A real scale application of a novel set of spatial and similarity features for detection and classification of natural seismic sources from Distributed Acoustic Sensing data

Camille Huynh, C. Hibert, C. Jestin, J. -P. Malet, V. Lanticq

▶ **To cite this version:**

Camille Huynh, C. Hibert, C. Jestin, J. -P. Malet, V. Lanticq. A real scale application of a novel set of spatial and similarity features for detection and classification of natural seismic sources from Distributed Acoustic Sensing data. Geophysical Journal International, 2024, ⟨10.1093/gji/ggae382⟩. ⟨insu-04763870⟩

HAL Id: insu-04763870

https://insu.hal.science/insu-04763870v1

Submitted on 4 Apr 2025

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

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

Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Geophysical Journal International

Geophys. J. Int. (2025) **240,** 462–482
Advance Access publication 2024 October 23
GJI Seismology

https://doi.org/10.1093/gji/ggae382

A real scale application of a novel set of spatial and similarity features for detection and classification of natural seismic sources from distributed acoustic sensing data

C. Huynh [1,2], C. Hibert [1,3], C. Jestin,[2,*] J.-P. Malet [1,3] and V. Lanticq[2]

[1]*Institut Terre et Environnement de Strasbourg (ITES), CNRS UMR 7063 Université de Strasbourg – 5 rue René Descartes, F-67084 Strasbourg, France. E-mail: camille.huynh@unistra.fr*
[2]*FEBUS Optics – 2 av. du Président Pierre Angot, F-64000 Pau, France*
[3]*Ecole et Observatoire des Sciences de la Terre (EOST), CNRS UAR 830 Université de Strasbourg – 5 rue René Descartes, F-67084 Strasbourg, France*

Accepted 2024 October 13. Received 2024 September 20; in original form 2024 May 30

SUMMARY

Distributed acoustic sensing (DAS) turns a fibre optic into a very dense network of equally distributed seismic sensors. We focused on the high-density sampling of the seismic wavefield, expressed in strain rates, measured by DAS. Classical approaches used to identify seismic signals rely on the recorded features at one station, but it is difficult to include spatial information in case of dense seismic station networks. This work aims at introducing new spatial and similarity features for seismic event classification suitable to analyse DAS observations. We propose a processing chain based on the XGBoost algorithm and the use of specifically designed spatiotemporal and similarity features for the event classification, and Markov random field for the spatial clustering. The methodology is designated to be applied on a continuous stream of DAS observations. We tested our processing chain to detect earthquakes and quarry blasts recorded in the region by permanent seismic networks and included in the RENASS catalogue. These events are part of a strain-rate seismic survey carried out during a 3 weeks campaign of DAS measurements along à 91 km fibre optic cable deployed in the central Pyrenees mountains (France). Despite the high anthropogenic activities along the fibre optic path, the proposed method succeeded in detecting earthquakes of magnitude >0.4 and quarry blasts of magnitude >1.0 while limiting the number of false alarms. This performance is particularly noteworthy for low-magnitude events, where detection is accomplished despite a lower signal-to-noise ratio compared to traditional seismometers. The methodology opens the door to real time detection and classification of seismic events measured with long-distance fibre optic systems.

Key words: Machine learning; Spatial analysis; Distributed acoustic sensing; Earthquake hazards; Seismic noise.

1 INTRODUCTION

The Pyrenees mountains, located in Western Europe between France and Spain, are the result of the collision of the Iberian microplate with the Eurasian plate at a rate of 0.85 mm yr^{-1} (Fernandes *et al.* 2007). Considered as the second most active seismic zone in France after the Alps, the area has been monitored for the past 25 yr by a geographically well-distributed seismological network, which has led to a better understanding of the tectonic processes in the region (Souriau & Pauchet 1998; Rigo *et al.* 2005; Lacan & Ortuño 2012; Sylvander *et al.* 2022). Nowadays, monitoring is carried out by a permanent network of about thirty seismic stations managed by the 'Réseau de Surveillance Sismique des Pyrénées' of the French Observatoire Midi-Pyrénées. Seismic events are continuously detected and catalogued by the French facilities 'Bureau Central et Sismologique Français' and 'REseau NAtional de Surveillance Sismique' (BCSF-RENASS). Within a 50 km radius around Lourdes, located in the Central Pyrenees, six permanent stations are available, making the use of new families of sensors like distributed fibre optic sensing (DFOS) an opportunity to create a more exhaustive catalogue of seismic events for the territory.

*Now at Département Acoustique Sous-Marine (ASM) Service Hydrographique et Oceanographique de la Marine (SHOM) – 13, rue du Chatellier, CS 92803, CEDEX 2, F-29228 Brest, France

© The Author(s) 2024. Published by Oxford University Press on behalf of The Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

Using a laser pulse propagating along a fibre optic cable, DFOS are able to measure various changes in the neighbourhood of the fibre. The intrinsic presence of asperities on the fibre core results in light backscattering effects that hold information about different physical properties. Among the DFOS family, distributed acoustic sensing (DAS), based on Rayleigh backscattering and phase shift, is sensitive to seismic acoustic waves crossing the interrogated fibre optic cable. Spatial sampling, corresponding to the spacing between each point of measurement along the fibre, can be set down to a few tens of centimetres and acquisition rates can be set up to hundreds of kHz. Previous studies have demonstrated the contribution of spatial analysis for the analysis of DAS records in order to take into account the spatial character of the measurement. Spatial filtering, such as F–K filtering (Hudson *et al.* 2021; Fukushima *et al.* 2022), or curvelet decomposition (Atterholt *et al.* 2022) can be applied for data denoising. Further, spatial redundancy makes the DAS system suitable to capture low-magnitude events in delimited geographical areas, and a high diversity of low magnitude environmental sources (landslides, avalanches, floodings,...) difficult to detect with permanent seismic stations but crucial to reach a better understanding of the regional seismicity of a territory. Further, DAS have proven to yield interesting new results in different contexts and for many different applications, for marine geophysics (Sladen *et al.* 2019; Spica *et al.* 2022), seismic imaging (Lindsey *et al.* 2017; Young *et al.* 2022; Zeng *et al.* 2022), volcanology (Jousset *et al.* 2022) and water reservoir monitoring (Zhu *et al.* 2021).

The dense measurement network and the high sensitivity of the DAS allow recording a large number and variety of seismic sources, making manual identification difficult and time-consuming. Classical automated detection and identification, based on short-time-average over long-time-average method (STA/LTA) or energy threshold, are complex to implement: the high diversity of events occurring along a fibre, as well as the increase of noise level with increasing distance, can generate high energy seismic signals making difficult the choice of suitable detection parameters. For natural event detection, instrumental noise and anthropogenic events should be filtered (Nayak *et al.* 2021; Chen *et al.* 2023).

Artificial intelligence (AI) techniques are the most promising solution for DAS data automated classification. Two main families of AI algorithms for classification purposes exist: feature-based algorithms, also named machine learning algorithms, and deep-learning-based algorithms. Deep learning algorithms rely on the structure of deep neural networks to build their own discriminating features for the classification and detection of certain types of events such as micro-earthquakes (Binder & Tura 2020) or footsteps (Jakkampudi *et al.* 2020). It is also possible to tackle topics difficult to solve with features, such as signal denoising (Ende *et al.* 2021; Zhao *et al.* 2021). Deep learning approaches require a large data set for proper model training, and do not allow an access to a deep understanding of the algorithm decision criteria. They also require appropriate data representation, both in terms of the choice of input data and the scale at which they are represented. In contrast, feature-based methods rely on the use of human engineered seismic signal features which have been defined in previous studies during different studies of seismic event discrimination. Temporal-related features had been widely used in seismic data analysis (Hibert *et al.* 2014, 2017; Maggi *et al.* 2017; Provost *et al.* 2017; Hibert *et al.* 2019; Chmiel *et al.* 2021; Domel *et al.* 2023). In addition to waveform analysis, features are often derived from signal transformation into the frequency domain using Fast Fourier Transform (Wiesmeyr *et al.* 2020; Tejedor *et al.* 2021), wavelet decomposition (Wang *et al.* 2019) or the use of the mel-frequency cepstral coefficient (Bublin 2021). Because of their ability to quantify individual contributions of each feature used for the classification, various feature-based AI algorithms are suited to build a must-have feature list adapted to the events we want to identify. This is the case for random forest and XGBoost machine learning algorithms (Breiman 2001; Chen & Guestrin 2016).

Huynh *et al.* (2022) proposed a methodology enabling to classify DAS data streams using seismic signal derived features that present interesting results when implemented for a 20 m-long trench for classifying anthropogenic event sources. In this article, we aim to apply a similar workflow for detecting and identifying earthquakes and quarry blasts under real-world field conditions. Expanding our preliminary processing chain to work the regional scale is challenging due to (1) the variety of sources in an area spanning dozens of kilometres; (2) the spatial extension of the array, which lead to the recording of numerous synchronous seismic sources and (3) the background noise level which can be very different at each part of the buried fibre optics (when going through dense habitation area, along roads, along rivers, etc.). We propose in this study a new processing chain, which use new seismic signals features designed specifically to benefit from the dense spatial distribution of DAS data, to overcome those difficulties.

2 FIBRE OPTIC DAS DATA

Data acquisition has been achieved along a 91 km long fibre optic deployed between Lannemezan and Luz-Saint-Sauveur in the Bigorre area, located in central French Pyrenees. The fibre is deployed close to several urban centres (Bagnères-de-Bigorre, Lourdes, Argelès-Gazost) and close to several quarries (for rock extraction) in activity. The cable also follows national and departmental roads (Fig. 1). Seismic data measured by the DAS, named strain rate (SR) and expressed in nanostrain per second (nstrain s^{-1}) are extracted from the optical raw data by setting the optical-acoustic processing parameters prior to acquisition, known as gauge length and derivation time (Hartog 2017). Setting these parameters is important because they have a direct impact on the measurement quality: the derivation time affects the maximum observable frequency, while the gauge length affects the spatial resolution and then the minimum detectable wavelengths. To fit with the observation of natural seismic events, SR has been recorded using a gauge length fixed to 10 m and a cut-off frequency of 200 Hz. Because of instrumental response biases, common mode frequency removal has to be carried out prior to data exploitation. Thus, average removal is performed on the strain rate raw data for each channel of measurement.

With a fixed gauge length of 10 m and assuming a 0.2 dB km^{-1} attenuation rate of the optical laser pulse along the fibre optic cable, the maximum distance range achievable is 50 km. To extend beyond this range while maintaining a favourable signal-to-noise ratio, additional optical devices can be used in combination with an interrogator such as repeaters, dual-channel interrogators or long-range systems. The FEBUS A1-R DAS has been installed in the town of Lannemezan, supplemented by a FEBUS range extension module connected at the middle of the fibre, in Lourdes. Recording has been carried out continuously over 3 weeks between 2022 Au-

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

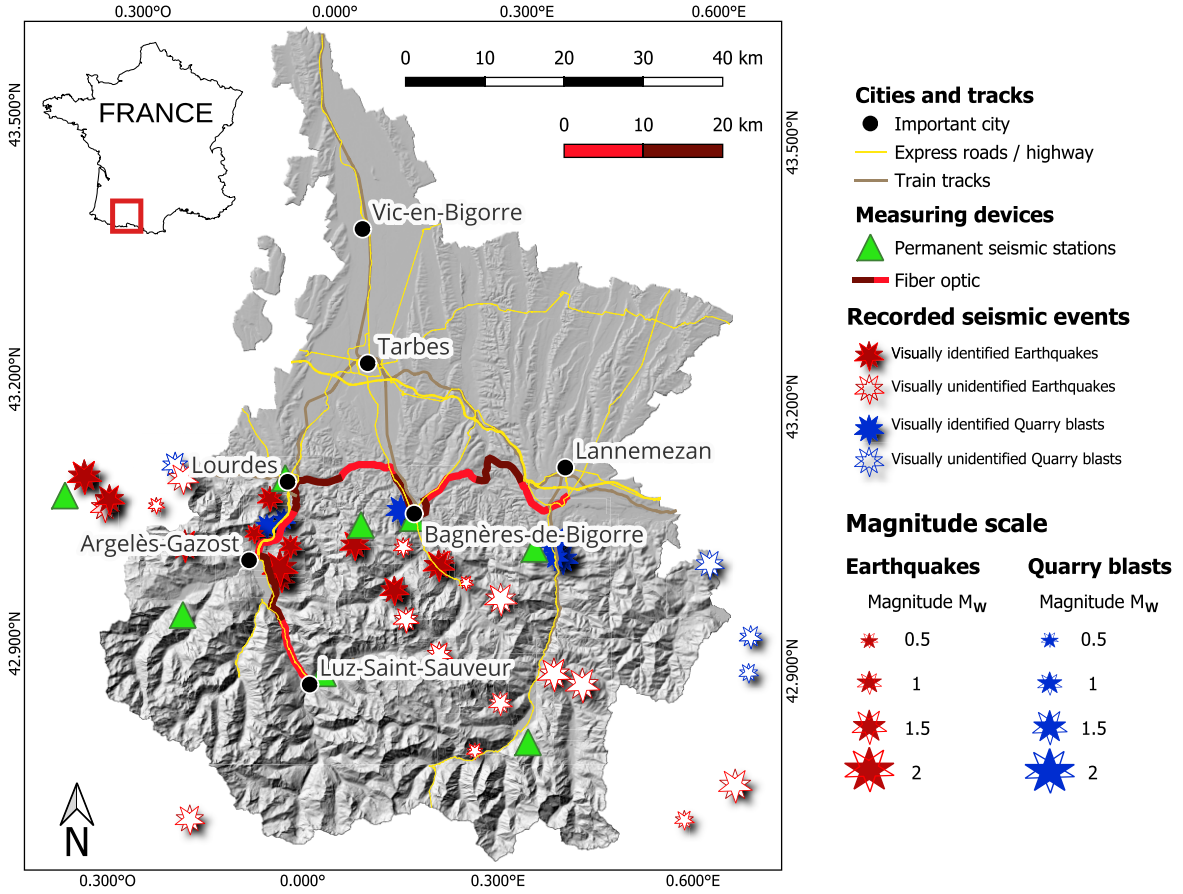

Figure 1. Fibre optic setup in the French area 'Hautes-Pyrénées'. The background corresponds to the topography of the area. The map includes location of main cities, roads and train tracks, as well as the positions of the catalogued earthquakes and quarry blasts from the BCSF-RENASS (using the permanent seismic stations represented by triangles). The visually unidentified events on DAS data are represented by unfilled stars.

gust 30 and September 20. The final acoustic DAS data are recorded with a spatial sampling of 4.8 m. This equates to 18 958 point sensors, called channels, equally spaced along the 91-km fibre optic. The DAS data size consists of 40 TB.

