Recurrent Scattering Network detects metastable behavior in polyphonic seismo-volcanic signals for volcano eruption forecasting

Angel Bueno, Randall Balestriero, Silvio De Angelis, Carmen Benítez, Luciano Zuccarello, Richard Baraniuk, Jesús M. Ibáñez and Maarten V. de Hoop

Abstract

We introduce an End-to-End (E2E) deep neural network architecture designed to perform seismo-volcanic monitoring focused on detecting change. Due to the complexity of volcanic processes, this requires a polyphonic detection, segmentation and classification approach. Through evolving epistemic uncertainty, invoking a Bayesian network strategy, we detect change and demonstrate its significance as an indicator for possible forecasting of eruptions using data from the Bezymianny and Etna volcanoes. Specifically, we propose morphing the scattering transform from previous work into a novel E2E hybrid and recurrent learnable deep scattering network to adapt to the multi-scale temporal dependencies from streaming data. The time-dependent scattering is in some sense physics informed, namely through time-frequency representation (TFR) of the data. At the same time, with a carefully designed deep convolutional-LSTM architecture, we learn intra-event, temporal dynamics from the scattering coefficients or features. We verify the effectiveness of transfer learning switching between volcanoes. Our experimental results set a new norm for semi-supervised seismo-volcanic monitoring.

Index Terms

Volcanoes, Seismology, Uncertainty, Recurrent Neural Networks, Wavelet transforms.

Corresponding author: A. Bueno. Contact email: angelbueno@ugr.es

IEEE Transactions on Geoscience and Remote Sensing accepted manuscript: https://ieeexplore.ieee.org/document/9645556

I. Introduction

We introduce an End-to-End deep neural network architecture designed to perform multisource seismo-volcanic monitoring on the one hand and detecting change on the other hand. The search for observable eruption precursory signals and their evolution has remained one of the challenges of modern volcanology and is critical for effective monitoring and the forecasting of eruptions. We address this search with a comprehensive deep learning framework.

A volcanic eruption, the emission of magma and gases as well as the exchange of various forms of energy, is the final consequence of a series of energetic physical and chemical processes in Earth's interior. Volcanic eruptions range from fluid lava flows to explosive emissions that inject large volumes of material into the atmosphere. Repose times in between individual eruptions also vary widely, even for individual volcanoes, extending from minutes to years, suggesting that eruptions are associated with quasi-stable processes [1], [2], complicating forecasting. From a seismic perspective, there are currently two main approaches to forecasting eruptions. The first one is based on the study of changes in the stress state of the volcanic edifice when a volcano evolves towards an eruption, as detectable in terms of variations in the propagation velocity of seismic waves [3]. This approach assumes that pre-eruptive changes in the stress state of the volcanic edifice reflect into corresponding variations of the propagation velocity of seismic waves. These latter ones, can be detected using the correlation of background seismic noise gathered at different sites on the volcano. Indeed, important progress in the analysis of seismic background noise at volcanoes has been made [4], [5]. The second approach is to detect and identify seismo-volcanic signals associated with physical processes within the volcano and to use these as precursors.

Here, we propose a third approach in which deep learning is used to relate changes in the seismic wave field to changes in the state of the volcano (presumably directly related to the underlying physics), including volcanic tremor. The results of this study show that the seismic wave field properties can reflect the evolution of volcano dynamics and can be used as precursor of volcanic eruptions.

Seismology has played a leading role in monitoring volcanoes and identifying precursors to eruptions [6], [7], [8]. The sources of seismic energy are diverse and include rock rupture through an accumulation of elastic energy, ground resonance phenomena, pressure changes due to the movement of fluids, conduit resonance, and many more. As a consequence, seismic signals

exhibit variable durations and spectral contents [9], [10], [11], [12], [13], [14]. One challenge is to identify classes of events causally related to distinct sources mechanisms, providing an association that can be used as an indicator of the physical processes occurring within the volcano. Identification is currently carried out using standard, supervised machine learning techniques [14], [15], [16], [17], [18], [19]. However, identifying signals does not provide information about the dynamics of the processes that lead to eruptions.

Our approach departs from those using handcrafted features to detect change [20], [21], [22], [23]; we suggest that subtle changes not currently identified in standard data analysis are important. Our deep neural network can identify such changes by dynamically learning a scattering representation of streaming data. The introduction of epistemic uncertainty revealed by Monte Carlo dropout [24] is pivotal in our approach, identifying a drift in uncertainty in the classification of events presumably related to the relevant processes.

State-of-the-art procedures for the detection, segmentation, and classification of signals or events in streaming data are often implemented as separate workflows by combining signal processing (to provide a priori representations) and traditional deep learning strategies (to probe these representations) [9], [16], [17], [18], [25], [26], [27]. The End-to-End approach proposed here addresses the shortcomings of past methods.

New approach: The main contributions of this paper can be summarized as follows:

1) We propose a scattering network that cascades learnable wavelet transforms and complex moduli, in its original form given by [28], to generate features with a learnable spline approximation in both central frequencies and wavelet shape for each layer. The central frequency of the mother wavelet in a given task can be calibrated by letting the dilation factor of the filters, along with the spline knots, be learnable parameters. Knots and filter learnability leads to a mother wavelet capable of shape morphing to better capture signal onsets, even in very challenging environments.
2) We introduce a novel integration of this learnable scattering network in a recurrent architecture. The learnable scattering network component produces a set of multiple-order representations (so-called scatter-grams) that contain a full range of wavelet scales. This range is converted to a sequence of temporal, structured representations of local scattering variations, with the scatter-grams as complementary channels, in which two stacked convolutional long-short term memories (conv-LSTMs) can extract intra-frequency temporal variation across multiple scales [29], [30]. In parallel, a learnable scattering denoising

operation is performed via convolutional skip connections, suppressing background noise and enhancing the scatter-grams from the data stream. The outputs of both components, conv-LSTMs and skip connections, are fused, captured by bidirectional LSTMs [31], and forwarded to a dense layer to output a probabilistic event detection and classification matrix highlighting multisource, coexistent seismic signals as *polyphonic* events.

3) We use the uncertainty to study the goodness of a classifying process of seismo-volcanic signals for the purpose of eruption early warning protocols that are exportable across volcanic systems. We define uncertainty as a measure of how little or how much one set of data resembles another and detect the variation or evolution of a physical system. Therefore, the concept of uncertainty can be used to control the quality of the physical measurements of a volcanic system and as an indicator of the evolution of these measurements over time; as such, this approach is suitable for forecasting volcanic eruptions. We employ epistemic uncertainty to detect change and reveal that the power-law drifts towards eruptions. The evolution of the seismic data in a volcano can be potentially associated with the appearance of new seismic signals or a change in the characteristics of those seismic signals. Therefore, an increase (or decrease) in the uncertainty can be used as a good indicator for early warning of volcanic eruptions and exported from one volcano to another [32].
4) We demonstrate through transfer learning that our implemented architecture can be exported across a range of different volcanoes and eruptive styles. With minimal hassle, we are able to reuse our system from one volcano to another one, even if there exists new or unknown signals in the target volcano. The use of uncertainty in a transfer learning approach permits establishing universal early warning and predictive protocols that are the same regardless of the volcano, thus giving an objective measure that can be communicated during volcanic crises.

We demonstrate the potential of our architecture using data from three well-studied 2007 eruptions of Bezymianny volcano, over a period of approximately 3 months. The three eruptions—those on 25 September, 14 October, and 5 November—were brief but very energetic, and included various pre-eruptive seismicity rates and eruption mechanisms. Moreover, application to data from Mt. Etna and Mt. St. Helens confirms that our approach can be applied to different volcanic systems. To the best of our knowledge, this is the first deep learning approach of its kind; that is, it is the first to address multisource detection, segmentation, and implicit classification

of physical processes in an eruptive sequence while detecting change through drift in epistemic uncertainty.

II. Prior work

Machine learning in seismology. Advanced machine learning techniques provide tools beyond human intuition to discover unusual signals and patterns, and have been applied to data analysis in the field of seismology [33], [34], [35], [36]. In volcano-seismology, the application of machine-learning methods has focused mostly on automated detection and classification of seismic signals through handcrafted signal properties [16], embedding vectors [37], Hidden Markov Models [26], and standard deep learning methodologies [17]. These works follow archetypal machine-learning pipelines: a set of features derived from the continuous seismo-volcanic data streams are selected to fine tune monitoring algorithms. However, catalogs remain incomplete leading to a partial assessment of the ongoing phenomena, and their significance in terms of eruption forecasting. This is partly due to the lack of uniformity in labeling events, and the occurrence of multiple sources at the same time in low SNR environments. Moreover, these methods have two main limitations when dealing with seismic data. First, they can only handle non-overlapping, monophonic seismic signals. Second, they cannot be applied to new seismic wave fields without re-training the entire network.

Study of volcano dynamics. Past attempts to the detection of volcanic precursors have concentrated on ideas adapted from engineering applications: the Failure Forecast Method (FFM) carries out a regressive estimation of the time to target (i.e., time to eruption or structural failure) based on handcrafted features from accelerated strain. Recent machine learning methods discern eruptive behavior from a set of handcrafted features. A filter-bank analysis of 6 years of data from Piton de la Fournaise volcano (Reunion, France) was performed to derive 990 features in the spectral band of 0.5–26 Hz, and forwarded to a decision-tree gradient boosting algorithm named XGBoost [38]. The empirical results showed feature evolution through time, with tracking capabilities of the variation in a narrow frequency range (3–5 Hz). A posterior analysis with spectral clustering provided a physical interpretation of the volcano dynamics, highlighting the dominance of tremor and high-frequency components during the eruptive behavior. An analysis similar to that of [38] but with different classifiers was carried out for Telica (Nicaragua) and Nevado del Ruiz (Colombia), linking eruptive and non-eruptive behavior to the temporal variations of features [39].

Despite the promising performance of integrated machine learning and handcrafted features, two issues arise for the real-time deployment of such systems in an observatory. The tabular set of features reflects a predetermined range based on analysts' experience; thus, unforeseen situations can saturate the classifier while underestimating volcanic unrest. Second, experimental results show that not all of these features contribute to the final prediction. Instead, current, traditional machine learning methods exhibit a predictive bias towards a specific frequency band that is considered essential to categorize all types of seismic transients (from pressure pulses in fluids to the fracturing of rocks) in a volcano. Finally, the consideration of parameterized windows over the seismic data stream can effectively trace its temporal evolution but do not identify which families of events contribute to those changes. The observable variables are often intertwined with a myriad of source-dependent seismic events, such as volcano-tectonic earthquakes or tremor. Given the different associations of these signals to physical mechanisms, distilling such knowledge is essential to understand the eruptive dynamics of a volcano.

Learnable filter banks. With the availability of large scale audio classification datasets, multiple methods have been developed to bring learnability to filter banks [40], [41], [42], [43], [44]. Those methods can be divided into two main categories. The first category [45] performs filter-bank learning by independently adapting the center frequencies and bandwidths of a collection of Morlet wavelets. Another family of methods [46] relies on learning the start and cutoff frequencies of a band-pass sinc filter apodized with a Hamming window. The second category of methods are based on the STFT. For example, [41] propose to learn Mel filters that are applied to a spectrogram (modulus of STFT). Those filters linearly combine adjacent filters in frequency which can be interpreted as learning a linear frequency sub-sampling of the spectrogram. Learning the apodization window used to produce the spectrogram has also been developed [43], [47]. Alternatively, one may select a transform a priori that can be adapted with learnable hyperparameters [48]. In monitoring applications where signals overlap in time, the intra-class separability can be extremely difficult, and using the correct transform can greatly impact predictive performance.

Convolution LSTM, 2D. Recurrent Neural Networks (RNNs) are highly successful in temporal modeling of time sequences. LSTMs and GRUs have been proposed as specialized architectures to refine temporal modeling. If applied over the scattering features, one of the LSTMs' limitations is that intra-frequency scatter-grams variations information is missing. In this regard, Convolutional-LSTM (ConvLSTM) is proposed as a mathematical modification able to model

scales variations through time by explicitly encoding this information into state tensors. The first application of ConvLSTM was for weather forecasting [30]. The network is trained on a forecasting problem with temporal sequences of radar echo maps. ConvLSTMs have been applied in speech recognition [29], gesture detection [49] and, as mentioned, weather forecasting [50], among many other disciplines. Modifications over the original structure have been developed to increase robustness in specific settings [51]. The applicability of ConvLSTMs in seismology remains to be explored. In our proposed framework, we modify the working regime of the ConvLSTM architecture to cope with a multi-scale time-frequency representation, adjusting the network to apply to waveform dynamics in a multi-source or polyphonic setting.

III. Problem description

Seismo-volcanic monitoring can be approached from a sequence prediction perspective. If multiple and simultaneous sources generate waveforms or signals, a polyphonic approach is needed to detect the presence and class-membership of overlapping events. We define our seismic data domain as the whole set of recorded waveforms and known seismo-volcanic event types for any given monitoring period, T. Within this defined data domain, our training dataset is $D = \{(X, Y)\}$, with $X = (x_0, x_1, \ldots, x_n)$, representing raw, sampled data streams of arbitrary time duration, and $Y = (y_0, y_1, ..., y_n)$ representing the labeled sequence, with $y_i, i = (0, \cdots, n)$ containing one or multiple categorical labels over K classes. From a machine learning perspective, finding the relationships between pairs of multi-source, framewise X and Y can be cast as polyphonic event detection and classification. To achieve this, the Y labels must be converted into a binary temporal matrix whose columns represent time and whose rows the frequencies which characterize the trace. This type of formulation allows the ubiquitous identification of seismic events whose sources may vary over time or for which enough data samples are not available for training purposes.

IV. Representation learning

A. Scattering Network

The deep scattering network or DSN [28], [52], [53], extracts locally invariant robust representations from a raw signal, by systematically applying a cascade of wavelet transforms and modulus operators. Consider the input signal of length N as $x \in \mathbb{R}^N$ where each time sample is accessed via $x(t), t = 1, \ldots, N$. Each DSN layer indexed by ℓ is parameterized by $J^{(\ell)}$ and

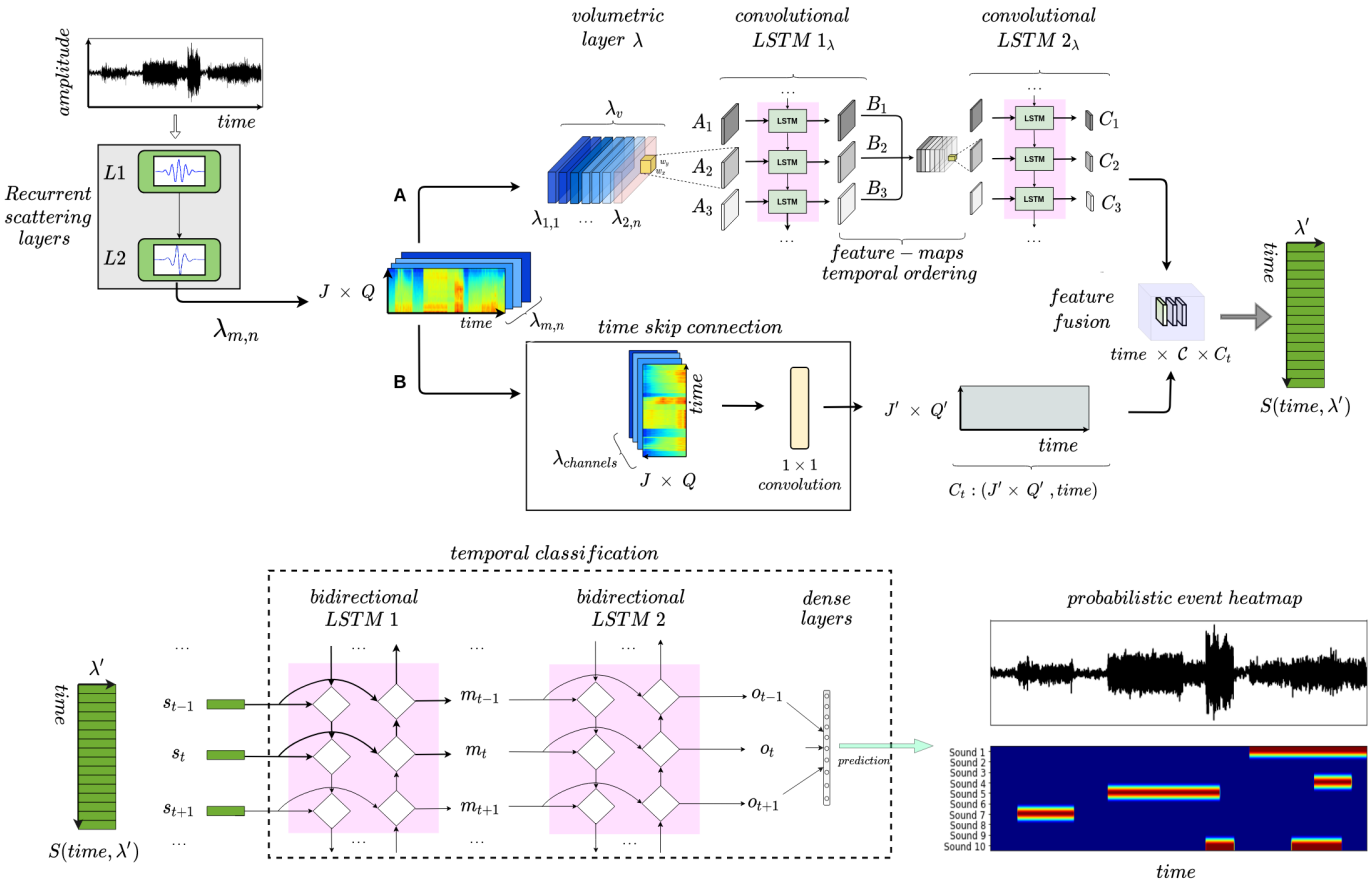

Fig. 1. **Network architecture for polyphonic sound event detection**: the recurrent learnable scattering network permits flexible temporal feature representation, while the joint modeling yields increased adaptability of the mother wavelet to the target sounds. Path A performs dynamic multi-scale convolutions, whereas path B performs dimensionality reduction and scattering compression yielding sensitivity to the onsets of events. After fusion, features are captured by a temporal classification module, with two bidirectional long-short term memory (LSTM) and an Fully Connected (FC) layer that outputs the probabilistic event detection matrix with per-class onsets.

