Supporting Information for "Machine Learning improves warning systems of debris flows"

Małgorzata Chmiel[1], Fabian Walter[1], Michaela Wenner[1,2], Zhen Zhang[1,3,4],

Brian W. McArdell[2], Clement Hibert[5]

[1]Laboratory of Hydraulics, Hydrology and Glaciology, ETH Zürich, Zürich, Switzerland

[2]Swiss Federal Institute for Forest, Snow and Landscape Research, Zürich, Switzerland

[3]Key Laboratory of Mountain Hazards and Surface Process, Institute of Mountain Hazards and Environment, Chinese Academy of Sciences, Chengdu, China

[4]University of Chinese Academy of Sciences, Beijing, China

[5]Institut de Physique du Globe de Strasbourg, CNRS UMR 7516, University of Strasbourg/EOST, 7 8 Strasbourg, France

Contents of this file

Additional Supporting Information (Files uploaded separately)

1. Captions for large Table S2.

Introduction

This file contains supplement text, tables, and figures for the manuscript "Machine Learning improves warning systems of debris flows". Supplemental Text1 provides a de-

tailed description of methods used for the development of the machine-learning detector. Supplemental Text2 provides additional information on the apparent total impact force spectrum (AIFS) calculation. Supplemental Figures S1 to S22 show vertical-component seismograms of 22 debris flow events recorded in 2017, 2018, and 2019. Supplemental Figure S23 shows debris-flow seismograms of 21 June 2019 event, and corresponding spectrograms. Supplemental Figure S24 shows machine-learning model evaluation for the first iteration. Supplemental Figure S25 shows receiver operating characteristic (ROC) curve analysis for the first and the second iteration. Supplemental Figure S26 shows the results of the detector run (from the second iteration) over 2017 and 2018 continuous data. Supplemental Figure S27 shows seismograms of a small debris flow event newly caught by the detector. Supplemental Figure S28 shows four photos of Illgraben debris-flow events with boulder-rich fronts detected in 2020. Supplemental Figure S29 shows results of a sensitivity test (accuracy of prediction vs number of debris-flow events used in a training set). Supplemental Figure S30 shows vertical apparent total impact force spectra (AIFS) for 2020 debris flows.

Supplemental Table S1 provides detailed information on the characteristics of 22 debris flows recorded in 2017, 2018, and 2019. Supplemental Table S2 (uploaded separately) shows 70 statistical features used as input in the machine-learning model. Supplemental Table S3 provides information on the characteristics of 13 debris flow events detected in 2020.

Text S1. Description of methods used in the development of the machine-learning debris flow detector.

Three classes of seismic signals are defined: pre-Check Dam 1 (CD1), post-CD1, and noise. We divide the dataset into 100 s time windows with an overlap of 50 %. We find that 100 s time window is long enough to contain enough of information to reliably extract signal statistics, and it is short enough to create a large catalog needed for machine learning.

We use a two-iteration training and testing approach, where in the first iteration we use a subset of 18 DF events with the cleanest seismic signature (15 for training and 3 for testing), and in the second iteration we use all 22 events (20 for training and 2 for testing) to increase the amount of information used to train the machine-learning model. In the second iteration we also inject false positives detections from the first iteration into the noise class. We use Scikit-learn (Pedregosa et al., 2011) implementation of the Random Forest classifier. When we refer to the machine-learning detector we refer to the overall system that involves pre-processing the data, extracting the features, and applying the machine-learning model to the new data.

We follow the following processing steps:

1. **Pre-processing of the seismic data**

The data is detrended and centred, and it is not corrected for instrument response to increase the computational performance. Further, the instrumental response is flat in the 1-50 Hz frequency range. The only pre-processing we apply to data is a low-cutoff filter at 1 Hz to focus on high frequency signals.

2. **Catalog preparation**

To extract debris-flow seismograms we manually pick the start time and the end time of the debris-flow events. The picks are based on Power Spectral Density (PSD) (Figure 1d,f in the manuscript) averaged over 1-50 Hz and root-mean squared (RMS) amplitudes calculated for each station and each event. An example of picks that are delimiting the three different classes is schematically presented on Figure S23. We choose the earliest and the latest picked times for each event as start times and end times making the picks uniform for all stations (Figure S23).

We then randomly choose 550 100-second long noise time windows from 2017, 2018 and 2019 and we manually select 41 rainfall events with a fixed duration of 30 min to compile the noise catalog. For the latter, we use rainfall data from the nearby precipitation station located in Montana, Switzerland (Swiss Meteorological Service). We choose long-duration rain events that do not overlap with the picked debris flow events (at least 45 min time difference between the debris flow events and the rainfall events). We define rain events as rainfall lasting at least 30 min with 1h moving (step=10 min) average higher than 1.4 mm. Note that in the second iteration the noise class contains also the false positive detections from the first iteration.

For the first iteration in total we obtain: 2,966 pre-CD1, 9,886 post-CD1, and 16,614 noise time windows in the catalog. For the second iteration we obtain: 3,631 pre-CD1, 13,046 post-CD1, and 46,355 noise time windows in the catalog. The number of time-windows per class in the second iteration is unbalanced, therefore we account for the non-balanced dataset in the training procedure (see details below).

3. **Extracting statistical features of the seismic data**

For statistical feature calculation, we follow works of Hibert et al. (2017); Maggi et al. (2017); Provost, Hibert, and Malet (2017). They defined a set of significant features that proved efficient in classification of other types of mass movements, including landslides and rockfalls. In total we use 70 features based on the waveform, spectral, spectrogram, and network attributes (see Table S2 for the list of features). We slightly modify the list of features presented in Provost et al. (2017) by adding the following features: (1) waveform features: RMS, and interquartile range, (2) network features: mean Wasserstein distance, and standard deviation of the Wasserstein distance calculated between stations in the frequency band of (1-50) Hz. We use the stations separately meaning that for each station and for each 100 s time window we calculate 59 individual attributes, and 11 network attributes are shared between the stations. We also tried different variations of combining the features and the stations (e.g., combining features from 4 stations or averaging features over stations), but treating the stations separately gives the best results. Finally, we eliminate time windows from the catalog when only one station was online. See Table S2 for the complete list of features, and for the formulas please refer to Provost et al. (2017).

4. **Train-test dataset split**

There are two ways we might approach the splitting of the dataset: (1) by the number of time windows, and (2) by the number of events. We use the first approach to optimize the hyperparameters of the machine-learning algorithm and the second one to assess the accuracy of the classification and the performance of the machine-learning model. Hyperparameters are settings of a machine learning algorithm used to control its behavior (Goodfellow et al., 2016).

In (1) we take the whole dataset as ensemble without considering debris-flow event separation. However, we need (2) to reliably and independently from the hyperparameter optimization estimate the accuracy of the classification.

- Hyperparameter optimization

We use randomly chosen 75 % of the dataset as a training set for the hyperparameter optimization. The hyperparameters are optimized by a 5-fold cross validation with a grid search, which allows for a full search over specified parameter values. The performance of a machine-learning model for different hyperparameter values is measured with accuracy, which is the proportion of sample for which the model produces the correct output.

We tune the following hyperparameters of the random forest algorithm: the number of trees in the forest, the minimum number of time windows required at node leaf, type of class balancing, the maximum depth of a tree, the minimum number of time windows required for an internal node split. The last two hyperparameters control the depth of trees. We use "balanced subsample" random forest mode to minimize a potential bias related to imblanced classes. It automatically adjusts class weights to be inversely proportional to class frequencies in each tree.

- Optimal dataset split per debris-flow event

Next, we search for the optimal number and combination of debris-flow events used for training and testing. We fist choose a subset of 8 debris-flow events with different characteristics (e.g., year of occurrence, volume, and waveforms). From these we choose 2-5 element event combinations that are used in the testing set. For each combination the testing set is independent from the training set. We use the accuracy score for each training-testing combination to assess the performance of the model.In each test we use

the optimized hyperparameters of the random forest algorithm. The accuracy score is normalized and its values are bounded between 0 and 1. The best subset reaches accuracy score of >0.9, meaning that more than 90% of labels are predicted correctly.

The debris-flow events used in the final testing set are marked in orange in Table S1 (3 events for the first iteration), and in green (2 events for the second iteration). For the noise class, we randomly select 1/3 of the samples for the testing process. In overall we use 80 % of the dataset for training and 20 % for testing.

5. **Building and evaluating the machine-learning model**

The model is then built in a training phase in which the machine-learning algorithm has access to seismic features and the corresponding classes which are used as labels. To evaluate the model we apply it to the testing set with restricted access to the features. We then compare the predicted labels to the true labels. The accuracy score might be misleading for the imbalanced dataset, so to better assess the performance of the model we calculate a confusion matrix and Receiver Operating Characteristic (ROC) curves with cross validation (5 different splits).

The confusion matrix from the first iteration is presented in Figure S24a. It indicates that the machine learning model performs less well in classification of pre-CD1 and noise signals with a score of 0.73.

Figure S25 shows the ROC curves for the three classes. ROC curves shows the true positive rate (TPR=TP/(TP/FN)) on the y axis and false positive rate (FPR=FP/(FP+TN)) on the x axis, where: TP=true positive, FN=false negatives, FP=false positives, TN=true negatives.

The higher the area under the curve (AUC), the better the accuracy of the classifier. The ROC curve is computed for each class separately by pairwise comparison (one class vs. all other classes) and the dotted line shows the baseline for a random guess. In general, the ROC curves shows good classification results, strongly above the random baseline with AUC values of 0.9, and with low standard deviation values meaning that the classifier output should not be too much affected by changes in the training data (Pedregosa et al., 2011). Moreover, the ROC curves indicate the improvement of the performance of the machine-learning model in the second iteration.

Figure S24b shows pairwise relationships between the three most important features. The three most important features belong to the network attributes: 1. ratio between the maximum RMS and the minimum RMS in the network, 2. station number with maximum RMS, and 3. maximum coherence between station pairs. The three most important features are the same in the first and in the second iteration.

6. **Applying the machine-learning model to new data**

We now apply the machine-learning model to 2019 continuous data. The results of the first iteration (Figure S24c) indicate that the model often classifies anthropogenic noise as post-CD1 class. This is interesting since it is not represented in the confusion matrix, hinting that the classification accuracy estimated over limited testing set might not fully represent the classification accuracy over the entire seismic dataset. These false positives are related to the presence of anthropogenic noise in Illgraben area because they appear only during week days, e.g., Monday-Friday:154-158, 182-186, 189-193 Julian days between ∼5 a.m. UTC (7 a.m. local time) and 10 a.m-11 a.m. until 15 (17 local time). The false

positives are consequently injected into the noise class in the second iteration to teach the model how to distinguish between this anthropogenic noise and debris-flow signals.

One can ask how many debris flow events are needed to train the machine-learning model and obtain a good performance in classification. Figure S29 shows a sensitivity test: balanced accuracy score [the average of recall TP/(TP+TN) obtained on each class] as a function of number of debris-flow events used in training set with cross-validation (5 folds). The results show that using even a single debris-flow event for training gives better results than a random guess, although higher values of balanced accuracy (>0.7) and stable prediction are obtained from 9 debris-flow events used in the training set.

Text S2. Apparent total impact force spectrum (AIFS) calculation

The vertical apparent total impact force spectrum (AIFS) of the debris flow events is calculated following Zhang et al. (2020). The peaks of seismic signals recorded at stations ILL11 and ILL2 (pink and yellow triangle on Figure 1) are used to estimate the average velocity of the debris flows. To calculate the AIFS, we assume that the velocity of the debris flow is constant during run-out and that the events originate upstream from CD1. The average debris flow velocities used in the AIFS calculations are presented in Table S3.

Two large peaks in the vertical AIFS at around 2,100 m and 3,100 m visible on Figure S30 are probably caused by large CDs located at these channel locations (Zhang et al., 2020). The AIFS is lower at 2,500 m, which might be related to a denser distribution of check dams at 2,500 m. If large particles (e.g., boulders) are present in a debris flow, they gradually gather at the flow front due to the particle sorting phenomena which corresponds

to an increase in AIFS (events on 4, 7, 8(2), 17, 29 June and 8 August 2020) . Other events seem to have lower volumes and strong deposits.

References

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. (`http://www.deeplearningbook.org`)

Hibert, C., Provost, F., Malet, J.-P., Maggi, A., Stumpf, A., & Ferrazzini, V. (2017). Automatic identication of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a random forest algorithm. *Journal of Volcanology and Geothermal Research*, *340*, 130-142.

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P., & Amemoutou, A. (2017). Implementation of a multistation approach for automated event classication at piton de la fournaise volcano. *Seismological Research Letters*, *88*(3), 878-891.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Provost, F., Hibert, C., & Malet, J.-P. (2017). Automatic classication of endogenous landslide seismicity using the Random Forest supervised classier. *Geophysical Research Letters*, *44*(1), 113. doi: 10.1002/2016GL070709

Schlunegger, F., Badoux, A., McArdell, B. W., Gwerder, C., Schnydrig, D., Rieke-Zapp, D., & Molnar, P. (2009). Limits of sediment transfer in an alpine debris-flow catchment, Illgraben, Switzerland. *Quat. Sci. Rev.*, *28*, 1097-1105. doi: doi:10.1016/j.quascirev.2008.10.025

Walter, F., Burtin, A., McArdell, B. W., Hovius, N., Weder, B., & M, T. J. (2017). Testing seismic amplitude source location for fast debris-flow detection at Illgraben, Switzerland. *Natural Hazards and Earth System Sciences*, *17*(6), 1939?955. doi: https://doi.org/10.5194/nhess?17?939?2017

Zhang, Z., Walter, F., McArdell, B. W., Wenner, M., Chmiel, M., & He, S. (2020). Extracting dynamics of debris flows from their seismic signature. *GRL*, *in submission*.

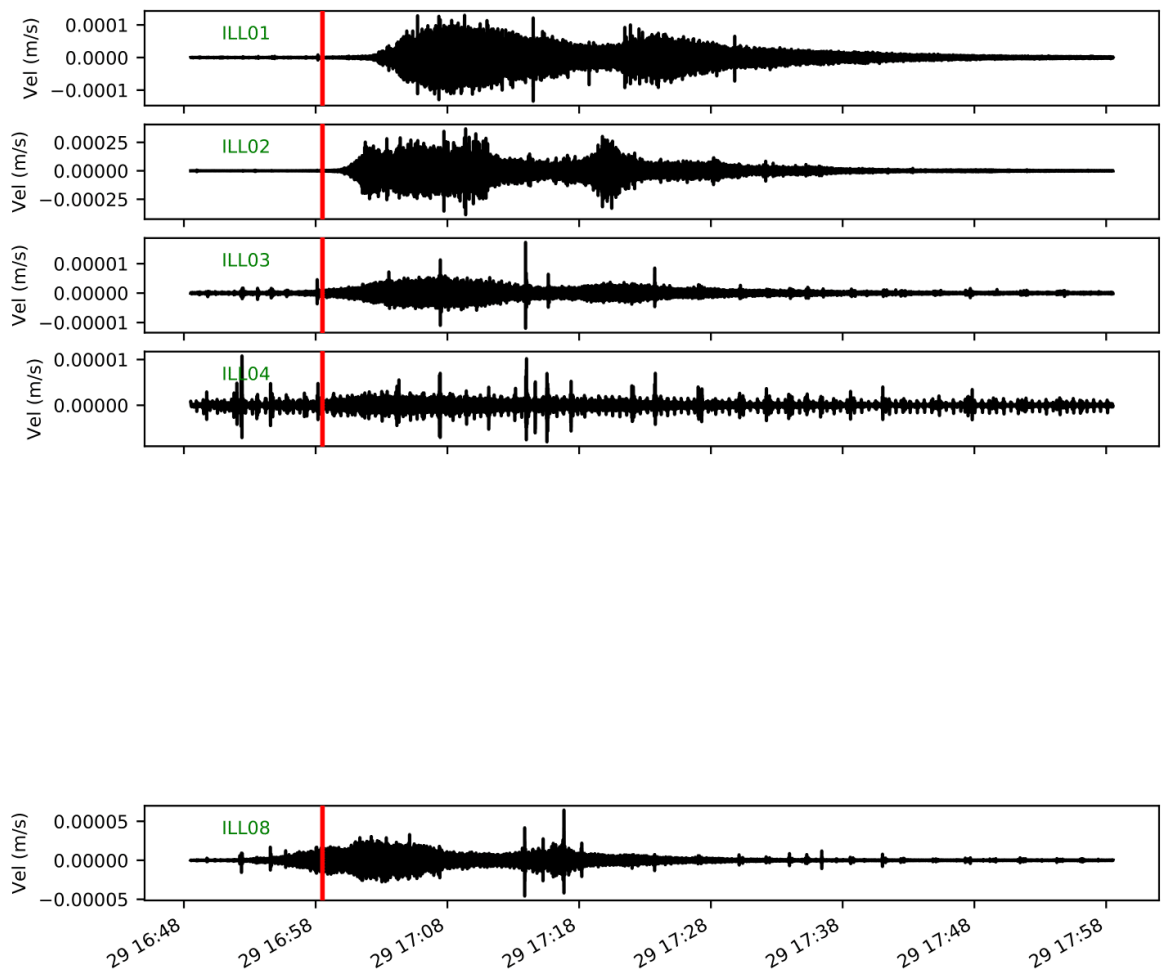

Figure S1. (a) Vertical-component seismograms generated by a debris-flow event on 29 May 2017. The arrival time of the debris flow front at CD1 is marked in red.

