Enhanced glacial earthquake catalogues with supervised machine learning for more comprehensive analysis

Emilie Pirot, Clément Hibert, Anne Mangeney

Geophysical Journal International

Geophys. J. Int. (2024) **236,** 849–871
Advance Access publication 2023 October 23
GJI Seismology

https://doi.org/10.1093/gji/ggad402

Enhanced glacial earthquake catalogues with supervised machine learning for more comprehensive analysis

Emilie Pirot [1], Clément Hibert [2] and Anne Mangeney[1]

[1] *Institut de Physique du Globe de Paris, CNRS, Université de Paris-Cité, Paris* 75005*, France. E-mail: emilie.pirot@gmail.com*
[2] *Institut Terre Environnement de Strasbourg, CNRS UMR* 7063*, University of Strasbourg/EOST, Strasbourg* 67000*, France*

Accepted 2023 October 10. Received 2023 August 7; in original form 2023 March 29

SUMMARY

Polar regions and Greenland in particular are highly sensitive to global warming. Impacts on Greenland's glaciers may be observed through the increasing number of calving events. However, a direct assessment of the calving activity is limited due to the remoteness of polar regions and the cloudy weather which makes impossible a recurrent observation through satellite imagery. To tackle this issue, we exploit the seismological network deployed in Greenland which actively records seismic signals associated with calving events, hereinafter referred to as glacial earthquakes. These seismic signals present a broad frequency range and a wide diversity of waveform which make them difficult to discriminate from tectonic events as well as anthropogenic and natural noises. In this study, we start from two catalogues of known events, one for glacial earthquake events which occurred between 1993 and 2013 and one for earthquakes which occurred in the same time period, and we implement a detection algorithm based on the STA/LTA method to extract signals' events from continuous data. Then, we train and test a machine learning processing chain based on the Random Forest algorithm which allows us to automatically associate the events respectively with calving and tectonic activity, with a certain probability. Finally, we investigate 844 selected days spanning time of continuous data from the Greenland regional seismic network which results in a new, more exhaustive, catalogue of glacial earthquakes expanded of 1633 newly detected glacial events. Moreover, we extensively discuss the choice of the features used to describe glacial earthquakes, in particular the 39 new features created in this study which have drastically improved our results with 7 of the 10 best features being in the added set. The perspective of further expansion of the glacial earthquake catalogue applying the processing chain discussed in this paper on different time spans highlights how combining seismology and machine learning can increase our understanding of the spatio-temporal evolution of calving activity in remote regions.

Key words: Arctic region; Machine learning; Glaciology; Seismology.

1 INTRODUCTION

The global rise in temperature due to climate change has an immediate impact on the polar ice sheets (Amundson *et al.* 2008; Aster & Winberry 2017). Indeed, the ice budget of the polar ice sheets is balanced between snow gain and ice loss, due to the melting of ice sheets and calving of icebergs. Iceberg discharge at the terminus of ice sheet glaciers is a major component of ice loss (Podolskiy & Walter 2016; Sergeant *et al.* 2019). Calving occur on average 20 times a year, according to the 1993–2013 catalogue of events from Nettles & Ekström (2010) and Olsen & Nettles (2017), but other phenomena such as ice avalanches and calving of small icebergs are more frequent and actively participate in Greenland's overall mass loss (Amundson *et al.* 2008). To overcome the difficulty of direct observations due to the remoteness of this region, seismology has been a preferred tool. It provides information on glacier behaviour, as well as on the physics of the source of particular events and on the composition of the ice cap (Aster & Winberry 2017; Sergeant *et al.* 2018).

In the 2000s, unidentified low frequency signals located near Greenland glaciers termini were discovered using a detection algorithm based on long-period surface waves (35–150 s; Ekström *et al.* 2003). These high-magnitude events (MSW in the order of of 4.5–5.5), recorded at very low frequency as teleseismic events, were unexpected in a non-tectonic region. Linked to the calving of large icebergs in the analysis conducted by Tsai & Ekström (2007), these events, generating signals named glacial earthquakes, were studied by modelling (Sergeant *et al.* 2018; Bonnet *et al.* 2020) and inverted to enable finer location of these events and creation of catalogues for different periods (Ekström *et al.* 2003; Ekström 2006;

Figure 1. (a) Picture modified from Chasing Ice Calving by Jeff Orlowski showing an event occurring on 28 May 2008 at Jakobshavn Isbrae. The black arrow represents the rotation of the detaching iceberg. The red arrow represents the force applied on the terminus. (b) Schematic side representation of the phenomenon where buoyancy and gravity forces are represented with blue arrows.

Tsai & Ekström 2007; Nettles *et al.* 2008; Olsen & Nettles 2019). However, these catalogues only include high-magnitude events (at least 4.9 MSW). A comprehensive catalogue is essential for a better understanding of the spatio-temporal evolution of calving events in Greenland, as the contribution of smaller events could increase the dynamic mass loss of Greenland glaciers not accounted for in current catalogues by as much as 10–30 per cent (Olsen & Nettles 2019). To address this issue, we have developed an algorithm to detect smaller events in order to create a more comprehensive catalogue.

Fig. 1 illustrates the phenomenon of glacial earthquakes. In this case, the top of the unstable iceberg pushes against the front of the glacier, creating a force on the iceberg terminus that varies in time with the rotation of the iceberg, generating waves in the glacier that are then transmitted to the Earth. These so-called 'bottom-out' events generate a seismic signal of lower amplitude than 'top-out' iceberg calving (the bottom of the iceberg reaches the glacier during rotation) of the same volume, showing that interpreting seismic amplitude or energy in terms of iceberg size may be misleading (Sergeant *et al.* 2018). However, iceberg calving simulation could be used to reproduce the inverted force from seismic data and lead to iceberg volume catalogues used to study the spatio-temporal evolution of ice mass loss in relation to climate change (Sergeant *et al.* 2019). Applying this method to smaller events would considerably improve quantification of ice mass loss on marine-terminating glaciers.

A capsizing iceberg is subject to the forces of buoyancy and gravity, as well as drag (Amundson *et al.* 2010; Burton *et al.* 2012; Sergeant *et al.* 2018; Bonnet *et al.* 2020). As it breaks away from the terminus, the iceberg can become unstable due to its height/width ratio, causing it to calve, that is to tend toward a more stable state. During this rotation, the iceberg reaches the terminus of the glacier. The force F_c represents the force at the source of the seismic event (Fig. 1).

A catalogue of glacial earthquakes that occurred between 1993 and 2013 has been presented by Tsai & Ekström (2007), Veitch & Nettles (2012) and Olsen & Nettles (2017), and it will be referred to as Columbia thereinafter. It groups 444 located glacial earthquakes, which are represented on Fig. 2 by stars of different colours, together with the seismic stations present in the perimeter at the time of the event (some stations are no longer operating today). These events are grouped into 16 zones, indicated on Fig. 2 and coloured according to the closest glacier. The most active glaciers in Greenland are Helheim Glacier (light green stars area on Fig. 2) and Jakobshavn Isbrae Glacier (purple stars on Fig. 2) on the east coast, and Kangerlussuaq Glacier (orange stars on Fig. 2) on the west coast. The GLISN regional network, whose stations are shown in red on Fig. 2, records cryo-seismic activity, but also tectonic activity in Iceland as well as anthropogenic noise (Podolskiy & Walter 2016; Aster & Winberry 2017).

The aim of this study is to extract glacial earthquake signals from continuous data and discriminate them from other recorded signals. The method used to date to detect and identify glacial earthquakes requires human supervision (Olsen & Nettles 2017) but the growing number of stations and data compromises the time/efficiency of this approach, particularly when analysing seismicity in frequency band above 1 Hz, with lots of local noise and numerous small natural events likely to generate a seismic signal. What's more, since denser networks mean larger quantities of data, it is all the more difficult to process them with a conventional tool. In this study, we develop a processing chain based on a supervised machine learning algorithm called Random Forest, in order to detect and identify new glacial earthquakes occurring in Greenland.

2 DATA AND METHODS

The first phase of our study aims at creating an automated processing chain able to detect signals of glacial earthquakes and earthquakes from continuous data. We start with gathering data from these two types of events from existing catalogues.

2.1 Data

2.1.1 Catalogues

For the glacial earthquake data set, we use events from the existing glacial earthquake catalogue created and then expanded by Tsai & Ekström (2007), Nettles & Ekström (2010) and Olsen & Nettles (2017). In this catalogue, 444 events occurring between 1993 and 2013 are collated with location coordinates and estimated magnitude (Fig. 2). The first glacial earthquake identified in this catalogue occurred on 24 January 1993 with a magnitude of 5.1 MSW, and the last one occurred on 27 December 2013 with a magnitude of 4.9 MSW. The MSW magnitude is calculated based on teleseismic

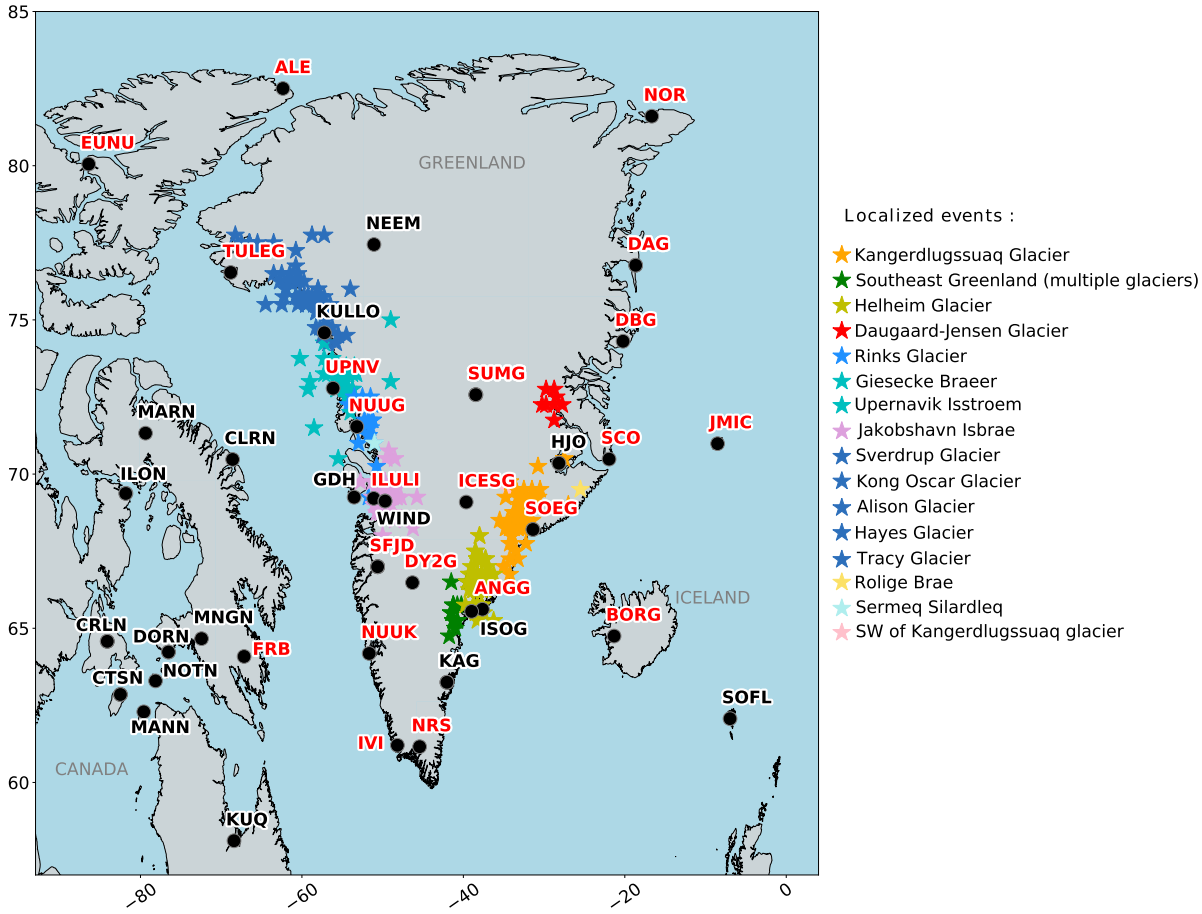

Figure 2. Seismic network around Greenland. Names of stations are written next to their location. Stations in red are part of the GLISN network. Stars of different colours correspond to different glaciers. Represented events are from the initial catalogue (Columbia 2007).

Rayleigh waves amplitudes (Ekström *et al.* 2003) and the magnitude range extends from 4.6 to 5.1 MSW.

We supplement the catalogue of glacial earthquakes with 400 earthquakes that occurred over the same period (1993–2013) and are of the same order of magnitude. Three examples of raw and filtered seismic signals generated by glacial earthquakes and earthquakes are shown in Fig. 3, together with spectrograms for each event. Events are highlighted by a colour (blue for glacial earthquakes and brown for earthquakes). The spectrogram covers frequencies from 4 to 0.001 Hz. Fig. 3(a) illustrates two events recorded by the same SOEG seismic station. The glacial earthquake occurred on 20 December 2001 at 7:27 a.m. and was located on the Kangerdlugssuaq glacier, 81 km from the station. The magnitude (MSW) of this event was 4.8. The earthquake also recorded by SOEG, with a magnitude of 5.4, occurred in Greenland on 26 August 2001 at 6:28 p.m. (1600 km from the station). Both events are indicated by stars on the map and on the station. In Fig. 3(b), the glacial earthquake occurred on 6 July 2004 at 10:20 a.m. on the Helheim glacier (85 km from ANGG) with a magnitude of 4.7 MSW. The earthquake also recorded at ANGG, with a magnitude of 5.1, occurred in the Iceland region on 19 July 2003 at 1:13 p.m. (685 km from the station). In Fig. 3(c), the glacial earthquake occurred on 25 December 2012 at 10:21 a.m. and was located at Rinks Glacier (85 km from NOR) with a magnitude of 4.7 MSW.

The duration of glacial earthquakes can vary from a few minutes to several tens of minutes. Their seismic signals also present a variety of waveforms, depending on iceberg geometry and volume. Very low frequencies ($>$ 100 s, $<$ 0.01 Hz) have been attributed to the movement of the seiche after the iceberg capsized (Sergeant et al. 2016). Earthquakes have clearer, more identifiable phases. In the last earthquake example shown in Fig. 3 (c), a very low frequency content is highlighted (pinkish brown) preceding a more impulsive signal. Glacial earthquakes tend to have a more emergent signal, often more identifiable in the 0.01-0.02Hz frequency band. These examples illustrate the need to investigate the different frequency bands for effective classification.

