters (the same is performed for all parameters of all wavelet layers). The ℓ superscript on the reconstruction loss represent the reconstruction loss for layer ℓ. Then, the parameter is updated following

$$\theta^{t+1} = \theta^t - \alpha G(\theta) \tag{S15}$$

with α the learning rate. Doing so in parallel for all the wavelet parameters concludes the gradient descent update of the current batch at time t. This is repeated multiple time over different mini-batches until convergence.

Within-cluster waveform analysis

The waveforms that belong to similar clusters are extracted from the continuous seismic data based on the starting t_i and ending dates $t_i + dt$of the scattering coefficients, where dt is the temporal resolution of the scattering coefficients. The time segments are extracted with an additional small time delay ϵdt in order to allow for cross-correlating the time segments. We align the M waveforms $w_m(t)$ belonging to the same cluster with respect to a reference waveform $w_r(t)$ by means of cross-correlation, and collect the maximal correlation coefficient

$$c_{mr} = \max_{\tau} \int_{t=0}^{T} w_m(t) w_r(t-\tau) dt \tag{S16}$$

Tests with different parameters

One key parameter is the number of knots used to learn the shape of the wavelet. This parameter is responsible for the wavelet duration in time, and inherently for the wavelet bank quality factor. Indeed, a small number of knots defines a wavelet localized in time with a large frequency bandwidth and vice-versa. We therefore vary the number of knots in Fig. 5A to C in order to observe both the clustering and reconstruction losses onto a small subset of the dataset (8.5 hours). These tests are also very helpful to show that the procedure still works with a small amount of data (9 hours), a situation where deep convolutional neural networks are known to fail easily. We see that taking a low number of 7 knots (case A) allows to better reconstruct the input data with a loss of 4.20 (Table 1), but have a relatively high clustering loss (3.79). We observe in Fig. 5A that the cumulative curves trends are not clearly separated between clusters 2 and 3, also indicating that the clustering may have not converged to a stable description of the data. As we can see on Table 1 for cases A to C, increasing the number of knots (from 7 to 15) improves the clustering quality, but lowers the reconstruction loss. Even if the detection results are highly similar between cases A to C, we consider 11 knots to be a good trade-off between a high clustering quality and a reasonable reconstruction loss. In any case the precursory signals are always recovered even with a small amount of data, a clear advantage of our clustering procedure over clustering strategies based on classical deep convolutional neural networks.

We then conduct 3 additional tests onto daylong data, where the number of knots is fixed to 11, and where we investigate the pooling factor of the scattering layer which defines a trade-off between the stability of the scattering coefficients and the final time resolution of the analysis. A very large pooling value (case F) could lead to a degraded time resolution, but will still be able to detect seismic events that are very localized in time, and therefore the number of clusters is similar in cases D and F because the pooling factor is large enough. In contrast, a smaller pooling could lead to a smaller time resolution, without being stable enough for clustering (case E). With this choice of pooling factor, we observe that a larger number of clusters are kept after training with, which is a sign of instability. The clustering loss is high (3.67) in comparison with other clustering results. The pooling factor therefore must be chosen with respect to the maximal duration of interest, and should be maximized if no *a priori* on the signal in search is available.

The case D presented in detail in the present study (Figs. 3 and 4) has an intermediate pooling factor leading to a $\sim$32-sec final time resolution with three layers. In addition, we tested in case D a larger number of octaves and wavelets per octaves at each layer. This test presents the lowest clustering and reconstruction losses, which is mostly due to the presence of more filters at each layer to describe the data. Note that increasing the number of wavelet per octave do not change the number of parameters to be optimized in the learning procedure since the filter bank of each layer is derived from the learnable mother wavelet only.