$Q^{(\ell)}$, the number of octaves and the number of wavelets per octave, respectively. In fact, the wavelet filter bank is derived as $\{\phi_\lambda, \lambda \in \{2^{i/Q}, i = 1, \ldots, JQ\}\}$ with $\phi_\lambda = \lambda^{-1/2}\phi(\lambda^{-1}t)$ where ϕ is a mother wavelet chosen a priori. This structure is similar to a convolutional neural network architecture. At layer ℓ, the wavelet filter-bank produces the representation $U_\lambda^{(\ell)}$ which corresponds to the application of each filter followed by application of taking the modulus as

$$U^{(\ell)}_{\lambda^{(\ell)},\lambda^{(\ell-1)},\ldots,\lambda^{(1)}}(t) = |(\phi_\lambda \star U^{(\ell-1)}_{\lambda^{(\ell-1)},\ldots,\lambda^{(1)}})(t)|, \tag{1}$$

where $|.|$ denotes the (complex) modulus, and the initialization at the first layer is defined as $U^{(0)} = x$. Note that the wavelet filter bank can vary depending on the layer ℓ. Once the

representations are obtained for all the layers, the scattering features or coefficients, which form the representation, are obtained by performing a time averaging. In the case of global average pooling this corresponds to

$$S^{(\ell)}_{\lambda^{(\ell)},\ldots,\lambda^{(1)}} = \frac{1}{N}\sum_{t=1}^{N} U^{(\ell)}_{\lambda^{(\ell)},\ldots,\lambda^{(1)}}(t), \tag{2}$$

with $S^{(\ell)}_{\lambda^{(\ell)},\ldots,\lambda^{(1)}}$ the scattering representation at layer ℓ. The time averaging can be made local based on the desired time invariance, thus defining the time resolution for the successive layers.

B. Learnable Scattering Transform

The scattering network introduced above consists of a succession of wavelet transforms and complex moduli, in a neural-network fashion with per-layer wavelet filter banks. Hermite Cubic Splines provide a learnable parametrization of the mother wavelet from which the wavelet filter bank per layer is derived. Formally, we introduce a partition of a compact support of the mother wavelet into R intervals using $K = R + 1$ knots, $t_r, r = 1, \ldots, K$, so that

$$\begin{aligned} f_{\Theta,\Gamma}(t) = \sum_{r=1}^{K-1} \Bigg(& \gamma_r + u_0\left(\frac{t-t_r}{t_{r+1}-t_r}\right) + \gamma_{r+1} + \\ & + u_1\left(\frac{t-t_r}{t_{r+1}-t_r}\right) + \theta_r + \\ & + v_0\left(\frac{t-t_r}{t_{r+1}-t_r}\right) + \theta_{r+1} + \\ & + v_1\left(\frac{t-t_r}{t_{r+1}-t_r}\right) 1_{\{t\in[t_r,t_{r+1}].\}} \end{aligned} \tag{3}$$

where $\Theta = \{\theta_1, \ldots, \theta_K\}$, $\Gamma = \{\gamma_1, \ldots, \gamma_K\}$ and the basis functions are given by

$$\begin{aligned} u_0(t) &= 2t^3 - 3t^2 + 1 \\ u_1(t) &= -2t^3 + 3t^2 \\ v_0(t) &= t^3 - 2t^2 + t \\ v_1(t) &= t^3 - 2t^2 \end{aligned}$$

subject to the following constraints,

$$\gamma_1 = \theta_1 = \theta_K = \gamma_K = 0, \qquad \text{(compact support)}$$

$$\gamma_2 = -\sum_{r \neq 1} \gamma_r, \qquad \text{(zero mean)}$$

$$\max_r |\gamma_r| < \infty, \max_r |\theta_r| < \infty \qquad \text{(boundedness)}.$$

The parameters γ and θ allows one to learn the shape while keeping a uniform partition for the spline with knots, t_r, equally spaced in time. In this work, we propose a further adaptivity of the mother wavelet by employing two critical extensions of the above formulation. First, we allow learning of the knots' positions of the mother wavelet. This enables learnability with varying instantaneous frequency. Even chirplets [54] can be learned in this scenario. Second, we allow learning of the scaling factors used to generate the filter bank from dilating the mother wavelet layerwise. This entails learning the center frequencies of the wavelets. Formally, the filter bank is obtained through dilation of the mother wavelet given a collection of scaling factors, Λ, as in

$$\text{FilterBank} = \{\psi_\lambda, \lambda \in \Lambda\} \;\; \text{with} \;\; \psi_\lambda(t) = \frac{1}{\sqrt{\lambda}} f_{\Theta,\Gamma}\left(\frac{t}{\lambda}\right) \tag{4}$$

As a result, the learnable parameters of the scattering transform are $\Theta \cup \Gamma \cup \{t_1, \ldots, t_K\} \cup \Lambda$, where the labeling of the DSN layers by ℓ is suppressed in the notation.

C. Time scattering learning stability

The learnable scattering transform coupled with RNNs generates learning instabilities due to variations of the knots. In fact, a translation in time of the knots results in the well-known phenomenon of *exploding gradients* [55]. We employ *gradient clipping*, an optimization constraint to maintain the gradient values of the knots within an interval. In our case, gradients are clipped between $[-1, 1]$. The optimization of the filters does not need additional constraints nor gradient clipping.

V. Architecture: Scattering Aware Recurrent Convolutional Layers

We now describe the modules of our recurrent architecture taking the scattering coefficients or features as input. These recurrent components comprise the volumetric layer, the pair of modified ConvLSTM 1_λ and ConvLSTM 2_λ, the time-skip connection and the temporal classification module in Figure 1.

A. Volumetric Layer

The recurrent DSN layers, $L1$ and $L2$ say, generate the first- and second-order scattering coefficients. The scattering coefficients generated by both DSN layers can be considered as an image upon identifying time-scale as spatial coordinates; the size of the image is $(J \times Q) \times$ (number of time samples). Thus an image is obtained as an alternative representation of the scattering coefficients of a signal while these vary with time. The number of time samples (and time interval) is determined by the time resolution after the pooling operation at the DSN layers $L1$ and $L2$ (see (2)). The volumetric layer arranges feature vectors of dimension n, say, into a single *volumetric representation*, λ_v. The volumetric representation consists of a multi-channel image with $\lambda_{2,n} + 1$ channels, arranged by scattering order. The first channel consists of the first-order scattering coefficients, and the $\lambda_{2,n}$ following channels constitute the second-order scattering coefficients. Thanks to the learnability of the filter banks in the DSN layers coupled with the subsequent recurrent networks, the multi-channel volume λ_v adjusts its sensitivity to the set of important frequencies, event shapes and transient dynamics.

B. Dynamic Convolutions

In the upper branch, denoted by A, two temporal LSTM networks take the volumetric representation and produces a sequence of tensors, $\mathcal{C} = (C_1, C_2, ..., C_n)$ as output. Instead of learning a unique representation from the scattering layers $L1$ and $L2$, the upper branch of the network learns the full sequence of scattering feature vectors from the volumetric layer. The multi-channel image will be codified into a set of sequential hidden states H and memory cells $\mathbf{C}$ that explicitly encode the intra-scale variations of the input feature vectors. To achieve this, the core components of *dynamic convolutions* are Convolutional-LSTMs (ConvLSTMs), a mathematically modified LSTMs units that have replaced its internal structure of standard multiplications with convolutions [30]. The incorporation of internal convolutions in ConvLSTMs remove the limitation of standard LSTMs that operate sequentially over single feature vectors [56]. Our branch of *dynamic convolutions* follows the governing equations described in [30], but are modified to accept the volumetric representation, λ_v, as input and performing a sequential analysis of n steps equal to the number of channels in λ_v.

The first ConvLSTM, (ConvLSTM 1_λ in figure 1) probes the volumetric representation and performs a sequential analysis with each feature vector. We note the sequence $\mathcal{A} = [A_1, A_2, A_3, \ldots, A_n]$ as the input-to-state transitions from λ_v to the memory cell $\mathbf{C}_n$ in ConvLSTM 1_λ. The memory

cell in a ConvLSTM has an embedded *read-and-write* functionality to select and store sequential meaningful information. This functionality is implemented by a set of matrices, also known as gates, that control the flux of information that has to be stored in $\mathbf{C}_n$ from the input feature vector sequence. In our ConvLSTM 1_λ, these gates control and select which time-scale dependencies are written and kept in memory cell $\mathbf{C}_n$. More precisely:

$$\mathbf{C}_n = f_n \circ C_{n-1} + i_n \circ \tanh(W_{\mathcal{A}\, c} * A_n + W_{hc} * H_{n-1} + b_c) \tag{5}$$

$$H_n = o_n \circ \tanh(\mathbf{C}_n) \tag{6}$$

with b_c the bias and H_{n-1}, H_n, the previous and current hidden states for the sequential step n. The symbol $\circ$ denotes the Hadamard product, whereas $*$ represents convolution. The term $W_{\mathcal{A}\, c}$ in (5) corresponds to the parameters of the input-to-state transitions, while W_{hc} corresponds to the internal state-to-state transitions, that is, the internal connections in ConvLSTM 1_λ from the hidden state H_n to the memory cell $\mathbf{C}_n$. The f_n or forget gate controls the amount of information that is erased, whereas i_n is the input gate that controls what must be written in the memory cell. The output gate o_n controls the amount of information that is passed from the memory to the output sequence. All these three gates, f_n, i_n and o_n are 3-dimensional tensors that for each sequential step n reduces to a matrix whose rows and columns are the processed values of A_n that must be stored or erased in the memory cell $\mathbf{C}_n$. As the gates operate over the sequence of time-scale variations, the internal memory preserves the number of time samples for all the elements in $\mathcal{A}$. ConvLSTM 1_λ is designed to learn and select the feature vectors that best inform the subsequent temporal classification modules. The output of ConvLSTM 1_λ is the sequence $\mathcal{B} = [B_1, B_2, \ldots, B_n]$, with each element signifying a new feature vector.

The second ConvLSTM, indicated by ConvLSTM 2_λ, is identical in its design to ConvLSTM 1_λ. The output of this ConvLSTM 2_λ is the feature vector sequence $\mathcal{C} = (C_1, C_2, ..., C_n)$. Its application results in a refined time-scale representation in which the interaction of the scattering coefficients across multiple scales is manifest. We observe that the physical, frequency transient phenomena are linked with the dynamics of the learnable scattering transform.

C. Time Skip Connection

In parallel to subjecting the volumetric layer to ConvLSTMs, we apply a single convolution. This is a *time-skip connection* (path B), that parses the multi-channel scattering coefficients into an image of size $(J'\, Q') \times (\text{number of time samples})$. The single convolution collapses the

multi-channel scattering coefficients into a single output value. We do not apply any pooling to preserve time resolution in the output representation.

D. Feature Fusion

The output of the branches A and B are concatenated in a new multi-channel image as follows. For each time we stack the *columns* from the images generated by branches A and B. The components for each time are labeled by λ'. The result we denote by S, essentially a matrix signifying feature fusion (in green in Figure 1).

E. Temporal Classification

The bidirectional configuration is designed to provide long-term contextual information from the entire input sequence S, one working as a forward and the other as a backward recurrence (see BiLSTM1 and BiLSTM2 in Figure 1). The number of sequential steps for the forward and backward LSTM unfolds is given by the length of the dimension $time$ in S. The bidirectionality of the recurrence allows incorporating contextual information from past and future frames, reinforcing the detection capacity of our network. Finally, the output of the BiLSTM2 is fed into an MLP to produce the prediction, that is, the predictive probability heatmap.

VI. Uncertainty and Change Quantification

The reliable detection of change in seismo-volcanic data streams remains a complex task given the polyphony associated with multiple sources, tremor background, and overlapping signals. The challenge is to find a universal approach, irrespective of the geological environment, to systematically detect anomalous behavior and change.

Changes in the distribution $D = \{(X, Y)\}$ are especially challenging to detect in streaming data [57]. Estimating the uncertainty in a monitoring framework with streaming data requires the generation of a distribution over the network's predicted outcomes, fitting with a Bayesian statistical framework [27]. The challenge in developing this in deep learning lies in providing an effective approximation to the high-dimensional parameter space of deep networks that is both fast and numerically reliable. Research work by [24] has related dropout neural networks with Bayesian statistics by sampling from multiple dropout masks to infer an approximation to the neural network's posterior distribution (for details, see appendix A). This approach has the advantages of scalability and integration with already well-established deep-learning training

methodologies. In seismology, Bayesian deep learning has been explored through a dropout approximation for phase picking, and earthquake localization [25]. Bayesian deep learning has also been applied to the probabilistic classification of events from filter-bank based features, characterizing pre- and post-eruptive periods based on different types of seismic events [27], [58].

A. *Bayesian Monitoring*

The sources driving change in a volcano are composed by an unknown number of latent, heterogeneous variables that contribute to the overall alteration during monitoring. We gather all sources of uncertainty into the *epistemic uncertainty* of our model at any given monitoring time [59]. Starting from the distilled scattering feature vectors S_t in the *feature fusion* module, MC dropout is invoked in the final two bi-directional LSTMs (BiLSTM1 and BiLSTM2) and the fully connected predictive layers (contained in the dashed square in Figure 1). That is, we fix both the scattering layers and the ConvLSTMs modules. The variational inference framework based on Monte Carlo numerical sampling and its connections with stochastic regularization techniques in deep learning are described in the Appendix A. During the training stage, invoking stochastic dropout on the last layers can simplify the variational procedure while still providing meaningful results from a Bayesian perspective. From a statistical perspective, it can be interpreted as a jointly-learning point estimation followed by a shallow BNN [60], [61]. When new data streams are presented to the neural network, the two bi-directional LSTMs and the fully connected predictive layers analyze the processed scattering feature vectors and compute an uncertainty estimate of the classification and segmentation for each of these processed vectors. This procedure permits quantifying signal variations due to any external mechanism presumably closely related to volcano dynamics.

VII. Experimental Methodology

This section introduces the studied eruptive period at Bezymianny volcano. We then provide the chronological dataset division used to obtain insight into how the recurrent scattering network performs multisource seismo-volcanic event recognition. Next, we describe the optimization procedure and the polyphonic metrics for multisource seismo-volcanic monitoring.