2.1.2 *Seismic network*

We work with raw data acquired by the Greenland Ice Sheet Monitoring Network (GLISN) and with all available stations, permanent or otherwise, over the period 1993–2013. The GLISN network, which began installation in 2005, now comprises 33 stations: 20 on Greenland and 13 off Greenland. Until 2003, there were only 7 permanent stations on the island, which were subsequently upgraded through integration into the GLISN network. Non-permanent stations located within a defined perimeter around the island ([57°N, −94°W; 83°N, −4°W]) were also used when data were available (Fig. 2). We only worked with stations equipped with three-component broad-band seismometers. Available and archived

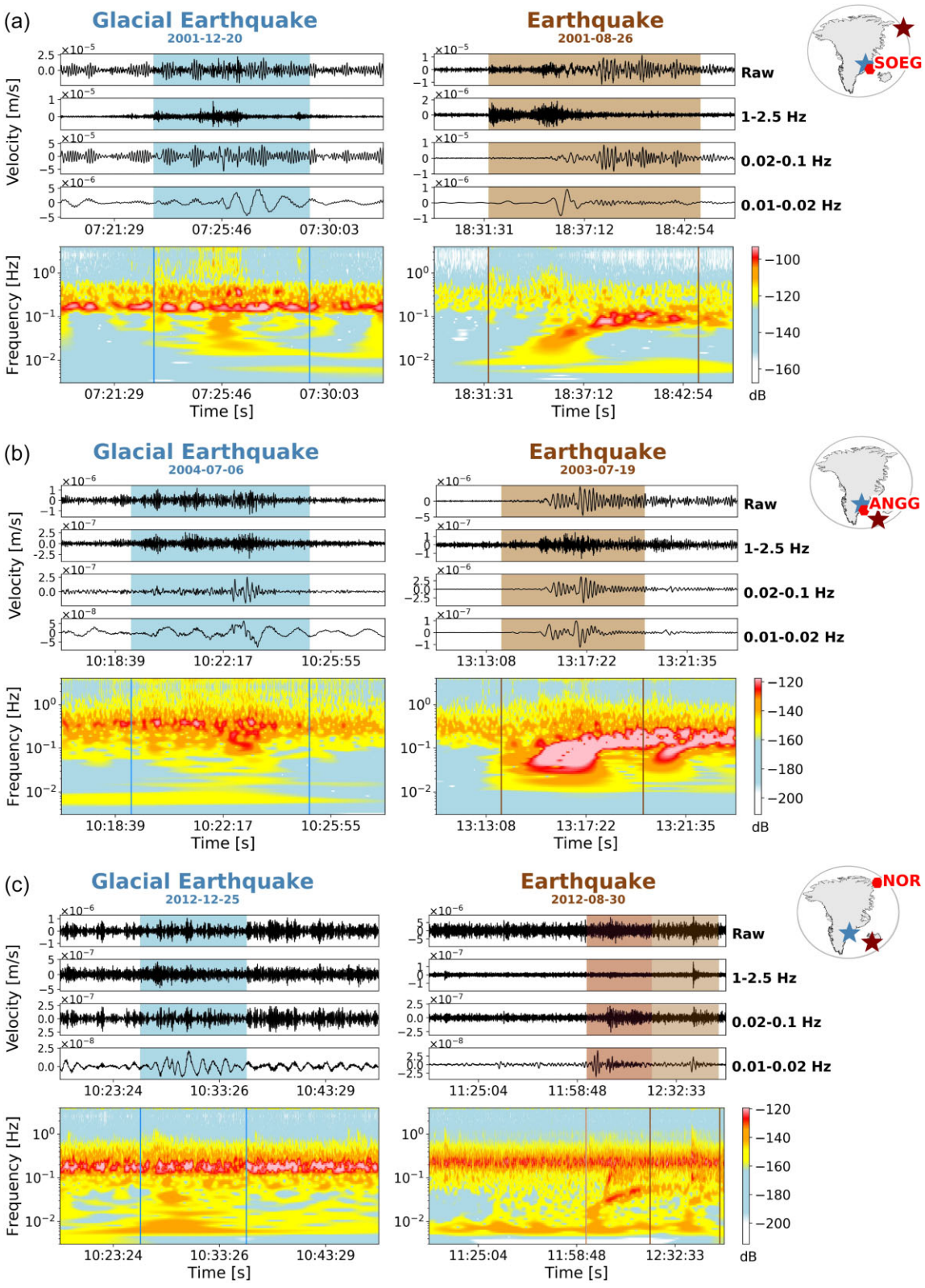

Figure 3. Example of seismic signal of glacial earthquakes and earthquakes from the two initial catalogues (Columbia 2007; USGS 2021) (raw data, filtered data in 1–2.5 Hz, in 0.02–0.1 Hz and in 0.01–0.02 Hz, and spectrogram). The location of events is represented by a brown star for earthquakes and a blue star for glacial earthquakes, and events on the same line are recordings from the same station (red hexagon on the map).

broad-band seismic data from these stations were downloaded from the IRIS and GFZ data centres. Restrictions on Iceland's temporary networks were added, as they are close to volcanic and tectonic zones, forcing us to process data not related to glacial activities.

2.1.3 Pre-processing

For each of the 844 events in the two catalogues (444 glacial earthquakes (Columbia 2007) and 400 earthquakes (USGS 2021)), we download available data 24 hr around the precise time of the event indicated in the catalogues. The instrumental response is removed before the data is processed. Several events may occur on the same day.

2.2 Methods

2.2.1 Detection algorithm

Automatic picking of glacial earthquake signals is complicated because of the diversity of waveforms. We decided to extract the signals from the raw 24-hr waveforms centred on each event in the catalogues. Our aim is to retrieve all signals from initial catalogues events (Columbia 2007; USGS 2021) as well as signals from new events, using the same detection algorithm. To achieve this, we have used a standard detection algorithm based on sliding windows of the short-term average/long-term average (STA/LTA) type, which we have adapted to glacial earthquake signals and earthquake signals. In this approach, a dimensionless ratio is calculated between the amplitude of the signal averaged over a short time window (STA) and that of the signal averaged over a long time window (LTA). When the ratio exceeds a user-defined threshold, the start time is recovered. When the ratio falls below another defined threshold, the signal end time is selected.

Glacial earthquake and earthquake signals have wide frequency contents (as shown in Fig. 3), so we calculated the classical STA/LTA over four frequency bands to cover a wide frequency range with different threshold values. The window lengths are set at 100 seconds for the short window and 1900 seconds for the long window. The frequency bands chosen are 1 Hz–Nyquist frequency, 0.02–0.1 Hz and 0.01–0.02 Hz with corresponding On/Off thresholds: 6/3, 3/1 and 3.5/2. Raw data are processed with 2/1 thresholds. The sum of all filtered STA/LTA, called 'Stack', are used with On/Off thresholds of 8/2. In Fig. 4(a), the triggers are shown in different colours corresponding to the frequency band in which the standard STA/LTA was performed.

The event signal can be detected by one or more STA/LTAs. For example, at station BORG, the STA/LTA calculated in the 0.01–0.02 Hz frequency band and the stack STA/LTA (triggers shown in pink and purple respectively in Fig. 4a) frame the catalogue event signal. Other signals are nevertheless detected, as on the SUMG station recording where the STA/LTA calculated from the raw data detects a signal after the known event (blue triggers in Fig. 4a).

In Fig. 4(b), the On/Off triggers (in blue and red) correspond to the merging of detections obtained in all intersecting STA/LTAs. The minimum and maximum trigger times of the intersecting detections are retained. With this method, all events in the original catalogues are detected (highlighted in pale red in Fig. 4b). Signals that are not filled in with a pale red band in Fig. 4(b) do not correspond to catalogue events: these may be new events, seismic signals linked to other sources or noise.

2.2.2 Association

When seismic signals are detected at several stations at a consistent time interval, they are grouped together to form an event. For example, in Fig. 4(b), the event will be considered as detected by 5 stations. In the same figure, other signals have been detected (Fig. 4b): they will be grouped together to form an event if the detections have a coherent arrival time at more than two stations. Seismic waves generated by glacial earthquakes are surface waves (Ekström 2006), so we use a propagation velocity for low-frequency surface waves in Greenland of $Vs = 3300\ \mathrm{m\,s^{-1}}$ given by Kumar *et al.* (2007) to calculate a theoretical arrival time, based on the distance between two stations. The arrival times of the detections are highly dependent on the STA/LTA triggers (as shown in Fig. 4b). The use of several frequency bands to detect a signal means that signal arrival times can vary greatly from station to station, as higher frequencies are rapidly attenuated. We added a 3-min buffer to group signals into events. This may appear to be a lot, but events last several tens of minutes. This choice of 3-min was tested on signals from known events detected with the STA/LTA and enabled us to associate at least two signals from different stations in order to retrieve all the initial events.

2.2.3 Random Forest

With a view to rapid classification of the newly detected events, we choose a supervised machine learning algorithm called 'Random Forest' developed by Breiman (2001). This algorithm has been used in the past to classify seismic signals of volcanic origin (Hibert *et al.* 2017; Malfante 2018; Falcin *et al.* 2020) as well as other seismic signals of natural origin (Dong *et al.* 2014; Provost *et al.* 2017; Malfante 2018; Hibert *et al.* 2019; Lin *et al.* 2020; Chmiel *et al.* 2021).

To process the data, the algorithm relies on a mathematical description of the waveforms, called features, which are, for example, signal kurtosis, envelope skewness, energy in certain frequency bands or signal duration. If the data is described with too few or irrelevant features, the algorithm may miss important information, resulting in poor performance. In this study, we work with 97 features, detailed in the appendix. The first 58 features are taken from previous studies aimed at discriminating seismic signals from various natural processes and can be divided into three families: waveform features, spectral features and spectrogram features. We have created 39 new features, designed to capture information on low frequencies and on the variation in intensity between two frequency bands. Combinations of energy differences and energy ratios in these frequency bands are calculated: 0.01–0.02 Hz, 0.01–0.05 Hz, 0.05–0.1 Hz, 0.1–1 Hz, 1–2 Hz and 2 Hz–Nyquist. Finally, we added a ratio based on the calculation of the standard deviation of the data, which can be assimilated to a signal-to-noise ratio.

The Random Forest algorithm is a supervised machine learning algorithm. In contrast to unsupervised algorithms, where no labels are predefined, the data set under study is labelled by user-defined classes. To assign a class to each element in the initial database, the algorithm uses decision trees. Each tree is created from a random subset of elements from the training set and a random selection of features to form a 'random forest'. Each decision tree in the 'forest' is therefore unique (Breiman 2001). Each tree assigns a class to each feature in the database, and the prediction assigned to the majority of the features is the final prediction for the feature. Each tree therefore contributes to the final prediction. We are working

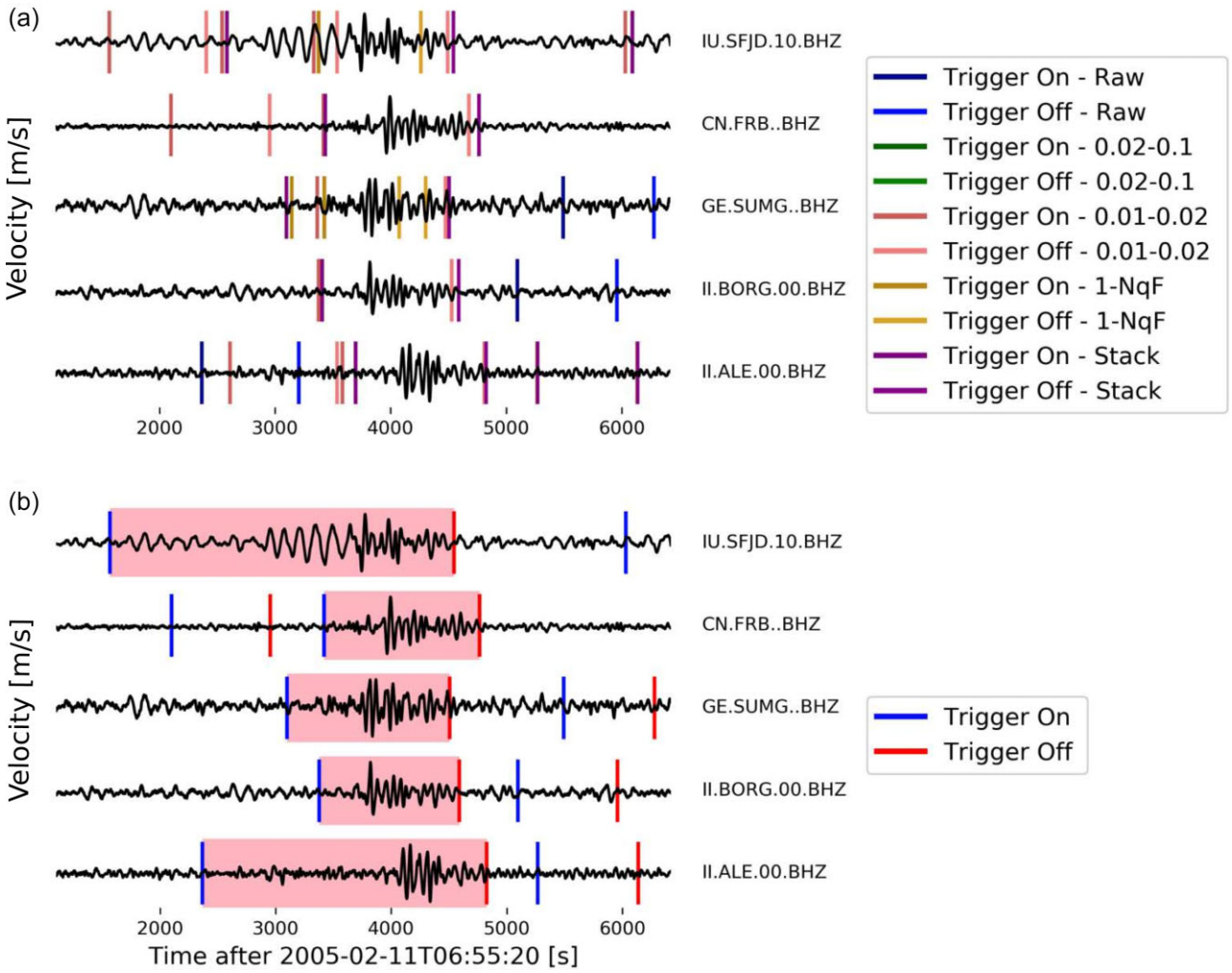

Figure 4. Filtered data in 35–150 s/0.01–0.02 Hz, after removing instrumental response for five stations: SFJD, FRB, SUMG, BORG and ALE, recorded on 11 February 2005 at 06:55, which correspond to a catalogue event, occurring at Helheim Glacier (Fig. 2, light green stars). All triggers (On and Off) for all frequency bands explored with the different STA/LTA (Raw, 0.02–0.1 Hz, 0.01–0.02 Hz, 1–NqF Hz, Stack). (b) Merging of triggers to form detections (final trigger: On-Off). In light red, seismic signals which corresponds to the catalogue event. Note that classical STA/LTA was not triggered in the 0.02-0.1Hz band for this signal, hence the absence of green triggers in the upper part of the figure.

with 500 trees, as this parameter has proved robust in the studies cited above.