Comparison of cluster detection rates and microseismic energy

We collect the spectral pressure calculated from the WAVEWATCH III model (CIET ARDHUIN) on a 0.5×0.5 degree grid globally, from 2017-06-01 to 2017-06-18. This pressure data cannot be directly used as a proxy for radiated seismic energy, because the radiation of body and surface waves depends on the bathymetric profile of the seefloor [45]. According to [45], the equivalent radiated spectral energy can be derived from the pressure with taking into account the resonnance of the water column at each point of the grid as amplification factor. We therefore used the amplification model presented in [45], where the global bathymetry is taken into account. We then considered the source time function of each points of a 4×4 degree grid, and correlated it with the temporal within-cluster detection. Because the pressure data is availble every 3 hours, we decimated the within-cluster detection on the same time basis.

The correlation is tested for several frequency bands (0.1 to 0.2, 0.2 to 0.3, 0.3 to 0.45 and 0.45 to 0.6) and seismic waves (P waves, S waves and Rayleigh waves). For each frequency band, the maximally correlated source time function and seismic wave type is identified and represented in Fig. S2 and S3. In addition to the water-column resonnance amplifications, we also apply different corrections for the different seismic wave types. The P-wave spectral energy is corrected from the shadowing of the Earth's core (no energy should be recorded between 104 and 140 degrees of epicentral distance). This first correction is applied as a mask on the correlation coefficients between within-cluster detection rates and source time functions. For Rayleigh waves, we also took into account the strong attenuation effects of the crust heterogeneities at these frequencies. We here considered an exponentially decaying attenuation with distance, with a decay of $1/500$ km^{-1}.

References

[1] Bergen, K. J., Johnson, P. A., Maarten, V. & Beroza, G. C. Machine learning for data-driven discovery in solid earth geoscience. *Science* **363**, eaau0323 (2019).

[2] Obara, K., Hirose, H., Yamamizu, F. & Kasahara, K. Episodic slow slip events accompanied by non-volcanic tremors in southwest japan subduction zone. *Geophysical Research Letters* **31** (2004).

[3] Perol, T., Gharbi, M. & Denolle, M. Convolutional neural network for earthquake detection and location. *Science Advances* **4**, e1700578 (2018).

[4] Ross, Z. E., Meier, M.-A., Hauksson, E. & Heaton, T. H. Generalized seismic phase detection with deep learning. *arXiv preprint arXiv:1805.01075* (2018).

[5] Scarpetta, S. *et al.* Automatic classification of seismic signals at mt. vesuvius volcano, italy, using neural networks. *Bulletin of the Seismological Society of America* **95**, 185–196 (2005).

[6] Esposito, A. M., D'Auria, L., Giudicepietro, F., Caputo, T. & Martini, M. Neural analysis of seismic data: applications to the monitoring of mt. vesuvius. *Annals of Geophysics* (2013).

[7] Maggi, A. *et al.* Implementation of a multistation approach for automated event classification at piton de la fournaise volcano. *Seismological Research Letters* **88**, 878–891 (2017).

[8] Malfante, M. *et al.* Machine learning for volcano-seismic signals: Challenges and perspectives. *IEEE Signal Processing Magazine* **35**, 20–30 (2018).

[9] Esposito, A. *et al.* Unsupervised neural analysis of very-long-period events at stromboli volcano using the self-organizing maps. *Bulletin of the Seismological Society of America* **98**, 2449–2459 (2008).

