Anatomy of Continuous Mars SEIS and Pressure Data from Unsupervised Learning

Salma Barkaoui* *et al.*

ABSTRACT

The seismic noise recorded by the Interior Exploration using Seismic Investigations, Geodesy, and Heat Transport (InSight) seismometer (Seismic Experiment for Interior Structure [SEIS]) has a strong daily quasi-periodicity and numerous transient microevents, associated mostly with an active Martian environment with wind bursts, pressure drops, in addition to thermally induced lander and instrument cracks. That noise is far from the Earth's microseismic noise. Quantifying the importance of nonstochasticity and identifying these microevents is mandatory for improving continuous data quality and noise analysis techniques, including autocorrelation. Cataloging these events has so far been made with specific algorithms and operator's visual inspection. We investigate here the continuous data with an unsupervised deep-learning approach built on a deep scattering network. This leads to the successful detection and clustering of these microevents as well as better determination of daily cycles associated with changes in the intensity and color of the background noise. We first provide a description of our approach, and then present the learned clusters followed by a study of their origin and associated physical phenomena. We show that the clustering is robust over several Martian days, showing distinct types of glitches that repeat at a rate of several tens per sol with stable time differences. We show that the clustering and detection efficiency for pressure drops and glitches is comparable to or better than manual or targeted detection techniques proposed to date, noticeably with an unsupervised approach. Finally, we discuss the origin of other clusters found, especially glitch sequences with stable time offsets that might generate artifacts in autocorrelation analyses. We conclude with presenting the potential of unsupervised learning for long-term space mission operations, in particular, for geophysical and environmental observatories.

KEY POINTS

- We apply unsupervised deep learning based on the scattering network to detect microevents in the InSight data.
- We track and cluster transient signals in both SEIS seismometer and pressure records.
- The Deep Scattering Network DSN identifies in seismic records nonstochastic glitches repetitions that can improve the SEIS data interpretations.

Supplemental Material

INTRODUCTION

The Interior Exploration using Seismic Investigations, Geodesy and Heat Transport (InSight) landed on Mars on 26 November 2018 (Banerdt *et al.*, 2020), and deployed the Seismic Experiment for Interior Structure (SEIS) experiment on the ground (Lognonné *et al.*, 2019). It records the Martian pressure with the Auxiliary Payload Sensors Suite (APSS) experiment (Banfield *et al.*, 2018, 2020a), and since February 2019, ground acceleration with SEIS almost continuously, detecting marsquakes (Giardini *et al.*, 2020; Lognonné *et al.*, 2020) and transient atmospheric signals (Garcia *et al.*, 2020; Kenda *et al.*, 2020; Charalambous *et al.*, 2021).

The SEIS background noise (Lognonné *et al.*, 2020; Stutzmann *et al.*, 2021) is much lower in amplitude than the Earth's seismic noise (Peterson, 1993). Because of the surface installation, atmospheric activity and surface temperature drive the noise fluctuations (Lognonné *et al.*, 2020; Charalambous *et al.*, 2021), leading to a strong daily trend and a significant nonstochastic character. This is, for example, illustrated by the relation in occurrences in time of the transient thermally induced microtilts (also denoted glitches) with the SEIS recorded temperature, as already observed by Scholz *et al.*

Full author list and affiliations appear at the end of this article.

*Corresponding author: barkaoui@ipgp.fr

Cite this article as Barkaoui, S., P. Lognonné, T. Kawamura, É. Stutzmann, L. Seydoux, M. V. de Hoop, R. Balestriero, J.-R. Scholz, G. Sainton, M. Plasman, *et al.* (2021). Anatomy of Continuous Mars SEIS and Pressure Data from Unsupervised Learning, *Bull. Seismol. Soc. Am.* **111**, 2964–2981, doi: 10.1785/0120210095

© Seismological Society of America

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

(2020). When not corrected, these glitches lead to artifacts in autocorrelation analyses, as demonstrated by Kim *et al.* (2021). During daytime, other frequent transients events are associated with pressure drops, analyzed and cataloged by Lorenz *et al.* (2020) and Spiga *et al.* (2021) from pressure data analysis and modeled by Lognonné *et al.* (2020), Banerdt *et al.* (2020), and Kenda *et al.* (2020). Mostly above 1 Hz, lander shaking events are also frequent, especially at lander resonance frequencies (Ceylan *et al.*, 2021).

All the required cataloging efforts in identifying these transient signals are time consuming, which might be critical for long-duration operations. In addition, the methods developed by Scholz *et al.* (2020) for glitches cannot identify easily nonstochastic patterns in the signal, such as sequences of glitches with stable offset time. Furthermore, the nonstochasticity can also be related to predictable changes in the color of the noise spectrum, such as those related to the daily variation of the atmospheric turbulences, even if not associated with observable transient signals in the time domain.

This study aims to identify families of signals in the continuous data recorded by SEIS to better understand the structure of the continuous data and its nonstochasticity using artificial intelligence. The analysis presented, here, does not focus on the detection of rare (on the time scale of a sol) seismic events (Clinton *et al.*, 2021), but investigates instrument or local (e.g., lander or instrument-related or environmental) sources, which might either generate single or repeating signals that are similar enough to be clustered. The associated clustering problem (Goodfellow *et al.*, 2016) fits in an unsupervised learning framework in a feature space generated with a deep scattering network (DSN) that has its roots in time–frequency analysis. The DSN (Andén and Mallat, 2014) has been made learnable, which allows our analysis to be fully adapted to unknown conditions, including those of Mars. The original algorithm was developed by Seydoux *et al.* (2020) for the Earth continuous seismic data. We modified it to detect and classify the transient signals in the very broadband (VBB)/SEIS continuous data (InSight Mars SEIS Data Service, 2019) or APSS pressure data (Mora, 2019). This work is the first study to apply deep learning on Martian seismic. During the course of this work, another study was made with atmospheric Curiosity data (Priyadarshini and Puri, 2021).

In this article, we first introduce the deep-learning strategy developed and applied on the Earth. We then applied it to SEIS and pressure data. For SEIS, two analyses are made—the first to identify how much the signal can be clustered and the second to detect repeating signals in the noise. For pressure, only the first step is made.

Finally, we compare the timing of the cluster's events with those reported in the already published catalog (Scholz *et al.*, 2020; Ceylan *et al.*, 2021; Lorenz *et al.*, 2021; Spiga *et al.*, 2021). For single events such as glitches and pressure drops, deep learning provides comparable (for glitches) or better (for pressure drops) detection results than the already published methods. This comparison depends of course on the various thresholds used by all techniques. More importantly, we show that unsupervised machine learning detects nonstochastics features, such as repeating series of glitches, and clusters the noise based on its color (or spectrum). This provides important feedback on the noise structure and a critical check on assumptions in scientific analysis, such as autocorrelations of the continuous data (Deng and Levander, 2020; Kim *et al.*, 2021; Schimmel *et al.*, 2021). This will also help better understand the impact of the atmospheric turbulence on SEIS data.

METHOD

Machine learning in seismology

Machine learning is a powerful approach for statistical data analysis and has had wide-ranging success in various fields (Jordan and Mitchell, 2015) such as seismology (e.g., Jia and Ma, 2017; Kong *et al.*, 2018; Malfante *et al.*, 2018; Hibert *et al.*, 2019; Seydoux *et al.*, 2020; Falcin *et al.*, 2021). Here, we distinguish supervised from unsupervised approaches. In a supervised approach, the algorithm learns the mapping between data samples and labels from labeled, training data.

Lack of labeled data requires unsupervised strategies. Cluster analysis is a common strategy used in unsupervised learning (e.g., Géron, 2019). Even if depending on hyperparameter values, unsupervised learning performs an information-based data analysis (Bergen and Beroza, 2018) not relying on former, human-based labeling of data. It can also reveal new classes, which is out of reach for supervised algorithms trained to recognize already-known classes. In the present study, we focus on noise, adopt the unsupervised strategy, and compare its efficiency on SEIS data with the existing catalogs. For future studies focusing on marsquakes, supervised learning (employing the recurrent scattering neural network) will be tested to automate the manual task performed by the Mars Quake Service (Clinton *et al.*, 2018) and possibly detect more events.

Selecting a relevant and stable representation of waveforms (or waveform features) is critical for the success of clustering, because the temporal representation of waveforms is sensitive to small deformations (Bruna and Mallat, 2013; Andén and Mallat, 2014). In seismic applications, the features have commonly been handcrafted (signal energy, spectral content; see, e.g., Malfante *et al.*, 2018), which implies having a priori knowledge of the data content. Here, we learn the relevant features, which is known as representation learning. The representation is formed by a learnable DSN.

Deep scattering network

A DSN extracts stable representations of continuous data. This network is built layerwise from wavelet transforms (convolutions), taking moduli and pooling, that is, decimation with prior low-pass filtering (see Fig. 1 and Seydoux *et al.*, 2020).

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

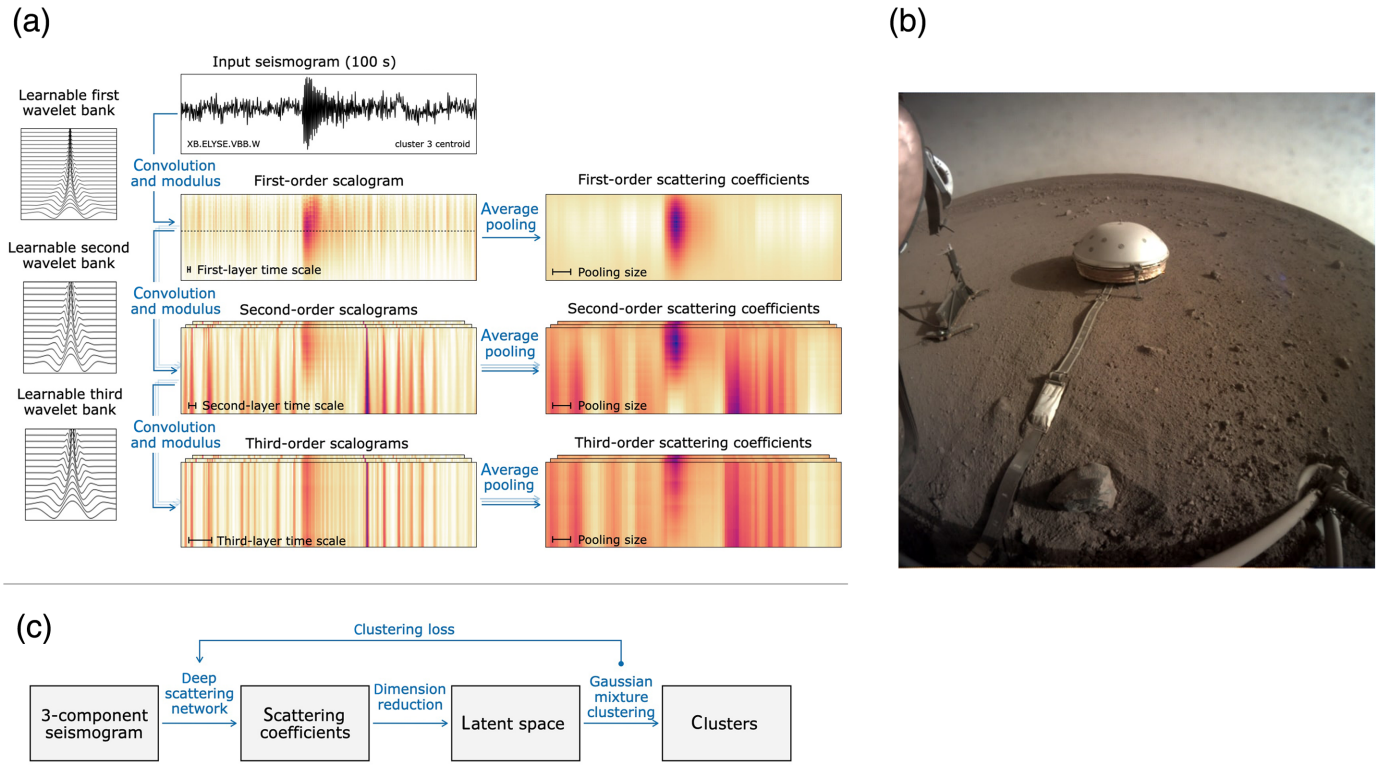


Figure 1. (a) Clustering continuous seismograms with deep scattering network (DSN) and Gaussian mixture model (GMM). DSN—the modulus of convolution between the input seismogram and a first learnable wavelet bank—defines the first-order scalogram. The average pooling in the time dimension of this scalogram provides the first-order scattering coefficients. The second-order scalograms are obtained from each scale of the first-order scalogram, similarly leading to the second-order scattering coefficients with average pooling (Andén and Mallat, 2014). This procedure can be performed at higher orders, and the collection of all-orders scattering coefficients define the scattering representation of the seismic data. The analysis of multiple channels is done by the concatenation of the scattering coefficients obtained for each channel. (b) Image of the SEIS instrument as installed on the ground and protected by the Wind and Thermal Shield, together with the tether joining the sensor to the lander. (c) Clustering workflow as defined in Seydoux *et al.* (2020): the scattering coefficients are extracted from the continuous multicomponent seismograms with a DSN (illustrated in panel (c)). A low-dimensional representation (latent space) of the continuous seismic data is obtained from the first few principal components of the scattering coefficients. The clustering is performed onto the data projected in the latent space with a GMM, allowing to assign a cluster to each segment of signal. The overall strategy optimizes the mother wavelets of each DSN layers to minimize the GMM clustering loss. The color version of this figure is available only in the electronic edition.

