Exploration of continuous seismic recordings with a machine learning approach to document 20 yr of landslide activity in Alaska

C. Hibert, D Michéa, F. Provost, J-P Malet, M Geertsema

Geophysical Journal International

Geophys. J. Int. (2019) **219,** 1138–1147
Advance Access publication 2019 July 31
GJI Seismology

doi: 10.1093/gji/ggz354

Exploration of continuous seismic recordings with a machine learning approach to document 20 yr of landslide activity in Alaska

C. Hibert [1], D. Michéa,[1,2] F. Provost [1], J.-P. Malet[1] and M. Geertsema[3]

[1] *Ecole et Observatoire des Sciences de la Terre / EOST, CNRS - University of Strasbourg, 67084 Strasbourg, France. E-mail: hibert@unistra.fr*
[2] *A2S / Application Satellite Survey - University of Strasbourg, 67084 Strasbourg, France*
[3] *Ministry of Forests, Lands, Natural Resource Operations and Rural Development, Prince George, BC V2N 4W5, Canada*

Accepted 2019 July 30. Received 2019 June 12; in original form 2019 April 25

SUMMARY

Quantifying landslide activity in remote regions is difficult because of the numerous complications that prevent direct landslide observations. However, building exhaustive landslide catalogues is critical to document and assess the impacts of climate change on landslide activity such as increasing precipitation, glacial retreat and permafrost thawing, which are thought to be strong drivers of the destabilization of large parts of the high-latitude/altitude regions of the Earth. In this study, we take advantage of the capability offered by seismological observations to continuously and remotely record landslide occurrences at regional scales. We developed a new automated machine learning processing chain, based on the Random Forest classifier, able to automatically detect and identify landslide seismic signals in continuous seismic records. We processed two decades of continuous seismological observations acquired by the Alaskan seismic networks. This allowed detection of 5087 potential landslides over a period of 22 yr (1995–2017). We observe an increase in the number of landslides for the period and discuss the possible causes.

Key words: North America; Numerical solutions.

1 INTRODUCTION

Recent observations have shown that massive landslides are impacting Alaskan mountain ranges and fjords (Geertsema *et al.* 2013; Coe *et al.* 2018). Some of these are among the largest historic landslides observed on the Earth. Although most of these occurred in underpopulated regions, the recent Taan–Tydall landslide, which generated a tsunami wave with an estimated height of almost 200 m (Dufresne *et al.* 2017; George *et al.* 2017; Gualtieri & Ekström 2018), has demonstrated the threat that such landslide could pose to human activity. Alaska (USA) is particularly prone to landslides due to its geology and the conjunction of tectonic (high-magnitude earthquakes) and environmental (glacial retreat, temperature rise, heavy precipitation and permafrost thaw) forcings. The latter are impacted by the global climate change that is expected to have a major influence on slope stability in Alaska and in other high- and low-latitude regions (Coe & Godt 2012; Huggel *et al.* 2012; Geertsema *et al.* 2013). Hence, emerge questions such as are we, in the next decades, going to observe more large landslides in this region? What are the dominating controlling factors? Is climate change impacting landslide activity in those high-latitude/altitude regions?

To answer these questions, building exhaustive landslide catalogues covering a long time period and a large spatial extent is compulsory. Remote sensing, geomorphological observations and direct witness allow us to detect and map the events (e.g. Kirschbaum *et al.* 2010; Coe *et al.* 2018) but at a low temporal resolution (weeks, months, years) and/or with a small geographical coverage. These limitations often impede accurate determination of the links between landslide triggering and short- and long-term meteorological patterns, especially at regional scale and for remote regions of the globe.

Over the last few years, environmental seismology aims at providing new insights on the dynamics of surface processes through the study of the seismic waves they generate. Those environmental processes are, for example, the seismic hum generated by storms (Gualtieri *et al.* 2018), the seismic waves generated by ice-calving events at glaciers (Sergeant *et al.* 2016) and those generated by landslides and other mass wasting processes (e.g. Allstadt *et al.* 2017). By studying the continuous seismic records, we can accurately determine the occurrence time (exact to the second) of landslides, and from the features of the seismic signals infer properties such as the mass, the distance travelled and the velocity (Brodsky *et al.* 2003; Favreau *et al.* 2010; Moretti *et al.* 2012; Ekström & Stark 2013; Yamada *et al.* 2013; Allstadt 2013; Dammeier *et al.* 2011; Hibert *et al.* 2014a,b,c, 2017a; Levy *et al.* 2015; Moore *et al.* 2017; Gualtieri & Ekström 2018). Further, one advantage of seismology is that some seismic stations have been active for decades, allowing the construction of event catalogues covering long time span.

We distinguish two approaches in landslide seismology. The first approach, which has been continuously developed since the 1980s

(Kanamori & Given 1982; Kanamori *et al.* 1984; Moretti *et al.* 2012; Allstadt 2013; Ekström & Stark 2013), is based on the analysis of the long-period (>10 s) seismic waves. This approach yielded unprecedented results such as the documentation of the dynamics of large landslides and the detection of new large events (Ekström & Stark 2013). However, this approach has a major limitation, as only rare and massive (>1 Mm^3) landslides generate long-period seismic waves, and thus most of the landslide events are missed. Hence, a second approach using high-frequency seismic waves has been developed (>1 Hz; Suriñach *et al.* 2005; Deparis *et al.* 2008; Dammeier *et al.* 2011; Hibert *et al.* 2011, 2014b, 2017a; Levy *et al.* 2015; Zimmer & Sitar 2015; Dietze *et al.* 2017; Fuchs *et al.* 2018; Schöpa *et al.* 2018). High-frequency waves are generated by most landslides and can be recorded by any seismic station located close enough to the source (from few kilometres to hundreds, depending on the mass and the dynamics of the landslide). However, this approach brings new processing difficulties. In the high-frequency bands, a significantly larger number of seismic event types are recorded and those can have diverse origins (tectonic, environmental and anthropogenic). This is a major limitation when analysing years of continuous seismological data from large networks (hundreds of sensors) because carrying out the detection and the source identification manually would take years. Recent significant advances in artificial intelligence and machine learning can help to overcome this difficulty. We used the 'Random Forest' (RF) (Breiman 2001) classifier that is based on the computation of a large number of decision trees (>500). In our work, we demonstrate how to use an implementation of the RF algorithm to explore 22 yr (starting in 1995) of continuous seismological data acquired by the Alaska seismic network.

2 DATA

In this work, we gathered two sets of data: (1) the first data set used to train the machine learning algorithm is the 'training set' that includes known events belonging to the class 'earthquakes' and to the class 'landslides'; (2) the second data set consists of continuous seismograms recorded by stations, which we explore with the proposed processing chain to build the landslide catalogue.

2.1 Training set

Before using the RF algorithm to process the continuous seismological data, it has to be trained and tested with a set of signals for which we know the source. We gathered a set of 3636 seismic signals generated by 290 earthquakes that occurred in January 2016, in Alaska but also in the whole America and Pacific regions. Those earthquakes are registered in the USGS catalogue with magnitudes ranging from 2.5 to 7.1 and were recorded by the broad-band stations we use in this study and that are located in Alaska and Canada, from the AK, AT, AV, CN, II, IU, IM, US and TA networks (Fig. 1). For each earthquake, automated detection and picking of the signal on each station were performed with the method described in Section 3.1, and manually checked before being integrated in the training set.