A first observation of the acquired data indicates the presence of a large number of anthropogenic events in the monitored area along the entire length of the fibre optic cable: we observe the presence of a day–night cycle with an increase in anthropogenic activity during the day, and clearly identify the presence of seismic signal corresponding to vehicles circulating on roads (Fig. 2). However, in the absence of appropriate detection techniques for DAS systems, the identification of natural earthquakes and other environmental natural and anthropogenic seismogenic events is difficult.

We use the BCSF-RENASS catalogue to build a reference catalogue of seismic events. BCSF-RENASS provides an open-access catalogue of natural and anthropogenic events (earthquakes, quarry blasts, induced seismicity, explosions) recorded by permanent seismic stations deployed in France and in neighbouring countries. By cross-referencing the seismological waveforms between the stations, events are detected and localized automatically. Human operators then validate the detection, the location and determine the natural (e.g. earthquakes, landslides, rockfalls) or anthropogenic (e.g. quarry blasts, explosion) origin of the event.

32 events (8 quarry blasts and 24 earthquakes of magnitude M_w estimated between 0.3 and 2.4) have been detected and listed in the BCSF-RENASS catalogue for the study period. Among these events, we visually confirmed the presence of 13 earthquakes and 6 quarry blasts in the DAS recordings (Fig. 1), with data available online (see 'Data Availability' section). It appears that most of the unobserved events on the processed DAS data occur beyond Bagnères-De-Bigorre in south-east direction. The absence of detection can have multiple explanations, based on the instrumental spatial configuration and geometry of the fibre optic cables (fibre optic is insensitive to seismic waves that arrive perpendicular to the fibre optic) or on the local geological conditions (nature of the geological media can modify the propagation of the seismic signal in the ground). For our analysis, we focus on the classification of three classes of events: 'earthquakes', 'quarry blasts' and 'noise'. 'Noise' groups all uninvestigated events (daily anthropogenic events as moving vehicles, wildlife events, aerial sections of fibre optic cable or instrumental noise). For purpose of detection, we will refer to earthquakes and quarry blasts as 'events', and the other sources, such as uninvestigated anthropogenic noise, as 'noise'.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

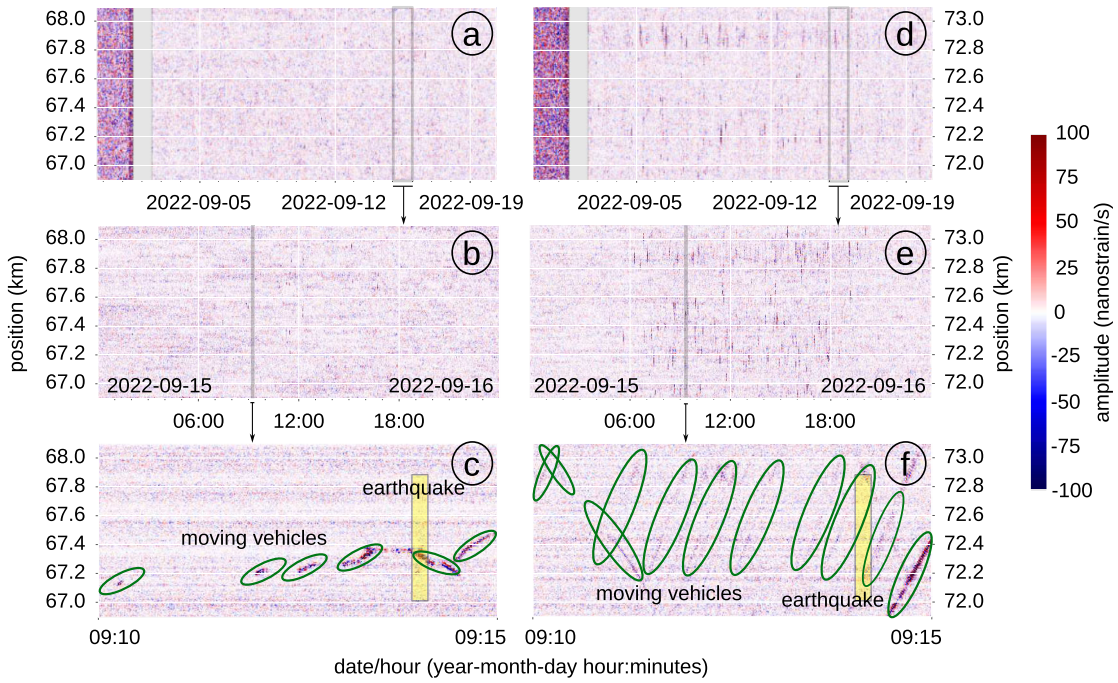

Figure 2. Samples of SR records on the Pyrenees DAS data set for several time windows, in the countryside and close to a city (GPS coordinate: 43.01298 °N, −0.08660 °E). (a) represents the data recorded over the whole acquisition period, at position between 66.9 and 68.1 km (countryside), (b) shows the data acquired over a period of 24 h around the same section and (c) the data recorded over a period of 300 s around the same section. (d, e and f) represent similar records at position between 71.9 and 73.1 km (close to Argelès-Gazost). Moving vehicles are delimited by circles and earthquakes by rectangles inside the 300 s view. The 0 km-position reference is located at the end of the fibre optic cable point close to Lannemezan.

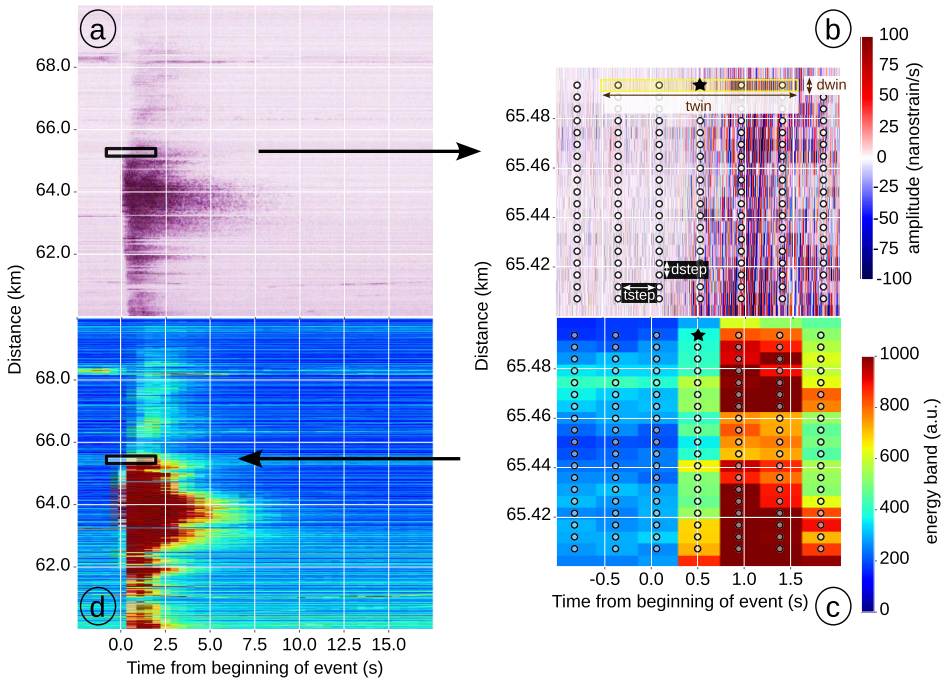

Figure 3. EB representation from SR. The mapping is described inside a small part of the DAS data, delimited in (a) and (d) by a rectangle. We define a new grid above the SR representation represented in (b) by circles. Each circle is separated by t_{step} and d_{step}. Taking one element of the grid (e.g. the black star in b), energy is computed using the SR contained inside a window centred on the element (represented by rectangle for the star element in b) and which size is given by (t_{win}, d_{win}). In the EB representation, each point corresponds to the computed energy at each grid position (c).

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

Table 1. Table listing all events identified on the DAS data set. An identifier (id) is assigned, and information on event type (Class), magnitude (Mw) and measurement time (Time) are provided. The projected position of the epicentre on the fibre ($D_{\parallel \text{fibre}}$) and the distance from the epicentre to the fibre ($D_{\perp \text{fibre}}$) are used to check the consistency of what is observed. The reader can refer to Appendix A for the SR and EB visualization of all the listed events.

id	Class	Mw	$D_{\parallel \text{fibre}}$ (km)	$D_{\perp \text{fibre}}$ (km)	Time (UTC)
QB1	QB	0.7	59	1.8	2022-08-31 at 08:02:39
QB2	QB	0.8	3.4	5.8	2022-09-01 at 09:26:00
EQ1	EQ	0.6	59	1.7	2022-09-03 at 03:50:17
EQ2	EQ	1.0	30	10	2022-09-03 at 13:13:41
EQ3	EQ	2.4	62	85	2022-09-03 at 18:27:47
EQ4	EQ	1.4	61	25	2022-09-05 at 00:51:58
EQ5	EQ	1.2	30	6.5	2022-09-06 at 07:58:54
QB3	QB	1.0	3.5	4	2022-09-06 at 10:10:58
QB4	QB	1.1	32	1	2022-09-08 at 10:03:36
EQ6	EQ	0.8	53	3.2	2022-09-08 at 17:10:59
EQ7	EQ	2.0	61	50	2022-09-09 at 07:07:40
EQ8	EQ	0.4	61	1.5	2022-09-09 at 17:37:59
QB5	QB	0.6	57	0.5	2022-09-12 at 10:07:43
QB6	QB	1.1	3.5	6.3	2022-09-14 at 09:26:07
EQ9	EQ	1.1	57.5	8.5	2022-09-15 at 09:12:16
EQ10	EQ	1.6	66	1.4	2022-09-16 at 11:16:50
EQ11	EQ	1.1	61	20	2022-09-16 at 23:12:33
EQ12	EQ	0.8	68	1.0	2022-09-18 at 04:18:12
EQ13	EQ	1.3	63	9.5	2022-09-20 at 04:39:26

3 METHODOLOGY

3.1 Event picking and identification for labelling

The massive seismological data set recorded during this study requires high-performance analysis tools to detect events of interest. Identifying them by direct observation of the SR is time-consuming especially when we are also interested in detecting low-magnitude events or in identifying certain types of event. We therefore work with a representation called energy band (EB), obtained for each channel along the fibre by summing the energy contained in the spectra (calculated using a Fast Fourier Transform, FFT) of a seismic trace window ($\text{FFT}_{\text{trace}}(f)$) between two frequency bounds f_0 and f_1:

$$E_{[f0,f1]} = \sum_{f \in [f0,f1]} \text{FFT}_{\text{trace}}(f). \quad (1)$$

The spectra are calculated using a sliding window of 2 s, with a shift of 0.5 s. f0 and f1 can be adapted using frequency response knowledge. In our analysis, values are chosen to cover the maximal range of the frequency content: f1 corresponds to the Nyquist frequency (100 Hz), whereas f0 corresponds to 1/window_width (0.5 Hz). The resulting representation can be displayed into a new grid whose general representation is given in Fig. 3: (d_{step}, t_{step}) refers to the spatial and temporal sampling of the EB representation, whereas (d_{win}, t_{win}) refers to the window on which is computed each point of the EB representation. In our case, as the energy is computed for each channel, spatial parameters d_{step} and d_{win} are taken as the distance between consecutive channels. t_{step} and t_{win} are respectively fixed to 0.5 and 2 s. EB representation improves the visual detection of low-magnitude events because of the integration of the spectrum content over a duration of 2 s. For this reason, we keep both EB and SR representations in the following figures. The reader should keep in mind that the EB with these window parameters is not involved in the processing chain.

19 of 32 seismic events recorded by BCSF-RENASS are visually detected using the EB representation. They are listed in Table 1 and segmented from the rest of the signal. Segmentation grid resolution matches the EB resolution (4.8 m, 0.5 s). Segmentation process is achieved by drawing a rectangle box around the EB where an event is located: the points inside the box are labelled as events, whereas the points outside are considered as noise.