Modulus of the convolution $|x * \varphi|$ of a time series $x(t)$ and a wavelet $\varphi(t)$ defines the time-series energy near the center frequency of this wavelet as a function of time. A wavelet transform is the convolution of a time series with a filter bank with various center frequencies (Fig. 1a). The wavelets of a given bank $\varphi_\lambda(t)$ are dilated versions of a mother wavelet $\varphi_0(t)$ with a scaling factor λ such as $\varphi_\lambda(t) = \lambda\varphi_0(t\lambda)$. The frequency range of a wavelet transform is controlled by the number of octaves J, and the frequency resolution is given by the number of wavelets per octave Q. The total number of wavelets in a bank is $F = JQ$.

A wavelet transform defines a time–frequency representation of a signal called a scalogram, as illustrated from a SEIS record in Figure 1a. The evolution of modulating signals or longer trends in the envelopes cannot be captured with a single wavelet transform when several orders of magnitude exist between the time scales, as usually observed in seismology. On the Earth, for instance, earthquakes often produce signals with sharp onsets and broad frequency contents. However, in the same frequency range, we can also observe nonvolcanic tremor signals without clear onsets (e.g., Obara, 2002). Similar observations are made on Mars for the seismic noise recorded by SEIS: In addition, we observe both localized pressure drop signals and continuous wind-generated noise in equal frequency bands (Kenda *et al.*, 2020; Lognonné *et al.*, 2020; Charalambous *et al.*, 2021). This motivates a design of a DSN with three layers (Fig. 1a). Each layer outputs scattering coefficients (see Fig. 1a, right), the order corresponding with the layer index. The scattering coefficients from all orders define the set of features used later in our clustering procedure. The invariance properties of this network promote robust clustering. The dimension of 1D data through the scattering network is summarized in Table 1. The time-pooling factor is adapted at each layer to allow for concatenating the scattering coefficients at all layers.

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

TABLE 1
Computational Dimensions of the Scattering Coefficients

Layer	Description	Dimension	After Pooling Dimension (Scattering Coefficient)	J_ℓ	Q_ℓ	f_ℓ (Nyquist) (Hz)	$F_\ell = J_\ell Q_\ell$
–1	Raw data (100 s, 20 samples per second)	3 channels × 2000 samples	N/A	N/A	N/A	10	N/A
0	Decimated data (100 s, 10 samples per second)	3 channels × 1000 samples	N/A	N/A	N/A	5	N/A
1	First layer	3 channels × F_1 filters × 512 samples	$3 \times F_1 \times 8$	6	6	1.25	36
2	Second layer	3 channels × ($F_1 \times F_2$) filters × 128 samples	$3 \times F_1 \times F_2 \times 8$	7	2	0.312	14
3	Third layer	3 channels × ($F_1 \times F_2 \times F_3$) filters × 32 samples	$3 \times F_1 \times F_2 \times F_3 \times 8$	7	2	0.079	14

Raw 20 samples per second data (layer –1) are first decimated to 10 samples per second for faster computation. A data window of 100 s from three channels contains 1000 samples per channel at the 0 layer (input layer). After convolving the signal with the F_1 filters of the first wavelet bank, the signal is decimated down to 512 samples. Then, successive decimations by four with a fourth-order Butterworth antialias filter are made from one layer to the next one, ending up to 512, 128, and 32 samples for layers 1, 2, and 3, respectively. Finally, we obtain the scattering coefficients with an adapted pooling operation performed on all layers at once. The pooling factor is larger at first layers (from 512 to 8 samples) and lower at last layers (from 32 to 8 samples). We finally end up with a number of eight samples in the time dimension, corresponding to a time resolution of 12.5 s in our case. Note that the dimension of the scattering coefficients grows exponentially with the number of filters per layers (F_ℓ) and the number of layers ℓ. The terms J_ℓ, Q_ℓ, and f_ℓ defined the network hyperparameters used in this study, and defined in both the Hyperparameters and the Comparison between SEIS Glitch Clusters and Glitch Catalog sections.

The three-axis SEIS data are individually transformed, and the scattering coefficients obtained from each component are concatenated to form a set of features for the three components within different time windows. The number of scattering coefficients obtained from a single time window can be significant depending on the number of wavelet filters and scattering orders. This full scattering representation is highly redundant, because the input signals may share similar properties at different frequencies, so there is no need to keep the entire scattering representation. For this reason, we perform a dimension reduction of the scattering coefficients with a projection on the first few principal components (Fig. 1c), corresponding to a low-dimensional representation (or latent space) for which the clustering is applied.

Clustering with Gaussian mixture models

The overall clustering procedure is depicted in Figure 1a. Once transformed into a low-dimensional latent space, the different time windows of seismic data are clustered with a Gaussian mixture model (GMM). The different groups of time windows are ultimately interpreted as clusters of events. As described in Seydoux *et al.* (2020), the mother wavelets at each scattering layer are learned by minimizing the clustering loss of the GMM. Learning the wavelets is indicated by the backpropagation arrow in Figure 1c.

The learning procedure involves two steps. First, we define the value and derivative of the mother wavelet on K knots at each layer of the scattering network. The full wavelet is then interpolated with Hermite cubic splines. The number of knots is low to minimize the number of parameters to learn (for instance, with a wavelet defined on five knots, we need to learn 10 parameters; seven for both amplitude and derivative). A three-layer scattering network with wavelets defined on five knots involves 30 learnable parameters. In the second step, we learn the three mother wavelets that maximize the clustering quality. Following Seydoux *et al.* (2020), we use the ADAM stochastic gradient descent to incrementally converge toward an optimal solution, backpropagating the GMM clustering loss. To prevent trivial solutions from being learned (e.g., zero-valued wavelets), we include a partial reconstruction loss to preserve the input signal's energy across the network. For each layer, the reconstruction loss is the quadratic error between the input signal and the partially reconstructed signal (see Seydoux *et al.*, 2020, for more formal details).

The DSN can be seen as a particular, regularized convolutional neural network (CNN), whereas the output is generated layerwise. In addition, the DSN filters are reminiscent of physically meaningful signal processing, because these involve multiple time and frequency analyses of the input data; this is illustrated in Table 1. This is an advantage over traditional CNNs, which was demonstrated in Andén and Mallat (2014) and Oyallon *et al.* (2017).

Hyperparameters

The overall clustering strategy involves several hyperparameters that define the network architecture, control the time and frequency scales and temporal resolution of the analysis based on the frequency content of the tracked event, and the maximum number of clusters found by the procedure. Here, we define these parameters:

- **The number of scattering layers L**: Andén and Mallat (2014) suggest that two layers are sufficient for audio signals, especially with a broad frequency spectrum. In our case, we use three layers, because the signals of interest span a narrow frequency band, with the second and third layers designed to

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

focus on the envelope oscillations at different frequencies (see the Deep Scattering Network section).

- **The decimation factor**: The main idea of the DSN is to capture the frequency content of the signals' envelope at different frequencies. For seismic data, we assume the envelopes vary smoothly with respect to time. Thus, we decimate the output of the different wavelet transforms at all layers by a given decimation factor. As a consequence, the temporal sampling of each layer reduces as depth increases (see Table 1).
- **The number of octaves** J_ℓ: Determines the frequency range of the wavelet filter bank at layer ℓ, from $\omega_{\min} = \pi 2^{-J_\ell}$ to $\omega_{\max} = \pi$ (in radians). In the first layer, J_1 defines the frequencies analyzed in the seismic data. The parameters J_2 and J_3 control the ranges of time scales seen in the signal envelopes.
- **The number of wavelets per octave** Q_ℓ: Controls the frequency resolution of each scalogram. Following Seydoux *et al.* (2020), we use a large Q in the first layer (dense representation) and a low Q in deeper layers (sparse representation) to maximize the separation between dissimilar events.
- **The number of wavelet knots** K: Controls the number of points to interpolate the wavelets and, therefore, the potential complexity in wavelet shape. Selecting more knots leads to a better description of the signal at reduced computational efficiency. To approximate standard mother wavelets such as the Gabor wavelet at affordable computation cost, we use $K = 5$ knots.
- **The latent space dimension**: Controls the number of components to keep in the dimension reduction with principal component analysis (PCA). There is a trade-off to consider between removing too much information (few principal components) and degrading the GMM clustering quality (too many principal components). Judging this trade-off, in the present study, we selected six components to perform the analysis.

APPLICATION TO SEIS CONTINOUS DATA

We focus on the continuous 20 samples per second VEL channels of the oblique VBB components U, V, and W. The two top panels of Figure 2 show one sol (184) of VBB U raw data and its spectrogram. Sols are Martian days (about 24 hr and 40 min) and are numbered since the landing date. A local mean solar time (LMST) hour is 1/24 of a sol.

As already described by Lognonné *et al.* (2020), Giardini *et al.* (2020), Stutzmann *et al.* (2021), and Ceylan *et al.* (2021), SEIS signals contain highly repetitive patterns in both noise amplitude and frequency of events (that we define here as short duration bursts of energy) from one sol to another. Because of this sol periodicity, we can expect to cover with a limited number of all sols patterns embedded in the noise,

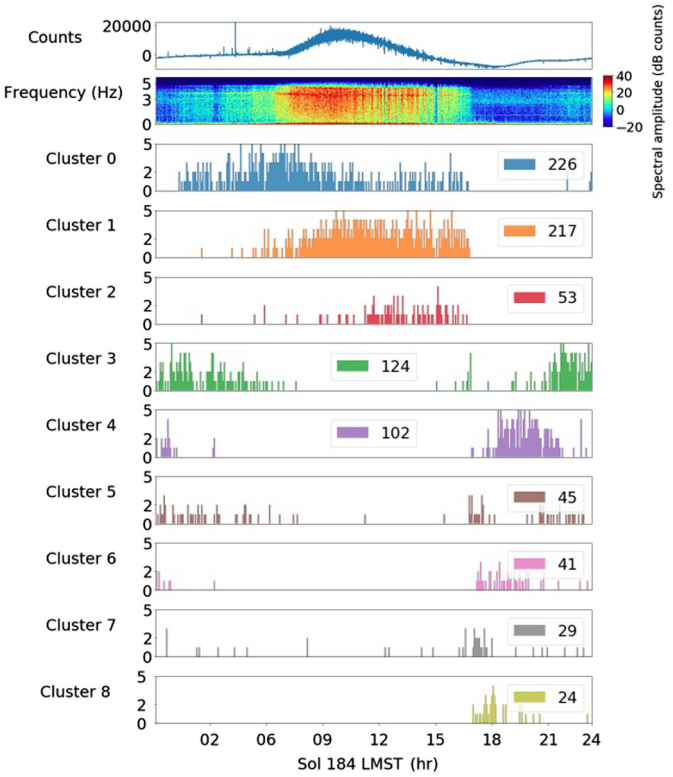


Figure 2. Cluster occurrence frequency on sol 184. The first panel from the top shows the raw very broadband (VBB) U data for Martian sol 184. The second panel is the associated spectrogram, computed with a window of 102.4 s, illustrating the evolution of the frequency content. The other panels show the histograms of cluster activities. They give the number of events occurring in an 11 min window as a function of Local Mean Solar Time (LMST). Numbers on the top right of each panel are the total number of events for that sol. The color version of this figure is available only in the electronic edition.

which will need only a few weeks for signal training. The first is from 3 to 11 June 2019, coinciding with the start of continuous 20 samples per second data. In addition, we selected three other weeks in 2019 (12–18 June, 23–30 June, and 7–14 July) to check that our unsupervised deep-learning algorithm (see the Deep Scattering Network section) can cluster the noise structure regardless of the duration and epoch of the time period. As clustering results are similar for the four weeks, we show only the results for the first week.

To interpret the clustering results, we have also used the temperature data from SEIS and the temperature and pressure data from the APSS experiment (Banfield *et al.*, 2018). See further details in Data and Resources.

