Geophysical Journal International

Geophys. J. Int. (2021) **225,** 416–444
Advance Access publication 2020 December 22
GJI Seismology

doi: 10.1093/gji/ggaa593

Machine-learning-based detection of volcano seismicity using the spatial pattern of amplitudes

Yuta Maeda, Yoshiko Yamanaka, Takeo Ito and Shinichiro Horikawa

Graduate School of Environmental Studies, Nagoya University, Nagoya 464-8601*, Japan. E-mail:* maeda@seis.nagoya-u.ac.jp

Accepted 2020 December 10. Received 2020 November 24; in original form 2020 June 26

SUMMARY

We propose a new algorithm, focusing on spatial amplitude patterns, to automatically detect volcano seismic events from continuous waveforms. Candidate seismic events are detected based on signal-to-noise ratios. The algorithm then utilizes supervised machine learning to classify the existing candidate events into true and false categories. The input learning data are the ratios of the number of time samples with amplitudes greater than the background noise level at 1 s intervals (large amplitude ratios) given at every station site, and a manual classification table in which 'true' or 'false' flags are assigned to candidate events. A two-step approach is implemented in our procedure. First, using the large amplitude ratios at all stations, a neural network model representing a continuous spatial distribution of large amplitude probabilities is investigated at 1 s intervals. Second, several features are extracted from these spatial distributions, and a relation between the features and classification to true and false events is learned by a support vector machine. This two-step approach is essential to account for temporal loss of data, or station installation, movement, or removal. We evaluated the algorithm using data from Mt. Ontake, Japan, during the first ten days of a dense observation trial in the summit region (2017 November 1–10). Results showed a classification accuracy of more than 97 per cent.

Key words: Neural networks, fuzzy logic; Volcano monitoring; Volcano seismology.

1 INTRODUCTION

Seismic records are fundamental data to monitor and understand volcanic activity. Several types of analyses can directly use continuous seismic records, including the analysis of long-term temporal variations in seismic amplitudes (e.g. Endo & Murray 1991) and seismic interferometry based on an ambient noise (e.g. Brenguier *et al.* 2008). However, most studies in volcano seismology focus on earthquakes and tremors, which usually occupy only a small fraction of the waveform records. Typical analyses of these types include locating the sources, measuring the magnitudes, investigating the source mechanisms, evaluating the peak frequencies, investigating the statistical characteristics such as *b*-values, or just counting daily numbers of these events. Clearly, none of these analyses can be initiated without identifying the earthquakes and tremors from continuous records. Event detection is therefore the starting point for most studies in volcano seismology.

It is relatively easy to find signals that are potentially earthquakes (hereafter called candidate events), owing to the well-established approach based on the ratio of a short-term signal average (STA) to a long-term signal average (LTA, Allen 1982). However, signals detected by this procedure are not limited to seismicity in a target region but also consist of regional and distant earthquakes from outside the study area and transient noise caused by, for example, the wind or human activity. Therefore, each candidate event detected by the STA/LTA method needs to be evaluated, often manually, to discern whether it is a true event. This manual evaluation is a difficult and time-consuming task at volcanoes where small but volcanologically important signals need to be detected from relatively noisy waveforms. The purpose of this study is to automate this evaluation, which allows the automatic detection of volcano seismic signals.

Some volcano seismic signals are difficult to distinguish from distant earthquakes or local noise based on a single waveform trace. The spatial pattern of amplitudes may be a useful metric to distinguish them. For small volcanic events, signals above the noise level are recorded only at nearby stations. This pattern is different from that of distant earthquakes, for which distinct signals are distributed over the entire network, and from that of local noise, which show a significant amplitude decay over an extremely short distance. This difference is one of the most useful features in manual evaluation for whether a candidate event is a true volcanic signal. However, spatial amplitude patterns have not been used in existing automatic earthquake detection algorithms. The difficulty in utilizing this information arises from the absence of an exact theoretical relation between a source location and an expected amplitude pattern. Without the theoretical relation, traditional optimization frameworks such as least-squares or maximum-likelihood approaches cannot be constructed. This difficulty can be overcome by

machine learning, which is a kind of optimization that does not require a theoretical relation between the data and model parameters. Instead, it uses a generalized mathematical model applicable to widespread fields. Because of this nature, machine learning can enable quantitative analysis of spatial amplitude patterns without requiring a theoretical relationship.

Machine-learning techniques have been extensively used to automatically detect or classify volcano seismic (e.g. Langer *et al.* 2003, 2006; Scarpetta *et al.* 2005; Curilem *et al.* 2009; Hibert *et al.* 2017; Killion *et al.* 2018; Malfante *et al.* 2018) or various other seismic signals (e.g. Li *et al.* 2018; Perol *et al.* 2018; Nakano *et al.* 2019). Most of these studies have used a single waveform trace at a specific station site, or multiple traces as independent data. It is expected that a geographical waveform feature mapping from the entire network would significantly ease classification. The simultaneous use of multiple stations would also stabilize the system against loss of data at a certain station. Several recent studies (e.g. Maggi *et al.* 2017; Reynen & Audet 2017; Bergen & Beroza 2018; Qu *et al.* 2019) have proposed multistation approaches to detect seismicity. The method by Reynen & Audet (2017) assumed distinct *P*- and *S*-phases, and that by Qu *et al.* (2019) assumed a regularly deployed dense array of sensors along a 1-D line. However, neither the distinct *S*-phase nor the regular dense array requirements are not necessarily met at volcanoes. The approach by Bergen & Beroza (2018), who detected earthquakes based on feature similarities from time window pairs (Yoon *et al.* 2015), may be applicable to volcanoes, but seems to require similar signals to be repeated for detection success. Maggi *et al.* (2017) proposed a multistation random forest algorithm to classify volcano seismic events. They prepare many features from many stations and then automatically select important features and stations. In their study, station locations are not directly used; rather, they use detailed characteristics of the waveforms and spectra at individual sites.

In this study, we propose a multistation machine-learning algorithm to automatically detect seismic events. A key improvement of the algorithm is a geographical mapping of features, which allows any station to be removed or moved to different places without requiring a re-learning work. The main targets of this study are shallow small magnitude volcano seismic events occurring near an active crater, especially long-period (LP) events and tremors, which are difficult to detect by traditional seismological approaches due to unclear onsets of *P*- and *S*-phases. Listing, counting, monitoring and analysing these events are important as they are closely linked to movement and volume changes of shallow volcanic fluids (e.g. Chouet & Matoza 2013). We initially tried to detect only the LP events and tremors, but did not obtain desirable results after several trials. We then changed our strategy to first detect all seismic events in a relatively wide region, including LP events and tremors, volcano-tectonic (VT) events, and local non-volcanic earthquakes, while excluding distant earthquakes and local noise. After detecting the events, we need to classify them to extract LP and tremor activity. The present study reports on the detection; the classification will be attempted in a future study.

2 DATA

2.1 Seismic networks

We used continuous records from seismic stations within a 40 km × 40 km region centred on the summit of Mt. Ontake, central Japan (Fig. 1). These stations were maintained by Nagoya University, the Japan Meteorological Agency, the National Research Institute for Earth Science and Disaster Resilience (NIED), and the Nagano and Gifu prefectures. Most of the seismic stations were located more than 2 km from the summit until 2017 November, when a dense observation trial in the summit region started (Horikawa *et al.* 2017; Yamanaka *et al.* 2018). In this study, we used the data from the first ten days of the summit observation [2017 November 1–10; dates and times are based on Japan Standard Time (JST) throughout this work]. Over this time, seismic records from nine broad-band stations with natural periods of 120–360 s and 31 short-period stations with a natural period of 1 s were available. Horizontal locations are duplicated for one pair of surface and borehole stations, for which we used the borehole sensor, resulting in 39 stations available for the analysis. All of the data were sampled at 100 Hz. Throughout the study, we used only the vertical component.

2.2 Earthquake catalogues

We used two independent earthquake catalogues, both created by automatic triggering based on STA/LTA ratios and hypocentre determinations based on manual picks of *P*- and *S*-phase arrival times. One was a routine catalogue created by two expert staff members in Nagoya University which covered a wide region around Mt. Ontake. It consisted of 182 VT and local tectonic earthquakes during the study period (2017 November 1–10) within a 60 km × 60 km region centred on the summit of Mt. Ontake (Fig. 2a). In this catalogue, stations installed by the summit observation trial were not used. The other one was a summit catalogue created using all the stations including the summit observation trial. This catalogue focused on the summit region seismicity and consisted of 26 VT events located within 4 km to the summit during the study period (Fig. 2b). Five VT events appeared in both catalogues.

2.3 Long-period and tremor activities in the continuous records

All of the events in the routine and summit catalogues were VT and local tectonic earthquakes. LP events and tremors are generally difficult to detect based on the phase-picking approach because of emergent onsets and unclear *S*-phases. Before developing an automated procedure, we manually evaluated the LP and tremor activities at Mt. Ontake. These activities were considered to be quite rare at this volcano before

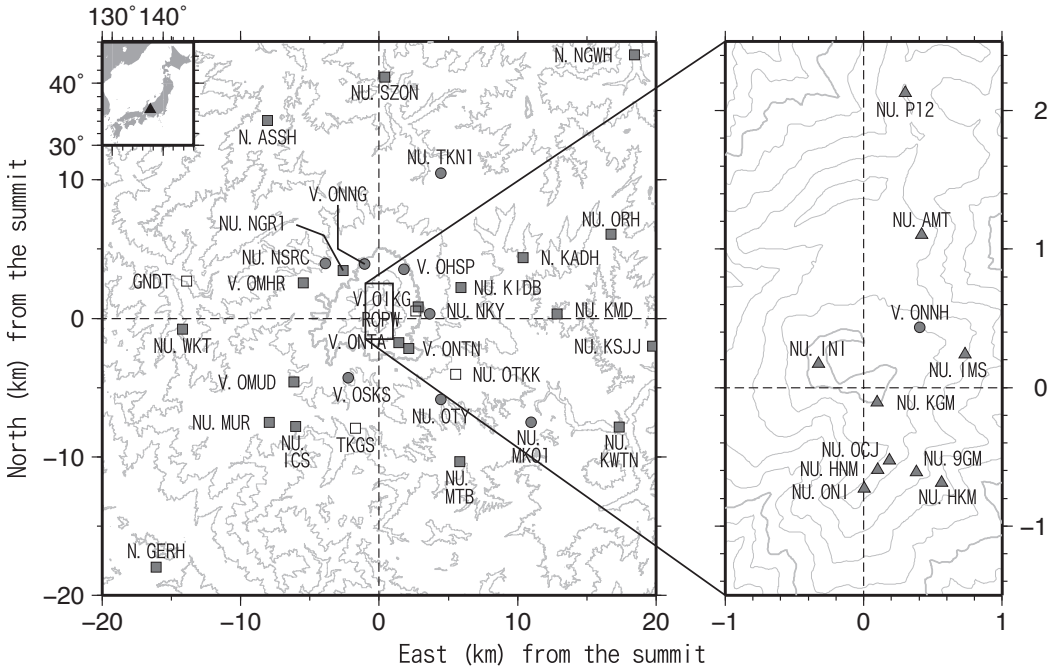

Figure 1. The seismometer network around Mt. Ontake. Squares and circles represent short-period (1 s) and broad-band stations, respectively. Triangles represent short-period (1 s) summit trial stations not used in the routine catalogue of earthquakes. Open symbols represent station defects in the main study period (2017 November 1–10); we used these stations in either of two subsequent analysis periods (2014 September 26–29 or 2017 June 24–27). Stations starting with 'NU.', 'V.' and 'N.' are operated by Nagoya University, the Japan Meteorological Agency and the NIED, respectively; the other stations are operated by the Nagano and Gifu prefectures. Contours represent the topography. Inset indicates the location of Mt. Ontake at a regional scale.