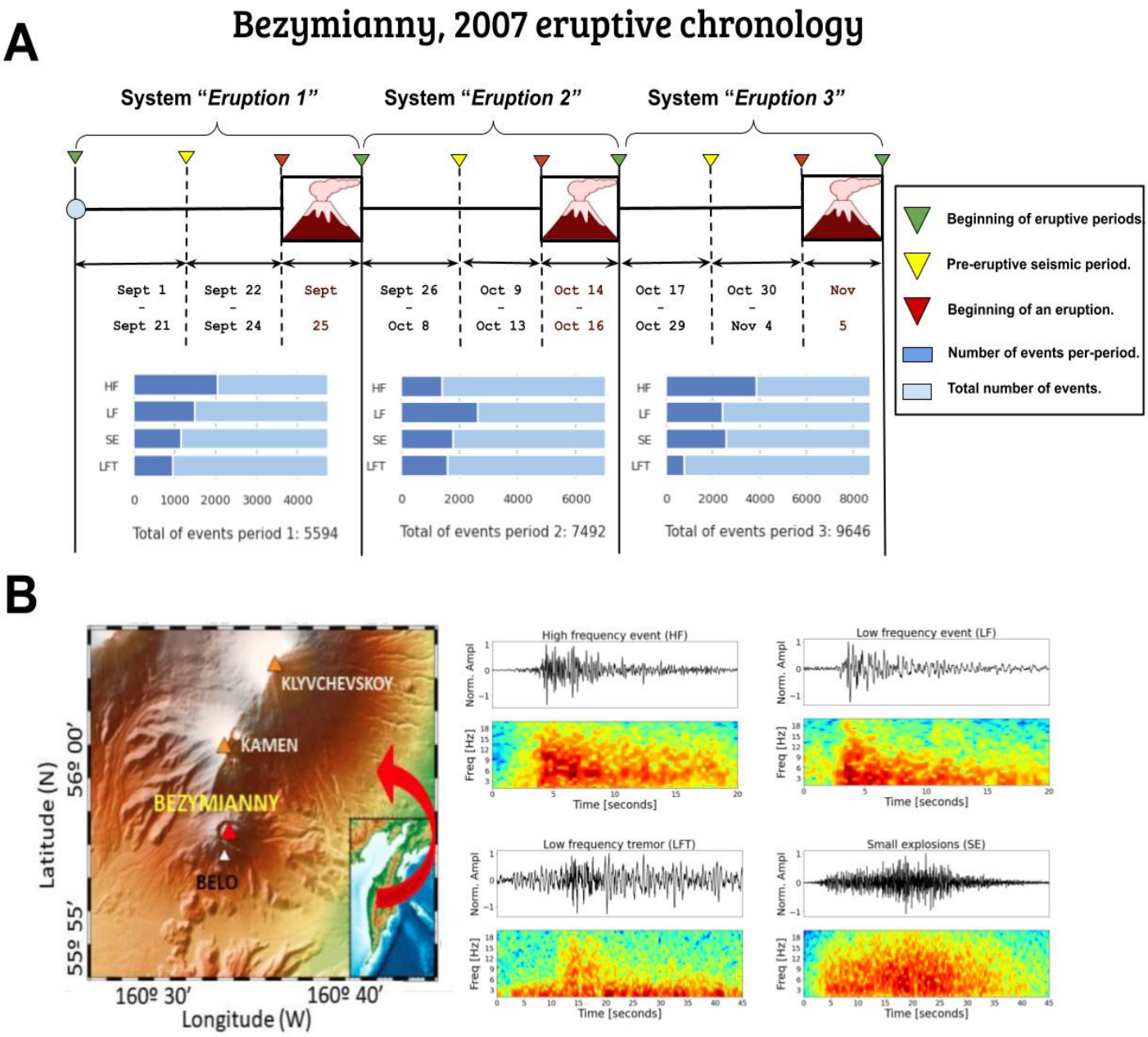

Fig. 2. **Bezymianny volcano (Kamchatka, Russia) dataset organization and geophysical signals.** (a) Dataset organization and total number of events following the 2007 eruptive chronology. Each trained system is trained with seismic data conditioned to the studied eruptive period, with green marks highlighting the beginning of each eruption. Yellow markings indicate the onset of pre-eruptive seismicity, whereas red markings indicate the beginning of an eruption. (b) Map of Bezymianny volcano representative examples of common waveforms recorded during the 2007 eruption. For each waveform type, the normalized waveform (in black) and spectrogram are depicted. For visualization purposes, all waveforms have been filtered between 1 and 20 Hz.

TABLE I

DATA-SET ORGANIZATION AND PRE-SEISMIC TEST PARTITIONS AND EVENTS

for the baseline systems.

System	Eruptive chronology datasets		Events eruption (tests sets)			
	Dev. set	Test (set)	HF	LF	SE	LFT
Eruption 1	Sept. 1 to Sept. 21	Sept. 22 to Sept. 25	233	246	374	62
Eruption 2	Sept. 26 to Oct. 8	Oct. 9 to Oct. 16	1121	1501	1289	1090
Eruption 3	Oct. 17 to Oct.29	Oct. 30 to Nov. 6	846	758	267	520

A. *Bezymianny volcano*

We selected an eruptive sequence from Bezymianny volcano (Kamchatka, Russia). The data catalog contains 3 months of daily records, from 1 September to 5 November 2007 [62]. During this period, three significant eruptive episodes (those on 25 September, 14 October, and 30 October) were reported by the Kamchatka Volcanic Eruption Response Team (KVERT), and confirmed via posterior geophysical studies [63]. Figure 2.(a) shows the chronological dataset organization and per-class data samples for Bezymianny volcano. The Figure 2.(b) shows the volcanic area and representative seismo-volcanic events recorded during this period and that comprises the studied dataset. Of the three eruptive episodes, which were all dominated by strong ash explosions and lava emissions, the second is considered the most energetic. This event continued for 2 full days, with the plume reaching 10 km in height and extending 1000 km to the southeast. Of the seismic stations monitoring the eruptive activity, we selected BELO, located near the Bezymianny dome and crater; BELO is a near-field station for which attenuation effects are expected to be minimal.

We categorized recorded waveforms according to the terminology proposed by [8]. This categorization scheme includes low-frequency (LF), high-frequency (HF), seismic volcanic tremor (SBT), surficial events (SE), and low-frequency tremor (LFT); (see figure 2). These labels presumably distinguish seismic source mechanisms induced by distinct volcanic processes. Moreover, although seismic networks can record several events per minute during volcanic crises (over periods of days to years), the association between the different event types (or labels) and the causative source process is still not uniform. Further complications arise when several processes co-occur, producing a suite of overlapping signals. Our neural network architecture utilizes all of available data at the quiescent periods, and through polyphonic detection, segmentation, and classification enables insight in the dynamics of the volcanic system.

TABLE II

BEZYMIANNY 2007 POLYPHONIC METRIC RESULTS AT ONE SECOND FRAME-WISE PRECISION.

System	Eruption 1				Eruption 2				Eruption 3			
	F1	**PR**	**RC**	**ER**	**F1**	**PR**	**RC**	**ER**	**F1**	**PR**	**RC**	**ER**
CNN	82.15	86.86	77.68	22.43	58.62	73.74	49.06	51.06	81.35	84.28	78.62	21.41
Hybrid CNN+RNN	84.18	87.22	81.36	18.71	59.48	71.37	51.11	48.72	84.48	84.13	78.99	21.04
Vanilla	85.86	87.09	84.66	**15.66**	61.97	69.07	56.19	**43.96**	84.28	86.75	81.96	**18.45**
Filters	85.35	86.72	84.07	**16.35**	61.45	69.24	55.24	**44.85**	84.62	86.23	83.06	**17.26**
Knots + Filters	85.29	86.64	83.98	**16.47**	59.71	65.09	55.15	**45.44**	84.45	86.36	82.62	**17.21**

We subdivided the dataset into time intervals based on periods of seismic unrest. During quiescence periods the level of seismic activity was very low. During periods of seismic unrest, occurring before each eruptive episode, seismic activity increased in number of events and energy. After each eruption, seismicity subsided, leading to another quiescence period (see Figure 2.a). This dataset division strategy also considered representative numbers of seismic events for training and testing. Table I shows the dataset composition and the pre-eruptive test partitions for the baseline systems. For the first eruption (on 25 September), the training data covered the period from 1 to 21 September. For the quiescence period between the first and second eruptive events, the training data covered the period from 26 September to 8 October. For the quiescence period between the second and third eruptive events, the training data covered the period from 17 October to 29 October. We tested our approach for each period independently, so as to gain deeper insight into the performance of our neural network as a monitoring tool.

B. Optimization procedure

The training is done, for all the data sets, with the raw seismic waveforms in samples. The number of knots k, the number of octaves J, and wavelets per octave Q are carefully chosen after some signal analysis of the input waveforms. We selected, for the *first scattering layer L1*, $k = 8$, $J = 6$ and $Q = 7$, and for the *second scattering layer L2*, $k = 2$, $J = 5$ and $Q = 5$. These parameters give enough bandwidth and frequency resolution to provide a scattering representation where the main frequencies of the events are discernible from the background tremor. After the concatenation of both scattering feature vectors, a volumetric pooling of size $(1, 1, 100)$ is applied to remove redundant information. Both $ConvLSTM_\lambda$s contain 16 filters, with a kernel size of

$(5, 5)$ and strides $(1, 1)$. The temporal skip connection is a single filter with a kernel size of $(1, 1)$ to perform dimensionality reduction. The successive BiLSTMs layers are 92 hidden units in each direction (128 in total for each BiLSTM). A per-class *sigmoid* activation function is applied to compute the final detection matrix: 5 hidden units for Bezymianny and Mount Saint Helens. The scattering transform module produces a sparse vector of 30000 samples pooled by 100 to achieve a 1s time resolution. Layer normalization is applied after all successive recurrent layers. The dense, predictive layer contains a sigmoid activation function in which a binarization threshold of 0.5 over the per-class event activity probabilities is applied to produce the event detection matrix allowing multiple class memberships to be detected. We adopt the polyphonic training framework in [44]; with binary cross-entropy loss function. We employ gradient clipping and recurrent weights regularization (0.001) only for the experiments involving joint learning of knots and filters [55]. We employ early-stopping with a patience interval of 5 epochs over 300 training epochs to prevent over-fitting. Furthermore, we compare the performance of the network introduced, here, with the performance of two baselines using the same input, that is, the raw waveform. First, we use a standard CNN architecture widely used in detection of seismic signals, with five convolutional layers ($64/64/128/128/256$ filters) followed by four dense layers ($512/512/512$ hidden units) and a sigmoid output layer (5 units). Second, we use an adaptation of the baseline CNN aiming to mimic the capabilities of our architecture to capture long-term seismic signal information [64]. This second model, which we refer to as *hybrid CNN+RNN*, keeps the five convolutional layers but replaces the previous three dense layers with three RNNs ($96/96/96$), followed by a sigmoid output layer (5 units).

The transfer learning experiments with Mount Etna entail the adaptation of the architecture to new conditions. First, we replace the temporal classification module with a new one, with the last per-class *sigmoid* activation function containing 4 hidden units, corresponding to the seismic categories in Mount Etna: SBT, HF, MX, and LF.

C. Monitoring metrics

In multi-source seismo-volcanic recognition, the reference at any given time is not a single class but multiple events simultaneously that must be accounted for individually. Therefore, one of the difficulties in measuring monitoring performance lies in recognizing multiple seismic events that can have different lengths from the ground truth labels. This is especially important in seismo-volcanic event recognition, with signals such as HF and LF spanning a couple of

seconds compared to LFT, sustained seismic signals with a duration from minutes to hours. The monitoring metrics need to consider the prediction alignment with ground truth annotations. To circumvent these problems, we adopt the event-based error rate (ER), derived from other acoustic domains such as speech recognition [65]. The event-based ER considers the temporal position of each sound event and the labels when comparing the system output to the ground truth reference. Seismic events with correct temporal position but incorrect seismic categories are counted as substitutions (S). Insertions (I) and deletions (D) are seismic events unaccounted for as correct or substituted in system output, respectively [65]. The total error rate is calculated considering the frame-wise I, D, and S over the total number of frames K, with $N(k)$ the number of seismic events that are active in the reference frame k:

$$ER = \frac{\sum_{k=1}^{K} S(k) + \sum_{k=1}^{K} D(k) + \sum_{k=1}^{K} I(k)}{\sum_{k=1}^{K} N(k)} \tag{7}$$

We include the polyphonic audio metrics from [65]: precision (PR), recall (RC), and the F-measure (F1), computed frame-wise. The PR quantifies the rate of positively classified frames, whereas RC measures the sensitivity of the system to recognize correct frames [27]. The F1 is a weighted average between the PR and RC:

$$F1 = \frac{2 \cdot PR \cdot RC}{RC + PR} \tag{8}$$

These metrics allow us to obtain the error made by the model for the specified temporal precision, in our case, one second, in addition to the classification prowess of the model, for the set of defined events.

VIII. Results

This section has been structured into three main subsections. The first subsection of this study demonstrates the monitoring capabilities of the implemented architecture and multisource (or polyphonic) monitoring tasks. To do this, we select three main eruptions of the Bezymianny volcano during the 2007 eruptive sequence. Then, using the polyphonic system, we present a complete application of how uncertainty and its adjustment with power-laws or exponential functions can predict eruptive processes successfully in Bezymianny volcano. We then study the exportability of the polyphonic system to other volcanic scenarios by doing a blind test with data from Mount Saint Helens (Mt. St. Helens), a volcano similar to Bezymianny.

The second part of this study demonstrates the ability of the system to perform predictive tasks in a completely different scenario, the Mt. Etna volcano. We first study how the uncertainty behave in a blind test after we switch the volcano type. We then perform a transfer learning approach and extend this study to use the uncertainty as a predictive element of eruptions in the Mt. Etna volcano.

A. Monitoring Results: Bezymianny

We follow the nomenclature vanilla, filters, and knots+filters to distinguish the learnable options of the scattering network module in our E2E approach. Table II shows the polyphonic metrics for the test (eruptive) partitions, trained using data from the quiescent periods before each of the respective eruptions, with a frame precision of 1 s, as well as for the baselines and the proposed architecture. Our approach yielded excellent performance for each of the eruptions. The PR, RC, and F1 metrics attained high values for all of the scattering network options. Furthermore, these results show that our proposed polyphonic approach yields better RC and lower ER during all three quiescent periods when compared to two traditional baselines. Notably, our proposed learnable configuration resulted in a high number of correctly detected frames at 1 s precision, along with the effective event categorization of localized scattering information (i.e., high PR). This performance is rooted in scattering learning stability inside the recurrent architecture. It is important to emphasize that the three eruptive processes (i.e., quiescence, unrest, and eruption) occur over a very small interval of time, but each involves different physical mechanisms that produce different seismic sources. In each process we have a first interval (quiescence) where the volcanic system seems to be in an apparent rest. The system quickly enters unrest and shows an acceleration of all the seismic observable culminating in an eruption. Finally there is a period of return to stability to enter the period of quiescence again.

Our polyphonic approach yielded high RC and PR during all three quiescent periods. The recurrent scattering architecture mitigates one of the main challenges in seismo-volcanic monitoring owing to its ability to adapt. Notably, in the second eruptive period, during which multiple physical mechanisms were active at the same time, the network's performance showed high PR and F1 scores, with RC being at an acceptable level for seismic signal detection and identification. The network successfully dealt with significant variability in the durations of events and with overlapping events. More than 50% of frames were correctly detected and assigned with high precision to multiple and co-existent seismo-volcanic classes. As an aside, we achieve favorable

performance metrics for event detection as a single task (see appendix B). These results confirm that our approach has significant advantages over more traditional monitoring protocols [14].

