



**HAL**  
open science

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

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

## ► To cite this version:

Léonard Seydoux, Randall Balestrieri, Piero Poli, Maarten De Hoop, Michel Campillo, et al.. Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature Communications*, 2020, 11 (1), 10.1038/s41467-020-17841-x . hal-03411505v1

**HAL Id: hal-03411505**

<https://hal.univ-grenoble-alpes.fr/hal-03411505v1>

Submitted on 3 Sep 2020 (v1), last revised 2 Nov 2021 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3 L. Seydoux<sup>1</sup>, R. Balestriero<sup>2</sup>, P. Poli<sup>1</sup>, M. de Hoop<sup>3</sup>, M. Campillo<sup>1</sup>,  
4 and R. Baraniuk<sup>2</sup>

5 <sup>1</sup>Institut des sciences de la Terre, Université Grenoble-Alpes, UMR CNRS 5375, France

6 <sup>2</sup>Electrical & Computational Engineering, Rice University, Houston, TX, 77005, USA

7 <sup>3</sup>Computational & Applied Mathematics, Rice University, Houston, TX, 77005, USA

8 May 19, 2020

9 **Abstract**

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

## 25 **Introduction**

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

38 Recently, techniques from machine learning have opened up new avenues for  
39 rapidly exploring large seismic datasets with minimum a priori knowledge. Ma-  
40 chine learning algorithms are data-driven tools that approximate non-linear re-  
41 lationships between observations and labels (supervised learning) or that reveal  
42 patterns from unlabeled data (unsupervised learning). Supervised algorithms  
43 rely on the quality of the predefined labels, often obtained via classical algo-  
44 rithms [3, 4] or even manually [5, 6, 7, 8]. Inherently, supervised strategies are  
45 used to learn how to detect or classify specific classes of already-known signals  
46 and, therefore, cannot be used for discovering new classes of seismic signals.  
47 Seismology can significantly benefit from the development of unsupervised ap-  
48 proaches because the data is mostly unlabeled. Unsupervised tools are likely the  
49 best candidates to explore seismic data without the need for any explicit signal  
50 model and hence, discover new classes of seismic signals. For this reason, un-  
51 supervised methods are more relevant for seismology, where the data is largely  
52 unlabeled and new classes of seismic signals should be sought. While supervised  
53 strategies are often easier to implement thanks to the evaluation of a prediction  
54 error, unsupervised strategies mostly rely on implicit models that are challeng-

55 ing to design. Unsupervised-learning based studies have mostly been applied to  
56 data from volcano monitoring systems, where a large variety of seismo-volcanic  
57 signals is usually observed [9, 10, 11, 12]. Some unsupervised methods have also  
58 been recently applied to induced seismicity [13, 14], global seismicity [15], and  
59 local-vs-distance earthquakes [16]. In both cases (supervised or unsupervised),  
60 the keystone to success lies in the data representation namely, we need to define  
61 an appropriate set of relevant waveform features for solving the task of interest.  
62 The features can be manually defined [17, 7, 18] or learned with appropriate  
63 techniques such as artificial neural networks [5, 3], the latter belonging to the  
64 field of deep learning.

65 In this paper, we develop a new unsupervised deep-learning method for clus-  
66 tering signals in continuous multichannel seismic time-series. Our strategy com-  
67 bines a deep scattering network [19, 20] for automatic feature extraction and  
68 a Gaussian mixture model for clustering. Deep scattering networks belong to  
69 the family of deep convolutional neural networks, where the convolutional filters  
70 are restricted to wavelets and with modulus activations [19]. The restriction to  
71 wavelets filters allows the deep scattering networks to have explicit and physics-  
72 related properties that greatly simplifies the architecture design in contrast with  
73 classical deep convolutional neural network (frequency band, time scales of in-  
74 terest, amplitudes). Scattering networks have shown to perform high-quality  
75 classification of audio signals [21, 20, 22] and electrocardiograms [23]. A deep  
76 scattering network decomposes the signal’s structure through a tree of wavelet  
77 convolutions, modulus operations and average-pooling, providing a stable rep-  
78 resentation at multiple time and frequency scales [20]. The resulting represen-  
79 tation is particularly suitable for discriminating complex seismic signals that  
80 may differ in nature (source and propagation effects) with several order of dif-  
81 ferent durations, amplitudes and frequency contents. After decomposing the  
82 time series with the deep scattering network, we exploit the representation in  
83 a two-dimensional feature space that results from a dimension reduction for vi-  
84 sualization and hence, interpretation purposes. The two-dimensional features  
85 are finally fed to a Gaussian mixture model for clustering the different time

86 segments.

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

## 105 Results

106 **Seismic records of the 2017 Nuugaatsiaq landslide.** We apply our strat-  
107 egy for blindly clustering and detecting the low-amplitude precursory seismic-  
108 ity [27] to the June 2017 landslide that occurred near Nuugaatsiaq, Greenland  
109 at 23:39 UTC. The volume of the rockfall was estimated between 35 to 51 mil-  
110 lion cubic meters by differential digital elevation models, forming thus a massive  
111 landslide [28]. This landslide triggered tsunami waves that impacted the small  
112 town of Nuugaatsiaq and caused four reported injuries [28].

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

115 (Fig. 1A). We select the daylong three-component seismograms from 2017-06-17  
116 00:00 to 2017-06-17 23:38 in order to remove the signal due to the mainshock  
117 and focus on the content of the seismic wavefield recorded before. A detailed  
118 inspection of the east component records revealed that a small event was occur-  
119 ring repetitively before the landslide, starting approximately 9 hours before the  
120 rupture and accelerating over time [27, 29]. The accelerating behavior of this  
121 seismicity suggests that an unstable initiation was at work before the massive  
122 landslide. This signal is not directly visible in raw seismic records; it is of very  
123 weak amplitude, has a smooth envelope, and most of its energy located in be-  
124 tween 2 and 8 Hz (see a zoom onto the last 4 hours of data in Fig. 1B) and was  
125 first highlighted with three-component template matching [27]. While some of  
126 these events may be visible in the seismograms filtered between 2 and 9 Hz at  
127 times close to the landslide, a large part are hidden in the background noise [27].  
128 The structure of such signal makes it hard to detect via traditional detection  
129 algorithms such as STA/LTA (the ratio between Short-Term Average and the  
130 Long-Term Average of the seismic signal [30]), because they are sensitive to  
131 brutal signal changes with decent signal-to-noise ratios [15]. Besides, STA/LTA  
132 only delivers an information about the presence of a signal in the continuous  
133 trace without any clue of the similarity with other signals, which is our primary  
134 goal here.

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

146 tions [15].