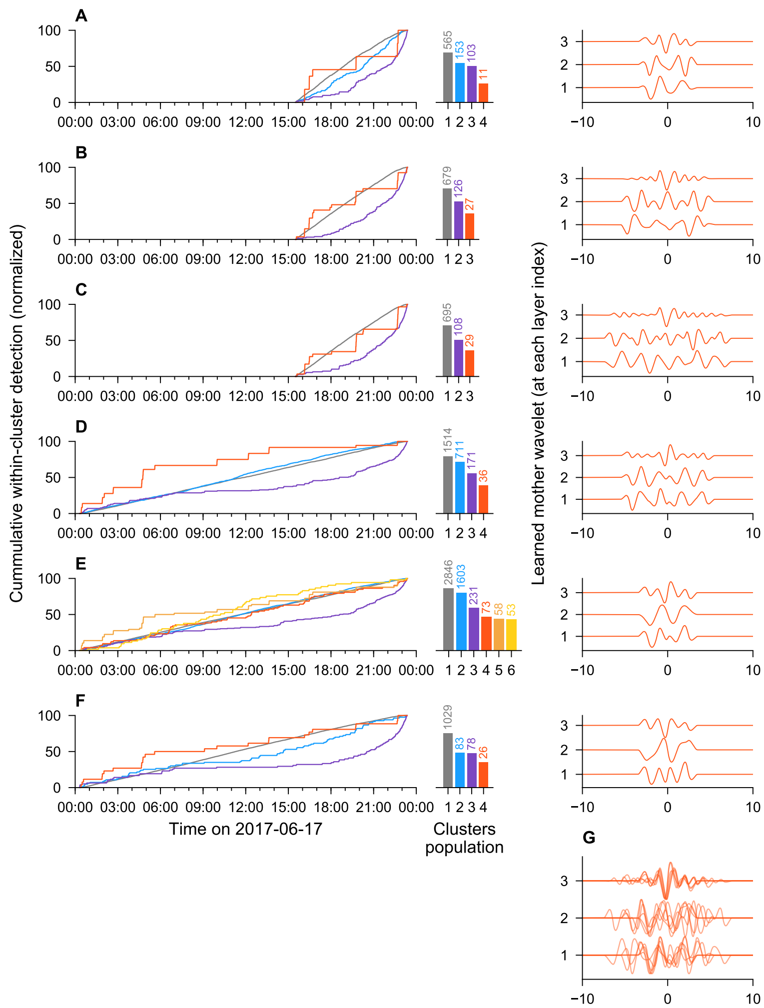

Figure S7: **(Supplementary material) Learning results with different parameters.** The different parameter sets are given in Table 1. The left and middle plots respectively show the within-cluster cumulative detections and the within-cluster number of samples after 10,000 training epochs. The right plots show the final learned wavelets at each layer. (**A** − **F**) results obtained with the parameters sets given in Table 1, (D) being the case analyzed in details in Fig. 3 and 4. (**G**) learned mother wavelet at each layer with all parameter sets.

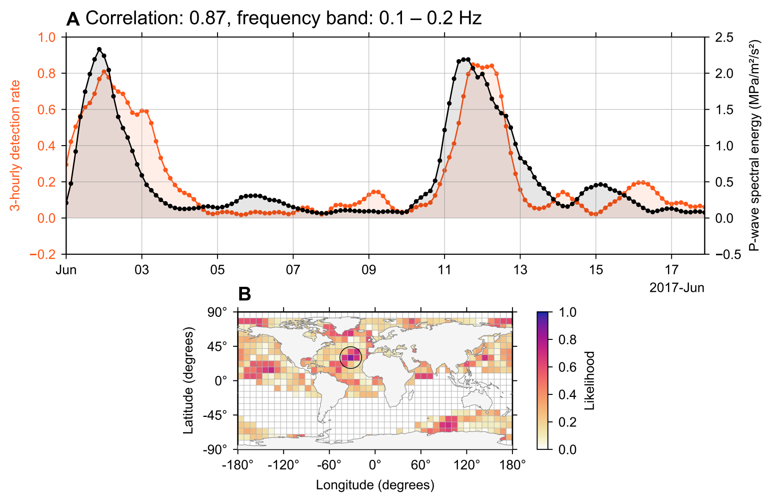

Figure S8: **(Supplementary material) Comparison of clusters D with P-wave microseismic energy.** (**A**) The whithin-cluster 3-hourly detection is presented in red curve over 17 days of 3-components seismic data. The best-matching radiated P-wave spectral energy in the frequency band 0.1 to 0.2 Hz is presented in black line. (**B**) Global matching likelihood of the spectral P-wave radiated energy between 0.1 and 0.2 Hz on a 4×4 degrees grid. The likelihood is corrected for theoretical P-wave shadow zones due to the presence of the core (between 104 and 140 degrees of epicentral distance), visible by the zero-likelihood zone. The highest likelihood from which the source-time function is extracted and presented in A is highlighted with a black circle in B.