The Random Forest algorithm can handle a large number of features and has the advantage of evaluating the impact of each feature by assigning importance scores. This importance score thus links the features of the seismic signal and the mathematical description of the source: it can therefore improve the physical interpretation of the phenomenon.

Once the features have been calculated for all the seismic signals, we divide the data set into a training set and a testing set in order to train and test the Random Forest model. The training set consists of a percentage of the main data set with signals from both classes (glacial earthquake and earthquake) and is used to build trees capable of discriminating between the two types of signal. Then, the model is applied to the remaining data (testing set). Results are presented in the form of confusion matrices that compare the number (or percentages) of data in the class with the data predicted to be in that class.

2.2.4 Workflows

Fig. 5 illustrates the algorithm steps in the processing chain, from the two initial catalogues (Columbia 2007; USGS 2021) to the final one with the new events found. After pre-processing the data and training the previously detailed Random Forest model, we focus here on the post-processing stage. Due to the expected large number of newly detected events, validation of all events predicted as glacial earthquakes would not be performed manually. We therefore add a step based on the scores obtained with the Random Forest algorithm, which reduces the number of events to be examined. The algorithm assigns a probability score to each seismic signal, and from this we have devised workflows to assign a class prediction and probability score to each event, composed of several seismic signals. These different workflows are presented in Fig. 5, with some example situations to illustrate their behaviour.

Workflow 1 (WF1) is based on the predictions of seismic signals that are generated by the same event. In this case, the final event prediction is the most represented prediction, regardless of the prob-

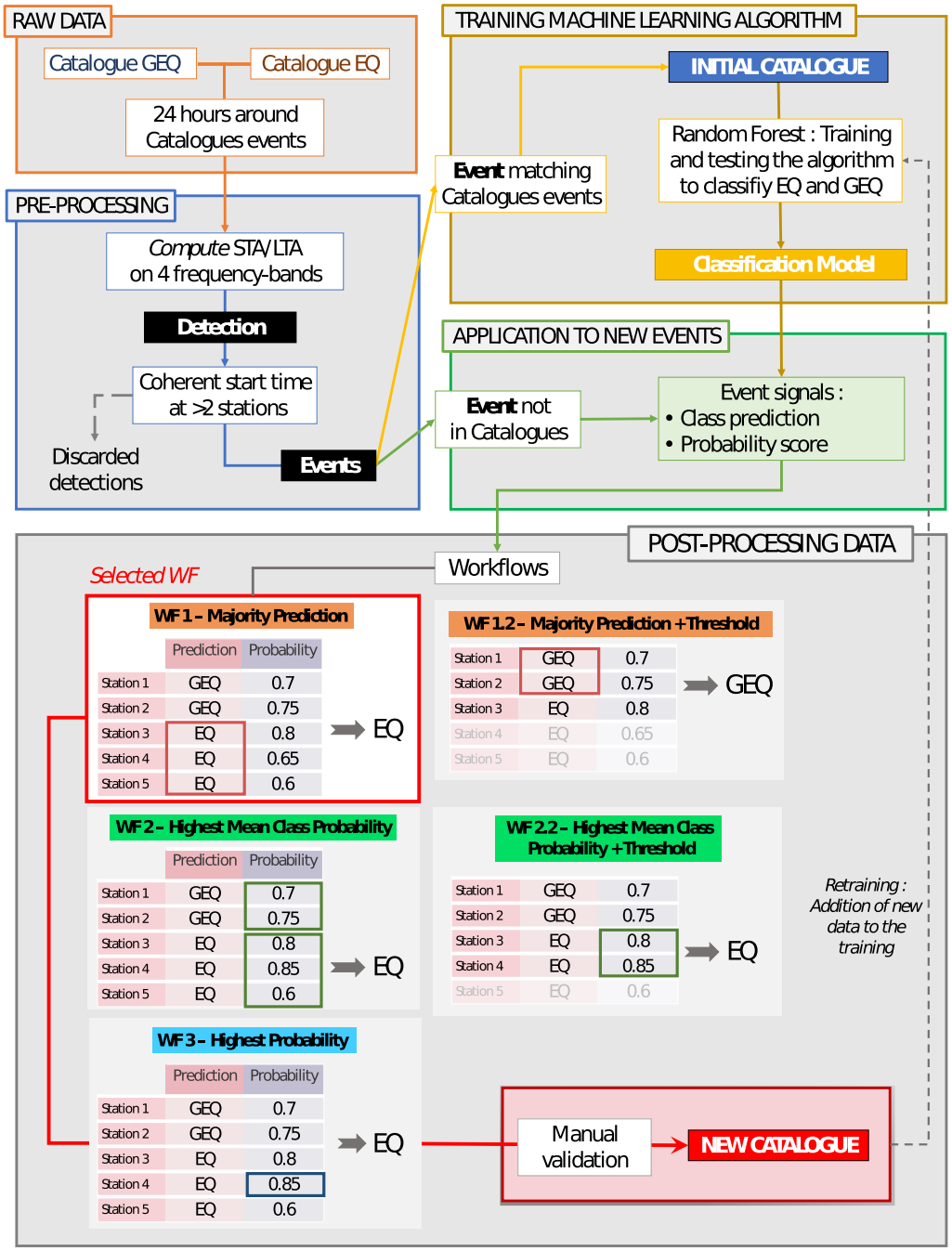

Figure 5. Global workflow. As input to this global workflow, we take existing catalogues of events (Raw data section). After the pre-processing step, which detects and extracts the signals of these events from the continuous data, the machine learning algorithm is trained and applied to signals that are not identified in pre-existing catalogues. In order to assign each signal a class and a probability score, signals belonging to the same unknown events are processed using a workflow (WF 1, 1.2, 2, 2.2 or 3). A manual validation is performed on this considerably reduced and refined selection of events.

ability scores (Fig. 5). The final probability score of the event is the average of the probability scores of the seismic signals labelled with the majority of the predicted class. Workflow 1.2 (WF1.2) is a variant of W1: it works in the same way, but the seismic signals taken into account at the start depend on their probability score, which must be above a threshold. In other words, a seismic signal whose probability score is below a defined threshold will be excluded from the calculation of the final probability of the event under consideration to belong to a given class. In Fig. 5, the threshold is strictly less than 0.7. For workflow 2 (WF2), the event probability score is the highest average of the probability scores within each class. Workflow 2.2 (WF2.2) works in the same way, with the threshold on signal probability scores. With workflow 3 (WF3), the event probability score is the highest probability score and the event prediction is the corresponding class. In summary, we obtain a classifier capable of discriminating between glacial earthquake and earthquake signals by following the protocol presented in Section 2.2.3 as well as the steps described in Fig. 5.

3 RESULTS

3.1 Random Forest training and testing results

The events included in the initial catalogue are used to train and test a Random Forest model. We evaluated the robustness of the approach by progressively increasing the number of events in the training set. Thus, 5, 10, 25 and 50 per cent of the seismic signals belonging to different events recorded in the initial catalogues are randomly and independently selected for each test. Each model obtained with a selected number of seismic signals on the training set is then tested on the remaining seismic signals. This procedure is repeated 10 times to gain a reliable overview of model behaviour as a function of the number of seismic signals used for training. Fig. 6 shows the percentages of well-predicted seismic signals as a function of the actual seismic signal label, averaged over 10 iterations of training and testing the model.

The good identification rate of glacial earthquake and earthquake seismic signals is relatively stable with variation in the number of seismic signals used for training. The percentage of well-predicted glacial earthquake signals varies only from 90.01 to 93.05 per cent when moving from 5 to 50 per cent of the catalogue events used in the training set. The percentage of well-predicted earthquake signals is similarly stable. The model is therefore able to correctly label around 90 per cent of seismic signals. All signals are then used for training, that is 100 per cent of the signals from both catalogues, to build the Random Forest model, applied to the 344 931 unlabelled seismic signals obtained with the automatic detection algorithm.

Model performance can be assessed using the accuracy score, which represents well-labelled prediction relative to total prediction. The overall accuracy of the model trained with 5 per cent of the catalogue is 88 per cent, rising to 91 per cent when 50 per cent of the events in the initial catalogue are used in the training set. The comparison with other studies using the Random Forest classifier is difficult as the number of events and of classes varies. For example, the correct identification rate reaches 99 per cent for landslide identification in Hibert *et al.* (2019) with much less events, the overall accuracy reaches 95.3 per cent in Malfante (2018) for volcano seismic signal classification and Maggi *et al.* (2017) have an accuracy score of 96 per cent for rockfall classification in a study with eight classes.

3.2 Results on 844 selected days

The STA/LTA-based algorithm, deployed on the 844 d of continuous data where an event occurs in the two initial catalogues (Section 2.1.3), yields 345 931 seismic signals that were detected on at least one of the five frequency bands used to calculate the STA/LTA detector (Section 2.2.1). As described in Section 2.2.2, the signals are combined into events: we obtain 60 933 events from the 345 931 signals. We then apply the Random Forest model trained with all the data available in the initial catalogues to the unlabelled seismic signals, predicting a class for each signal detected. Of these 60 933 events, around 40 335 events are classified as glacial earthquakes (65 per cent) using Workflow 1. Given the large number of events to be examined, we decided to use a threshold on the probability of scoring an event before adding it to the new catalogue.

In Fig. 7, we compare the number of events in the initial catalogues (Columbia 2007) that are well labelled by the model and the number of new events predicted as glacial earthquakes for the five workflows. These numbers are given as a function of the value of the event probability score, which ranges from 0.60 to 0.95, below which we discard the event. By increasing the probability score threshold, events with a low probability score are discarded from consideration. At a threshold of 0.6, all workflows tend to assign the correct label to known events (different coloured squares at 0.6). The higher the threshold, the fewer events are retained. For example, with workflow 2 (dark green), at 0.95, only 170 events from the initial catalogue of 444 glacial earthquakes (Columbia 2007) are retained in the final selection for manual validation. Workflow 1 minimizes the number of events predicted as glacial earthquakes (right axis) while maximizing the number of known well-predicted events (left axis). By choosing a threshold of 0.8, only 5460 events are classified as glacial earthquakes instead of 40 000. 90 per cent of the events in the initial catalogue are also well identified. Furthermore, by only considering events with a probability score above 0.8, there is a greater probability of excluding noisy events that have been classified as glacial earthquakes, as we assume they have a lower probability score than glacial earthquakes. Workflow 1 combined with the threshold of 0.8 seems to be the best balance between the number of events to be examined, the hypothetical false positive rate and the number of well-identified glacial earthquakes in the initial catalogue (Columbia 2007). With this choice of parameters, we have 5460 events to check manually in order to validate the predicted label.

During the manual validation phase, we decided to be conservative and validate as glacial earthquakes only those events of which we were convinced of, by comparing them with proven glacial earthquakes (see Fig. 3). We note that some events are only recorded by stations in Iceland and Canada, but are not always detected by stations in Greenland. This may be a bias in the algorithm which forms the events. These events were labelled as glacial earthquakes by the model and obtained a high event probability score with the chosen workflow. We discard them during manual validation. We sometimes observe the classification of teleseismic events as glacial earthquakes due to their similar low-frequency content (an example is presented in Fig. 9).

Of the 5460 events classified as glacial earthquakes, 1633 (28 per cent) are validated as new glacial earthquakes after manual verification; 758 are reclassified as earthquakes and the remaining events are discarded because we cannot be sure of the source of these signals, as these events only often present glacial earthquake characteristics on one station, or none at all, but are still not earthquake signals. Over a period of 844 non-consecutive days, we therefore found 1633 new glacial earthquakes, that is 3.6 times more events than in the initial catalogue (Columbia 2007). These new events constitute the new catalogue of glacial earthquakes.

3.3 Preliminary analysis of the new catalogue

The catalogue presented here does not cover a continuous period of data, and does not allow us to study the evolution of the number of events as a function of time, which is the ultimate aim of deploying the algorithm over the entire period from 1993 to the present day. However, the catalogue obtained for the 844 selected days allows us to make some initial observations.

Fig. 8(a) shows the number of events per year of the original glacial earthquake catalogue (blue) and the new catalogue (red), as well as the evolution of the number of stations per year (black triangles). Events are represented year by year over the period 1993 to 2013. Changes in the number of stations seem to have a limited impact after the years 1999–2000: between 2008 and 2011, there is a gap in the new catalogue and no sudden increase in the

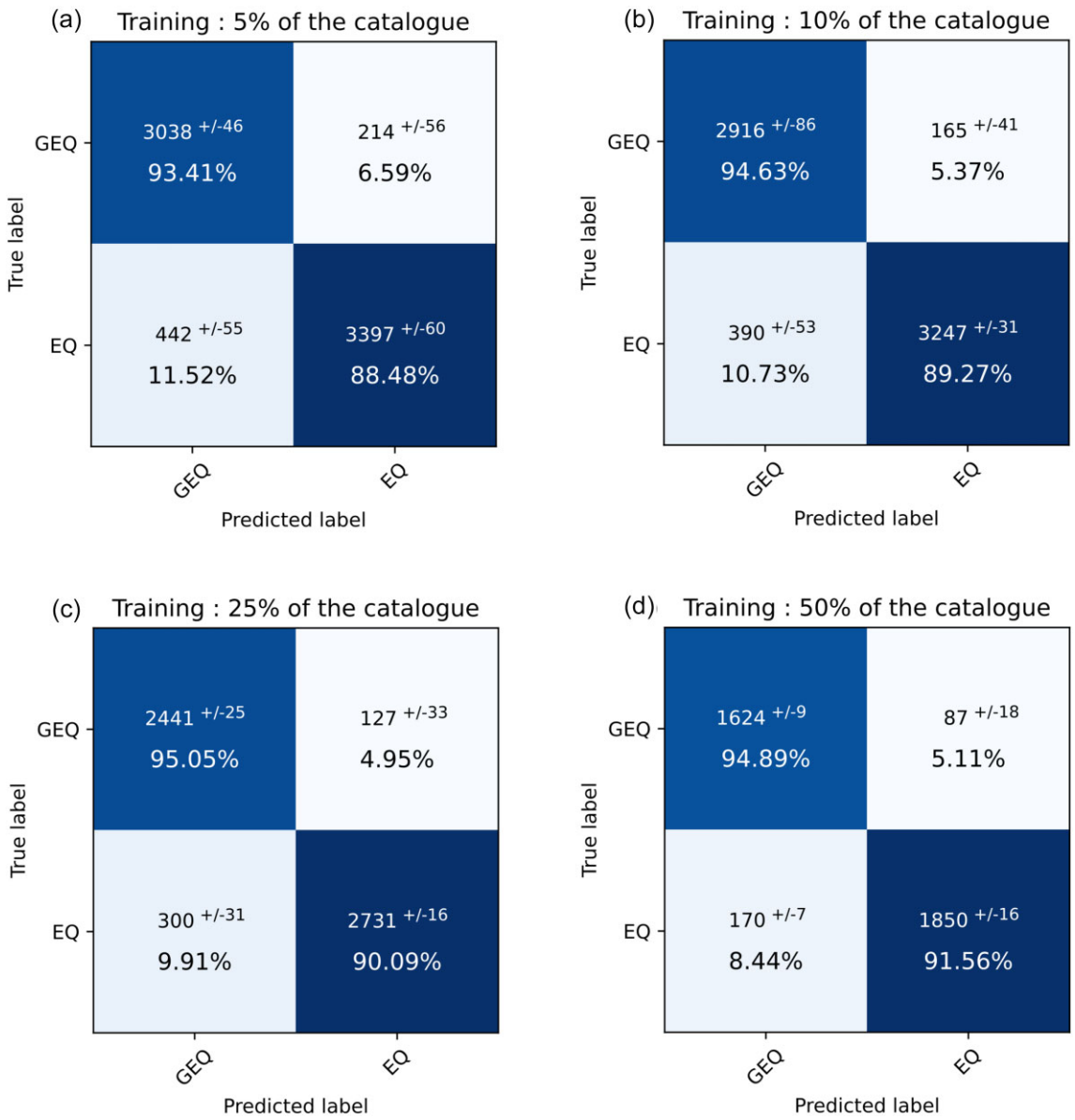

Figure 6. Four confusion matrices with various number of training signals at training phase (from 5 to 50 per cent). Percentages are the means of the accuracy within the class, obtained on 10 tries.