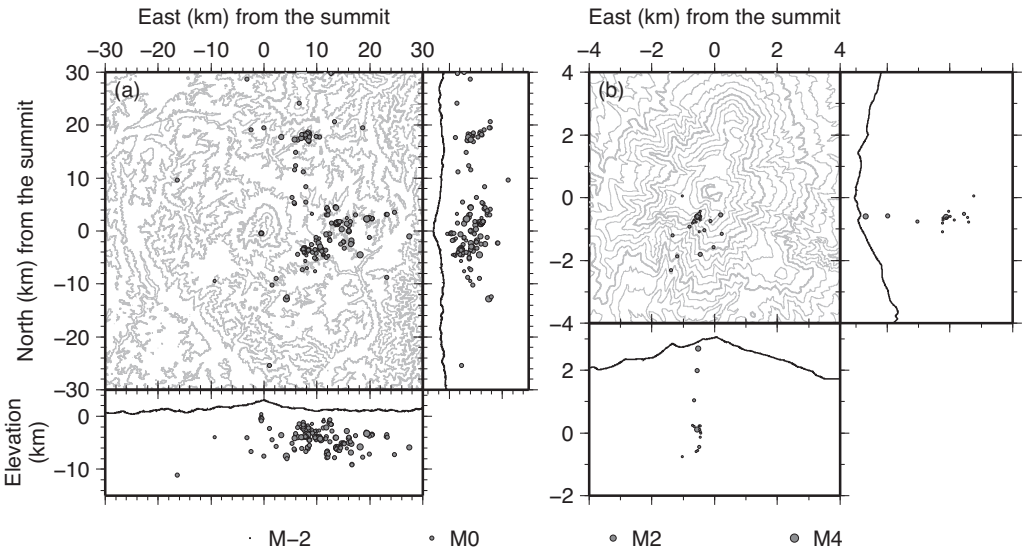

Figure 2. (a) Hypocentres of earthquakes during the study period in the routine and (b) summit catalogues. Some hypocentres in (b) were located above the ground (out of the vertical axis range) due to a poorly constrained velocity structure in the summit region used for hypocentre determination.

the beginning of the summit observation trial except for a period before the 2007 eruption (Nakamichi *et al.* 2009). We carefully looked at the continuous waveforms during the first seven days of the summit observation (2017 November 1–7) and identified five LP events and 11 tremors (arrows in Fig. 3). We noted several additional potential LP events and tremors, although we were not confident whether they were true volcanic signals. These results indicated that LP events and tremors occur at Mt. Ontake, although the occurrence rates were unclear.

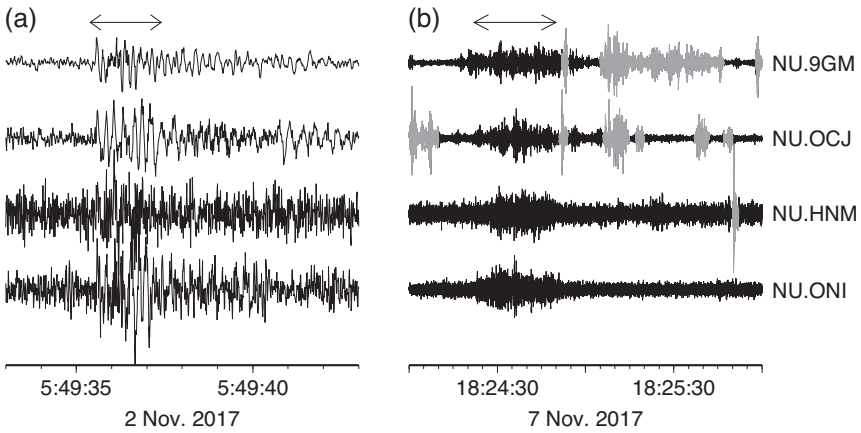

Figure 3. (a) Waveforms of an LP event and (b) a tremor (shown by arrows) at Mt. Ontake during the study period identified manually in advance of the study. Vertical waveforms high-pass filtered at 1 Hz are shown. A common amplitude scale is used for the four traces in each panel. Grey portions are considered to be local noise.

3 METHOD

3.1 Outline of the proposed method

The aim of the proposed method is to classify candidate events, detected in advance based on signal-to-noise ratios, into true and false events based on the spatial amplitude patterns with a supervised machine-learning technique. Here, the true events point to shallow volcanic or local tectonic earthquakes and tremors that occurred within an analysis region; the false ones consist of all the remaining signals, including regional and distant earthquakes from outside the analysis region and temporal increases in noise. We note that the summit records are sometimes contaminated by frequent spike-like increases in amplitude (grey portions in Fig. 3b). The occurrence times of these spike-like signals are inconsistent among the stations regardless of their proximity (∼100 m), suggesting that these are not true volcanic signals but local noise. Similar local noise events are detected as candidate events according to signal-to-noise ratios. The main purpose of the machine learning is to discriminate these from true volcano seismic events.

The analysis flow is illustrated in Fig. 4. Like most previous machine-learning studies, we first extracted features from continuous waveforms. We use only one feature, the ratio of the number of time samples with amplitudes greater than the background noise level (large amplitude ratios), for each 1 s interval of each station (Sections 3.2.2 and 3.2.3; Fig. 4). This ratio is a stable measure of the spatial amplitude pattern as the amplitudes were compressed to a [0, 1] range by keeping only the information on whether the amplitudes were greater than the background noise level, thereby reducing the effects of local structural amplifications and abnormally large electronic noise. The small number of features is essential to supervised machine learning with a relatively small teaching data set. In earthquake detection problems, the teaching data are lists of events that must be created manually and thus, cannot be extensive. Also, both true and false events have large varieties in waveform and spectral features, making it difficult to take into account these features for the classification.

Using the large amplitude ratios, we automatically detect candidate events (Section 3.2.4; Fig. 4) for which we further assign 'true' or 'false' flags based on manual waveform evaluations (Section 3.4.2; Fig. 4). Using these teaching data, we apply a two-step machine-learning approach (Sections 3.3 and 3.4.3; Fig. 4). In the first step, the spatial distribution of large amplitude ratios given at all the stations are used to investigate the continuous spatial distribution of the probability that each location on the ground had an amplitude greater than the background noise level (a large amplitude probability) over 1 s intervals (Section 3.3). A neural network model is used to express the continuous distribution. Several features are then defined to characterize the spatial distribution of large amplitude probabilities (Section 3.4.1; Fig. 4). In the second step, a relationship between these features and a classification of true or false events is learned (Section 3.4.3; Fig. 4) referring to the manual classifications prepared above. This two-step approach is essential to account for temporal loss of data, or station installation, movement, or removal.

Details of individual steps are described below, where the notation used is listed in Table 1. To simplify the description in the main text, we show only the parameters that were finally adopted. Details of the parameter choices are given in Appendix A.

3.2 Preparation of input data for machine learning

3.2.1 Filtering

From each 1-hr record, we removed the median amplitude and corrected the instrumental response to that of a short-period (1 s) seismometer with a damping constant of 0.7. We then applied a two-pole zero-phase 1 Hz Butterworth high-pass filter to suppress micro seismic noise around 0.5 Hz, which was strong during the study period. Our manual check of the continuous records suggested that volcano seismic events

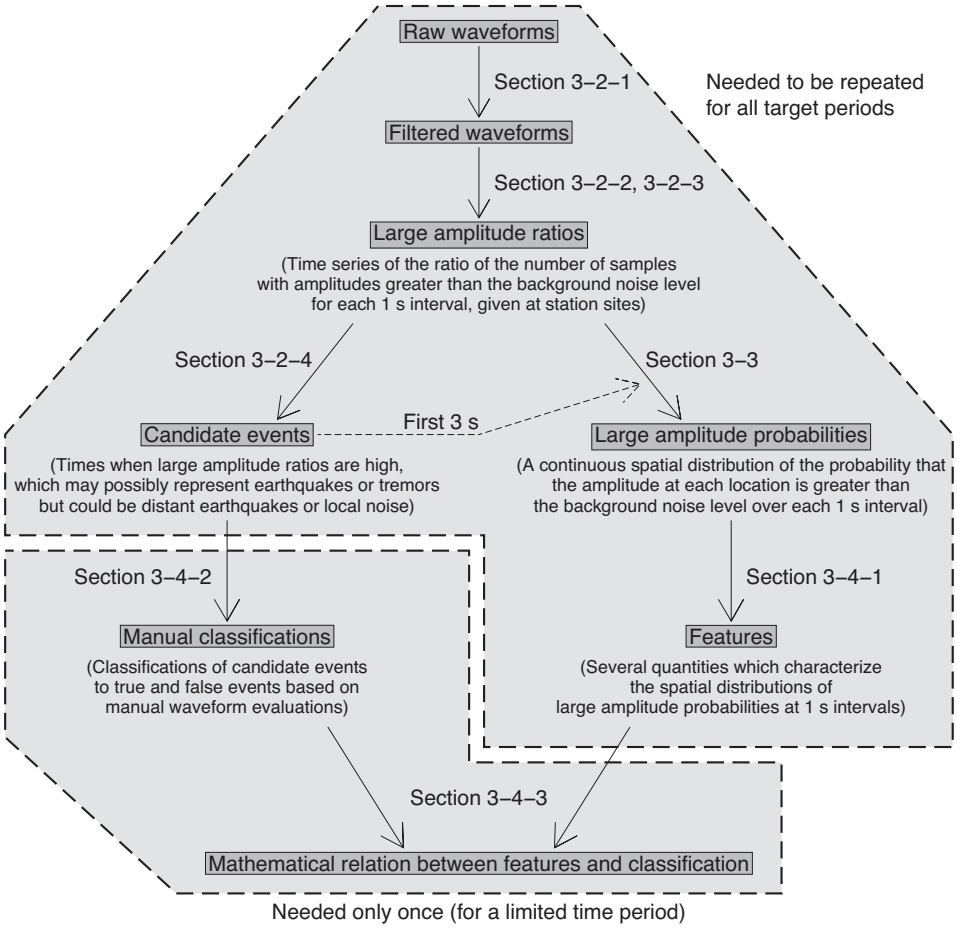

Figure 4. A schematic analysis flow of the analysis method.

with a peak frequency less than 1 Hz were quite rare at Mt. Ontake. Also, very long period (VLP; 0.01–0.5 Hz) events are sometimes associated with high frequency (> 1 Hz) oscillations (e.g. fig. 2a in Chouet *et al.* 2003; fig. 3a in Nakamichi *et al.* 2009; fig. 4a in Lyons & Waite 2011; fig. 2c in Maeda *et al.* 2019), suggesting that the high-pass filter does not exclude detections of these events. We note that the uses of median and short-period responses were essential to stabilize the results; a standard operation of removing an average and deconvolving an instrumental response resulted in artificial signals associated with static offsets, which were not suppressed enough by high-pass filtering for several traces at the time of a large (M5.6) local earthquake (Supporting Information).

3.2.2 Evaluating the background noise level

To define the large amplitude ratios, a background noise level needs to be set. A frequently used measure is an LTA multiplied by a constant value (around 3). This works if the LTA is not seriously affected by signals. Indeed, the summit data at Mt. Ontake are sometimes severely contaminated by large or frequent spike-like local signals (Fig. 3b). In this case, the LTA no longer represents the background noise level. Therefore, in this study, we use the cumulative frequency distribution of absolute amplitudes to investigate the background noise level. Our proposal is to evaluate the level by a threshold amplitude, below which the cumulative data distribution matches with that of Gaussian noise, and above which the two distributions deviate.

We quantify this idea as follows. Let $\bar{T}$ be the length of a time window to evaluate the background noise level (for which we used 5 min throughout; Appendix A.2), and N be the number of data samples in the window. We assume that the smallest $N'(\leq N)$ samples of the data in $\bar{T}$ obey a Gaussian distribution with zero average:

$$G\left(v; N'\right) = \exp\left[-v^2/2\sigma\left(N'\right)^2\right]/\sqrt{2\pi}\sigma\left(N'\right), \tag{1}$$

where $\sigma(N')$ is a standard deviation of the distribution given by

$$\sigma\left(N'\right) = \sqrt{\sum_{n=1}^{N'} v_n^2/N'}, \tag{2}$$

Table 1. Symbols used in this study in alphabetic order. SVM: support vector machine.