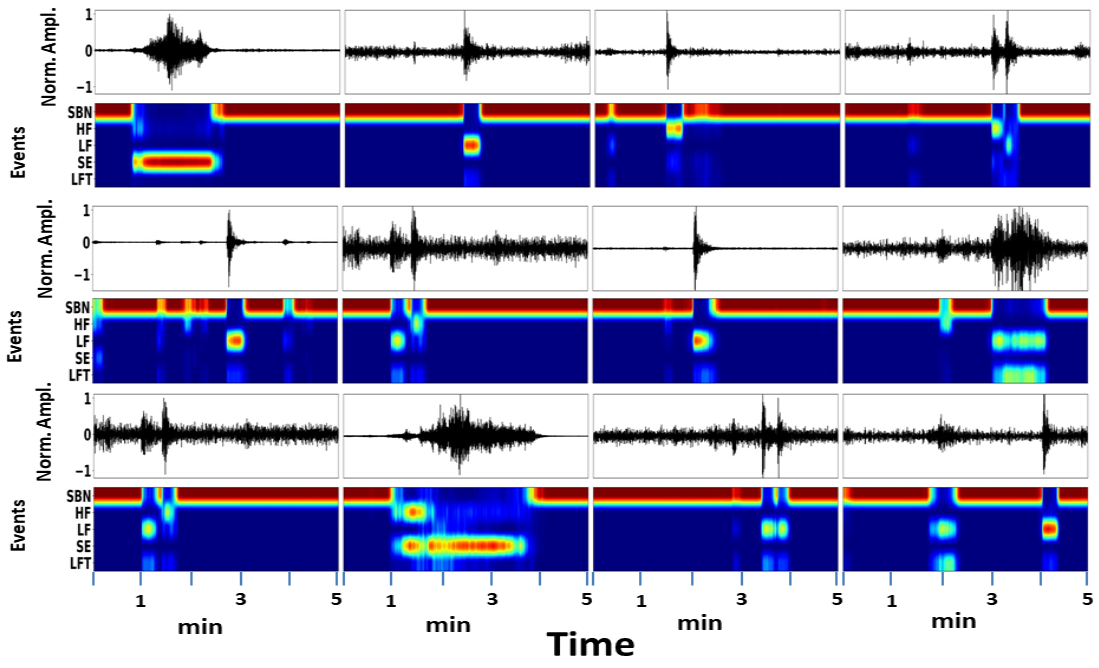

Fig. 3. **Pre-eruptive heat maps for Bezymianny volcano (Kamchatka, Russia) from 21 to 24 September 2007**. Normalized waveforms and per-class probabilistic heat-maps for the pre-eruptive sequence at Bezymianny volcano from 21 to 24 September 2007. The neural network systematically isolates fundamental seismo-volcanic events, even when they occur in rapid succession. The scattering transform can identify multiple active sources—including low frequency (LF) earthquakes—and uses the recurrent structure to unmask the coupled mechanisms from the background noise or tremor. The polyphonic multi-output provides invaluable geophysical information about potential sources and refines our understanding of volcanic unrest.

B. Predictive Heatmaps: Bezymianny

We use probabilistic heat-maps as an intuitive representation to verify if the learnable scattering transform has effectively adapted to performing polyphonic event segmentation; that is, these maps provide visual evidence that our neural network can detect, segment, and classify simultaneous seismic waveforms. Figure 3 shows probabilistic heat-map predictions with 1 s resolution for the pre-eruptive seismo-volcanic data stream of 22–24 September 2007. Attaining accurate segmentation is critical for seismo-volcanic monitoring; the duration of a recorded event is directly linked to the type of source mechanism that exchanges energy with its surroundings. For example, low and high-frequency events have short durations (up to 30 s), while volcanic

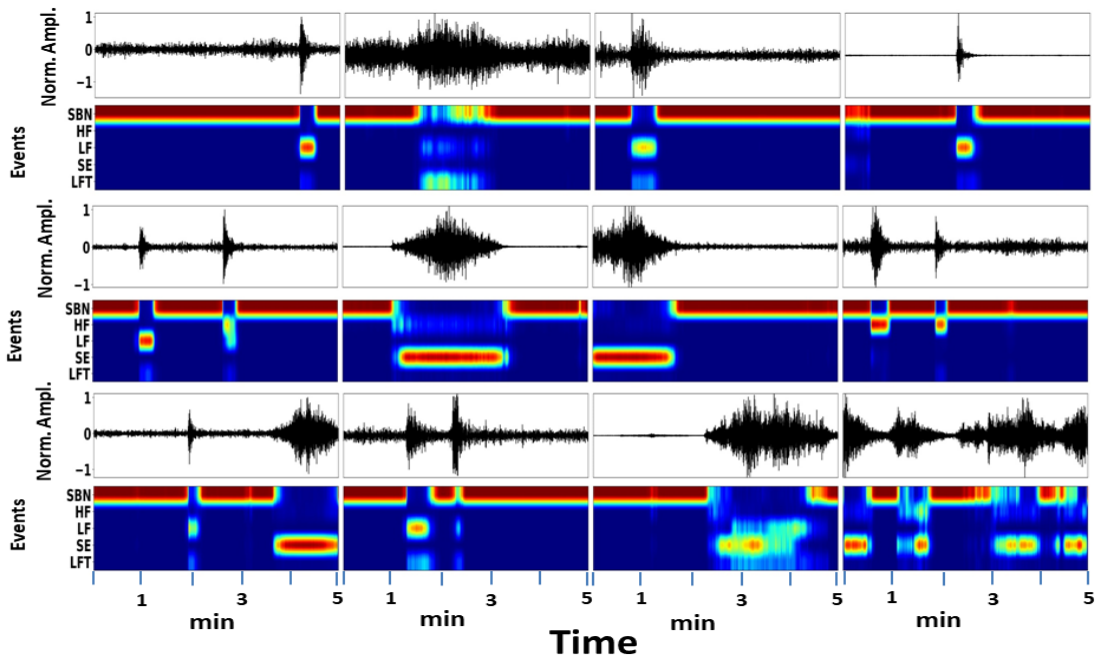

Fig. 4. **Heat maps for the 24–25 September 2007 eruptive period at Bezymianny volcano (Kamchatka, Russia).** Normalized waveforms and per-class probabilistic heat-maps for the eruptive sequence at Bezymianny volcano from 24 to 25 September 2007. The first two rows represent reported mild pre-eruptive seismicity (24 September). Earthquakes are classified as low frequency (LF) and high frequency (HF) events, with occasional surficial debris processes (SE) identified during dome inflation. The bottom row (25 September eruption) represents co-existent debris processes and low frequency tremor (LFT) during the main eruptive episode; the LFT component is recognized over the recorded exogenous signal, effectively demonstrating how a polyphonic approach enhances volcano monitoring by providing all potential sources active at any given time.

tremor can last from minutes to hours. Isolated events are generated by sudden fluid (water or magma) pressure changes (e.g., LF events) or the sudden release of energy due to rock rupture (e.g., HF events). Meanwhile, volcanic tremor arises from the sustained exchange of energy with the surroundings as the result of multiple pressure pulses within a fluid owing to bubbles, energy transients due to fluid flow, and/or other dynamics. Given their relationships with the physical processes acting at the source, information inferred from each seismic event is equally important but is only partially encoded in its duration and frequency content. Our recurrent scattering network captures long- and short-term dependencies at 1 s resolution, departing from standard approximations that often require a posteriori analysis. The multi-resolution analysis underlying scatter-gram variation enables the network to learn the set of scattering coefficients with maximum intra-event information. The polyphonic approach is evident in the multiple frequency

components isolated as sparse probabilistic maps, concentrating higher output probabilities to higher spectral content. For example, the waveforms of HF and LF have durations that are longer than the actual duration of the source because they are composed of direct waves incoming from the source (being the most energetic portion of the signal) and successive arrivals of waves coming from the resonance of the system, or scattered waves generated by heterogeneities of the medium, among other possible phenomena. Monitoring challenges at Bezymianny include the energetic background tremor and noise level. However, our results show that the effect of these on intra-class separability is negligible.

In Figure 4, we show seismic traces from the first eruption (25 September). Our approach succeeded in isolating and categorizing the presence/absence of volcano-seismic signals despite the per-class probabilistic heat maps showing many recorded concurrent processes. As an example, elevated confusion occurs within the lower left seismic trace, in which there exist many simultaneous processes with mixed frequencies, influencing intra-class separability and predictive performance. The recorded SE activity could be caused by increased deformation through dome growth paired with energetic rockfalls recorded prior to an eruption.

In Figure 5, we show a set of polyphonic heatmaps for data streams from 9 to 16 October (first two rows) and 30 October to 5 November (last two rows). Polyphony is evident in the heat maps for all of the earthquake transients; our neural network detected LF and HF events co-existent with LFT. Furthermore, for the third eruptive period (the last two rows), polyphony is also clearly visible, with both HF and energetic SE classes. As mentioned, in each eruptive period, a series of changes in the seismic wavefield are used for forecasting. Alongside every data stream that has been extracted and classified is the ever-present volcanic tremor. The tremor wavefield also changes as the volcano evolves, and in some cases can become highly energetic and relevant from an identification and classification point of view. For example, in the second eruptive period, such a background signal was manifested and identified as the LFT. These results show that our neural network succeeds in intra-class separability of earthquake transients and background signals, even under intense seismic fluctuations. Our neural network learns, without supervision, the energetic onset/offset times of events, being sensitive to subtle fluctuations in seismic transients from background noise and surficial events. Indeed, our neural network can operate as an automatic event detector with the maximum probabilities of temporally aligned and concentrated seismic signals. An interesting observation is that when our approach switches the highest probabilities from volcanic tremor to a different class, a low probability emission is

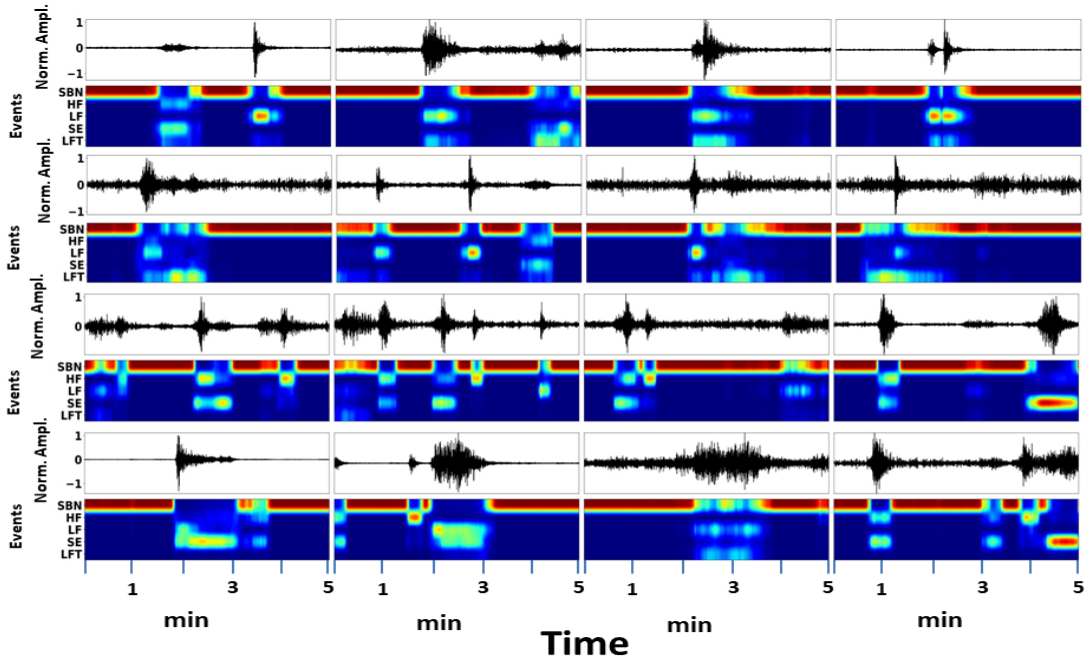

Fig. 5. **Heat maps for Bezymianny volcano (Kamchatka, Russia) from 9 October to 5 November.** Normalized waveforms and per-class probabilistic heat-maps for Bezymianny volcano from 9 October to 5 November 2007. Note that the waveforms present concurrent processes. Multisourced signals arise in the low-frequency components, as the frequencies of low frequency (LF) and high frequency (HF) events (i.e., earthquakes) overlap with the those of low frequency tremor (LFT). Similarly, surficial events (SE) are correctly segmented, as their arrivals and subsequent envelopes differ from those of earthquakes.

generated before the potential arrival of the event in this different class, as it can be noticed from figures 3, 4 and 5. The learned scattering transform provides a physics-informed identification of the distortions in the near seismic wave field. Such information serves as an a *a-priori* estimate of the phase arrival time, covering an extensive low-probability range in time until the energetic arrival. These probability switches are consistently present despite signal-to-noise ratio (SNR) and spectral differences between seismic signals and background tremor.

C. Drifts in Pre-eruptive Epistemic Uncertainty

Epistemic uncertainty is uncertainty in model predictions reflecting the degree of data knowledge of the model [27]; here, the model is the classification module of our designed recurrent scattering network. The data stream is mapped to the tensor S_t, which is forwarded to the temporal classification module (corresponding to BiLSTM1 and BiLSTM2 in Figure 1). The tensor S_t contains, for any given frame time t, the mixture of all the scattering coefficients that

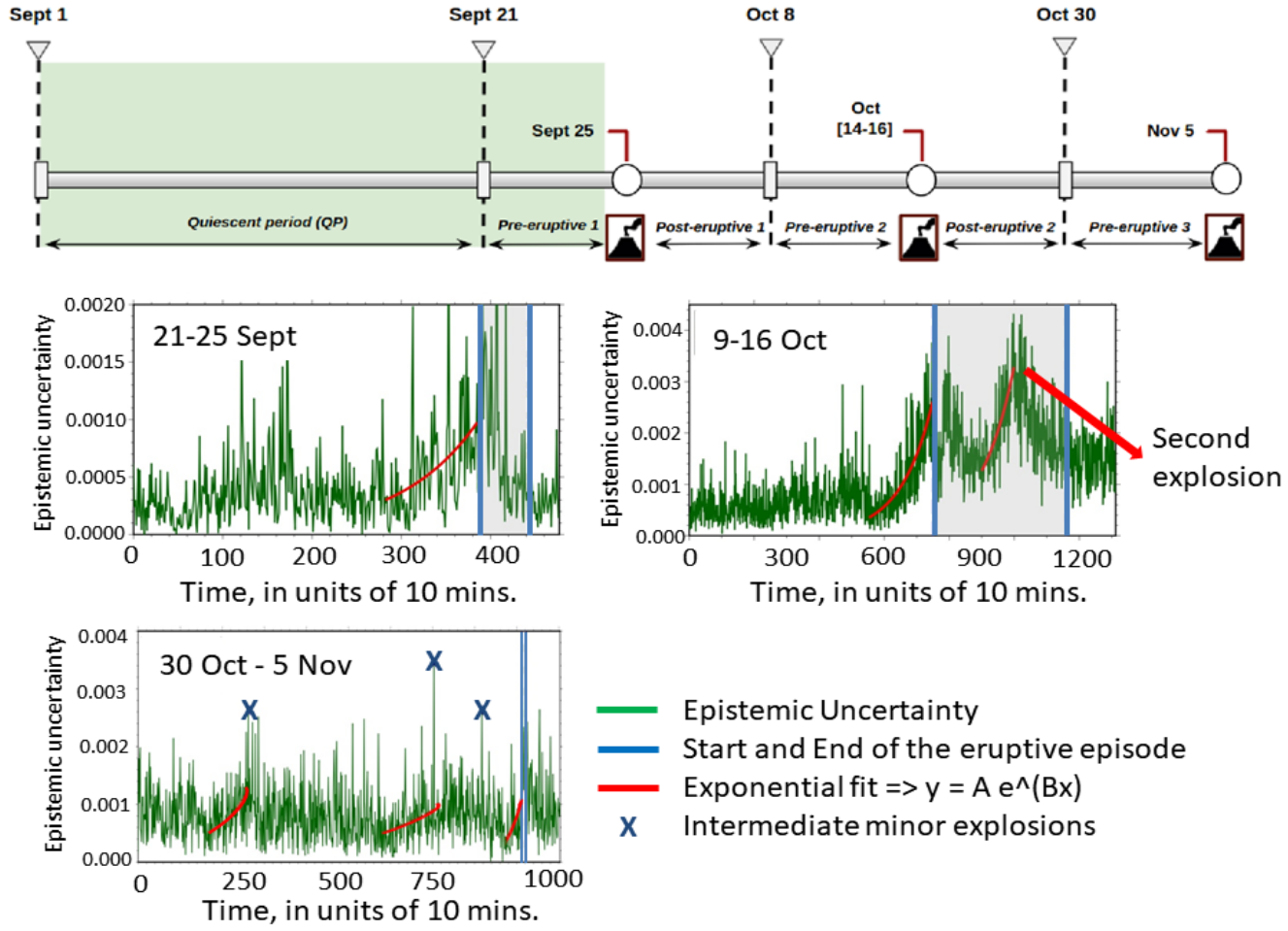

Fig. 6. **Temporal epistemic uncertainty variation and detected drifts in pre-eruptive and eruptive sequences at Bezymianny volcano.** The main reported eruptions are indicated in gray. Multiple explosions for the third eruptive episode are indicated by ×. Fits to power laws preceding eruptions and explosions are indicated in red. The Bayesian recurrent scattering network successfully detected change prior to the main eruptions, revealing that changes in the uncertainty are related to unrest.

best explain the raw observable signal. From a Bayesian perspective, S_t is used as an embedding in which compressed information can be used to approximate epistemic uncertainty. The greater the data availability, the better we can approximate the training data distribution, with lower epistemic uncertainty values. Regardless how the support data distribution is approximated, the Bayesian approach is always conditioned to the training data distribution, in which probabilistic shifts are detectable via epistemic uncertainties [27]. For our learned scattering recurrent network, and for each day in the validation/test streaming data sets, we computed the predictive map from each of the 20 stochastic Monte-Carlo dropout samples, chronologically at 10 min time intervals.