3.2 Processing chain

Fig. 4 details the processing chain. The input corresponds to the measured SR after baseline correction (average removal of each channel) (Fig. 4a) and the output is the associated classification map encoded in three colours (Fig. 4d). Initially, the continuous data stream is processed with a sliding window approach (Wenner *et al.* 2021). The sliding window uses two parameters that define the spatial and temporal sizes (d_{win} and t_{win}) and that defines the shift (d_{step} and t_{step}). The choice of the window size and the impact on the outcome of the processing chain is discussed in Section 4.3.

Data contained in a window is labelled and pre-processed to extract a vector of 111 features described in Section 3.3. Concatenating the vectors obtained at every spatiotemporal window centroid position outputs a feature matrix that is injected into the machine learning algorithm XGBoost for event classification. XGBoost produces a classification score for each window that can be combined to create a score map (Fig. 4b). An initial classification map is generated using a threshold of 0.95 on the score map to reduce the amount of misclassification (Fig. 4c). Then a final post-processing step (Markov random field) aggregates the classification, the score of each individual window and the classification result of neighbouring windows, to keep only the most significant classification results (Fig. 4d).

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

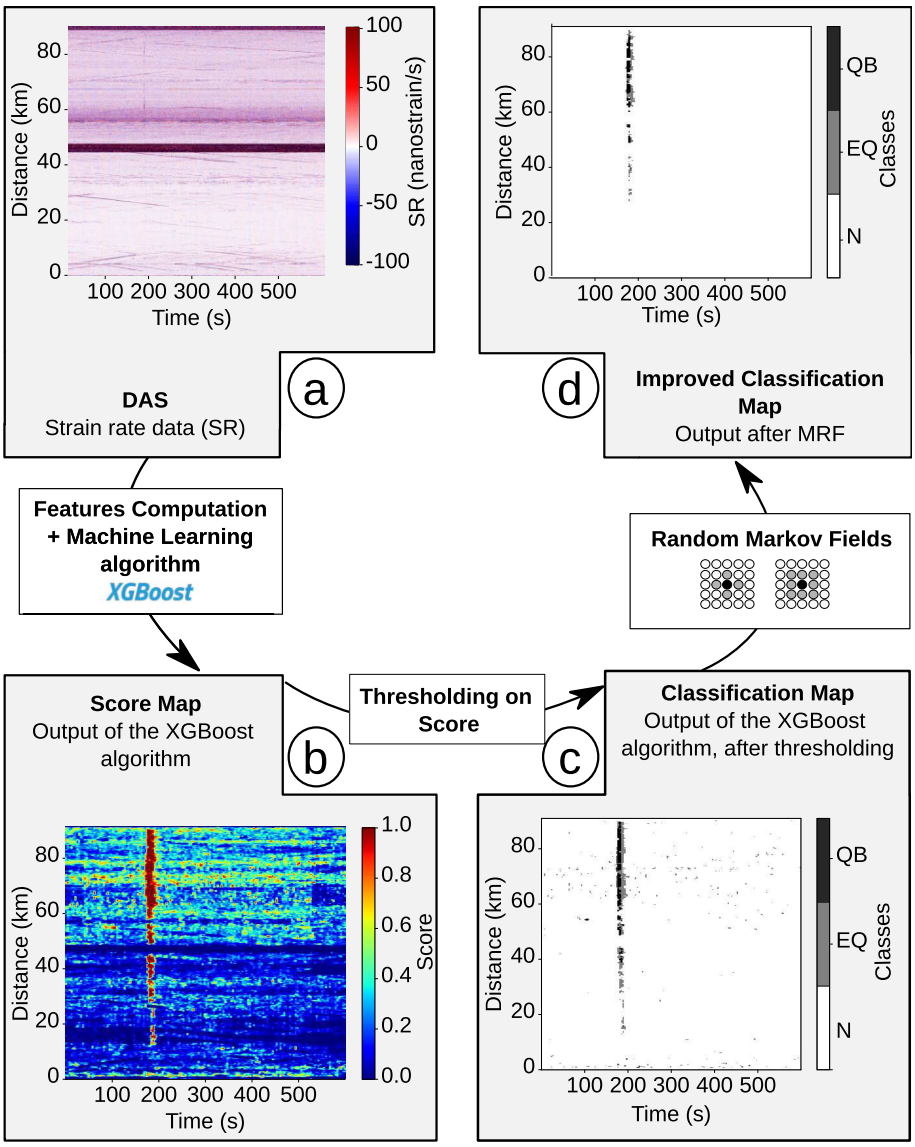

Figure 4. Overview of the processing chain. The SR data (a) is introduced in a machine learning model and produces one score map per class (b). Points with a score value higher than 0.95 are kept and produces a classification map (c). A post-processing algorithm, Markov random field (MRF), is applied on the classification map to reduce the amount of noises (d). *N* refers to noise class, EQ to earthquake and QB to quarry blast. To simplify the representation, the score map represented here corresponds to the sum of non-noise classes.

3.3 Data pre-processing: signal description using features and labellizations

Machine learning algorithms rely on the use of features. Features are specific measurable properties or characteristics extracted from the data that help the algorithm to make predictions. Our work focuses on a set of 111 features divided into three major categories: temporal features (63), spatial features (24) and similarity features (24).

Temporal features are derived from previous works on seismic signal classification using traditional seismometers recordings (Hibert *et al.* 2017, 2019; Maggi *et al.* 2017; Provost *et al.* 2017; Malfante *et al.* 2018; Chmiel *et al.* 2021; Falcin *et al.* 2021; Wenner *et al.* 2021; Domel *et al.* 2023) and DAS recordings (Huynh *et al.* 2022). The temporal features are divided in three families, named waveform, spectral and spectrogram features. Waveform features describe the evolution of the signal and its envelope in the time domain, using tools such as energy, skewness and autocorrelation measurements. Spectral features provide information on the frequency content of each of the windowed channels and measure, for example, the value of the median FFT, energy between different frequency bands or the width of the spectral centroid. Spectrogram features measure changes in spectral content over time, such as the variation of the median FFT over time.

Spatial features are introduced in this work to account for the distributed nature of the DAS measurements. Unlike the use of a dense array of seismometers, the distance between two consecutive virtual sensors is constant. Fig. 5(b) illustrates the waveform profile of a trace measured at one channel, and Fig. 5(c) the measured signal at a fixed time along the fibre. Observation of the SR (Fig. 5a) shows that, depending on the magnitude of the source, a larger or smaller portion of the fibre may record the event. Low energy sources, such as cars, low magnitude earthquakes and quarry blasts, are recorded only by a limited portion of the fibre but larger energetic sources as

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

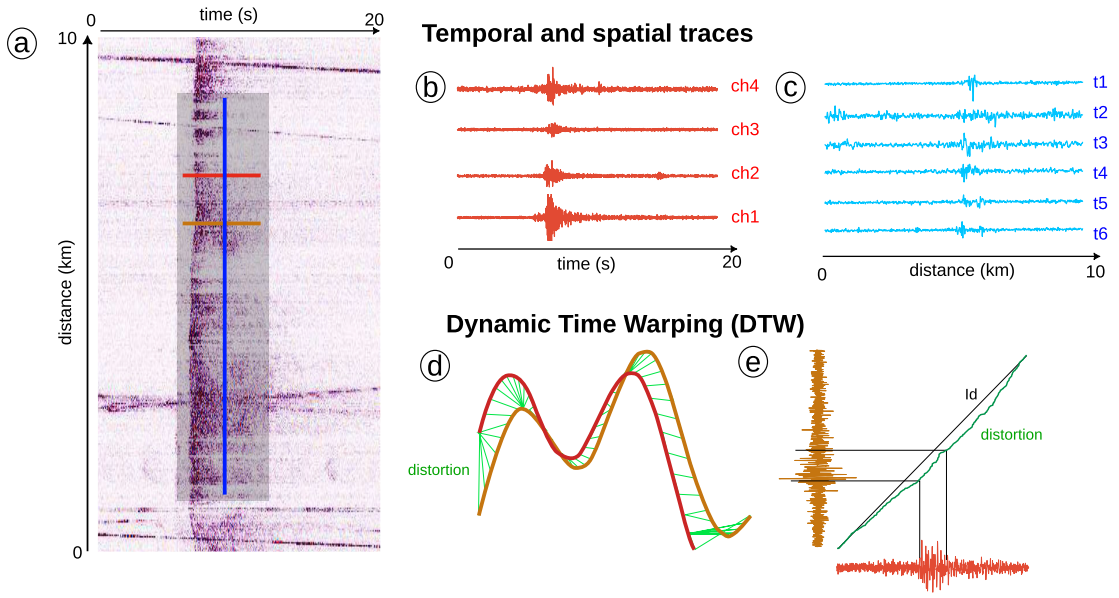

Figure 5. From the SR (a), features can be built by considering temporal trace observed at fixed position ((b), represented in (a) with horizontal line) and spatial trace observed at fixed time ((c), represented in (a) with vertical line). Similarity features include DTW, built using an estimated distortion function that stretch two traces to fit together (d and e).

earthquakes can be recorded over several tens of kilometres along the fibre.

The proposed spatial features are based on the shape of the seismic signal measured at a given time along all channels contained in the window. It includes average, standard deviation and autocorrelation such as the one computed for the temporal waveform feature category. These features are designed to account for the spatial coherence of seismic signals, the apparent velocity of different phases and dispersion effects. All these seismic signals characteristics help distinguishing visually between an earthquake and a moving vehicle, for example. We will thereafter refer to temporal trace for seismic signals recorded at one channel along the fibre (red in Fig. 5b), and to spatial trace for a seismic signal recorded at a given time on all the channels of the fibre (blue in Fig. 5c).

Similarity features are computed from the comparison of traces taken at different positions along the fibre. We choose two different methods to compute the similarity: cross-correlation and dynamic time warping (DTW). Cross-correlation estimates the optimal time-shift between the signal arrival time at the compared traces. DTW constructs an optimal mapping (warping path) between corresponding points of two sequences, allowing for nonlinear stretching or compression of the time axes to achieve the best alignment, thereby accommodating temporal distortions (Sakoe & Chiba 1978; Berndt & Clifford 1994; Keogh & Ratanamahatana 2005; Müller 2007). Kumar *et al.* (2022b) have shown the interest of using DTW for geophysical applications, including a higher sensitivity to minor variation compared to classical cross-correlation. Examples of applications are associated with full-waveform inversion in reflection seismology (Ma & Hale 2013), GNSS displacement time-series analysis (Kumar *et al.* 2022a) and classification of volcano-seismic events (Ida *et al.* 2022). The reader may refer to Appendix B for an item-by-item description of 111 used features introduced in the machine learning model.

Labelling each feature vectors is not trivial due to the resolution difference between the segmentation grid, as described in Section 3.1, and the grid used for feature vector computation. The labelling of each feature vector is determined by considering all the labels within the window used for feature computation. Given the segmentation grid resolution of 4.8 m by 0.5 s, and a window size of 998.4 m by 8 s, the feature vector labelling accounts for 3328 labels. If at least one of these labels corresponds to a 'earthquake' or 'quarry blast,' the feature vector is labelled as either 'earthquake' or 'quarry blast,' depending on which of these two classes has the majority of samples within the label grid.

3.4 Data processing: machine learning with XGBoost

We choose the machine learning algorithm called XGBoost as the identification tool in our workflow (Chen & Guestrin 2016). XGBoost is a supervised ensemble learning method based on the use of a large number of weak learners (e.g. decision trees) to predict a class. Based on the boosting technique, decision trees are sequentially ordered. Misclassified samples made by one decision tree are kept for training the next one. Compared with other machine learning algorithms, features that poorly characterize the event do not reduce the performance of the XGBoost model. This algorithm is also able to deliver quantitative information about the influence of features in the classification. Several applications in geophysical contexts include rock facies classification (Zhang & Zhan 2017), density log estimation for reservoir characterization (Zhong *et al.* 2020) and seismic source classification (Wang *et al.* 2023).

Once the training is completed, the XGBoost model uses a method based on vote analysis to provide a classification: each weak learner classifies the sample, and the class returned by the majority of them constitutes the output of the XGBoost model. A score can be deduced for each possible class by dividing the number of trees voting for one class by the total number of trees in the XGBoost model: a score close to 1 for a particular class indicates that most of the weak learners voted for this class. Conversely, for a two-class problem, a score close to 0.5 indicates that the classification of an event has a very high uncertainty, as there are no clear majority in

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

the vote casted by the weak learners. Scores are then a powerful indicator of the uncertainty of the classification and can be used to reduce the amount of false alarms.