Symbol	Description	Defined in:
A	An area of the analysis region $(x^{\max} - x^{\min})(y^{\max} - y^{\min})$	Section 3.4.1
$\tilde{b}_i$	A binary time-series for a station i (a function of time t)	Section 3.2.3
b_{ik}	A large amplitude ratio for a station i and a time window k	Eq. (14)
b^{th}	A threshold value of b_{ik} for a candidate event detection	Appendix A.3
c_k	A circularity of R_k, defined as $c_k = S_k/(\pi r_k^2)$	Section 3.4.1
C_k	A correlation between b_{ik} and $P_k(x_i, y_i)$	Section 3.4.1
d_k	A distance from the weight centre of the distribution of $P_k(x, y)$ to the nearest boundary of the analysis region	Section 3.4.1
d_k^r	A normalized value of d_k, defined as $d_k^r = d_k/\sqrt{A}$	Section 3.4.1
D_k	A typical distance from the weight centre of the distribution of $P_k(x, y)$ to nearby stations	Eq. (27)
D_k^r	A normalized value of D_k, defined as $D_k^r = D_k/\sqrt{A}$	Section 3.4.1
$\tilde{E}_j$	The cross-entropy error for a time step j	Eq. (24)
E_k	The cross-entropy error for a time window k	Eq. (23)
f_m	An activation function for an mth layer in a neural network model	Eq. (18)
G	A probability distribution of v for a Gaussian noise (a function of v and N')	Eq. (1)
h	An unknown parameter for an SVM that characterizes the boundary plane as $\boldsymbol{u}^T\boldsymbol{\phi}(\boldsymbol{\xi}_l) + h = 0$	Eq. (29)
i	An index of stations	Section 3.2.3
I	The number of stations	Eq. (23)
I_k^{all}	The number of stations in R_k	Section 3.4.1
I_k^f	$\tanh(I_k^{\text{large}} \ln(3)/4)$	Section 3.4.1
I_k^{large}	The number of stations having a value $b_{ik} \geq 0.5P_k^{\max}$ in R_k	Section 3.4.1
I_k^r	$I_k^{\text{large}}/I_k^{\text{all}}$	Section 3.4.1
I^{th}	The threshold number of stations for a candidate event detection	Appendix A.3
j	An index of time steps ($t = j\Delta t_i$)	Eq. (14)
J_i	The number of samples in T for a station i	Eq. (14)
k	An index of time windows ($kT \leq t < (k+1)T$)	Section 3.2.3
l	An index of teaching data (candidate events) for an SVM	Eq. (28)
l'	An index of teaching data (candidate events) for an SVM	Eq. (30)
L	The number of teaching data (candidate events) for an SVM	Eq. (28)
m	An index of layers in a neural network model	Eq. (17)
M	The number of intermediate layers in a neural network model	Eq. (17)
n	An index of data samples in an ascending order of absolute amplitudes	Eq. (2)
N	The number of samples in $\bar{T}$	Section 3.2.2
N'	The number of samples in $\bar{T}$ corresponding to a Gaussian noise	Section 3.2.2
p	An index of neurons in a neural network model	Eq. (17)
$\tilde{P}_j$	An instantaneous large amplitude probability for a time moment $t = j\Delta t_i$ (a function of location (x, y))	Section 3.3
P_k	A large amplitude probability for a time window k (a function of location (x, y))	Section 3.3
$P_k^{\max}$	The maximum value of $P_k(x, y)$	Section 3.4.1
q	An index of neurons in a neural network model	Eq. (17)
Q_m	The number of neurons in an mth layer of a neural network model	Eq. (17)
r_k	A distance from the weight centre of the distribution of $P_k(x, y)$ to the farthest point in R_k	Section 3.4.1
R_k	A region within which $P_k(x, y) \geq 0.5P_k^{\max}$	Section 3.4.1
S_k	An area of R_k	Section 3.4.1
S_k^r	A normalized value of S_k, defined as $S_k^r = S_k/A$	Section 3.4.1
t	Time	Section 3.2.3
Δt_i	A sampling interval for station i	Eq. (14)
T	The length of a time window to compute a large amplitude ratio (1 s throughout this study)	Section 3.2.3
$\bar{T}$	The length of a time window to evaluate a background noise level	Section 3.2.2
$\boldsymbol{u}$	An unknown parameter for an SVM (a vector) that characterizes the boundary plane as $\boldsymbol{u}^T\boldsymbol{\phi}(\boldsymbol{\xi}_l) + h = 0$	Eq. (28)
v	The velocity value of a data sample	Eq. (1)
v'	The velocity value of a data sample	Eq. (3)
v_n	A velocity whose absolute value is nth smallest in $\bar{T}$	Eq. (2)
v_n^{even}	A velocity whose absolute value is nth smallest among even-numbered data samples in $\bar{T}$	Eq. (6)
v_n^{odd}	A velocity whose absolute value is nth smallest among odd-numbered data samples in $\bar{T}$	Eq. (6)
$W_{qp}^{(m)}$	The weight coefficient for a qth linear combination of $X_p^{(m)}$ in a neural network model	Eq. (17)
x	A location (an eastward distance from the summit)	Section 3.3
x_i	The x-coordinate of a station i	Section 3.3
$x^{\max}$	The eastern end of an analysis region	Eq. (15)
$x^{\min}$	The western end of an analysis region	Eq. (15)
$X_p^{(m)}$	An output signal intensity from a pth neuron of an mth layer in a neural network model	Eqs (15-17)
y	A location (a northward distance from the summit)	Section 3.3

Table 1. Continued

Symbol	Description	Defined in:
y_i	The y-coordinate of a station i	Section 3.3
$y^{\max}$	The northern end of an analysis region	Eq. (16)
$y^{\min}$	The southern end of an analysis region	Eq. (16)
$Y_q^{(m)}$	An input signal intensity to a qth neuron of an mth layer in a neural network model	Eq. (17)
α_l	An unknown parameter for an SVM	Eq. (31)
γ	A tuning parameter for an SVM	Eq. (30)
η_l	An unknown parameter for an SVM that characterizes an extrusion distance of lth teaching data	Eq. (28)
μ	The maximum value of μ^{eo} and μ^{oe} (a function of N')	Eq. (5)
μ^{eo}	A misfit between Θ^{even} and Π^{odd} (a function of N')	Eq. (6)
μ^{oe}	A misfit between Θ^{odd} and Π^{even} (a function of N')	Eq. (7)
$\boldsymbol{\xi}_l$	A feature vector of lth teaching data for an SVM	Eq. (29)
Θ	A cumulative probability distribution of $\lvert v\rvert$ for the observed data of smallest N' samples (a function of $\lvert v\rvert$ and N')	Eq. (4)
Θ^{even}	The distribution Θ for even-numbered data (a function of $\lvert v\rvert$ and N')	Eq. (8)
Θ^{odd}	The distribution Θ for odd-numbered data (a function of $\lvert v\rvert$ and N')	Eq. (9)
Π	A synthetic cumulative probability distribution of $\lvert v\rvert$ corresponding to G (a function of $\lvert v\rvert$ and N')	Eq. (3)
Π^{even}	The distribution Π for even-numbered data (a function of $\lvert v\rvert$ and N')	Eq. (10)
Π^{odd}	The distribution Π for odd-numbered data (a function of $\lvert v\rvert$ and N')	Eq. (11)
ρ_{ki}	A distance from the weight centre of the distribution of $P_k(x, y)$ to a station i	Eq. (27)
σ	A standard deviation of G (a function of N')	Eq. (2)
σ^{even}	A standard deviation of G for even-numbered data (a function of N')	Eq. (12)
σ^{odd}	A standard deviation of G for odd-numbered data (a function of N')	Eq. (13)
$\boldsymbol{\phi}$	A mathematical function (a vector) of a feature vector for an SVM	Eq. (29)
χ	A tuning parameter for an SVM	Eq. (28)
ψ_l	The correct class of lth teaching data for an SVM; 1 for true event, -1 for false one	Eq. (29)

where $|v_n|$ is the nth smallest absolute amplitude in $\bar{T}$. The corresponding cumulative distribution of absolute amplitudes is:

$$\Pi\left(|v|;N'\right)=\int_0^{|v|}2G\left(v';N'\right)dv'=\operatorname{erf}\left[\frac{|v|}{\sqrt{2}\sigma\left(N'\right)}\right], \tag{3}$$

where *erf* represents the error function. This expression is based on the assumption of Gaussian noise, whereas the cumulative distribution of real data is given by

$$\Theta\left(|v_n|;N'\right)=(n-1)/\left(N'-1\right). \tag{4}$$

A search for the optimal N' that minimizes the misfit between $\Pi(|v_n|; N')$ and $\Theta(|v_n|; N')$ produces the background noise level. In practice, small adventitious misfits may be obtained in case of small N'. To avoid this, we divide the time-series data to even- and odd-numbered samples, and calculate

$$\mu\left(N'/2\right)=\max\left\{\mu^{eo}\left(N'/2\right),\mu^{oe}\left(N'/2\right)\right\}, \tag{5}$$

where

$$\mu^{eo}\left(N'/2\right)=\sqrt{\sum_{n=1}^{N'/2}\left[\Theta^{\text{even}}\left(\left|v_n^{\text{even}}\right|;N'/2\right)-\Pi^{\text{odd}}\left(\left|v_n^{\text{odd}}\right|;N'/2\right)\right]^2/(N'/2)}, \tag{6}$$

$$\mu^{oe}\left(N'/2\right)=\sqrt{\sum_{n=1}^{N'/2}\left[\Theta^{\text{odd}}\left(\left|v_n^{\text{odd}}\right|;N'/2\right)-\Pi^{\text{even}}\left(\left|v_n^{\text{even}}\right|;N'/2\right)\right]^2/(N'/2)}, \tag{7}$$

$$\Theta^{\text{even}}\left(\left|v_n^{\text{even}}\right|;N'/2\right)=(n-1)/\left(N'/2-1\right), \tag{8}$$

$$\Theta^{\text{odd}}\left(\left|v_n^{\text{odd}}\right|;N'/2\right)=(n-1)/\left(N'/2-1\right), \tag{9}$$

$$\Pi^{\text{even}}\left(|v|;N'/2\right)=\operatorname{erf}\left(|v|/\sqrt{2}\sigma^{\text{even}}\left(N'/2\right)\right), \tag{10}$$

$$\Pi^{\mathrm{odd}}\left(|v|\,;N'/2\right)=\ \mathrm{erf}\left[|v|\,/\sqrt{2}\sigma^{\mathrm{odd}}\left(N'/2\right)\right], \tag{11}$$

$$\sigma^{\mathrm{even}}\left(N'/2\right)\ =\sqrt{\sum_{n=1}^{N'/2}\left(v_n^{\mathrm{even}}\right)^2/\left(N'/2\right)}, \tag{12}$$

$$\sigma^{\mathrm{odd}}\left(N'/2\right)\ =\sqrt{\sum_{n=1}^{N'/2}\left(v_n^{\mathrm{odd}}\right)^2/\left(N'/2\right)}, \tag{13}$$

and $|v_n^{\mathrm{even}}|$ and $|v_n^{\mathrm{odd}}|$ are the nth smallest absolute amplitudes in the even- and odd-numbered data, respectively. Here, μ^{eo} represents the misfit between the observed cumulative distribution of the even-numbered data (Θ^{even}) and synthetic (Gaussian) cumulative distribution of the odd-numbered data (Π^{odd}); μ^{oe} does *via versa*. Taking the maximum of μ^{eo} and μ^{oe}, the adventitious good fit is avoided. We examine eq. (5) for all the even N' from 4 to N to find the optimal value that minimizes $\mu(N'/2)$. The background noise level is defined as $|v_{N'}|$ for the optimal N'.