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

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

174 We define a deep scattering network as the sequence of convolution-modulus  
175 operations performed at higher orders, allowing to scatter the signal structure

176 through the tree of time and frequency analyses. We finally obtain a locally  
177 invariant signal representation by applying an average-pooling operation to the  
178 all-order pooling layers [19, 21, 20]. This pooling operation is adapted for con-  
179 catenation, with an equal number of time samples at each layer (Fig. 2). The  
180 scattering coefficients are invariant to local time translation, small signal de-  
181 formations and signal overlapping. They incorporate multiple time scales (at  
182 different layers) and frequencies scales (different wavelets). The tree of op-  
183 erations performed in a scattering network forms a deep convolutional neural  
184 networks, where the convolutional filters are restricted to wavelets, and where  
185 the activation function is the modulus operator [19]. Scattering networks are  
186 located in between (1) classical time and frequency analysis routinely applied  
187 in seismology that is often limited to a typical time scale, and (2) deep con-  
188 volutional neural networks where the unconstrained filters are often hard to  
189 interpret, and where the network architecture is often challenging to define. In  
190 contrast, deep scattering networks can be designed in a straightforward way,  
191 thanks to the analytic framework defined in [19].

192 From one layer to another, we increase the frequency range of the filter banks  
193 in order to consider at the same time small-duration details of the waveform,  
194 and larger-duration histories (see Table 1, case D for the selected architecture in  
195 the present study). The number of wavelets per octaves and number of octaves  
196 defines the frequency resolution and bandwidth of each layer, and the depth  
197 (total number of layers) of the scattering network controls the final temporal  
198 resolution of the analysis. Following the recommendations cross-validated onto  
199 audio signal classification [20], we use a large number of filters at the first layer,  
200 and we gradually increase the number of octaves while reducing the number of  
201 wavelets per octave from the first to the last layer (Table 1, case D). That way,  
202 the representation is highly redundant at the layer `conv1` and gets sparser at  
203 the higher-order layers `conv2` and `conv3`, where fewer filters are used at each  
204 frequency to decompose the signal. This has the main effect of improving the  
205 contrast between signals of different nature [20]. We finally choose the network  
206 depth based on the range of time scales of interest. In the present study, we aim



207 at investigating mostly impulsive earthquake-like signals that may last between  
208 several seconds to less than one minute. A deeper scattering network could be  
209 of interest in order to analyze the properties of longer-duration signals such  
210 as seismic tremors [36] or background seismic noise. Finally, with our choice of  
211 pooling factors, we obtain a temporal resolution of 35 seconds for each scattering  
212 coefficient.

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

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

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

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

266       The clustering loss decreases with the learning epochs (Fig. 3C). We declare  
267 the clustering to be optimal when the loss stagnates (reached after approxi-

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

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

285 **Analysis of clusters.** An analysis of the temporal evolution of the clusters  
286 is presented in Fig. 4. The within-cluster cumulative detections obtained af-  
287 ter training (epoch 10,000) are presented in Fig. 4A for clusters 1 and 2, and  
288 in Fig. 4B for clusters 2 and 3. The two most populated clusters 1 and 2  
289 (Fig. 4A) gather more than 90% of the overall data (observed on the histograms  
290 in Fig. 3B). They both show a linear detection rate over the day with no par-  
291 ticular concentration in time and, therefore, relate to the background seismic  
292 noise. Clusters 3 and 4 (Fig. 4B) show different non-linear trends that include  
293 10% of the remaining data.

294 The temporal evolution of cluster 4 is presented in Fig. 4B. The time seg-  
295 ments that belong to cluster 4 are extracted and aligned to a reference time  
296 segment (at the top) with local cross-correlation for better readability (see fur-  
297 ther details about the strategy in the supplementary materials). We see that

298 these time segments contain seismic events localized in time with relatively high  
299 signal-to-noise ratio and sharp envelope. These events do not show a strong  
300 similarity in time, but they strongly differ from the event belonging to other  
301 clusters, explaining why they have been gathered in the same cluster. The de-  
302 tection rate is sparse in time, indicating that cluster 4 is mostly related to a  
303 random background seismicity or other signals which interest is beyond the  
304 scope of the present manuscript.

305 The temporal evolution of cluster 3 shows three behaviors. First, we observe  
306 a nearly-constant detection rate from the beginning of the day to approximately  
307 07:00. Second, the detection rate lowers between 07:00 and 13:00 where only 4%  
308 of the within-cluster detections are observed. An accelerating seismicity is finally  
309 observed from 13:00 up to the landslide time (23:39 UTC). The time segments  
310 belonging to cluster 3 are reported on Fig. 4D in gray colorscale, and aligned  
311 with local cross-correlation with a reference (top) time segment. The correlation  
312 coefficients obtained for the time lag that maximizes the alignment are indicated  
313 in orange color in Fig. 4E. As with the template matching strategy, we clearly  
314 observe the increasing correlation coefficient with the increasing event index [27],  
315 indicating that the signal-to-noise ratio increases towards the landslide rupture.  
316 This suggests that the repeating event may still exist earlier in the data even  
317 before 15:00, but that the detection threshold of the template matching method  
318 is limited by the signal-to-noise ratio [27]. In contrast, we observe that the  
319 probability of these 171 events remains high in our approach, with 97% of the  
320 precursory events previously found [27] recovered.

321 A interesting observation is the change of behavior in the detection rate of  
322 this cluster at nearly 07:00 (Fig. 4B). The events that happened before 07:00  
323 have all a relatively high probability to belong to cluster 3, refuting the hy-  
324 pothesis that noise samples have randomly been misclassified by our strategy  
325 (Fig. 4E). The temporal similarity of all these events in Fig. 4D is particularly  
326 visible for later events (high index) because the signal-to-noise ratio of these  
327 events increases towards the landslide [27]. The two trends may be whether  
328 related to similar signals generated at same position (same propagation) with a

329 different source, or by two types of alike-looking events that differ in nature, but  
330 that may have been gathered in the same cluster because they strongly differ  
331 from the other clusters. This last hypothesis can be tested with using hierarchi-  
332 cal clustering [40]. Our clustering procedure highlighted those 171 similar events  
333 in a totally unsupervised way, without the need of defining any template from  
334 the seismic data. The stack of the 171 waveforms is shown in black solid line in  
335 Fig. 4D, indicating that the template of these events is defined in a blind way  
336 thanks to our procedure. In addition, these events have very similar properties  
337 (duration, seismic phases, envelope) in comparison with the template defined  
338 in [27].

## 339 Discussion and conclusions

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

355 Our strategy is capable of blindly recovering the small-amplitude precur-  
356 sory signal reported in [27, 29]. This indicates that waveform templates can  
357 be recovered from our method without the need of any manual inspection of

358 the seismic data prior to the clustering process, and tedious selection of wave-  
359 form template in order to perform high-quality detection. Such unsupervised  
360 strategy is of strong interest in the exploration of seismic datasets, where the  
361 structure of seismic signals can be complex (low-frequency earthquakes, non-  
362 volcanic tremors, distant vs. local earthquakes, etc.), and where some class of  
363 unknown signals is likely to be disregarded by a human expert.