We included in the training set 205 seismic signals generated by 11 large landslides (of volumes above 1 million cubic meters) that occurred worldwide (Hibert *et al.* 2017a) and were confirmed by geomorphological observations. We choose to not only include seismic signals generated by landslides that occurred in Alaska to increase the number of signals in the training set for the landslide class. We also selected events for which seismic signals have been recorded at regional distances (from few to hundreds of kilometres) in order to integrate in the training set seismic signals differently impacted by propagation effects. The name, date, location, number of signals selected and the range of distances of the stations from the source are given for the 11 landslides in Table 1. As demonstrated by numerous studies (Deparis *et al.* 2008; Vilajosana *et al.* 2008; Dammeier *et al.* 2011, 2016; Hibert *et al.* 2011; Levy *et al.* 2015; Hibert *et al.* 2017a; Allstadt *et al.* 2018), the general features of the high-frequency seismic signals generated by landslides (emergent onset, no phase, spindle-shaped spectrogram, long coda) are the same for events of different sizes, occurring in different contexts and recorded by different networks. Therefore, we assume that, because the machine learning approach used in this study is based on these features including seismic signals of landslides that did not occur in Alaska in the training, data set should not reduce its capacity to detect new landslide seismic signals.

2.2 Continuous seismic observations

We analysed continuous seismic records acquired by 243 broad-band seismometers from the AK, AT, AV, CN, II, IU, IM, US and TA networks, located within a rectangular zone between latitudes of 48° and 68° and longitudes −124° and −144° (Fig. 1). We focused our study on the BHZ channel of each station (vertical component and sample rate of 40 or 50 Hz) to limit the number and the size of the seismic records to process. The stations for which such channel did not exist were excluded from our data set. For each station selected, we collected all the continuous records available between 1995 and 2017. We did not select any triggered records. Each record was deconvolved from instrumental response and high-pass filtered above 1 Hz.

3 METHODS

Most of the machine learning algorithms use the complete seismic signal (from the onset to the end) generated by an event. Hence, before identifying the events, we have to extract their seismic signals from the continuous seismic records acquired by each station. The processing chain to build the landslide catalogue is based on two modules. The first module allows detecting and extracting the waveforms from the continuous seismograms. The second module computes a set of features from the extracted waveform and injects them to the machine learning algorithm, which then predicts the class (earthquake or landslide) to which the source might belong.

3.1 Detection and extraction of signals from continuous seismic records

Detection and extraction of signals from continuous seismic records can be difficult, especially when seeking landslide seismic signals. Classical methods such as STA/LTA (short-time average over long-time average) applied directly on the waveform of the seismic signal (Allen 1982) often miss the onset of those signals as they emerge slowly from the noise. To overcome this difficulty, we use a spectral-based approach that allows detecting seismic signals with low signal-to-noise ratios and emerging onsets (Helmstetter & Garambois 2010). Each continuous record is transformed into a spectrogram by computing the fast Fourier transform (FFT) on a moving window. The spectrograms are then transformed into spectral pseudo-envelope by integrating the amplitude of the FFT

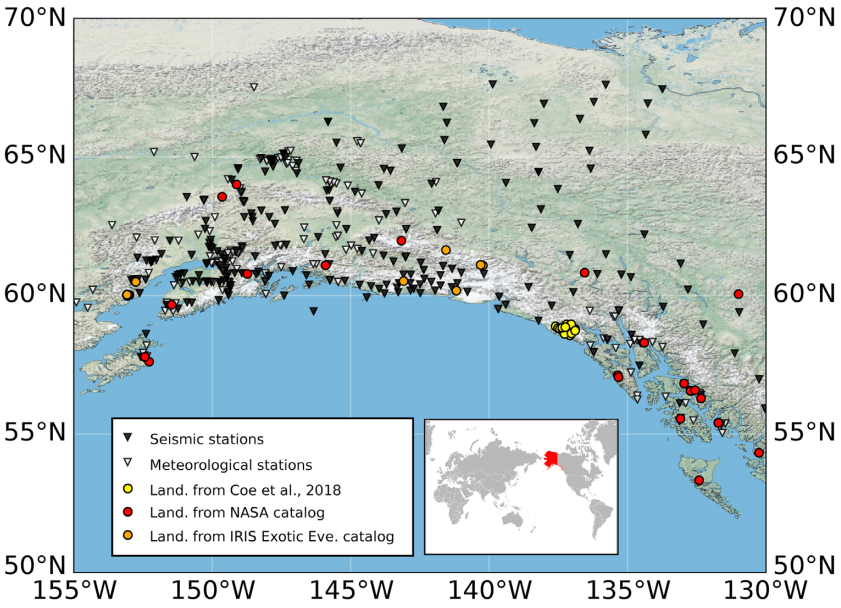

Figure 1. Map of the study region with seismic stations used represented as the dark triangle, meteorological stations used to compute yearly averages as the white triangle and location of landslides given in the catalogues provided by Coe *et al.* (2018) as the yellow circles, by Kirschbaum *et al.* (2010) as the red circles and Allstadt *et al.* (2017) as the orange circles.

Table 1. Landslides used to constitute the training set.

Name	Date (UTC)	Coordinates (Lat° / Lon°)	Number of signals selected	Stations distance range (km)
Akatani	04-09-2011 07:22	34.13° / 135.72°	7	17–66
Bingham #1	11-04-2013 03:31	40.53° / −112.14°	23	13–351
Bingham #2	11-04-2013 05:06	40.53° / −112.14°	19	13–351
Hubbard	21-05-2012 14:25	60.07° / −139.53°	19	55–249
Mount Dall	04-09-1999 15:15	62.58° / −152.46°	8	115–348
Mount Lituya	11-06-2012 22:23	58.80° / −137.43°	36	140–474
Mount Steele	25-07-2007 00:57	61.11° / −140.30°	41	49–822
Mount Steller	14-09-2005 19:59	60.49° / −143.09°	20	26–485
Mount Wrangell	25-07-2013 10:15	61.98° / −143.17°	20	69–203
Oso-Steelhead	22-03-2014 17:37	48.28° / −121.84°	7	12–124
Sheemahant Glacier	09-07-2010 07:35	51.87° / −125.95°	5	152–469

spectrum at each time step. Finally, a classical STA/LTA detection is performed on this spectral pseudo-envelope. In this study, we choose a moving window length of 100 samples to compute the FFT spectra, an overlap of the moving window of 90 per cent, and an STA/LTA ratio threshold of 0.5, with an STA window of 20 samples and an LTA window of 1000 samples.

Once an event is detected by the spectral detector, we refined the onset picking with a kurtosis-based picker (Baillard *et al.* 2014; Hibert *et al.* 2014b) on the waveform of the signal. The refined picking is performed on a window starting 5 s before the detection time given by the spectral detection and ending 30 s after. The characteristic functions are computed using three frequency bands: 1–3, 3–10 and 10–15 Hz. They are then stacked, and the global minimum of this stacked function gives the onset time of the signal. The values of the other parameters used in the parametrization of the kurtosis-based picker are the same as the ones given in Hibert *et al.* (2014b). This permits an accurate picking even for emerging signals, with difference between manual and automated picks that are usually less than 1 s. The end of the signal was picked when a dynamic threshold on the seismic signal envelope equal to two times the signal-to-noise ratio computed on a window of 100 samples before the onset time determined by the kurtosis-based picker is reached.