Data preprocessing and learning convergence

Minimal preprocessing was performed on the continuous data, limited to (1) a decimation by two due to available graphic processing unit memory limitations and (2) a 0.001 Hz high-pass filtering to remove the very-long-period thermal

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

signal. Therefore, all data are expressed in digital unit, which correspond to about 1.25×10^{-11} m/s at 1 Hz. Our various tests confirmed that the resulting 0.001–5 Hz 10 samples per second three-axis continuous was sufficient for the clustering task.

Table 1 summarizes the choice of hyperparameters and dimension of the feature vectors used in our clustering approach. The continuous data were segmented in 100 s duration intervals viewed as samples without overlap due to processing limitation. Even if DSN has been able to track and cluster events close from the window's borders, future optimization can easily be made with overlapping.

The most frequent events in the data (such as those generated by pressure drops or glitches) do not need many interpolation points for their reconstruction by the mother wavelets in each layer. As mentioned before, we set K to five and the latent space dimension to six. J_ℓ, Q_ℓ values, and Nyquist frequencies, determining the bandwidth of each layer, are provided in Table 1.

Up to 9000 iterations (or epochs) were possible for the learning, but a stopping criterion after 2000 epochs was introduced, based on training's track and reconstruction losses and with awareness of the plateau phenomenon (Seydoux *et al.*, 2020), found generally in the first 500 epochs. We tested our DSN with different depths and concluded that three layers were sufficient to extract information from noisy data and guarantee stability during learning. Although a maximum of 15 clusters was allowed for the clustering, the clustering converged toward the smaller number of nine clusters. This convergence to nine clusters was found for different maxima tested (from 10 to 20) and was also found for training on either one week of data or only one sol.

Clusters and centroid for one sol

Figure 2 shows how the 885 100 s data samples cluster during sol 184—one of the sol of the week-learning period. It provides the number of data samples per hour for each cluster as a function of the LMST at the InSight location. Cluster's occurrence frequency is described in Figure 2. For example, cluster 0 is the most frequent cluster with 226 samples found, whereas cluster 8 is the least frequent one with 24 samples. This already shows that all clusters are associated with specific LMST and are, therefore, thermally triggered or associated with specific temperature and pressure conditions. Clusters 0, 1, and 2 occur during day time, clusters 4, 6, 7, and 8 during early night, and clusters 3 and 5 during late night. This will be confirmed by results for one week, presented later in the discussion and illustrated in the supplemental material to this article.

These data samples are clusterized either due to similar events occurring in the 100 s window or due to similar noise properties (e.g., level or color) in these windows. Before continuing the interpretation, we first briefly review the types of events already identified on SEIS data. These can be divided into two families:

- **Frequent events**: These appear every sol and are either associated with the Martian environment, the lander, and/or the SEIS instrument. First examples are the pressure drops generating ground deformations. See Banerdt *et al.* (2020), Lognonné *et al.* (2020), and Kenda *et al.* (2020) for their signal on SEIS and Banfield *et al.* (2020a) for the pressure signal on APSS. Spiga *et al.* (2021) and Lorenz *et al.* (2020) cataloged them, based on the pressure signal shape. Other wind bursts examples appear through lander vibrations. See among other (Ceylan *et al.*, 2021; Charalambous *et al.*, 2021). Finally, due to thermoelastic stress release and also to pressure drops, glitches are very frequent on all SEIS records, with a visual repeatability sol by sol. See Lognonné *et al.* (2020) and Scholz *et al.* (2020) for more details. They generate microtilts, leading to high-amplitude instrument responses in the raw data.
- **Rare events**: For our analysis, rare events are the seismic events (Giardini *et al.*, 2020). With a rate of a few events per sol (Clinton *et al.*, 2021), they are much less frequent than those listed earlier. In the framework of this article, these will not be captured by clustering. Furthermore, for all the weeks analyzed, no correlation between reported P or S arrival times (as given by InSight Marsquake Service, 2020) and any cluster's event origin time was found. The frequency of event clusters during the seismic events was similar to the one found at the same LMST but for sols without events.

To better quantify common waveform similarities between samples of the same cluster, we extracted for each cluster their centroid waveform and compared them to the best similarity (BS) waveform. The later is by definition the closest to the covariance ellipsoid's center in the scattering manifold, the distance corresponding in the machine-learning vocabulary to the similarity coefficient. For a given cluster, the centroid waveform is the waveform stack of all events of the cluster and is obtained as follows. First, all events from the cluster are aligned with the BS waveform and sorted with increasing correlation coefficient. See Figure 3 for cluster 6. Alignment is made by maximum correlation time lag, and correlation is computed in a 100 s window. The weighted stack is then obtained from the aligned waveforms as follows:

$$X(t) = \sum_{i=0}^{N} \omega_i x_i(t - \tau_i), \tag{1}$$

in which ω_i is the correlation weight, and τ_i is the correlation time lag of the event i with respect to the reference.

For nine clusters, centroid waveforms are shown in Figure 4, whereas the data waveforms are shown in Section 1 of the supplemental material (Figs. S1–S9, respectively).

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

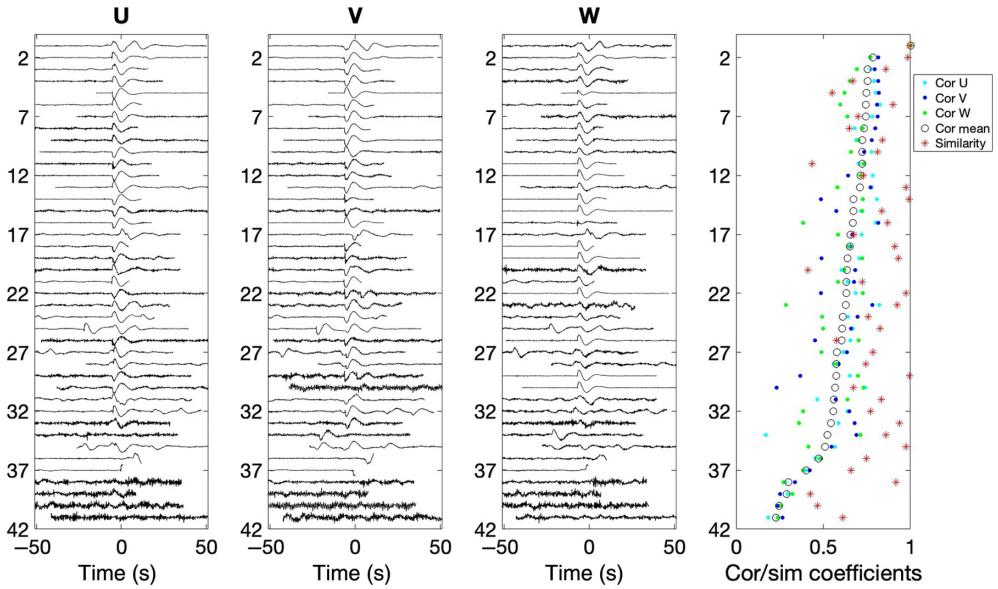


Figure 3. Cluster 6 events aligned on their largest amplitude for the three components. Starting from left to right panels: U, V, and W. The top traces are the reference waveforms and the ranking from top to bottom corresponds to decreasing correlation. All waveforms are normalized with respect to their maximum amplitude, and event start time is at $t = 0$ s. Correlation is defined as the mean value of the three correlations—each obtained for each axis, and it is shown in the right panel as a circle. The three values of correlations for U, V, and W are also shown as colored dots on this right panel, together with similarity with a red stars, plotted to the power 1/6 due to the six dimensions of the manifold. As clustering is done with a mixture of noise level and waveform similarities, correlations and similarities are not correlated. The color version of this figure is available only in the electronic edition.

The BS waveforms for these nine clusters are provided in Figure S10 of Section 1 of the supplemental material. For each cluster, the correlation and similarity values between the BS and centroid waveforms are shown in Figure 5, and support a classification in three families, listed A, B, and C.

Clusters A (numbered 6, 7, 8) are characterized by both a high correlation and a high similarity between the BS waveform and the centroid. Although having no significant similarity, the B-type centroids (numbered 3, 4, 5) conserve a high correlation with the BS waveform for clusters 4 and 5. Correlation drops to about 0.5 for cluster 3. Clusters C are those numbered 0, 1, 2. They have a very low correlation but with a significant similarity for 1 and 2.

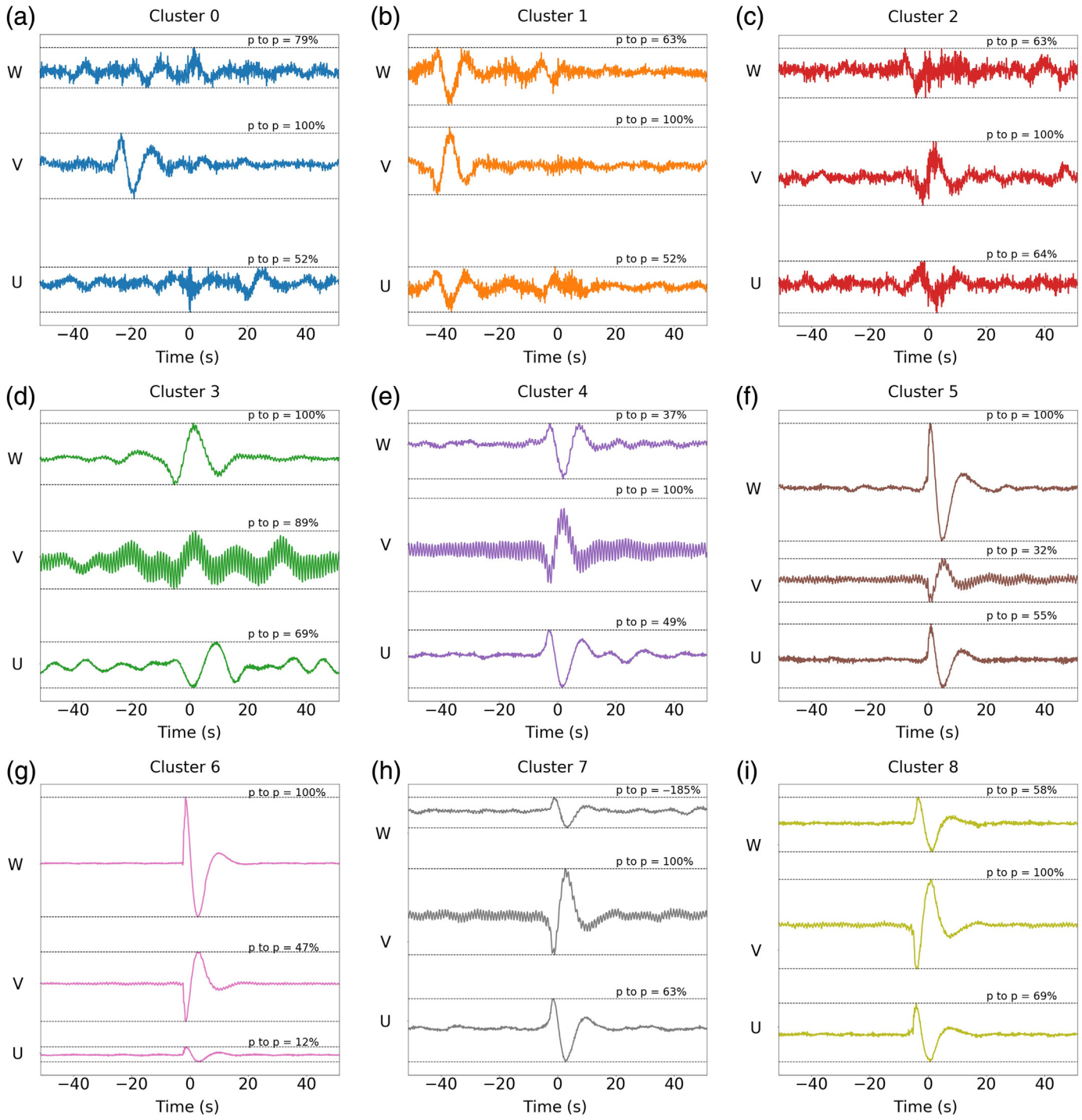
Both centroids and BS waveforms of families A and B are characterized by VBB glitches, as identified by Lognonné *et al.* (2020) and Scholz *et al.* (2020). They appear on these raw data as the instrument response to an acceleration step. The glitches are less clear for cluster 3 BS waveform and appear mostly on the corresponding centroid. The centroids and BS waveforms of family C are all occurring during windy activity, either during the day regime for clusters 1 and 2 with their larger spectral amplitudes or in the second part of the night, continuing to the morning for cluster 0. We will see later that these clusters are associated either with pressure drops or wind burst, generating in both cases SEIS signal.