initial catalogue, despite a network of 20 stations. Before 2000, very few events were found in both catalogues. Fig. 8(b) plots the number of events per month to see how they correlate with the seasons. The yellow vertical bands represent summer periods, from June to the end of September. A periodicity in the occurrence of events can be observed, most pronounced for new catalogue events in red, but also visible for initial catalogue events in blue during this period. Some summers (2003, 2004, 2011, 2012 and 2013) also show a correlation in the number of events in the two catalogues.

Fig. 8(c) represents the number of events over the 844 d studied. A day on which a glacial earthquake from the original catalogue occurred is represented by a blue line, as shown in Figs 8(e) and (g). Days with red and blue lines represent days when an event from the initial glacial earthquake catalogue occurred and at least one new glacial earthquake was identified. Days with only a red line correspond to days when an earthquake event from the initial earthquake catalogue (Columbia 2007) occurred and a new glacial earthquake was identified. We observe three gaps in the number of new events, in 2009, 2010 and 2011, so far without explanation. Future analysis of the continuous data will enable us to better analyse whether or not there are any gaps, and what causes them.

In 2004, we see in Fig. 8(d), which is a zoom on the selected year, up to 50 new events identified during the summer on a day when an event from the initial catalogue occurred. Fig. 8(e) shows only events from the initial catalogue for that year. In 2012 (Fig. 8f), we observe between 1 and 19 new events recorded on days when an event from the initial catalogue of glacial earthquakes occurred. On a day when no glacial earthquake event occurred, 20 new events were classified as glacial earthquakes (red bar without blue at top). The same year, in February 2012, 3 events were identified in the initial catalogue and five new events were identified on the same day, underlining the trend in the initial catalogue.

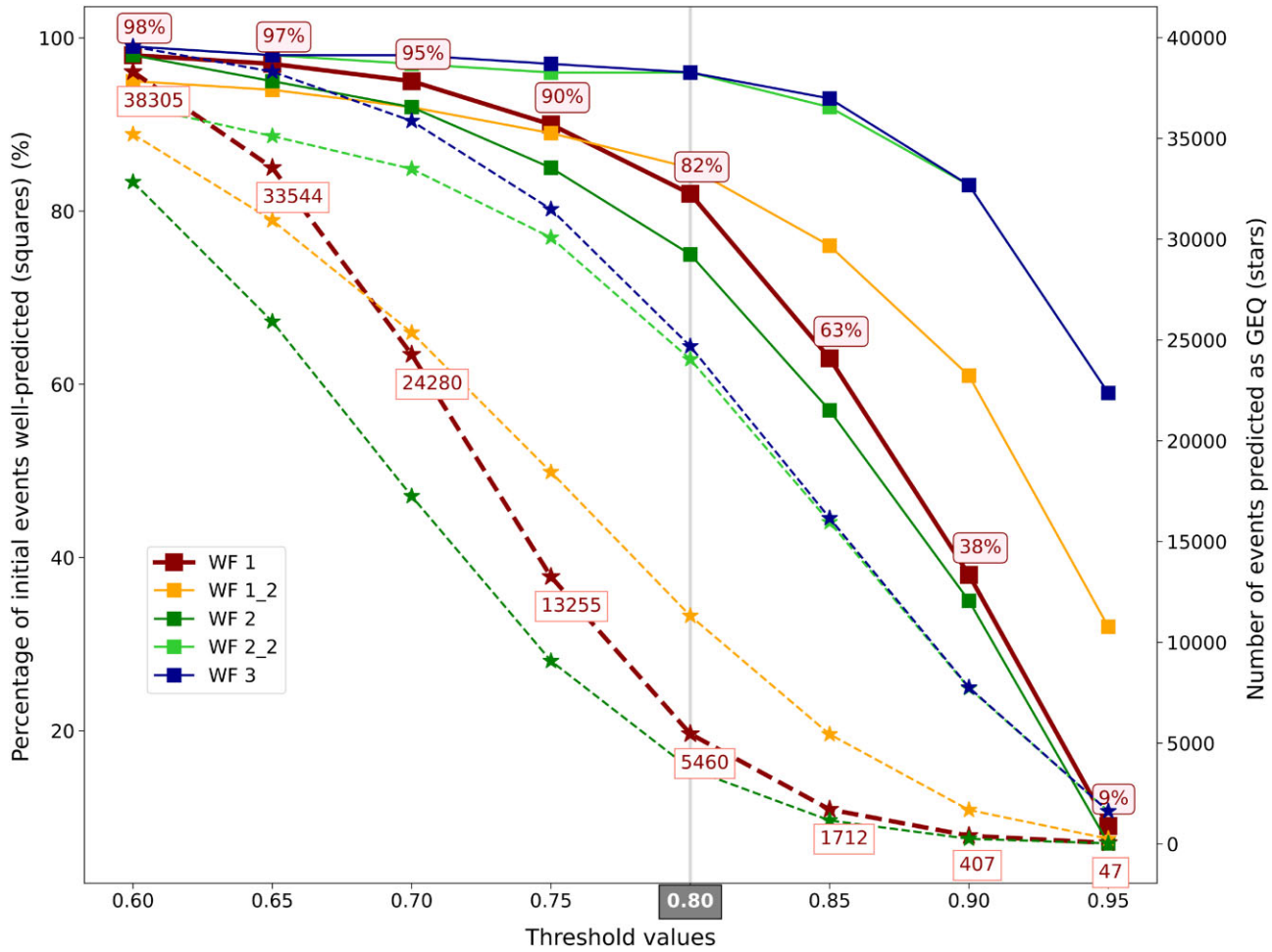

Figure 7. The performance of five workflows is compared on the basis of the number of events in the initial catalogues (solid lines with squares) and the overall number of events predicted as glacial earthquakes (dashed lines with stars), as a function of the threshold value. The chosen configuration is Workflow 1 (red) with a threshold of 0.8 (grey vertical line).

3.4 Retraining with new events

3.4.1 Reclassified signals as earthquakes

At the stage of manual event validation, we note a misclassification of some teleseismic events as glacial earthquakes. These teleseismic events are not regional events like those chosen to train the random forest model, but teleseismic events occurring all around the globe. Their seismic signature therefore differs from that of regional earthquakes, depending on the station recording the signal (Fig. 3). In Fig. 9, signal energy increases with time towards higher frequencies, corresponding to surface waves. The closer the tectonic activity, the stronger the increase. *P* and *S* waves are not easy to identify, as they are often attenuated, and depend on the magnitude and location of the event. The teleseismic events were identified one by one by using the time of signal recording at each station that had detected it. The event shown in Fig. 9, mislabeled as a glacial earthquake, is a teleseismic event of magnitude 5.2 that occurred in Serbia at 7:40 a.m. on 1 July 1999. Teleseismic events have been added to the learning set as earthquakes. As our aim is to extract glacial earthquake signals from other signals (earthquakes, noise, anthropological activities, etc.) and not to label new signals, we have excluded the creation of a third class with teleseismic signals.

3.4.2 Comparison of model trained with catalogue events and with the addition of reclassified events

The reclassified events are added to the random forest training set. We compare the performance of the model using only data from the initial catalogue, that is 444 glacial earthquakes and 400 earthquakes (Columbia 2007; USGS 2021), and the model using both initial catalogues plus reclassified events, i.e. new glacial earthquakes and teleseismic and other tectonic events in the glacial earthquakes and earthquakes classes respectively. In Fig. 10, we compare the detection approach (Fig. 10a) and the event approach (Fig. 10b) of the two models (orange and red) by examining the model accuracy score obtained with a training set varying from 5 to 100 per cent of the data set. The accuracy score is the number of correctly identified events divided by the number of predicted events in that class. We also use it to measure the false alarm rate in a class. The median score is the median of all event probability scores, whatever the final prediction. It can be an indicator of the model's level of certainty.

In Fig. 10(a), the accuracy score of the model trained with initial catalogues (Columbia 2007; USGS 2021; orange line) is better than the accuracy score of the model with the addition of reclassified events (red line), whatever the percentage of events on the training set. With the event-based approach (Fig. 10b) at high percentages of the trained number, both models show the same accuracy

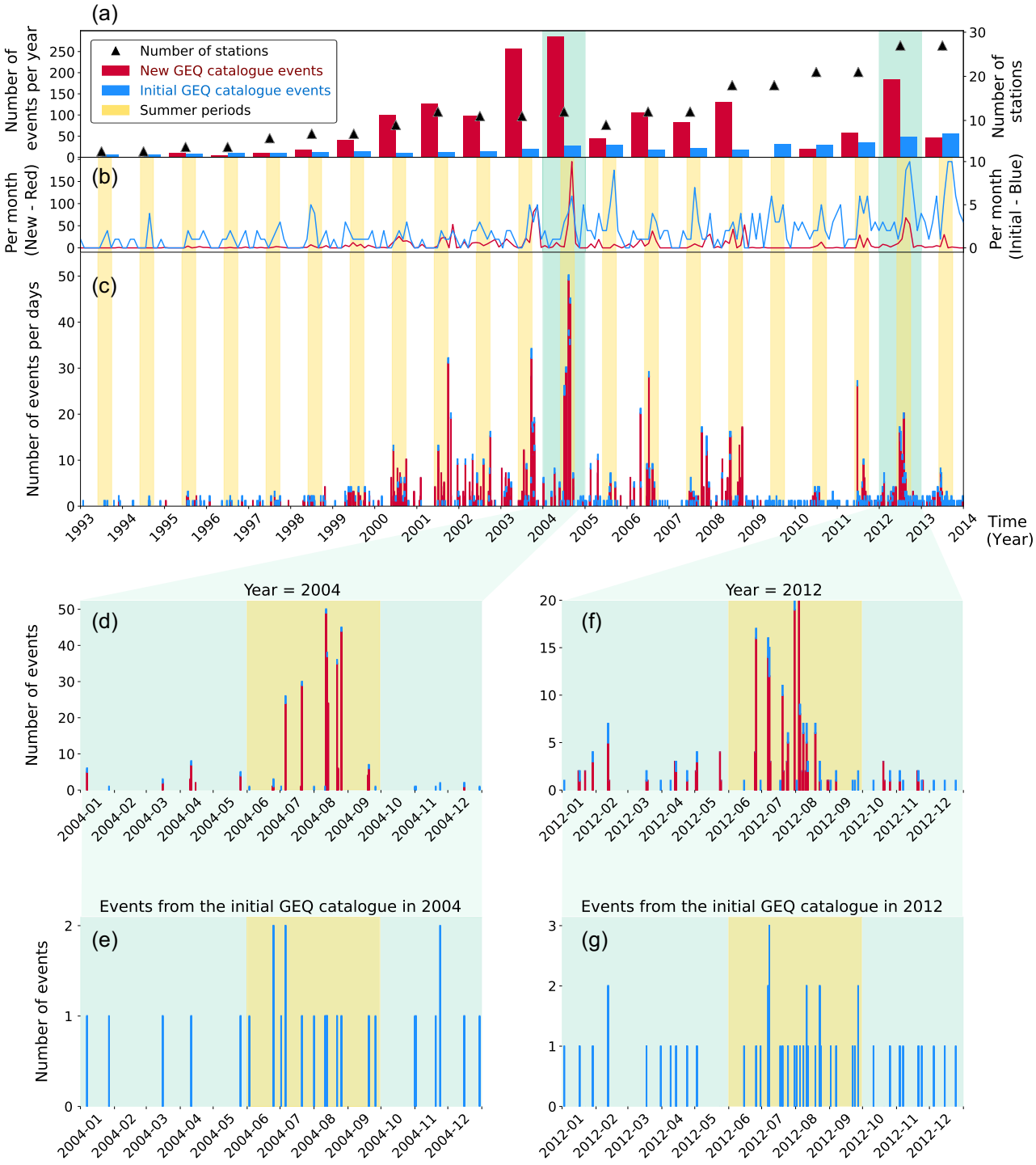

Figure 8. Distribution of occurrence of glacial earthquakes in the original glacial earthquake catalogue (blue) and glacial earthquakes in the new catalogue (red) by year (a), month (b) and day (c). Distribution of events in 2004 (d and e) and 2012 (f and g). Evolution of the number of stations (a). Yellow area indicate summer periods.

score (>80 per cent). We note that accuracy scores are improved by 10 per cent when working with the event-driven approach, and that scores are relatively stable as a function of the number of trained events. Median probability scores are lower in both approaches, with a difference of 0.2 in the detection approach and 0.1 in the event approach. Accuracy and median scores were expected to be lower with the model featuring reclassified events, as new events introduce difficulties during the training phase: reclassified events are often noisier and have less energy in certain frequency bands. Adding reclassified events to the training data does not improve the performance of the machine learning algorithm, but one of its advantages is to add diversity to the signals used to train the model, making it capable of identifying events that are more dissimilar to those in the initial catalogue.