3.2.3 Computing large amplitude ratios

Once the background noise level of station i is obtained, we can define a binary time-series $\tilde{b}_i(t)$ for this station, which is 1 if the waveform amplitude at a time t is greater than the background noise level, or 0 otherwise. Using $\tilde{b}_i(t)$, the large amplitude ratio for the kth window of length T (i.e. $kT \le t < (k+1)T$) is expressed as

$$b_{ik}=\frac{1}{T}\int_{kT}^{(k+1)T}\tilde{b}_i\left(t\right)dt=\frac{1}{J_i}\sum_{j\,=\,kJ_i}^{(k+1)J_i-1}\tilde{b}_i\left(j\,\Delta t_i\right), \tag{14}$$

where Δt_i is the sampling interval at this station, and $J_i = T/\Delta t_i$ is the number of samples in T. We use $T = 1$ s throughout the study. In this case, J_i represents the sampling frequency. Note that b_{ik} can be defined using a common T even for a network of mixed sampling rates J_i, although all the data at Mt. Ontake are sampled at 100 Hz.

3.2.4 Detection of candidate events

A candidate event is defined as a continuous time period during which the b_{ik} values of at least five stations exceed 0.3 simultaneously (see Appendix A.3 for more detail); however, two consecutive time periods that meet this criterion are regarded as a single candidate event if the later one starts within 5 s of the end of the earlier one, to avoid duplicated detections of a single event.

3.3 Learning (1): estimating a large amplitude probability for each 1 s

The first learning step is to investigate the probability, $P_k(x, y)$, that the amplitude at each ground-surface location (x, y) is greater than the background noise level over 1 s intervals (a large amplitude probability, Fig. 4). We define x and y to be east and north, respectively, from the summit of Mt. Ontake (N35°53′34″, E137°28′49″). The teaching data of this step are large amplitude ratios b_{ik} given at the station sites (x_i, y_i). We normalize x and y by

$$X_0^{(0)}=\ -1+2\left(x-x^{\min}\right)/\left(x^{\max}-x^{\min}\right), \tag{15}$$

and

$$X_1^{(0)}=\ -1+2\left(y-y^{\min}\right)/\left(y^{\max}-y^{\min}\right), \tag{16}$$

respectively, where $[x^{\min}, x^{\max}] \times [y^{\min}, y^{\max}]$ is a target region of the analysis. We then use a multilayer perceptron neural network model (e.g. Goodfellow *et al.* 2016) to calculate $P_k(x, y)$. In this model, the relationship between $(X_0^{(0)}, X_1^{(0)})$ and $P_k(x, y)$ is expressed by a chain of intermediate variables (neurons) in several layers, where each neuron receives signals from the previous layer and sends a signal to the next layer. The greater the input signal intensity, the greater the output intensity. This relation is expressed as

$$Y_q^{(m+1)}=\sum_{p=0}^{Q_m-1}W_{qp}^{(m)}X_p^{(m)}\ +W_{qQ_m}^{(m)}\quad\left(m\ =\ 0,1,\cdots,M;q\ =\ 0,1,\cdots,Q_{m+1}-1\right), \tag{17}$$

$$X_q^{(m)}=f_m\left(Y_q^{(m)}\right)\quad\left(m\ =\ 1,2,\cdots,M+1;q\ =\ 0,1,\cdots,Q_m-1\right), \tag{18}$$

$$P_k\ (x, y) = X_1^{(M+1)}\ , \tag{19}$$

where M is the number of intermediate layers, Q_m is the number of neurons in the mth layer ($m = 0$ and $m = M + 1$ represent the input and output layers, respectively), $Y_q^{(m)}$ and $X_q^{(m)}$ are input and output signal intensities, respectively, for the qth neuron of the mth layer, and f_m is an activation function that determines the neuron response in the mth layer. The input signal to each neuron is a weighted sum of the signals from the neurons in the previous layer (eq. 17), and the weights $W_{qp}^{(m)}$ are unknown parameters to be investigated. For f_m, a sigmoid function:

$$f_m\ \left(Y_q^{(m)}\right) = \frac{1}{1 + \exp\left(-Y_q^{(m)}\right)} \tag{20}$$

is commonly used, which is 0 for $Y_q^{(m)} = -\infty$ and 1 for $Y_q^{(m)} = \infty$. Other frequently used choices for f_m are a rectified linear unit (ReLU):

$$f_m\ \left(Y_q^{(m)}\right) = \begin{cases} Y_q^{(m)} & \left(Y_q^{(m)} > 0\right) \\ 0 & \left(Y_q^{(m)} \leq 0\right) \end{cases}\ , \tag{21}$$

and a tangent hyperbolic (tanh) function:

$$f_m\ \left(Y_q^{(m)}\right) = \frac{\exp\left(Y_q^{(m)}\right) - \exp\left(-Y_q^{(m)}\right)}{\exp\left(Y_q^{(m)}\right) + \exp\left(-Y_q^{(m)}\right)}\ ; \tag{22}$$

see Goodfellow *et al.* (2016) for more detail.

We investigate $W_{qp}^{(m)}$ by minimizing cross-entropy error:

$$E_k = \frac{1}{I} \sum_{i=1}^{I} \left\{(1 - b_{ik})\left[1 - P_k\left(x_i, y_i\right)\right] + b_{ik} P_k\left(x_i, y_i\right)\right\}, \tag{23}$$

where I is the number of stations. Minimizing the cross-entropy error would be equivalent to maximizing the likelihood that the teaching data are realized by the estimated model if the teaching data attributes were either 0 or 1. Indeed, the attributes b_{ik} are real numbers between 0 and 1. Given that, the question is what quantity is optimized by minimizing E_k in eq. (23). To address this question, imagine investigating an instantaneous large amplitude probability $\tilde{P}_j(x, y)$ at a time moment $t = j\Delta t_i$ using the binary trace $\tilde{b}_i(j\Delta t_i)$. In this case, minimizing cross-entropy error:

$$\tilde{E}_j = \frac{1}{I} \sum_{i=1}^{I} \left\{\left[1 - \tilde{b}_i\left(j\Delta t_i\right)\right]\left[1 - \tilde{P}_j\left(x_i, y_i\right)\right] + \tilde{b}_i\left(j\Delta t_i\right)\tilde{P}_j\left(x_i, y_i\right)\right\} \tag{24}$$

is equivalent to maximizing the likelihood because all $\tilde{b}_i(j\Delta t_i)$ are either 0 or 1. If we assume that $\tilde{P}_j(x, y)$ is constant over a time window of length T:

$$\tilde{P}_j\ (x, y) = P_k\ (x, y)\quad (j = kJ_i, kJ_i + 1, \cdots, (k+1)J_i - 1), \tag{25}$$

then E_k in eq. (23) is equal to the average of $\tilde{E}_j$ in eq. (24) over the window:

$$E_k = \frac{1}{J_i} \sum_{j=kJ_i}^{(k+1)J_i - 1} \tilde{E}_j. \tag{26}$$

We can, therefore, conclude that minimizing E_k in eq. (23) is equivalent to minimizing $\tilde{E}_j$ in eq. (24) under an assumption of constant $\tilde{P}_j(x, y)$ over a 1 s window.

We performed a neural network analysis of large amplitude ratios to investigate large amplitude probabilities around Mt. Ontake over the entire study period. We used $x^{\min} = y^{\min} = -30$ km and $x^{\max} = y^{\max} = 30$ km. We generated initial models by Gaussian random values with an average of 0 and a standard deviation of 1, and used $M = 2$, $Q_1 = 5$, $Q_2 = 2$, $f_m = \tanh$ for $m \leq M$ and $f_{M+1} = \text{sigmoid}$, which were the best choices based on our tests using ideal data (Appendix A.4). To stabilize the analysis results, we used dummy stations of zero values on the target region edges with a 2 km interval (Appendix A.5). We thus had real stations within $\pm$ 20 km in each direction from the summit (Fig. 1) and dummy stations with zero values along the boundaries at $\pm$ 30 km, meaning that the large amplitude probability was gradually suppressed to zero between 20 and 30 km. We investigated the best neural network parameters ($W_{qp}^{(m)}$) by the gradient method with an Adadelta algorithm (Zeiler 2012), where a decay constant of 0.95 and the other constant (ϵ in Zeiler 2012) of 1×10^{-8} were used. We used 1000 iterations from 20 initial models, which are shown to be enough to avoid falling to a local minimum (Appendix A.5).

3.4 Learning (2): classification as true and false events

3.4.1 Feature extraction

The second learning step is to investigate a relationship between the spatial distribution of a large amplitude probability and the classification as true or false events (Fig. 4). In principle, the distribution is completely described by $W_{qp}^{(m)}$. However, our preliminary tests using $W_{qp}^{(m)}$ did not perform well. We thus instead calculated eight features from the large amplitude probability, which well characterized the difference between true and false events. We computed the features for the first 3 s of each candidate event because the difference between true and false events was unclear in later time periods. Therefore, a total of 24 features were available for each candidate event.

The eight features are: (1) the maximum value, $P_k^{\max}$, of the large amplitude probability $P_k(x, y)$; (2) the area, S_k, of a high probability region $R_k = \{(x, y);\ P_k(x, y) \geq 0.5P_k^{max}\}$; (3) the circularity of R_k defined as $c_k = S_k/(\pi r_k^2)$, where r_k is the distance from the weight centre of the distribution of $P_k(x, y)$ to the farthest point in R_k; (4) the distance, d_k, between the weight centre and the nearest boundary of the analysis region; (5) a typical distance from the weight centre to nearby stations, defined by

$$D_k = \left[\frac{1}{I}\sum_{i=1}^{I}\rho_{ki}^{-2}\right]^{-1/2}, \tag{27}$$

where ρ_{ki} is the distance from the weight centre to ith station; (6) the number, I_k^{large}, of stations with value $b_{ik} \geq 0.5P_k^{\max}$ in R_k; (7) the correlation, C_k, between b_{ik} and $P_k(x_i, y_i)$; and (8) $I_k^r = I_k^{\text{large}}/I_k^{\text{all}}$, where I_k^{all} is the total number of stations in R_k. Here, the circularity c_k is 1 if the region R_k is a perfect circle and becomes smaller as R_k elongates. In the definition of D_k (eq. 27), the inverse average is used to emphasize stations close to the weight centre. Instead of S_k, d_k, D_k and I_k^{large}, we use normalized quantities $S_k^r = S_k/A$, $d_k^r = d_k/\sqrt{A}$, $D_k^r = D_k/\sqrt{A}$ and $I_k^f = \tanh(I_k^{\text{large}}\ln(3)/4)$, respectively, where $A = (x_{\max} - x_{\min})(y_{\max} - y_{\min})$. The definition of I_k^f is designed to emphasize small values of I_k^{large}; $I_k^f = 0$ for $I_k^{\text{large}} = 0$, $I_k^f = 0.5$ for $I_k^{large} = 2$, and I_k^f approaches to 1 for $I_k^{\text{large}} \to \infty$.

True and false events are expected to be distinguished by these features. As will be shown in Section 4.2, a high probability in a relatively small region of the circular shape was typical for true events; thus, a large $P_k^{\max}$, a small S_k^r, and a large c_k are expected. However, these characteristics were also observed for some earthquakes from outside the analysis region and local noise recorded by a small number of stations in a sparse area. In these cases, the high probability region was near an analysis domain edge or in an area with low station density, suggesting smaller d_k^r, larger D_k^r, or smaller I_k^f than those for true events. Additionally, true events tended to show better matches between observed large amplitude ratios and predicted large amplitude probabilities, which would lead to larger values of C_k and I_k^r than those for false ones.