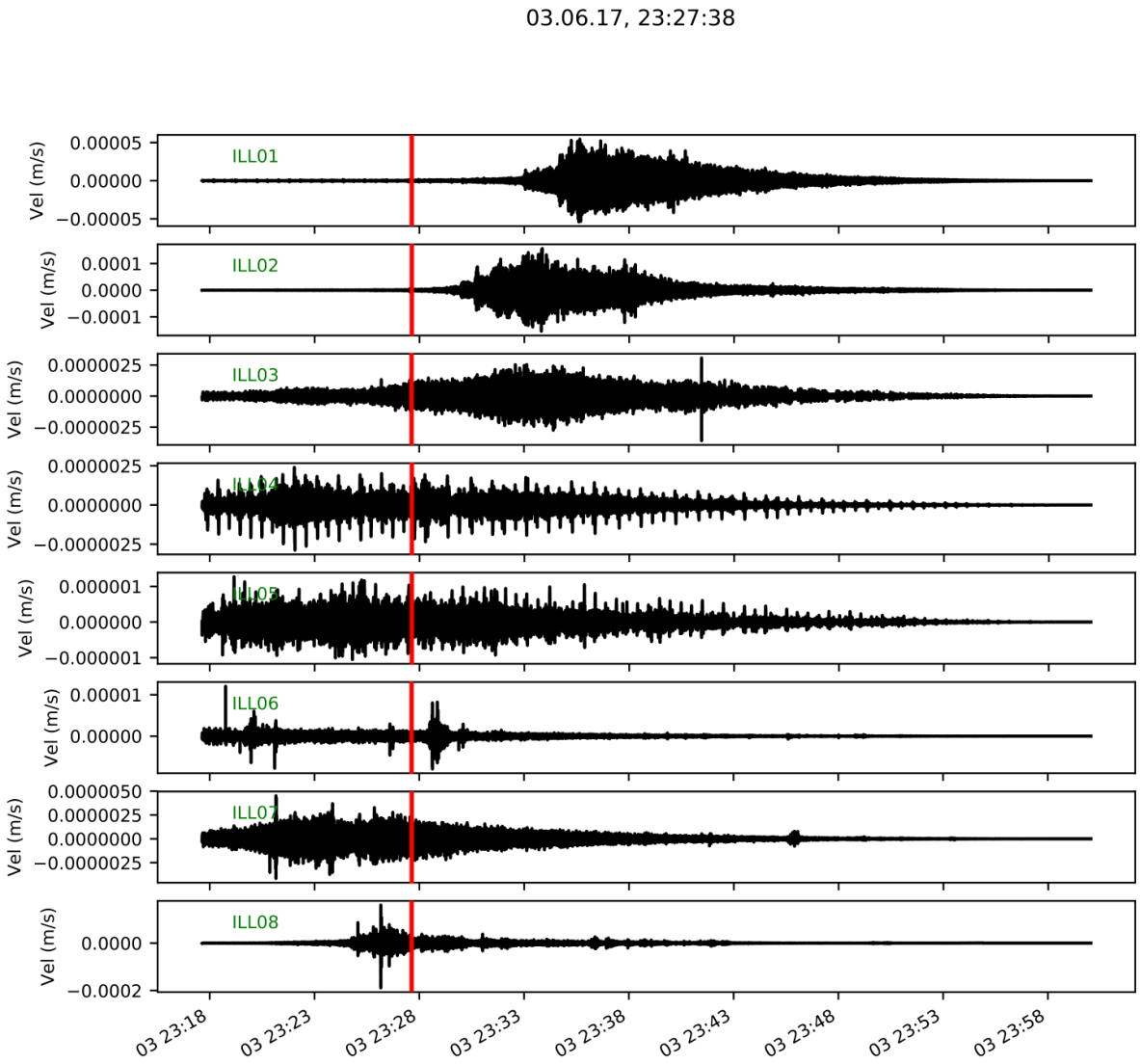

Figure S2. (a) Vertical-component seismograms generated by a debris-flow event on 3 June 2017. The arrival time of the debris flow front at CD1 is marked in red.

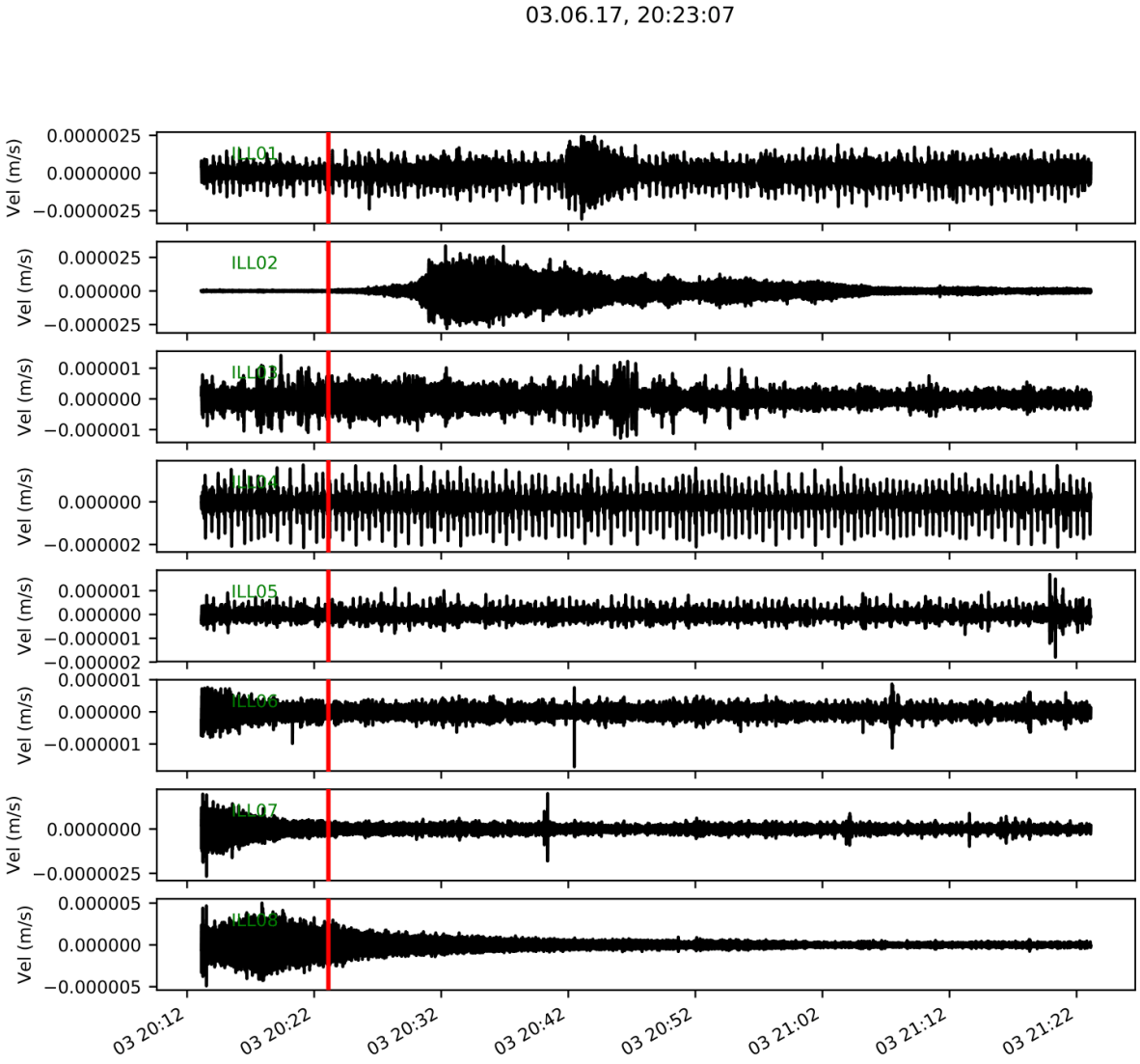

Figure S3. (a) Vertical-component seismograms generated by a debris-flow event on 3 June 2017. The arrival time of the debris flow front at CD1 is marked in red.

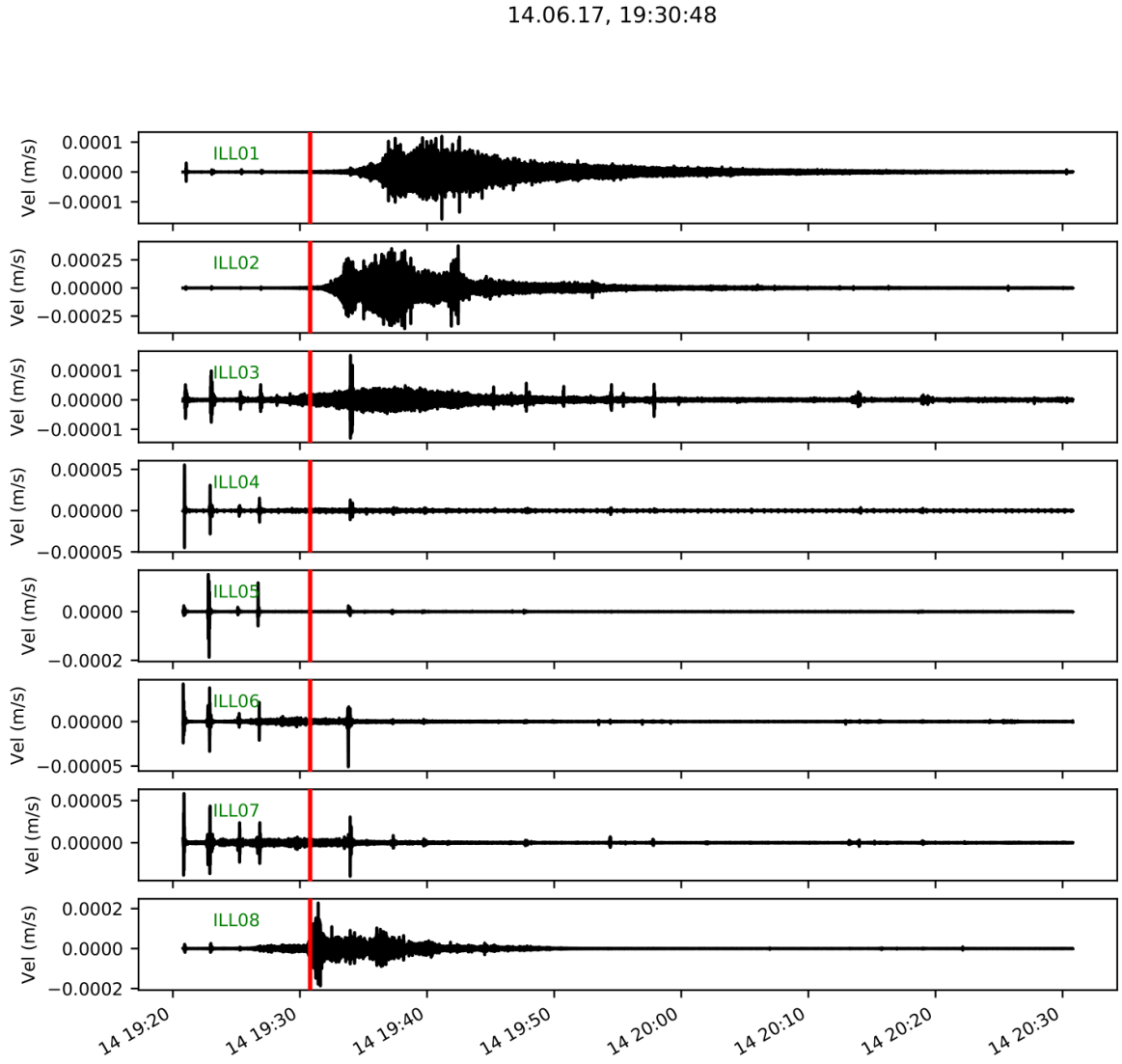

Figure S4. (a) Vertical-component seismograms generated by a debris-flow event on 14 June 2017. The arrival time of the debris flow front at CD1 is marked in red.

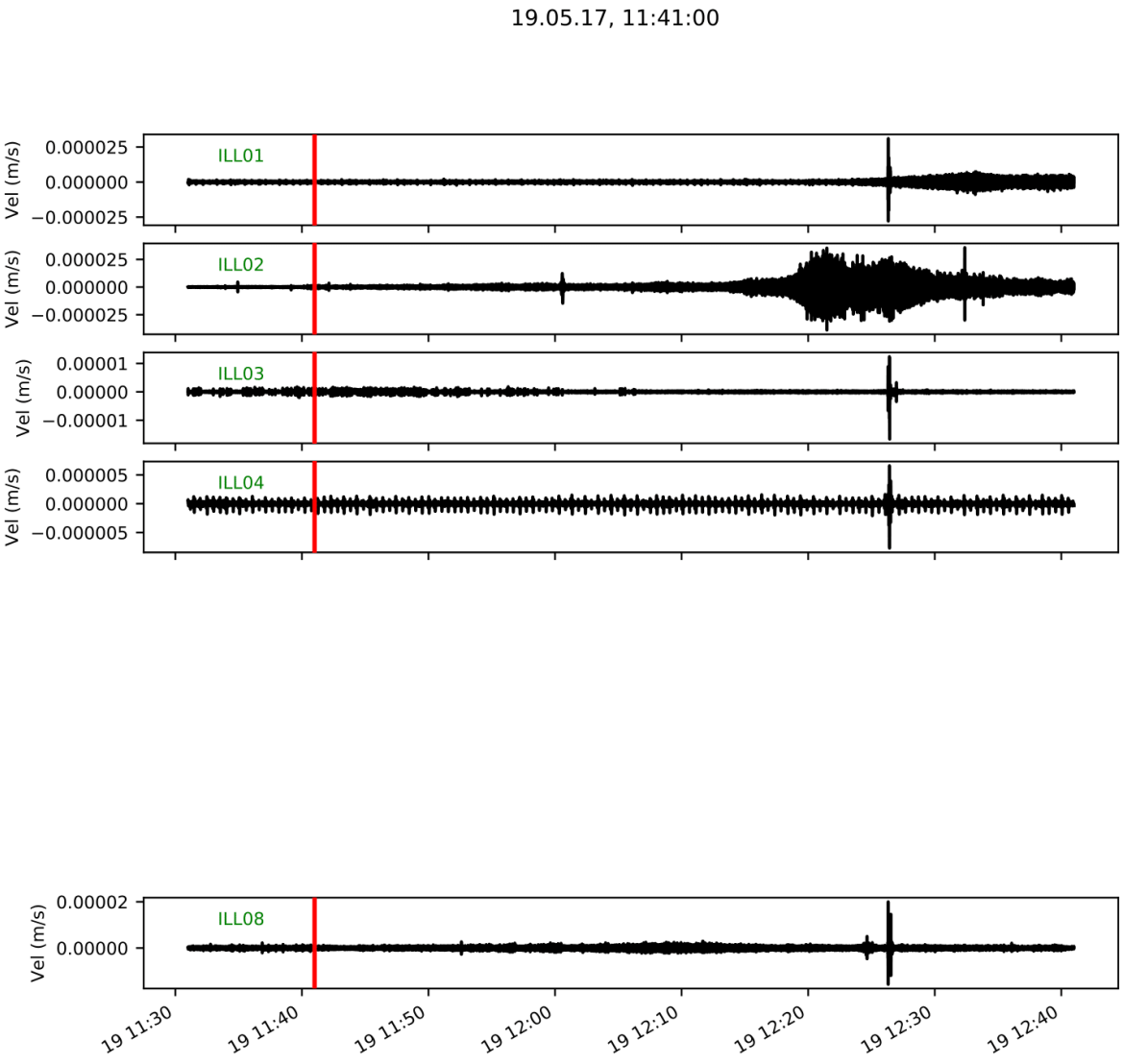

Figure S5. (a) Vertical-component seismograms generated by a debris-flow event on 19 May 2017. The arrival time of the debris flow front at CD1 is marked in red.