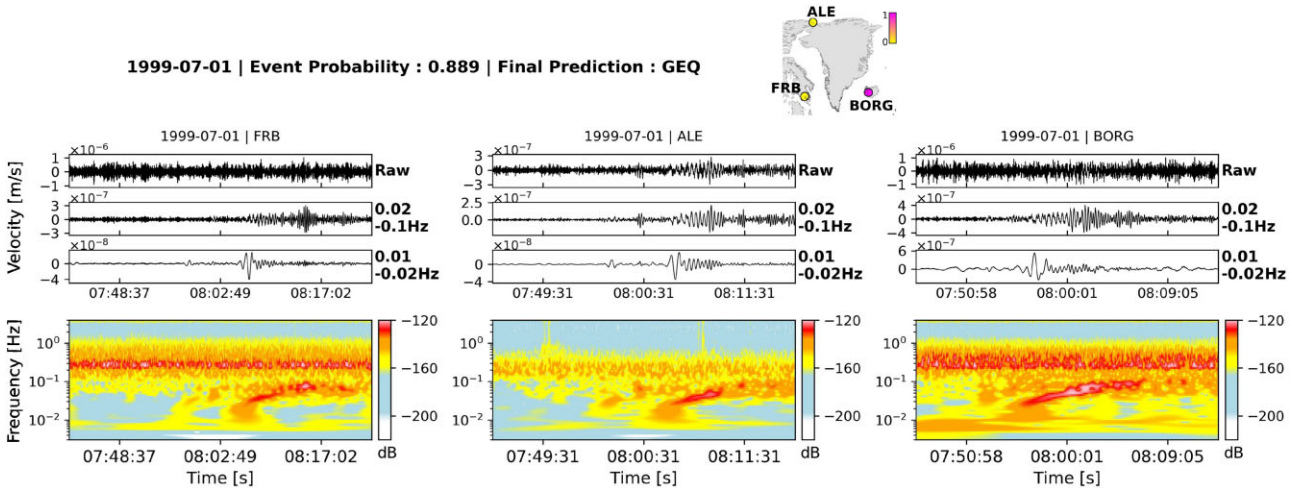

Figure 9. A teleseismic event occurred in Serbia at 7:40 a.m. on 1 July 1999, with a magnitude of 5.2 recorded by stations FRB, ALE and BORG.

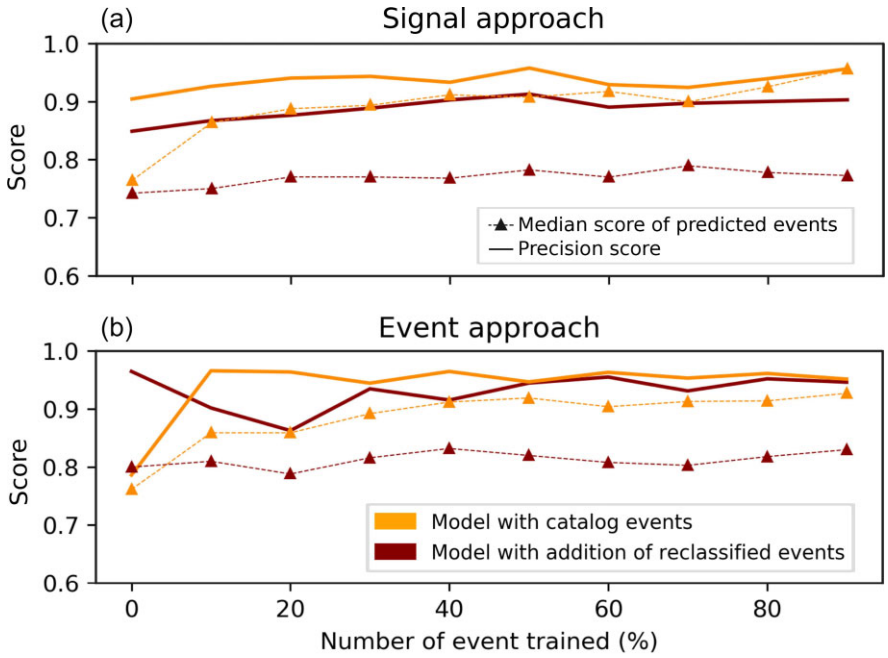

Figure 10. Sensitivity to the number of events used to train the Random Forest model with initial catalogue events (Columbia 2007; USGS 2021) and the Random Forest model with both initial and reclassified events (reclassified glacial earthquakes and reclassified earthquakes), with the single signal approach (a) and the event approach (b). The accuracy score of each model is represented by a single line. The median probability score of all events, assigned after the W1 workflow, is represented by a dotted line and triangles.

4 DISCUSSION

The importance of the features used in the Random Forest algorithm is examined here, as are the benefits of the addition the 39 new features (see Section 2.2.3). These features allow us to understand why the model works well and may give us insights into the source of glacial earthquakes. All features are detailed in Table C1 in Appendix C. The weight in the identification done by the Random Forest algorithm is computed based on the decision trees which yields importance scores. The relevance of a feature's choice is assessed using the importance score: a high importance score means that the specific feature has a greater effect on the classification model. In Fig. 11, the 97 features are represented according to their importance score. Scores are normalized so that the sum of scores is equal to 1. The 10 most important features are shown in red. Importance scores decrease rapidly: the best feature (feature 64—kurtosis of the signal in the 0.05–0.1 Hz frequency band) has a score of 0.112 and the tenth feature (feature 95—energy ratio 10–20 Hz/0.05–0.1 Hz) has a score of 0.029. The number of peaks in the discrete Fourier transform of the average signal (feature number 47) is the least discriminating feature, with a score of 2.16×10^{-6}. The 39 new features, framed in light brown in Fig. 11, which are mainly based on kurtosis, ratio and energy differences in specified frequency bands, significantly influenced the accuracy of our model: 7 of the 10 highest-scoring features come from these 39 features, including the most discriminating (64). The 3 features that are among the top 10 but not among the 39 new features are signal flattening (feature 5), the skewness of the envelope (feature 8) and kurtosis in the 0.1–1 Hz frequency band (feature 18).

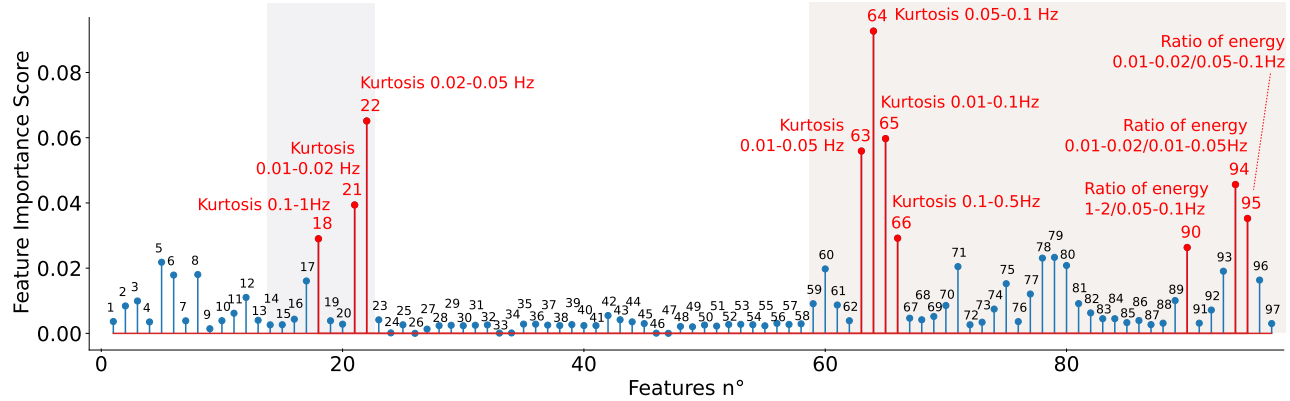

Figure 11. Importance scores for the 97 features used to describe waveforms. In red, the 10 best features. The light brown square frames the 39 features added for this study. The grey square frames features from the original study that have been modified, in particular by changing the frequency bands to suit the sampling rate of the stations used.

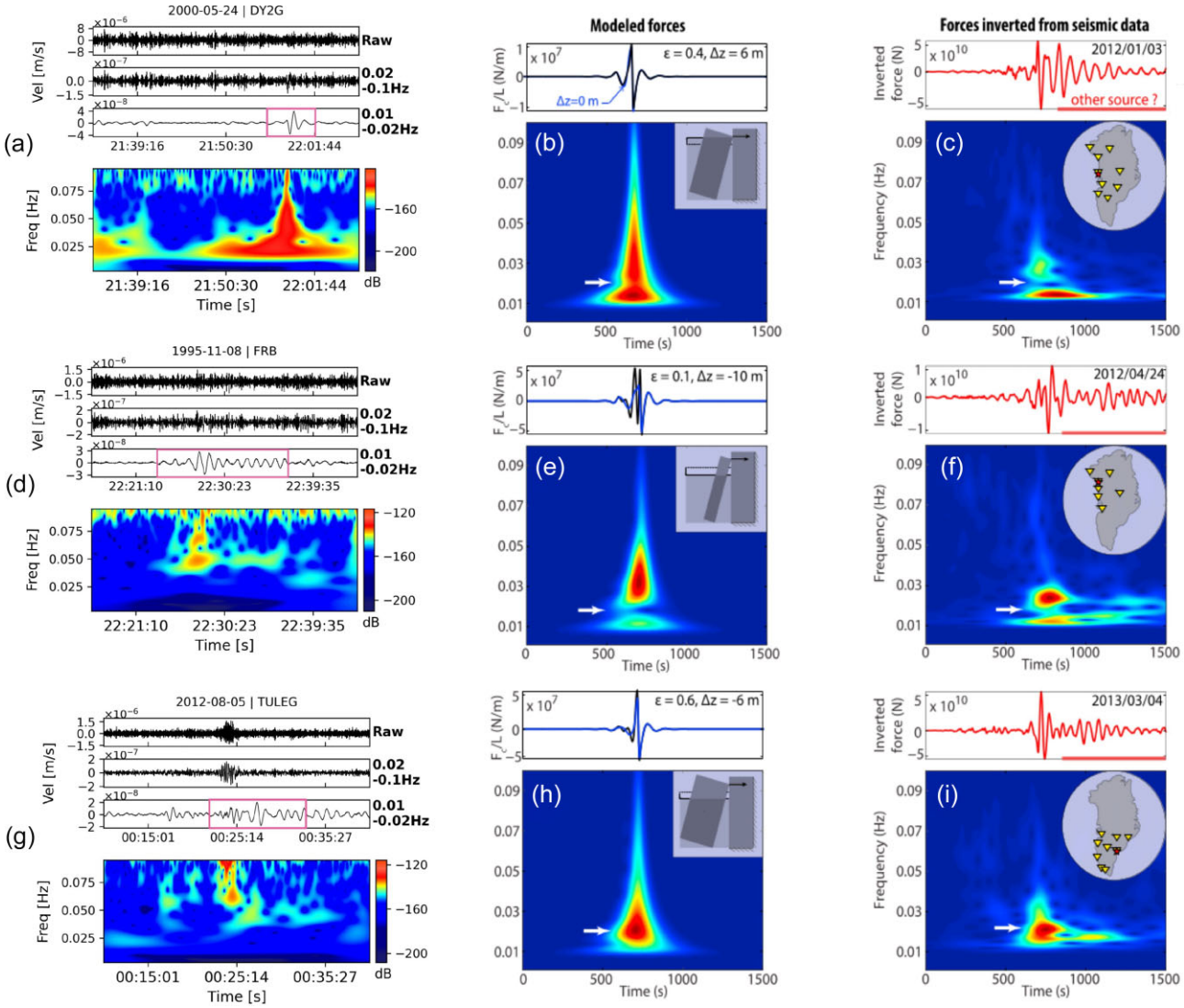

Figure 12. Comparison of 3 new glacial earthquakes (a., d. and g.) with three synthetic forces (b, e and h) and three inverted forces from the initial catalogue (c, f and i) adapted from Sergeant *et al.* (2018). For the new events, the signal is decomposed into different frequency bands and represented with a linear spectrogram, with the same frequency limits as the spectrogram limits of the synthetic signals.

To confirm the final labelling of the newly identified glacial earthquakes, we visually compare the seismic signals of the new events with synthetic signals obtained by Sergeant *et al.* (2018). This approach also enables us to better understand the role of the best features identified by our algorithm. In Fig. 12, the spectrograms of the seismic signals (Figs 12a, d and g), not deconvoluted from Green's function, show the same patterns as the spectrogram of forces simulated by a mechanical model describing the calving of an iceberg against the tip of a glacier (Figs 12b, e and h). Different iceberg configurations (different ϵ aspect ratios and different buoyancy conditions related to the initial iceberg height represented by Δz), illustrated in Figs 12(b), (e) and (h), lead to different synthetic signals, as explained in Sergeant *et al.* (2018). In Figs 12(c), (f), (i), the inverted forces are waveform inversions of the original catalogue recorded by the GLISN network. The new seismic signals illustrated share some features with the forces, in each configuration. The higher frequency range of the new events suggests to us that the events are smaller than the catalogue events used to invert the forces. The signals are very similar, particularly at low frequencies (35–100 s - 0.02–0.01 Hz). We observe the three best features : kurtosis in 0.05–0.1 Hz, 0.01–0.1 Hz and 0.01–0.05 Hz. They do indeed appear to be relevant features for describing synthetic signals. The attributes we have chosen therefore carry information about the source and are less affected by propagation. This sensitivity to source parameters and less to propagation effects is probably what makes this approach so successful.

5 CONCLUSION

In the context of ongoing global warming and its major impact on polar regions, we highlight the importance of creating comprehensive catalogues of glacial earthquakes to refine the understanding and quantification of ice mass loss in the Greenland ice sheet and its link to climate change. In this study, we have developed a semi-automatic processing chain capable of detecting and identifying new seismic signals of glacial earthquakes from continuous data, adaptable to longer time periods over which seismic data are available, that is up to 30 yr.

We first design a detection algorithm based on several standard STA/LTAs to extract glacial earthquakes and seismic signal earthquakes from continuous data. Deployed over 844 d from the catalogue of 444 glacial earthquakes (Columbia 2007) and the catalogue of 400 earthquakes (USGS 2021), we extract 344 931 seismic signals of potential cryoseismic events. This huge number of signals calls for machine learning methods to obtain a pre-selection of new glacial earthquake events. Consequently, with a Random Forest algorithm and a workflow designed to reduce the number of misidentified events, 1633 events were identified as new glacial earthquakes in a period when only 444 glacial earthquakes had previously been recorded. The addition of 39 new features considerably improved the efficiency of the random forest model, since 7 of the 10 best features belonged to this group. The similarities between the seismic signals from the new events and the synthetic signals obtained in previous studies confirm the discovery of new glacial earthquakes.