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

387 Besides being adapted to small amount of data, our strategy can also work  
388 with large amount of data, as scalability is guaranteed by batch processing, and

389 using only small-complexity operators (convolution and pooling). Indeed, batch  
390 processing allows to control the amount of data seen by the scattering network  
391 and GMM at a single iteration, each epoch being defined when the whole dataset  
392 have been analyzed by the algorithm. There is no limitation to the total amount  
393 of data being analyzed because only the selected segments at each iteration are  
394 fed to the network. At longer time scales, the number of clusters needed to fit  
395 the seismic data must change, however, with an expectation that the imbalance  
396 between clusters should increase. We illustrate this point another experiment  
397 performed on the continuous seismogram recorded at the same station over 17  
398 days, including the date of the landslide (from 2017-06-01 to 2017-06-18). With  
399 this larger amount of data, the clustering procedure still converges and exhibit  
400 9 new clusters. The hourly within-clusters detections of these new clusters  
401 are presented in Fig. 5. Among the different clusters found by our strategy,  
402 we observe that more than 93% of the data is identified in slowly evolving  
403 clusters, most likely related to fluctuations of the ambient seismic noise (Fig. 5,  
404 clusters A to E). The most populated clusters (A and B) occupy more than  
405 61% of the time, and are most likely related to diffuse wavefield without any  
406 particular dominating source. Interestingly, we observe two other clusters with  
407 large population with a strong localisation in time (clusters C and D in Fig. 5).  
408 A detailed analysis of the ocean-radiated microseismic energy [44, 45] allowed us  
409 to identify the location and dominating frequency of the sources responsible for  
410 these clusters to be identified (illustrated in Fig. S2 and S3 in the supplementary  
411 material). The source time function of the best-matching microseismic sources  
412 have been reported on clusters C and D in Fig. 5.

413 Compared with these long-duration clusters, the clustering procedure also  
414 reports very sparse clusters where less than 7% of the seismic data is present.  
415 Because of clustering instabilities caused by the large class imbalance of the seis-  
416 mic data, we decided to perform a second-order clustering on the low-populated  
417 clusters. This strategy follows the idea of hierarchical clustering [40], where  
418 the firstly identified clusters are analyzed several consecutive times in order to  
419 discover within-cluster families. For the sake of brevity, we do not intend to per-

420 form a deep-hierarchical clustering in the present manuscript, but to illustrate  
421 the potential strength of such strategy in seismology, where the data is essen-  
422 tially class-imbalanced. We perform a new clustering from the data obtained in  
423 the merged low-populated clusters (F to I in Fig. 5). This additional clustering  
424 procedure detected two clusters presented in Fig. 6A. These two clusters have  
425 different temporal cumulated detections and exhibits different population sizes.  
426 A zoom of the cumulated within-cluster detections is presented in Fig. 6B, and  
427 show a high similarity with clusters 3 and 4 previously obtained in Fig. 3 from  
428 the daylong seismogram. This result clearly proves that the accelerating pre-  
429 cursor is captured by our strategy even when the data is highly imbalanced. If  
430 the scattering network provide highly relevant features, clustering seismic data  
431 with simple clustering algorithms can be a hard task that can be solved with hi-  
432 erarchical clustering, as illustrated in the present study. This problem can also  
433 be better tackled by other clustering algorithms such as spectral clustering [41]  
434 which has the additional ability to detect outliers. Clustering the outlier signals  
435 may then be an alternative to GMM in that case. Another possibility would be  
436 to use the local similarity search with hashing functions [15] in order to improve  
437 our detection database onto large amount of seismic data.

438 The structure of the scattering network shares some similarities with the  
439 FAST algorithm (for Fingerprint And Similarity Search [15]) from a architec-  
440 tural point of view. FAST uses a suite of deterministic operations in order to  
441 extract waveforms features and feed it to a hashing system in order to per-  
442 form a similarity search. The features are extracted from the calculation of  
443 spectrogram, Haar wavelet transforms and thresholding operations. While be-  
444 ing similar, the FAST algorithm involves a number of paramaters that are not  
445 connected to the underlying physics. For instance, the thresholding operation  
446 has to be manually inspected [15], as well as the size of the analyzing window.  
447 In comparison, our alrotihm’s parameters are based on physical intuition, and  
448 does not imply any signal windowing (only the resolution of the final result can  
449 be controlled). FAST is not a machine learning strategy because no learning  
450 is involved; in contrast, we do learn the representation of the seismic data that



451 best solves the task of clustering. While FAST needs a large amount of data to  
452 be run in an optimal way [15], our algorithm still works with a few number of  
453 samples.

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

## 463 **Aknowldgements**

464 L.S., P.P. and M.C. acknowledge support from the European Research Council under  
465 the European Union Horizon 2020 research and innovation program (grant agree-  
466 ment no. 742335, F-IMAGE). M.V.d.H. gratefully acknowledges support from the  
467 Simons Foundation under the MATH + X program and from NSF under grant DMS-  
468 1815143. R.B. and R.G.B. were supported by NSF grants IIS-17-30574 and IIS-18-  
469 38177, AFOSR grant FA9550-18-1-0478, ONR grant N00014-18-12571, and a DOD  
470 Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047. L.S. thanks Ro-  
471 main Cosentino for very helpful discussions and comments. The facilities of IRIS  
472 Data Services, and specifically the IRIS Data Management Center, were used for  
473 access to waveforms and related metadata used in this study. IRIS Data Services  
474 are funded through the Seismological Facilities for the Advancement of Geoscience  
475 and EarthScope (SAGE) Project funded by the NSF under Cooperative Agreement  
476 EAR-1261681. The authors declare that they have no competing financial inter-  
477 ests. Correspondence and requests for materials should be addressed to L.S. (email:  
478 leonard.seydoux@univ-grenoble-alpes.fr).

<sup>479</sup> **Figures and tables**

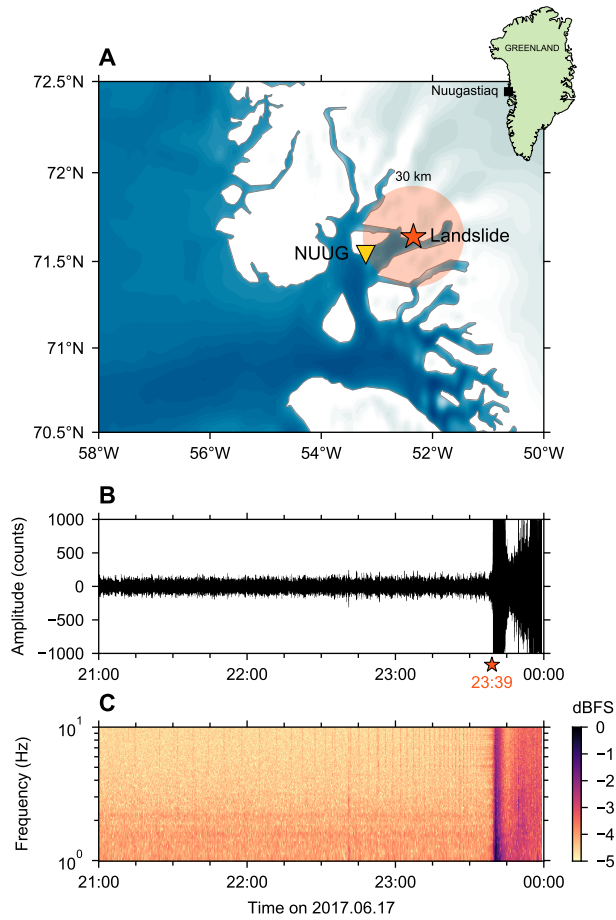


Figure 1: **Geological context and seismic data.** **A** Location of the landslide (red star) and the seismic station NUUG (yellow triangle). The seismic station is located in the vicinity of the small town of Nuugaatsiaq, Greenland (top-right inset). **B** Raw record of the seismic wavefield collected 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.