The clustering is, however, not made only on the waveform similarities, but also on their spectral properties, including spectra color and ratio between high-frequency and low-frequency amplitudes. This explains why these families have several clusters and not only one, and is illustrated by the spectra of the nine centroid waveforms shown in Figure 6 for the V component and in Section 1 of the supplemental material (Fig. S11) for the associated BS spectra.

The large differences in the ratio between low-frequency and high-frequency amplitudes for families A and B confirm differences between five clusters. For example, centroid's spectra of clusters 3, 4, and 7 are comparable above 1 Hz but have growing amplitudes between 0.1 and 0.2 Hz, whereas the high-frequency amplitudes and color of centroid's spectra 6, 8, and 5 are red, white, and blue, respectively. Likely, the clustering is also sensitive to the 1 Hz tick noise (and associated 2–3 Hz harmonics) that acts as an amplitude reference. While being an artifact related to the interference of the house keeping data inrush current on the VBB feedback analog signal, its amplitude is indeed stable over time (Ceylan *et al.*, 2021). Another interesting feature is the 2.4 Hz resonance peak, proposed as a ground resonance by Giardini *et al.* (2020). It has a very comparable amplitude for the three low-noise clusters (3, 4, 7), which confirms the stability of its amplitude. Clusters 3 and 4–8 all occur during the night. But clusters 5 and 6 have a much larger noise level, covering the 2.4 Hz resonance, whereas cluster 8 has the highest background noise of both families A and B. For the family C, spectra are on the other hand much more comparable after scaling. However, the 1 Hz tick noise allows to understand that clusters 0, 2, 1 are associated with growing noise level, the tick noise being for example observed on both the BS waveform and the centroid for cluster 0, and absent for cluster 1. We will discuss this family later, in more detail, after having compared these clusters with pressure drop statistics.

Other VBB components provide similar results (Fig. S12 for W and Fig. S13 for U) but with amplitude differences of the 1, 2, 3, and 2.4 Hz peaks.

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user



Clustering stability

The stability of the clustering is tested with training on different sols selected from the middle of northern spring to the middle of northern summer. Between 8 and 10 clusters are commonly identified. Four clusters (1, 2, 3, 7) are stable over time, representing families A, B, and C for clusters 7, 3, and 1–2, respectively. A similarity larger than 95% between events from different sols but from that same stable cluster is found. The centroid spectra are shown in Figure 7 for all these sols.

Figure 4. Centroid waveforms of the nine clusters. Panels (a), (b), (c), (d), (e), (f), (g), (h) and (i) show cluster 0, 1, 2, 3, 4, 5, 6, 7 and 8, respectively. For each cluster, the waveforms of the three components U, V, and W are plotted together with the corresponding peak to peak percentage computed with respect to the component with the maximum amplitude. This provides the relative amplitude of the three components—a feature taken into account in the clustering process. Amplitudes are normalized by the mean-squared norm L_2 applied on the three axes. The cluster event starts at 0 s (centered using the same procedure explained in Fig. 3). The color version of this figure is available only in the electronic edition.

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

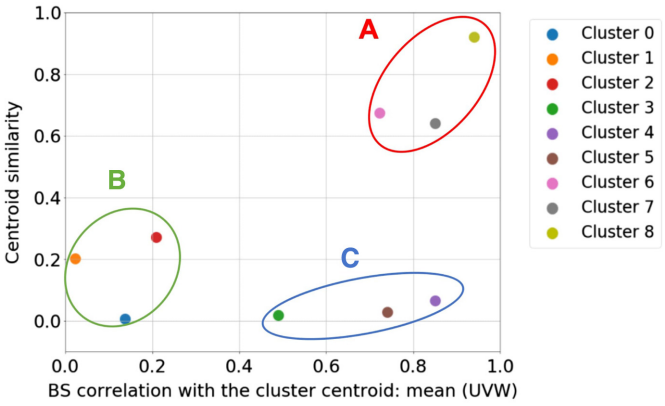


Figure 5. Cluster's centroid similarity distribution in function of its correlation with the best similarity (BS) event for each cluster. This figure highlights three families A, B, and C. Family A in red: during the clustering procedure, the waveform shape is the dominant feature. Family B in green: the background noise is the dominant feature during the clustering procedure. It is more related to the response of the external Martian sources in the seismic data, such as the background noise generated by pressure drops (clusters 0 and 1) or wind burst (cluster 3). Family C in blue: the waveform is not the only main feature used in the clustering (e.g., the background noise, the relative amplitude, and so on). The color version of this figure is available only in the electronic edition.

The two clusters with the lowest spectral amplitudes have clear 1 Hz (and overtones) tick noise peaks plus the 2.4 Hz resonance, and correspond to clusters 7 and 3 from Figures 4 and 6. As already said, they are associated with glitches occurring during the early and late night, respectively, and therefore with decreasing cooling rate, which might explain the smaller glitch amplitude of cluster 3 as compared with 7, as illustrated by the almost 20 dB differences in Figure 6. The two other clusters are those with the largest spectral amplitudes, have large excitation levels for the lander resonances above 3 Hz, and correspond to clusters 1 and 2 from Figures 4 and 6. These two clusters appear during the day or the second half of the night, respectively. We will later see that cluster 1 is associated with pressure drops, whereas cluster 2 is associated with bursts of energy generating ringing associated with the lander resonances. This ringing is also found when examining the spectrograms of individual events.

These clustering stabilities from sol to sol are likely related to both the waveform and spectra similarities, including for spectra resonances. The invariance properties of the DSN contribute to these stabilities.

COMPARISON BETWEEN SEIS GLITCH CLUSTERS AND GLITCH CATALOG

Families A and B are characterized by powerful (for clusters 4–8) and weak (for cluster 3) glitches. Glitches from clusters 4 and 8 are dominating the signal in the time domain (Fig. 2). During sol 184, they sum up to 364 events. These glitches are frequent enough to occur during marsquakes and perturb the recorded marsquakes signals (Lognonné *et al.*, 2020).

Cataloging the glitches and possibly removing them was an early effort (Lognonné *et al.*, 2020) and is detailed by Scholz *et al.* (2020). We compare here the detection timing of the DSN with those provided by the more classical glitch detection techniques. On sol 184 (3 June 2019), the number of glitches reported in these catalogs ranges from 50 to more than 200, depending on the detection algorithm and threshold parameters, and we use for comparison a catalog of 127 glitches obtained for a middle threshold value (see folder GlitchListing of the e-supplemental zip file for the associated listing and further details in Scholz *et al.*, 2020). The histogram of the results is shown in Figure 8. The zero-centered distribution confirms the matching between the two approaches. In fact, DSN retrieves 117 glitches from the catalog out of 127 with a timing error smaller than 2 s, which corresponds to 92% of the cataloged glitches and therefore 8% of false negative.

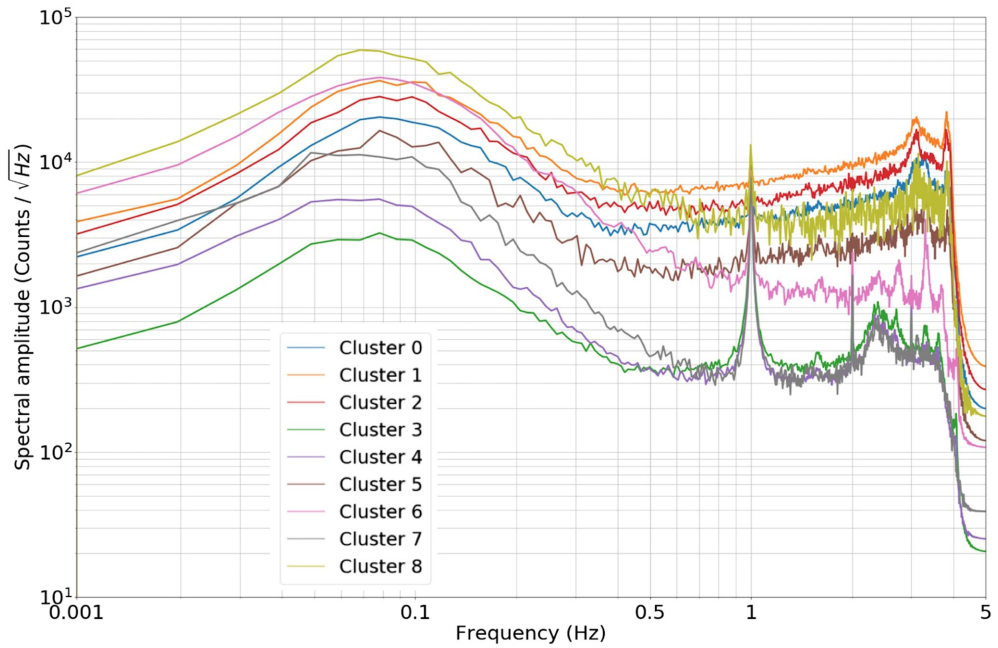


Figure 6. Amplitude spectral density of the nine cluster centroid waveforms for the V component. The color version of this figure is available only in the electronic edition.

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

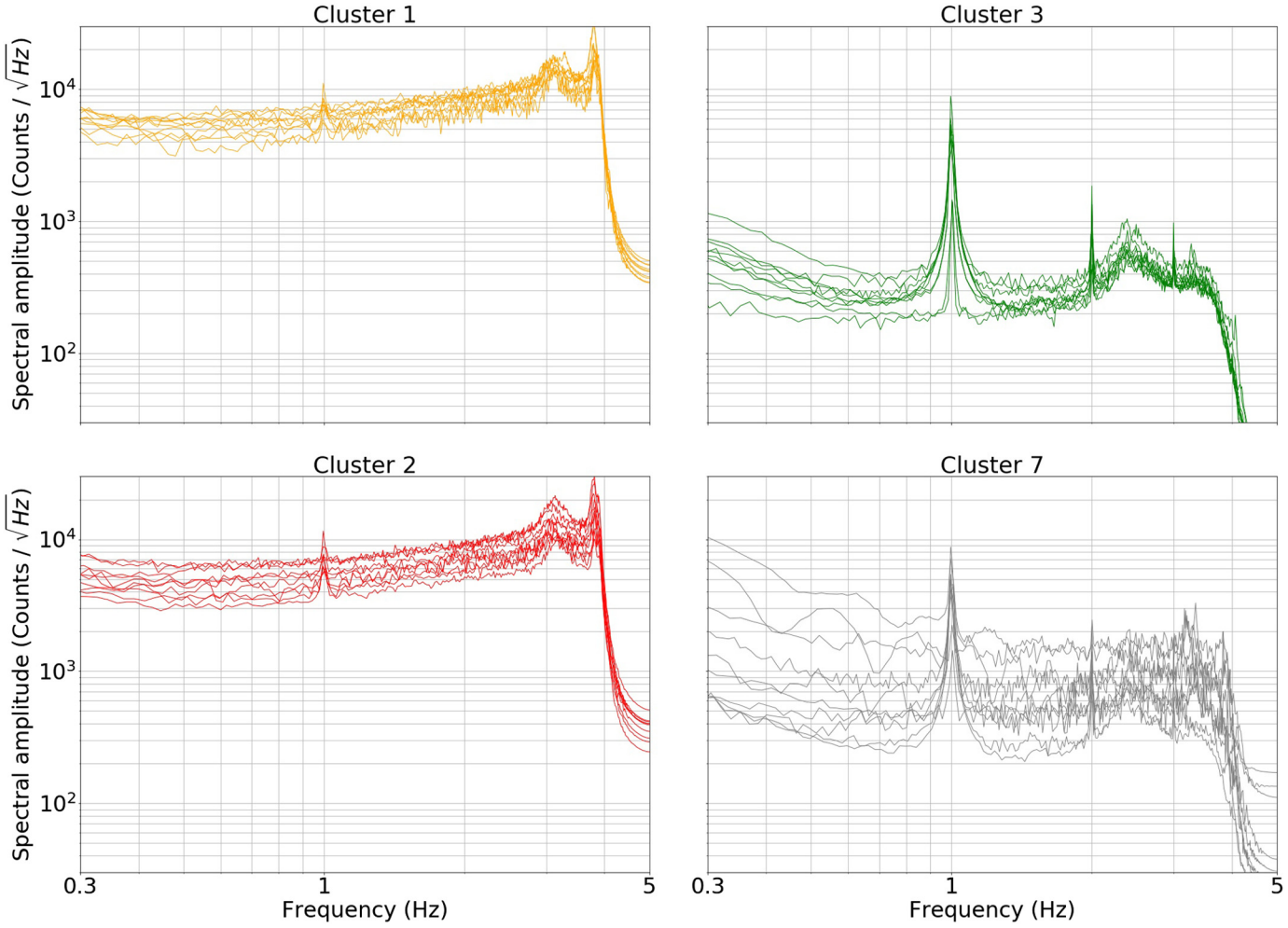


Figure 7. Stable cluster spectrum. Each plot shows the centroid spectra of the clusters 1, 2, 3, and 7, as obtained from learning on the following Martian sols: sol 193, sol 203, sol 213, sol 223, sol 234, sol 243, sol 253, sol 363, sol 372, and sol 393. These cluster's events have 95% similarity between each others. The color version of this figure is available only in the electronic edition.