3.5 Data post-processing: Markov random field (MRF)

Using a threshold on the score yielded by the XGBoost algorithm is a simple and efficient technique to improve the machine learning classification results. However, this technique does not consider the spatial and temporal relationships between feature vectors for data stream processing (Fig. 3). As the classification is achieved for each SR windows which spacing is defined by the parameters t_{step} and d_{step} (Fig. 3), we can assume that, in the case of a resolution at least two time smaller than the duration and the spatial footprint of an event, several neighbouring windows include the same event. Neighbouring classification is therefore a source of information that can be used in the Markov random field (MRF) method to mitigate uncertainties about event identification. MRF, originally used for image processing, is a spatial clustering algorithm relying on the classification score and the classification of neighbour points (Cross & Jain 1983). The combination of these two parameters is achieved through a loss function C that gathers class neighbourhood loss C_{neigh} and class likelihood loss C_{likeli}, defined as follows:

$$C_{\text{neigh}}(x) = \sum_{S \in V} \mathbb{1}_{\tilde{x}}(\text{Class(S)}) \tag{2}$$

$$C_{\text{likeli}}(x) = \ln L(\theta|x) \tag{3}$$

$$C(x) = \text{Potential} \times C_{\text{neigh}}(x) + C_{\text{likeli}}(x), \tag{4}$$

where V i. the neighbourhood, x is one class among the output classes of XGBoost (here x takes a value among 'noise', 'quarry blast' or 'earthquake'), $\mathbb{1}$ is the indicator function, θ is the parameter vector describing the corresponding observation (here associated with the features vector used in the XGBoost model) and L is the likelihood function. 'Potential' weights the contribution of C_{neigh} and C_{likeli} in the loss function C. In the case of a neighbourhood defined by the eight nearest neighbours, $C_{\text{neigh}} \in [0, 8]$ is equal to 0 if all neighbours are of class x, and to eight if none of the neighbours is of class x.

The efficiency of the method has been demonstrated in previous studies for the real-time classification of anthropogenic events recorded by DAS interrogator on a controlled test bench (Huynh *et al.* 2022). In particular, it helps reducing the number of false alarms by avoiding isolated misclassified points on the classification map.

3.6 Performance of the method

To define the performance of the method, we divided the DAS data set in two parts: a training set, used to build the machine learning model and to define the parameters for feature computation; and a test set used to measure the performance of the processing chain. During the three weeks of data acquisition, we recorded 19 events with several magnitudes and with hypocentres located at different geographical positions. Because of the small size of the data set, we choose a leave-one-out cross-validation (LOOCV) evaluation technique, consisting in keeping a single event as a test set for each cross-validation. Evaluation is performed for each event using the machine learning algorithm trained on the remaining events.

The method can be quantitatively validated by comparing the class predictions for each feature vector yielded by the machine learning algorithm to the manual classification of the corresponding portion of the DAS data. Quantitative criteria include multiclass precision, multiclass recall and F1-score. These values are defined by the equations:

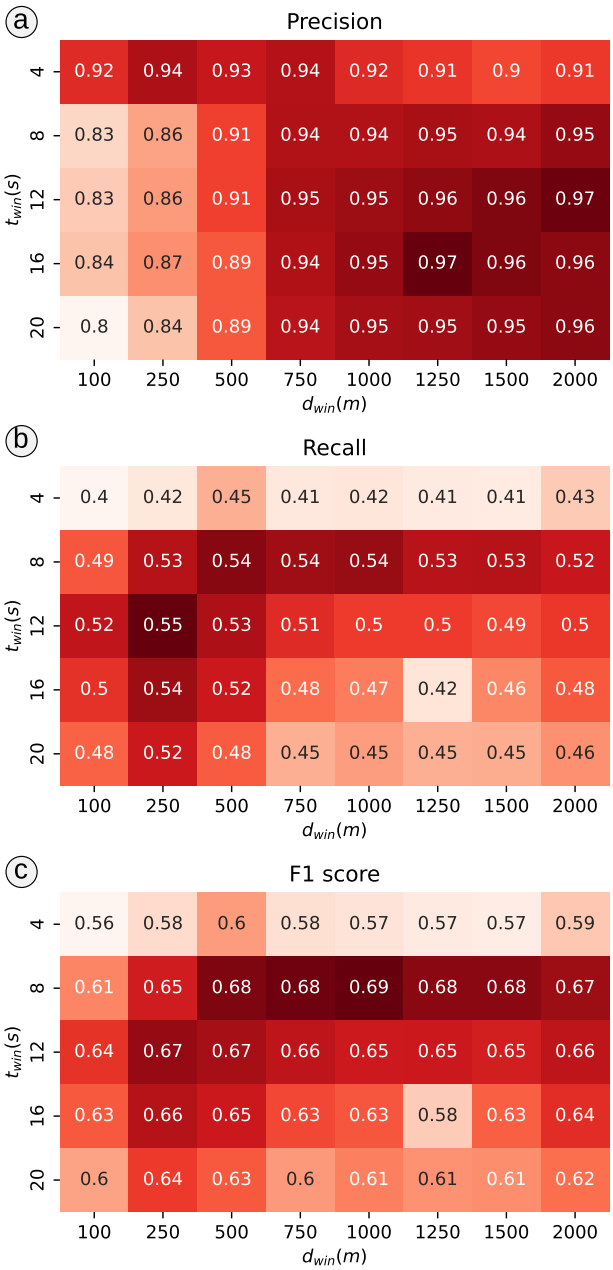

Figure 6. Performance measurements for several time and spatial windows. Results are presented in terms of precision (a), recall (b) and F1 score (c).

$$\text{Precision} = \frac{1}{N} \sum_{i \in \text{Classes}} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{5}$$

$$\text{Recall} = \frac{1}{N} \sum_{i \in \text{Classes}} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{6}$$

$$F1\text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{7}$$

where N denotes the total number of classes, TP_i the amount of true positive considering class i, FP_i the amount of false positive and FN_i the amount of false negative.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

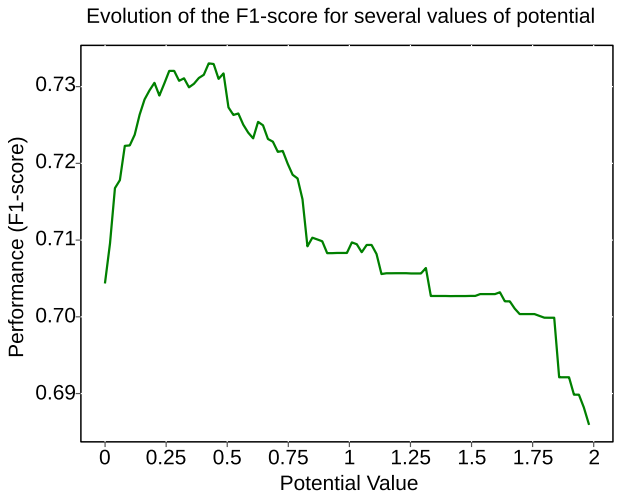

Figure 7. Performance estimation based on F1 score on the output of the Markov random field for several values of Potential. The optimal parameter which maximize this metric is Potential = 0.45.

Qualitative validation consists of a visual inspection performed after reconstructing the classification map. Incorporating MRF as the final step in the processing chain helps to filter out areas with sparse detection and low classification scores, while preserving those with few detection but high relevance. As visual validation is prone to subjectivity, we present the results of each classification in Appendix B.

4 RESULTS

4.1 Impact of DAS window size parameters

The dimensions of the window (t_{win}, d_{win}) used to compute the features have an impact on the results of the classification. We thus estimate the performance of our machine learning algorithm for several (t_{win}, d_{win}) pairs. For each pair, we use the LOOCV method for data set splitting and the three criteria presented in Section 3.6. The goal is to find the best pair that maximizes the F1 score (Fig. 6c) derived from precision (Fig. 6a) and recall (Fig. 6b). Results given in Fig. 6 show that the choice of a temporal window t_{win} smaller than 4 s, or higher than 20 s, degrades the recall. Similar conclusions can be drawn for the choice of a spatial window d_{win} smaller than 100 m. We note that the optimization of precision values is obtained for spatial windows d_{win} higher than 250 m. In the range $t_{\text{win}} \in [8, 12]$ s and $d_{\text{win}} \in [500, 1250]$m, F1-score value is the highest with values between 0.65 and 0.69. Choosing large windows implies a longer computation time. With our data set, we measure that doubling the duration or the spatial length of the window doubles the computation time. Looking for a compromise between model performance and computation time, we selected for the rest of our work the parameters: (t_{win}, d_{win}) = (8 s, 1000 m), and (t_{step}, d_{step}) = (4 s, 100 m). We note that, for these values, recall is very low compared to precision. This implies that our model does not detect a part of the points that belong to events but the alarm is reliable when a point corresponding to an event is detected. Fig. A1 in Appendix A for several events indicates that some of the points associated with an event are undetected because of their low energy.

4.2 Impact of post-processing parameters

The post-processing, based on the thresholding of the scores and on the application of the MRF algorithm, relies on the use of two parameters named *Score_threshold* and Potential. The *Score_threshold* is applied to the XGBoost score (Section 3.4) and is chosen by visual inspection to reduce the number of false alarms. For our analysis, its value is set equal to 0.95. The Potential value is chosen to maximize the F1 score on the training set after applying the MRF method on the training database. Fig. 7 shows the calculated value of the normalized accuracy for several Potential values; it indicates that 0.45 is the optimal value according to F1 score.

4.3 Impacts of spatial and similarity features

The influence of each feature on the model can be directly determined with the feature importance yielded by the XGBoost model (Fig. 8). Because of evaluation with LOOCV, the final feature importance is averaged from each computed model.

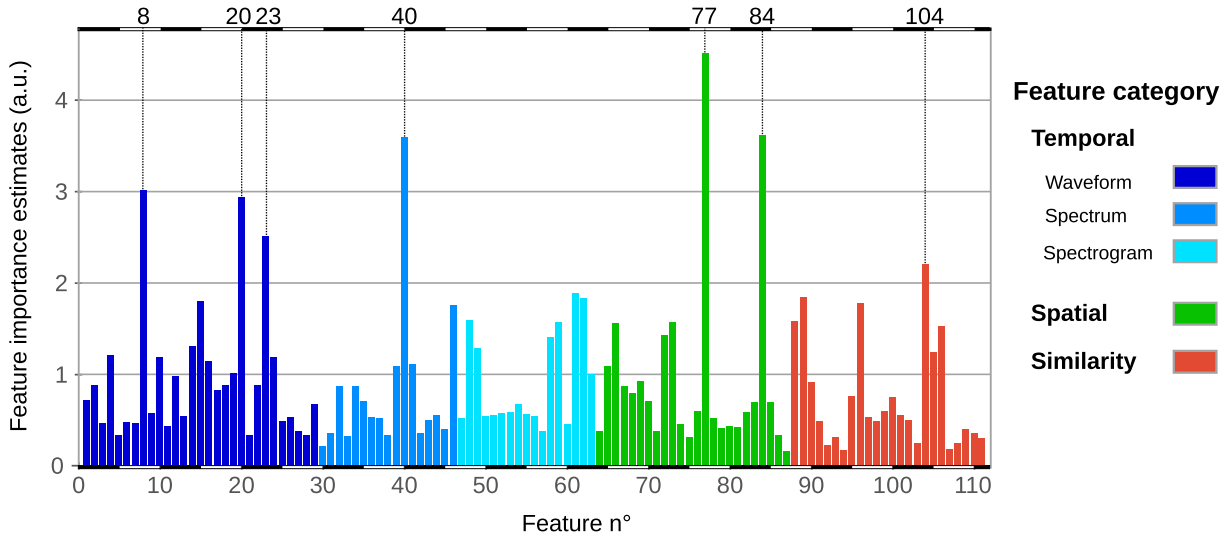

Figure 8. Feature importance evaluated with XGBoost algorithm. Features 1 to 63 are temporal features, 64 to 87 are spatial features and 88 to 111 are similarity features.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