Last but not least, the creation of a widely extended catalogue of glacial earthquakes will contribute to a better study of this phenomenon. The machine-learning model can extract information on the source property from the characteristics of glacial earthquakes. In addition, the model can be improved by adding new identified events to the training set and by restricting temporary networks from Iceland. These improvements could reduce the hypothetical false-positive rate while enabling the identification of a greater number of glacial earthquakes. This workflow from continuous data to the identification of seismic signals as glacial earthquakes can be deployed over other time periods to enrich the catalogue. Deployment over 30 yr of continuous data, from 1993 to 2023, will be the subject of future work.

ACKNOWLEDGMENTS

This work was supported by the Doctoral School 560 STEP'UP (Sciences de la Terre et de l'Environnement et Physique de l'Univers de Paris).

We thank the operators of the GLISN networks for data collection and the IRIS Data Management System for providing easy access to the data. We thank all the people who worked to create the initial catalogue of glacial earthquakes (Tsai & Ekström 2007; Veitch & Nettles 2012; Olsen & Nettles 2017) referred as Columbia (2007) in this paper. The catalogue of earthquakes have been selected from the USGS database (https://www.usgs.gov/) and is referred as USGS (2021) in this article.

All authors shared ideas, contributed to the interpretation of the results and to the writing of the manuscript. The codes developed for this study can be accessed by contacting authors.

DATA AVAILABILITY

All seismic data were downloaded through the EarthScope Consortium, supported by the National Science Foundation under Cooperative Agreements EAR-1261681 and are openly available at: http://ds.iris.edu/SeismiQuery/.

REFERENCES

Amundson, J.M., Truffer, M., Lüthi, M.P., Fahnestock, M., West, M. & Motyka, R.J., 2008. Glacier, fjord, and seismic response to recent large calving events, Jakobshavn Isbræ, Greenland, *Geophys. Res. Lett.,* **35**(22), doi:10.1029/2008GL035281.

Amundson, J.M., Fahnestock, M., Truffer, M., Brown, J., Lüthi, M.P. & Motyka, R.J., 2010. Ice mélange dynamics and implications for terminus stability, Jakobshavn Isbræ, Greenland, *J. geophys. Res.,* **115**(F1), doi:10.1029/2009JF001405.

Aster, R.C. & Winberry, J.P., 2017. Glacial seismology, *Rep. Prog. Phys.,* **80**(12), 126801 , doi:10.1088/1361-6633/aa8473.

Bonnet, P. *et al.*, 2020. Modelling capsizing icebergs in the open ocean, *J. geophys. Int.,* **223**(2), 1265–1287.

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

Burton, J.C. *et al.*, 2012. Laboratory investigations of iceberg capsize dynamics, energy dissipation and tsunamigenesis, *J. geophys. Res.,* **117**(F1) , doi:10.1029/2011JF002055.

Chmiel, M., Walter, F., Wenner, M., Zhang, Z., McArdell, B. & Hibert, C., 2021. Machine learning improves debris flow warning, *Geophys. Res. Lett.,* **48** , doi:10.1029/2020GL090874.

Dong, L., Li, X. & Xie, G., 2014. Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive bayes classification, *Abstr. Appl. Anal.,* **2014,** e459137 , doi:10.1155/2014/459137.

Ekström, G., Nettles, M. & Abers, G.A., 2003. Glacial earthquakes, *Science,* **302**(5645), 622–624.

Ekstrom, G., 2006. Global detection and location of seismic sources by using surface waves, *Bull. seism. Soc. Am.,* **96**(4A), 1201–1212.

Falcin, A. *et al.*, 2020. A machine-learning approach for automatic classification of volcanic seismicity at La Soufrière Volcano, Guadeloupe,

J. Volc. Geotherm. Res., **411,** 107151 , doi:10.1016/j.jvolgeores.2020.107151.

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

Hibert, C., Michéa, D., Provost, F., Malet, J.-P. & Geertsema, M., 2019. Exploration of continuous seismic recordings with a machine learning approach to document 20 years of landslide activity in Alaska, *Geophys. J. Int.,* **219** , doi:10.1093/gji/ggz354.

Kumar, P., Kind, R., Priestley, K. & Dahl-Jensen, T., 2007. Crustal structure of Iceland and Greenland from receiver function studies, *J. geophys. Res.,* **112** , doi:10.1029/2005JB003991.

Lin, G.-W., Hung, C., Chang Chien, Y.-F., Chu, C.-R., Liu, C.-H., Chang, C.-H. & Chen, H., 2020. Towards automatic landslide-quake identification using a random forest classifier, *Appl. Sci.,* **10**(11), doi:10.3390/app10113670.

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., 2018. Automatic classification of natural signals for environmental monitoring, *These de doctorat,* Université Grenoble Alpes (ComUE).

Nettles, M. & Ekström, G., 2010. Glacial earthquakes in Greenland and Antarctica, *Ann. Rev. Earth planet. Sci.,* **38,** 467–491.

Nettles, M. *et al.* , 2008. Glacier acceleration, glacial earthquakes, and ice loss at Helheim Glacier, Greenland, in *Proceedings of the AGU Fall Meeting,* Abstract, 15–19 December 2008, San Francisco, CA, USA.

Olsen, K.G. & Nettles, M., 2017. Patterns in glacial-earthquake activity around Greenland, 2011–13, *J. Glaciol.,* **63**(242), 1077–1089.

Olsen, K.G. & Nettles, M., 2019. Constraints on terminus dynamics at Greenland glaciers from small glacial earthquakes, *J. geophys. Res.,* **124**(7), 1899–1918.

Podolskiy, E. & Walter, F., 2016. Cryoseismology, *Rev. Geophys.,* **54**(4), 708–758.

Provost, F., Hibert, C. & Malet, J.-P., 2017. Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier, *Geophys. Res. Lett.,* **44**(1), 113–120.

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

Sergeant, A., Yastrebov, V.A., Mangeney, A., Castelnau, O., Montagner, J.-P. & Stutzmann, E., 2018. Numerical modeling of iceberg capsizing responsible for glacial earthquakes, *J. geophys. Res.,* **123**(11), 3013–3033.

Sergeant, A. *et al.*, 2019. Monitoring Greenland ice sheet buoyancy-driven calving discharge using glacial earthquakes, *Ann. Glaciol.,* **60**(79), 75–95.

Tsai, V.C & Ekström, G., 2007. Analysis of glacial earthquakes, *J. geophys. Res.,* **112**(F3), doi:10.1029/2006JF000596.

Veitch, S.A. & Nettles, M., 2012. Spatial and temporal variations in Greenland glacial-earthquake activity, 1993-2010: Greenland Glacial Earthquakes, 1993-2010, *J. geophys. Res.,* **117**(F4), doi:10.1029/2012JF002412.

APPENDIX A: INVENTORY OF NEW EVENTS

A1 Two new events occurring in 1998 and 1999

Before 2000, there were only 6 permanent seismic stations, including FRB and SFJ (Fig. 2). Station SFJ was replaced at the same location by SFJD in 2005, and is now part of the GLISN network. Two events occurring in 1998 (Fig. A1a) and 1999 (Fig. A1b) are shown. On the map, stations are coloured with a scale corresponding to the amplitude of the raw signal recorded: the pinker the colour, the greater the amplitude. Events from the initial catalogue, glacial earthquakes (Columbia 2007) or earthquakes, occurring on the days of the events shown are located by blue or brown stars, respectively. New events are framed by a dotted line, and sometimes several events are identified in the same record. The new event in Fig. A1(a) was detected on a day when a glacial earthquake was detected 5 hr earlier and located at Jakobshavn Isbrae, which is 240 km from the SFJ station. Two events seem to have occurred one after the other, with similar waveforms. The event in Fig. A1(b) was detected on a day when an earthquake occurred but no glacial earthquake from the initial catalogue of glacial earthquakes was detected. These two events can be compared with an event from the initial catalogue occurring on 2013-06-17 at Jakobshavn Isbrae (see Figs B3a and c in the appendices). The amplitude of the filtered signals in 0.01–0.02 Hz of these new events ($\approx 1 \times 10^{-8}$) is slightly lower than that of the catalogue events ($\approx 4 \times 10^{-8}$). We note that these two events were detected by the same two stations, SFJ and FRB, using our seismic signal algorithm.

A2 Example of new events occurring the same day

We then present a day on which a catalogue event occurred at 03:02 a.m. on the Jakobshavn Isbrae glacier (east coast of Greenland in Fig. 2 - purple stars), illustrated in Fig. A3. In Fig. A2(a), the event occurred 1 hr before the catalogue event, and the event in Fig. A2(b) occurred 4 hr after. The event represented in Fig. A2(b) has similar characteristics to those of the catalogue event (Fig. A3), particularly on the SFJD record. The amplitude of the filtered signals in 0.01–0.02 Hz is of the order of 2×10^{-8}, whereas the catalogue events at Jakobshavn Isbrae have amplitudes of between 1×10^{-8} and 4×10^{-8}. The new events appear to have an amplitude of the same order as the catalogue events. The waveforms of the initial catalogue signal and the waveforms observed on the same day, a few hours before and after, are sufficiently similar for us to identify these new signals as new glacial earthquakes. The signals, recorded at different stations, show similarities in the 0.02–0.1 and 0.01–0.02 Hz frequency bands. We also observe a sequence that seems to repeat itself [Fig. A2b (ILULI) and Fig. A3] with two consecutive signals (the second being of lesser amplitude).

A3 Example of two new events (2013)

Finally, we look at two new events that occurred in 2013 in Fig. A4. Only the two signals with the largest amplitude are shown. In Fig. A4(a), the signals recorded at ILULI and SFJD have amplitudes of the same order. The waveforms show the same pattern as signals recorded during an event a few days earlier (2013-06-17) at Jakobshavn Isbrae, illustrated in the appendices. The amplitudes of the signals filtered in 0.01–0.02 Hz are of the same order, which may reflect a new event of similar magnitude (4.9 for the event in the initial catalogue). The event shown in Fig. A4(a) appears to have occurred on the West Coast, due to the location of the station with the highest amplitude, while the event (b) appears to have occurred on the East Coast.

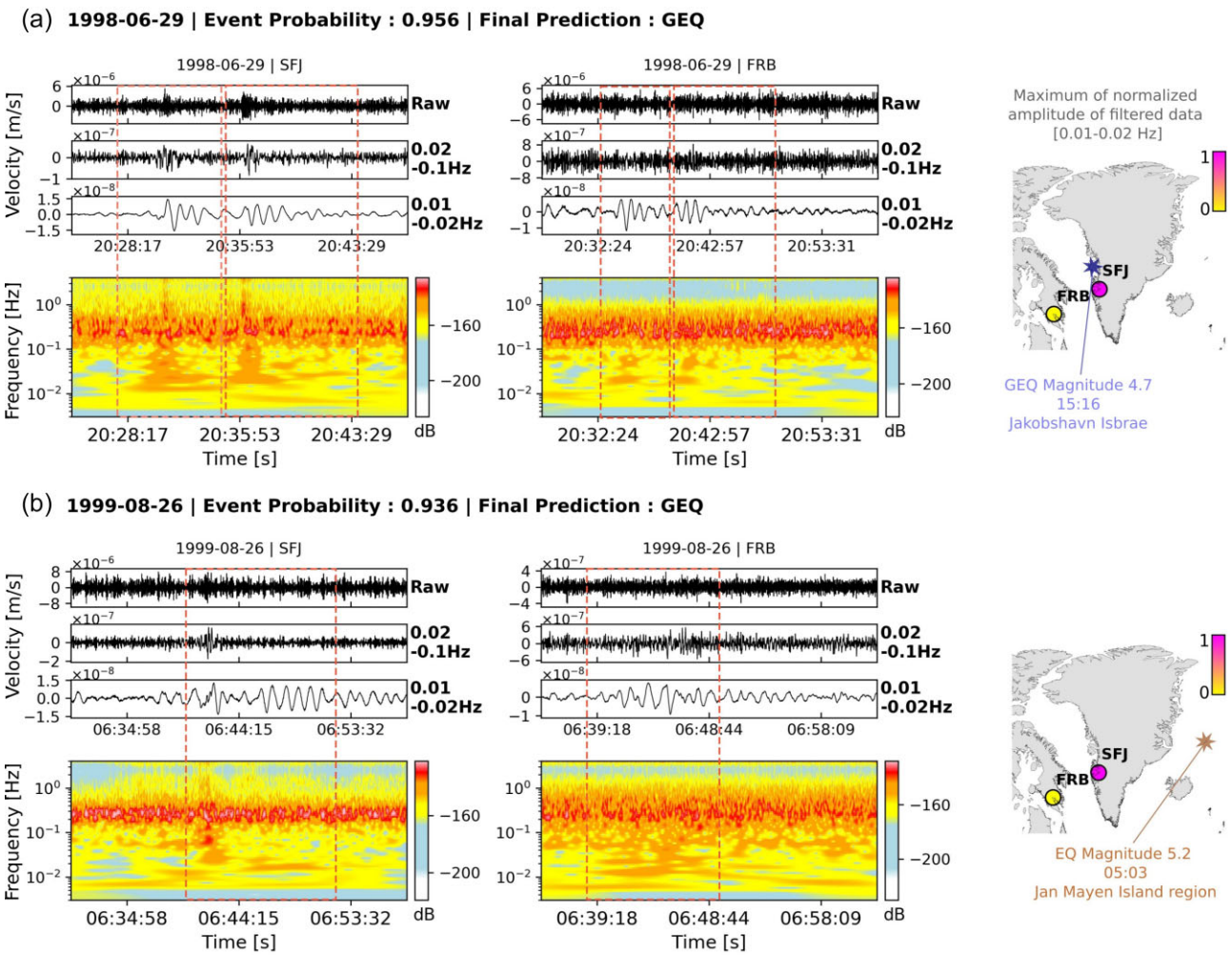

Figure A1. Two events occurred in 1998 (a) and 1999 (b), both recorded by stations SFJ and FRB. Each subfigure shows the raw signal, the filtered signal in two frequency bands (0.02–0.1 and 0.01–0.02 Hz) and a spectrogram with signal intensity in decibels. Stations are shown on a map of Greenland, coloured according to the normalized amplitude of the filtered signal in 0.01–0.02 Hz. Catalogue events of the day are represented by a blue or brown star for glacial earthquakes or earthquakes, respectively. The signal from new events is framed by two red dotted lines.

APPENDIX B: EVENTS FROM THE INITIAL CATALOGUE

We present some events from the initial catalogue of glacial earthquakes (Columbia 2007), grouped by glacier: we show examples from the three most iceberg-producing glaciers, Helheim Glacier (Fig. B1), Kangerlussuaq Glacier (Fig. B2) and Jakosbshavn Isbrae (Fig. B3). These events occurred at different times and were recorded by a varied number of stations. They were used for visual comparison to validate the new glacial earthquakes.