When looking on the similarity coefficient, 170 of these events have a very low similarity coefficient, smaller than 10^{-8}, and can therefore be either false detection or weak detection. The total of glitches detected by DSN with high similarity is therefore 194 (152% of the cataloged glitches). These additional events are likely glitches with amplitude lower than the catalog threshold.

Cluster polarizations

Polarization provides additional information on the origin of clusters, and we determined the azimuth and dip for all events as follows. We first high-pass filter all components with a 0.1 Hz cutoff. We then normalized all components with the transfer function of the U component (a correction made by the ratio of the U transfer function with the transfer function of the component) and then rotated the data to obtain the N, E, and Z components (based on SEED dataless information). The event's azimuth and dip are then computed using the Ppol software (Fontaine *et al.*, 2009; Scholz, 2017). For all clusters except 1 and 3, we use a ±5 s time window around the event center, as defined by the correlation with the centroid. For clusters 3 and 1, a window of ±40 and ±20 s, respectively, is used.

Figure 9 shows the back azimuths and dips of all events from clusters 1, 3, and 8, whereas those of the other clusters are only discussed later. Clusters 0 and 1 have subvertical dip, as observed by Kenda *et al.* (2020) for pressure drops. Cluster 2 has a horizontal dip but with a relatively large azimuth scatter, even if some clustering toward north is observed with large amplitude. We retrieve here observations from Stutzmann *et al.* (2021), Charalambous *et al.* (2021), and an interpretation based on wind-induced lander noise. Cluster 8 is typical for SEIS glitches (such as clusters 4–8). Its dip departure from horizontal is very small (as for cluster 7), and its azimuth points to the north (as for clusters 4, 6, and 7) and seems related to longitudinal microtilts with respect to the tether. The not shown cluster 5 has on its side an azimuth close to orthogonal from the Load Shunt Assembly (LSA)/tether in line with their peak occurrence during the cooling of the early night

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

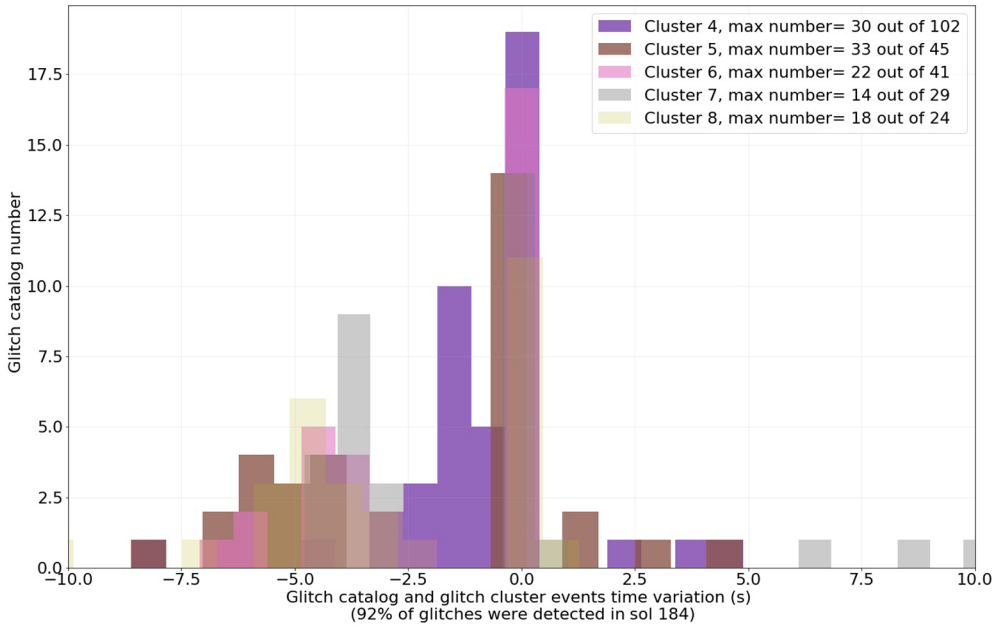


Figure 8. Glitch detection timing. Histogram showing the time difference of glitches cataloged by Scholz *et al.* (2020) and the glitch clusters. Only differences smaller than 10 s are shown in sol 184. In total, 127 glitches are listed in catalog. In the figure's legend, we mention the total number of glitches for each cluster out of its total event's number. The color version of this figure is available only in the electronic edition.

(Fig. 2), whereas those with smaller amplitude have still a subhorizontal dip. Cluster 3 has more vertical component but a relatively stable azimuth toward the lander and correspond to either low-amplitude glitches or internal SEIS glitches, known as source of vertical signal (Scholz *et al.*, 2020).

CORRELATION OF SEIS CLUSTERS WITH TEMPERATURE AND SOL QUASI-PERIODICITY

Both the dependency on LMST and the clustering stability over sols suggest that our clusters are driven by daily temperature variations. This is confirmed by Figure 10, which shows for five days, the number of events per hour together with the outside temperature, the scientific temperature, and the VBB temperature. The latter two have a delay related to 3 and 5.5 hr time constant of the VBB enclosure and wind thermal shield (Lognonné *et al.*, 2019). None of these temperature data were input of the learning.

First, we discuss the correlation of occurrence rate with temperature starting from midnight (0:00 LMST). During the nighttime cooling and associated decrease of atmospheric temperature, cluster 3 (green) is dominant, with a low-noise level (Fig. 6) and a clear 2.4 Hz resonance. Glitches from cluster 5 are also present in background, especially when noise levels are larger.

Cluster 0 (blue), with a much larger noise above 1 Hz, hiding the 2.4 Hz resonance (Fig. 6), increases in intensity when night winds rise up; and this cluster replaces cluster 3 and becomes the most abundant in the early morning. Cluster 1 (orange) starts when the temperature increases after sunrise and reaches its maximum occurrence rate in the late morning. During the daytime's atmospheric activity, this cluster dominates. The occurrence rate of cluster 2 (red) is increasing in the late morning and has a plateau between 12:00 and 17:00 LMST. Clusters 0, 2, and 1 have increasing background noise levels (Fig. 6). This behavior is related to afternoon wind bursts, the signature of which is also found in the large variations in atmospheric temperature. Clusters 0 and 1 have both long-period events and short-period events. They seem to be associated with the conjunction of pressure drops and wind burst. Cluster 2 is mostly a high-frequency event and is likely associated with wind bursts.

Thermal glitches, identified with clusters 3–8 are mostly occurring for 4–8 during the cooling phase of the late afternoon, reaching maximum activity between 18:00 and 20:00 LMST and a diffuse activity all the night. Cluster 3 glitches are on their side observed almost all the night.

For all clusters, the sol-by-sol repetition suggests that the clustering is able to capture the waveform and noise differences of these events, and that these are directly related to LMST and/or to a physical processes depending on LMST. Although the sequence is found every sol, differences in amplitude and in start/end times for each cluster are observed in sol-to-sol comparisons, as shown in Figure 10. Climatic variation will need further analysis, but we can expect these to generate mostly a drift of the occurrence LMST of the temperature correlated clusters.

CHARACTERIZATION OF SEIS MULTIGLITCHES

Our approach detected another type of events in the data: these are glitches appearing in pairs, repeating with a stable time offset within the event window, and also repeating as the previous one every sol. We refer to these as doublet and, more generally, tuplet glitches.

To identify these, we simply increased the basic time window from 100 to 1200 s. We now obtained six clusters. For a test made on two weeks of data in April 2019, the learning process converged after 8000 epochs and detected sequences of tuplet glitches with quasi-periodic recurrence times of 83,

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

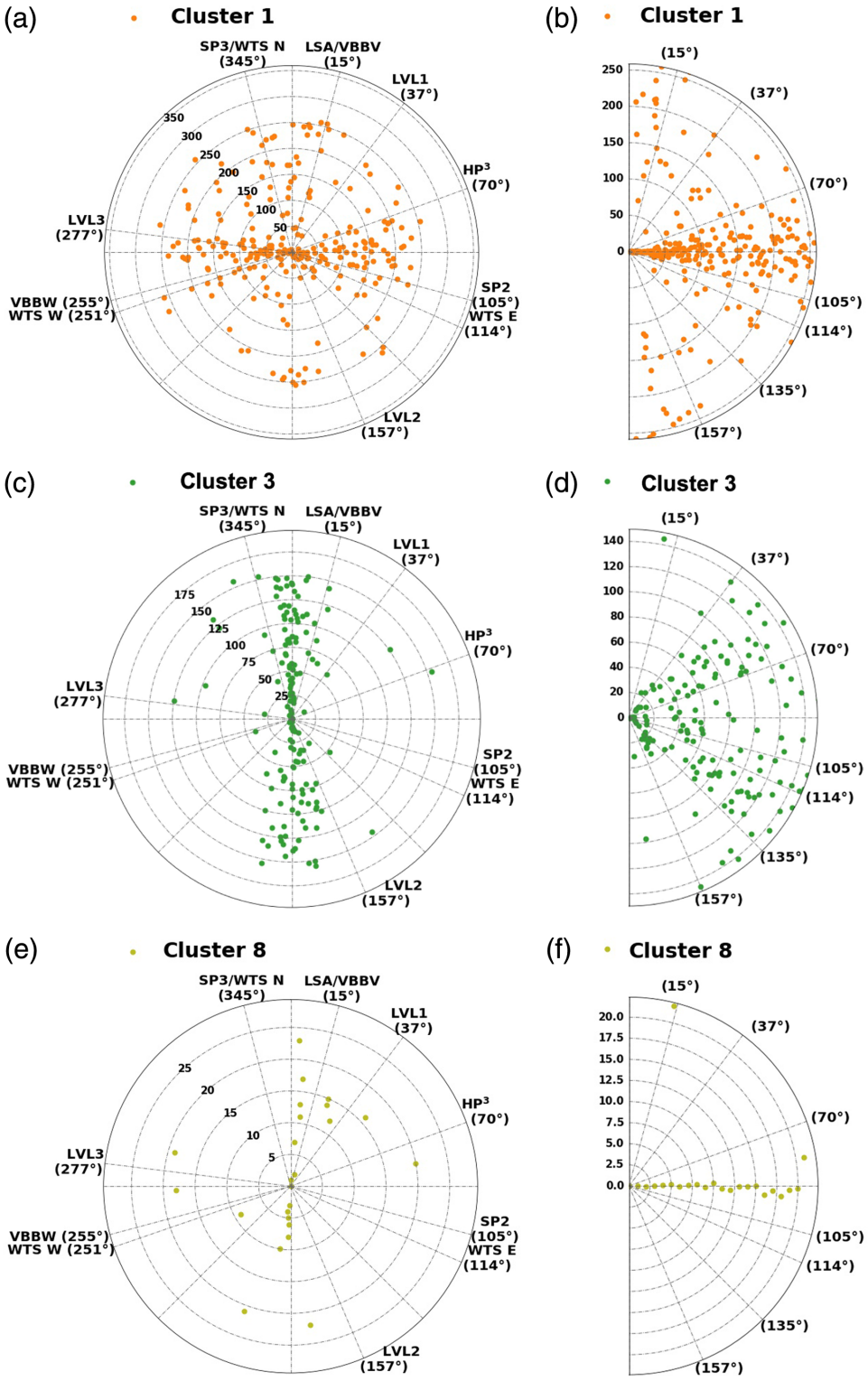


Figure 9. Back azimuths and dip of the events of clusters 1, 2, and 8 recorded on sol 183. Figures (a), (c), (e) show the back azimuth of the clusters 1, 2, and 8, respectively. The first three plots on the left show the back azimuth of the clusters 1, 2, and 8. For each cluster, the corresponding events are plotted with points as a function of their back azimuth from 0° to 360° along the outer circle and as a function of their index along the radius. The events are assumed to be linearly polarized. The inner dashed circles give the event indices. Events in the center have the BS with the cluster centroid. Note that these numbers are different for each cluster. Azimuths related to the Seismic Experiment for Interior Structure (SEIS) instrument feature are given on the outer circle and include: the sensitivity azimuth of the VBB (U, V, W) and SP sensors (SP2, SP3), the feet of the Leveling System (LVL) (LVL1-2-3), the feet of the Wind Thermal Shield subsystem (WTS E, W, N), and the Load Shunt Assembly (LSA). SP1 is not listed, because this is the vertical-component SP sensor. Figures (b), (d), and (f) illustrate the dip of 1, 2, and 8 clusters, respectively, following the same representation as the azimuth. The color version of this figure is available only in the electronic edition.