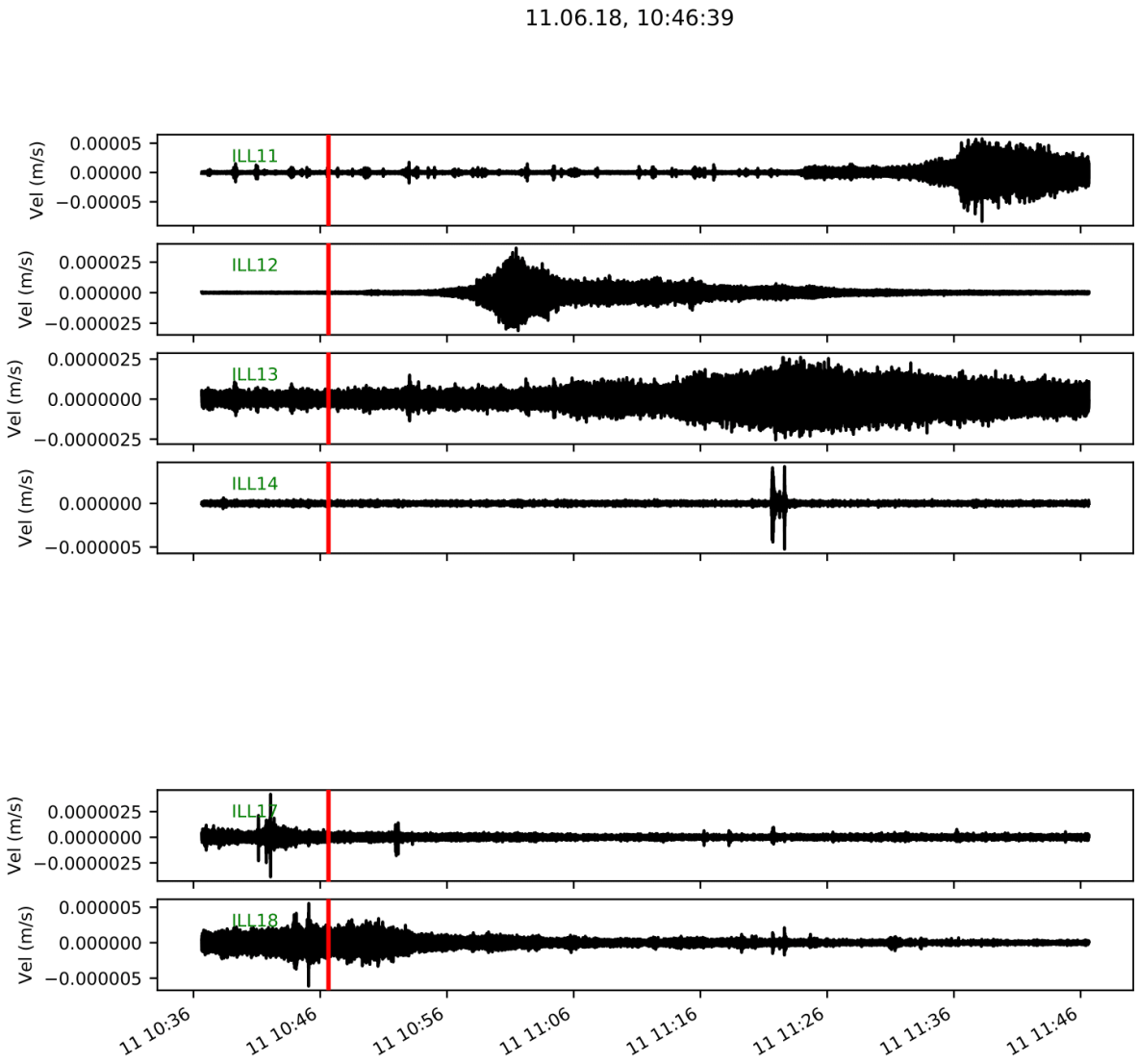

Figure S6. (a) Vertical-component seismograms generated by a debris-flow event on 11 June 2018. The arrival time of the debris flow front at CD1 is marked in red.

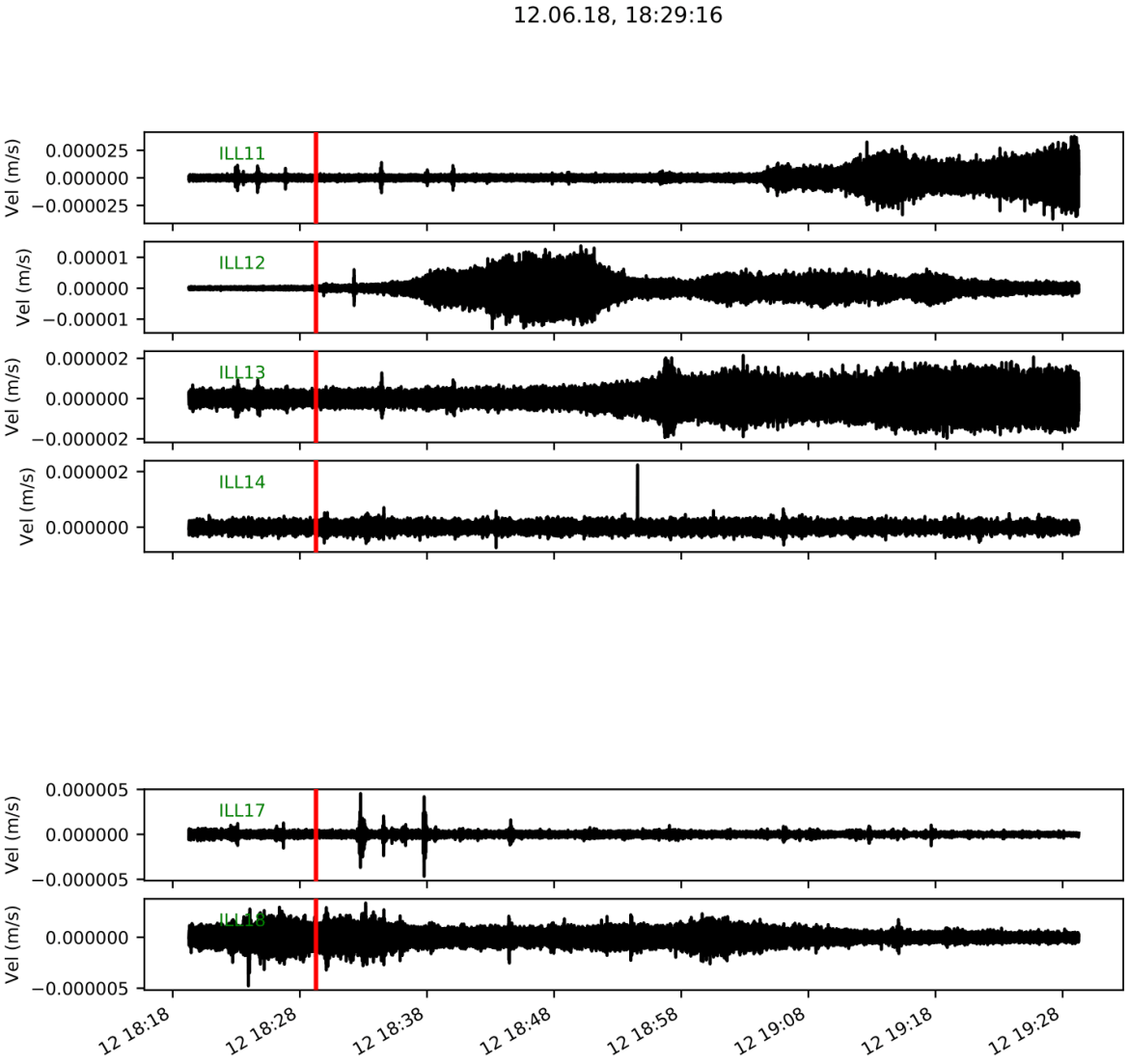

Figure S7. (a) Vertical-component seismograms generated by a debris-flow event on 12 June 2018. The arrival time of the debris flow front at CD1 is marked in red.

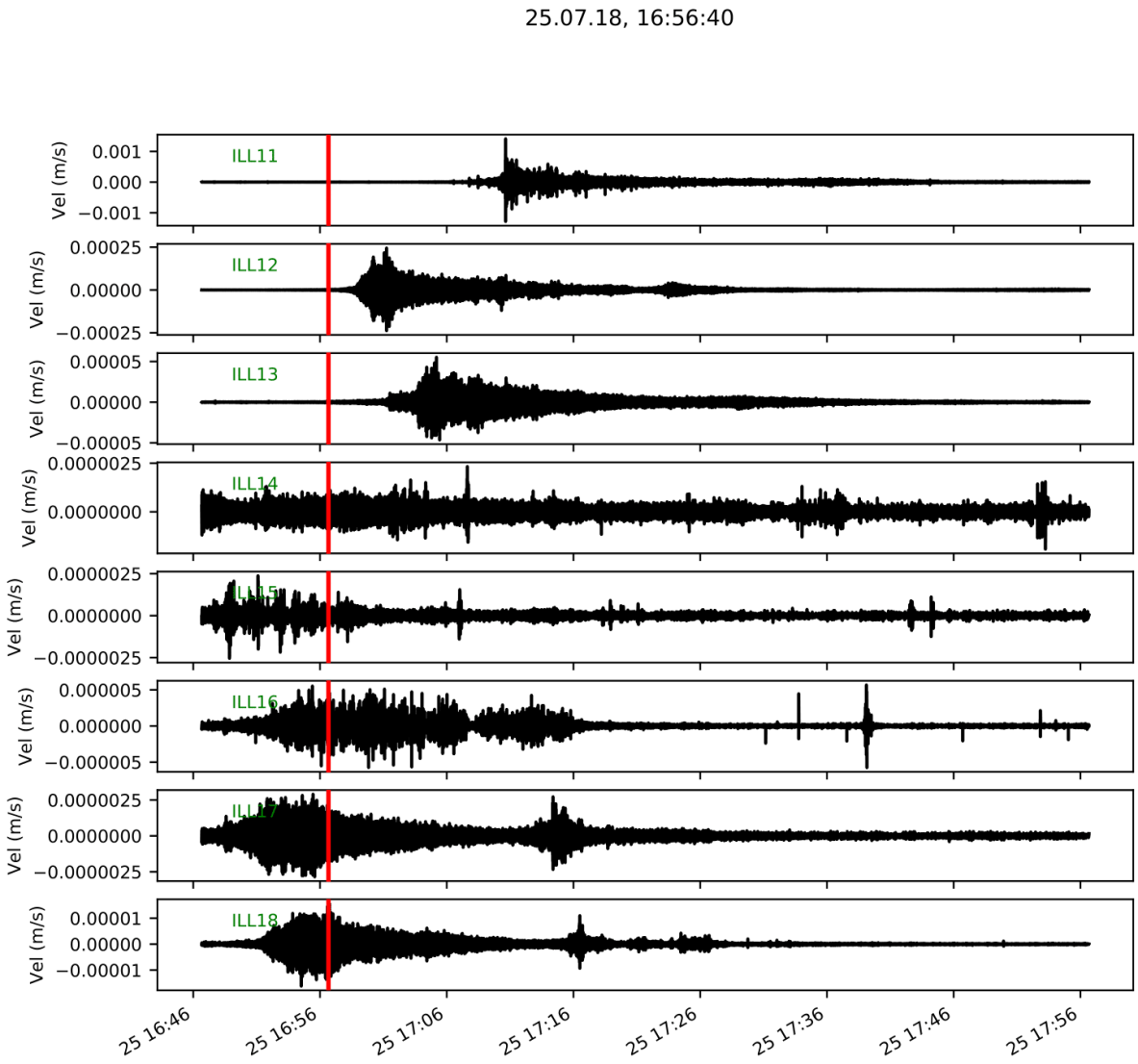

Figure S8. (a) Vertical-component seismograms generated by a debris-flow event on 25 July 2018. The arrival time of the debris flow front at CD1 is marked in red.

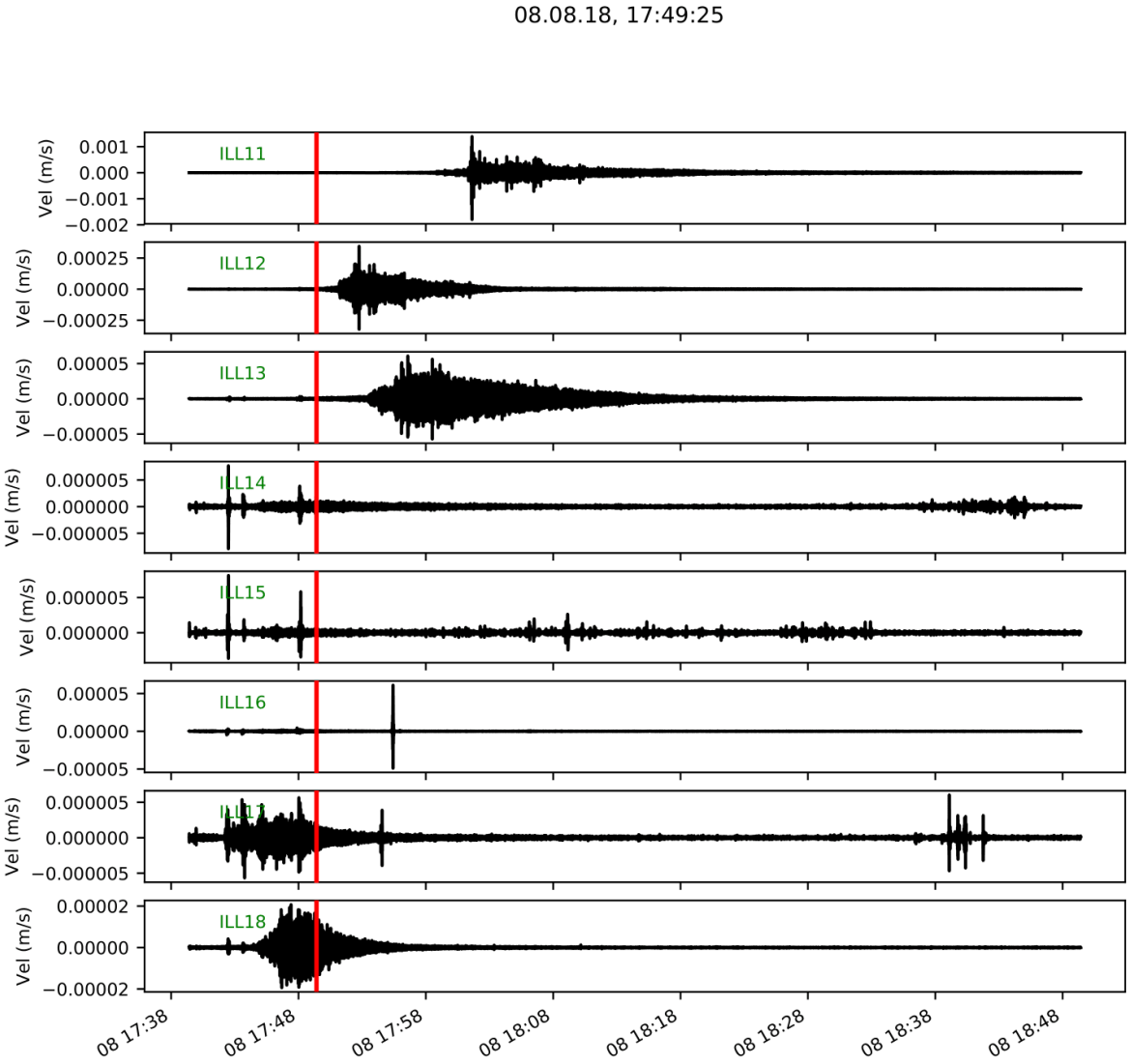

Figure S9. (a) Vertical-component seismograms generated by a debris-flow event on 08 August 2018. The arrival time of the debris flow front at CD1 is marked in red.