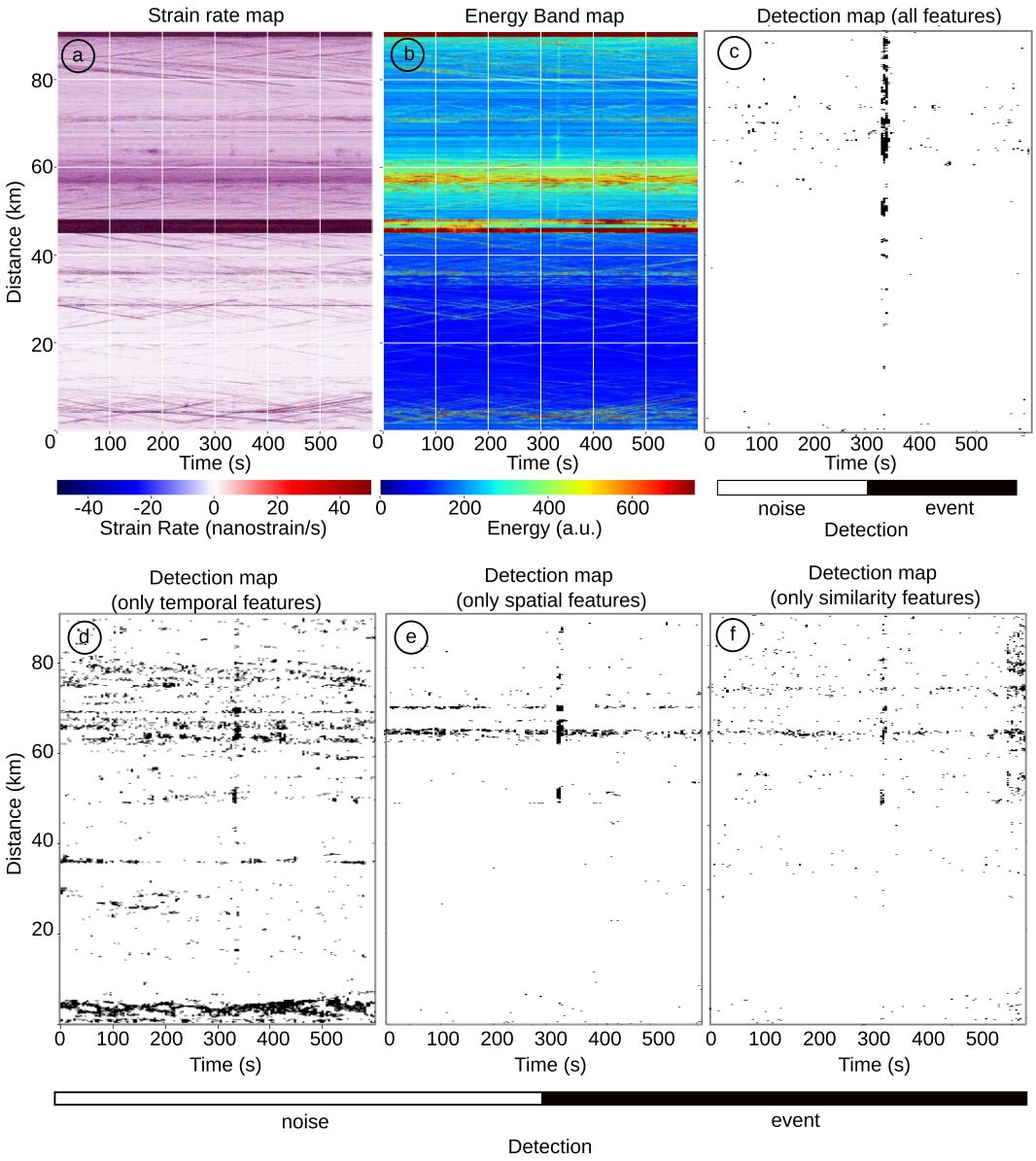

Figure 9. Detection map representation of an earthquake of magnitude $M_w = 1.1$ that happens close to Argelès-Gazost, on 2022 September 15 at 9:12:16 UTC. Detection map is obtained on the score map with a threshold of 0.95. (a) represents the SR, (b) represents the EB, (c) the detection map obtained with a classifier that use all features, (d) the detection map obtained with a classifier that is only trained on temporal features, (e) the detection map obtained with a classifier that is only trained on spatial features, (f) the detection map obtained with a classifier that only uses similarity features.

Features presented in Section 3.3 are arbitrarily numbered from 1 to 111, and are named with their categories (temporal, spatial or similarity) followed by their number detailed in table A1 in appendix B. For example, (Spatial77) refers to the feature n°77, which belongs to the spatial features category. Fig. 8 shows that the eight most important features are sorted in decreasing order of importance as follows: average of the raw spatial trace (Spatial77), energy filtered between $[0, \frac{1}{4}]$ of the Nyquist frequency band (Temporal40), number of peaks in autocorrelation function of spatial and temporal traces (Spatial84 and Temporal8), ratio between the energy in [10, 30] and [30, 50] Hz (Temporal20), Kurtosis of the trace filtered in [5,10] Hz (Temporal23), indexes of maximum of cross-correlation function computed for two consecutive stacks of temporal traces (Similarity104) and maximum of cross-correlation function computed for two consecutive stacks of temporal traces (Similarity89).

Considering the eight most important features, two are spatial features (Spatial77, 84), two are similarity features (Similarity104, 89) and four are temporal features (Temporal40, 8, 20, 23). The contribution of spatial and similarity features can be qualitatively observed in the classification map (Fig. 9) after using a score threshold of 0.95 on the output of XGBoost for each event. The choice is discussed in Section 4.1. Taking the example of an earthquake that occurred on 2022 September 15 at 9:12:16 UTC (Figs 9a and b), we observe that relying solely on temporal features is insufficient for achieving accurate classification results, as anthropogenic events generate a lot of false alarms in Fig. 9(d). Spatial and similarity features have a lower sensitivity to anthropogenic events as depicted

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

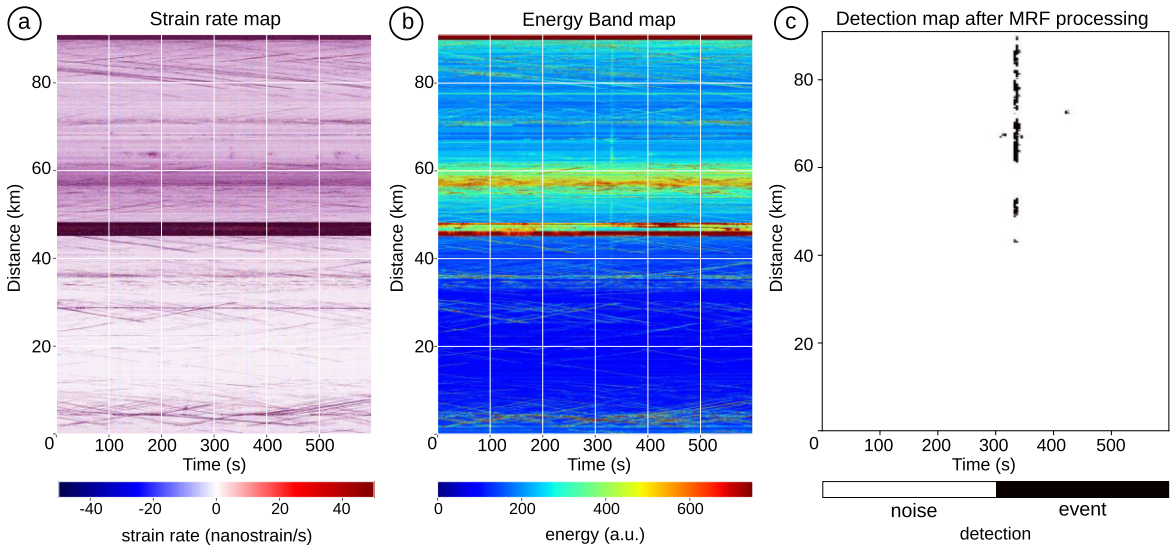

Figure 10. Detection map representation of an earthquake of magnitude $M_w = 1.1$ that happens close to Argelès-Gazost, on 2022 September 15 at 9:12:16 UTC. Detection map is obtained on the score map with application of MRF model, with a threshold of 0.95 and with use of MRF. (a) represents the SR, (b) the EB, (c) and the detection map.

in Figs 9(e) and (f). The combination of all the presented features results in a detection map with less false detections (Fig. 9c). Each detection map is obtained using a grid defined as (t_{win}, d_{win}, t_{step}, d_{step}) = (8 s, 1000 m, 4 s, 100 m), as discussed in Section 4.1. Appendix A provides the detection maps for the 19 catalogued events recorded during the study period.

4.4 Performance of the processing chain

Using the different window parameters and features defined in the previous sections, we compute the features for all the identified events, train various models with several training sets using LOOCV technique and post-process the results with score thresholding and MRF method.

Fig. 10 contains an example of the detection of an earthquake of magnitude $M_w = 1.1$ (Fig. 10c) using the SR map (Fig. 10a). We also plot the EB map (Fig. 10b) to help the reader identify the event location. The detection and identification maps resulting from our workflow, along with the SR and the EB maps, are presented for each event recorded during the study period in Appendix A.

The XGBoost algorithm combined with the proposed spatial and similarity features is able to correctly detect an earthquake of magnitude $M_w = 0.4$ at a maximum distance of 1.5 km (EQ8 in Table 1) and quarry blasts of magnitude $M_w = 1.1$ at a maximum distance of 1.0 km (QB4 in Table 1). Detection is achieved despite the surrounding anthropogenic noise, and quantified using the LOOCV splitting for which each event is processed by a model trained with the other events from our database. Amongst the 19 events visible on the DAS data, 13 of the 13 earthquakes are detected, as well as 3 of the 6 quarry blasts. One of the detected earthquakes is close to non-detection and corresponds to an event of magnitude $M_w = 0.8$ located at 1.0 km from the closest point along the fibre (EQ12 in Table 1). Undetected quarry blasts are located at distances higher than 4 km from the closest point sensor of the fibre (QB2, QB3 and QB6 in Table 1). Compared to the detected ones, they present lower energy and are difficult to detect above the noise.

5 DISCUSSION AND CONCLUSION

This paper introduces novel DAS-oriented features—spatial and similarity features—complementing the temporal features established in Huynh *et al.* (2022). These features enhance event detection and identification within a comprehensive processing chain. Our findings confirm that the integration of these new features significantly improves classification outcomes, even under field conditions with a standard fibre optic telecommunication cable spanning nearly 100 km. The promising results yielded by our processing chain despite the presence of environmental and anthropogenic noises are showing the relevance of AI based processing chains for the detection of seismic events of interest with the DAS acquisition system for the monitoring of natural seismic sources at regional scales. We achieved effective earthquake detection, highlighting DAS capability to identify such events even in challenging conditions. Quarry blasts were accurately detected when located within 1.8 km of the fibre optic network, showcasing the method utility in near-field applications.

However, several limitations must be acknowledged. The relatively small data set and the scarcity of natural seismic events constrained our analysis, impacting the interpretability of the results. The reliance on a supervised machine learning algorithm necessitated accurate data labelling, which was challenging given the potential mislabelling of low-magnitude events. Consequently, we opted to train our algorithm using only 10-min segments of DAS data associated with catalogued events, as opposed to utilizing the full three weeks of continuous data.

To address these limitations, future research could explore semisupervised or self-supervised machine learning approaches, which would leverage the extensive amount of data collected during the measurement campaign (Wang *et al.* 2022; Zhu *et al.* 2023; Rimpot *et al.* 2024). Additionally, extending the acquisition campaign or employing data sets from other locations could enhance the data set size and diversity. Another promising avenue is to utilize the existing seismic knowledge from the six permanent seismometers in the region to inform DAS monitoring through transfer learning (Titos *et al.* 2019; Lapins *et al.* 2021; Donnadille *et al.* 2024).

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

The geometric configuration of the fibre optic network and its varying signal-to-noise ratio (SNR) based on position also influenced our results. The non-proportional increase in distance between event sources and fibre points affects the spatial and similarity features measured. Hence, incorporating the geographic distribution of events through localization when possible is crucial for minimizing the influence of network geometry on model performance in different study areas.

The ability to detect very low magnitude earthquakes, starting from $M_w = 0.4$ at a distance of 1.5 km from the fibre, and quarry blasts within 4 km, underscores the high sensitivity and efficacy of DAS for event monitoring. This high sensitivity, coupled with the distributed nature of the sensor network, establishes DAS as an excellent solution for long-term monitoring of natural processes over regional scales with unprecedented spatial resolution. Future research should address data set limitations, explore semi- or self-supervised and transfer learning methods, and consider the implications of fibre geometry on detection performance.

ACKNOWLEDGMENTS

This work has been supported by the Institut Terre et Environnement de Strasbourg (CNRS-ITES) and FEBUS Optics as part of a CIFRE-ANRT research contract N°2021/0001. The authors thank Total Energies for being partner of this project, and Gaëtan Calbris (FEBUS Optics) for his expertise on the FEBUS A1-R system during the measurement period. The FO-DAS dataset processing has been carried out on the EOST-A2S HPC facility of University of Strasbourg, part of the Data-Terra Research Infrastructure. The authors would like to acknowledge David Michéa for supporting this work by providing technical support and access to the computing resources. Part of the computing resources were funded by the Equipex+ GAIA-DATA project (Programme Investissements d'Avenir) and the CPER A2S (Application Satellite Survey). The authors also thank the two anonymous reviewers and editors who have helped to improve the clarity and the readability of the manuscript.

DATA AVAILABILITY

The strain rate data used in this study is available at https://doi.org/10.57932/5b1302d6-57cd-44e4-81ac-5d585a7f8951 (Huynh *et al*. 2024), and the machine learning code on gitlab at: https://gitlab.com/eost/seis-learning-spatial/. The machine learning code includes the feature computation code, the model training code and the classification map builder code.

REFERENCES

Atterholt, J., Zhan, Z., Shen, Z. & Li, Z., 2022. A unified wavefield-partitioning approach for distributed acoustic sensing, *Geophys. J. Int.,* **228**(2), 1410–1418.

Berndt, D.J. & Clifford, J., 1994. Using dynamic time warping to find patterns in time series, in *Proceedings of the 3rd international conference on knowledge discovery and data mining,* AAAI Press, Seattle WA, pp. 359–370.

Binder, G. & Tura, A., 2020. Convolutional neural networks for automated microseismic detection in downhole distributed acoustic sensing data and comparison to a surface geophone array, *Geophys. Prospect.,* **68**(9), 2770–2782.

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

Bublin, M., 2021. Event detection for distributed acoustic sensing: combining knowledge-based, classical machine learning, and deep learning approaches, *Sensors,* **21**(22), 7527, doi:10.3390/s21227527.

Chen, T. & Guestrin, C., 2016. XGBoost: a scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 785–794, ACM, San Francisco California USA.

Chen, Y. *et al.*, 2023. Denoising of distributed acoustic sensing seismic data using an integrated framework, *Seismol. Soc. Am.,* **94**(1), 457–472.

Chmiel, M., Walter, F., Wenner, M., Zhang, Z., McArdell, B.W. & Hibert, C., 2021. Machine learning improves debris flow warning, *Geophys. Res. Lett.,* **48**(3), doi:10.1029/2020GL090874.