3.2 Source identification: RF and seismic signal features

The RF algorithm (Breiman 2001) is based on the computation of a large number of decision trees. The trees are built from a training data set including signals for which we know the origin of the seismic source. The seismic signals forming the training set are described by features that are statistically analysed to build the decision trees. The originality of this algorithm is that each tree is built from a subset of the seismic signals and their features selected randomly. Once the RF model is trained, it provides an identification of the source of a new signal by injecting the values of its features within each tree. Each tree will then provide a predicted class for the event. The final decision on the class is given by the majority vote of all the trees. The major advantage of the RF algorithm is that it yields a measure of the uncertainty on the classification

processes by giving scores, which are the normalized number of trees that vote for the predicted class. For example, a score of 1 means that 100 per cent of the trees agreed on a class; a score of 0.5 that 50 per cent of the trees agreed on a class.

When implementing machine learning algorithms for seismic signal processing, the most important step is to select relevant features describing the signals. The choice of the classifier is secondary and several have proven to work well with seismic data (Langer *et al.* 2006; Cortés *et al.* 2009; Ibáñez *et al.* 2009; Hammer *et al.* 2012; Langet 2014; Hibert *et al.* 2014b; Dammeier *et al.* 2016; Malfante *et al.* 2018). We choose to use the features proposed by Provost *et al.* (2017), Hibert *et al.* (2017b) and Maggi *et al.* (2017), along with the RF algorithm, as this methodology has proven to be accurate, robust and versatile when applied to seismic signals generated by landslides. Features include waveform properties (such as the duration, the ratio of the maximum and the mean of the envelope, the rising and decreasing duration of the envelope, the kurtosis of the envelope, etc.), spectral features (such as mean frequency, centroids of the spectra, number of peaks in the spectrum, energy in different frequency band, etc.) and spectrograms attributes (such as the envelope of the evolution of the maximum or the mean frequency with time, the kurtosis of the spectrogram envelope, etc.). For the full list of the 61 features used, we invite the readers to refer to Provost *et al.* (2017) and Hibert *et al.* (2017b).

For the processing, we used the Scikit-learn package implementation of the RF algorithm (Python 3.5). The RF algorithm has inherent qualities such as being able to work efficiently even when trained with a limited number of events (which is not the case of conventional neural network approaches), to not be spuriously influenced by ill-chosen features (conversely to fuzzy logic or Support Vector Machine), to avoid overfitting, and to be highly portable (as opposed to hidden Markov model that requires a specific training for each station, for example). All these qualities fulfil the requirements we identified as being critical to warranty the success of this work. We choose to grow an RF with 1000 trees. We use the typical number of features randomly selected at each split of the RF, which is $\sqrt{N}$ with N being the number of attributes; $N = 61$ in our case. We used the standard number of feature samples of each class randomly selected to build each node corresponding to 2/3 of the samples in the training set. The split feature and threshold are selected by minimizing the Gini impurity index in our case. For a complete description of the identification method, we invite the reader to refer to the appendix A of Hibert *et al.* (2017b).

3.3 Assessment of the RF algorithm accuracy with the training set

To assess the capability of the RF algorithm to discriminate between landslide and earthquake seismic signals, we choose to train the algorithm with an equal number of signals of each class from the training set and then used the trained model to classify the rest of the signals in the training set. This process is repeated 100 times by each time randomly selecting signals in the training set to train the algorithm. We increase gradually the number (from 10 to 100 signals of each class) of signals used to train the algorithm to investigate the sensibility of the classification to this parameter. Both tests allow evaluating the robustness of the algorithm. A common approach to quantify the rate of good identification of automated identification methods is to compute the sensitivity, which is, for a two-class problem, the ratio between the signals identified as belonging to a class on the true number of signals from this class (i.e.

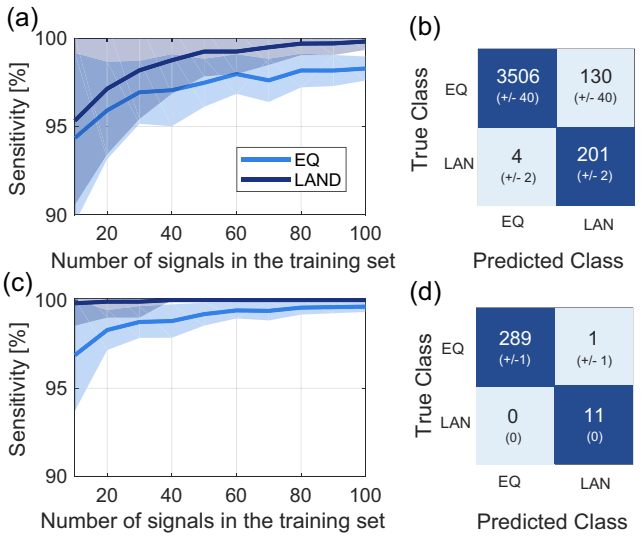

Figure 2. (a) Averaged total sensitivity as a function of the number of signals used in the training set for 100 instances of selecting randomly data of the earthquake (EQ) and the landslide (LAND) classes. The coloured patches indicate the standard deviation around the computed average sensitivities. (b) Confusion matrix for the average of the 100 instances of selecting randomly 100 signals of each class and identifying the rest of the signals in the training set. (c) Averaged total sensitivity as a function of the number of signals used in the training set for 100 instances of selecting randomly data of the EQ and the LAND classes and identifying events from the majority vote amongst all the signals associated with a given event. (d) Confusion matrix for the average of the 100 instances of selecting randomly 100 signals of each class and identifying events with the multisignal and score strategy.

if 50 true signals over 100 are correctly identified, the sensitivity is 50 per cent).

Fig. 2 shows the result of the instances of training the algorithm with a subset of the training set and identifying the rest of signals in the training set. With only 10 signals used to train the algorithm, a sensitivity between 90 and 100 per cent is reached (Fig. 2a). When using more than 40 signals of each class in the training set, the sensitivity is always above 95 per cent (Fig. 2a). Finally, over the 100 instances of training the algorithm with 100 signals of each class and identifying the signals from the training set we obtained an average sensitivity of 98.2 ± 0.7 per cent for the earthquake class and 99.8 ± 0.5 per cent for the landslide class (Fig. 2a).

Those rates of good identification are high, but do not reflect the capability the algorithm has to perform operationally on unlabelled data. We have to take into account the rate of false identification, which is, for the landslide class, the ratio of the number of earthquakes seismic signals identified as landslides over the number of true landslide seismic signals in the training set. As indicated by the confusion matrix of Fig. 2(b), for the 100 instances of training the algorithm with 100 signals of each class, the average ratio between real landslide seismic signals and earthquakes seismic signals identified as landslide seismic signals can be high. When using 100 seismic signals of each class to train the algorithm, the average number of earthquakes identified as landslides is 130 ± 40 (Fig. 2b), which is small compared to the 3636 earthquake seismic signals, but large compared to the 205 real landslide seismic signals in the training set. Those results show that more than a third of the seismic signal sources labelled as 'landslide' by the RF algorithm are misidentified. This comes mainly from imbalance in the classes of the training set as we have one order of magnitude more earthquake signals than landslide signals in the training set. This class

imbalance will probably be even more important when analysing continuous seismic records of the stations deployed in Alaska, as we can expect a lot more earthquake detection than landslide detection in this tectonically active region. Finally, the number of landslides falsely identified as earthquakes is 4 ± 2 over 3636 earthquakes seismic signals (Fig. 2b), which shows that almost no real landslide seismic signals are missed.