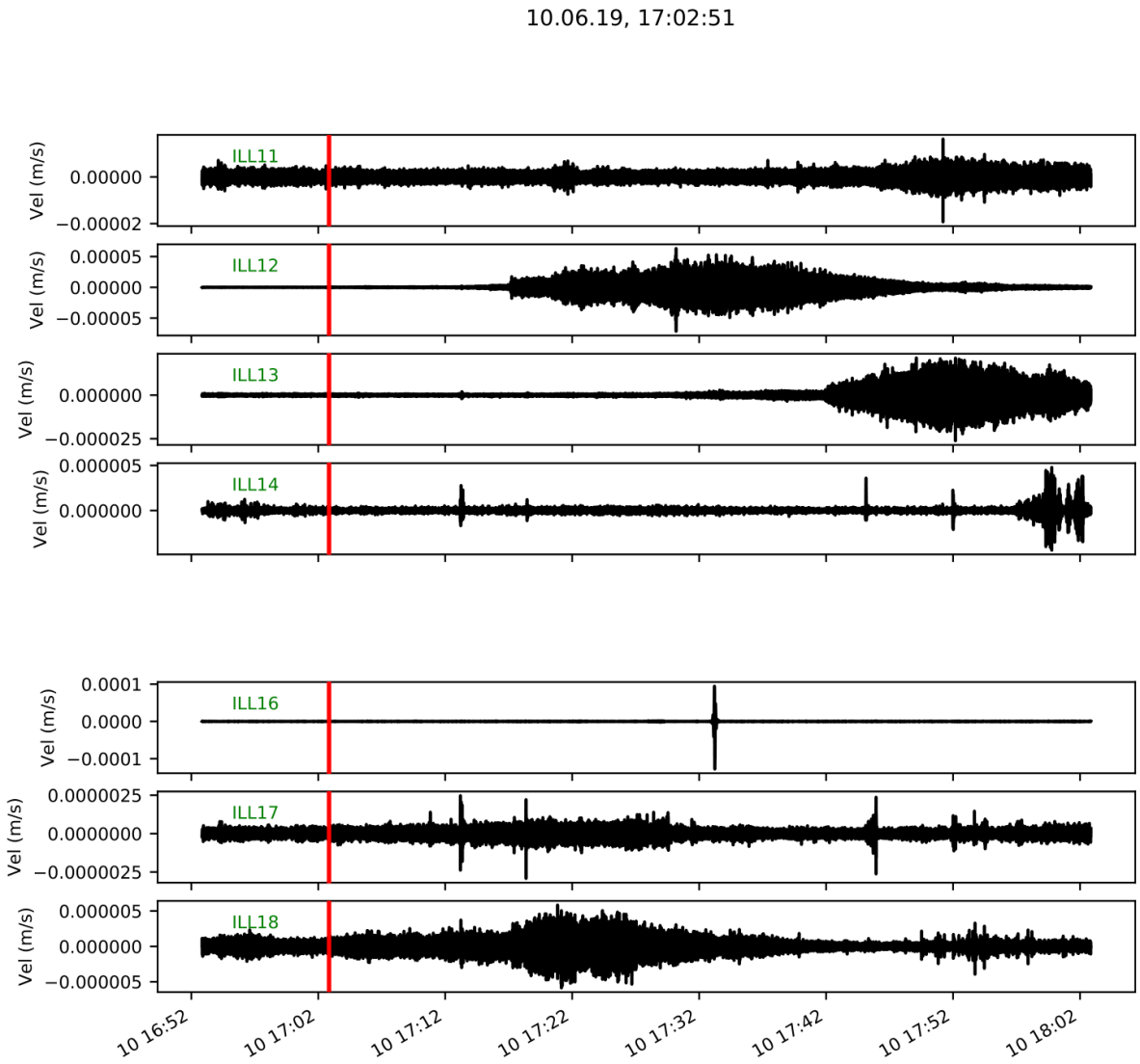

Figure S10. (a) Vertical-component seismograms generated by a debris-flow event on 10 June 2019. The arrival time of the debris flow front at CD1 is marked in red.

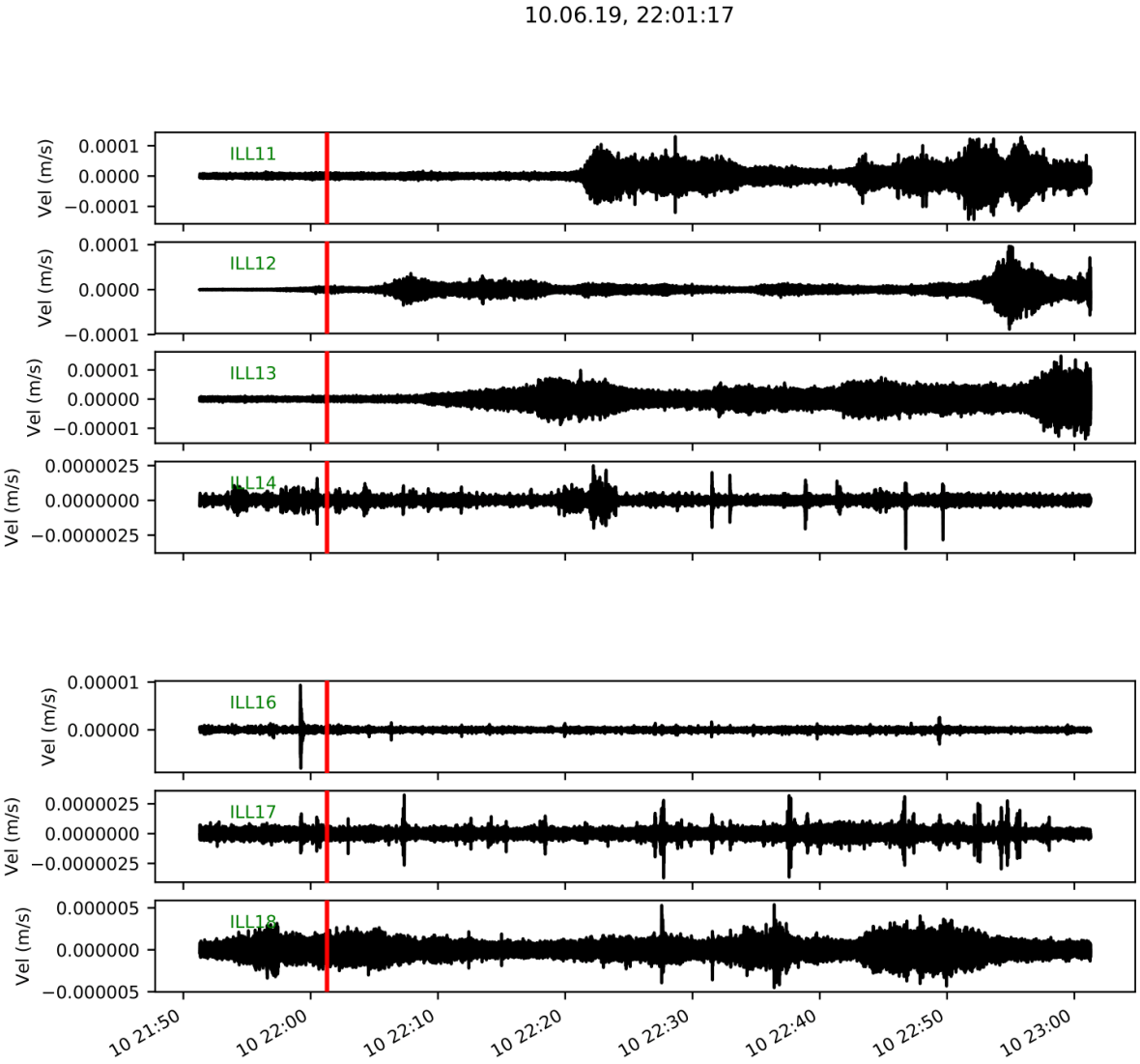

Figure S11. (a) Vertical-component seismograms generated by a debris-flow event on 10 June 2019. The arrival time of the debris flow front at CD1 is marked in red.

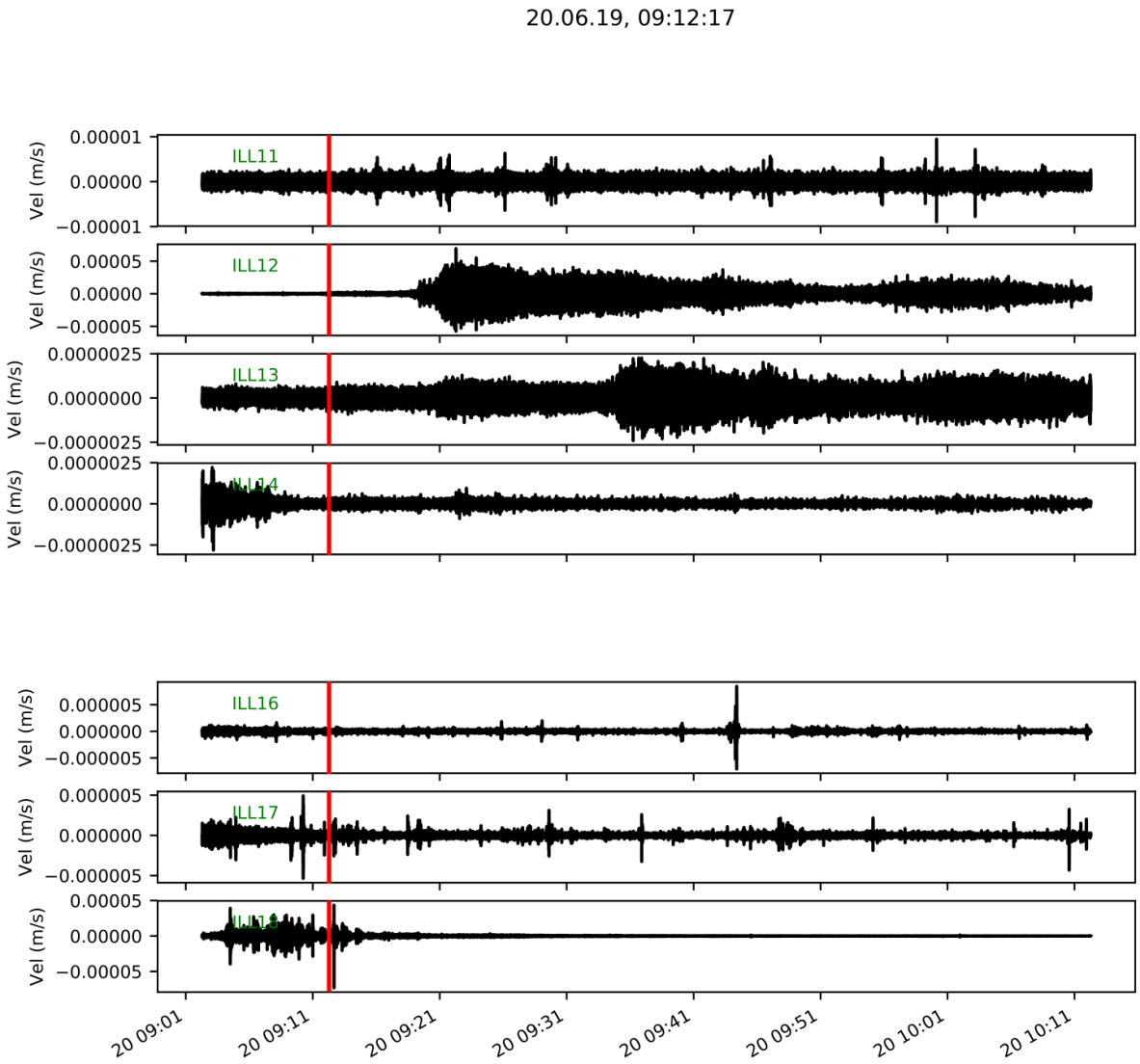

Figure S12. (a) Vertical-component seismograms generated by a debris-flow event on 20 June 2019. The arrival time of the debris flow front at CD1 is marked in red.

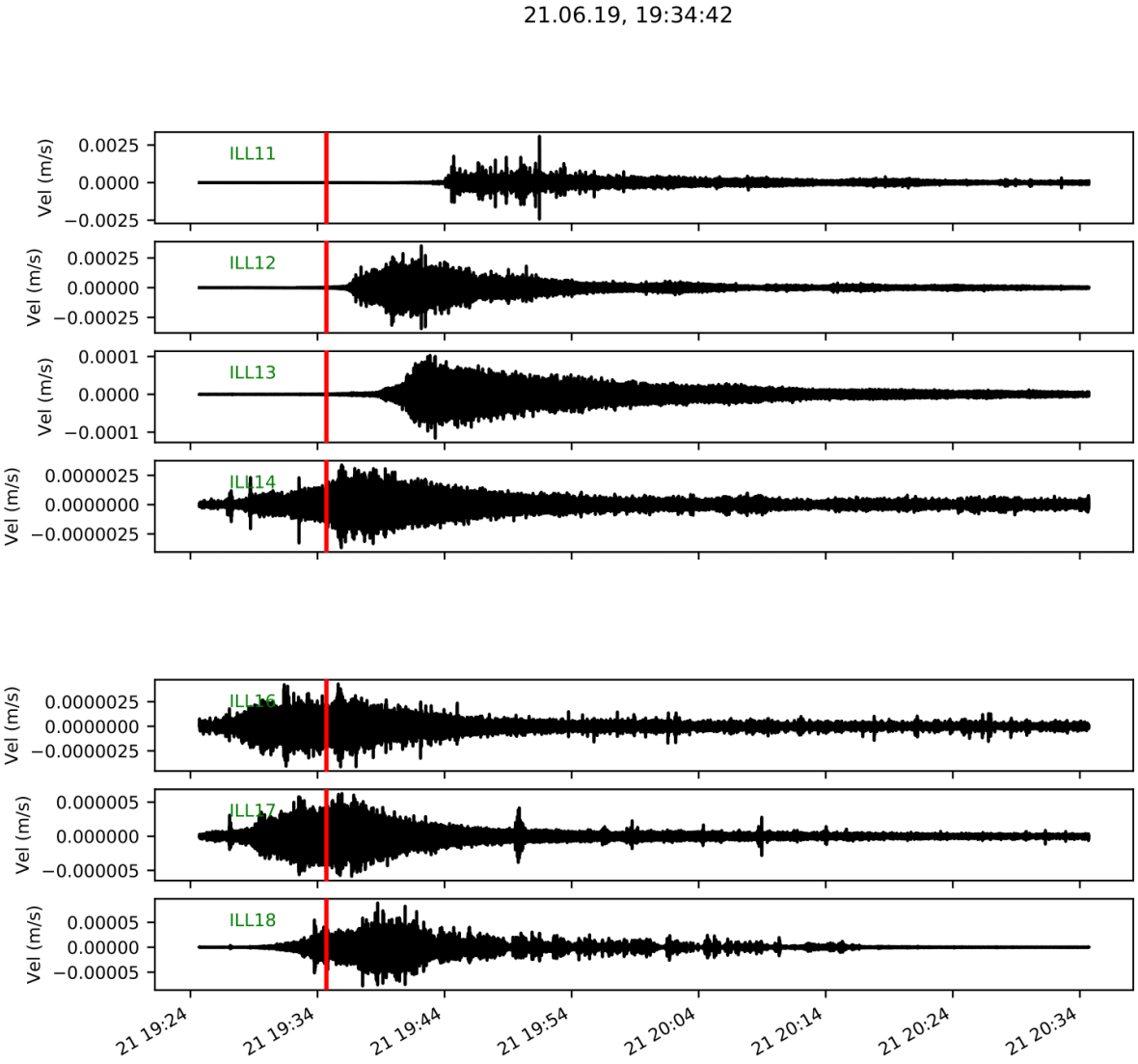

Figure S13. (a) Vertical-component seismograms generated by a debris-flow event on 21 June 2019. The arrival time of the debris flow front at CD1 is marked in red.