3.4.2 Classification and selection of candidate events

We manually classified the candidate events in the study period into true and false events. The true events point to seismic events that occurred within the analysis target area, including LP and VT events, volcanic tremors and local tectonic earthquakes occurring beneath the flank of the volcano. The false events consist of all the other kind of signals, including regional and distant earthquakes from outside the analysis region, electronic noise and local noise caused by, for example, human activity or the wind. Small volcano seismic events were difficult to distinguish from local wind noise by looking at a single waveform trace. Given the proximity of stations ($\sim$ 100 m) in the summit region (Fig. 1), true volcano seismic events should have consistent waveforms, amplitudes, timings of distinct wave packets and spectral characteristics among the stations. We examined these consistencies to distinguish true and false events. Even after these examinations, some candidate events could not be identified whether they were true or false. To preserve the teaching data quality, we used only the candidate events that could be confidently classified. We also did not use earthquakes between $\pm$ 20 and $\pm$ 40 km in each direction from the summit as teaching data. This means that tectonic earthquakes from less than $\pm$ 20 km and more than $\pm$ 40 km are regarded confidently to be true and false events, respectively, whereas the region from $\pm$ 20 to $\pm$ 40 km is used as a transition zone from true to false. We do not include background noise as false events; however, local wind noise exhibit overall small values of the large amplitude probabilities (Section 4-2), similar to the background noise which shows almost zero values in the probabilities.

After the manual classification, the number of false events was significantly greater than the true ones. We selected the false events randomly to ensure the teaching data were composed of equal numbers of true and false events. We then randomly divided the teaching data into training (80 per cent) and test (20 per cent) data sets. We created 200 data sets by using different random choices for the false events and separations to the training and test data.

3.4.3 Learning the relation between the features and classification

Using the 24 features (Section 3.4.1) and manual classifications (Section 3.4.2) of candidate events as teaching data, we investigate a boundary plane in a 24-D feature space, which separates true and false events. A neural network model is not the best choice for this step, because the model in this large dimension requires many free parameters whereas the number of teaching data is limited to the number of manually classified candidate events. A support vector machine (SVM) model would be more stable in case of a relatively small data set (e.g. Qu *et al.* 2019). It investigates the boundary plane by maximizing the distance between the boundary and the nearest teaching data in each class. In

this study, we use a C-support vector classification (C-SVC), a kind of SVM algorithm in which extrusions of several data to the opposite side of the boundary are allowed. This is realized by minimizing:

$$\frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} + \chi \sum_{l=1}^{L} \eta_l \tag{28}$$

under constraints:

$$\psi_l \left[\boldsymbol{u}^T \boldsymbol{\phi}(\boldsymbol{\xi}_l) + h\right] \geq 1 - \eta_l, \quad \eta_l \geq 0, \tag{29}$$

where $\boldsymbol{\xi}_l$ is the feature vector of the lth training data; ψ_l is the correct data class, which is 1 if the candidate event is a true one, and -1 if it is a false one; $\chi(>0)$ is a tuning parameter; $\boldsymbol{\phi}(\boldsymbol{\xi}_l)$ is a mathematical function used to realize a nonlinear boundary plane; the superscript T represents the transpose of a matrix; and $\boldsymbol{u}$, h and η_l are unknown parameters (Chang & Lin 2011). The boundary plane is expressed by $\boldsymbol{u}$ and h as $\boldsymbol{u}^T\boldsymbol{\phi}(\boldsymbol{\xi}_l) + h = 0$, and η_l represents an extrusion distance of the lth training data to the opposite side. The first term in eq. (28) represents the square inverse of the distance between the boundary plane and nearest teaching data, and the second term measures the total extrusion distance of the data to the opposite side of the plane; therefore, the larger the χ value, the more tightly the extrusion is suppressed. The boundary plane is flat if $\boldsymbol{\phi}(\boldsymbol{\xi}_l)$ is linear, and bent if a nonlinear $\boldsymbol{\phi}(\boldsymbol{\xi}_l)$ is used. Following Hsu *et al.* (2003), we use a Gaussian kernel:

$$\boldsymbol{\phi}(\boldsymbol{\xi}_l)^T\boldsymbol{\phi}(\boldsymbol{\xi}_{l'}) = \exp\left(-\gamma\left|\boldsymbol{\xi}_l - \boldsymbol{\xi}_{l'}\right|^2\right) \tag{30}$$

for the choice of $\boldsymbol{\phi}(\boldsymbol{\xi}_l)$, where γ is a positive constant. Note that only the product $\boldsymbol{\phi}(\boldsymbol{\xi}_l)^T\boldsymbol{\phi}(\boldsymbol{\xi}_{l'})$ is needed without requiring an explicit form of $\boldsymbol{\phi}(\boldsymbol{\xi}_l)$. This is because minimizing eq. (28) under constraints (29) is equivalent to minimizing:

$$\sum_{l=1}^{L} \alpha_l - \frac{1}{2}\sum_{l,\,l'=1}^{L} \alpha_l\alpha_{l'}\psi_l\psi_{l'}\boldsymbol{\phi}(\boldsymbol{\xi}_l)^T\boldsymbol{\phi}(\boldsymbol{\xi}_{l'}) \tag{31}$$

under constraints

$$\sum_{l=1}^{L} \alpha_l\psi_l = 0, \quad 0 \leq \alpha_l \leq \chi, \tag{32}$$

where α_l are unknown parameters. In this framework, χ and γ are tuning parameters. Hsu *et al.* (2003) proposed to search for the optimal choices for χ and γ by exponentially varying them. Following their study, we first varied $\log_2\chi$ and $\log_2\gamma$ from -5 to 5 at increments of 1, and then varied them around the optimal value of the first search at increments of 0.1. For each combination of χ and γ, we count the number of wrong classifications in the 200 test data sets (Section 3.4.2). The values of χ and γ that minimize the wrong classifications are regarded as the optimal choices. We used the libSVM library (Chang & Lin 2011) for our implementation of the C-SVC analysis.

4 RESULTS

4.1 Results for preparation of input data for machine learning

Our approach to the background noise level estimation (Section 3.2.2) assumed that the background noise obeys a Gaussian distribution and spike-like signals do not. We first evaluated this assumption using several records with different activity levels. When there is no distinct signal (Fig. 5aA), the observed cumulative distribution was well fitted by the synthetic (Gaussian-based) one (Fig. 5bA). However, in a record with intense earthquake signals, a significant deviation occurred between the two distributions (Fig. 5B). We see smaller deviations in records with smaller earthquakes (Fig. 5C) and still smaller spikes of noise (Fig. 5D). These observations indicate the validity of our assumption in the background noise level estimation.

In Fig. 6, we summarize the background noise level estimations for the four traces in Fig. 5. Fig. 6(a) shows the misfits, $\mu(N'/2)$ (eq. 5), between observed and synthetic (Gaussian-based) cumulative distributions plotted against N'. For the record with large earthquake signals, the misfit was minimal at $N' = 28010$ (Fig. 6aB). This means that among all the 30 000 samples in this 5 min, the smallest 28 010 data obeyed the Gaussian distribution, whereas the remaining 1990 samples deviated from it. The optimal N' was greater for records with smaller signals (Fig. 6a). In Fig. 6(b), we compare the two distributions of optimal N', showing excellent fits. Thick dashed lines in Fig. 6(c) indicate the amplitudes corresponding to the optimal N'. For comparison, we plot average envelope amplitudes multiplied with 3 (thin lines in Fig. 6c), which are typically used as a detection threshold in the STA/LTA method. The proposed and conventional (STA/LTA) methods provide similar thresholds when the record consists of no distinct signal (Fig. 6cA). When a distinct signal is present, the threshold by the proposed method is still consistent with the background noise, whereas a significant overestimation occurs in the conventional one (Fig. 6cB). The large amplitude ratios (eq. 14) in every 1 s, calculated using the estimated background noise levels, are consistent with distinct signals in the original waveforms (Fig. 6d). Based on the large amplitude ratios, we identified 2030 candidate events, which satisfied the detection criteria which we already have explained in Section 3.2.4.

We further examine whether the background noise levels could alternatively be estimated by LTA in long time windows. The background noise levels are used for the detection of candidate events and the calculation of large amplitude ratios. Fig. 7(a) shows that if we used three

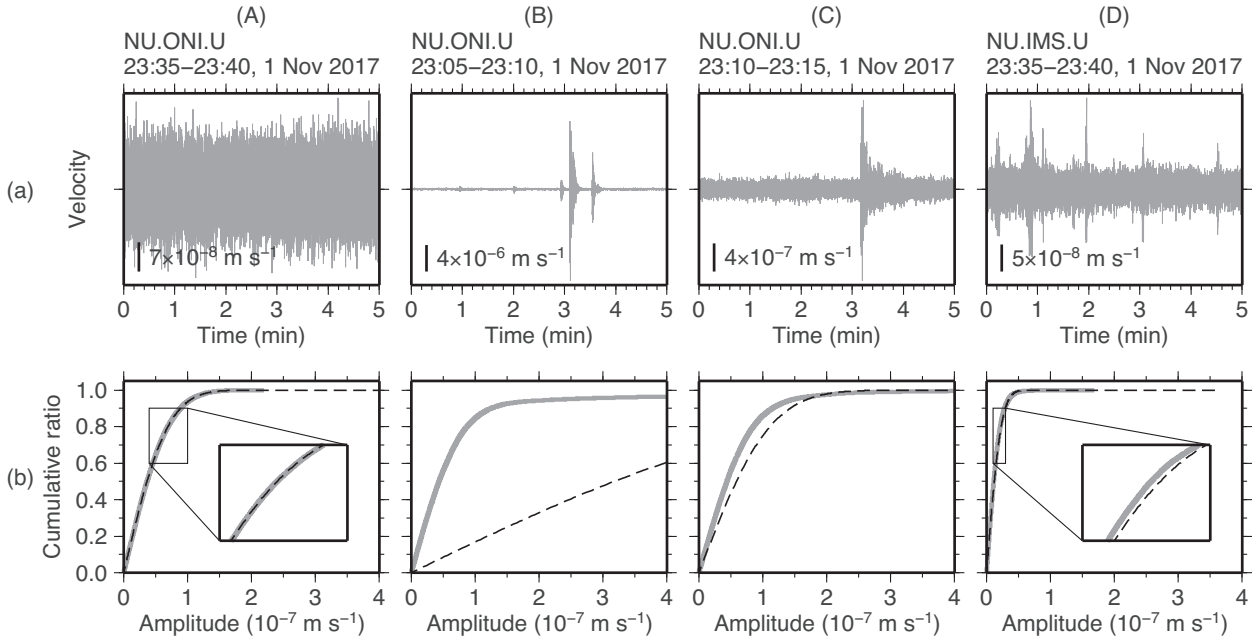

Figure 5. (a) Waveforms in 5-min window consisting of (A) no distinct signal, (B) intense seismic signals, (C) a weaker seismic signal and (D) still weaker pulse-like local noise, from three different time periods and two different stations. (b) Comparisons of cumulative amplitude distributions for the data (eq. 4; grey lines) and Gaussian noise (eq. 3; dashed lines).