To overcome this class imbalance problem, we propose a new approach by considering not individual signals, but all the signals generated by a given event. An event, which is either a landslide or an earthquake in our case, is the gathering of all the seismic signals generated by the source and recorded at the different stations. For a given event, we consider all the seismic signals recorded, and perform the identification of each individual signal. Then, we look at the majority vote from the identification returned at each seismic station. With this voting strategy, we exclude individual signals for which identification by the RF algorithm yielded a score below 0.6 (i.e. at least 60 per cent of the tree have voted for the majority class). This strategy gives sensitivities of 99 ± 1 and 100 per cent for the earthquake and the landslide classes, respectively, when using 50 seismic signals of each class or more to train the algorithm (Fig. 2c). Over the 290 earthquakes in the data set, less than 2 are falsely identified as landslides in average (mean of 1.08 earthquakes over 100 iterations—Fig. 2d). Considering that we have 11 landslides in the data set, this gives a minimum effective rate of good identification of more than 82 per cent (Fig. 2d).

From this test on the training set, we defined a strategy to explore the continuous seismological observations: once a signal has been detected and picked at a station, if it is classified by the RF algorithm with a score above 0.6 on more than one station in a time window of 4 min around the onset of the detected signal, the event is declared as a landslide and included in the catalogue. Hence, to include an event in the landslide catalogue at least two stations must independently have detected an event and classified it as having a landslide source in this time window of 4 min. Having an event included in the catalogue only if it is detected by two or more stations also allows exclusion of most of local source of noises, as usually seismic signals generated by noise do not propagate over more than few kilometres for frequencies above 1 Hz, whereas interstations distances of the networks we used is generally of the order of dozens of kilometres. For the exploration of the continuous records, the algorithm was trained with all the seismic signals in the training set (3636 earthquakes signals and 205 landslide signals).

4 RESULTS

Processing years of continuous data is highly demanding in terms of computing time. We deployed the processing chain at the High Performance Computing (Datacentre/Mesocentre) facility of the University of Strasbourg. Parallelization of the code allowed reducing the computing time from approximately 1 yr on a consumer grade personal computer to 10 hr for 22 yr of data recorded on a maximum of 243 stations.

Over the period, we detected 6213 possible landslides. A manual inspection of the waveforms and spectrograms of the signals associated with each event was carried out to confirm the origin of the events. We conclude that 5087 events have generated seismic signals that have features consistent with those usually observed for landslides, implying a strong probability of a landslide source. This yields an effective rate of good identification of 81.8 per cent, close to the one estimated from the test on the training set. The misclassified events are mostly earthquakes originating from the West Pacific area, which have high-frequency signals resembling those generated by landslides (in particular a long duration and a general spindle shape), due to their long propagation through the Earth before reaching Alaskan seismic stations. A peak of false identifications is observed in March 2011, corresponding to the month of the occurrence of the Tohoku earthquake (Japan, 2011 March 11) and its aftershocks (Fig. 4a).

Among the 5087 landslides, we recover all the previously known major landslides that occurred in the region during the study period. Among these landslides that were not included in the training set but are known because of long-period seismic detection (Ekström & Stark 2013), are detected the La Perouse (16/02/2014 14:24 UTC, Lat.: 58.561°; Lon.: −137.062°), the Lamplugh glacier (28/06/2016, Lat.: 58.775°; Lon.: −136.935°) and the Taan–Tydall (18/10/2015 05:19 UTC, Lat.: 60.177°; Lon.: −141.187°) landslides. Finally, we also detected all the landslides listed in the IRIS Exotic Seismic Events catalogue (Allstadt *et al.* 2017) that occurred in Alaska and that were detected by at least two of the 243 stations used in our study (19 over 34 events listed as 'Rock/ice/debris avalanche and slide' as of April 2019).

We were also able to constrain the exact time of the Orville–Wilbur landslide (Lat.: 58.736°; Lon.: −137.272°) to the 07/03/2015 at 19:19 UTC (Fig. 3a). Previously, the timing for this large landslide, deduced from satellite images and airborne observations (Fig. 3b), was estimated to be between 2015 February 25 and 2015 March 8. The general features of the seismic signals and spectrograms recorded at each station (emergent onset, no phase, long coda, spindle-shaped spectrogram, spectrum energy between 1 and 10 Hz) strongly suggest a landslide source for these seismic signals. The arrival of the seismic waves at 17 stations located from 136 to 385 km from the location of the observed Orville–Wilbur landslide deposit (Fig. 3c) is consistent with a localization of the seismic source at this location (Fig. 3d).

Fig. 4(a) shows the monthly and yearly distribution of the 5087 potential landslides detected, revealing a gradual increase in the number of events occurring during the studied period. In 1996, less than 10 landslides were detected. This number reaches 100 landslides in 2003, 300 in 2008 and almost 500 landslides per year in 2013, 2014 and 2015. We observe a seasonal variation of the number of landslides as the months with the highest number of landslides are in spring (March, April, May) and late summer (July, August; insert in Fig. 4a). The distribution of the number of landslides as a function of the hour of the day is uniform.

5 DISCUSSION

Before interpreting the possible causes of the increase in landslide rates, we have to discard possible bias related to the evolution of the seismic network: as more stations are deployed in a region more landslides can potentially be detected. Hence, the question is are we observing an increase in the number of landslides or are we just detecting more landslides? A first approach to analyse this potential bias is to analyse the correlation between the increases in the number of landslides per month and in the number of stations used. The monotonic dependence of two variables is quantified by computing the Spearman's rank correlation coefficient (r_s). We compute this correlation coefficient on a window (Fig. 5a) by reducing the time window by 1 yr at each iteration. When considering the whole period (1995–2017), the Spearman's rank correlation coefficient is