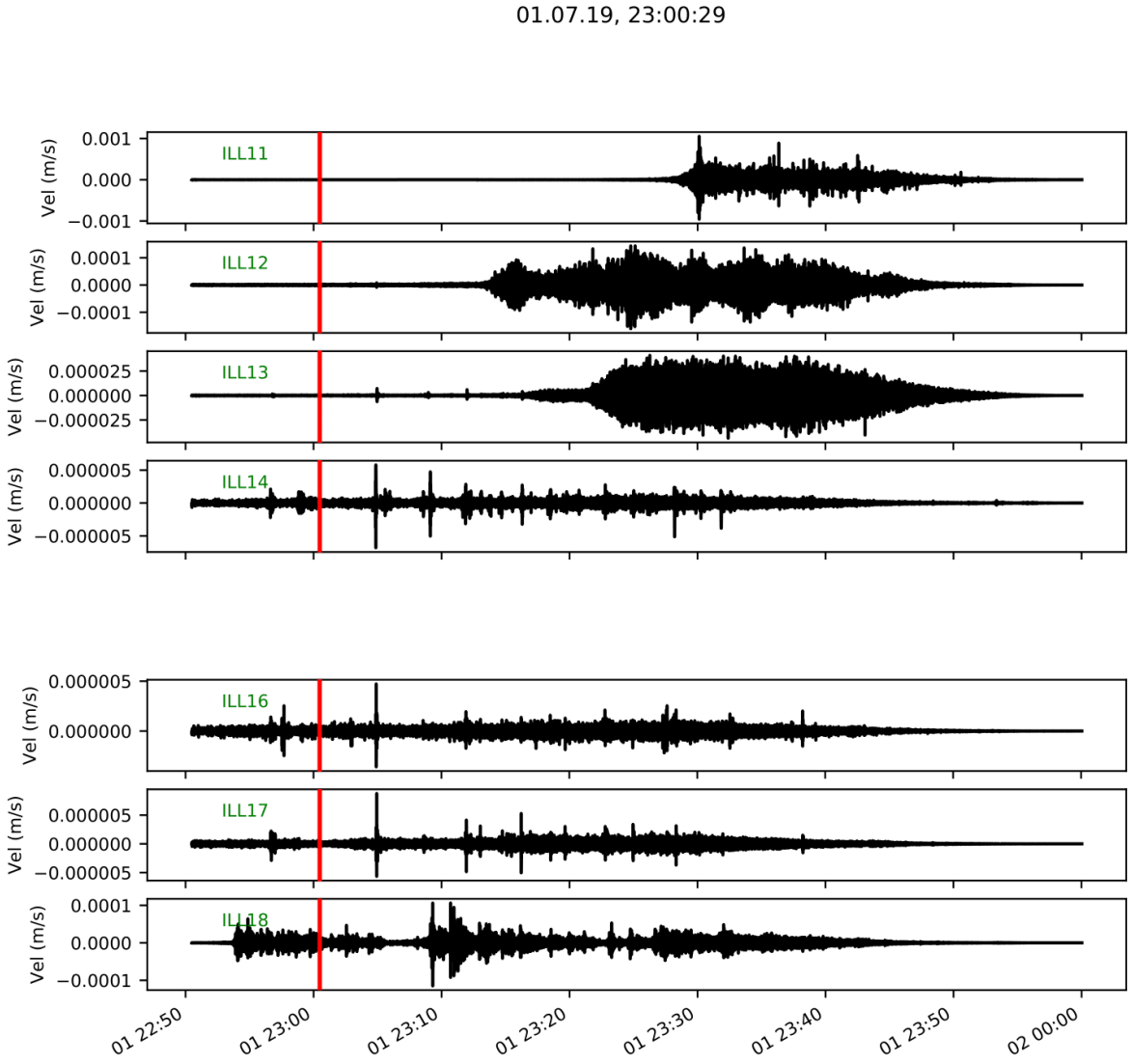

Figure S14. (a) Vertical-component seismograms generated by a debris-flow event on 01 July 2019. The arrival time of the debris flow front at CD1 is marked in red.

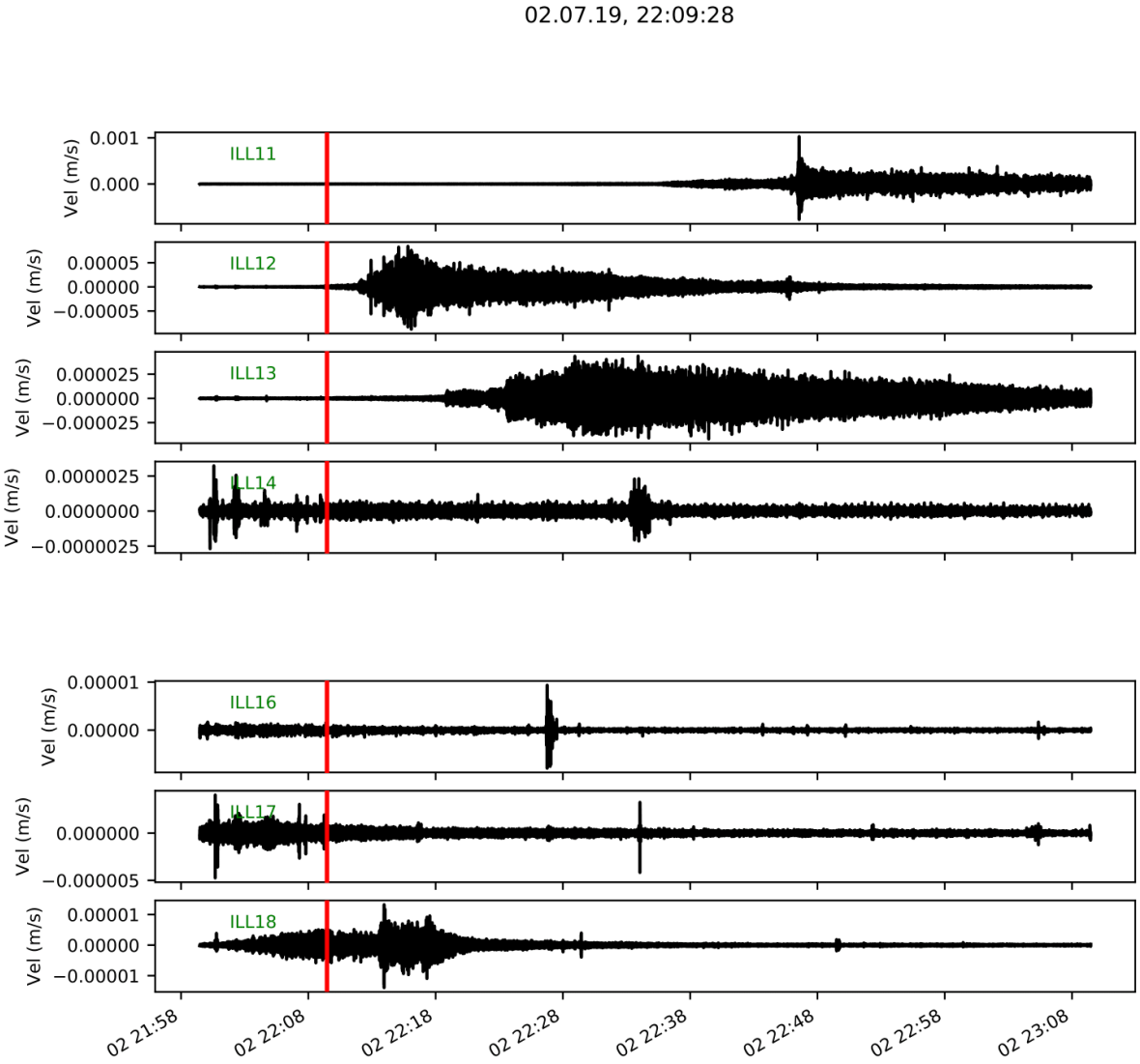

Figure S15. (a) Vertical-component seismograms generated by a debris-flow event on 02 July 2019. The arrival time of the debris flow front at CD1 is marked in red.

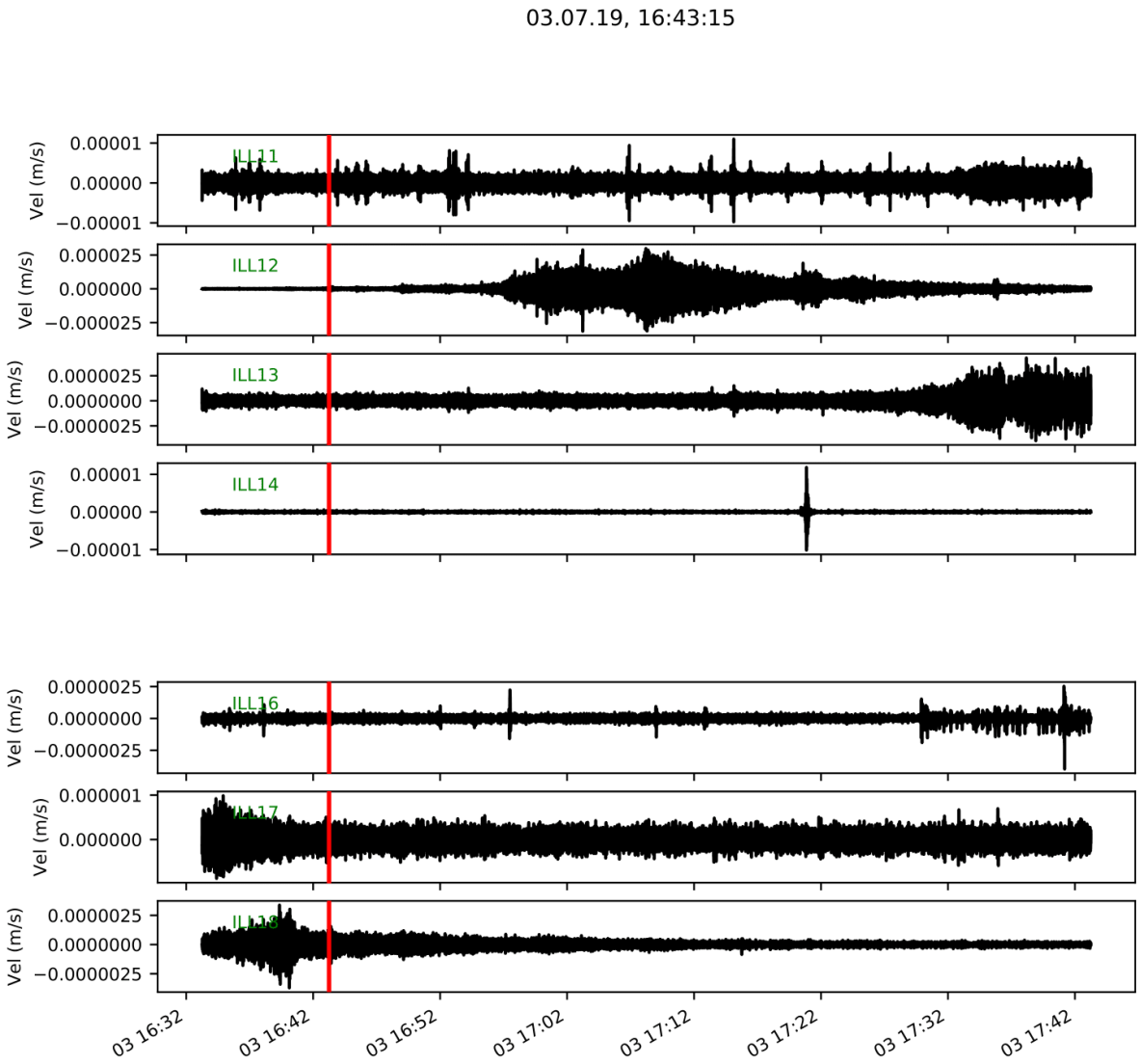

Figure S16. (a) Vertical-component seismograms generated by a debris-flow event on 03 July 2019. The arrival time of the debris flow front at CD1 is marked in red.

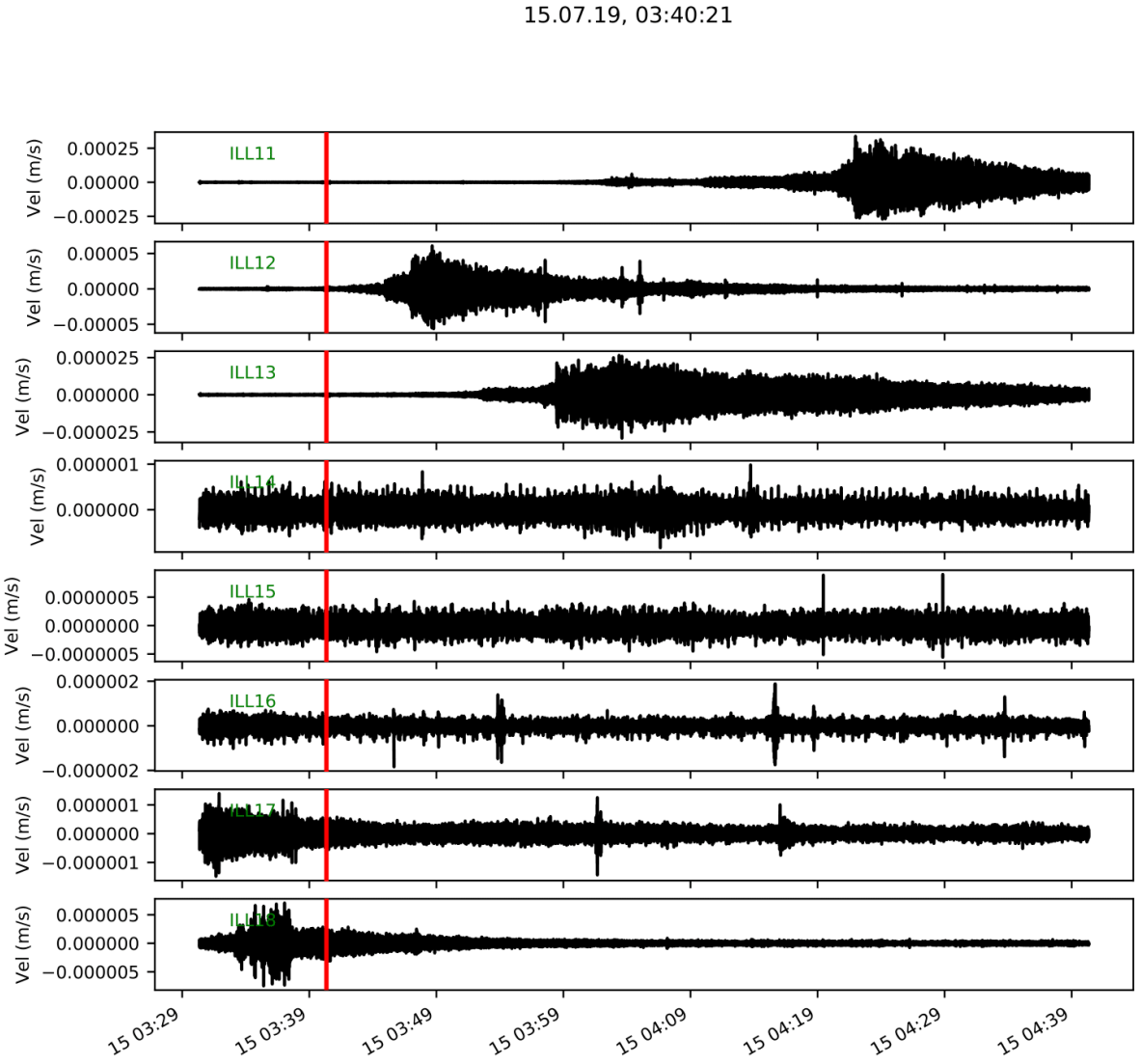

Figure S17. (a) Vertical-component seismograms generated by a debris-flow event on 15 July 2019. The arrival time of the debris flow front at CD1 is marked in red.