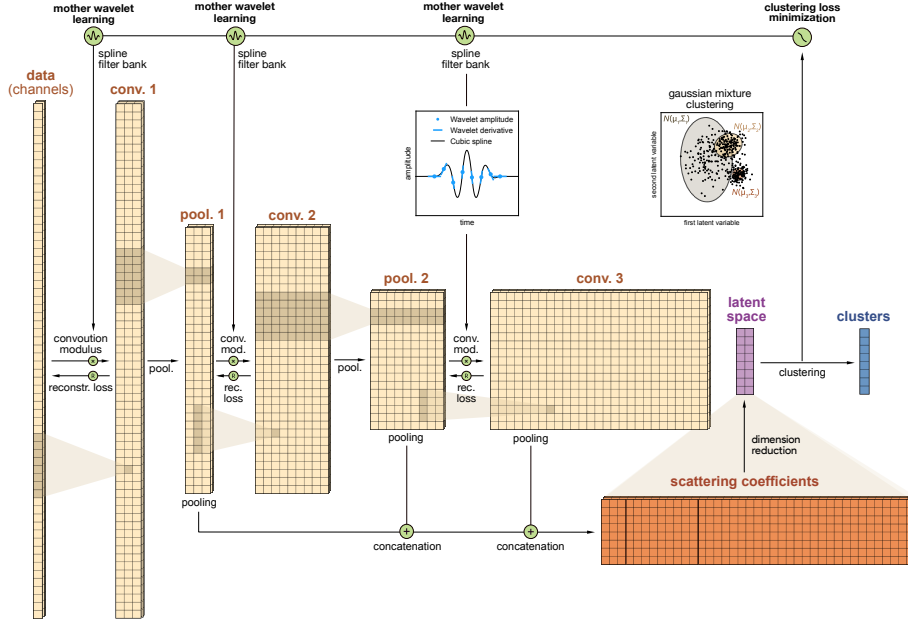
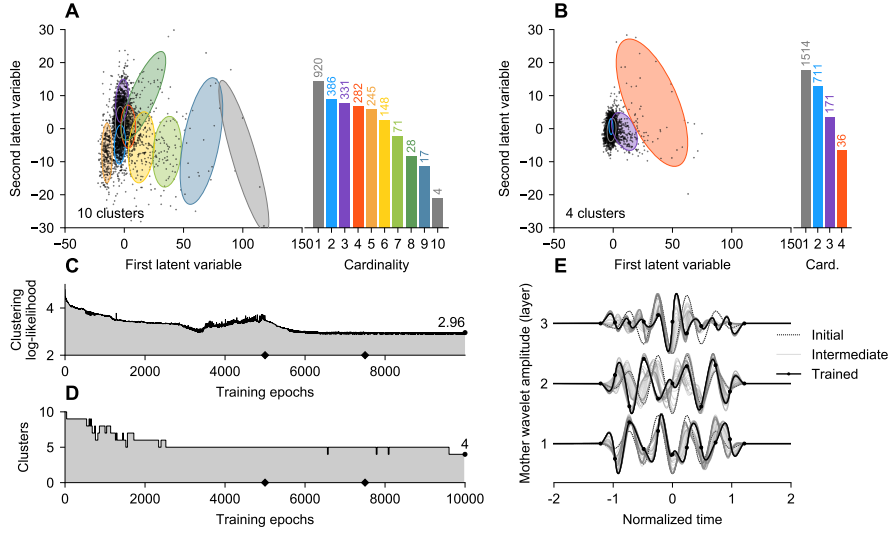


Figure 2: **Deep learnable scattering network with Gaussian mixture model clustering.** The network consists in a tree of convolution and modulus operations successively applied to the multichannel time series (layers conv 1 – 3). A reconstruction loss is calculated at each layer in order to constrain 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  $\ell$  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.

Title	Data		Scattering network				Learning		
	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 → 4	3.79	4.20
B	15:00	23:30	3, 6, 6	8, 2, 1	11	$2^{10}$	10 → 3	3.42	5.40
C	15:00	23:30	3, 6, 6	8, 2, 1	15	$2^{10}$	10 → 3	3.17	5.49
* D	<b>00:30</b>	<b>23:30</b>	<b>4, 6, 6</b>	<b>8, 4, 3</b>	<b>11</b>	<b><math>2^{10}</math></b>	<b>10 → 4</b>	<b>2.96</b>	<b>3.06</b>
E	00:30	23:30	3, 6, 6	8, 2, 1	11	$2^9$	10 → 6	3.67	1.76
F	00:30	23:30	3, 6, 6	8, 2, 1	11	$2^{11}$	10 → 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.

## 480 Supplementary materials

### 481 Deep scattering network

482 A complex wavelet  $\psi \in \mathcal{L}$  is a filter localized in frequency with zero average,  
483 center frequency  $\omega_0$  and bandwidth  $\delta\omega$ . We define the functional space  $\mathcal{L}$  of any  
484 complex wavelet  $\psi$  as