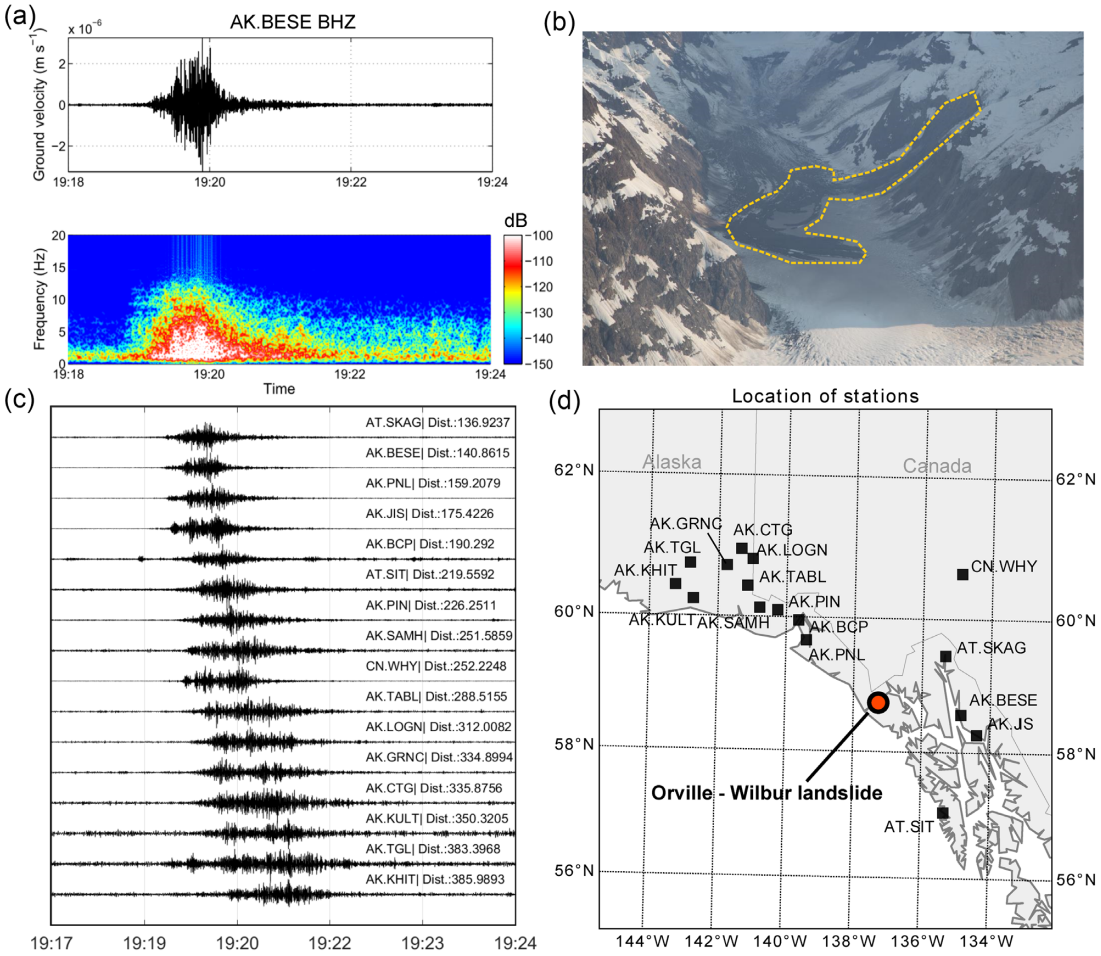

Figure 3. (a) Seismic signal filtered between 1 and 10 Hz and spectrogram of the seismic signal generated by the Orville–Wilbur landslide and recorded at station AK.BESE. (b) Aerial photograph of the Orville–Wilbur landslide deposit taken in 2015. (c) Seismic signals generated by the Orville–Wilbur landslide recorded by 16 stations and (d) map of the location of those 16 stations and of the location of the Orville-Willbur landslide.

higher than 0.8, which indicates that the increase in the number of landslides and seismic stations deployed might be correlated. However, when considering more recent periods, both time-series start to decorrelate, and after 2005, r_s falls below 0.5. This statistical analysis suggests that the densification of the seismic network cannot explain alone the increase in the number of landslides after 2004–2005.

To investigate the influence of the number of stations used on the number of landslides detected, we further conducted several statistical analyses. First, we looked at the number of landslide detections achieved uniquely by a given pair of stations. Because we decided to include a landslide in our catalogue when at least 2 stations have detected and identified a seismic signal as being generated by a landslide source, this removes any bias of detection related to the addition of stations in the network. We selected the 115 possible couples of stations (a given station belongs to several pairs) that have been active for the longest period. All the pairs include stations that were conjointly active at least since 2005. For each pair of stations, we can determine the evolution with time of the number of landslides detected. We can then compute the slope of the regression lines between the number of landslides detected and the date of detection. If the coefficient of this regression line is positive, it indicates that the number of landslides is increasing over the period. For 71 of 115 pairs of stations, we observed slopes with values above 0.1, 14 values under −0.1 and 30 between −0.1 and 0.1 (Fig. 5b). Hence, for two-thirds (65 per cent) of the pairs of stations that have been active before 2005, we observe positive slopes indicating an increase in the number of landslides with time.

We also investigated the evolution of the number of landslides detected only by all the stations that were deployed before 2000 (Fig. 5c) and before 2005 (Fig. 5d). For both cases, the number of landslides is clearly increasing after the chosen dates. Because we do not consider the stations included in the network after those dates, these increasing trends cannot be linked to the network densification, and can only be explained by an effective increase in the number of landslides. Finally, Figs 5(e) and (f) were produced by removing randomly (bootstrapping) 50 and 75 per cent of the stations in the data set and by computing the number of landslides per year detected by two or more of the remaining stations. This processing was repeated 100 times for each case (50 and 75 per cent). We observe consistently an increase in the number of landslides over time. The bootstrapping approach shows that the increasing trend is not controlled by few stations but is observed through all the zone covered by the seismic network.

All the elements discussed previously suggest that the densification of the seismic network is not the dominant factor explaining the observed increase in the number of landslides. Therefore, we looked at the impact of the variation of the average air temperature

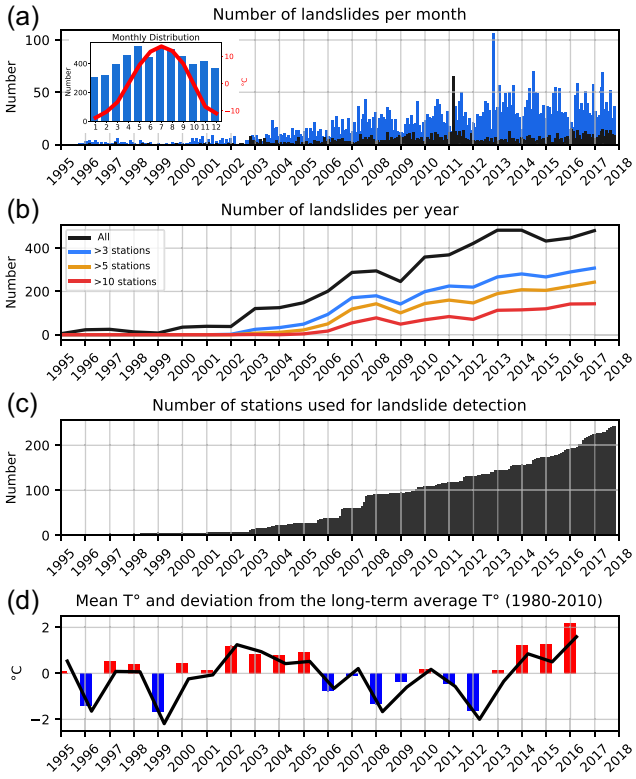

Figure 4. (a) Number and distribution of landslides per month in blue and of other sources wrongly identified as landslides in black. The insert shows the distribution of the number of landslide recorded for each month and the average temperature of each month; (b) Number of landslides per year, detected by more than 2 stations (the black line), by more than 3 stations (the blue line), by 5 stations (the orange line) and by 10 stations (the red line); (c) Number of stations included in the processing; (d) Average temperature (the black line) and deviation from the climatic normal (1980–2010) per year.

in Alaska. Meteorological data were obtained from the NOAA's National Climatic Data Center (NCDC). The deviation of annual temperatures is computed from a reference normal temperature provided at each meteorological station for the period 1981–2010. We then averaged the deviation from normal temperature recorded at each station. Public data for the year 2017 are not available at the time of this study. The warmest years post-2005 were 2013 and 2016. The largest number of landslides is observed during this period. Increase of the number of landslides corresponds most of the time to periods transitioning from cold years to warm years. For example, the strong peak of activity observed on Fig. 4(b) for the year 2013 is concomitant with the highest difference of average air temperature between two consecutive years (2012 and 2013—Fig. 4d) observed after 2005. We observe the same qualitative correlation between the two quantities in 2006–2007 and 2009–2010 (Figs 4b,d, 5c and d). We explored other possible links between long-term landslide activity and environmental factors, such as the yearly maximum snow cover, yearly average precipitation, yearly number of days with temperature above 0 °C, 21 °C and 32 °C and total sunshine, but none yielded any significant correlation at the regional scale.

A quantitative analysis looking at the details of the spatio–temporal evolution of the landslide activity is, therefore, required and will constitute the focus of future works based on the seismology-derived landslides catalogue we produced. It will help to understand the potential multiple causes of monthly and daily sharp increase in the number of landslides. For example, the highest monthly number of landslide is observed in October 2012 (Fig. 4a). A major earthquake occurred south of the Glacier Bay on the 27/10/2012 (Haida Gwaii Earthquake, M_w 7.8) and might have caused this peak of landslide activity.