Cross, G.R. & Jain, A.K., 1983. Markov random field texture models, *IEEE Trans. Pattern Anal. Mach. Intell.,* **PAMI-5**(1), 25–39.

Domel, P., Hibert, C., Schlindwein, V. & Plaza-Faverola, A., 2023. Event recognition in marine seismological data using Random Forest machine learning classifier, *Geophys. J. Int.,* **235**(1), 589–609.

Donnadille, M., Turquet, A., Hibert, C. & Richard, C., 2024. *Distributed Acoustic Sensing Automated Classifiers Design via Transfer Learning for Seismology, Tech. rep., Copernicus Meetings,* European Geosciences Union General Assembly 2024.

Ende, M. V.D., Lior, I., Ampuero, J.-P., Sladen, A., Ferrari, A. & Richard, C., 2021. *A Self-Supervised Deep Learning Approach for Blind Denoising and Waveform Coherence Enhancement in Distributed Acoustic Sensing data,* IEEE, **34,** 3371–3384, *IEEE Transactions on Neural Networks and Learning Systems*.

Falcin, A. *et al.*, 2021. 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.

Fernandes, R.M.S., Miranda, J.M., Meijninger, B.M.L., Bos, M.S., Noomen, R., Bastos, L., Ambrosius, B.A.C. & Riva, R.E.M., 2007. Surface velocity field of the Ibero-Maghrebian segment of the Eurasia-Nubia plate boundary, *Geophys. J. Int.,* **169**(1), 315–324.

Fukushima, S., Shinohara, M., Nishida, K., Takeo, A., Yamada, T. & Yomogida, K., 2022. Detailed S-wave velocity structure of sediment and crust off Sanriku, Japan by a new analysis method for distributed acoustic sensing data using a seafloor cable and seismic interferometry, *Earth Planets Space,* **74**(1), 92, doi:10.1186/s40623-022-01652-z.

Hartog, A.H., 2017. *An Introduction to Distributed Optical Fibre Sensors, Series in Fiber Optic Sensors,* CRC Press, Taylor & Francis Group,

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

Hibert, C., Michea, D., Provost, F., Malet, J.-P. & Geertsema, M., 2019. Exploration of continuous seismic recordings with a machine learning approach to document 20 yr of landslide activity in Alaska, *Geophys. J. Int.,* **219**(2), 1138–1147.

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.

Hudson, T.S. *et al.*, 2021. Distributed acoustic sensing (DAS) for natural microseismicity studies: a case study from Antarctica, *J. geophys. Res.: Solid Earth,* **126**(7), e2020JB021493, doi:10.1029/2020JB021493.

Huynh, C. *et al.*, 2024. DAS-BIGORRE-2022: Fiber-Optics Distributed Acoustic Sensing records (Bigorre, Hautes-Pyrénées, France), *EaSy Data,* doi:10.57932/5b1302d6-57cd-44e4-81ac-5d585a7f8951.

Huynh, C., Hibert, C., Jestin, C., Malet, J.-P., Clement, P. & Lanticq, V., 2022. Real-time classification of anthropogenic seismic sources from distributed acoustic sensing data: application for pipeline monitoring, *Seismol. Res. Lett.,* **93**(5), 2570–2583.

Ida, Y., Fujita, E. & Hirose, T., 2022. Classification of volcano-seismic events using waveforms in the method of k-means clustering and dynamic time warping, *J. Volc. Geotherm. Res.,* **429,** 107616, doi:10.1016/j.jvolgeores.2022.107616.

Jakkampudi, S., Shen, J., Li, W., Dev, A., Zhu, T. & Martin, E.R., 2020. Footstep detection in urban seismic data with a convolutional neural network, *Leading Edge,* **39** 654–660.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

Jousset, P. *et al.*, 2022. Fibre optic distributed acoustic sensing of volcanic events, *Nat. Commun.,* **13**(1), 1753.
Keogh, E. & Ratanamahatana, C.A., 2005. Exact indexing of dynamic time warping, *Knowl. Inf. Syst.,* **7**(3), 358–386.
Kumar, U., Legendre, C.P., Lee, J.-C., Zhao, L. & Chao, B.F., 2022a. On analyzing GNSS displacement field variability of Taiwan: hierarchical agglomerative clustering based on dynamic time warping technique, *Comput. Geosci.,* **169,** 105243, doi:10.1016/j.cageo.2022.105243.
Kumar, U., Legendre, C.P., Zhao, L. & Chao, B.F., 2022b. Dynamic time warping as an alternative to windowed cross correlation in seismological applications, *Seismol. Soc. Am.,* **93**(3), 1909–1921.
Lacan, P. & Ortuño, M., 2012. Active Tectonics of the Pyrenees: a review, *J. Iberian Geol.,* **38**(1), 9–30.
Lapins, S., Goitom, B., Kendall, J.-M., Werner, M.J., Cashman, K.V. & Hammond, J.O., 2021. A little data goes a long way: automating seismic phase arrival picking at Nabro volcano with transfer learning, *J. geophys. Res.: Solid Earth,* **126**(7), e2021JB021910, doi:10.1029/2021JB021910.
Lindsey, N.J., Martin, E.R., Dreger, D.S., Freifeld, B., Cole, S., James, S.R., Biondi, B.L. & Ajo-Franklin, J.B., 2017. Fiber-optic network observations of earthquake wavefields, *Geophys. Res. Lett.,* **44**(23), doi:10.1002/2017GL075722.
Ma, Y. & Hale, D., 2013. Wave-equation reflection traveltime inversion with dynamic warping and full-waveform inversion, *Geophysics,* **78**(6), R223–R233.
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**(3), 878–891.
Malfante, M., Dalla Mura, M., Metaxian, J.-P., Mars, J.I., Macedo, O. & Inza, A., 2018. Machine learning for volcano-seismic signals: challenges and perspectives, *IEEE Signal Process. Mag.,* **35**(2), 20–30.
Müller, M., 2007. Dynamic time warping, in *Information Retrieval for Music and Motion,* pp. 69–84, Springer, Berlin, Heidelberg.
Nayak, A., Ajo-Franklin, J. *et al.*, 2021. Distributed acoustic sensing using dark fiber for array detection of regional earthquakes, *Seismol. Res. Lett.,* **92**(4), 2441–2452.
Provost, F., Hibert, C. & Malet, J.-P., 2017. Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier: seismic sources automatic classification, *Geophys. Res. Lett.,* **44**(1), 113–120.
Rigo, A., Souriau, A., Dubos, N., Sylvander, M. & Ponsolles, C., 2005. Analysis of the seismicity in the central part of the Pyrenees (France), and tectonic implications, *J. Seismol.,* **9**(2), 211–222.
Rimpot, J., Hibert, C., Malet, J.-P., Forestier, G. & Weber, J., 2024. *Self-supervised learning strategies for clustering continuous seismic data,* European Geosciences Union General Assembly 2024, Vienna, Austria.
Sakoe, H. & Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. Acoustics Speech Signal Process.,* **26**(1), 43–49.
Sladen, A., Rivet, D., Ampuero, J.P., De Barros, L., Hello, Y., Calbris, G. & Lamare, P., 2019. Distributed sensing of earthquakes and ocean-solid Earth interactions on seafloor telecom cables, *Nat. Commun.,* **10**(1), 5777, doi:10.1038/s41467-019-13793-z.
Souriau, A. & Pauchet, H., 1998. A new synthesis of Pyrenean seismicity and its tectonic implications, *Tectonophysics,* **290**(3-4), 221–244.
Spica, Z.J., Castellanos, J.C., Viens, L., Nishida, K., Akuhara, T., Shinohara, M. & Yamada, T., 2022. Subsurface imaging with ocean-bottom distributed acoustic sensing and water phases reverberations, *Geophys. Res. Lett.,* **49**(2), doi:10.1029/2021GL095287.
Sylvander, M. *et al.*, 2022. Seismicity patterns in southwestern France, *C. R. Geosci.,* **353**(S1), 79–104.
Tejedor, J., Macias-Guarasa, J., Martins, H.F., Martin-Lopez, S. & Gonzalez-Herraez, M., 2021. A multi-position approach in a smart fiber-optic surveillance system for pipeline integrity threat detection, *Electronics,* **10**(6), 712, doi:10.3390/electronics10060712.
Titos, M., Bueno, A., García, L., Benítez, C. & Segura, J.C., 2019. Classification of isolated volcano-seismic events based on inductive transfer learning, *IEEE Geosci. Remote Sens. Lett.,* **17**(5), 869–873.
Wang, S., Liu, F. & Liu, B., 2022. Semi-supervised deep learning in high-speed railway track detection based on distributed fiber acoustic sensing, *Sensors,* **22**(2), 413, doi:10.3390/s22020413.
Wang, T., Bian, Y., Zhang, Y. & Hou, X., 2023. Classification of earthquakes, explosions and mining-induced earthquakes based on XGBoost algorithm, *Comput. Geosci.,* **170,** 105242, doi:10.1016/j.cageo.2022.105242.
Wang, Y. *et al.*, 2019. Pattern recognition using relevant vector machine in optical fiber vibration sensing system, *IEEE Access,* **7,** 5886–5895.
Wenner, M., Hibert, C., van Herwijnen, A., Meier, L. & Walter, F., 2021. Near-real-time automated classification of seismic signals of slope failures with continuous random forests, *Nat. Hazards Earth Syst. Sci.,* **21**(1), 339–361.
Wiesmeyr, C., Litzenberger, M., Waser, M., Papp, A., Garn, H., Neunteufel, G. & Döller, H., 2020. Real-time train tracking from distributed acoustic sensing data, *Appl. Sci.,* **10**(2), 448, doi:10.3390/app10020448.
Young, C., Shragge, J., Schultz, W., Haines, S., Oren, C., Simmons, J. & Collett, T.S., 2022. Advanced distributed acoustic sensing vertical seismic profile imaging of an Alaska North Slope Gas Hydrate Field, *Energy Fuels,* **36**(7), 3481–3495.
Zeng, X., Bao, F., Thurber, C.H., Lin, R., Wang, S., Song, Z. & Han, L., 2022. Turning a telecom fiber-optic cable into an ultradense seismic array for rapid postearthquake response in an urban area, *Seismol. Res. Lett.,* **93**(2A), 853–865.
Zhang, L. & Zhan, C., 2017. Machine learning in rock facies classification: an application of XGBoost, in *International Geophysical Conference, Qingdao, China, 17-20 April 2017,* pp. 1371–1374, Society of Exploration Geophysicists and Chinese Petroleum Society.
Zhao, Y., Li, Y. & Wu, N., 2021. Distributed acoustic sensing vertical seismic profile data denoiser based on convolutional neural network, *IEEE Trans. Geosci. Remote Sens.,* **60,** 1–11.
Zhong, R., Johnson, R., Jr & Chen, Z., 2020. Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost), *Int. J. Coal Geol.,* **220,** 103416.
Zhu, T., Shen, J. & Martin, E.R., 2021. Sensing Earth and environment dynamics by telecommunication fiber-optic sensors: an urban experiment in Pennsylvania, USA, *Solid Earth,* **12**(1), 219–235.
Zhu, W., Biondi, E., Li, J., Yin, J., Ross, Z.E. & Zhan, Z., 2023. Seismic arrival-time picking on distributed acoustic sensing data using semi-supervised learning, *Nat. Commun.,* **14**(1), 8192, doi:10.1038/s41467-023-43355-3.