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

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

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

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

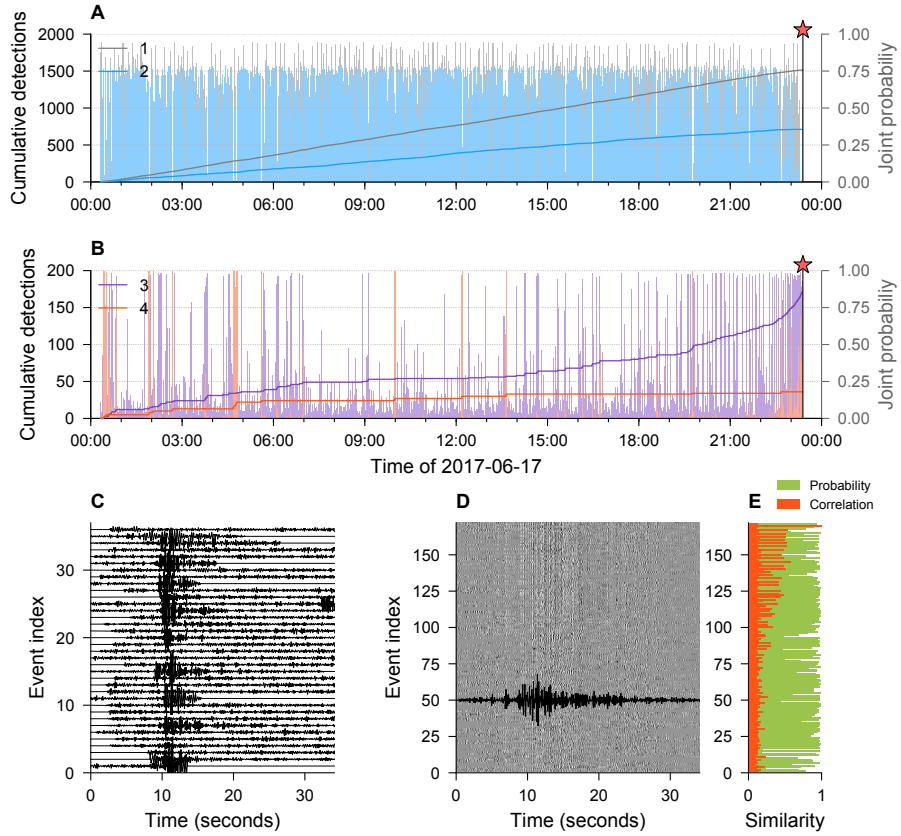


Figure 4: **Analysis of clusters in the time domain.** Within-cluster cumulative number of detection of events in clusters 1 and 2 (A) and clusters 3 and 4 (B) at epoch 10,000. The relative probability for each time window to belong to each cluster is represented with lighter bars. The waveforms extracted 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.

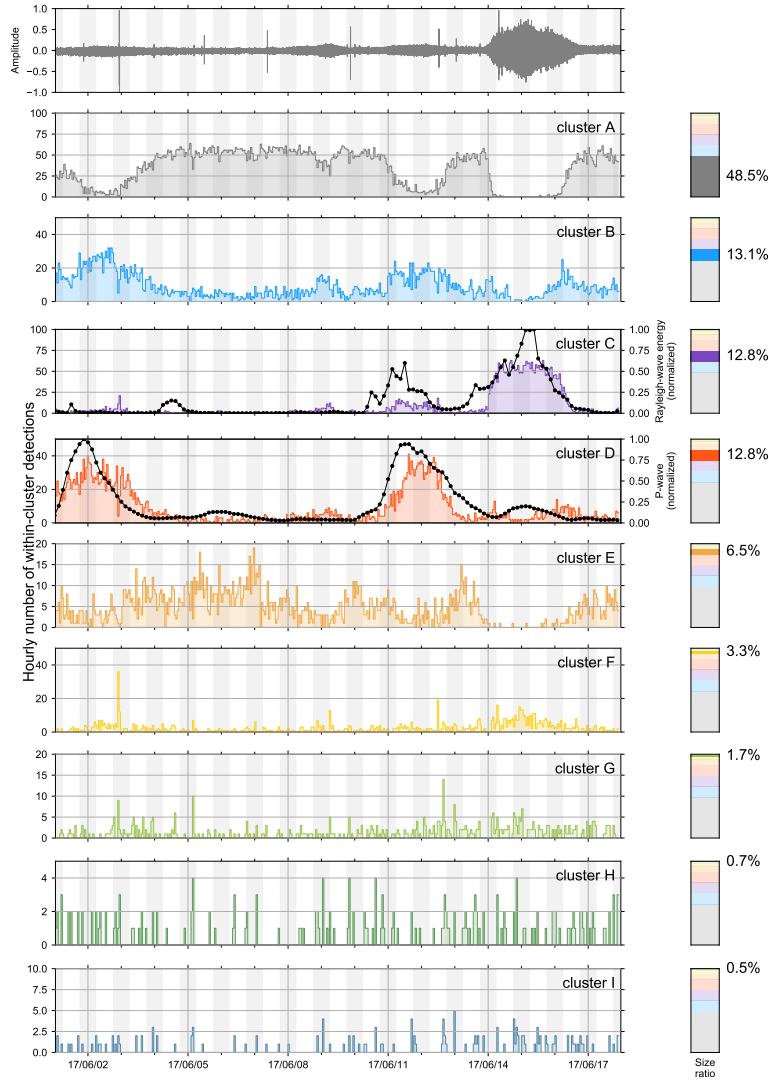


Figure 5: **Clustering results obtained long-duration seismic data.** The broadband seismogram recorded by the station NUUG (Fig. 1) from 2017-06-01 to 2017-06-18 is presented in the top plot. The hourly within-cluster detection rate is presented for each of the 9 clusters (A to I). The right-hand side 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).



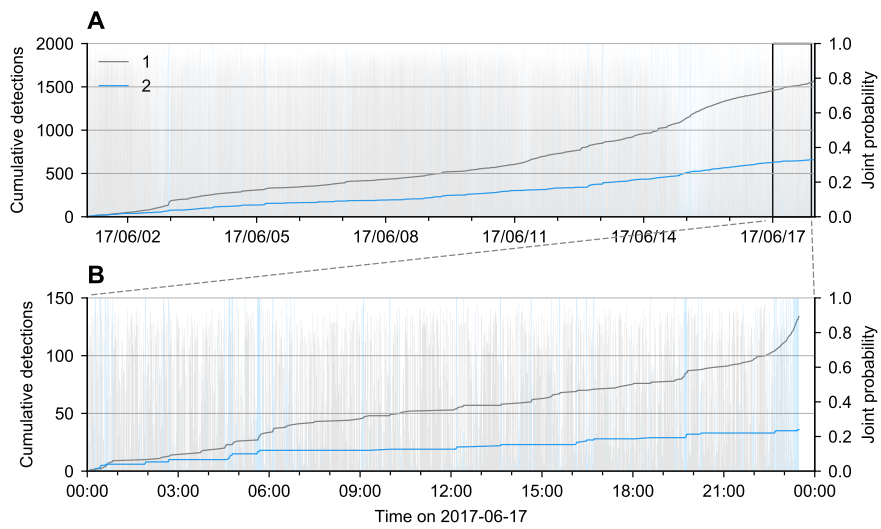


Figure 6: **Hierarchical clustering of long-duration seismic data.** (A) Within-cluster cumulative detection overserved for second-order clustering of former clusters F to I presented in Fig. 5 from 2017-06-01 to 2017-06-18. (B) Zoom on the day 2017-06-17 from the detections presented in A. Similarly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

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

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

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

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

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

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

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

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

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

521 for observing larger and larger time scales in the structure of the input signal  
 522 at reasonable computational cost.

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

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

525 with  $M$  standing for the depth of the scattering network, and  $F = J^{(1)}Q^{(1)}(1 +$   
 526  $\dots(1 + J^{(M)}Q^{(M)}))$  is the total number of scattering coefficients (or features).

527 When dealing with multiple-channel data, we also concatenate the scattering  
 528 coefficients obtained at all channels. The feature space therefore is a high-  
 529 dimensional representation that encodes multiple time-scales properties of the  
 530 signal over a time interval  $[t, t + \delta t]$ . The time resolution  $\delta t$  of this representation  
 531 then depends on the size of the pooling operations. The choice of the scattering  
 532 network depth thus should be chosen so that the final resolution of analysis is  
 533 larger than maximal duration of the analyzed signals.