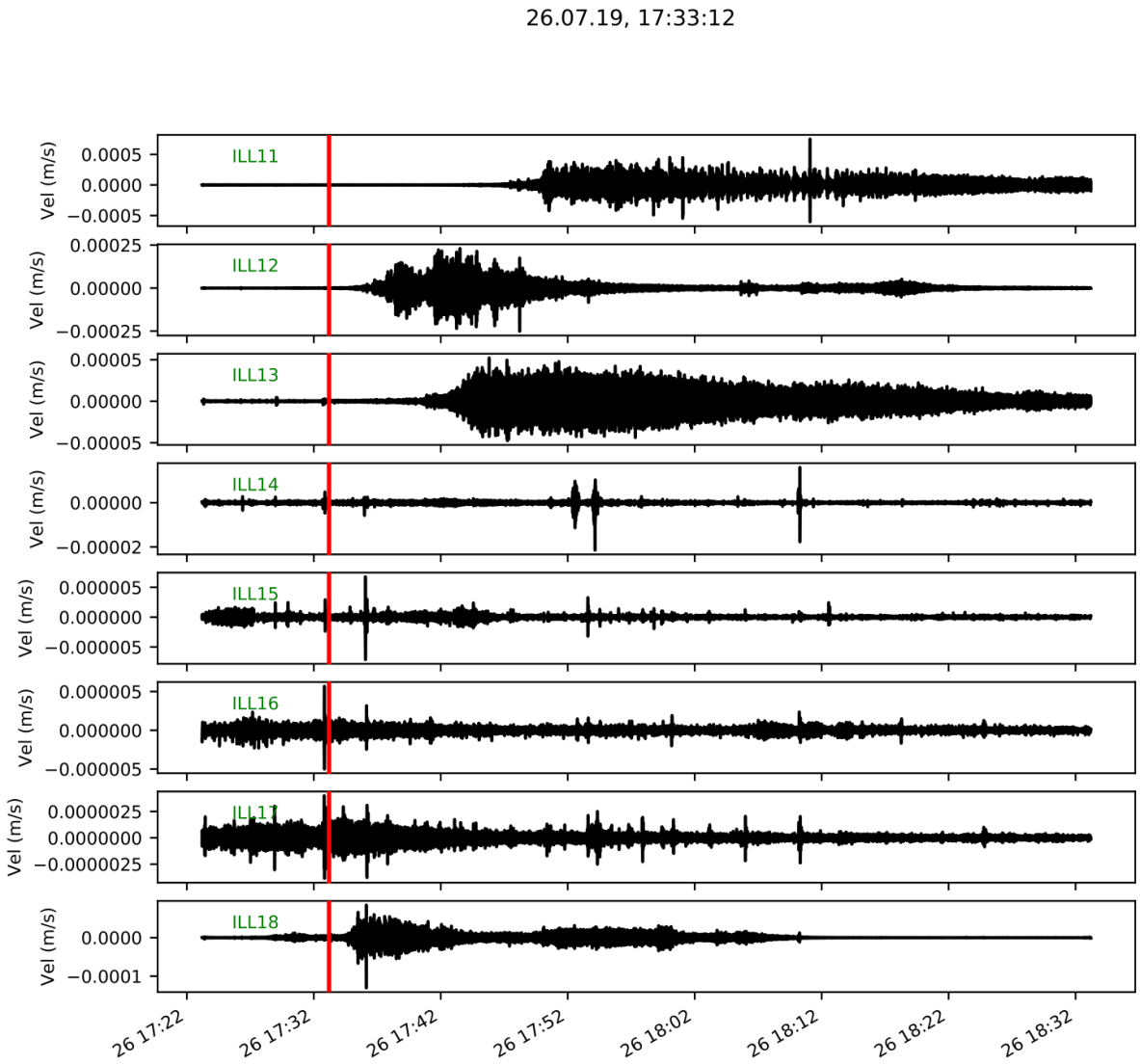

Figure S18. (a) Vertical-component seismograms generated by a debris-flow event on 26 July 2019. The arrival time of the debris flow front at CD1 is marked in red.

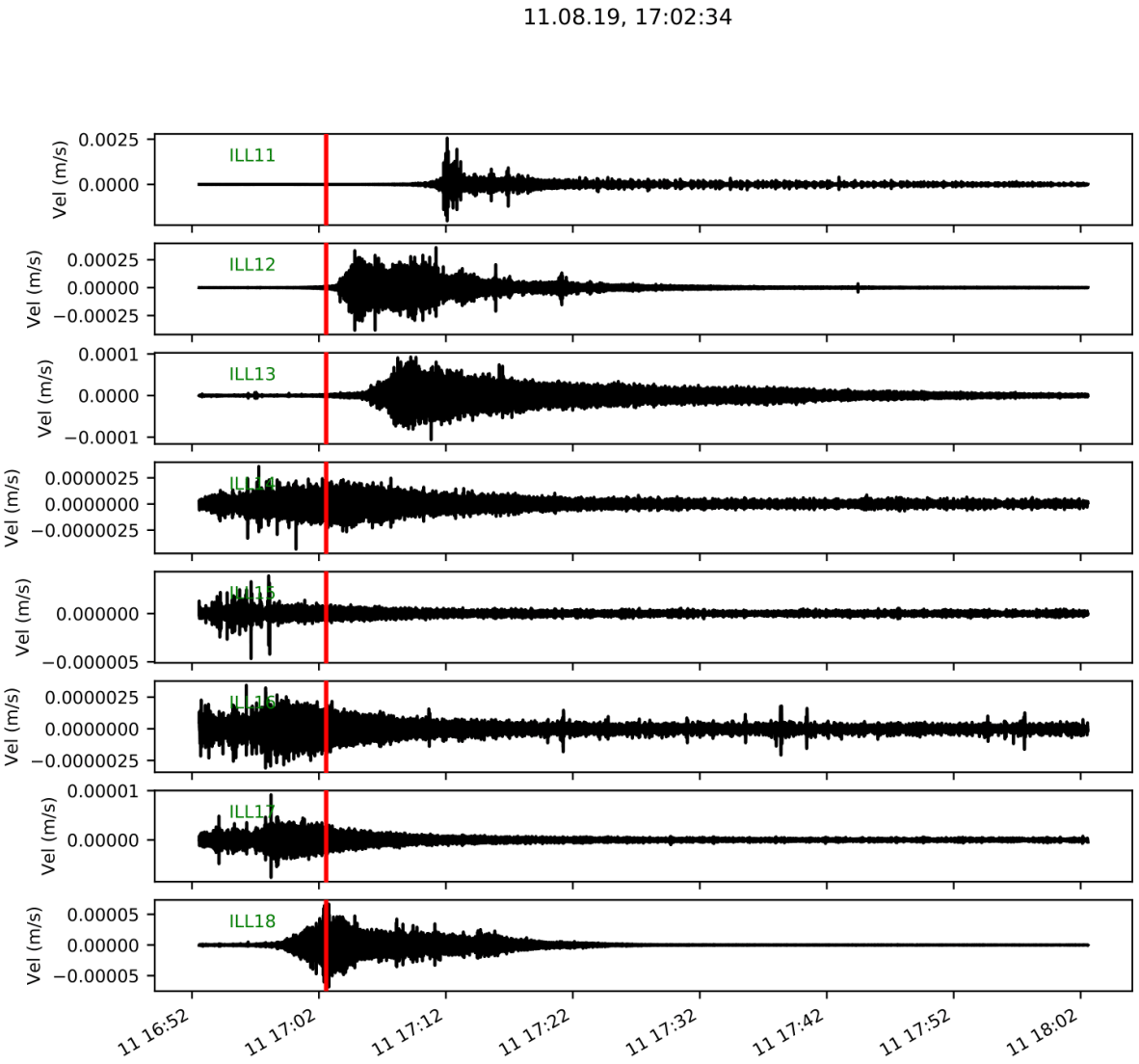

Figure S19. (a) Vertical-component seismograms generated by a debris-flow event on 11 August 2019. The arrival time of the debris flow front at CD1 is marked in red.

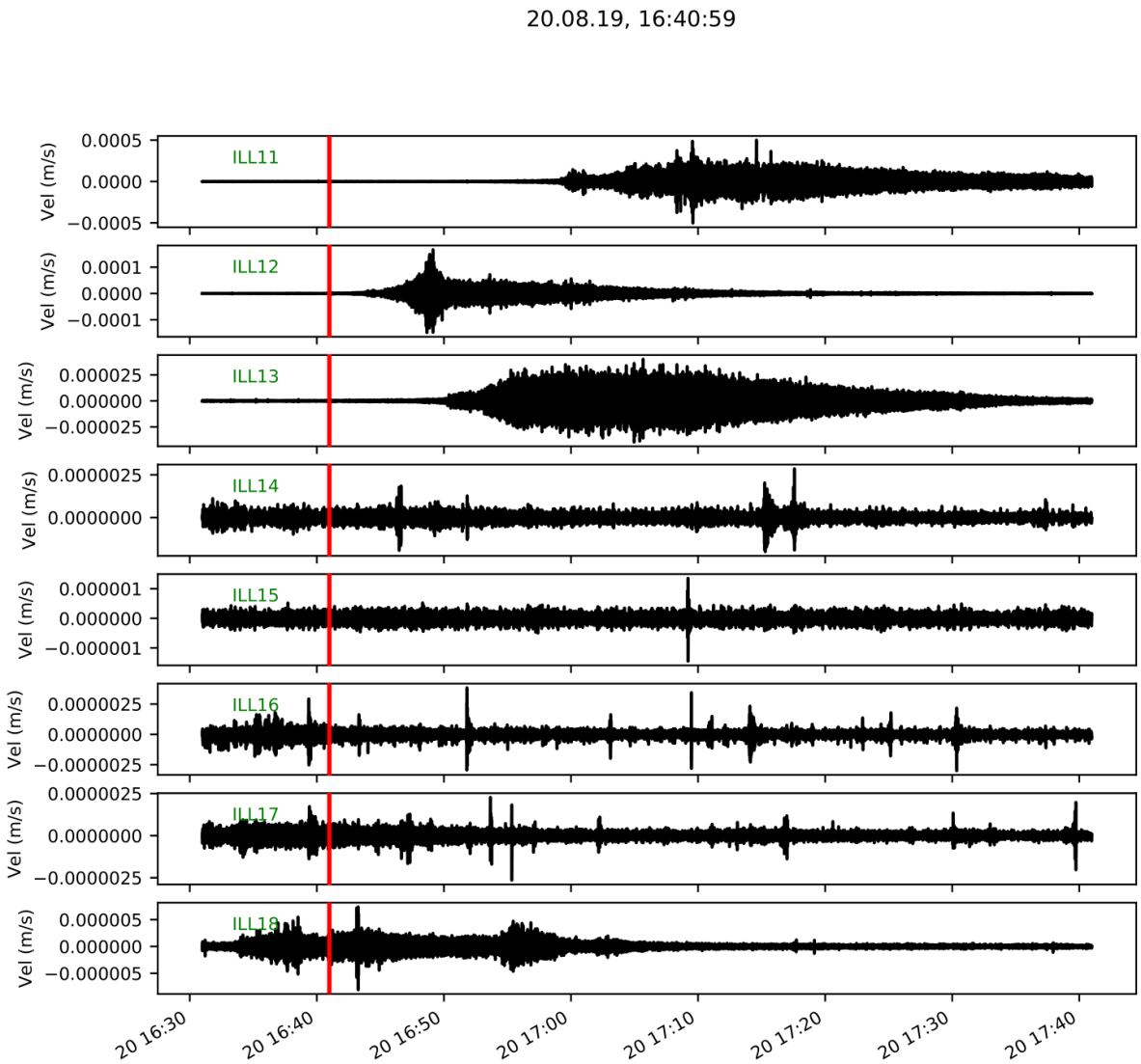

Figure S20. (a) Vertical-component seismograms generated by a debris-flow event on 20 August 2019. The arrival time of the debris flow front at CD1 is marked in red.

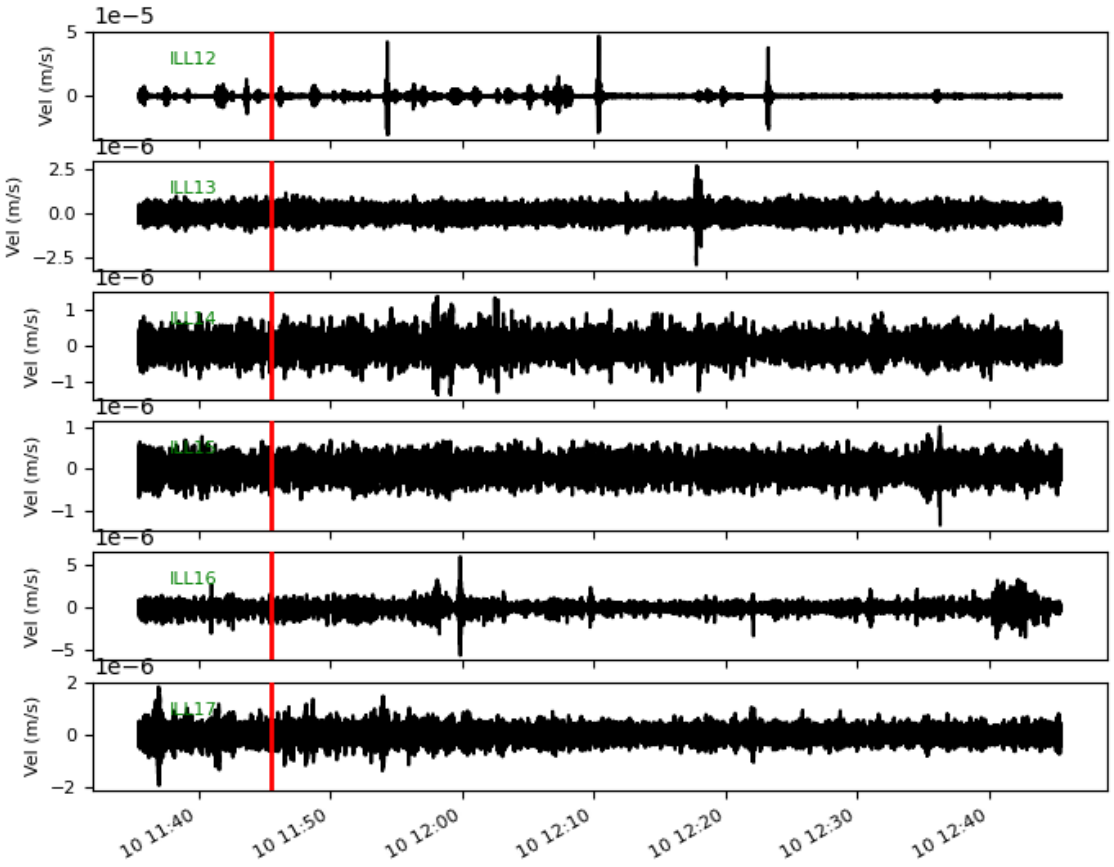

Figure S21. (a) Vertical-component seismograms generated by a debris-flow event on 10 October 2019. The arrival time of the debris flow front at CD1 is marked in red.