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

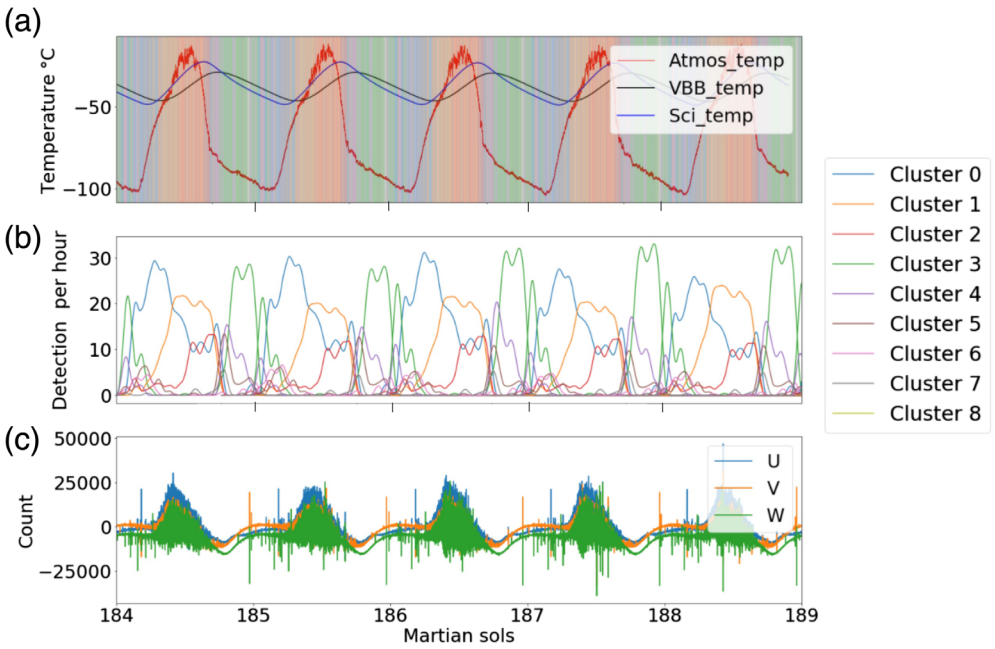


Figure 10. Temperature correlation. (a) Temperature (in Celsius) recorded at three different locations—on the lander (outside temperature in red), under seismometer thermal shielding (scientific temperature in blue), and next to the VBB U sensors (VBB temperature in black). For each local hour, the color in the background corresponds to the cluster that has the maximum number of detection, as shown in the bottom plot. (b) Number of detection per hour for the nine clusters. Each color line corresponds to one cluster with the same color code, as shown in Figure 7 (0, blue; 1, orange; 2, green; 3, red; 4, purple; 5, brown; 6, pink; 7, gray; 8, gold). Both plots are a function of local time in sols, from sols 183 to 189. (c) U, V, and W raw data presented from sols 184 to 189. The color version of this figure is available only in the electronic edition.

208, 280, 295, 327, and 374 s. We provide further details in Figures S2 and S14. For a two-week period in June 2019 (from sols 183 to 197), the quasi-period recurrence times found are 91, 198, 208, 218, 280, 368, and 385 s. We provide further details in Figures S2 and S15. Although the recurrence times vary slightly, they concentrate in the ranges 80–90, 195–220, 280, and 365–385 s. Figure 11 shows examples of glitches repeating with about 368 s delay, with 574 events found during a long sequence, from sols 184 to 198, and the aligned weighted stack of all these events. A mean rate of about 28 events per sol is therefore found for these events, and the root mean square of the 368 s time offset is only 2.3 s. An interesting finding is that these signals are not present during some periods of the night, roughly between 1 and 7 LMST. Further works, outside the scope of this article, will be necessary to understand if they are instrument or lander generated.

Clearly, the occurrence of these glitches cannot be assumed as random, and this might impact analysis assuming stochastic ambient noise. In this regard, the timing of mantle and core signals proposed by Deng and Levander (2020) from autocorrelation of the raw (and nondeglitched) SEIS data are coinciding with the 280 and 380 s delays found in doublet clusters. An in-depth analysis of the impact of glitches has been made by Kim *et al.* (2021), confirming that these signals must be handled with care for any geophysical interpretation. To our knowledge, the clustering analysis proposed here is at this time the only proposed method enabling the identification in the SEIS data of doublets and, generally speaking, of multiglitches with nonstochastic timing. Furthermore, it allows us to find periods in the data, during which these signals disappear, and which might be more adequate for autocorrelation analysis.

PRESSURE DROP CLUSTERS AND CATALOG

We now analyze the correlations between SEIS microevent clusters and pressure drops induced by atmospheric vortices, very frequent in Elysium Planitia (Banfield *et al.*, 2020a) and more generally the efficiency of clustering for pressure signals. We do it first with a clustering analysis of the pressure signal alone and then compare the timing of pressure clusters and the SEIS clusters obtained in the previous sections with pressure drop catalogs (Lorenz *et al.*, 2020; Spiga *et al.*, 2021).

Cataloging pressure drops with DSN

For the first step, we use only two layers in the DSN structure and limit the PCA to three components instead of six. This focuses on the most frequent and clearest events. The 10 samples per second calibrated pressure data (Banfield *et al.*, 2020b) were used. The training was made with pressure data starting on 2 June 2019 00:00:00 UTC and ending on 11 June 2019 00:00:00 UTC, covering sols 182–191.

We obtained seven clusters, described in Section 4 of supplemental material, and focussed here on the three clusters clearly associated with pressure drops. Their stacks are shown in Figure 12, and these clusters differ by their frequency content and waveform shape. The less frequent events from cluster 2 have, for example, a more pronounced peak shape than those of clusters 0 and 1.

Pressure drop catalog correlation with the pressure clusters

To confirm the link with pressure drops, we use the published pressure drop catalog. It reports 278 pressure drops larger than

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

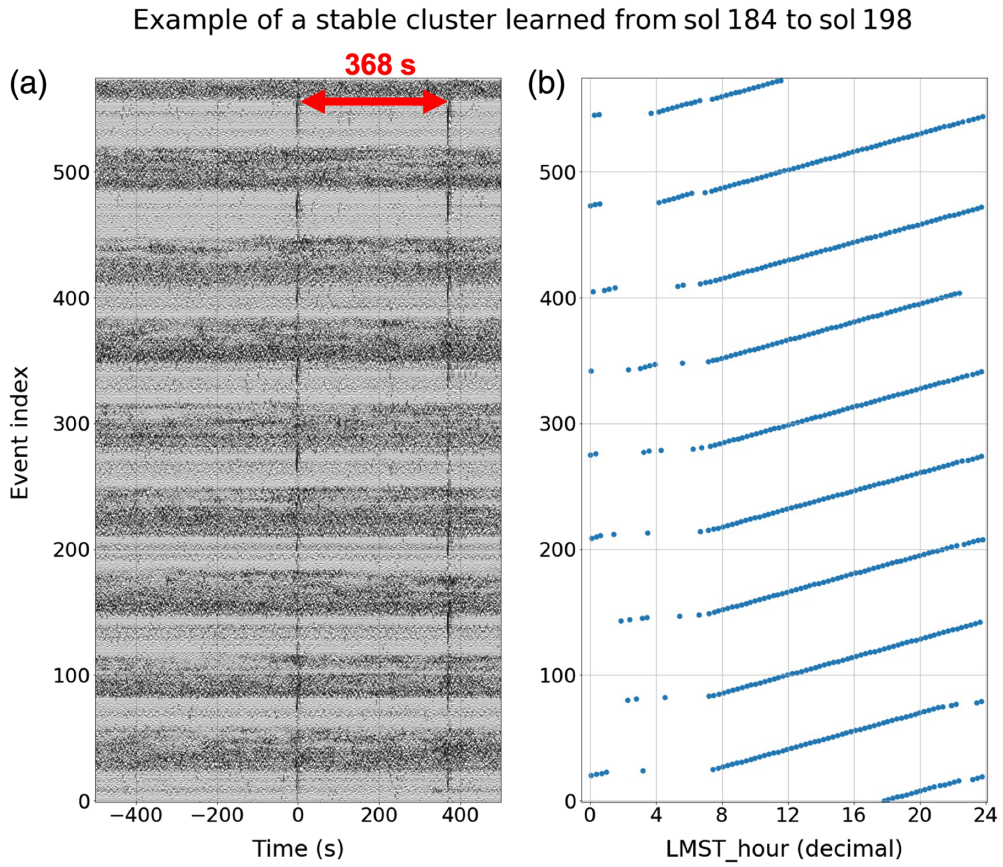


Figure 11. (a) Cluster of doublet glitches with 368 s time delay on component W. Amplitudes are normalized as shown in Figure 4, and root mean square (rms) is 2.3 s. (b) The LMST of these glitches show that an interruption is observed during the coldest time of the night. The color version of this figure is available only in the electronic edition.

−0.3 Pa during the learning period. For each event of the three clusters, we determine first its time through cross correlation with the cluster centroid waveform and compute then the time difference with the closest cataloged pressure drop. Events with a time difference larger than the learning window (100 s) are rejected (about 8%). The learning detected 341 events for cluster 0, 566 events for cluster 1, and 24 events for cluster 2, respectively. Similarity coefficients were larger than 0.00018 for cluster 0, 5.3×10^{-7} for cluster 1, and 0.0019 for cluster 2, respectively. In total, 111.9% (311 learned events out of 278 in the pressure drop catalog) of the reported dust devils are within this 100 s window, and 92% have furthermore a time difference of less than 20 s. DSN can, therefore, catalog the pressure drops directly from pressure data, and in addition improve the detection of smaller and not yet reported ones. For these two clusters, we detected 3.34 times more events than those found manually, with pressure amplitudes of −0.015 Pa for cluster 0, −0.017 Pa for cluster 1, and −0.2 Pa for cluster 2, respectively.

Figure 13 shows the occurrence time difference between the DSN-detected pressure drop and those of the catalog. Most of the events timing are within ±5 s from those of the catalog. Note also the secondary peaks, mostly 25 s prior, which is likely related to double pressure drop structure.

Figure 14 summarizes the results with a frequency–amplitude log–log cumulative histogram. This shows that the power −2 slope proposed by Lorenz *et al.* (2021) can be extended to lower amplitudes and at least down to 0.2 Pa. This doubles the number of pressure drops. The cluster 2, with a sharp pressure drop, seems more sensitive to noise and is found only for the largest pressure drop, whereas the cluster 0 might complete, for low amplitude, the cluster 1. Therefore, machine learning is efficient, and likely better at detecting and classifying pressure drops than previous studies made with InSight data.

VBB clusters correlation with the pressure clusters

Let us now compare the occurrence time of the seismic VBB clusters and those of the published dust devils catalog to identify VBB events related to pressure drops. Cluster numbers are those from the Comparison between SEIS Glitch Clusters and Glitch Catalog, the Correlation of SEIS Clusters with Temperature and Sol Quasi-Periodicity, and the Characterization of SEIS Multiglitches sections.

Let us first focus on pressure drops found with a time delay within ±100 s to a cluster event. For a cluster with a rate of N event per sol, a fraction of these might be coincident just by chance. In such a random process, the probability to get n pressure drops in the N windows of $\Delta T = 200$ s is $p(n) = a_0^n C(N, n)$, in which $C(N, n)$ is the binomial coefficient of n combinations over N and $a_0 = \Delta T/\text{sol}$, in which sol is the duration of one sol. This provides the 1σ threshold for all clusters, respectively, equal to n = 13, 13, 3, 8, 3, 6, 3, 2, 2 for clusters 0–8. These numbers were all computed for the reference period detailed in Section 1 of supplemental material for which the list of all clusters can be found.

Only five clusters are found above the 1σ threshold, with associated histograms in Figure 15: clusters 0, 1, 2, 4, and 8,

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

Figure 12. Waveforms of pressure drop clusters. The stacked waveforms are obtained using the approach outlined in the Comparison between SEIS Glitch Clusters and Glitch Catalog section. The color version of this figure is available only in the electronic edition.

with a number of pressure drops, respectively, 14×, 7×, 6×, 1.5×, and 8.5× those of the 1σ thresholds. For the 201 pressure drops reported in the test period from 3 to 10 June 2019, 90% (respectively 50%) of these pressure drops can be associated with an event of cluster 0 (respectively, 1).