Overall, we show that there is an increase in the number of landslides over 22 yr, which is not linked to the seismic network densification at least after 2005, and is possibly related to the increase in the yearly average temperature. How and through which processes the global climatic forcing is impacting the landslide activity is an open question. Future developments will allow to thoroughly analyse the catalogue we constructed to provide an extensive knowledge of the spatio–temporal evolution of landslide activity in Alaska. According to our tests on the training set, this catalogue should be as exhaustive as possible. We have shown that our method tends to produce some false landslide detection but do not miss any real landslide, which is confirmed by the detection of all the seismogeneic landslides that occurred in the region and are not included in our training set. However, this comes at the cost of having a landslide recorded by at least two seismic stations. This hinders the detection of the smallest events. To detect the smallest events, we must be capable to use single-station detection and identification and/or incorporate more stations in our analyses. Both require reducing the rate of false landslide detection. Our work suggests that this is possible by reducing the class imbalance by increasing the number of landslides seismic signals in the training set. This possibility will be explored in future work.

6 CONCLUSIONS

We develop a new approach to explore continuous seismic records that focus on automatically detecting and identifying landslide seismic sources. Our results demonstrate that the RF machine learning algorithm, associated with a set of selected seismic signal features, is able to discriminate between earthquake and landslide-generated seismic signals with an effective success rate of 99 per cent on the training set. This high success rate can be achieved with a small number of signals to train the algorithm. An overall sensitivity of approximately 90 per cent is already obtained when using only 10 signals of each class to train the algorithm. This demonstrates an important strength of our approach, that is, to yield a high accuracy score with a small training set. We also show that using seismic signals of landslides recorded by stations deployed in a specific region can be used with success to identify landslide seismic signals recorded by stations from a seismic network deployed in another region of the world. Finally, we consistently reach high rates of good identification while selecting randomly a small fraction of different seismic signals to train the RF algorithm, showing that it is possible to successfully discriminate between earthquakes and landslides seismic signals even with signals recorded at a wide range of distances. Those observations suggest that the majority of the features we selected are controlled by the physics of the source and not the effects of the propagation of the seismic waves. Our results demonstrate the robustness and the versatility of the RF algorithm associated with the features we defined for landslide seismic sources detection.

The new processing chain, based on the high-frequency seismic waves generated by landslides, allows detecting new events and potentially smaller events than the very large landslides discovered through long-period seismic waves detection. This is an important step forward for the production of exhaustive landslide catalogues

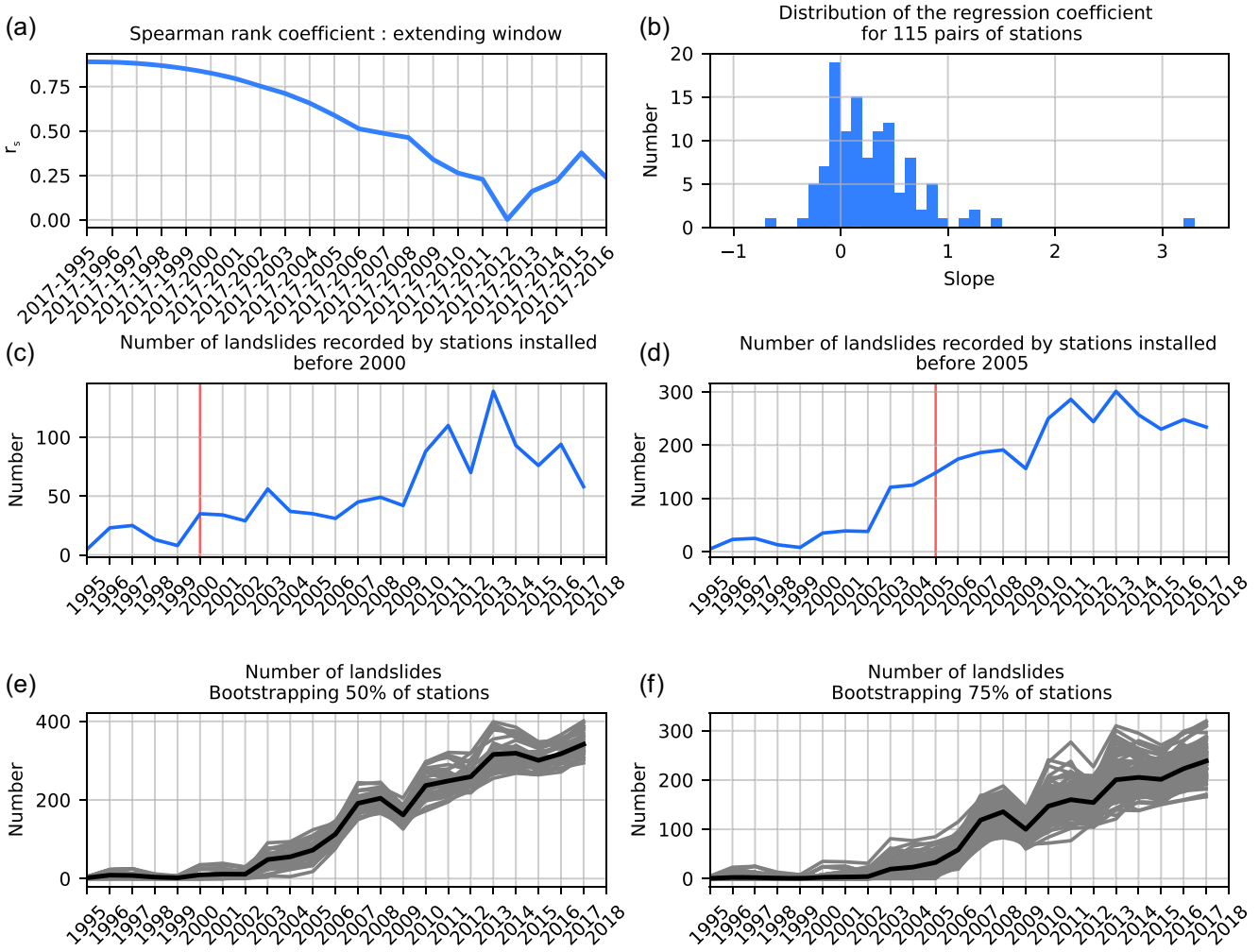

Figure 5. (a) Spearman's rank correlation coefficient r_s computed on an extending window; (b) distribution of the values of the slope of the regression line of the number of landslides per year and per pair of stations that are active since 2005; (c) number of landslides per year detected by 2 or more stations deployed before 2000; (d) number of landslides per year detected by 2 or more stations deployed before 2000 and number of landslides per year detected by 2 or more stations while removing randomly; (e) 50 per cent of stations and (f) 75 per cent of stations of the data set. This process was repeated 100 times for each case, producing at each iteration a grey line. The black line is the mean over the 100 iterations of the number of landslides detected per year.

through seismic observation. Having catalogues that integrate not only very large events is important to better assess and apprehend the interaction between forcings and landslide triggering at the local, regional and even global scale.

The application of the processing chain to the exploration of 22 yr (from 1995 to 2017) of continuous data recorded by 243 stations of seismic networks in Alaska yielded a catalogue of 6213 landslide detections, among which we identify 5087 events for which the seismic signals have the general features of those generated by landslide sources, thus giving an effective success rate of identification of 82 per cent. We observe that the number of detected potential landslides is increasing through the period studied. The different analyses we conducted on the link between the increasing number of stations deployed and the increasing number of landslides with time indicate that the former might not be the primary driver of the latter. We also show that the increase in the number of landslides is distributed through the whole zone covered by the seismic network.