APPENDIX A: THE NINETEEN IDENTIFIED EVENTS ON DAS RECORDING

Appendix A gives an overview of all the events recorded by BCSF-RENASS and visually inspected on the DAS recordings. Fig. A1 presents the SR, EB, score map of the machine learning algorithm and detection map obtained using our processing chain for each event.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

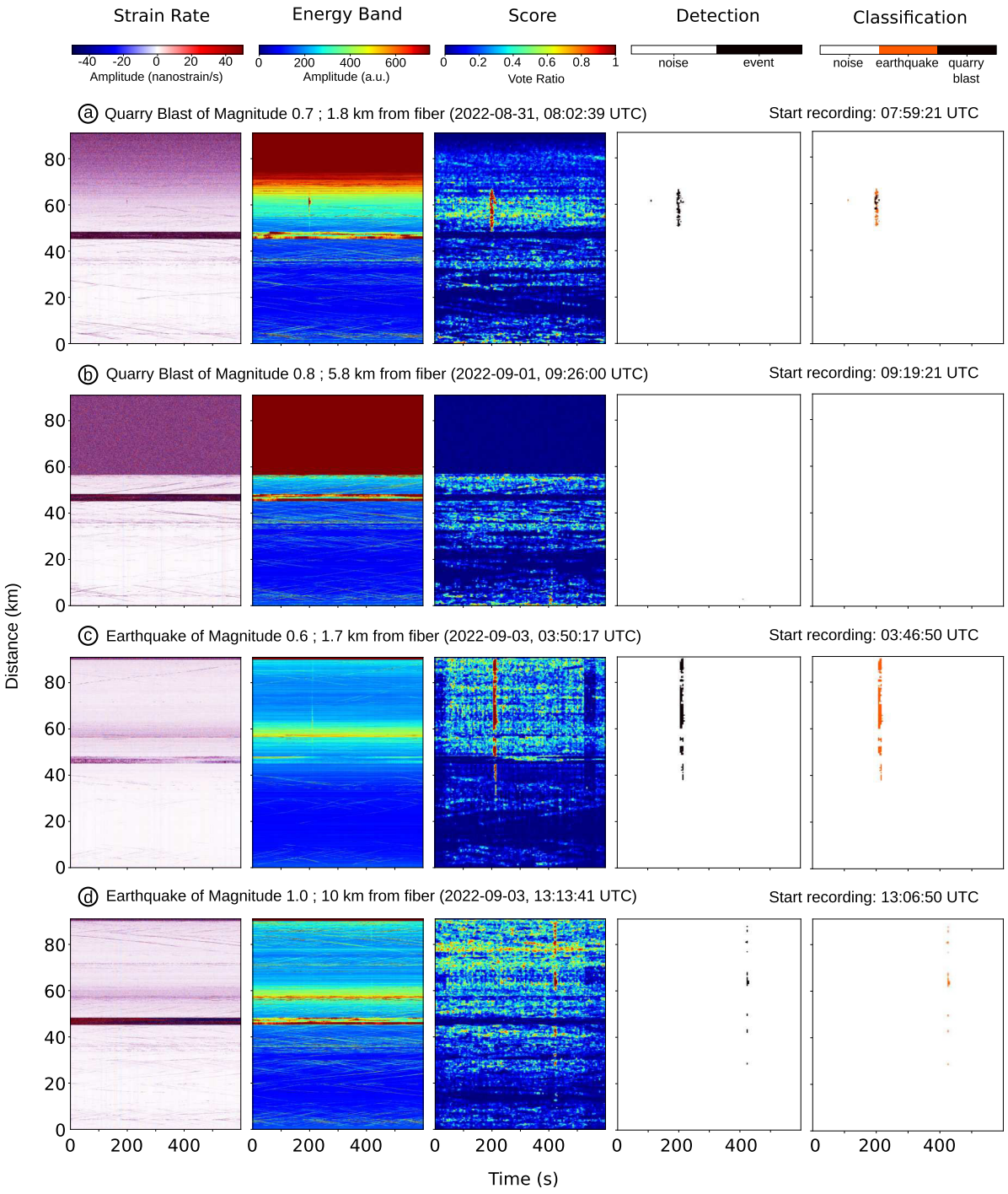

Figure A1. Strain rate, energy band, score map and detection map of the 19 recorded events along the fibre. Strain rate and energy band are computed directly from acquired data, whereas score map and detection map are obtained using our classification processing chain.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

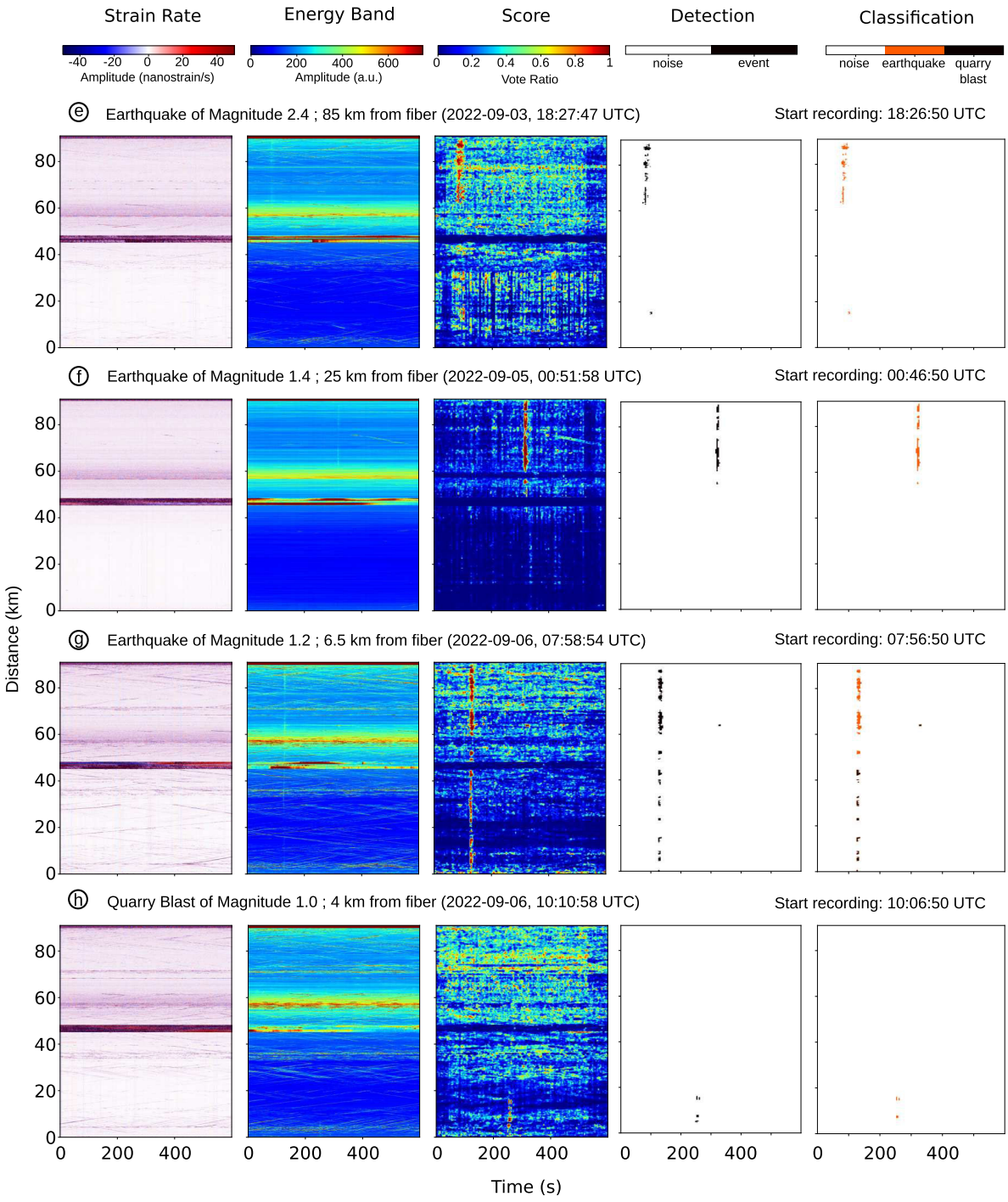

Figure A1. (continued) strain rate, energy band, score map and detection map of the 19 recorded events along the fibre. Strain rate and energy band are computed directly from acquired data, whereas score map and detection map are obtained using our classification processing chain.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

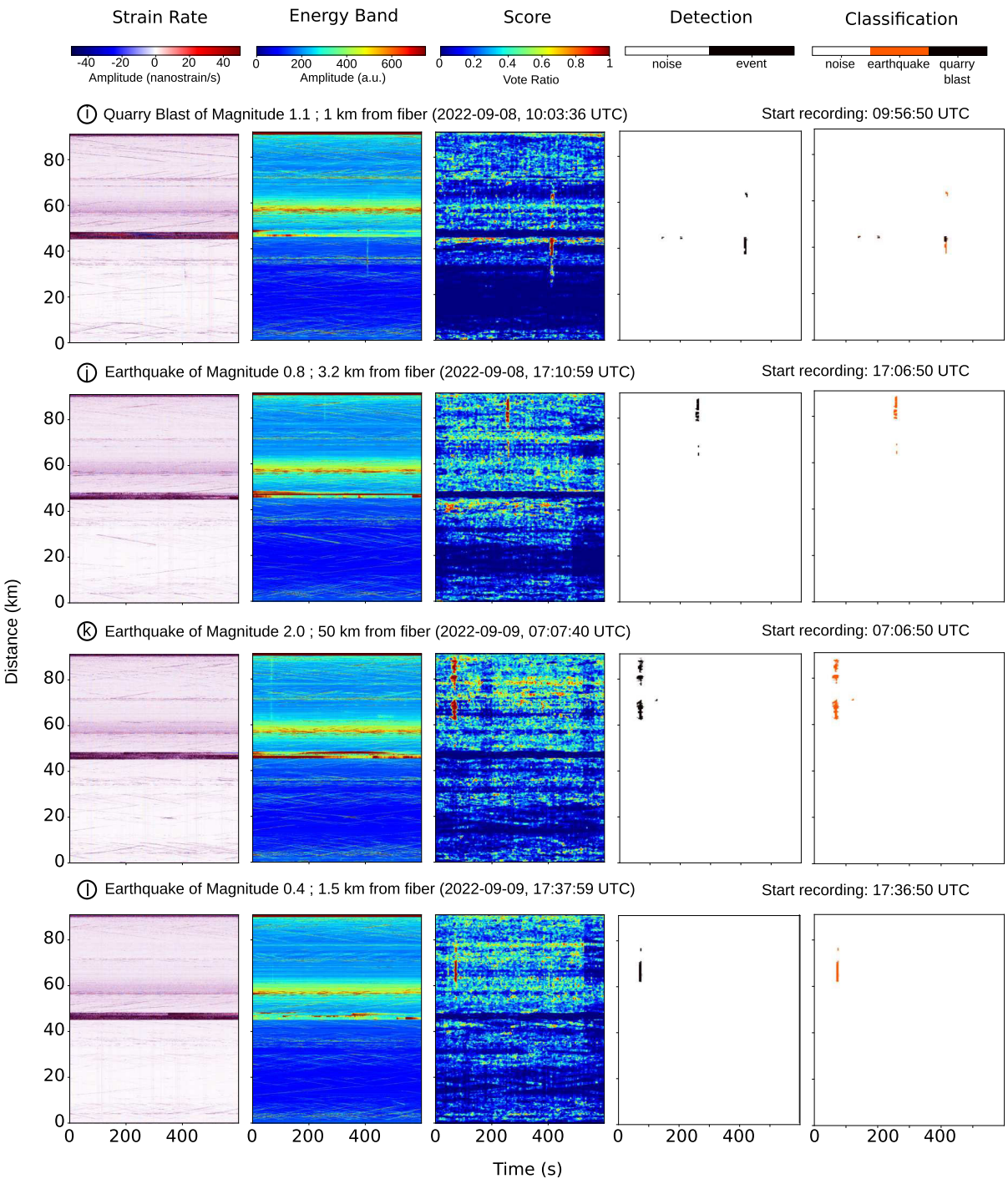

Figure A1. (continued) Strain rate, energy band, score map and detection map of the 19 recorded events along the fibre. Strain rate and energy band are computed directly from acquired data, whereas score map and detection map are obtained using our classification processing chain.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

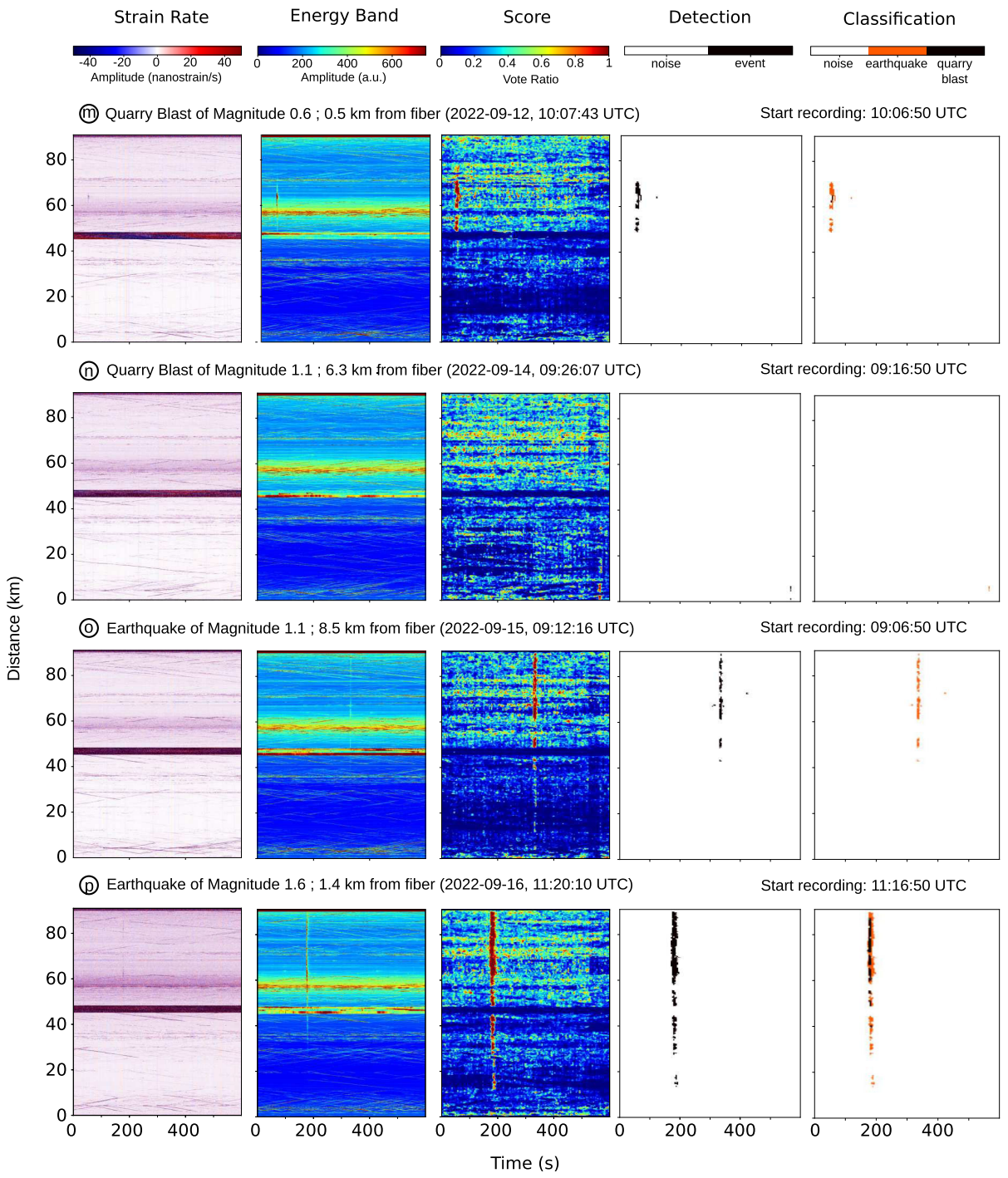

Figure A1. (continued) Strain rate, energy band, score map and detection map of the 19 recorded events along the fibre. Strain rate and energy band are computed directly from acquired data, whereas score map and detection map are obtained using our classification processing chain.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