Our hypothesis was that a drift in epistemic uncertainty would be directly correlated with changes in the volcanic processes.

Figure 6 depicts the results for the pre-eruptive and eruptive sequences for each of the eruptive periods considered here; each point on the horizontal axis corresponds to a 10 min interval and eruptions are marked in dark blue. In the pre-eruptive sequences, the predictions are remarkably robust, reflecting the invariances encoded by the scattering transform module of the network. Drifts in epistemic uncertainty preceding eruptions signify detectable changes in volcanic processes. Remarkably, these drifts exhibit power-law behavior in time, directly indicative of metastable behavior. Prominent peaks in epistemic uncertainty correlate with high-energy events that are, clearly, not present in the training data sets. The general time fluctuations in the uncertainty mostly relate to the complexity associated with polyphony in the predictions. The uncertainty may not return to values encountered in a pre-eruptive period; this can be attributed to changes in material properties and structural conditions after eruption. In this case, our deep neural network might need to be retrained, for example, through transfer learning. The plots in Figure 7 are consistent with those of previous studies [62], [66], [67].

For the first sequence (21–25 September), only minor pre-eruptive seismicity was recorded, with no significant post-eruptive changes. Our epistemic uncertainly shows a sustained drift in the form of a power law (in red) for 208 time units (or 34.6 h) prior to the eruption; the uncertainty peak coincides with the main explosion. This first eruption ends at 442 time units (73.6 h), while the uncertainty returns values encountered in the pre-eruptive period. For the second sequence, the eruption of 14 October was preceded by increased seismicity and tremor, including relatively large magnitude earthquakes. The eruption was characterized as explosive and lasted until the 17 October. The uncertainty again shows a clear drift following a power law (in red) for 197 time units (32.83 h) prior to the eruption on 14 October. A second, short-lived but high-energy eruption is clearly visible in the uncertainty on 15 October. Preceding this second eruption, the drift in epistemic uncertainty is attributed to an increment of tremor amplitude. Finally, the third eruption at Bezymianny, that on 5 November, was reported as weak with minor fumarole activity. The activity could not be confirmed owing to weather conditions, but it was postulated that a sequence of explosive events occurred. We can confirm such explosive activity based on two spikes in the uncertainty at 700 and 810 time units along with other minor explosive activity that could have preceded the main explosion at 914 time units. Again, the individual explosions were preceded by a power-law drift in the uncertainty (in red) at 63 and 123 time units (10.5

and 20.5 h) for the first and second individual eruptions, respectively.

D. Power-law and Quasi-exponential Uncertainty Curves

In volcanology, eruption prediction methodologies include one based on modeling the rate of change of selected observables using a differential equation with exponent, α [68]. Exponential behavior occurs if $\alpha = 1$, and power-law behavior occurs otherwise. We carried out careful fitting [69] of α to the epistemic uncertainty obtained for the three eruptive periods (Figure 6; note that the exponents of the eruptions are not 1), and argue that α estimated from epistemic uncertainty is indicative of eruptive type.

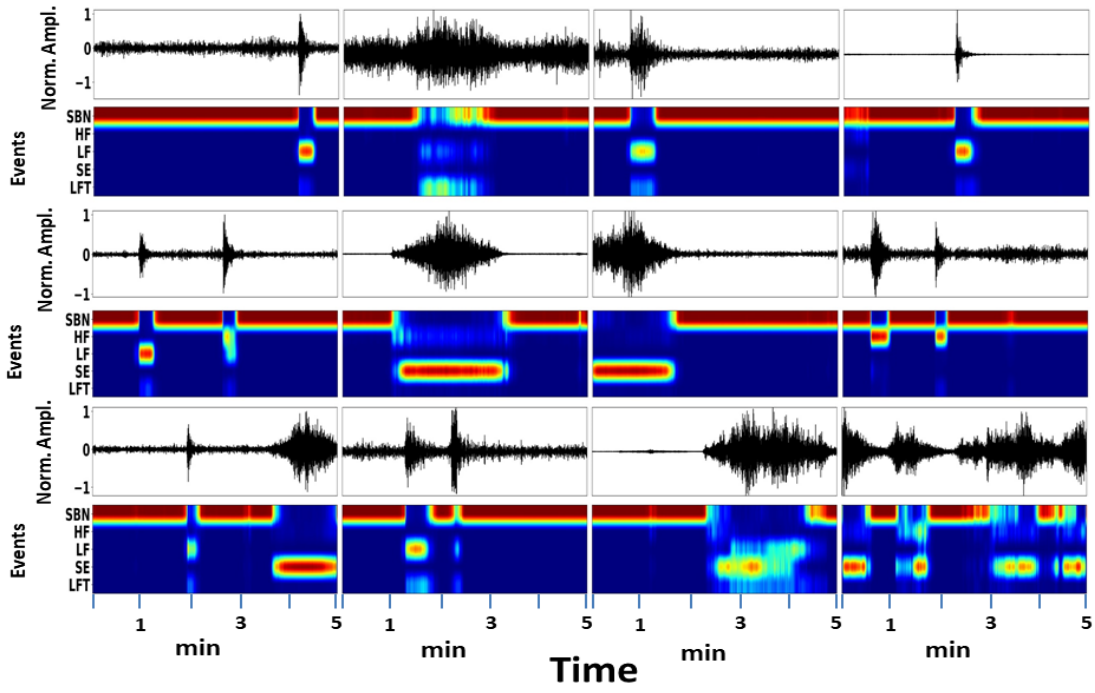

Fig. 7. **Generalization and exportability test of the implemented system using data from Mt. Saint Helens.** Normalized waveforms and per-class probabilistic heat-maps for Mt. St. Helens volcano from 0000 PDT to 23:59 PDT on 9 September 2005. Seismic waveforms are from station JUN. Seismic signals are characterized by regularly spaced events (or "drumbeats").

A power law defines a polynomial relationship between two quantities, in which relative changes in one leads to a proportional change in the other, modeled by exponent α. In our volcanic setting, the quantities are time and uncertainty. Power-law models are often referred to as a "material failure forecast" and can be used for the prediction of volcanic eruptions

[70]. Using this approach, it is more important to determine if there is an acceleration in the change than if there is a change in itself; therefore, the second derivative is critical. As such, an exponential greater than one is often interpreted as a precursor element. Power laws are universal across volcanoes, permit scaling, and implicitly encoding acceleration. Their use is not new [8], [70]; however, by using uncertainty as an early warning element, our method can reduce the uncertainty whether or not there is an increase in the homogeneity of the data. If the uncertainty grows, it implies the appearance of new classes of events or that the events within a class have begun to differ from each other. Therefore, an increase in uncertainty, which initially reflects a lack of classifier quality, can be used as a good indicator for early warning of volcanic eruptions. It is important to note that this approach -measuring the variation of the uncertainty- can be exported between volcanoes. While the nominal values of the uncertainty will differ among volcanoes, observing the derivatives is more important than the absolute nominal value of the uncertainty.

Figure 6 shows changes in uncertainty before each of the three eruptive events. The nominal values of uncertainty differ, but similarities can be observed in the acceleration processes immediately before each eruption. Among the three eruptions, field observations indicated that the second was the most energetic, while the third was the less energetic. The shapes of change for each uncertainty curve differ and are associated with the energy of the eruption. Both the forecasting time and exponent of the second eruption exceeded those of the other two events. For the third eruption, we identified potential previous failures and a law much closer to 1. These observations are promising as they support the use of this approach for early warning.

E. Blind Test: Mount Saint Helens, 2005

Bezymianny and Mount Saint Helens (Mt. St. Helens) share a similar composition and eruptive style; both primarily andesitic and present a classic stratovolcano-type morphology. Both produce explosive eruptions of high viscosity magma associated with a high rate of volcano seismicity.

The seismic data from Mt. St. Helens were gathered during a vent-clearing phase recorded from 9 September (starting from 0000 PDT) to 5 October 2005 at the JUN station. This seismic sequence comprises regularly spaced earthquakes, of very similar magnitude and waveform also known as drumbeats, resulting from stick-slip motion of a conduit spine. During the vent-clearing episode, the event rate increased to up to 3 earthquakes per minute, towards a sustained eruption. Figure 7 shows the polyphonic results for the blind-test.

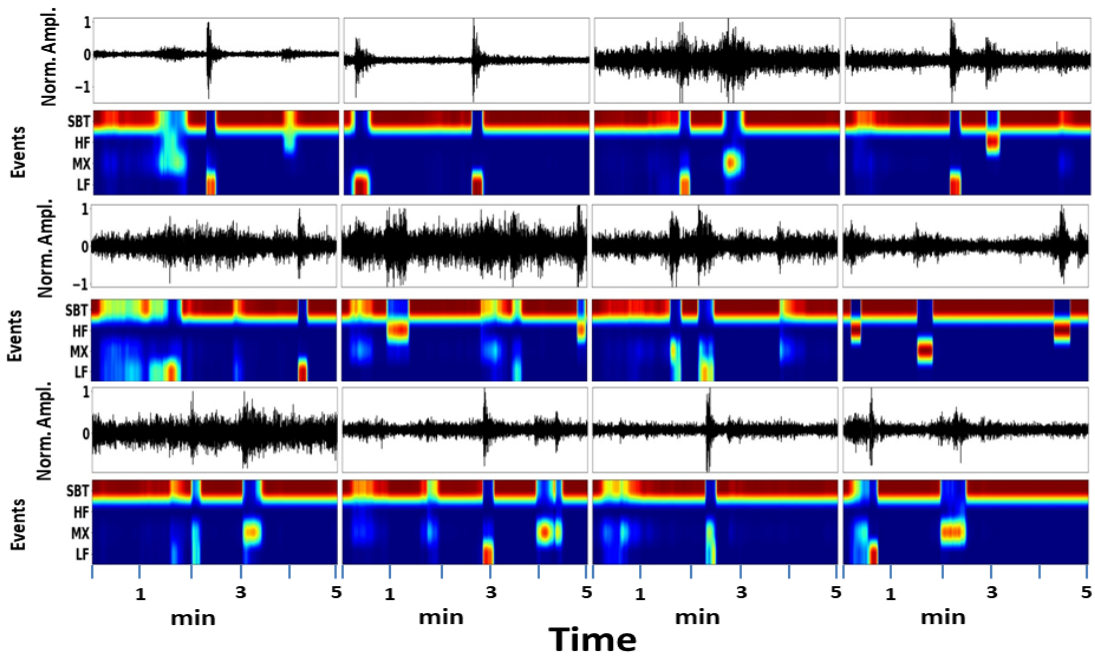

Fig. 8. **Generalization and exportability test of the implemented system using data from Mt. Etna.** Normalized waveforms and per-class probabilistic heat-maps for Mt. Etna from 14 to 24 July 2019. Seismic waveforms are from station ENCR. The polyphonic segmentation of the events permitted the identification of low frequency (LF) and MX frequency events, demonstrating the existence of coupled mechanisms, such as hybrid-type events.

For explosions registered on 10 September 2005, although the arrival of the waveform is clearly visible and detectable by less sophisticated algorithms [71], it is within the polyphonic categorization where similarities across volcanoes are perceivable. Our pre-trained network can detect events that share waveform properties with those at Bezymianny, and, despite the differences in geophysical processes, it successfully highlights the current mechanisms via probabilistic heatmaps. We note that with the pre-trained network we reliably detected the earthquakes in a polyphonic setting, assigning a high probability to LF, HF, and SE classes. These experimental results demonstrate the success of the proposed methodology for volcanoes that share similar geophysical features, and thus being easily exportable between volcanoes without re-training the whole architecture with a pre-existing dataset.

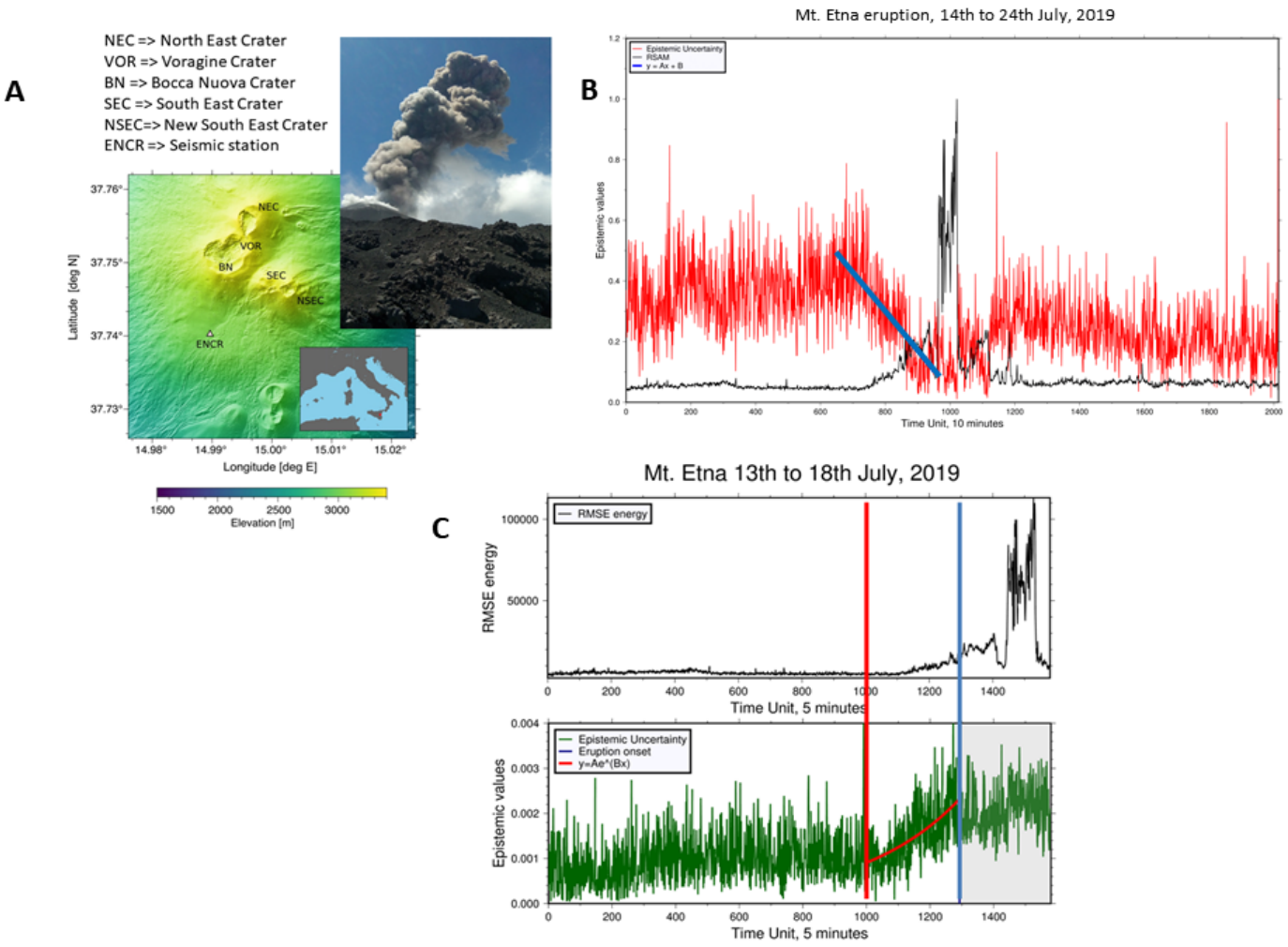

Fig. 9. **Switching volcano and eruption style with transfer learning:** (A) Image of the Mt. Etna eruptive column from on 18 July 2019, along with a map of seismic stations and the location of the ENCR station shown. (B) Temporal evolution of the uncertainty for an exportability test to Mt. Etna using the pre-trained system on Bezymianny. The blue line represents the linear fit when uncertainty drops to zero. Observing this line indicates in advance when the more explosive episode of the eruption will start. (C) RMSE energy and temporal uncertainty variation for the 2019 eruption after the system has been trained. The red and blue lines represent the start and end times in which change is detectable.