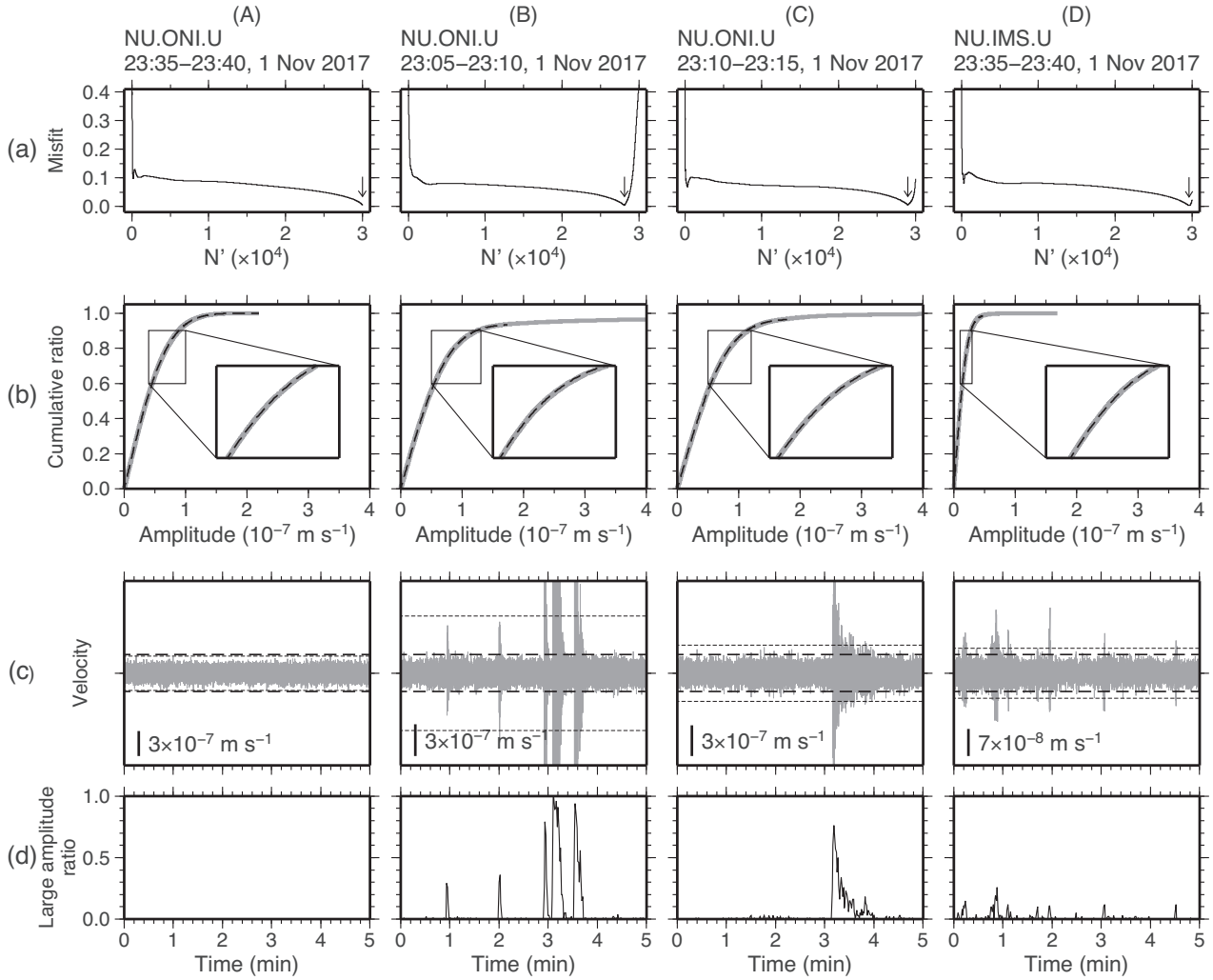

Figure 6. Investigations of the background noise levels and large amplitude ratios for the data in Fig. 5. (a) Misfits between cumulative amplitude distributions for the data and Gaussian noise calculated with the smallest N' samples (eq. 5), plotted against N'. Arrows represent the minimum misfits. (b) Comparisons of the two distributions (grey: data, black: Gaussian) for the optimal N'. (c) Comparisons of the waveforms (grey), estimated background noise levels (thick lines) and three times the average envelope amplitudes of the data (thin lines). (d) The large amplitude ratios are calculated for every 1 s.

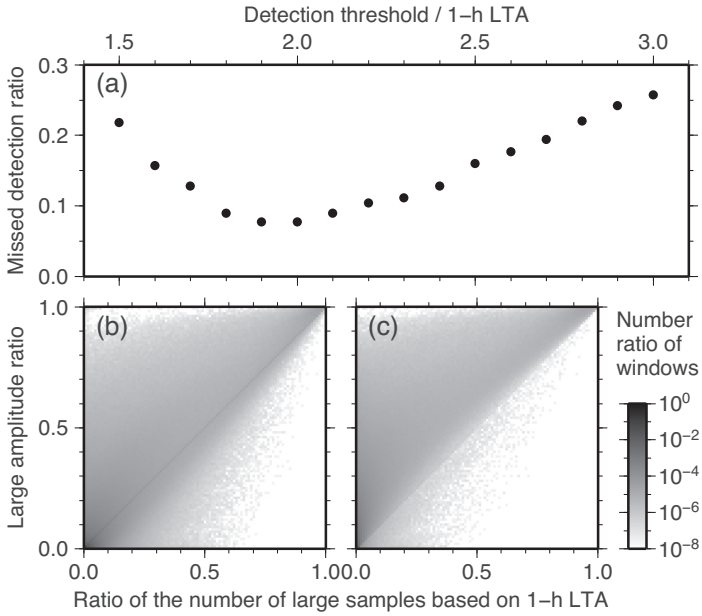

Figure 7. Comparisons of the results from proposed and conventional (LTA-based) methods for the background noise level estimation. (a) The number ratios of true events which were not detected when we used LTA multiplied by the horizontal axis values. (b) and (c) The number ratios of 1-s windows plotted against the combinations of large amplitude ratios calculated with conventional (the lateral axis) and proposed (the vertical axis) methods. In the conventional method, two and three times LTA are used in (b) and (c), respectively, for the background noise level. The average envelope amplitudes of 1-hr data are used as LTA.

times 1-hr LTA for the detection threshold, more than 25 per cent of events would be missed, and this ratio would decrease to less than 10 per cent if we lowered the threshold to two times LTA. Further lowering the threshold results in an increase of the missed ratio as some events are obscured by small false events. In Figs 7(b) and (c), we compare the large amplitude ratios calculated with the proposed method (the vertical axis) and those calculated as the ratios of the numbers of samples with amplitudes greater than two times (Fig. 7b) and three times (Fig. 7c) 1-hr LTA (lateral axes). Using three times LTA, the background noise levels tend to be overestimated (Fig. 6c), resulting in overall underestimates of the ratios (Fig. 7c). Using two times LTA, the background noise levels are underestimated for quiet traces and overestimated for active traces, resulting in a broad scatter (Fig. 7b). These results suggest that, although two times 1-hr LTA could alternatively be used for the detection part, the background noise levels need to be investigated by the proposed method to ensure stable estimates of large amplitude ratios which are used directly in the neural network model.

4.2 Results for learning (1): estimating a large amplitude probability for each 1 s

Although we investigated the large amplitude probabilities for all of the 2030 candidate events, we show typical results from seven candidate events of different types, which are LP (Fig. 8a) and VT (Fig. 8b) events, a local tectonic earthquake beneath the flank of the volcano (Fig. 8c), a regional earthquake from outside the target region (Fig. 8d), a distant earthquake (Fig. 8e), electronic noise caused by sensor check signals at NIED stations (Kunitomo 2014) (Fig. 8f) and local noise (Fig. 8g). The VT event in Fig. 8(b) is in the summit catalogue (800 m west and 934 m south of the summit of Mt. Ontake; magnitude: −0.4). The earthquake in Fig. 8(c) is in the routine catalogue (9033 m east and 2628 m south of the summit, 2618 m below sea level, magnitude: 0.2). The earthquake in Fig. 8(d) is reported by the Japan Meteorological Agency as located at 36.859° N, 140.548° E (∼300 km northeast of Mt. Ontake) and 8.6 km below sea level, with a magnitude of 2.9 and an origin time at 3:14:41. The distant earthquake in Fig. 8(e) is from 13.851° N, 144.822° E (near Guam island, U.S.), and 143.7 km below sea level, with a magnitude of 4.8 and an origin time at 23:46:19 JST on 2017 October 31, according to the U.S. Geological Survey.

The results from these seven typical candidate events are shown in Figs 9–10. Here, the darknesses in the circles represent the large amplitude ratios at stations, and the background gray scales represent the continuous spatial distributions of large amplitude probabilities. True events (Figs 8a–c) are characterized by a high probability in a relatively small region of circular shapes (Fig. 9), whereas false events (Figs 8d–g) are characterized by an elongated region of high probabilities (Fig. 10a) or a widespread region of low probabilities (Figs 10b–d). This difference for the first 3 s of each candidate event is useful to distinguish true and false events. We note that the difference was unclear in later time periods because the seismic wave of true events was spread over the entire network (Fig. 9c), leading to patterns similar to Fig. 10.

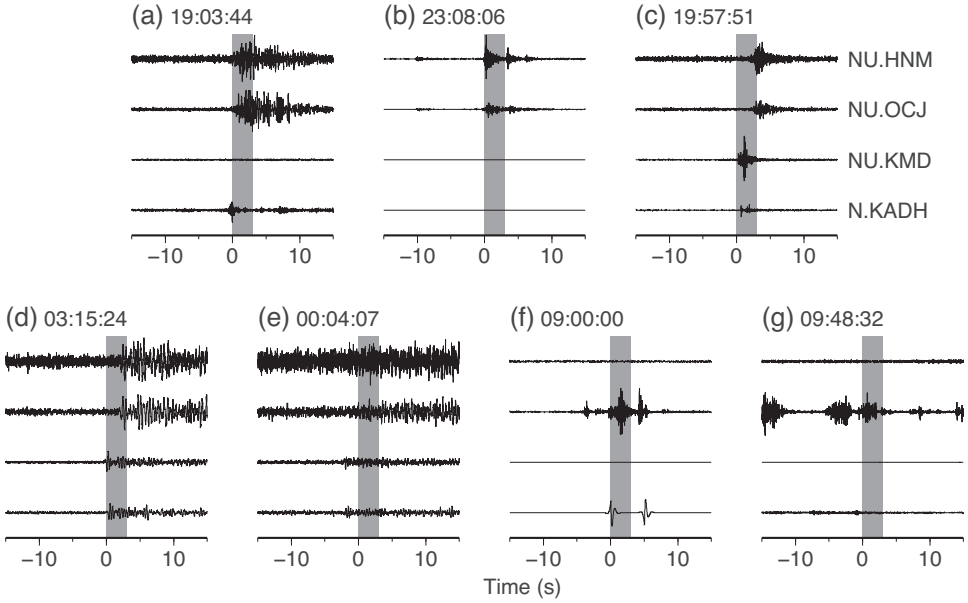

Figure 8. Waveforms of seven events that showed typical amplitude patterns at Mt. Ontake. (a) An LP event, (b) a VT event, (c) a local tectonic earthquake beneath the eastern flank, (d) a regional earthquake from outside the analysis region, (e) a distant earthquake, (f) electronic noise and (g) local noise. Vertical waveforms high-pass filtered at 1 Hz are shown. A common amplitude scale is used for the four traces in each panel. Time 0 in each panel corresponds to the start time of each candidate event on 2017 November 1 shown on the top. Grey bars represent the time frames used in Figs 9–10.

4.3 Results for learning (2): classification to true and false events

The 2030 detected candidate events (Section 4.1) consisted of 412 true and 1452 false events based on our manual classification (Section 3.4.2); the remaining 166 were difficult to classify. After a random selection of the false events (Section 3.4.2), we had 824 samples in each data set, which were divided randomly into 659 (80 per cent) training and 165 (20 per cent) test data.

A grid search for χ and γ (eq. 28 and 30) showed that $\log_2\chi = 1.2$ and $\log_2\gamma = 1.9$ were the best choices (Fig. 11). Using these values, 911 true and 897 false events were mislabelled as false and true ones, respectively, among a total of 33 000 candidate events in the 200 random test data sets (Table 2). Note that the former and latter types of mislabelling would result in missing and misdetections, respectively, if we use a list of true events created by the SVM model as the final list of events. The total number of missing and misdetections was 1808 (5.5 per cent).

From the 200 independent SVM-based candidate models corresponding to the 200 random data sets, one was chosen as the final model. To do this, we applied each model independently to classify all the candidate events, and compared the results with the manual classifications. The model that yielded the smallest number of wrong classifications was selected as the final model. This best model resulted in 6 (1.5 per cent) missing detections among 412 true events and 44 (3.0 per cent) misdetections among 1452 false ones (Table 3). The total number of wrong classifications was 50 (2.7 per cent) among the 1864 candidate events for which the manual classifications were given, indicating that more than 97 per cent of the candidate events were correctly classified by the optimal model.

5 DISCUSSION

Using the methods and results presented so far, we can construct an automated event detection system as described below. From the continuous waveform at each station, a time-series of large amplitude ratios is first created. Using these ratios, candidate events are detected. For the first 3 s of each candidate event, large amplitude probabilities are investigated using a neural network model. The eight features of the probabilities at 1 s intervals are used as the inputs of the optimal SVM model, obtained in Section 4.3, which gives a true or false label to each candidate event and creates an automatic list of true events.