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

## 541 Adaptive Hermite cubic splines

542 Instead of learning all the coefficients of the mother wavelet  $\psi_0^{(\ell)}$  at each layer  
 543 in the frequency domain, as one would do in a convolutional neural network,  
 544 we restrict the learning to the amplitude and the derivative on a specific set of  
 545  $\mathcal{K}$  knots  $\{t_k \in c\}_{k=1\dots\mathcal{K}}$  laying in the compact temporal support  $c$  (see Eq. S1).  
 546 The mother wavelet  $\psi_0^{(\ell)}$  can then be approximated with Hermite cubic splines  
 547 [23], a third-order polynomial defined on the interval defined by two consecutive

548 knots  $\tau_k = [t_k, t_{k+1}]$ . The four equality constraints

$$\left\{ \begin{array}{l} \psi_0^{(\ell)}(t_k) = \gamma_k \\ \psi_0^{(\ell)}(t_{k+1}) = \gamma_{k+1} \\ \dot{\psi}_0^{(\ell)}(t_k) = \theta_k \\ \dot{\psi}_0^{(\ell)}(t_{k+1}) = \theta_{k+1} \end{array} \right., \quad (\text{S7})$$

uniquely determine the Hermite cubic spline solution piecewise on the consecutive time segments  $\tau_k$ , given by

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

549 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  
 550 value and derivative of the wavelets on the knots, where  $x(t) = \frac{t-t_k}{t_{k+1}-t_k}$  is the  
 551 normalized time on the interval  $\tau_k$ , and where the Hermite cubic functions  $f_i(t)$   
 552 are defined as

$$\left\{ \begin{array}{l} f_1(t) = 2t^3 - 3t^2 + 1, \\ f_2(t) = -2t^3 + 3t^2, \\ f_3(t) = t^3 - 2t^2 + t, \\ f_4(t) = t^3 - 2t^2. \end{array} \right. \quad (\text{S9})$$

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

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

558 The parameters  $\gamma_k$  and  $\theta_k$  solely control the shape of the mother wavelet  
 559 and are the only parameters that we learn in our strategy. Notice that thanks to  
 560 the above constraints, for any value of those parameters, the obtained wavelet

561 is guaranteed to belong into the functional space of wavelets  $\mathcal{L}$  defined in Eq. S1  
 562 with compact support. By simple approximation argument, Hermite cubic  
 563 splines can approximate arbitrary functions with a quadratically decreasing error  
 564 with respect to the increasing number of knots  $\mathcal{K}$ . Once the mother filter  
 565 has been interpolated, the entire filter-bank is derived according to Eq. S2.

## 566 Clustering in a low-dimensional space

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

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

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

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

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

## 596 **Learning the wavelets with gradient descent**

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

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

615 **Reconstruction loss**

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

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

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

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

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

623 **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

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

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

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

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

## 627 **Within-cluster waveform analysis**

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

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

## 635 **Tests with different parameters**

636 One key parameter is the number of knots used to learn the shape of the wavelet.  
 637 This parameter is responsible for the wavelet duration in time, and inherently  
 638 for the wavelet bank quality factor. Indeed, a small number of knots defines a  
 639 wavelet localized in time with a large frequency bandwidth and vice-versa. We  
 640 therefore vary the number of knots in Fig. 5A to C in order to observe both  
 641 the clustering and reconstruction losses onto a small subset of the dataset (8.5  
 642 hours). These tests are also very helpful to show that the procedure still works  
 643 with a small amount of data (9 hours), a situation where deep convolutional  
 644 neural networks are known to fail easily. We see that taking a low number  
 645 of 7 knots (case A) allows to better reconstruct the input data with a loss of  
 646 4.20 (Table 1), but have a relatively high clustering loss (3.79). We observe in  
 647 Fig. 5A that the cumulative curves trends are not clearly separated between  
 648 clusters 2 and 3, also indicating that the clustering may have not converged to  
 649 a stable description of the data. As we can see on Table 1 for cases A to C,



650 increasing the number of knots (from 7 to 15) improves the clustering quality,  
651 but lowers the reconstruction loss. Even if the detection results are highly similar  
652 between cases A to C, we consider 11 knots to be a good trade-off between a  
653 high clustering quality and a reasonable reconstruction loss. In any case the  
654 precursory signals are always recovered even with a small amount of data, a  
655 clear advantage of our clustering procedure over clustering strategies based on  
656 classical deep convolutional neural networks.

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

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

680 **Comparison of cluster detection rates and microseismic en-**  
681 **ergy**

682 We collect the spectral pressure calculated from the WAVEWATCH III model  
683 (CIET ARDHUIN) on a  $0.5 \times 0.5$  degree grid globally, from 2017-06-01 to 2017-  
684 06-18. This pressure data cannot be directly used as a proxy for radiated seismic  
685 energy, because the radiation of body and surface waves depends on the bathy-  
686 metric profile of the seefloor [45]. According to [45], the equivalent radiated  
687 spectral energy can be derived from the pressure with taking into account the  
688 resonance of the water column at each point of the grid as amplification factor.  
689 We therefore used the amplification model presented in [45], where the global  
690 bathymetry is taken into account. We then considered the source time func-  
691 tion of each points of a  $4 \times 4$  degree grid, and correlated it with the temporal  
692 within-cluster detection. Because the pressure data is availble every 3 hours, we  
693 decimated the within-cluster detection on the same time basis.

694 The correlation is tested for several frequency bands (0.1 to 0.2, 0.2 to 0.3,  
695 0.3 to 0.45 and 0.45 to 0.6) and seismic waves (P waves, S waves and Rayleigh  
696 waves). For each frequency band, the maximally correlated source time func-  
697 tion and seismic wave type is identified and represented in Fig. S2 and S3. In  
698 addition to the water-column resonance amplifications, we also apply different  
699 corrections for the different seismic wave types. The P-wave spectral energy is  
700 corrected from the shadowing of the Earth's core (no energy should be recorded  
701 between 104 and 140 degrees of epicentral distance). This first correction is ap-  
702 plied as a mask on the correlation coefficients between within-cluster detection  
703 rates and source time functions. For Rayleigh waves, we also took into account  
704 the strong attenuation effects of the crust heterogeneities at these frequencies.  
705 We here considered an exponentially decaying attenuation with distance, with  
706 a decay of  $1/500 \text{ km}^{-1}$ .