APPENDIX C: LIST OF FEATURES

Features used for the Random Forest model are detailed in the table given in Table C1.

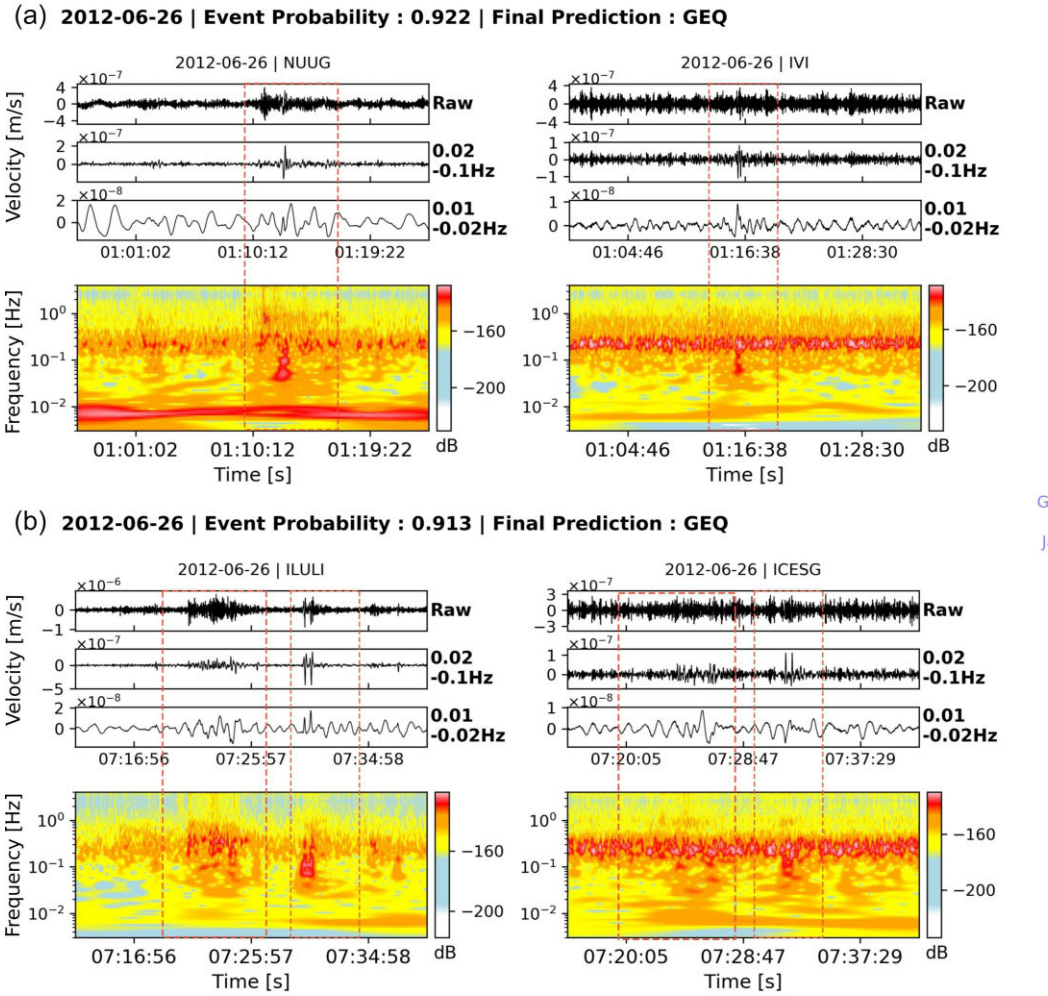

Figure A2. Two new events occurred on the same day as a catalogue event (see Fig. A3). The known event occurred at 03:02 a.m. on the Jakobshavn Isbrae glacier, located by a dark blue star. Event a. occurred 1 hr before and event b. 4 hr after the catalogue event. Each subfigure shows the raw signal, the filtered signal in two frequency bands (0.02–0.1 and 0.01–0.02 Hz) and a spectrogram with signal intensity in decibels. Stations are shown on a map of Greenland, coloured according to the amplitude of the filtered signal in 0.01–0.02 Hz. Catalogue events of the day are represented by a blue or brown star for glacial earthquakes or earthquakes, respectively. The signal from events is framed by red or orange dotted lines.

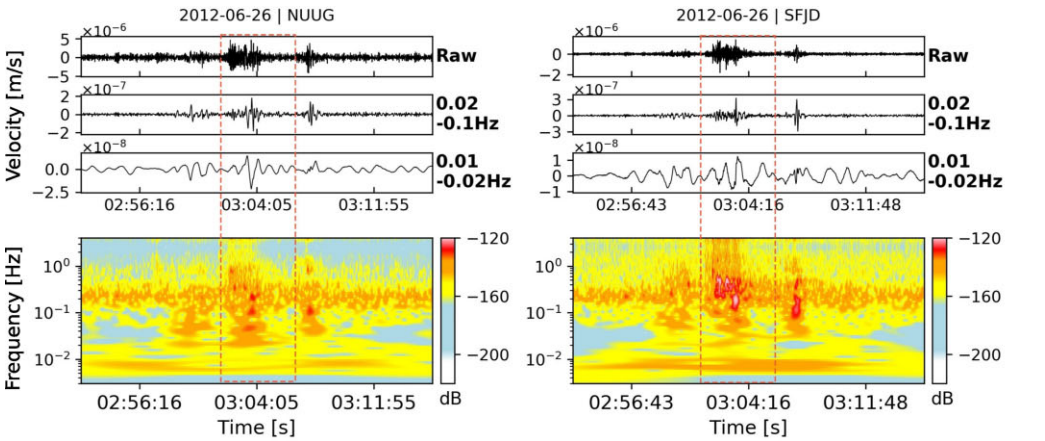

Figure A3. The catalogue event occurred on 2012-06-26 at 03:02 a.m. at Jakobshavn Isbrae, magnitude 4.7. Each subfigure shows the raw signal, the filtered signal in two frequency bands (0.02–0.1 and 0.01–0.02 Hz) and a spectrogram with signal intensity in decibels. Stations are shown on a map of Greenland, coloured according to the amplitude of the filtered signal in 0.01–0.02 Hz. The event is represented by a blue star. The signal from events is framed by red dotted lines.

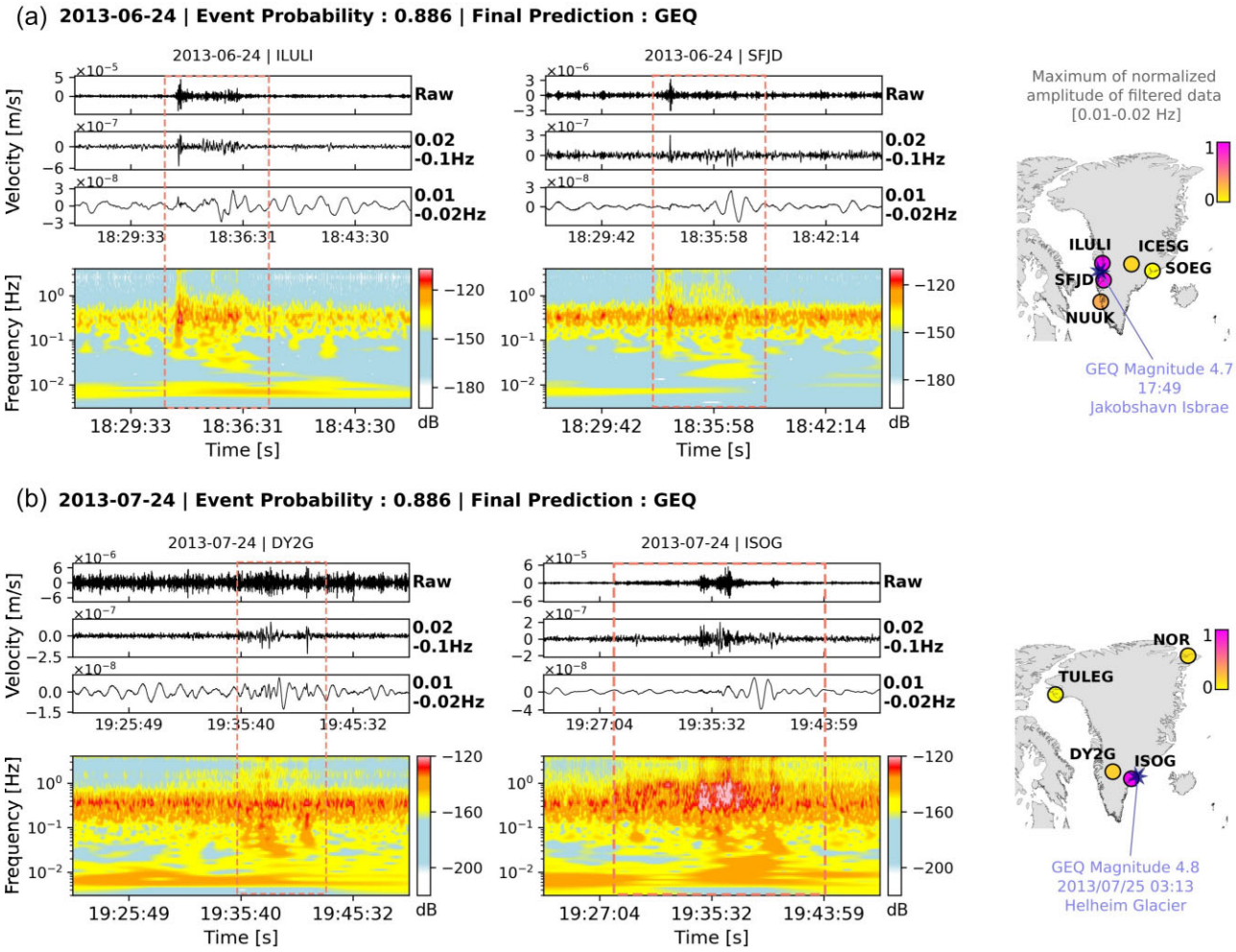

Figure A4. Two events that occurred in June and July 2013, recorded at respectively 5 and 4 stations. Each subfigure shows the raw signal, the filtered signal in two frequency bands (0.02–0.1 and 0.01–0.02 Hz) and a spectrogram with signal intensity in decibels. Stations are shown on a map of Greenland, coloured according to the amplitude of the filtered signal in 0.01–0.02 Hz. The event is represented by a blue star. The signal from events is framed by red dotted lines.

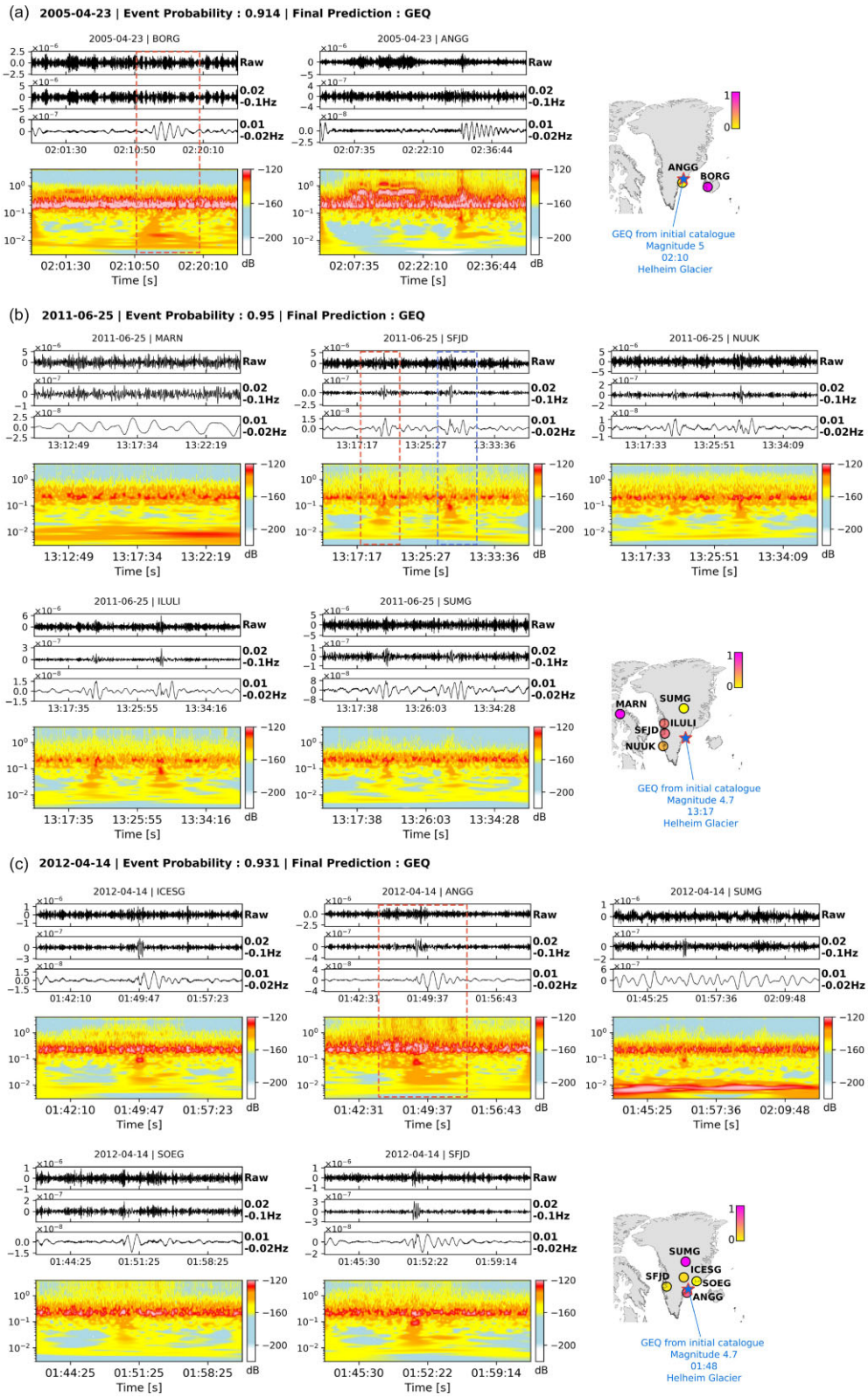

Figure B1. Helheim Glacier. Events are indicated by stars on the map on the right, with estimated magnitude and time. The signal is framed by a dotted line on one of the stations to highlight it. The stations are located and marked by a colour gradient corresponding to the amplitude of the signal in the frequency band of the filtered signal between 0.01 and 0.02 Hz, the station with the largest amplitude being in pink (1) and the smallest in yellow (0). (a) Event of 2005-04-23 recorded on 2 stations BORG and ANGG. (b) Event of 2011-06-25 recorded on 5 stations MARN (station far from the event and showing no trace of the signal), SFJD, NUUK, ILULI and SUMG. Two signals follow each other a few minutes apart (13:17 and 13:25). (c) 2012-04-14 event recorded on 5 stations: ICESG, ANGG, SUMG, SOEG and SFJD.