With a closer look at Figure 15, we find pressure drops within ±25 s for cluster 0 and ±40 s for cluster 1. In addition, we observed on the centroid waveforms (Fig. 4), long-period oscillations, 25 s before the event's center for cluster 0 and 40 s before for cluster 1. VBB events are, therefore, detected in advance of the drop in pressure data. The glitch clusters 4 and 8 are also correlated with pressure drops. Their dip is close to horizontal, suggesting generation by the pressure drop of a microtilt on the instrument, in contrary to clusters 0 and 1 for which the signal has a significant vertical component. Finally, some of the events of cluster 2, related to wind bursts, are also associated with pressure drop. They account only for about 10% of the pressure drops and are likely related to the high winds observed during the pressure drop events.

CONCLUSION

The DSN method has proven to be powerful and effective when applied to the Martian dataset gathered by the InSight mission. It has successfully classified the dynamics of the noise in an automatic and unsupervised way. DSN is capable of extracting multiple features in a large dimensional space, to which the noise is mapped. This allows us to better understand and identify the properties in each time window of SEIS and pressure data. Naturally, the DSN approach can be generalized to other time series. With the multiple wavelets cascade and activation functions, patterns that cannot be easily identified are retrieved.

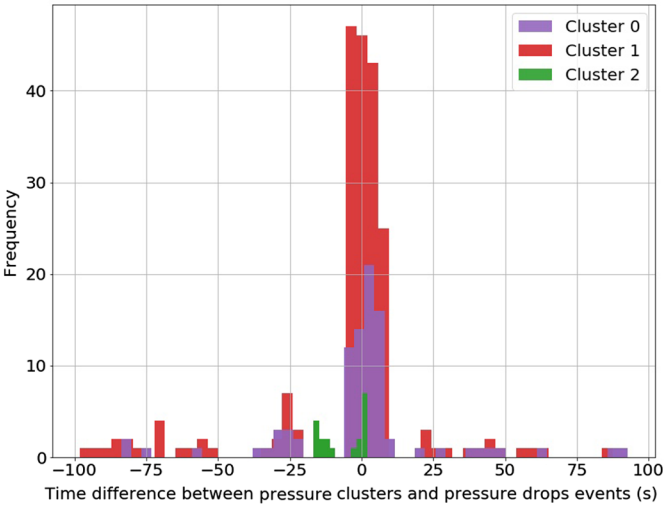


Figure 13. Timing of pressure drops. Histogram showing the number of pressure drops as a function of time difference between the pressure drop center, as reported in the pressure drop catalog of Spiga *et al.* (2021) and the center of the pressure drop, as event of clusters 0, 1, and 2. The bin size is 4 s for clusters 0, and 1 and 2 s for cluster 2. The learning window is 100 s, and the difference is reported when within ±100 s. The color version of this figure is available only in the electronic edition.

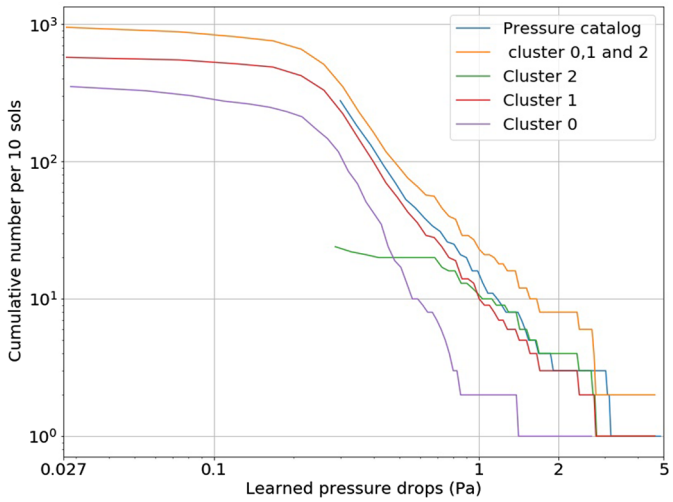


Figure 14. Statistics of pressure drops. Cumulative histogram of the pressure drops from Spiga *et al.* (2021) catalog (blue) and for the combined clusters 0-1-2 (orange). The histograms for each pressure drop cluster are also provided, with colors purple, red, and green for 0, 1, and 2, respectively. The color version of this figure is available only in the electronic edition.

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

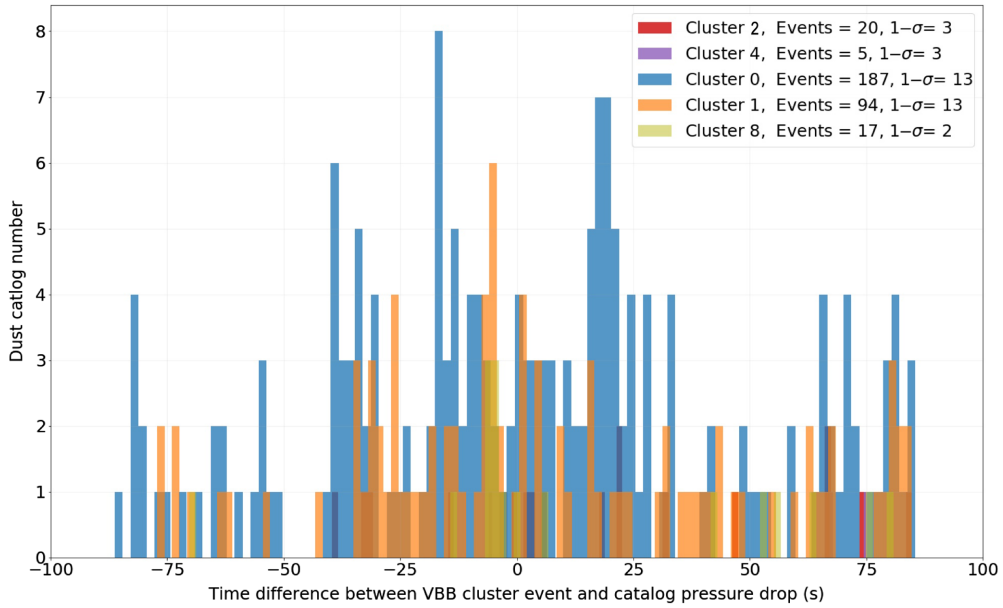


Figure 15. VBB pressure drop statistics. Histogram showing the number of pressure drops as a function of time difference between the pressure drop center, as reported in the pressure drop catalog and the center of the VBB events of several clusters. The learning window is 100 s, and the difference is only reported when within ±100 s. Only clusters for which the number of coincidence is larger than the $1-\sigma$ value obtained for random process are shown. The color version of this figure is available only in the electronic edition.

As a result, we detected multiple environmental Martian events such as glitches, pressure drops, and wind bursts with efficiency and sensitivity comparable with the published catalogs but in a full unsupervised way. More importantly, the DSN was able to discover and characterize, for the first time, tuplets of glitches, for which stable separation in time must be integrated in future autocorrelation analysis to ensure that these quasi-periodic events are not misinterpreted in terms of deep interior seismic phases. Therefore, DSN appears as a powerful tool for studying the nonstochasticity of seismic noise and finding noise structures both in terms of waveform and spectra. When implemented on continous data, this will allow possible misinterpretation between seismic phases and microseismic noise bursts, especially for low signal-to-noise ratio events.

This analysis also shows that unsupervised deep learning efficiently identifies clusters of microevents in seismic data. If used in parallel with more classical seismic detection algorithms, this could prevent detection saturation and select noise samples for future planetary or the Earth's ocean-bottom geophysical observatories unable to fully transmit their data. In this regard, DSN can not only enhance the robotic system performance, but also increase science return.

DATA AND RESOURCES

Seismic Experiment for Interior Structure (SEIS) data used are available in SEED format (InSIght Mars SEIS Service, 2019a) or PDS4 format (InSIght Mars SEIS Service, 2019b) at the respective dois: 10.18715/SEIS.INSIGHT.XB_2016 and 10.17189/1517570. Pressure and atmospheric data are available at National Aeronautics and Space Administration (NASA) Planetary data System (PDS) at the respectives dois: 10.17189/1518939 and 10.17189/1518950. Namely, we used in addition to seismic data the very broadband (VBB) sensor's temperatures (03.VKU, 03.VKV, 03.VKW), the Leveling System (LVL) temperature (VKI), and for Auxiliary Payload Sensors Suite (APSS), the atmospheric temperature (VKO) and pressure (03.BDO). Catalogs are available for Mars Quake Service (MQS) event in InSight Marsquake Service (2020), for pressure drops in Spiga *et al.* (2021), and for glitches in Scholz *et al.* (2020). The deep scattering network clustering algorithm used in this study is the original repository made on July 2020 at https://github.com/leonard-seydoux/scatnet. Ppol software was downloaded on February 2019 at https://ppol.readthedocs.io/en/latest/. This article is accompanied by two supplementary materials. Supplementary material A that contains additional figures to enhance and complete the article results. Supplementary material B contains three folders: The first folder "GlitchListing" is the list of glitches provided by Scholz *et al.* (2020). The second folder "PressureClusters" corresponds to the learned pressure drop data (three files) and documents all events of the three pressure drop clusters shown in the article. In each file, columns provide the event index, the UTC time, the similarity coefficient, the correlation coefficient, the amplitude, and finally the amplitude of the event in Pa. The final folder "VBBClusters" describes the VBB cluster's data and documents all events of the nine VBB clusters shown in the article. In each file, columns provide the event index, the UTC time, the similarity coefficient, the correlation coefficient, and the amplitude of, respectively, U, V, and W channels. All correlation coefficients are those with the event having the highest similarity coefficient, as explained in the article.

DECLARATION OF COMPETING INTERESTS

The authors declare no conflict of interest that could influence the work reported in this article.

ACKNOWLEDGMENTS

The authors acknowledge National Aeronautics and Space Administration (NASA), Centre National d'Etudes Spatiales (CNES), their partner agencies and Institutions (United Kingdom Space Agency [UKSA], Swiss Space Office [SSO], Deutsches Zentrum für Luft- und Raumfahrt [DLR], Jet Propulsion

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

Laboratory [JPL], Institut du Physique du Globe de Paris [IPGP]-Centre National de la Recherche Scientifique [CNRS], Eidgenössische Technische Hochschule Zürich [ETHZ], Imperial College London [ICL], Max Planck Institute for Solar System Research [MPS], Max–Planck–Gesellschaft [MPG]), and the flight operations team at JPL, SEIS on Mars operation Center (SISMOC), Mars SEIS data service (MSDS), Incorporated Research Institutions for Seismology–Data Management Center (IRIS-DMC), and Planetary Data System for providing SEED Seismic Experiment for Interior Structure (SEIS) data. Salma Barkaoui acknowledges CNES and the Ecole Doctorale 560 STEP'UP for her Ph.D. support. French authors are supported by Agence Nationale de la recherche (ANR) (ANR-19-CE31-0008-08) and by CNES for SEIS science support. Maarten V. de Hoop was supported by U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences and Biosciences Division under Grant Number DE-SC0020345 and the Simons Foundation under the MATH + X program. The authors thank Guest Editor Victor C. Tsai, and the two anonymous reviewers for their fruitful reviews which have improved greatly the article, as well as Renée C. Weber for her review and reading. This is the InSight Contribution Number 83 and IPGP Contribution Number 4248.

REFERENCES

Andén, J., and S. Mallat (2014). Deep scattering spectrum, *IEEE Trans. Signal Process.* **62,** no. 16, 4114–4128.

Banerdt, W. B., S. E. Smrekar, D. Banfield, D. Giardini, M. Golombek, C. L. Johnson, P. Lognonné, A. Spiga, T. Spohn, C. Perrin, *et al.* (2020). Initial results from the InSight mission on Mars, *Nature Geosci.* **13,** no. 3, 183–189.

Banfield, D., J. A. Rodriguez-Manfredi, C. T. Russell, K. M. Rowe, D. Leneman, H. R. Lai, P. R. Cruce, J. D. Means, C. L. Johnson, A. Mittelholz, *et al.* (2018). InSight auxiliary payload sensor suite (APSS), *Space Sci. Rev.* **215,** no. 1, 1–33.

Banfield, D., A. Spiga, C. Newman, F. Forget, M. Lemmon, R. Lorenz, N. Murdoch, D. Viudez-Moreiras, J. Pla-Garcia, R. F. Garcia, *et al.* (2020a). The atmosphere of Mars as observed by InSight, *Nature Geosci.* **13,** no. 3, 190–198.

Banfield, D., A. Spiga, C. Newman, F. Forget, M. Lemmon, R. Lorenz, N. Murdoch, D. Viudez-Moreiras, J. Pla-Garcia, R. F. Garcia, *et al.* (2020b). Insight APSS *PS* data product bundle, urn:nasa:pds:insightps, doi: 10.17189/1518939.

Bergen, K. J., and G. C. Beroza (2018). Earthquake fingerprints: Extracting waveform features for similarity-based earthquake detection, *Pure Appl. Geophys.* **176,** no. 3, 1037–1059.