We observe a possible correlation between the year-to-year increase in average temperature and the increase in the number of potential landslides at the regional scale. However, a more quantitative approach focusing on smaller temporal and spatial scales, which is now possible because of the high spatio–temporal resolution of the catalogue derived from our seismological analysis, is needed. This short-term analysis might provide insights to understand the chronology of landslide triggering and to quantify the healing of the affected slopes during episodes of intense landsliding in relation with different forcings (heat waves, glacial retreat, intense precipitations or strong earthquakes) but also with the local geological and geomorphological contexts, meteorological data and regional climate models. This will require locating the landslides that will also permit to infer some of their properties (volume, dynamics, etc.) and study their potential evolution as a function of external forcings. However, due to the complex nature of landslide generated seismic waves, locating landslides from high-frequency seismic signals is still challenging, and will be addressed in future works.

Applying automated machine learning techniques to seismological observations constitutes a promising step forward to study environmental processes at regional scales such as landslides, but also

ice-calving events, pit crater formation in high-latitude regions, and possibly triggering of submarines slumps. The approach allows documenting chronicles of events over extended time periods and might provide a new mean for quantifying long-term changes. It also opens the possibility for near-real-time detection of events from the global and regional seismic networks deployed all over the Earth.

ACKNOWLEDGEMENTS

This work was carried with the support of the French National Research Agency (ANR) through the projects HYDROSLIDE 'Hydrogeophysical Monitoring of Clayey Landslides, ANR-15-CE04-0009' and TIMES 'High-performance processing techniques for mapping and monitoring environmental changes from massive, heterogeneous and high-frequency data times series, ANR-17-CE23-0015-01'. Support of the TelluS-ALEAS program of the French National Institute of Sciences of the Universe (CNRS-INSU) is acknowledged. We thank the operators of the AK (http://dx.doi.org/10.7914/SN/AK), AT (http://dx.doi.org/10.7914/SN/AT), AV (http://dx.doi.org/10.7914/SN/AV), CN (http://dx.doi.org/10.7914/SN/CN), II (http://dx.doi.org/10.7914/SN/II), IU (http://dx.doi.org/10.7914/SN/IU), IM, US (http://dx.doi.org/10.7914/SN/US) and TA (http://dx.doi.org/10.7914/SN/TA) seismic networks for data collection and the IRIS Data Management System for providing easy access to the data. We also thank the operators of the NOAA's NCDC for the acquisition and the sharing of the meteorological data.

The authors gratefully acknowledge Florian Fuchs and an anonymous reviewer for insightful reviews and very interesting discussions. We also would like to thank the editors for their comments, which helped to improve this manuscript.

Author contributions: CH collected the data, developed the processing chain, validated and analysed the catalogue and produced the figures. DM performed the HPC implementation of the processing chain. FP contributed to the implementation of the machine learning algorithm. J-PM and MG provided the climatic and meteorological data. All the authors shared ideas, contributed to the interpretation of the results and to the writing of the manuscript.

Seismic data are available through the IRIS portal (http://ds.iris.edu/SeismiQuery/). Meteorological data are available through the NOAA's NCDC website (https://www.ncdc.noaa.gov/cdo-web/search). The codes developed for this study can be accessed by contacting the first author.

REFERENCES

Allen, R., 1982. Automatic phase pickers: Their present and future prospects, *BSSA,* **72**(6B), S225–S242.

Allstadt, K., 2013. Extracting source characteristics and dynamics of the August 2010 Mount Meager landslide from broad-band seismograms, *J. geophys. Res.,* **118**(3), 1472–1490.

Allstadt, K.E., McVey, B. & Malone, S., 2017. Seismogenic landslides, debris flows, and outburst floods in the western United States and Canada from 1977 to 2017: US geological survey data release, doi:10.5066/F7251H3W.

Allstadt, K.E., Matoza, R.S., Lockhart, A., Moran, S.C., Caplan-Auerbach, J., Haney, M., Thelen, W.A. & Malone, S.D., 2018. Seismic and acoustic signatures of surficial mass movements at volcanoes, *J. Volcanol. Geotherm. Res.,* **364,** 76–106.

Baillard, C., Crawford, W.C., Ballu, V., Hibert, C. & Mangeney, A., 2014. An automatic kurtosis-based p-and s-phase picker designed for local seismic networks, *Bull. seism. Soc. Am.,* **104**(1), 394–409.

Breiman, L., 2001. Random forests, *Mach. Learn.,* **45**(1), 5–32.

Brodsky, E.E., Gordeev, E. & Kanamori, H., 2003. Landslide basal friction as measured by seismic waves, *Geophys. Res. Lett.,* **30**(24), 2236.

Coe, J.A. & Godt, J., 2012. Review of approaches for assessing the impact of climate change on landslide hazards, in *Landslides and Engineered Slopes, Protecting Society Through Improved Understanding: Proceedings of the 11th International and 2nd North American Symposium on Landslides and Engineered Slopes,* Vol. **1,** pp. 371–377, eds Eberhardt, E., Froese, C., Turner, A.K. & Leroueil, S., Taylor & Francis Group.

Coe, J.A., Bessette-Kirton, E.K. & Geertsema, M., 2018. Increasing rock-avalanche size and mobility in 479 Glacier Bay National Park and Preserve, Alaska detected from 1984 to 2016 Landsat imagery, *Landslides,* **15,** 393–407.

Cortés, G., Arámbula, R., Gutiérrez, L., Benítez, C., Ibánez, J., Lesage, P., Alvarez, I. & Garcia, L., 2009. Evaluating robustness of a HMM-based classification system of volcano-seismic events at colima and popocatepetl volcanoes, in *2009 IEEE International Geoscience and Remote Sensing Symposium,* Vol. 2, Cape Town, South Africa, pp. 1012–1015.

Dammeier, F., Moore, J.R., Haslinger, F. & Loew, S., 2011. Characterization of alpine rockslides using statistical analysis of seismic signals, *J. geophys. Res.,* **116,** F04024.

Dammeier, F., Moore, J.R., Hammer, C., Haslinger, F. & Loew, S., 2016. Automatic detection of alpine rockslides in continuous seismic data using hidden Markov models, *J. geophys. Res.,* **121**(2), 351–371.

Deparis, J., Jongmans, D., Cotton, F., Baillet, L., Thouvenot, F. & Hantz, D., 2008. Analysis of rock-fall and rock-fall avalanche seismograms in the French Alps, *Bull. seism. Soc. Am.,* **98**(4), 1781–1796.

Dietze, M., Mohadjer, S., Turowski, J.M., Ehlers, T.A. & Hovius, N., 2017. Seismic monitoring of small alpine rockfalls—validity, precision and limitations, *Earth Surf. Dyn.,* **5**(4), 653–668.

Dufresne, A. *et al.*, 2017. Sedimentology and geomorphology of a large tsunamigenic landslide, Taan Fiord, Alaska, *Sedimentary Geol.,* **364,** 302–318.

Ekström, G. & Stark, C.P., 2013. Simple scaling of catastrophic landslide dynamics, *Science,* **339,** 1416–1419.

Favreau, P., Mangeney, A., Lucas, A., Crosta, G. & Bouchut, F., 2010. Numerical modeling of landquakes, *Geophys. Res. Lett.,* **37,** L15305, doi:10.1029/2010GL043512.

Fuchs, F., Lenhardt, W. & Bokelmann, G., 2018. Seismic detection of rockslides at regional scale: examples from the eastern alps and feasibility of kurtosis-based event location, *Earth Surf. Dyn.,* **6**(4), 955–970.