## 707 References

- 708 [1] Bergen, K. J., Johnson, P. A., Maarten, V. & Beroza, G. C. Machine  
709 learning for data-driven discovery in solid earth geoscience. *Science* **363**,  
710 eaau0323 (2019).
- 711 [2] Obara, K., Hirose, H., Yamamizu, F. & Kasahara, K. Episodic slow slip  
712 events accompanied by non-volcanic tremors in southwest japan subduction  
713 zone. *Geophysical Research Letters* **31** (2004).
- 714 [3] Perol, T., Gharbi, M. & Denolle, M. Convolutional neural network for  
715 earthquake detection and location. *Science Advances* **4**, e1700578 (2018).
- 716 [4] Ross, Z. E., Meier, M.-A., Hauksson, E. & Heaton, T. H. Generalized seis-  
717 mic phase detection with deep learning. *arXiv preprint arXiv:1805.01075*  
718 (2018).
- 719 [5] Scarpetta, S. *et al.* Automatic classification of seismic signals at mt. vesu-  
720 vius volcano, italy, using neural networks. *Bulletin of the Seismological*  
721 *Society of America* **95**, 185–196 (2005).
- 722 [6] Esposito, A. M., D’Auria, L., Giudicepietro, F., Caputo, T. & Martini,  
723 M. Neural analysis of seismic data: applications to the monitoring of mt.  
724 vesuvius. *Annals of Geophysics* (2013).
- 725 [7] Maggi, A. *et al.* Implementation of a multistation approach for automated  
726 event classification at piton de la fournaise volcano. *Seismological Research*  
727 *Letters* **88**, 878–891 (2017).
- 728 [8] Malfante, M. *et al.* Machine learning for volcano-seismic signals: Challenges  
729 and perspectives. *IEEE Signal Processing Magazine* **35**, 20–30 (2018).
- 730 [9] Esposito, A. *et al.* Unsupervised neural analysis of very-long-period events  
731 at stromboli volcano using the self-organizing maps. *Bulletin of the Seis-*  
732 *mological Society of America* **98**, 2449–2459 (2008).

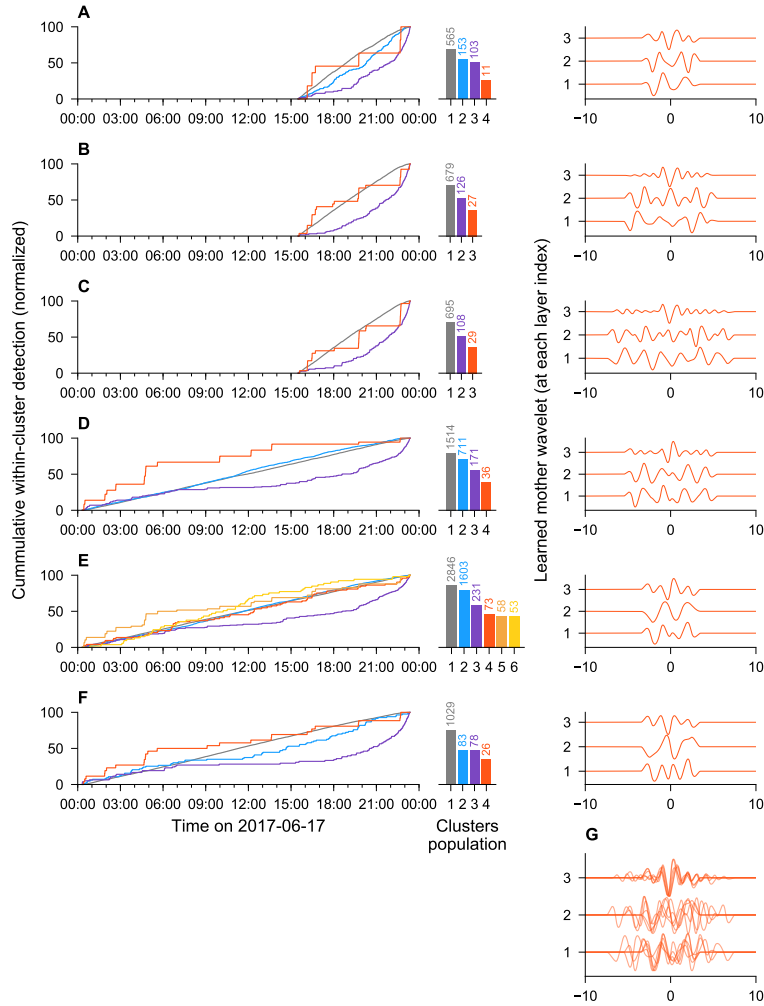


Figure S7: **(Supplementary material) Learning results with different parameters.** The different parameter sets are given in Table 1. The left and middle plots respectively show the within-cluster cumulative detections and the within-cluster number of samples after 10,000 training epochs. The right 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.

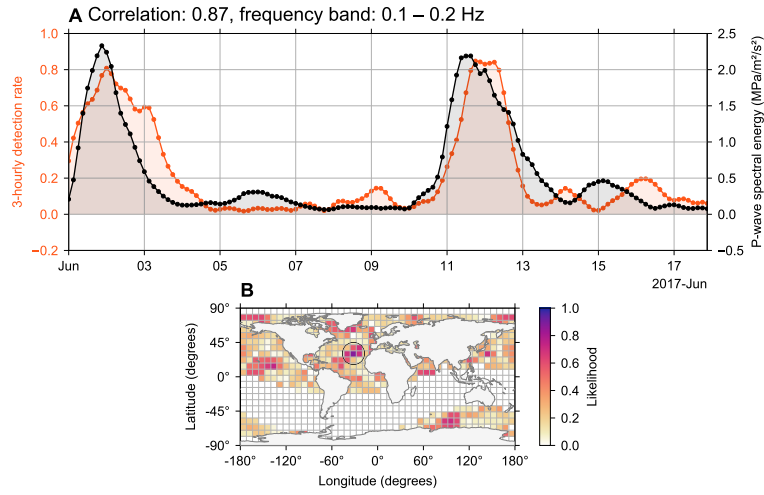


Figure S8: **(Supplementary material) Comparison of clusters D with P-wave microseismic energy.** (A) The within-cluster 3-hourly detection is presented in red curve over 17 days of 3-components seismic data. The best-matching radiated P-wave spectral energy in the frequency band 0.1 to 0.2 Hz is presented in black line. (B) Global matching likelihood of the spectral P-wave radiated energy between 0.1 and 0.2 Hz on a  $4 \times 4$  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.

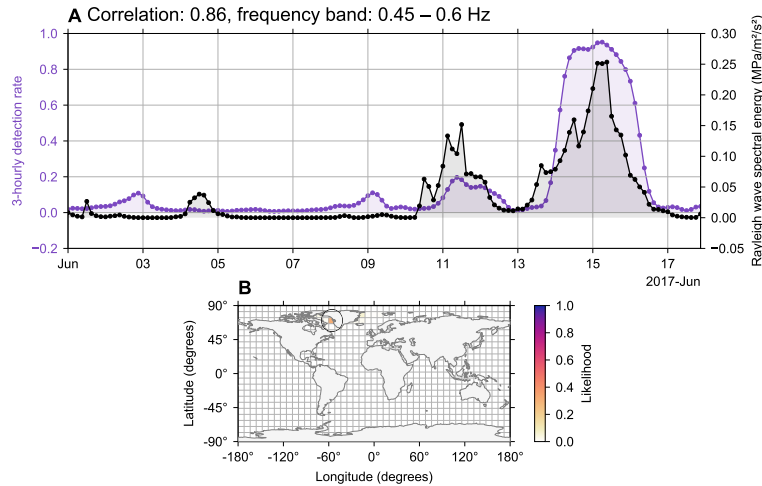


Figure S9: **(Supplementary material) Comparison of clusters C with Rayleigh wave microseismic energy.** **(A)** The within-cluster 3-hourly detection is presented in purple curve over 17 days of 3-components seismic data. The best-matching radiated Rayleigh-wave spectral energy in the frequency band 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  $4 \times 4$  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.