Bruna, J., and S. Mallat (2013). Invariant scattering convolution networks, *IEEE Trans. Pattern Anal. Mach. Intell.* **35,** no. 8, 1872–1886.

Ceylan, S., J. F. Clinton, D. Giardini, M. Böose, C. Charalambous, M. van Driel, A. Horleston, T. Kawamura, A. Khan, G. Orhand-Mainsant, *et al.* (2021). Companion guide to the Marsquake catalog from InSight, sols 0478: Data content and non-seismic events, *Phys. Earth Planet. In.* **310,** 106597.

Charalambous, C., A. E. Stott, T. Pike, J. McClean, T. Warren, A. Spiga, D. Banfield, R. F. Garcia, J. Clinton, S. C. Stäler, *et al.* (2021). A comodulation analysis of atmospheric energy injection into the ground motion at InSight, Mars, *J. Geophys. Res.* **126,** e2020JE006538.

Clinton, J., S. Ceylan, M. van Driel, D. Giardini, S. C. Stähler, M. Böose, C. Charalambous, N. L. Dahmen, A. Horleston, T. Kawamura, *et al.* (2021). The Marsquake catalogue from InSight, sols 0478, *Phys. Earth Planet. In.* **310,** 106595.

Clinton, J., D. Giardini, M. Böose, S. Ceylan, M. Van Driel, F. Euchner, R. F. Garcia, S. Kedar, A. Khan, S. C. Stähler, *et al.* (2018). The Marsquake Service: Securing daily analysis of SEIS data and building the Martian seismicity catalogue for InSight, *Space Sci. Rev.* **214,** Article Number 133, 1–33.

Deng, S., and A. Levander (2020). Autocorrelation reflectivity of Mars, *Geophys. Res. Lett.* **47,** no. 16, e2020GL089630.

Falcin, A., J.-P. Métaxian, J. Mars, É. Stutzmann, J.-C. Komorowski, R. Moretti, M. Malfante, F. Beauducel, J.-M. Saurel, C. Dessert, *et al.* (2021). A machine-learning approach for automatic classification of volcanic seismicity at La Soufriere Volcano, Guadeloupe, *J. Volcanol. Geoth. Res.* **411,** 107151.

Fontaine, F. R. R., G. Barruol, B. L. N. Kennett, G. H. R. Bokelmann, and D. R. Reymond (2009). Upper mantle anisotropy beneath Australia and Tahiti from *P* wave polarization: Implications for real-time earthquake location, *J. Geophys. Res.* **114,** no. B3, doi: 10.1029/2008JB005709.

Garcia, R. F., B. Kenda, T. Kawamura, A. Spiga, N. Murdoch, P. H. Lognonné, R. Widmer-Schnidrig, N. Compaire, G. Orhand-Mainsant, D. Banfield, *et al.* (2020). Pressure effects on the SEIS-InSight instrument, improvement of seismic records, and characterization of long period atmospheric waves from ground displacements, *J. Geophys. Res.* **125,** no. 7, e2019JE006278.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensor-Flow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, Sebastopol, California.

Giardini, D., P. Lognonné, W. B. Banerdt, W. T. Pike, U. Christensen, S. Ceylan, J. F. Clinton, M. van Driel, S. C. Stähler, M. Böose, *et al.* (2020). The seismicity of Mars, *Nature Geosci.* **13,** no. 3, 205–212.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*, MIT Press, Boston, Massachusetts.

Hibert, C., D. Michéa, F. Provost, J.-P. Malet, and M. Geertsema (2019). Exploration of continuous seismic recordings with a machine learning approach to document 20 yr of landslide activity in Alaska, *Geophys. J. Int.* **219,** no. 2, 1138–1147.

InSight Marsquake Service (2020). Mars seismic catalogue, InSight mission; v12/1/2020, ETHZ, IPGP, JPL, ICL, ISAE-Supaero, MPS, Univ. Bristol, doi: 10.12686/A6.

InSight Mars SEIS Data Service (2019a). SEIS raw data, InSight mission, IPGP, JPL, CNES, ETHZ, ICL, MPS, ISAE-Supaero, LPG, MFSC, doi: 10.18715/seis.insight.xb_2016.

InSight SEIS Data Service (2019b). InSight SEIS Data Bundle, InSight SEIS Science Team, NASA Planetary Data System, doi: 10.17189/1517570.

Jia, Y., and J. Ma (2017). What can machine learning do for seismic data processing? An interpolation application, *Geophysics* **82,** no. 3, V163–V177.

Jordan, M. I., and T. M. Mitchell (2015). Machine learning: Trends, perspectives, and prospects, *Science* **349,** no. 6245, 255–260.

Kenda, B., M. Drilleau, R. F. Garcia, T. Kawamura, N. Murdoch, N. Compaire, P. Lognonné, A. Spiga, R. Widmer-Schnidrig, P. Delage, *et al.* (2020). Subsurface structure at the InSight landing

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user

site from compliance measurements by seismic and meteorological experiments, *J. Geophys. Res.* **125,** no. 6, e2020JE006387.

Kim, D., P. Davis, V. Leki, R. Maguire, N. Compaire, M. Schimmel, E. Stutzmann, J. Irving, P. Lognonné, J.-R. Scholz, *et al.* (2021). Potential pitfalls in the analysis and structural interpretation of seismic data from the Mars InSight mission, *Bull. Seismol. Soc. Am.* doi: 10.1785/0120210123.

Kong, Q., D. T. Trugman, Z. E. Ross, M. J. Bianco, B. J. Meade, and P. Gerstoft (2018). Machine learning in seismology: Turning data into insights, *Seismol. Res. Lett.* **90,** no. 1, 3–14.

Lognonné, P., W. B. Banerdt, D. Giardini, W. T. Pike, U. Christensen, P. Laudet, S. de Raucourt, P. Zweifel, S. Calcutt, M. Bierwirth, *et al.* (2019). SEIS: InSight's seismic experiment for internal structure of Mars, *Space Sci. Rev.* **215,** no. 1, 12.

Lognonné, P., W. B. Banerdt, W. T. Pike, D. Giardini, U. Christensen, R. F. Garcia, T. Kawamura, S. Kedar, B. Knapmeyer-Endrun, L. Margerin, *et al.* (2020). Constraints on the shallow elastic and anelastic structure of mars from InSight seismic data, *Nature Geosci.* **13,** no. 3, 213–220.

Lorenz, R. D., M. T. Lemmon, J. Maki, D. Banfield, A. Spiga, C. Charalambous, E. Barrett, J. A. Herman, B. T. White, S. Pasco, *et al.* (2020). Scientific observations with the insight solar arrays: Dust, clouds, and eclipses on Mars, *Earth Space Sci.* **7,** no. 5, e2019EA000992.

Lorenz, R. D., A. Spiga, P. Lognonné, M. Plasman, C. E. Newman, and C. Charalambous (2021). The whirlwinds of Elysium: A catalog and meteorological characteristics of "dust devil" vortices observed by InSight on Mars, *Icarus* **355,** 114119.

Malfante, M., M. D. Mura, J.-P. Metaxian, J. I. Mars, O. Macedo, and A. Inza (2018). Machine learning for volcano-seismic signals: Challenges and perspectives, *IEEE Signal Process. Mag.* **35,** no. 2, 20–30.

Mora, L. (2019). APSS PS data, Atmosphere's node, doi: 10.17189/1518939.

Obara, K. (2002). Nonvolcanic deep tremor associated with subduction in southwest Japan, *Science* **296,** no. 5573, 1679–1681.

Oyallon, E., E. Belilovsky, and S. Zagoruyko (2017). Scaling the scattering transform: Deep hybrid networks, *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, Venice, Italy, 22–29 October.

Peterson, J. (1993). *Observations and modeling of seismic background noise*, *U.S. Geol. Surv. Open-File Rept. 93-322.*

Priyadarshini, I., and V. Puri (2021). Mars weather data analysis using machine learning techniques, *Earth Sci. Inf.* doi: 10.1007/s12145-021-00643-0.

Schimmel, M., E. Stutzmann, P. Lognonné, N. Compaire, P. Davis, M. Drilleau, R. Garcia, D. Kim, B. Knapmeyer-Endrun, V. Lekic, *et al.* (2021). Seismic noise autocorrelations on Mars, *Earth Space Sci.* **8,** no. 6, e2021EA001755, doi: 10.1029/2021ea001755.

Scholz, J.-R., G. Barruol, F. R. Fontaine, K. Sigloch, W. Crawford, and M. Deen (2017). Orienting ocean-bottom seismometers from P-wave and Rayleigh wave polarisations, *Geophys. J. Int.* **208,** no. 3, 1277–1289, doi: 10.1093/gji/ggw426.

Scholz, J.-R., R. Widmer-Schnidrig, P. Davis, P. Lognonné, B. Pinot, R. F. Garcia, K. Hurst, L. Pou, F. Nimmo, S. Barkaoui, *et al.* (2020). Detection, analysis, and removal of glitches from InSight's seismic data from Mars, *Earth Space Sci.* **7,** no. 11, e2020EA001317.

Seydoux, L., R. Balestriero, P. Poli, M. de Hoop, M. Campillo, and R. Baraniuk (2020). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning, *Nat. Comm.* **11,** no. 1, 3972.

Spiga, A., N. Murdoch, R. Lorenz, F. Forget, C. Newman, S. Rodriguez, J. PlaGarcia, D. V. Moreiras, D. Banfield, C. Perrin, *et al.* (2021). A study of daytime convective vortices and turbulence in the Martian planetary boundary layer based on half-a-year of insight atmospheric measurements and large-eddy simulations, *J. Geophys. Res.* **126,** no. 1, e2020JE006511.

Stutzmann, E., M. Schimmel, P. Lognonné, A. Horleston, S. Ceylan, M. van Driel, S. Stahler, B. Banerdt, M. Calvet, C. Charalambous, *et al.* (2021). The polarization of ambient noise on Mars, *J. Geophys. Res.* **126,** no. 1, e2020JE006545.

AUTHORS AND AFFILIATIONS

Salma Barkaoui: Institut de Physique du Globe de Paris, Université de Paris, CNRS, Paris, France, https://orcid.org/0000-0001-7266-0815; **Philippe Lognonné:** Institut de Physique du Globe de Paris, Université de Paris, CNRS, Paris, France, https://orcid.org/0000-0002-1014-920X; **Taichi Kawamura:** Institut de Physique du Globe de Paris, Université de Paris, CNRS, Paris, France, https://orcid.org/0000-0001-5246-5561; **Éléonore Stutzmann:** Institut de Physique du Globe de Paris, Université de Paris, CNRS, Paris, France, https://orcid.org/0000-0002-4348-7475; **Léonard Seydoux:** Institut des Sciences de la Terre, Université Grenoble-Alpes, UMR CNRS, Gières, France, https://orcid.org/0000-0002-6596-5896; **Maarten V. de Hoop:** Rice University, Houston, Texas, U.S.A.; **Randall Balestriero:** Rice University, Houston, Texas, U.S.A., https://orcid.org/0000-0002-5692-4187; **John-Robert Scholz:** Max Planck Institute for Solar System Research, Göttingen, Germany, https://orcid.org/0000-0003-1404-2335; **Grégory Sainton:** Institut de Physique du Globe de Paris, Université de Paris, CNRS, Paris, France, https://orcid.org/0000-0002-9375-4877; **Matthieu Plasman:** Institut de Physique du Globe de Paris, Université de Paris, CNRS, Paris, France, https://orcid.org/0000-0002-5630-2089; **Savas Ceylan:** Institute for Geophysics, ETH Zürich, Zürich, Switzerland, https://orcid.org/0000-0002-6552-6850; **John Clinton:** Swiss Seismological Service, ETH Zürich, Zürich, Switzerland, https://orcid.org/0000-0001-8626-2703; **Aymeric Spiga:** Laboratoire de Météorologie Dynamique/Institut Pierre Simon Laplace (LMD/IPSL), Sorbonne Université, Centre National de la Recherche Scientifique (CNRS), Paris, France, https://orcid.org/0000-0002-6776-6268; **Rudolf Widmer-Schnidrig:** University of Stuttgart, Institute of Geodesy, Stuttgart, Germany, https://orcid.org/0000-0001-9698-2739; **Francesco Civilini:** California Institute of Technology, Pasadena, California, U.S.A., https://orcid.org/0000-0003-0669-0404; and **W. Bruce Banerdt:** Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A., https://orcid.org/0000-0003-3125-1542

Manuscript received 30 March 2021
Published online 9 November 2021

Downloaded from http://pubs.geoscienceworld.org/ssa/bssa/article-pdf/111/6/2964/5475350/bssa-2021095.1.pdf
by CNRS INSU user