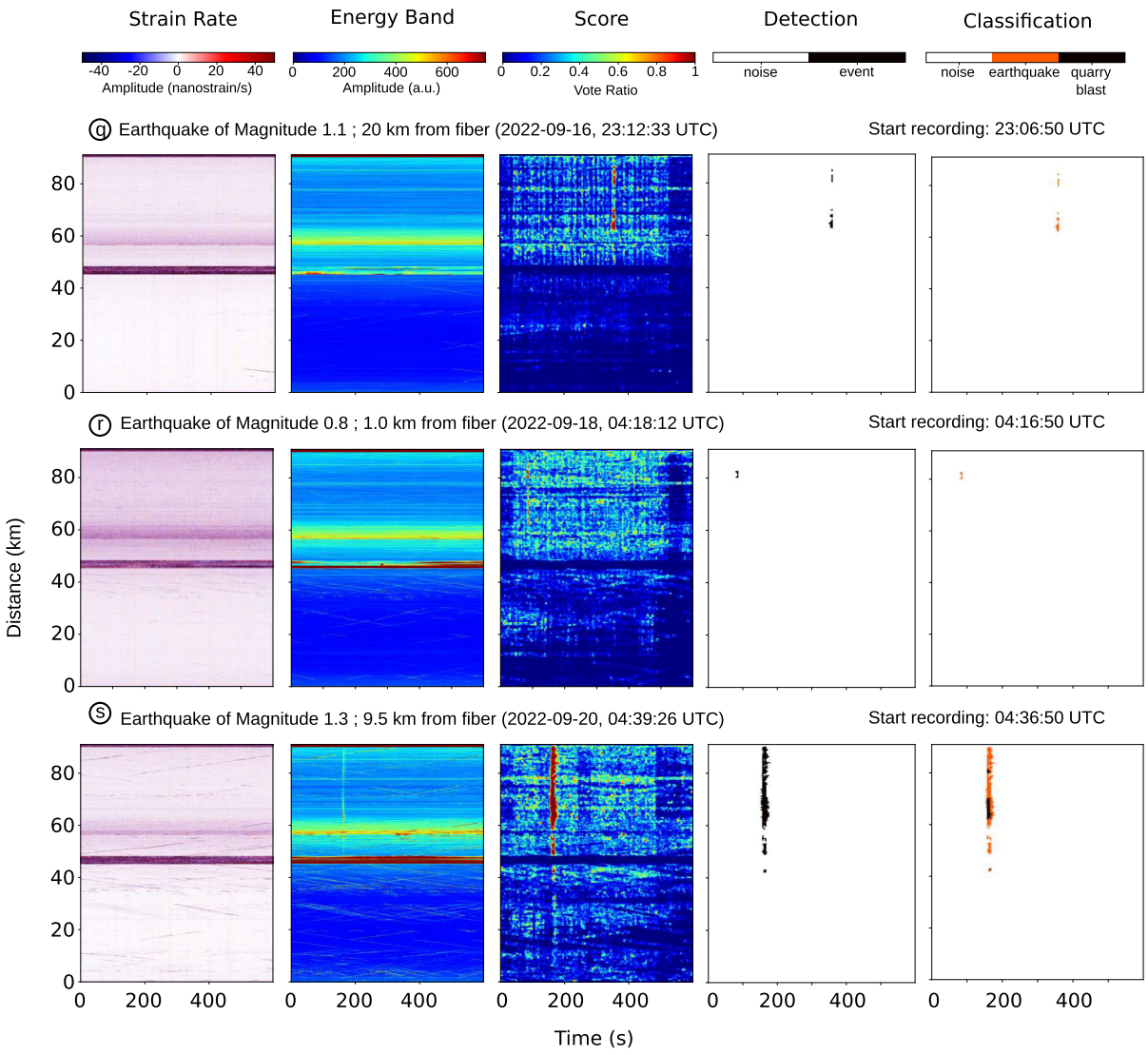

Figure A1. (continued) Strain rate, energy band, score map and detection map of the 19 recorded events along the fibre. Strain rate and energy band are computed directly from acquired data, whereas score map and detection map are obtained using our classification processing chain.

APPENDIX B: FEATURES USED TO DESCRIBE THE SR SIGNAL

Appendix B gives an overview of all the features used for the machine learning process. Table B1 gathers the temporal features related to waveform, the temporal features related to spectrum, the temporal features related to spectrogram, the spatial features and the similarity features. Spatial and similarity features were developed in this study specifically for DAS recordings.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

Table B1. Features used to describe the SR signal.

Number	Description	Formula
	Temporal features—Waveform	
Averaging in window		
1	Ratio of the mean over the maximum of the envelop signal	–
2	Ratio of the median over the maximum of the envelop signal	–
3	Ratio between ascending and descending time	$\frac{t_{\max}-t_i}{t_f-t_{\max}}$ with $t_{\max}$: time of the largest amplitude
4	Kurtosis of the raw signal (peakness of the signal)	$\frac{m_4}{\theta^4}$ with m_4: fourth moment, θ: standard deviation
5	Kurtosis of the envelop	see 4
6	Skewness of the raw signal	$\frac{m_3}{\theta^3}$ with m_3: third moment, θ: standard deviation
7	Skewness of the envelop	see 6
8	Number of peaks in the autocorrelation function	–
9	Energy in the first third part of the autocorrelation function	$\int_0^{T/3} C(\tau)d\tau$, with T: signal duration, C: autocorrelation function
10	Energy in the remaining part of the autocorrelation function	see 9
11	Ratio of 10 and 9	–
12–16	Energy of the signal filtered in 5–10 Hz, 10–30 Hz, 30–50 Hz, 50–75 Hz, 75–100 Hz	$\int_0^T y_f(t)dt$, with y_f: filtered signal in the frequency range $[f_1, f_2]$
17–22	Ratio between (12,13), (12,14), (12,15), (13,14), (13,15), (14,15)	–
23–27	Kurtosis of the signal filtered in 5–10 Hz, 10–30 Hz, 30–50 Hz, 50–75 Hz, 75–100 Hz	see 4
28	RMS between the decreasing part of the signal and $I(t) = Y_{\max} - \frac{Y_{\max}}{t_f-t_{\max}}t$	$\sqrt{\overline{Y(t)-I(t)^2}}$, with Y: envelop of the signal spectral features
29	Maximum of envelope	–
	Temporal features—Spectral	
Averaging in window		
30	Mean of the DFT	DFT: discrete Fourier transform
31	Max of the DFT	–
32	Frequency at the maximum of the DFT	–
33	Frequency of spectrum centroid	–
34	Central frequency of the first quartile	–
35	Central frequency of the third quartile	–
36	Median of the normalized DFT	–
37	Variance of the normalized DFT	–
38	Number of peaks ($> 0.75\text{DFT}_{\max}$)	$\text{DFT}_{\max}$: maximum of the DFT
39	Mean peaks value for peaks > 0.7	–
40–43	Energy in $[0\,\frac{1}{4}]$ NyF, $[\frac{1}{4}\,\frac{1}{2}]$ NyF, $[\frac{1}{2}\,\frac{3}{4}]$ NyF, $[\frac{3}{4}\,1]$ NyF	$\int_{f_1}^{f_2} \text{DFT}(f)df$, with f_1, f_2: considered frequency ranges
44	Spectral centroid	$\gamma_1 = \frac{m_2}{m_1}$, with m_1, m_2: first and second moment
45	Gyration radius	$\gamma_2 = \sqrt{\frac{m_3}{m_2}}$, with m_3: third moment
46	Spectral centroid width	$\gamma_3 = \sqrt{\gamma_1^2 + \gamma_2^2}$
	Temporal features—Spectrogram	
Averaging in window		
47	Kurtosis of the maximum of all discrete Fourier transforms (DFTs)	kurtosis[$\max_{t\in[0,T]}$(SPEC(t, f))], with SPEC(t, f): spectrogram as a function of time t
48	Kurtosis of the maximum of all DFTs as a function of time t	see 47
49	Mean ratio between the maximum and the mean of all DFTs	mean($\frac{\max(\text{SPEC})}{\text{mean}(\text{SPEC})}$)
50	Mean ratio between the maximum and the median of all DFTs	see 49
51	Number of peaks in the curve showing the temporal evolution of the DFTs maximum	–
52	Number of peaks in the curve showing the temporal evolution of the DFTs mean	–
53	Number of peaks in the curve showing the temporal evolution of the DFTs median	–
54	Ratio between 51 and 52	–
55	Ratio between 51 and 53	–
56	Number of peaks in the curve of the temporal evolution of the DFTs central frequency	–
57	Number of peaks in the curve of the temporal evolution of the DFTs maximum frequency	–
58	Ratio between 56 and 57	–
59	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and mean frequency	–
60	Mean distance between the curves of the temporal evolution of the DFTs maximum frequency and median frequency	–

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

Table B1. Continued

Number	Description	Formula
61	Mean distance between the 1st quartile and the median of all DFTs as a function of time	–
62	Mean distance between the 3rd quartile and the median of all DFTs as a function of time	–
63	Mean distance between the 3rd quartile and the 1st quartile of all DFTs as a function of time	–
	Spatial features	
Averaging in window		
64	Mean of the envelope of spatial trace	–
65	Mean of the raw spatial trace	–
66	Standard deviation of the envelope of spatial trace	–
67	Standard deviation of the raw spatial trace	–
68	Kurtosis of the envelope of spatial trace	see 4
69	Kurtosis of the raw spatial trace	see 4
70	Skewness of the envelope of spatial trace	see 6
71	Skewness of the raw spatial trace	see 6
72	Number of peaks of the autocorrelation function of spatial trace	–
73	Energy in the 1/3 around the origin of the autocorrelation function of spatial trace	see 9
74	Energy in the last 2/3 of the autocorrelation function of spatial trace	see 9
75	Ratio of 74 and 73	–
Standard deviation in window		
76	Mean of the envelope of spatial trace	–
77	Mean of the raw spatial trace	–
78	Standard deviation of the envelope of spatial trace	–
79	Standard deviation of the raw spatial trace	–
80	Kurtosis of the envelope of spatial trace	see 4
81	Kurtosis of the raw spatial trace	see 4
82	Skewness of the envelope of spatial trace	see 6
83	Skewness of the raw spatial trace	see 6
84	Number of peaks of the autcorrelation function of spatial trace	–
85	Energy in the 1/3 around the origin of the autocorr function of spatial trace	see 9
86	Energy in the last 2/3 of the autocorr function of spatial trace	see 9
87	Ratio of 86 and 85	–
	Similarity features	
Averaging in window		
88	Index of Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	–
89	Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	–
90	Similarity measurement of two consecutive stacks of temporal trace provided from DTW function	–
91	Integral of the DTW temporal distortion function*	–
92	Energy in the first third part of the autocorrelation function of the DTW temporal distortion function*	–
93	Energy in the remaining part of the autocorrelation function of the DTW temporal distortion function*	–
94	Ratio of 93 and 92	–
95	Difference of 93 and 92	–
Standard deviation in window		
96	Index of Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	–
97	Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	–
98	Similarity measurement of two consecutive stacks of temporal trace provided from DTW function	–
99	Integral of the DTW temporal distortion function*	–
100	Energy in the first third part of the autocorrelation function of the DTW temporal distortion function*	–
101	Energy in the remaining part of the autocorrelation function of the DTW temporal distortion function*	–
102	Ratio of 101 and 100	–

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

Table B1. Continued

Number	Description	Formula
103	Difference of 101 and 100	–
Median in window		
104	Index of Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	–
105	Maximum of cross-correlation function computed for two consecutive stacks of temporal trace	–
106	Similarity measurement of two consecutive stacks of temporal trace provided from DTW function	–
107	Integral of the DTW temporal distortion function*	–
108	Energy in the first third part of the autocorrelation function of the DTW temporal distortion function*	–
109	Energy in the remaining part of the autocorrelation function of the DTW temporal distortion function*	–
110	Ratio of 109 and 108	–
111	Difference of 109 and 108	–

*DTW temporal distortion function corresponds to the estimated stretch to apply to points of a trace A to obtain a trace closest to trace B.

Downloaded from https://academic.oup.com/gji/article/240/1/462/7833573 by guest on 04 April 2025

© The Author(s) 2024. Published by Oxford University Press on behalf of The Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.