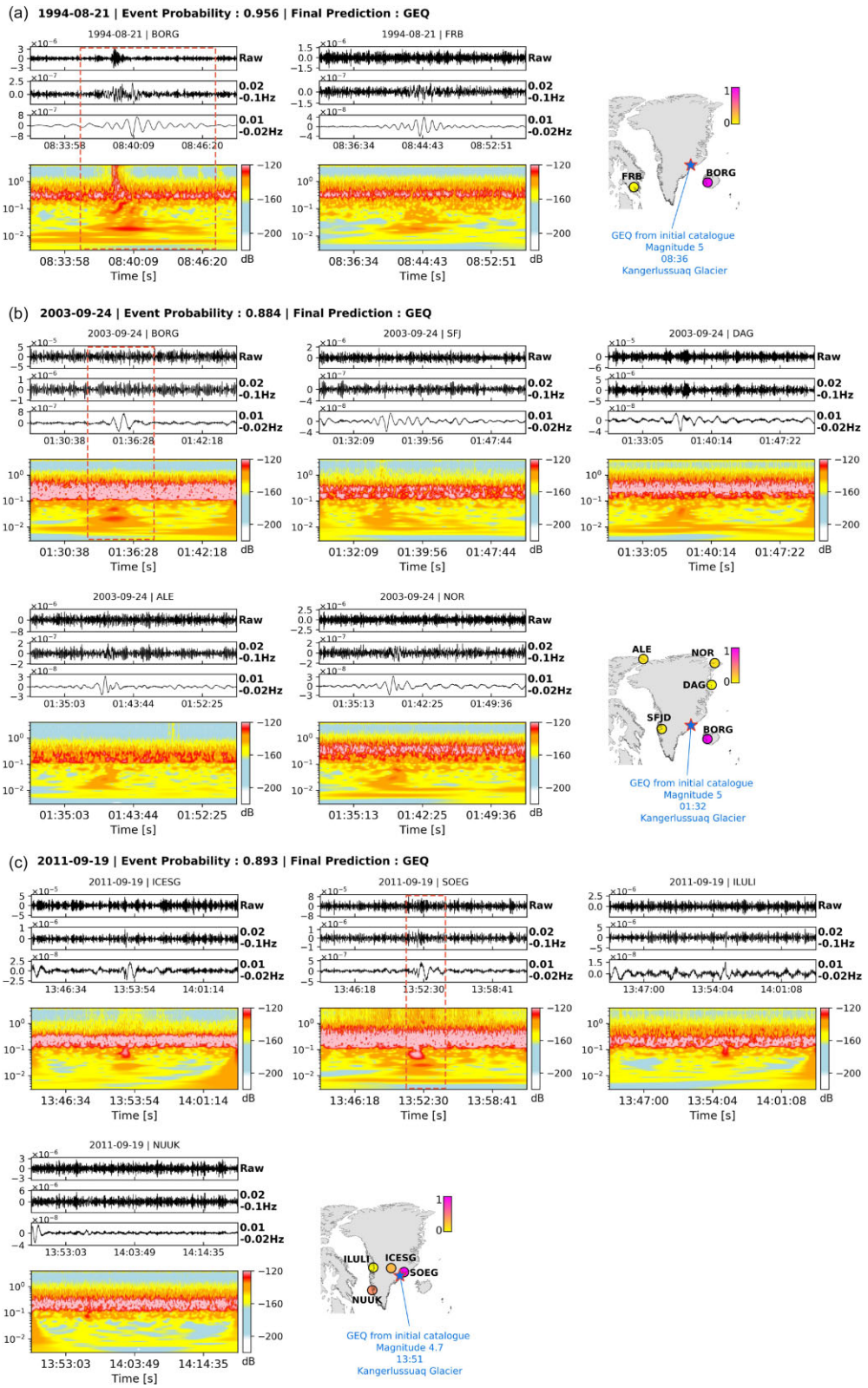

Figure B2. Three events at the Kangerlussuaq glacier. (a) Event of 1994-08-21 recorded at 2 stations: BORG and FRB. (b) Event of 2003-09-24 recorded at 5 stations: BORG, SFJ, DAG, ALE and NOR. (c) Event of 2011-09-19 recorded at 4 stations: ICESG, SOEG, ILULI and NUUK.

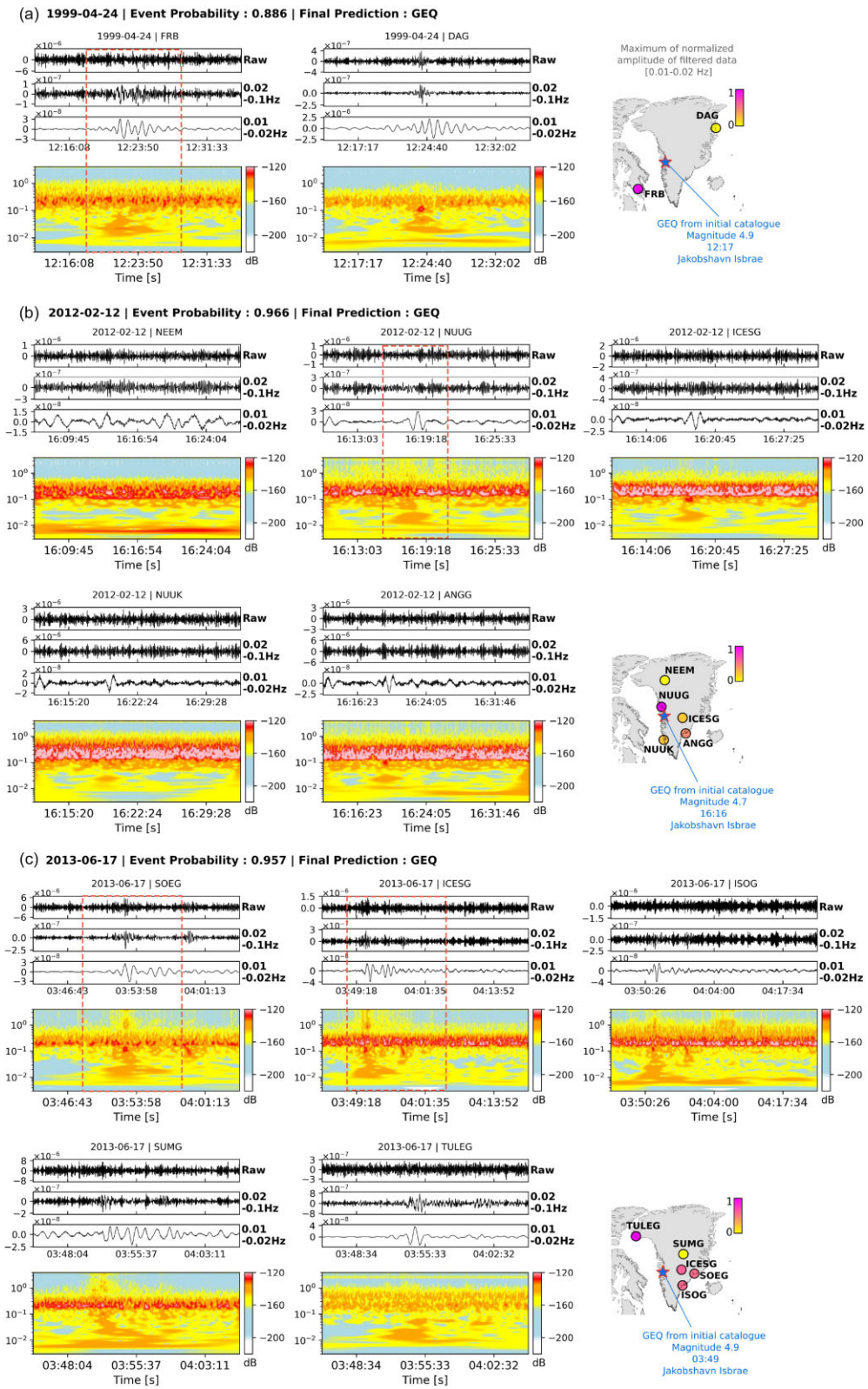

Figure B3. Three events at Jakosbshavn Isbrae. (a) Event of 1999-04-24 recorded on 2 stations FRB and DAG. (b) Event of 2012-02-12 recorded on 5 stations NEEM, NUUG, ICESG, NUUK and ANGG. (c) Event of 2013-06-17 recorded on 5 stations: SOEG, ICESG, ISOG, SUMG and TULEG.

Table C1. Features and corresponding description.

Number	Name
1	Duration of the signal
2	Ratio of the Max and the Mean of the normalized envelope
3	Ratio of the Max and the Median of the normalized envelope
4	Ascending time/Decreasing time of the envelope
5	Kurtosis Signal
6	Kurtosis Envelope
7	Skewness Signal
8	Skewness envelope
9	Number of peaks in the autocorrelation function
10	Energy in the 1/3 around the origin of the autocorrelation function
11	Energy in the last 2/3 of the autocorrelation function
12	Ratio of the energies calculated in 10 and 11
13	Energy of the seismic signal in the 0.1–1 Hz frequency band
14	Energy of the seismic signal in the 1–2 Hz frequency band
15	Energy of the seismic signal in the 2–Nyquist Hz frequency band
16	Energy of the seismic signal in the 0.01–0.02 Hz frequency band
17	Energy of the seismic signal in the 0.02–0.05 frequency band
18	Kurtosis of the signal in the 0.1–1 Hz frequency band
19	Kurtosis of the signal in the 1–2 Hz frequency band
20	Kurtosis of the signal in the 2–Nyquist Hz frequency band
21	Kurtosis of the signal in the 0.01–0.02 Hz frequency band
22	Kurtosis of the signal in the 0.02–0.05 Hz frequency band
23	Difference between decreasing coda amplitude and straight line
24	Ratio between max envelope and duration
25	Mean FFT
26	Max FFT
27	Frequency at Max(FFT)
28	Frequency of spectrum centroid
29	Frequency of 1st quartile
30	Frequency of 3rd quartile
31	Median Normalized FFT spectrum
32	Var Normalized FFT spectrum
33	Number of peaks in normalized FFT spectrum
34	Mean peaks value for peaks >0.7
35	Energy in the 1 – NyF/4 Hz band
36	Energy in the NyF/4 – NyF/2 Hz band
37	Energy in the NyF/2–3*NyF/4 Hz band
38	Energy in the 3*NyF/4 – NyF/2 Hz band
39	Spectrum centroid
40	Spectrum gyration radio
41	Spectrum centroid width
42	Kurtosis of the envelope of the maximum energy on spectro
43	Kurtosis of the envelope of the median energy on spectro
44	Ratio Max DFT(t)/ Mean DFT(t)
45	Ratio Max DFT(t)/ Median DFT(t)
46	Number of peaks Max DFTs(t)
47	Number of peaks Mean DFTs(t)
48	Number of peaks Median DFTs(t)
49	Ratio Max/Mean DFTs(t)
50	Ratio Max/Median DFTs(t)
51	Number of peaks X centroid Freq DFTs(t)
52	Number of peaks X Max Freq DFTs(t)
53	Ratio Freq Max/X Centroid DFTs(t)
54	Mean distance between Max DFT(t) Mean DFT(t)
55	Mean distance between Max DFT Median DFT
56	Distance Q2 curve to Q1 curve (QX curve = envelope of X quartile of DTFs)
57	Distance Q3 curve to Q2 curve
58	Distance Q3 curve to Q1 curve
59	Energy of the seismic signal in the 0.01–0.05 Hz frequency band
60	Energy of the seismic signal in the 0.05–0.1 Hz frequency band
61	Energy of the seismic signal in the 0.01–0.1 Hz frequency band
62	Energy of the seismic signal in the 0.1–0.5 Hz frequency band
63	Kurtosis of the signal in the 0.01–0.05 Hz frequency band
64	Kurtosis of the signal in the 0.05–0.1 Hz frequency band
65	Kurtosis of the signal in the 0.01–0.1 Hz frequency band

Table C1. Continued

Number	Name
66	Kurtosis of the signal in the 0.1–0.5 Hz frequency band
67	Difference of energy 0.1–1 Hz/1–2 Hz
68	Difference of energy 0.1–1 Hz/2 Hz–Nq
69	Difference of energy 0.1–1 Hz/0.01–0.02 Hz
70	Difference of energy 0.1–1 Hz/0.01–0.05 Hz
71	Difference of energy 0.1–1 Hz/0.05–0.1 Hz
72	Difference of energy 1–2 Hz/2–Nq Hz
73	Difference of energy 1–2 Hz/0.01–0.02 Hz
74	Difference of energy 1–2 Hz/0.01–0.05 Hz
75	Difference of energy 1–2 Hz/0.05-0.1 Hz
76	Difference of energy 2 Hz-Nq/0.01–0.02 Hz
77	Difference of energy 2 Hz-Nq/0.01–0.05 Hz
78	Difference of energy 2 Hz-Nq/0.05–0.1 Hz
79	Difference of energy 0.01–0.02 Hz/0.01–0.05 Hz
80	Difference of energy 0.01–0.02 Hz/0.05–0.1 Hz
81	Difference of energy 0.01–0.05 Hz/0.05–0.1 Hz
82	Ratio of energy 0.1–1 Hz/1–2 Hz
83	Ratio of energy 0.1–1 Hz/2 Hz–Nq
84	Ratio of energy 0.1–1 Hz/0.01–0.02 Hz
85	Ratio of energy 0.1–1 Hz/0.01–0.05 Hz
86	Ratio of energy 0.1–1 Hz/0.05–0.1 Hz
87	Ratio of energy 1–2 Hz/2 Hz–Nq
88	Ratio of energy 1–2 Hz/0.01–0.02 Hz
89	Ratio of energy 1–2 Hz/0.01–0.05 Hz
90	Ratio of energy 1–2 Hz/0.05–0.1 Hz
91	Ratio of energy 2 Hz–Nq/0.01–0.02 Hz
92	Ratio of energy 2 Hz–Nq/0.01–0.05 Hz
93	Ratio of energy 2 Hz–Nq/0.05–0.1 Hz
94	Ratio of energy 0.01–0.02 Hz/0.01–0.05 Hz
95	Ratio of energy 0.01–0.02 Hz/0.05–0.1 Hz
96	Ratio of energy 0.01–0.05 Hz/0.05–0.1 Hz
97	SNR