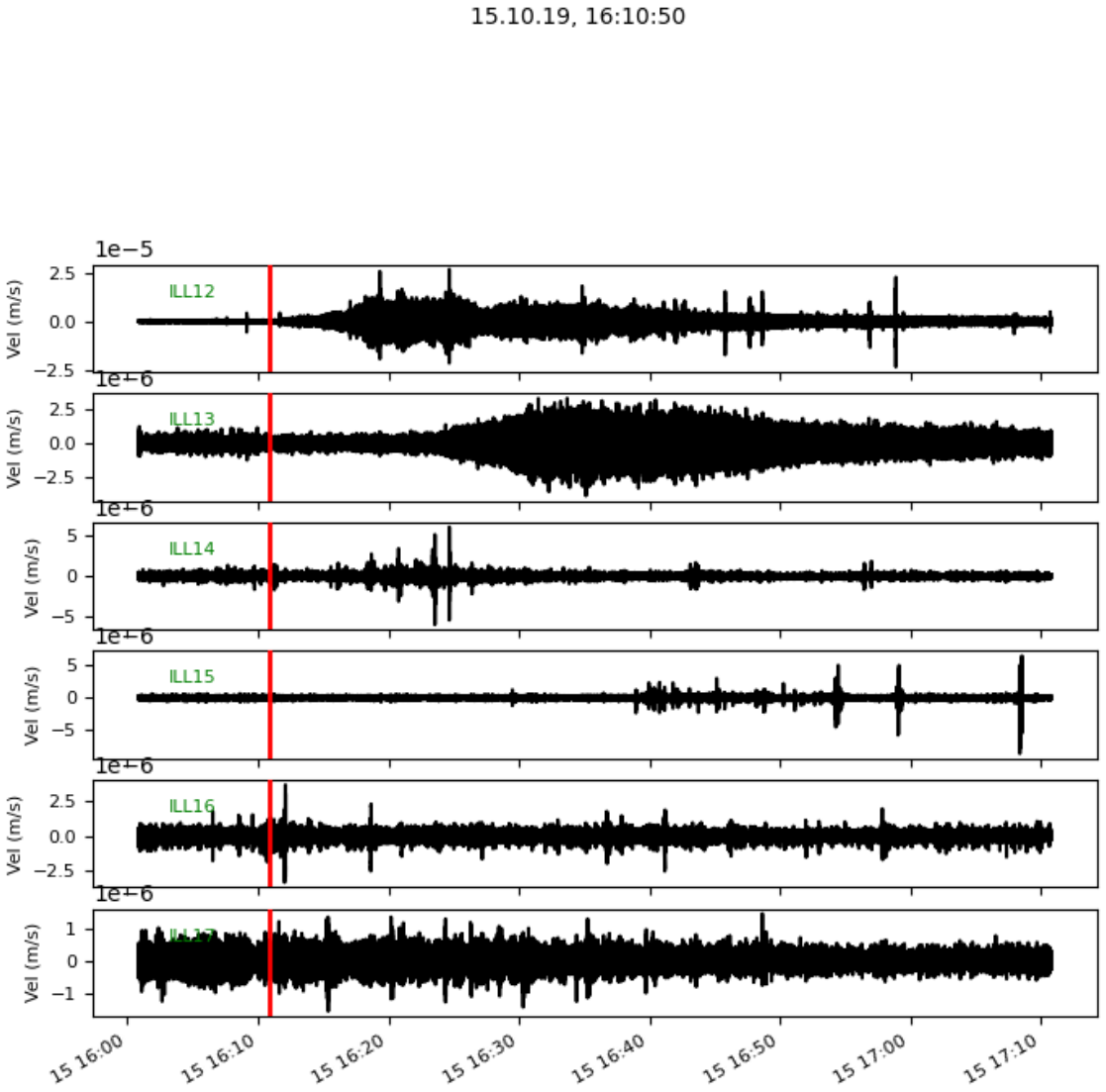

Figure S22. (a) Vertical-component seismograms generated by a debris-flow event on 15 October 2019. The arrival time of the debris flow front at CD1 is marked in red.

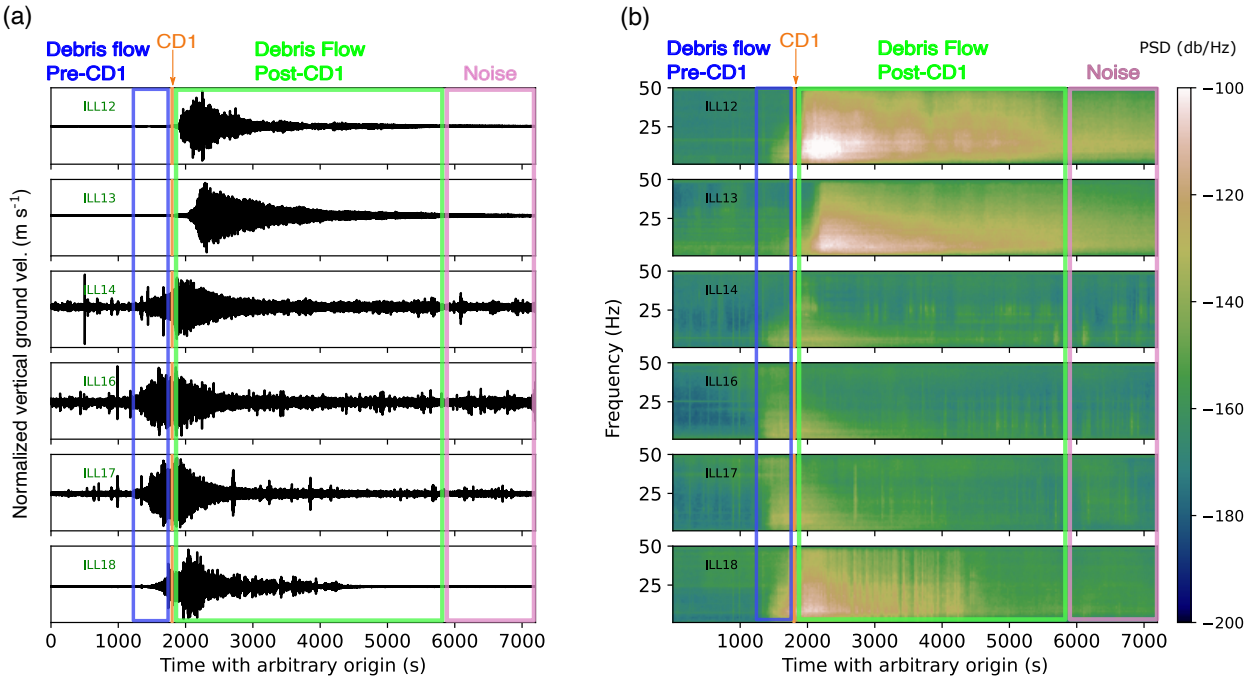

Figure S23. (a) Debris-flow seismograms generated by 21 June 2019 event with vol.= 83,000 m^3 recorded over six stations in the network. Corresponding spectrograms are showed in panel b. Three classes of seismic events used in machine-learning detector are schematically represented with different colours: pre-CD1 (blue), post-CD1 (green) and noise (pink). The arrival time of the debris flow front at CD1 is marked in orange.

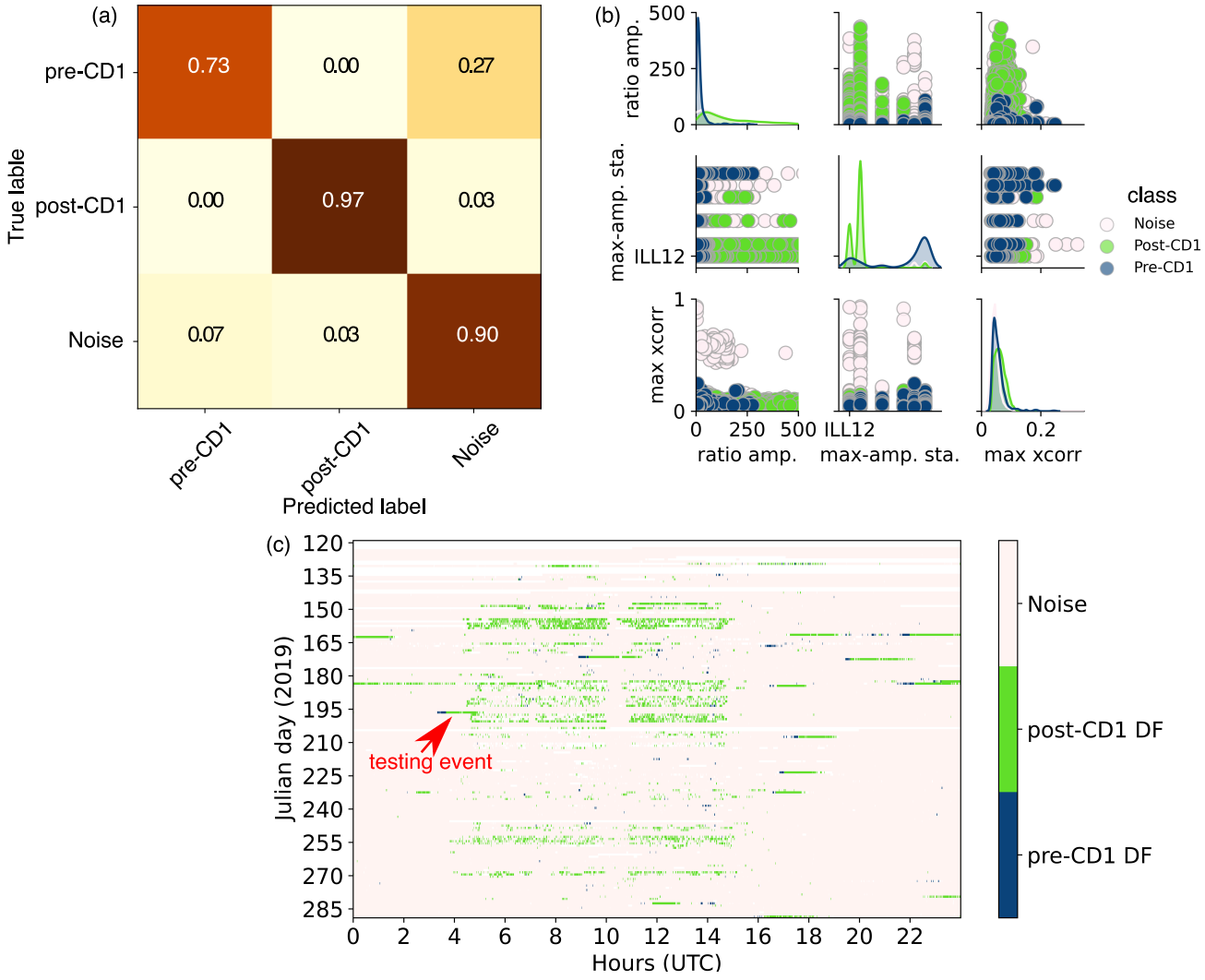

Figure S24. Machine-learning model evaluation for the first iteration. (a) Normalized confusion matrix with true labels as columns and predicted labels as rows. (b) Pairwise relationships of the three most important features. In each subplot, two features are plotted against each other (the same features are plotted on diagonal, which show univariate distribution of features). Features from each class are marked in different colors. (c) Results of the machine-learning detector executed on 2019 continuous data.

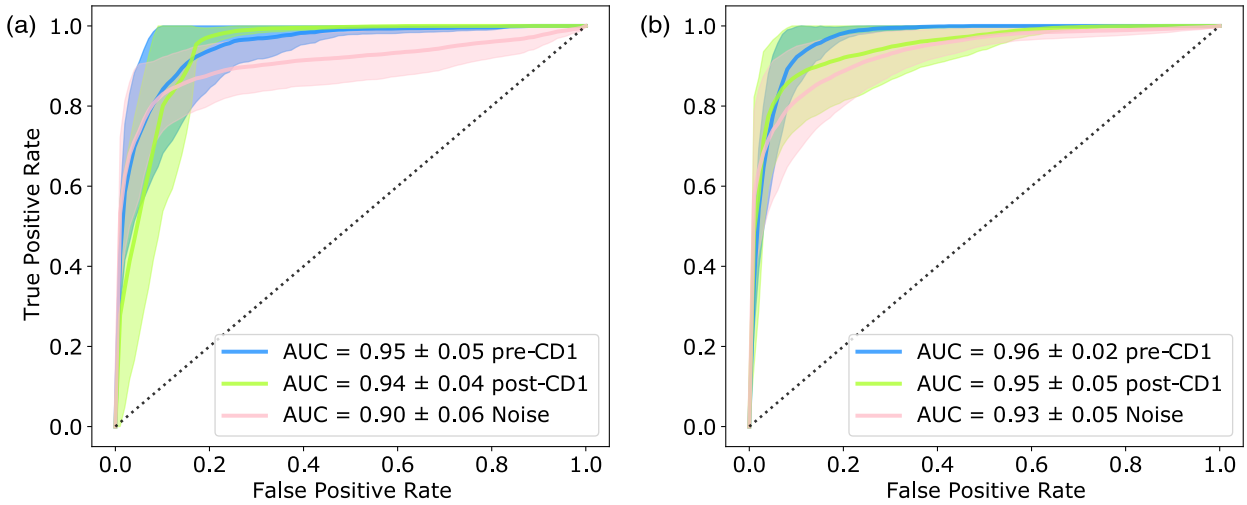

Figure S25. Mean ROC curves for the first (a), and the second iteration (b), calculated using 5-fold cross-validation. The mean ROC curves are marked in solid lines with shaded standard deviations. The true positive rate (TPR) is presented on the y axis and false positive rate (FPR) on the x axis. The area below the curve (AUC) measures model accuracy.

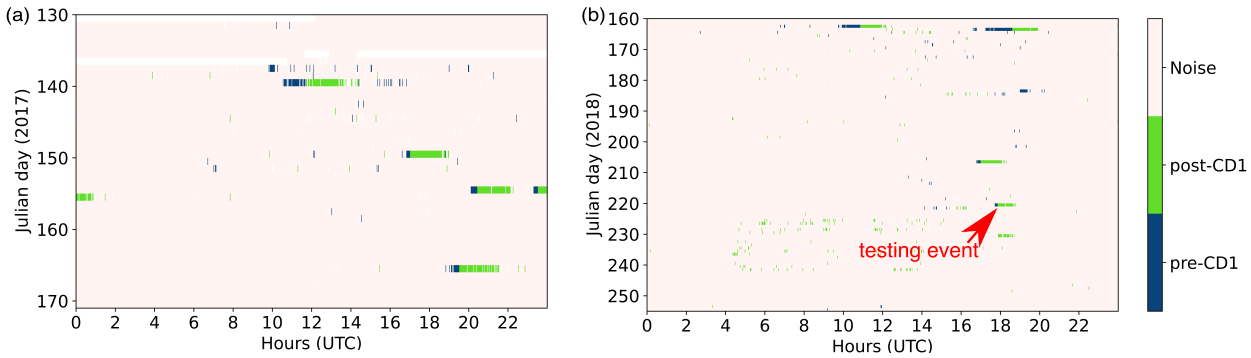

Figure S26. Results of the machine-learning detector run (from the second iteration) over 2017 (a), and 2018 (b) continuous data.