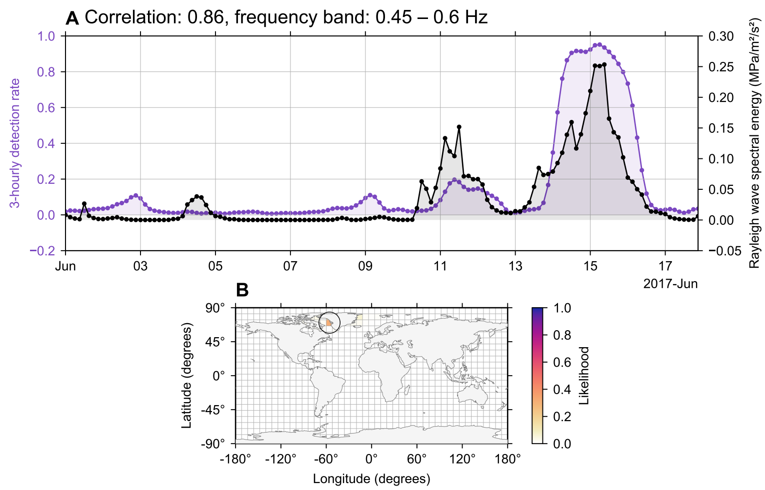

Figure S9: **(Supplementary material) Comparison of clusters C with Rayleigh wave microseismic energy.** (**A**) The whithin-cluster 3-hourly detection is presented in purple curve over 17 days of 3-components seismic data. The best-matching radiated Rayleigh-wave spectral energy in the frequency band 0.45 to 0.6 Hz is presented in black line. (**B**) Global matching likelihood of the spectral Rayleigh-wave radiated energy between 0.45 to 0.6 Hz on a 4×4 degrees grid. The likelihood is corrected from theoretical Rayleigh wave attenuation due to strong scattering at these frequencies. The highest likelihood from which the source-time function is extracted and presented in A is highlighted with a black circle in B.

[10] Unglert, K. & Jellinek, A. Feasibility study of spectral pattern recognition reveals distinct classes of volcanic tremor. *Journal of Volcanology and Geothermal Research* **336**, 219–244 (2017).

[11] Hammer, C., Ohrnberger, M. & Faeh, D. Classifying seismic waveforms from scratch: a case study in the alpine environment. *Geophysical Journal International* **192**, 425–439 (2012).

[12] Soubestre, J. *et al.* Network-based detection and classification of seismovolcanic tremors: Example from the klyuchevskoy volcanic group in kamchatka. *Journal of Geophysical Research: Solid Earth* **123**, 564–582 (2018).

[13] Beyreuther, M., Hammer, C., Wassermann, J., Ohrnberger, M. & Megies, T. Constructing a hidden markov model based earthquake detector: application to induced seismicity. *Geophysical Journal International* **189**, 602–610 (2012).

[14] Holtzman, B. K., Paté, A., Paisley, J., Waldhauser, F. & Repetto, D. Machine learning reveals cyclic changes in seismic source spectra in geysers geothermal field. *Science advances* **4**, eaao2929 (2018).

[15] Yoon, C. E., O'Reilly, O., Bergen, K. J. & Beroza, G. C. Earthquake detection through computationally efficient similarity search. *Science advances* **1**, e1501057 (2015).

[16] Mousavi, S. M., Zhu, W., Ellsworth, W. & Beroza, G. Unsupervised clustering of seismic signals using deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters* (2019).