Our study shows that if we apply this procedure to the period of 2017 November 1–10, a list of 547 events would be created, 44 (8.0 per cent) of which are false (Table 3). This ratio is higher than the incorrect classification ratio of 2.7 per cent (Section 4.3) because of the large total number of false events, but is still small enough to use the automatically created event list for evaluating a temporal change in seismicity. The number of events detected by this procedure (547, or 503 excluding false events; Table 3) was more than twice that of previously known earthquakes (182 in the routine catalogue and 26 in the summit catalogue). The detected events consisted of 166 (91.2 per cent) earthquakes in the routine catalogue and 25 (96.2 per cent) VT events in the summit catalogue; two earthquakes in the routine catalogue were mislabelled as false events by the SVM model, and the remaining 14 in the routine catalogue and a VT event in the summit catalogue were not detected as candidate events (Fig. A1; Appendix A.3). The detected events also consisted of approximately 10 LP events, defined as events with peak frequencies less than 5 Hz and confined to the summit region (Fig. 12a), including those not identified manually in advance (Section

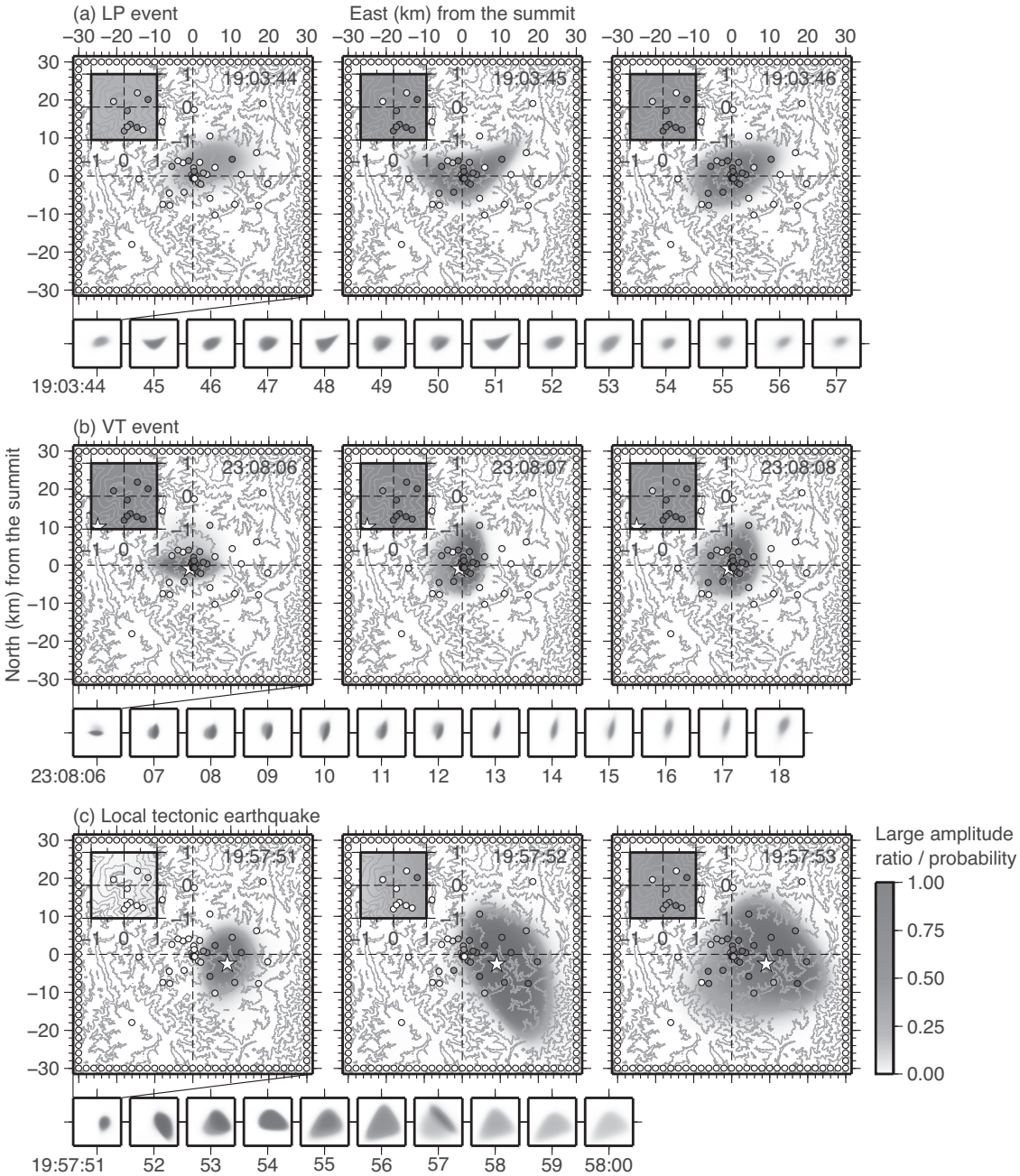

Figure 9. Comparison of the large amplitude ratios at station sites used as the teaching data (circles) and the large amplitude probabilities investigated by a neural network model (the background grey scale) for the first 3 s of true events (Figs 8a–c). White circles along ± 30 km in east or north represent dummy stations with zero values (Section 3.3 and Appendix A.5). Stars in (b) and (c) represent epicentres. Contours represent the topography. The inset in each panel represents an extension around the summit region. The small frames below show the large amplitude probabilities in the entire analysis region during the time period of each candidate event. The main point of this figure is that true events show high probabilities in relatively small regions of circular shapes.

2.3). Their exact number was difficult to count because it was often the case that some stations showed a peak frequency in the LP band (< 5 Hz) whereas the others did not. Additionally, four of the detected events seem to come from a northern region (near station P12; Fig. 12b) where seismicity has not been detected before. These results indicate that both the number and variety of detected seismicity improves by the proposed procedure. We note that a majority of candidate events detected solely by signal-to-noise ratios are false ones (Table 3), especially in the day time (Fig. 13a) when many spike-like local noise signals (Fig. 3b) are present in the records. Therefore, an automatic classification of the candidate events into true and false events is important to significantly reduce manual work to check the waveforms.

We can apply the same procedure to different time periods. Fig 13(b) shows the hourly number of detected events for a time period around an eruption in 2014, which started at 11:52:30 on 2014 September 27 (Maeda *et al.* 2015) and lasted for several hours, with the most active period continuing until 12:15 (see Oikawa *et al.* 2016 for more detail). A corresponding increase in the number of candidate events,

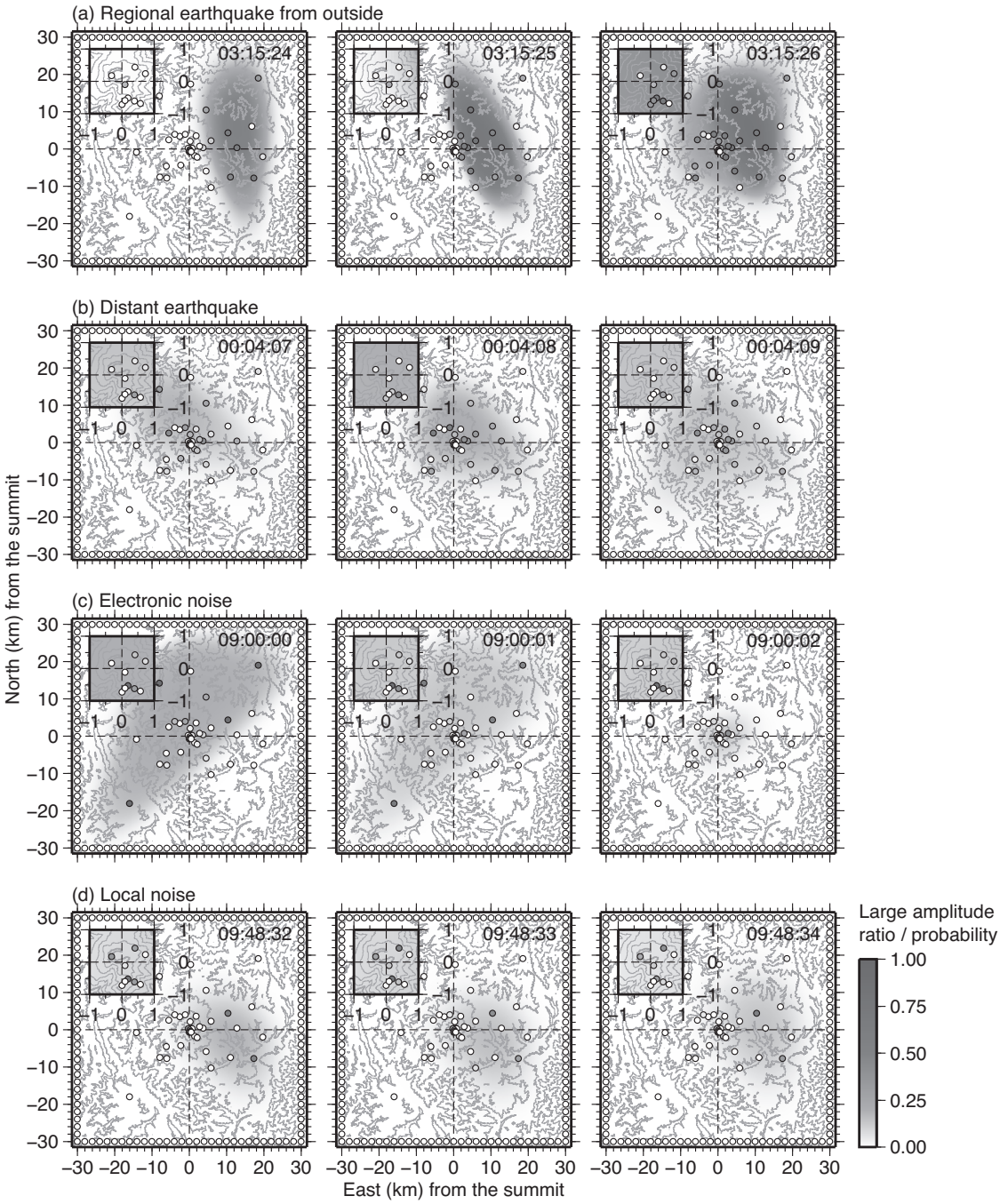

Figure 10. Comparison of the large amplitude ratios at station sites used as the teaching data (circles) and the large amplitude probabilities investigated by a neural network model (the background grey scale) for the first 3 s of false events (Figs 8d–g). The plotting formats are same as those for Fig. 9. The main point is that false events show either (a) elongated high probability regions or (b)–(d) widespread regions of low probabilities.

most of which were labelled as true events based on the SVM model, is visible in Fig. 13b. An earthquake immediately after the onset of the eruption (Fig. 14a) was also identified as a true event.

Fig. 13(c) shows the hourly number of detected events for a time period around a large (M5.6) earthquake, which occurred at 7:02:15 on 2017 June 25. The main shock (Fig. 14b) and most aftershocks (Fig. 13c) were identified as true events. In Fig. 15(a), we compared the hourly numbers of events detected by the proposed method and in the routine catalogue. Immediately after the main shock, the proposed method detected a smaller number of events than the routine catalogue because many consecutive aftershocks were identified as single events (Fig. 15b). Several hours later, the proposed method resulted in comparable or greater numbers of events than the routine catalogue (Fig. 15a), and the detected events were consistent with distinct earthquake signals in the waveform (Fig. 15c).