F. Transfer Learning, Switching Volcano Type: Mt. Etna

We also applied our Bezymianny-trained deep neural network to data from Mt. Etna. The internal structure of Mt. Etna frequently changes owing to intra-crater volcanic activity [72], [73], [74], [75], which provides a challenge for our approach of detecting change with epistemic uncertainty. The significant changes in eruptive activity and correlation with other volcanic data

are described in [76]. We used data gathered at the south-east of Mt. Etna (Bocca Nuova, BC) during the paroxysm recorded from 4 to 24 July 2019 at the ENCR station. The main eruption occurred at 23:09 UTC on 18 July. The seismic events and style of the eruption were fundamentally different from those at Bezymianny and Mt. St. Helens. The activity comprised explosive degassing preceding ash-rich explosions. Signals were characterized by HF and LF events, along with a new class of MX events, a hybrid frequency event that shares the spectra of LF and HF events. This new volcanic scenario has the following classes: HF, LF, MX, and SBT; spanning a total number of events, from 4th to 18th July 2019, 2528 HFs, 3118 LFs, and 5315 MXs events.

G. Uncertainty blind-test exportability

As observed with Mt. St. Helens, our architecture and approach can be exportable and produce heatmaps to another volcano when no prior knowledge is available. Now we wonder if the study of the temporal evolution of uncertainty can be exportable without having prior knowledge of the new volcanic scenario. To do this, we performed a blind test for the selected Mt. Etna data using the pre-trained system on the quiescent period (system "eruption 1") in Bezymianny volcano. We first window the data stream as in the Bezymianny case but average the estimated uncertainty for 10 minutes. No training or additional fine-tuning of the system is further performed in this test.

Figure 9.B represents the temporal evolution of this exported uncertainty along time. The main observation is the uncertainty drops to zero when the main energetic moment of the eruption will happen (observing the sudden increase of the energy), i.e. the network is effectively tracking the frequency bands of the seismic data stream through time, which is why the uncertainty remains relatively high at the beginning and drops to zero towards the eruption. The results of the blind test do not indicate when the eruptive process will begin but when the most energetic event will occur. If a simple linear fit is performed to the decay of the uncertainty in the blind test (blue line in figure 9.B), we observe that this line marks precisely the moment of greatest energy of the eruptive process. This observation is fundamental because although it does not say when the eruption will occur, it does forecast when the paroxysmal moment of the eruption will occur. This observation cannot be generalized yet, but it is a significant result and opens a new venue to the forecasting processes of volcanic eruptions. From the perspective of physical models that could explain this behaviour, we do not yet have a conclusive answer. We believe that it is

TABLE III

MOUNT ETNA 2019 PRE-ERUPTIVE PARTITION AND POLYPHONIC METRIC RESULTS

(a)

Pre-eruptive	**Hours**	**HF**	**LF**	**MX**
From 14th to 18th July	96	234	746	3934

(b)

System	**F1 (1s)**	**PR (1s)**	**RC (1s)**	**ER (1s)**
Full-training	**73.12**	72.46	73.78	**30.97**
Transfer learning	**73.67**	73.98	73.35	**27.48**

associated with the fact that at the paroxysmal moment of the eruption, the system begins to register again a high content of high-frequency signals, compared to the pre-eruptive moment where low-frequency events dominate. Therefore, in the blind text, as we approach the explosive paroxysm, high frequencies dominate again in the seismic data stream, similar to those already-known already known by the system. Hence, the exported experience with the uncertainty tends towards zero. We must insist that this test of exportability and the success as a predictor element (onset of the eruption and paroxysmal moment) is a promising result, and tests should continue in new volcanic scenarios to see it as a universal tool.

H. Transfer learning on Mount Etna

Applying our developed solution on different volcanoes can be done very simply by using our pre-trained model and only adapting the parameters of the last layers in the architecture. We adopted a transfer learning approach by replacing the whole temporal classification module with two new bi-directional LSTMs and a dense layer with the number of classes at this volcano (HF, LF, HYB, and SBT). Hence, the temporal classification module (dashed-square) in Figure 1 has been substituted by a new temporal classification module; but changing the number of hidden units to the same number of classes in this volcano. The rest of the seed network remained unmodified, that is, the recurrent scattering layers, the volumetric layer, the ConvLSTM 1_λ and ConvLSTM 2_λ; the time-skip connection and the feature fusion module. We followed the same training procedure as that used for Bezymianny and retrained the new temporal classification layers using the quiescent period of 10 days at Mt. Etna (4–14 July). We also train a system from scratch with this available data, that we refer as *full-training*.

Table III shows recognition results for the full-training, and with the transfer learning procedure. The transfer learning system shows better performance compared to the model trained from scratch. This is expected as the network is reusing accumulated geophysical knowledge from previous seismic waveforms at Bezymianny volcano, yielding higher PR and F1 scores with lower ER. Therefore, the transfer learning system is particularly suitable for new recognition tasks in different eruptive styles, where fine-grained information is required. Figure 8 shows per-class heatmaps for the eruptive sequence recorded at Mt. Etna. Our system detected LFT and MX classes attributed to fluid interactions (i.e., tremor and hybrid events). Note that the MX class is well recognized and distinguished from the SBT class. The principled recognition of MX events is present for all of the depicted waveforms, confirming that the network has adapted to a new domain in which the broad spectra of seismo-volcanic events is different from that of the seed dataset. As Mt. Etna volcano has different volcanic dynamic compared with Bezymianny or Mt. St. Helens volcanoes, there can be different geophysical interpretations for seismic signals with the same label. However, while the target dataset is significantly different from the original dataset from a machine learning perspective, the pre-trained model already has learned features relevant to the monitoring problem. These results show that the proposed network can reliably detect seismo-volcanic signals from different origins after transfer learning.

Figure 9.C depicts the temporal variation of the epistemic uncertainty for period before the main eruption. The root mean square (RMSE) energy is computed for the filtered trace between 0.5 and 12 Hz. First, it is noticeable that the uncertainty shows a gradual drift, following a power law (in red), at 1005 time units (or 23.91 h). Note that the energy does not exhibit such a clear drift from the usual background energy until 200 time units (7.6 h) prior to the main eruption. As discussed, it is not possible to compare nominal values of uncertainty from one volcanic system to another. However, we use uncertainty to define when our system can be considered sufficiently well trained and adapted to the seismic data domain. One possible indication is that the uncertainty value remains stable regardless of whether new events are used within the trainer system. If the uncertainty remains stable when increasing the number of events, the system is considered well trained. For Mt. Etna, the uncertainty at the main eruption is similar to that of the second eruption at Bezymianny. The fact that it remains constant prior to the eruption implies that the system is sufficiently well trained for use with the Mt. Etna dataset.

1) The evolution of uncertainty is a good indicator for volcanic early warning.
2) The laws of change for this uncertainty seems exportable from one volcanic system to

another.

3) The form and timing of eruptions could be related to the apparent energy in each eruptive process.
4) Comparing figure 6, figure 9.B and figure 9.C, we can confirm that the temporal behavior of the uncertainty is opposite to when the system is trained and adjusted from one volcano to a new, different one.
5) In terms of eruption pre-warning time, uncertainty is ahead of other observables, rendering it potentially more effective for volcanic alert systems. For example, at Mt. Etna, the uncertainty started to increase 16.6 hours before the eruption, that is, 200 time units before the RMSE. The acceleration of uncertainty with such an hourly anticipation constitutes the pillars to create exportable and universal early-warning systems in poorly-monitored regions, and other volcanoes world-wide. Finally, we have shown that our technique was able to detect many hours in advance when the most energetic explosive process of this eruption was expected.

IX. Conclusion

We present a new neural network architecture rooted in a learnable scattering transform to perform temporal modeling and introduce a multi-modular recurrent architecture to implement polyphonic detection, segmentation and classification of seismic signals with the purpose of seismo-volcanic data exploration. Our architecture demonstrates that the flexibility introduced in learning layer-wise knots and filter-bank design, jointly with specialized recurrent dynamic convolutions, yields optimal, robust features or representations in a frame-wise fashion. The application of our architecture to data from three volcanoes, of different types and on different continents, shows that our approach generalizes well and properly adapts to different environments. The non-uniform data taxonomies in seismo-volcanic applications are collapsed into a generic but well-known categorization scheme to enable the computation of invariant, robust, and universal scatter-grams. Our neural network guarantees the rapid recognition of events, and is robust against sparse data taxonomies and the presence of background noise, which is an active topic of research in machine learning applications [77].

Nonetheless, the designed recurrent scattering networks allows for seamless integration with modern seismic data workflows, and includes an online streaming data approach that can provide direct warning system statements. This requires a straightforward adjustment of the architecture,

namely, changing the bidirectionality to an unidirectional frame-wise sequential recurrent network. Hence, our system is a universal framework that bridges the gap between deep learning and online monitoring; moreover, despite the myriad of complex physical mechanisms involved in volcanic unrest, our framework allows for simultaneous seismic events to be fully detected and characterized.

In addition to studying the complexities in mechanisms that drive volcano dynamics, the timing of eruptions remains an active research topic. Established forecasting algorithms are designed to find mathematical relationships involving accelerated strain, energy, or frequency variations; they assume handcrafted features that do not consider the class-membership of events and discard hidden information that might offer insight in the source mechanisms that generate seismic signals and drive volcanism. Our deep recurrent scattering network departs from the traditional perspective and opens new directions for designing forecasting methods based on the connections between epistemic uncertainty and volcano dynamics. The power-law drift in epistemic uncertainty associated with seismic data streams implies that volcanic processes preceding eruptions are detectable. With no prior assumptions about signal distribution, deep learning can identify such behavior without supervision or parametrization by data alone. The epistemic uncertainty generated by our deep neural network holds promise for forecasting eruptions, although challenges remain. As our approach can be modified to act on real-time streaming data, this goes in concert with the development of a novel early-warning strategy.

APPENDIX A
BAYESIAN NEURAL NETWORKS

Mathematical notation: First, we establish the mathematical notation that we will follow in the formulation of Bayesian modelling for neural networks. We define our dataset (assuming we are working with seismic data) as a set of N points, $D = (\mathbf{X}, \mathbf{Y})$ where $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ the matrix whose rows correspond to a set of seismic events (in samples) extracted from a continuous seismic record, and $\mathbf{Y} = \{y_1, y_2, \ldots, y_N\}$ is the matrix whose rows contain the corresponding labels that are assigned to every seismic event and that are categorized over a set of C classes. We refer as L to the total number of layers in the neural network, being i the sub-index of any of its layers, $i = \{1, \cdots, L\}$. Therefore, in this appendix, we specify our neural network as a function parameterized by its weights, $\mathbf{y} = f^{\omega}(\mathbf{x})$, where ω represents all weights matrices associated with the hidden layers of the neural network, that is, $\omega = \{W_i\}_{i=1}^{L}$, $\mathbf{x}$ the

input feature vector of the network (i.e. a row of matrix $\mathbf{X}$) and $\mathbf{y}$ the associated polyphonic vector labels (i.e. a row of matrix $\mathbf{Y}$). In this appendix, we refer as θ to the parameters that define the approximate distribution $q_\theta(\omega)$.

Bayesian modelling in neural networks: It is well known that Bayesian methods provide a measure of uncertainty for each input and output of a given model, based on all observed data. In most of the so-called *frequentist approaches*, commonly used in neural networks, the final optimization result is a set of best-fitting parameters. Unlike frequentist methods, the result of a Bayesian fit is a probability distribution of each parameter of the model, called the *posterior distribution*. For a given set of parameters in our neural network, and the dataset D defined, the posterior $p(\omega|\mathbf{X},\mathbf{Y})$ is determined by using the Bayes Theorem:

$$p(\omega|\mathbf{X},\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X},\omega)p(\omega)}{p(\mathbf{Y}|\mathbf{X})}, \tag{9}$$

with $p(\mathbf{Y}|\mathbf{X},\omega)$ is defined as the *model likelihood distribution*, that is, the knowledge of the model on the data distribution, and therefore, the assignment of probabilities for each $\mathbf{X}$ and $\mathbf{Y}$, given the parameters of the model. The term $p(\omega)$ is known as *prior* and constitutes the initial, known probability distribution of the parameters of the network. Hence, in a Bayesian neural network (BNN), the prior distribution is specified as a set of probability distributions located on their weights [78]. The denominator of (9) corresponds to the *model evidence* or *marginal likelihood*, a normalizing constant that can be obtained by marginalizing the likelihood over the parameters ω:

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{X},\omega)p(\omega)\,\mathrm{d}\omega. \tag{10}$$

Theoretically, the marginalization in (10) involves the average with respect to all possible parameters of the model ω, weighted by $p(\omega)$. For complex models, such as BNNs, an approximation is required [79]. Defining all the terms in the numerator and denominator of (9), a BNN can predict the outputs y^* for any new input x^* through the predictive function by integration over the parameters of the network ω:

$$p(y^*|x^*,\mathbf{X},\mathbf{Y}) = \mathbb{E}_{p(\omega|\mathbf{X},\mathbf{Y})} = \int p(y^*|x^*,\omega)p(\omega|\mathbf{X},\mathbf{Y})\,\mathrm{d}\omega, \tag{11}$$

where $p(y^*|x^*,\omega)$ is the data likelihood for this new point x^*. The prediction of new seismic events for multiple instances in y^* is known as *inference*. However, the exact inference in (11) is impossible given that the posterior is part of the integral. The computation of (11) with $p(\omega|\mathbf{X},\mathbf{Y})$ is equivalent to evaluate an infinite number of neural networks with all the possible

parameter configurations. This is computationally intractable for neural networks of any size. For this reason, in Bayesian modelling, an *approximate inference* procedure is required. This type of inference entails an optimization conditioned to the training of the architecture, that is, the approximation of this integral. Variational inference methods are used to approximate $p(\omega|\mathbf{X}, \mathbf{Y})$ and therefore the equation (11).