- 733 [10] Unglert, K. & Jellinek, A. Feasibility study of spectral pattern recogni-  
734 tion reveals distinct classes of volcanic tremor. *Journal of Volcanology and*  
735 *Geothermal Research* **336**, 219–244 (2017).
- 736 [11] Hammer, C., Ohrnberger, M. & Faeh, D. Classifying seismic waveforms  
737 from scratch: a case study in the alpine environment. *Geophysical Journal*  
738 *International* **192**, 425–439 (2012).
- 739 [12] Soubestre, J. *et al.* Network-based detection and classification of seismo-  
740 volcanic tremors: Example from the klyuchevskoy volcanic group in kam-  
741 chatka. *Journal of Geophysical Research: Solid Earth* **123**, 564–582 (2018).
- 742 [13] Beyreuther, M., Hammer, C., Wassermann, J., Ohrnberger, M. & Megies,  
743 T. Constructing a hidden markov model based earthquake detector: ap-  
744 plication to induced seismicity. *Geophysical Journal International* **189**,  
745 602–610 (2012).
- 746 [14] Holtzman, B. K., Paté, A., Paisley, J., Waldhauser, F. & Repetto, D.  
747 Machine learning reveals cyclic changes in seismic source spectra in geysers  
748 geothermal field. *Science advances* **4**, eaao2929 (2018).
- 749 [15] Yoon, C. E., O’Reilly, O., Bergen, K. J. & Beroza, G. C. Earthquake detec-  
750 tion through computationally efficient similarity search. *Science advances*  
751 **1**, e1501057 (2015).
- 752 [16] Mousavi, S. M., Zhu, W., Ellsworth, W. & Beroza, G. Unsupervised clus-  
753 tering of seismic signals using deep convolutional autoencoders. *IEEE Geo-*  
754 *science and Remote Sensing Letters* (2019).
- 755 [17] Köhler, A., Ohrnberger, M. & Scherbaum, F. Unsupervised pattern recog-  
756 nition in continuous seismic wavefield records using self-organizing maps.  
757 *Geophysical Journal International* **182**, 1619–1630 (2010).
- 758 [18] Rouet-Leduc, B. *et al.* Machine learning predicts laboratory earthquakes.  
759 *Geophysical Research Letters* **44**, 9276–9282 (2017).

- 760 [19] Bruna, J. & Mallat, S. Invariant scattering convolution networks. *IEEE*  
761 *transactions on pattern analysis and machine intelligence* **35**, 1872–1886  
762 (2013).
- 763 [20] Andén, J. & Mallat, S. Deep scattering spectrum. *IEEE Transactions on*  
764 *Signal Processing* **62**, 4114–4128 (2014).
- 765 [21] Andén, J. & Mallat, S. Scattering representation of modulated sounds.  
766 *15th DAFX* **9** (2012).
- 767 [22] Peddinti, V. *et al.* Deep scattering spectrum with deep neural networks.  
768 In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE Inter-*  
769 *national Conference on*, 210–214 (IEEE, 2014).
- 770 [23] Balestrieri, R., Cosentino, R., Glotin, H. & Baraniuk, R. Spline filters  
771 for end-to-end deep learning. In Dy, J. & Krause, A. (eds.) *Proceedings of*  
772 *the 35th International Conference on Machine Learning*, vol. 80 of *Proceed-*  
773 *ings of Machine Learning Research*, 364–373 (PMLR, Stockholmsmässan,  
774 Stockholm Sweden, 2018).
- 775 [24] Ahuja, N., Lertrattanapanich, S. & Bose, N. Properties determining choice  
776 of mother wavelet. *IEE Proceedings-Vision, Image and Signal Processing*  
777 **152**, 659–664 (2005).
- 778 [25] Meyer, Y. *Wavelets and operators*, vol. 1 (Cambridge university press,  
779 1992).
- 780 [26] Coifman, R. R. & Wickerhauser, M. V. Entropy-based algorithms for best  
781 basis selection. *IEEE Transactions on information theory* **38**, 713–718  
782 (1992).
- 783 [27] Poli, P. Creep and slip: Seismic precursors to the nuugaatsiaq landslide  
784 (greenland). *Geophysical Research Letters* **44**, 8832–8836 (2017).
- 785 [28] Chao, W.-A. *et al.* The large greenland landslide of 2017: Was a tsunami  
786 warning possible? *Seismological Research Letters* (2018).



- 787 [29] Bell, A. F. Predictability of landslide timing from quasi-periodic precursory  
788 earthquakes. *Geophysical Research Letters* **45**, 1860–1869 (2018).
- 789 [30] Allen, R. Automatic phase pickers: Their present use and future prospects.  
790 *Bulletin of the Seismological Society of America* **72**, S225–S242 (1982).
- 791 [31] Gibbons, S. J. & Ringdal, F. The detection of low magnitude seismic events  
792 using array-based waveform correlation. *Geophysical Journal International*  
793 **165**, 149–166 (2006).
- 794 [32] Brown, J. R., Beroza, G. C. & Shelly, D. R. An autocorrelation method  
795 to detect low frequency earthquakes within tremor. *Geophysical Research*  
796 *Letters* **35** (2008).
- 797 [33] Estivill-Castro, V. Why so many clustering algorithms: a position paper.  
798 *SIGKDD explorations* **4**, 65–75 (2002).
- 799 [34] Chakraborty, A. & Okaya, D. Frequency-time decomposition of seismic  
800 data using wavelet-based methods. *Geophysics* **60**, 1906–1916 (1995).
- 801 [35] Dumoulin, V. & Visin, F. A guide to convolution arithmetic for deep  
802 learning. *arXiv preprint arXiv:1603.07285* (2016).
- 803 [36] Shelly, D. R., Beroza, G. C. & Ide, S. Non-volcanic tremor and low-  
804 frequency earthquake swarms. *Nature* **446**, 305 (2007).
- 805 [37] Reynolds, D. Gaussian mixture models. *Encyclopedia of biometrics* 827–  
806 832 (2015).
- 807 [38] Mallat, S. *A wavelet tour of signal processing: the sparse way*, chap. 4,  
808 111–112 (Elsevier, 2009), 3 edn.
- 809 [39] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization.  
810 *arXiv preprint arXiv:1412.6980* (2014).
- 811 [40] Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **32**, 241–254  
812 (1967).

- 813 [41] Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing*  
814 **17**, 395–416 (2007).
- 815 [42] Sifre, L., Kapoko, M., Oyallon, E. & Lostanlen, V. Scatnet: a matlab  
816 toolbox for scattering networks (2013).
- 817 [43] Xu, L. & Jordan, M. I. On convergence properties of the em algorithm for  
818 gaussian mixtures. *Neural computation* **8**, 129–151 (1996).
- 819 [44] Arduin, F. *et al.* Ocean wave sources of seismic noise. *Journal of Geo-*  
820 *physical Research: Oceans* **116(C9)** (2011).
- 821 [45] Li, L., Boue, P. & Campillo, M. Spatiotemporal connectivity of noise-  
822 derived seismic body waves with ocean waves and microseism excitations  
823 *Eartharxiv* (2019).