Geertsema, M., Clague, J.J. & Hasler, A., 2013. Influence of climate change on geohazards in Alaskan parks, *Alaska Park Sci.,* **12,** 80–85.

George, D., Iverson, R. & Cannon, C., 2017. New methodology for computing tsunami generation by subaerial landslides: application to the 2015 Tyndall Glacier landslide, Alaska, *Geophys. Res. Lett.,* **44**(14), 7276–7284.

Gualtieri, L. & Ekström, G., 2018. Broad-band seismic analysis and modeling of the 2015 Taan Fjord, Alaska landslide using instaseis, *Geophys. J. Int.,* **213**(3).

Gualtieri, L., Camargo, S.J., Pascale, S., Pons, F.M. & Ekström, G., 2018. The persistent signature of tropical cyclones in ambient seismic noise, *Earth planet. Sci. Lett.,* **484,** 287–294.

Hammer, C., Beyreuther, M. & Ohrnberger, M., 2012. A seismic-event spotting system for volcano fast-response systems, *Bull. seism. Soc. Am.,* **102**(3), 948–960.

Helmstetter, A. & Garambois, S., 2010. Seismic monitoring of Séchilienne rockslide (French Alps): analysis of seismic signals and their correlation with rainfalls, *J. geophys. Res.,* **115,** F03016, doi:10.1029/2009JF001532.

Hibert, C., Mangeney, A., Grandjean, G. & Shapiro, N.M., 2011. Slope instabilities in Dolomieu crater, Réunion Island: from seismic signals to rockfall characteristics, *J. geophys. Res.,* **116,** F04032, doi:10.1029/2011JF002038.

Hibert, C., Ekström, G. & Stark, C.P., 2014a. Dynamics of the Bingham Canyon Mine landslides from seismic signal analysis, *Geophys. Res. Lett.,* **41,** 4535–4541.

Hibert, C. *et al.*, 2014b. Automated identification, location, and volume estimation of rockfalls at Piton de la Fournaise volcano, *J. geophys. Res.,* **119**(5), 1082–1105.

Hibert, C., Stark, C.P. & Ekström, G., 2014c. Seismology of the Oso-Steelhead landslide, *Nat. Hazards Earth Syst. Sci. Discuss.,* **2,** 7309–7327.

Hibert, C., Ekström, G. & Stark, C.P., 2017a. The relationship between bulk-mass momentum and short-period seismic radiation in catastrophic landslides, *J. geophys. Res.,* **122**(5), 1201–1215.

Hibert, C., Provost, F., Malet, J.-P., Maggi, A., Stumpf, A. & Ferrazzini, V., 2017b. Automatic identification of 535 rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a Random Forest 536 algorithm, *J. Volcanol. Geotherm. Res.,* **340,** 130–142.

Huggel, C., Clague, J.J. & Korup, O., 2012. Is climate change responsible for changing landslide activity in high mountains? *Earth Surf. Process. Landf.,* **37**(1), 77–91.

Ibáñez, J.M., Benítez, C., Gutiérrez, L.A., Cortés, G., García-Yeguas, A. & Alguacil, G., 2009. The classification of seismo-volcanic signals using hidden Markov models as applied to the Stromboli and Etna volcanoes, *J. Volcanol. Geotherm. Res.,* **187**(3), 218–226.

Kanamori, H. & Given, J.W., 1982. Analysis of long-period seismic waves excited by the May 18, 1980, eruption of Mount St. Helens—A terrestrial monopole? *J. geophys. Res.,* **87**(B7), 5422–5432.

Kanamori, H., Given, J.W. & Lay, T., 1984. Analysis of seismic body waves excited by the Mount St. Helens eruption of May 18, 1980, *J. geophys. Res.,* **89**(B23), 1856–1866.

Kirschbaum, D.B., Adler, R., Hong, Y., Hill, S. & Lerner-Lam, A., 2010. A global landslide catalogue for hazard 548 applications: method, results, and limitations, *Nat. Hazards,* **52,** 561–575.

Langer, H., Falsaperla, S., Powell, T. & Thompson, G., 2006. Automatic classification and a-posteriori analysis of seismic event identification at Soufriere Hills volcano, Montserrat, *J. Volcanol. Geotherm. Res.,* **153**(1), 1–10.

Langet, N., 2014. Détection et caractérisation massives de phénomènes sismologiques pour la surveillance d'événements traditionnels et la recherche systématique de phénomènes rares, *PhD thesis,* Université de Strasbourg.

Levy, C., Mangeney, A., Bonilla, F., Hibert, C., Calder, E.S. & Smith, P.J., 2015. Friction weakening in granular flows deduced from seismic records at the Soufrière Hills Volcano, Montserrat, *J. geophys. Res.,* **120**(11), 7536–7557.

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P. & Amemoutou, A., 2017. Implementation of a multistation approach for automated event classification at Piton de la Fournaise volcano, *Seismol. Res. Lett.,* **88,** 878–891.

Malfante, M., Dalla Mura, M., Mars, J.I., Métaxian, J.-P., Macedo, O. & Inza, A., 2018. Automatic classification of volcano seismic signatures, *J. geophys. Res.,* **123,** 12, 10 645–10 658.

Moore, J.R., Pankow, K.L., Ford, S.R., Koper, K.D., Hale, J.M., Aaron, J. & Larsen, C.F., 2017. Dynamics of the Bingham Canyon rock avalanches (Utah, USA) resolved from topographic, seismic, and infrasound data, *J. geophys. Res.,* **122**(3), 615–640.

Moretti, L., Mangeney, A., Capdeville, Y., Stutzmann, E., Huggel, C., Schneider, D. & Bouchut, F., 2012. Numerical modeling of the Mount Steller landslide flow history and of the generated long period seismic waves, *Geophys. Res. Lett.,* **39,** L16402, doi:10.1029/2012GL052511.

Provost, F., Hibert, C. & Malet, J.-P., 2017. Automatic classification of endogenous landslide seismicity using 570 the Random Forest supervised classifier, *Geophys. Res. Lett.,* **44,** 113–120.

Schöpa, A., Wei-An, C., Lipovsky, B.P., Hovius, N., White, R.S., Green, R.G. & Turowski, J.M., 2018. Dynamics of the Askja Caldera July 2014 landslide, Iceland, from seismic signal analysis: precursor, motion and aftermath, *Earth Surf. Dyn.,* **6**(2), 467.

Sergeant, A., Mangeney, A., Stutzmann, E., Montagner, J.-P., Walter, F., Moretti, L. & Castelnau, O., 2016. Complex force history of a calving-generated glacial earthquake derived from broad-band seismic inversion, *Geophys. Res. Lett.,* **43**(3), 1055–1065.

Suriñach, E., Vilajosana, I., Khazaradze, G., Biescas, B., Furdada, G. & Vilaplana, J.M., 2005. Seismic detection and characterization of landslides and other mass movements, *Nat. Hazards Earth Syst. Sci.,* **5,** 791–798.

Vilajosana, I., Suriñach, E., Abellán, A., Khazaradze, G., Garcia, D. & Llosa, J., 2008. Rockfall induced seismic signals: case study in Montserrat, Catalonia, *Nat. Hazards Earth Syst. Sci.,* **8,** 805–812.

Yamada, M., Kumagai, H., Matsushi, Y. & Matsuzawa, T., 2013. Dynamic landslide processes revealed by broadband seismic records, *Geophys. Res. Lett.,* **40,** 2998–3002.

Zimmer, V.L. & Sitar, N., 2015. Detection and location of rock falls using seismic and infrasound sensors, *Eng. Geol.,* **193,** 49–60.