Variational inference in BNNs: Variational Inference (VI) focuses on obtaining an approximation to $p(\omega|\mathbf{X}, \mathbf{Y})$ by using optimization procedures [80]. Formally, this optimization aims at determining a probability density, $q_\theta(\omega)$, that should be as close as possible to $p(\omega|\mathbf{X}, \mathbf{Y})$. The measure of closeness is given by the Kullback-Leibler (KL) divergence between both distributions:

$$KL(q_\theta(\omega) \,||\, p(\omega|\mathbf{X}, \mathbf{Y})) = \int q_\theta(\omega) \, \log \left\{ \frac{q_\theta(\omega)}{p(\omega|\mathbf{X}, \mathbf{Y})} \right\} d\omega. \tag{12}$$

with $q_\theta(\omega)$ known as the *variational distribution*. We minimize the (12) by optimizing the variational parameters θ of our variational distribution $q_\theta(\omega)$:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \; \mathbb{E}_{q_\theta(\omega)}[log \; q_\theta(\omega) - log \; p(\omega|\mathbf{X}, \mathbf{Y})] \tag{13}$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \; KL(q_\theta(\omega) \,||\, p(\omega|\mathbf{X}, \mathbf{Y})) \tag{14}$$

with $\hat{\theta}$ the parameters that results in the minimum KL divergence. Once we obtained our variational approximation $q_\theta(\omega)$, and the KL in (12) has been minimized, the predictive distribution is given as:

$$p(y^*|x^*, D) \approx \int p(y^*|x^*, \omega) q_{\hat{\theta}}(\omega) \, \mathrm{d}\omega =: q_{\hat{\theta}}(y^*|x^*). \tag{15}$$

However, we can verify that the evaluation of the KL divergence in (12) requires the computation of the posterior distribution for our network's parameters, which are precisely the distribution that we want to approximate. To circumvent this, we can minimize a function similar to (12) added to a constant term. This function is known as Evidence Lower Bound (ELBO). The mathematical relationship between the ELBO and KL divergence, $KL(q_\theta(\omega) \,||\, p(\omega|\mathbf{X}, \mathbf{Y}))$ can be derived from (12) by invoking Bayes rule and taking logarithm, yielding [80]:

$$log \, p(\mathbf{Y}|\mathbf{X}) - \underbrace{\int q_\theta(\omega) \; log \left\{ \frac{p(\mathbf{Y}|\mathbf{X}, \omega) \, p(\omega)}{q_\theta(\omega)} \right\} d\omega}_{\text{ELBO or } \mathcal{L}_{ELBO}(\theta)} \tag{16}$$

Therefore it is observable from the above equations that the KL divergence is equal to the ELBO ($\mathcal{L}_{ELBO}(\theta)$) and a constant which is given by the marginal log-likelihood of our data. Since the KL divergence is a probabilistic distance and always positive, we can thus write:

$$log\, p(\mathbf{Y}|\mathbf{X}) \geq \mathcal{L}_{ELBO}(\theta) + KL(q_\theta(\omega) \,||\, p(\omega|\mathbf{X}, \mathbf{Y})). \tag{17}$$

with $\mathcal{L}_{ELBO}(\theta)$ becoming the objective of our optimization problem. In addition, minimizing the divergence of KL is also equivalent to maximizing ELBO with respect to the variational parameters of the distribution $q_\theta(\omega)$. We can expand the term $\mathcal{L}_{ELBO}(\theta)$, and obtain its closed-form expression:

$$\mathcal{L}_{ELBO}(\theta) := -\int q_\theta(\omega)\; log\, p(\mathbf{Y}|\mathbf{X}, \omega) d\omega \;+\; KL(q_\theta(\omega) \,||\, p(\omega)). \tag{18}$$

The optimization of the first integral term conditions the Bayesian model to better fit our data. The second KL term acts as a regularizer, keeping $q_\theta(\omega)$ from extreme deviations of $p(\omega)$. This analytical representation can be used to rewrite (12), $KL(q_\theta(\omega) \,||\, p(\omega|\mathbf{X}, \mathbf{Y}))$ in approximative terms for the parameters of our neural network:

$$-\sum_{n=1}^{N} \int q_\theta(\omega)\; \log\, p(y_n|f^\omega(x_n)) d\omega + KL(q_\theta(\omega) \,||\, p(\omega)) \tag{19}$$

where $f^\omega(x_n)$ is the output of the neural network for a given arbitrary input x_n, and the summatory term defined as the *expected log likelihood*. Once all the parameters for variational optimization in a BNN have been established, it is necessary to choose the prior and explicitly define the variational $q_\theta(\omega)$ distribution to optimize in (19). In a BNN, $q_\theta(\omega)$ is always conditioned to the distribution given by the matrices of its neural connections. The multiple non-linearities and the evaluation of the first integral in (19) with N events of a dataset entails a prohibitive, non-scalable computation. However, we can consider the Monte Carlo sampling estimators and their connections to regularization techniques in deep neural networks. Monte Carlo estimators permit an approximation of the expected log-likelihood for neural network models with multiple hidden layers and their derivatives with respect to the variational parameters θ. The so-called Monte-Carlo dropout (MC-dropout) is thus a variational estimation that connects dropout regularization and standard neural network optimization with the inference procedure in (19) [24].

Monte Carlo dropout: The dropout technique can be used in a BNN as a Bayesian approximation of the posterior distribution of the network parameters. Initially, the dropout is formulated by [81] as a stochastic regularization technique for deep neural networks, randomly deactivating the parameters of a neural network with a given probability, p_i. A Bernoulli distribution can

model this probability p_i, selecting which of the hidden units remain active in the network. The key result for this reasoning is derived by [24] and [82]: the integral and KL terms in (19) can be linked to standard dropout training in deep neural networks. This permits scalable and robust inference for large datasets in very complex networks.

Formally, our neural network is composed of a set of weight matrices in all its layers, $\omega = \{W_i\}_{i=1}^{L}$. Each weight matrix has a dimension $K_i \times K_{i-1}$. We define the variational distribution $q_\theta(\omega)$ as the factorization over the weight matrices of all the hidden layers conditioned to the dropout technique:

$$W_i = M_i \cdot \text{diag}([z_{i,j}])_{j=1}^{K_i}, \tag{20}$$

$$z_{i,j} \sim \text{Bernoulli}(p_l), \quad i = 1, \ldots, L, \; j = 1, \ldots, K_{l-1}, \tag{21}$$

where $z_{i,j}$ represents the dropout masks (matrices of zeros and ones drawn from the Bernoulli distribution) which disable the hidden element j on layer $i-1$. The term M_i is a *mean weight matrix*, whose set $\theta = \{M_i\}_{i=1}^{L}$ are the variational parameters. Finally, having defined the variational distribution, we can use Monte-Carlo estimation to approximate the integral of the expected log-likelihood in (19):

$$-\int q_\theta(\omega) \; \log \, p(\mathbf{Y}|\mathbf{X}, \omega) d\omega = \frac{1}{N} \sum_{n=1}^{N} - log \, p(y_n | f^{\hat{\omega}}(x_n)) \tag{22}$$

where $\hat{\omega}$ is not a maximum posterior estimate, but multiple realisations of random variables from the Bernoulli distribution, $\hat{\omega} \sim q_\theta(\omega)$. This reasoning is identical to applying successive dropout masks to the network weights. Hence, the averaged sum of $log \, p(y_n | f^{\hat{\omega}}(x_n))$ represents, by definition, the cost function of a neural network.

In order to link the variational inference optimization $\mathcal{L}_{ELBO}(\theta)$ to the optimization objective of standard neural networks with dropout, $\mathcal{L}_{dropout}(\theta)$, it is necessary that the mathematical relation known as KL-condition in the regularizer term is fulfilled. The KL-condition links the derivatives of the optimization objective in (12) with standard loss functions in neural networks. In this Appendix, we do not cover the entire proof in detail and refer the reader to the original work by [24], pages 150-152, Appendix A. The KL-condition establishes that the regularizer KL term in (12) can be approximated as a standard dropout regularizer weighted by a normalization constant λ. Our objective of variational minimization is defined as:

$$\mathcal{L}_{dropout}(\theta) = \frac{1}{N} \sum_{n=1}^{N} - log \, p(y_n | f^{\hat{\omega}}(x_n)) + \lambda \sum_{l=1}^{L} (\|M_l\|_2^2 + \|b_l\|_2^2) \tag{23}$$

Therefore, approximate inference procedures result in an optimization goal identical to that of a neural network using the loss function $\mathcal{L}_{dropout}(\theta)$. This function is defined to optimize the parameters of the neural network and find the best $q_{\hat{\theta}}$ that minimizes the KL divergence, $KL(q_\theta(\omega) \,||\, p(\omega|\mathbf{X}, \mathbf{Y}))$ in equation (19). Finally, we can use the approximation learned by our network to evaluate the predictive function in (15), using Monte-Carlo sampling with T sampling steps:

$$q_{\hat{\theta}}(y^*|x^*) = \int p(y^*|f^{\hat{\omega}}(x^*))q_{\hat{\theta}}(\omega)d\omega \approx \frac{1}{T}\sum_{t=1}^{T} p(y^*|f^{\hat{\omega}_t}(x^*)) \tag{24}$$

or equivalently $\hat{\omega}_t \sim q_\theta(\omega)$. Therefore, at the time of inference, the dropout layers are applied to the M_i matrices, generating a Monte-Carlo sample from the posterior distribution (see equation 11). In practice, the average of these samples can be interpreted as the prediction of the network, although a single estimate is not obtained, as many as T sampling steps are performed. We can use the probabilities obtained by MC-dropout to estimate the uncertainty in the application of seismo-volcanic recognition.

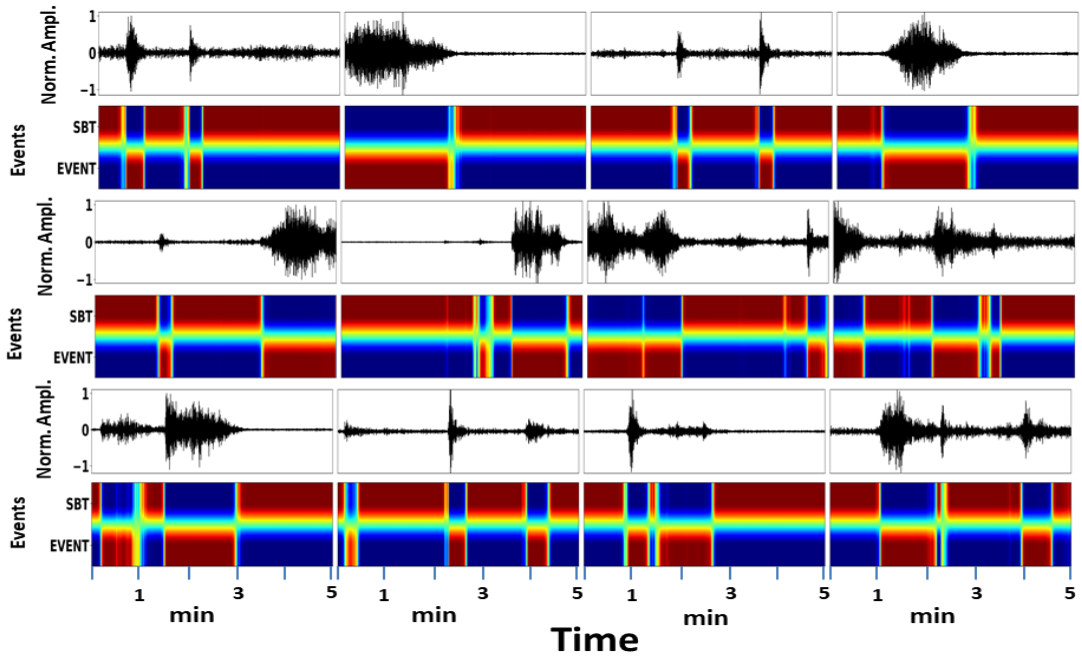

Fig. 10. **Pre-eruptive event detection heat maps**. Probabilistic heat maps for the first eruptive sequence of Bezymianny volcano. Data were obtain on 25 September 2007.

TABLE IV

SEISMIC EVENT DETECTION PERFORMANCE AT ONE SECOND FRAME-WISE PRECISION ON SEPTEMBER 25TH, BEZYMIANNY ERUPTION.

System	F1	PR	RC	ER
Vanilla	88.17	88.16	88.18	11.83
Filters	88.11	88.12	88.09	11.86
Knots + Filters	88.18	88.18	88.19	11.82

APPENDIX B
EVENT DETECTION

Seismo-volcanic event detection is a monophonic task. Our neural network is trained to detect and segment the presence or absence of events. We use the same hyper-parameter configuration of our network that in the multisource problem, except that we change the number of hidden units in the *sigmoid* class to two, event/no event (see section VII). Table IV contains the attained metrics for each learnable configuration on the test data set associated with the Bezymianny eruption of 25 September 2007. The results improve on those obtained at the first eruption using a multisource setting. Implicitly, the neural network becomes more sensitive to the onset of events and the duration. Remarkably, the ER is lower, which shows a better alignment of the prediction with the ground truth. It can thus be applied to seismology as an alternative to traditional algorithms [71], feature-based algorithms [83], or data-mining based pipelines [84]. Traditional STA/LTA methods require some prior knowledge about the signals to which they are applied, which is avoided by using a learning-based approach. Feature-based algorithms improve the segmentation boundaries in the time of events but require prior knowledge about the frequency range in which events are distinguished from noise. An algorithm based on fingerprint similarity, where fingerprints are generated through a network analogue from signals using data mining [84], relies on the assumption that all events are sufficiently frequently represented in the data set. Our neural network for event detection can be pre-trained and exportable to be applied ubiquitously to many volcanic systems, regardless of eruptive style. Results on test data from Bezymianny are illustrated in Figure 10. The pre-eruptive event detection heat maps show that our neural network can systematically detect and segment seismo-volcanic events even in the presence of significant background noise.

ACKNOWLEDGMENT

The facilities of IRIS Data Services, and specifically the IRIS Data Management Center, were used for access to the waveforms, related metadata, and/or derived products used in this study. IRIS Data Services were funded through a Seismological Facilities for the Advancement of Geoscience (SAGE) Award of the National Science Foundation under Cooperative Support Agreement EAR-1851048.

The Mt. Etna data were collected within the framework of the project: Volcanic Emissions Analysis through Seismic and Infrasound Advanced monitoring (VOSSIA), supported by the Trans-National Access component of the EUROVOLC project (European Network of Observatories and Research Infrastructures for Volcanology, EU Horizon 2020 Research Infrastructure Project grant No 731070).

This work is supported by TEC2015-68752 (KNOWAVES), PID2019-106260GB-I00 (FEMALE), Department of Energy grant DE-SC0020345, Simons Foundation MATH + X program and the Geo-Mathematical Imaging Group at Rice University.

REFERENCES

[1] Mie Ichihara. “Seismic and infrasonic eruption tremors and their relation to magma discharge rate: A case study for sub-Plinian events in the 2011 eruption of Shinmoe-dake, Japan”. In: *Journal of Geophysical Research: Solid Earth* 121.10 (2016), pp. 7101–7118.

[2] Azusa Mori and Hiroyuki Kumagai. “Estimating plume heights of explosive eruptions using high-frequency seismic amplitudes”. In: *Geophysical Journal International* 219.2 (2019), pp. 1365–1376.

[3] F. Brenguier, N. Shapiro, M. Campillo, V. Ferrazzini, Z. Duputel, O. Coutant, and A. Nercessian. “Towards forecasting volcanic eruptions using seismic noise”. In: *Nature Geoscience* 1.2 (2008), pp. 126–130.

[4] C. Gómez-García, F. Brenguier, P. Boué, N. Shapiro, DV. Droznin, SY. Droznina, SL. Senyukov, and EI. Gordeev. “Retrieving robust noise-based seismic velocity changes from sparse data sets: synthetic tests and application to Klyuchevskoy volcanic group (Kamchatka)”. In: *Geophysical Journal International* 214.2 (2018), pp. 1218–1236.

[5] T. Takano, F. Brenguier, M. Campillo, A. Peltier, and T. Nishimura. “Noise-based passive ballistic wave seismic monitoring on an active volcano”. In: *Geophysical Journal International* 220.1 (2020), pp. 501–507.

[6] Bernard Chouet. “Volcano seismology”. In: *Pure and applied geophysics* 160.3 (2003), pp. 739–788.

[7] RSJ. Sparks, J. Biggs, and JW. Neuberg. “Monitoring volcanoes”. In: *Science* 335.6074 (2012), pp. 1310–1311.

[8] S.R. McNutt, G. Thompson, J. Johnson, S. De Angelis, and D. Fee. “Seismic and infrasonic monitoring”. In: *The Encyclopedia of Volcanoes (Second Edition)*. Elsevier, 2015, pp. 1071–1099.

[9] J.M. Ibáñez, E. Pezzo, J. Almendros, M. La Rocca, G. Alguacil, R. Ortiz, and A. García. “Seismo-volcanic signals at Deception Island volcano, Antarctica: Wave field analysis and source modeling”. In: *Journal of Geophysical Research: Solid Earth* 105.B6 (2000), pp. 13905–13931.

[10] M Palo, JM Ibáñez, M Cisneros, M Bretón, E Del Pezzo, E Ocana, J Orozco-Rojas, and AM Posadas. "Analysis of the seismic wavefield properties of volcanic explosions at Volcan de Colima, Mexico: insights into the source mechanism". In: *Geophysical Journal International* 177.3 (2009), pp. 1383–1398.

[11] Vyacheslav M Zobin. *Introduction to volcanic seismology*. Vol. 6. Elsevier, 2012.

[12] Kostas I Konstantinou. "Tornillos modeled as self-oscillations of fluid filling a cavity: Application to the 1992–1993 activity at Galeras volcano, Colombia". In: *Physics of the Earth and Planetary Interiors* 238 (2015), pp. 23–33.

[13] R.S. Matoza, A. Arciniega-Ceballos, R.W. Sanderson, G. Mendo-Pérez, A. Rosado-Fuentes, and B.Chouet. "High-Broadband Seismo-acoustic Signature of Vulcanian Explosions at Popocatépetl Volcano, Mexico". In: *Geophysical Research Letters* 46.1 (2019), pp. 148–157.