[17] Köhler, A., Ohrnberger, M. & Scherbaum, F. Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps. *Geophysical Journal International* **182**, 1619–1630 (2010).

[18] Rouet-Leduc, B. *et al.* Machine learning predicts laboratory earthquakes. *Geophysical Research Letters* **44**, 9276–9282 (2017).

[19] Bruna, J. & Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence* **35**, 1872–1886 (2013).

[20] Andén, J. & Mallat, S. Deep scattering spectrum. *IEEE Transactions on Signal Processing* **62**, 4114–4128 (2014).

[21] Andén, J. & Mallat, S. Scattering representation of modulated sounds. *15th DAFx* **9** (2012).

[22] Peddinti, V. *et al.* Deep scattering spectrum with deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 210–214 (IEEE, 2014).

[23] Balestriero, R., Cosentino, R., Glotin, H. & Baraniuk, R. Spline filters for end-to-end deep learning. In Dy, J. & Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, 364–373 (PMLR, Stockholmsmässan, Stockholm Sweden, 2018).

[24] Ahuja, N., Lertrattanapanich, S. & Bose, N. Properties determining choice of mother wavelet. *IEE Proceedings-Vision, Image and Signal Processing* **152**, 659–664 (2005).

[25] Meyer, Y. *Wavelets and operators*, vol. 1 (Cambridge university press, 1992).

[26] Coifman, R. R. & Wickerhauser, M. V. Entropy-based algorithms for best basis selection. *IEEE Transactions on information theory* **38**, 713–718 (1992).

[27] Poli, P. Creep and slip: Seismic precursors to the nuugaatsiaq landslide (greenland). *Geophysical Research Letters* **44**, 8832–8836 (2017).

[28] Chao, W.-A. *et al.* The large greenland landslide of 2017: Was a tsunami warning possible? *Seismological Research Letters* (2018).

[29] Bell, A. F. Predictability of landslide timing from quasi-periodic precursory earthquakes. *Geophysical Research Letters* **45**, 1860–1869 (2018).

[30] Allen, R. Automatic phase pickers: Their present use and future prospects. *Bulletin of the Seismological Society of America* **72**, S225–S242 (1982).

[31] Gibbons, S. J. & Ringdal, F. The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International* **165**, 149–166 (2006).

[32] Brown, J. R., Beroza, G. C. & Shelly, D. R. An autocorrelation method to detect low frequency earthquakes within tremor. *Geophysical Research Letters* **35** (2008).

[33] Estivill-Castro, V. Why so many clustering algorithms: a position paper. *SIGKDD explorations* **4**, 65–75 (2002).

[34] Chakraborty, A. & Okaya, D. Frequency-time decomposition of seismic data using wavelet-based methods. *Geophysics* **60**, 1906–1916 (1995).

[35] Dumoulin, V. & Visin, F. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* (2016).

[36] Shelly, D. R., Beroza, G. C. & Ide, S. Non-volcanic tremor and low-frequency earthquake swarms. *Nature* **446**, 305 (2007).

[37] Reynolds, D. Gaussian mixture models. *Encyclopedia of biometrics* 827–832 (2015).

[38] Mallat, S. *A wavelet tour of signal processing: the sparse way*, chap. 4, 111–112 (Elsevier, 2009), 3 edn.

[39] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[40] Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967).

[41] Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing* **17**, 395–416 (2007).

[42] Sifre, L., Kapoko, M., Oyallon, E. & Lostanlen, V. Scatnet: a matlab toolbox for scattering networks (2013).

[43] Xu, L. & Jordan, M. I. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation* **8**, 129–151 (1996).

[44] Ardhuin, F. *et al.* Ocean wave sources of seismic noise. *Journal of Geophysical Research: Oceans* **116(C9)** (2011).

[45] Li, L., Boue, P. & Campillo, M. Spatiotemporal connectivity of noise-derived seismic body waves with ocean waves and microseism excitations *Eartharxiv* (2019).