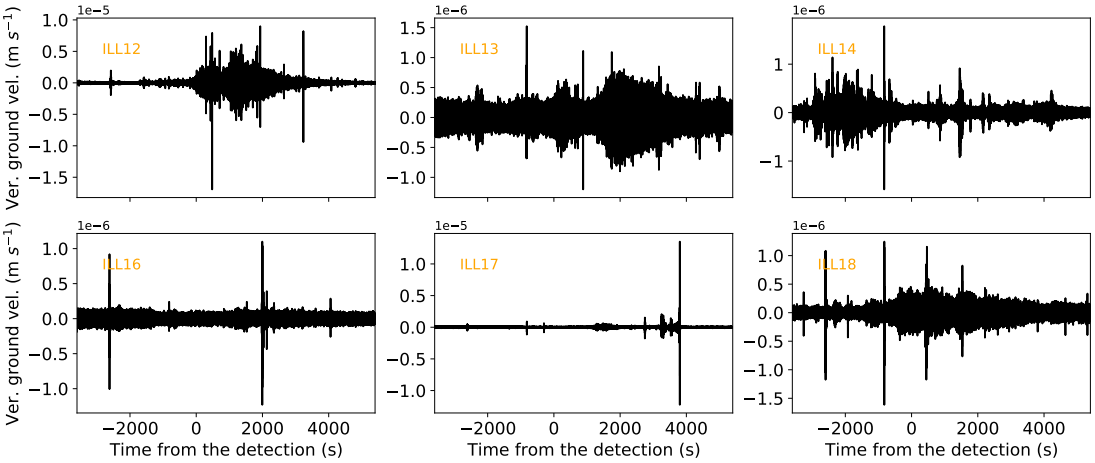

Figure S27. Waveforms generated by a small debris flow event found by the ML-based DF detector, and detected on 18 August 2018, 17:53:20.

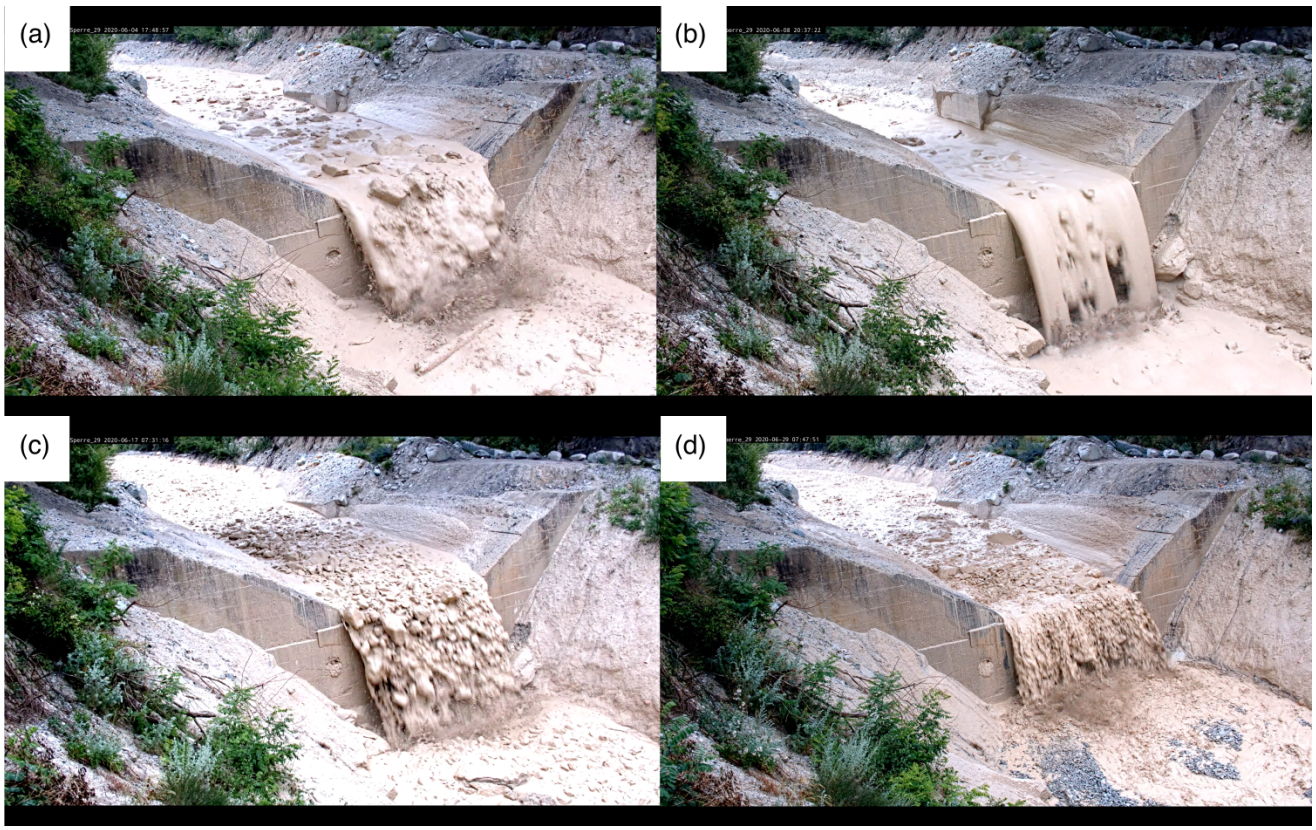

Figure S28. Photos of Illgraben debris-flow events with boulder-rich fronts passing through CD29 detected by the machine-learning detector on 4, 8(2), 17, and 29 June 2020 (Source: WSL).

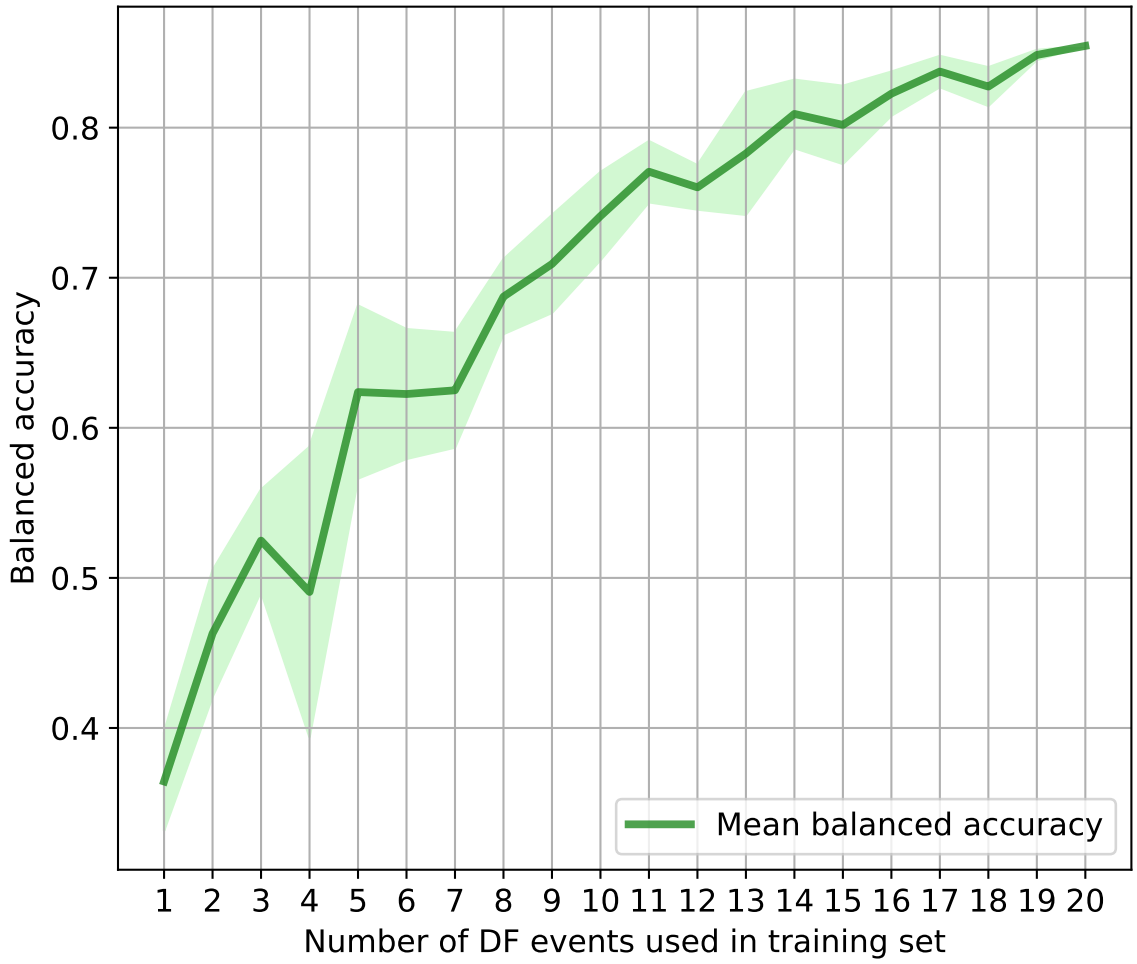

Figure S29. Sensitivity test: balanced accuracy score as a function of n number of debris-flow (DF) events used in training set with cross-validation (5 folds). For each test a subset of n events is chosen from 20 events to train the ML-model and two events marked in green in TableS1 are used as testing set. The mean balanced accuracy is marked in solid green line with shaded standard deviation. The values are normalized between 0 and 1. Even a model trained on a single event gives better classification results than a random guess (for a 3-class classification problem balanced accuracy of a random guess converges towards 1/3). Higher values of balanced accuracy (>0.7) and stable prediction are obtained from n=9 events used in the training set.

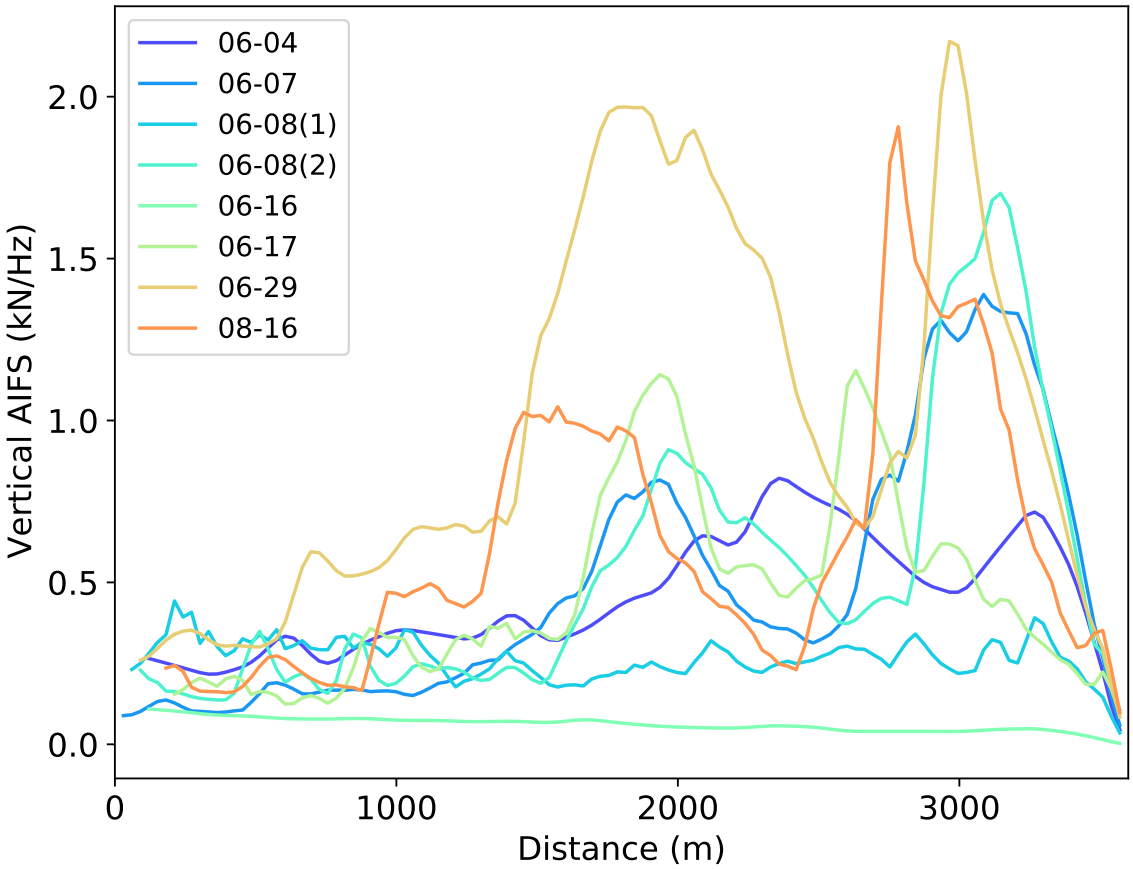

Figure S30. Vertical apparent total impact force spectra (AIFS) for 2020 debris flows.

N°	Date	Arrival time CD1 (UTC)	Vol.(m^3)	Vel.(m s^{-1})	h_{99} (m)
0	2017-05-19	11:41:00	n.a.	n.a.	n.a.
1	2017-05-29	16:58:31	100,000	6.67	4.80
2	2017-06-03 (1)	20:23:07	n.a.	n.a.	n.a.
3	2017-06-03 (2)	23:27:38	25,000	5.10	3.30
4	2017-06-14	19:30:48	n.a.	7.10	3.40
5	2018-06-11	10:46:39	35,000	7.00	3.50
6	2018-06-12	18:29:16	n.a.	n.a.	n.a.
7	2018-07-25	16:56:40	<50,000	4.69	2.00
8	2018-08-08	17:49:25	<100,000	6.70	n.a.
9	2019-06-10 (1)	17:02:51	3,300	0.90	0.64
10	2019-06-10 (2)	22:01:17	6,600	2.38	0.59
11	2019-06-20	09:12:17	n.a.	n.a.	n.a.
12	2019-06-21	19:34:42	83,000	5.60	2.45
13	2019-07-01	23:00:29	78,000	3.80	1.62
14	2019-07-02	22:09:28	39,000	2.50	0.71
15	2019-07-03	16:43:15	n.a.	n.a.	n.a.
16	2019-07-15	03:40:21	16,000	5.00	0.68
17	2019-07-26	17:33:12	64,000	6.97	1.21
18	2019-08-11	17:02:34	53,000	5.56	n.a.
19	2019-08-20	16:40:59	13,000	0.95	0.89
20	2019-10-09	11:45:28	n.a.	n.a.	n.a.
21	2019-10-15	16:10:50	n.a.	n.a.	n.a.

TableS1. Characteristics of 22 debris flow events recorded in 2017, 2018, and 2019. Volume is the integrated sum of discharge over the entire debris-flow wave. Flow velocity is calculated from the travel time between in-channel sensors as described in (Schlunegger et al., 2009). Flow depth h_{99} is the depth where 99 % of the depth values are smaller. n.a. denotes values that were not estimated. Volume and flow depth are estimated at the instrumented wall, CD29. The arrival times at CD1 come mostly from the measurements of a geophone installed at CD1, although the arrival times of events 0 and 2 were estimated based on the ASL results (Walter et al., 2017). Events used in the first iteration in the testing set are marked in orange, and events used in the testing set in the second iteration are marked in green.

N°	Date	Alarm	ILL11 arrivals	Warning time increase (min:s)	Peak ampl. (n° counts)	DF Vel. (m s^{-1})
0	2020-06-04	14:55:08	15:41:51	43:43	250,849	3.2
1	2020-06-07	07:33:28	09:29:32	136:04	117,228	0.7
2	2020-06-08 (1)	13:43:28	15:51:59	128:31	28,383	0.6
3	2020-06-08 (2)	16:35:08	17:59:10	84:02	132,500	0.8
4	2020-06-09	23:35:08	00:55:43(+1d)	90:35	84,089	2.1
5	2020-06-16	20:20:08	23:56:18	216:10	2,486	3.2
6	2020-06-17	03:11:48	04:06:58	55:10	7,468	0.7
7	2020-06-29	04:33:29	05:49:13	75:44	159,970	1.5
8	2020-07-22	15:41:50	n.a.	n.a.	n.a.	n.a.
9	2020-07-28	16:05:10	17:58:55	113:45	2,461	n.a.
10	2020-08-16	21:15:08	23:04:11	109:03	96,098	0.6.
11	2020-08-30	04:54:24	05:52:23	57:59	23,455	n.a.
12	2020-09-01	04:54:24	n.a.	n.a.	n.a.	n.a.

TableS3. Characteristics of 13 alarms in 2020. n.a. denotes values that were not estimated. 22 July 2020 and 01 September 2020 debris flows stopped before ILL11, and 30 August 2020 event had multiple surges which makes difficult reliable debris-flow velocity estimation and AIFS calculation.

Captions for large Table S2.

TableS2. 70 statistical features including waveform, spectral, spectrogram, and network attributes used as input in the machine-learning model. The table is uploaded separately as an excel file.