[14] Gilberto Saccorotti and Ivan Lokmer. "A review of seismic methods for monitoring and understanding active volcanoes". In: *Forecasting and Planning for Volcanic Hazards, Risks, and Disasters* (2020), pp. 25–73.

[15] G. Cortés, R. Carniel, M. Ángeles Mendoza, and P. Lesage. "Standardization of Noisy Volcanoseismic Waveforms as a Key Step toward Station-Independent, Robust Automatic Recognition". In: *Seismological Research Letters* 90.2A (Jan. 2019), pp. 581–590. ISSN: 0895-0695. DOI: 10.1785/0220180334.

[16] M. Malfante, M. Dalla Mura, J. Metaxian, J. I. Mars, O. Macedo, and A. Inza. "Machine Learning for Volcano-Seismic Signals: Challenges and Perspectives". In: *IEEE Signal Processing Magazine* 35.2 (2018), pp. 20–30. ISSN: 1053-5888.

[17] M. Titos, A. Bueno, L. García, M. C. Benítez, and J. M. Ibáñez. "Detection and Classification of Continuous Volcano-Seismic Signals With Recurrent Neural Networks". In: *IEEE Transactions on Geoscience and Remote Sensing* (2018), pp. 1–13. ISSN: 0196-2892. DOI: 10.1109/TGRS.2018.2870202.

[18] M.S. Khan, M. Curilem, F. Huenupan, Muhammad M.F. Khan, and N.B. Yoma. "A Signal Processing Perspective of Monitoring Active Volcanoes [Applications Corner]". In: *IEEE Signal Processing Magazine* 36.6 (2019), pp. 125–163.

[19] N. Perez, P. Venegas, D. Benítez, R. Lara-Cueva, and M. Ruiz. "A new volcanic seismic signal descriptor and its application to a data set from the cotopaxi volcano". In: *IEEE Transactions on Geoscience and Remote Sensing* 58.9 (2020), pp. 6493–6503.

[20] J. Lahr, B.A. Chouet, C.D. Stephens, JA Power, and R. Page. "Earthquake classification, location, and error analysis in a volcanic environment: Implications for the magmatic system of the 1989–1990 eruptions at Redoubt Volcano, Alaska". In: *Journal of Volcanology and Geothermal Research* 62.1-4 (1994), pp. 137–151.

[21] B.A. Heath, E. Hooft, and D.R. Toomey. "Autocorrelation of the seismic wavefield at Newberry Volcano: Reflections from the magmatic and geothermal systems". In: *Geophysical Research Letters* 45.5 (2018), pp. 2311–2318.

[22] H. Nakamichi, M. Iguchi, H. Triastuty, M. Hendrasto, and I. Mulyana. "Differences of precursory seismic energy release for the 2007 effusive dome-forming and 2014 Plinian eruptions at Kelud volcano, Indonesia". In: *Journal of Volcanology and Geothermal Research* 382 (2019), pp. 68–80.

[23] A. Bevilacqua, M. Bursik, A. Patra, E. Pitman, and R. Till. "Bayesian construction of a long-term vent opening probability map in the Long Valley volcanic region (CA, USA)". In: *Statistics in Volcanology* 3 (Apr. 2017). DOI: 10.5038/2163-338X.3.1.

[24] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning". In: *international conference on machine learning*. 2016, pp. 1050–1059.

[25] S Mostafa Mousavi and Gregory C Beroza. "Bayesian-Deep-Learning Estimation of Earthquake Location From Single-Station Observations". In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).

[26] M. C. Benitez, J. Ramirez, C. Segura, J.M. Ibanez, J. Almendros, A. Garcia-Yeguas, and G. Cortes. "Continuous HMM-based seismic-event classification at Deception Island, Antarctica". In: *IEEE Transactions on Geoscience and remote sensing* 45.1 (2006), pp. 138–146.

[27] A. Bueno, C. Benítez, L. Zuccarello, S. de Angelis, and J. M. Ibáñez. "Bayesian Monitoring of Seismo-volcanic Dynamics". In: *IEEE Transactions on Geoscience and Remote Sensing* (2021). DOI: 10.1109/TGRS.2021.3076012.

[28] Joan Bruna and Stéphane Mallat. "Invariant scattering convolution networks". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1872–1886.

[29] Y. Zhang, W. Chan, and N. Jaitly. "Very deep convolutional networks for end-to-end speech recognition". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 4845–4849.

[30] S. Xingjian, Z. Chen, H. Wang, DT. Yeung, W-K Wong, and W c Woo. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting". In: *Advances in neural information processing systems*. 2015, pp. 802–810.

[31] Alex Graves and Jürgen Schmidhuber. "Framewise phoneme classification with bidirectional LSTM networks". In: *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*. Vol. 4. IEEE. 2005, pp. 2047–2052.

[32] Ángel Bueno Rodríguez. "Bayesian transfer learning for continuous monitoring of active volcanoes". In: (2021). URL: http://hdl.handle.net/10481/70452.

[33] K.J. Bergen, T. Chen, and Z. Li. "Preface to the Focus Section on Machine Learning in Seismology". In: *Seismological Research Letters* 90.2A (2019), pp. 477–480.

[34] K.J. Bergen, P.A. Johnson, M.V. de Hoop, and G. Beroza. "Machine learning for data-driven discovery in solid Earth geoscience". In: *Science* 363.6433 (2019). ISSN: 0036-8075. DOI: 10.1126/science.aau0323. URL: https://science.sciencemag.org/content/363/6433/eaau0323.

[35] D.C. Bolton, P. Shokouhia, B.Rouet-Leduc, C. Hulbert, J. Rivière, C. Marone, and P.A. Johnson. "Characterizing acoustic signals and searching for precursors during the laboratory seismic cycle using unsupervised machine learning". In: *Seismological Research Letters* 90.3 (2019), pp. 1088–1098.

[36] Y. Wu, Y. Lin, Z. Zhou, D.C. Bolton, J.Liu, and P.A Johnson. "DeepDetect: A cascaded region-based densely connected network for seismic event detection". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.1 (2018), pp. 62–75.

[37] G. Cortés, L. García, I. Álvarez, M.C. Benítez, Á. de la Torre, and J.M. Ibáñez. "Parallel System Architecture (PSA): An efficient approach for automatic recognition of volcano-seismic events". In: *Journal of Volcanology and Geothermal Research* 271 (2014), pp. 1 –10. ISSN: 0377-0273. DOI: https://doi.org/10.1016/j.jvolgeores.2013.07.004.

[38] C.X. Ren, A. Peltier, V. Ferrazzini, B. Rouet-Leduc, P.A. Johnson, and F. Brenguier. "Machine learning reveals the seismic signature of eruptive behavior at piton de la fournaise volcano". In: *Geophysical Research Letters* 47.3 (2020), e2019GL085523.

[39] G.F. Manley, D.M. Pyle, T.A. Mather, M. Rodgers, D.A. Clifton, B.G. Stokell, G. Thompson, J.M. Londoño, and D.C. Roman. "Understanding the timing of eruption end using a machine learning approach to classification of seismic time series". In: *Journal of Volcanology and Geothermal Research* (2020), p. 106917.

[40] T. Sainath, R.J. Weiss, K.W. Wilson, A. Narayanan, and M. Bacchiani. "Factored spatial and spectral multichannel raw waveform CLDNNs". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 5075–5079.

[41] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve, and E. Dupoux. "Learning filterbanks from raw speech for phone recognition". In: *ICASSP*. IEEE. 2018, pp. 5509–5513.

[42] Mirco Ravanelli and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2018, pp. 1021–1028.

[43] Florent Jaillet and Bruno Torrésani. "Time-frequency jigsaw puzzle: Adaptive multiwindow and multilayered Gabor expansions". In: *Int. J. of Wavelets, Multiresolution and Inf. Proc.* 5.02 (2007), pp. 293–315.

[44] E. Cakir, E. C. Ozan, and T. Virtanen. “Filterbank learning for deep neural network based polyphonic sound event detection”. In: *2016 International Joint Conference on Neural Networks (IJCNN)* (2016), pp. 3399–3406.

[45] Haidar Khan and Bulent Yener. “Learning filter widths of spectral decompositions with wavelets”. In: *Advances in Neural Inf. Proc. Sys. 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. 2018, pp. 4601–4612.

[46] Mirco Ravanelli and Yoshua Bengio. “Interpretable convolutional filters with sincnet”. In: *arXiv preprint arXiv:1811.09725* (2018).

[47] Soo-Chang Pei and Shih-Gu Huang. “STFT with adaptive window width based on the chirp rate”. In: *IEEE Transactions on Signal Processing* 60.8 (2012), pp. 4065–4080.

[48] Randall Balestriero, Herve Glotin, and Richard Baraniuk. “Interpretable and Learnable Super-Resolution Time-Frequency Representation”. In: *Proceedings of Machine Learning Research*. Ed. by Joan Bruna, Jan Hesthaven, and Lenka Zdeborova. Vol. 107. Conference on Mathematical and Scientific Machine Learning. JMLR, 2021, pp. 1–25.

[49] W Ye, J Cheng, F Yang, and Y Xu. “Two-Stream Convolutional Network for Improving Activity Recognition Using Convolutional Long Short-Term Memory Networks”. In: *IEEE Access* 7 (2019), pp. 67772–67780.

[50] A. Chattopadhyay, P. Hassanzadeh, and S. Pasha. “Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data”. In: *Scientific Reports* 10.1 (2020), pp. 1–13.

[51] Z. Liang, Z. Guangming, M. Lin, S. Peiyi, S. Syed, and B. Mohammed. “Attention in Convolutional LSTM for Gesture Recognition”. In: *NIPS* (2018).

[52] Joakim Andén and Stéphane Mallat. “Deep scattering spectrum”. In: *IEEE Transactions on Signal Processing* 62.16 (2014), pp. 4114–4128.

[53] Stéphane Mallat. “Understanding deep convolutional networks”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150203.

[54] Steve Mann and Simon Haykin. “The chirplet transform: Physical considerations”. In: *IEEE Transactions on Signal Processing* 43.11 (1995), pp. 2745–2761.

[55] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. “On the difficulty of training recurrent neural networks”. In: *International conference on machine learning*. 2013, pp. 1310–1318.

[56] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. “Speech recognition with deep recurrent neural networks”. In: *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE. 2013, pp. 6645–6649.

[57] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. “Learning under concept drift: A review”. In: *IEEE Transactions on Knowledge and Data Engineering* 31.12 (2018), pp. 2346–2363.

[58] A. Bueno, C. Benítez, S. de Angelis, A. Díaz-Moreno, and J. M. Ibáñez. “Volcano-Seismic Transfer Learning and Uncertainty Quantification with Bayesian Neural Networks”. In: *IEEE Transactions on Geoscience and Remote Sensing* (2019), pp. 1–13. DOI: 10.1109/TGRS.2019.2941494.

[59] Armen Der Kiureghian and Ove Ditlevsen. “Aleatory or epistemic? Does it matter?” In: *Structural safety* 31.2 (2009), pp. 105–112.

[60] Jiaming Zeng, Adam Lesnikowski, and Jose M. Alvarez. “The Relevance of Bayesian Layer Positioning to Model Uncertainty in Deep Bayesian Active Learning”. In: (2018). arXiv: 1811.12535 [cs.LG].

[61] N. Brosse, C. Riquelme, A. Martin, S. Gelly, and É. Moulines. “On Last-Layer Algorithms for Classification: Decoupling Representation from Uncertainty Estimation”. In: (2020). arXiv: 2001.08049 [stat.ML].

[62] W. Thelen, M. West, and S. Senyukov. “Seismic characterization of the fall 2007 eruptive sequence at Bezymianny Volcano, Russia”. In: *Journal of Volcanology and Geothermal Research* 194.4 (2010), pp. 201 –213. ISSN: 0377-0273. DOI: https://doi.org/10.1016/j.jvolgeores.2010.05.010.

[63] J. Weston, F. Ratle, H. Mobahi, and R. Collobert. “Deep learning via semi-supervised embedding”. In: *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 639–655.

[64] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. “Scaper: A library for soundscape synthesis and augmentation”. In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2017, pp. 344–348.

[65] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. “TUT Database for Acoustic Scene Classification and Sound Event Detection”. In: *24th European Signal Processing Conference 2016 (EUSIPCO 2016)*. Budapest, Hungary, 2016.

[66] Michael E West. “Recent eruptions at Bezymianny volcano—A seismological comparison”. In: *Journal of Volcanology and Geothermal Research* 263 (2013), pp. 42–57.

[67] Olga A Girina. “Chronology of Bezymianny volcano activity, 1956–2010”. In: *Journal of Volcanology and Geothermal Research* 263 (2013), pp. 22–41.

[68] B. Voight. “Voight, B. A method for prediction of volcanic eruptions. Nature 332, 125-130”. In: *Nature* 332 (Feb. 1988), pp. 125–130. DOI: 10.1038/332125a0.

[69] A.F. Bell, J. Greenhough, M.J. Heap, and I.G. Main. “Challenges for forecasting based on accelerating rates of earthquakes at volcanoes and laboratory analogues”. In: *Geophysical Journal International* 185.2 (May 2011), pp. 718–723. ISSN: 0956-540X. DOI: 10.1111/j.1365-246X.2011.04982.x.

[70] A. Boué, P. Lesage, Guillermo G. Cortés, B. Valette, and G. Reyes-Dávila. “Real-time eruption forecasting using the material Failure Forecast Method with a Bayesian approach”. In: *Journal of Geophysical Research: Solid Earth* 120.4 (2015), pp. 2143–2161.

[71] Amadej Trnkoczy. “Topic Understanding and parameter setting of STA/LTA trigger algorithm”. In: *New Manual of Seismological Observatory Practice* 2 (1999).

[72] B. Behncke, S. Branca, R. Corsaro, E. De Beni, L. Miraglia, and C. Proietti. “The 2011–2012 summit activity of Mount Etna: Birth, growth and products of the new SE crater”. In: *Journal of Volcanology and Geothermal Research* 270 (2014), pp. 10–21.

[73] A. Cappello, G. Ganci, G. Bilotta, C. Corradino, A. Hérault, and C. Del Negro. “Changing eruptive styles at the south-east crater of Mount Etna: Implications for assessing lava flow hazards”. In: *Frontiers in Earth Science* 7 (2019), p. 213.

[74] S. Scudero, G. De Guidi, and A. Gudmundsson. “Size distributions of fractures, dykes, and eruptions on Etna, Italy: Implications for magma-chamber volume and eruption potential”. In: *Scientific reports* 9.1 (2019), pp. 1–9.

[75] E. Giampiccolo, O. Cocina, P. De Gori, and C. Chiarabba. “Dyke intrusion and stress-induced collapse of volcano flanks: The example of the 2018 event at Mt. Etna (Sicily, Italy)”. In: *Scientific reports* 10.1 (2020), pp. 1–8.

[76] S. De Angelis, M.M. Haney, J.J. Lyons, A. Wech, D. Fee, A. Diaz-Moreno, and L. Zuccarello. “Uncertainty in Detection of Volcanic Activity Using Infrasound Arrays: Examples From Mt. Etna, Italy”. In: *Frontiers in Earth Science* 8 (2020), p. 169. ISSN: 2296-6463. DOI: 10.3389/feart.2020.00169.

[77] L. Jiang, D. Huang, M. Liu, and W. Yang. “Beyond synthetic noise: Deep learning on controlled noisy labels”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4804–4815.

[78] David JC MacKay. “Bayesian methods for backpropagation networks”. In: *Models of neural networks III*. Springer, 1996, pp. 211–254.

[79] Geoffrey E Hinton and Drew Van Camp. “Keeping the neural networks simple by minimizing the description length of the weights”. In: *Proceedings of the sixth annual conference on Computational learning theory*. 1993, pp. 5–13.

[80] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. “Variational inference: A review for statisticians”. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[81] N. Srivastava, G.E. Hinton, A. Krizhevsky I. Sutskever, and R. Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting.” In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[82] Yarin Gal. “Uncertainty in deep learning”. In: (2016).

[83] A. Bueno, A. Díaz-Moreno, S. De-Angelis, C. Benítez, and J. M. Ibáñez. “Recursive Entropy Method of Segmentation”. In: *Seism. Res. Lett.* 90.4 (2019), pp. 1670–1677.

[84] C.E. Yoon, O. O’Reilly, K.J. Bergen, and G.C. Beroza. “Earthquake detection through computationally efficient similarity search”. In: *Science advances* 1.11 (2015), e1501057.