Note that the SVM model was created by the teaching data from 2017 November 1–10, and consisted of only small ($\leq$ M2.0) earthquakes with no eruption. Nevertheless, the model was able to identify seismicity associated with the eruption and large earthquake. This success may

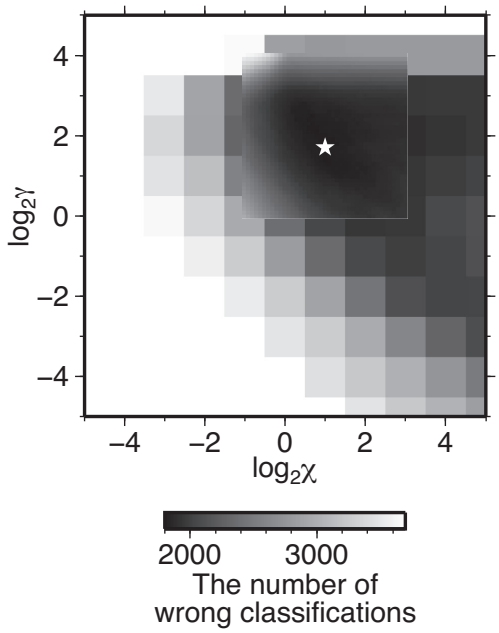

Figure 11. The number of wrong classifications in the 200 random test data (33 000 samples in total) plotted against $\log_2\chi$ and $\log_2\gamma$ (eqs 28 and 30). The star indicates the optimal values that minimized the number of wrong classifications.

Table 2. Summary of the classification results of 200 random test data for $\chi = 1.2$ and $\gamma = 1.9$ (the optimal choice). The number of candidate events in the study period (2017 November 1–10) classified by each type is shown.

		Prediction by the best SVM model		
		True events	False events	Total
Manual classification	True events	15 520	911	16 431
	False events	897	15 672	16 569
	Total	16 417	16 583	33 000

Table 3. Summary of the best classification results from the SVM analysis. The number of candidate events in the study period (2017 November 1–10) classified by each type is shown.

		Prediction by the best SVM model		
		True events	False events	Total
Manual classification	True events	406	6	412
	False events	44	1408	1452
	Unknown	97	69	166
	Total	547	1483	2030

be related to the use of large amplitude ratios, in which only the information on whether the amplitude is greater than the background noise level is kept, losing the absolute amplitude information. Additionally, because we used only the first 3 s of each candidate event, the signal did not reach the entire network even for large events (Fig. 14b), resulting in a similar pattern between large and small events.

Detection of long-lasting tremors may be more challenging as shown in Fig. 16. Because we investigate the background noise level for every 5 min period, a tremor longer than 5 min results in an overestimation of the background noise level (Fig. 16a), and thus, an underestimation of the large amplitude ratios (Fig. 16b). Nevertheless, we can detect several time sections of tremor when the amplitude varies with time (e.g. 11:48:44–11:49:59 in Fig. 16).

One question is if hypocentres of detected events be estimated. Figs 9(b) and (c), and 14(b) suggest that when a large amplitude probability shows a circular shape, the centre of the circle may be close to the epicentre. To examine this hypothesis, we compared the weight centres of large amplitude probabilities, evaluated over a 1 s window, which showed the highest circularity c_k among the first 3 s of each event, with the epicentres for the earthquakes in the routine and summit catalogues. The comparison showed that the weight centres were systematically inward for earthquakes near a margin of the analysis region (Fig. 17a) and eastward for summit VT events (Fig. 17b), probably due to an inhomogeneous azimuthal station density distribution from the epicentres. The average distances between the epicentres and weight centres were 3.5 and 2.9 km for the routine and summit catalogues, respectively. We also compared the events from 2017 June 24–27, most of which were aftershocks of the M5.6 earthquake. For these data, the weight centres (Fig. 18b) were more scattered than the epicentres (Fig. 18a; see Fig. 18c for a comparison of locations). The weight centres showed a relatively small scatter when we focused on

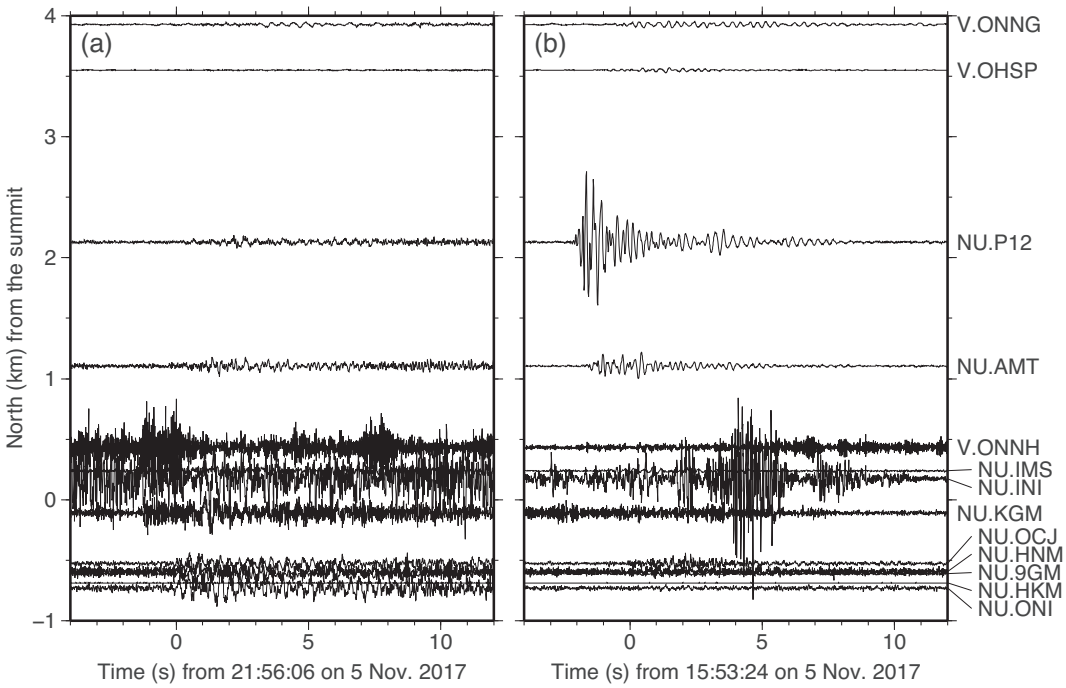

Figure 12. Examples of (a) an LP event and (b) an event coming from a northern area of the volcano detected by the proposed method. High-pass filtered (1 Hz) waveforms at stations within 2 km east/west of the summit are plotted along the N–S coordinates.

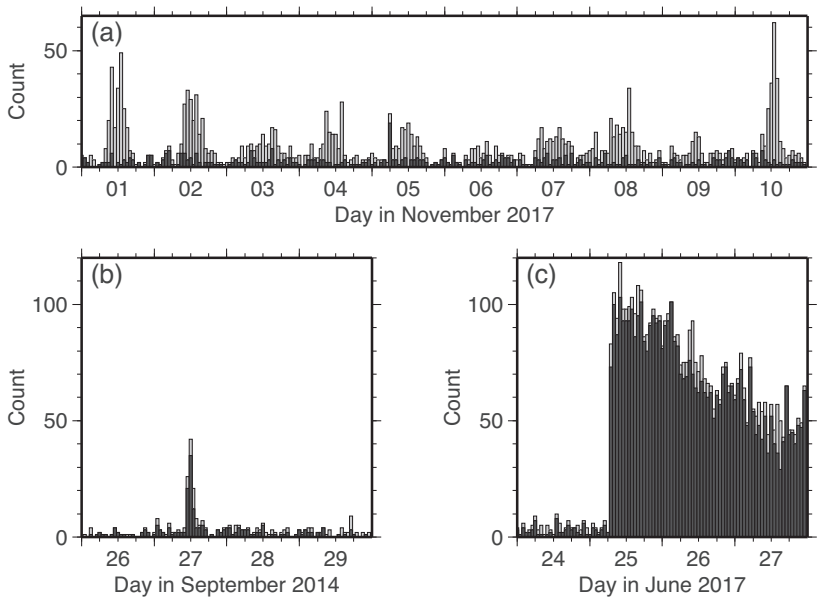

Figure 13. Hourly numbers of candidate events (a) from 2017 November 1–10 (the main study period), (b) from 2014 September 26–29 (around an eruption) and (c) from 2017 June 24–27 (around a large (M5.6) earthquake). Dark and light grey bars represent the numbers of true and false events based on the SVM classifications, respectively.

the events with $c_k \geq 0.9$ (black circles in Fig. 18b). Furthermore, the number density of the event weight centres was consistent with the epicentre distribution of aftershocks (Fig. 18d). These results suggest that the spatial distribution of the weight centres could be used for a rapid investigation of epicentral distributions for abundant seismicity such as aftershocks, although the locations need to be evaluated later.

A key point of the proposed algorithm is a high flexibility of the trained classifier to future changes in the station network. This is because the SVM model refers to continuous spatial distributions of large amplitude probabilities without assuming a specific station layout. Owing to this nature, the SVM model needs not be re-trained after addition, removal, or movement of stations. This is shown by the application of the SVM model, trained by the data in 2017 November 1–10, to the two different time periods when the station networks were also different (Fig. 14). An implication of this flexibility is a potential transferability of the SVM model, obtained at Mt. Ontake, to other volcanoes. Given

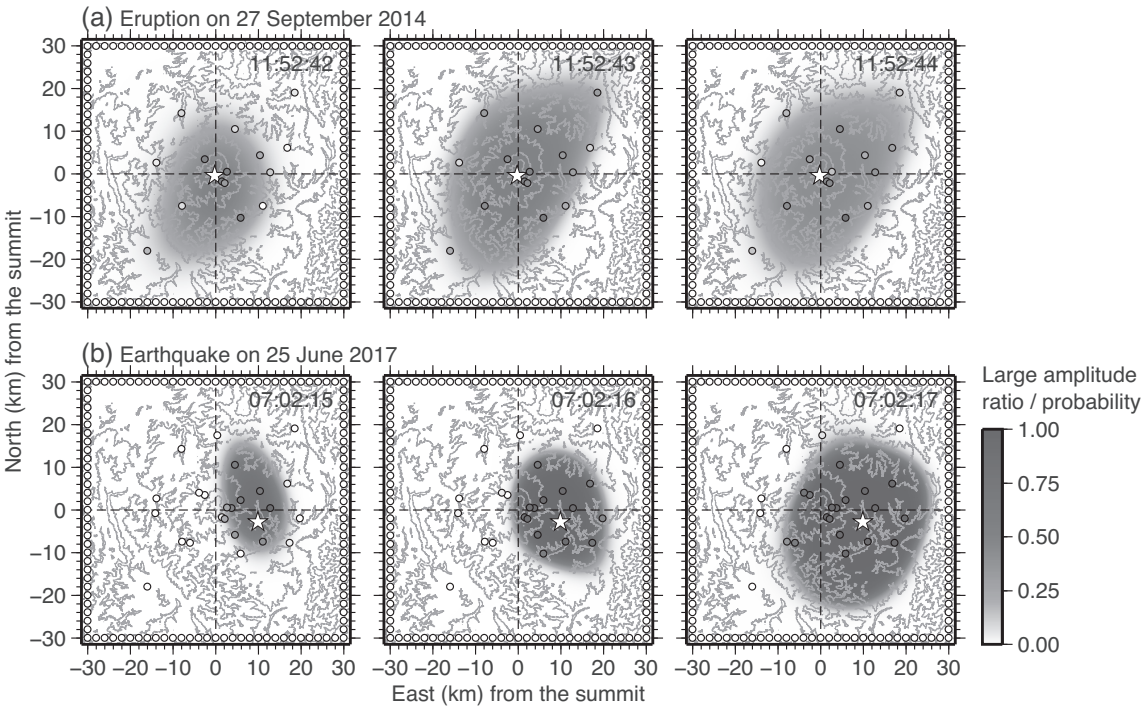

Figure 14. Comparison of the large amplitude ratios at station sites used as the teaching data (circles) and the large amplitude probabilities investigated by a neural network model (the background gray scale) for the first 3 s of (a) an event immediately after the onset of a phreatic eruption on 2014 September 27, and (b) a large (M5.6) earthquake on 2017 June 25. The plotting formats are same as those in Fig. 9. Stars in (a) represents a VLP source of the eruption (Maeda *et al.* 2015). The main point of this figure is that these large events show similar probability patterns to smaller true events (Fig. 9).

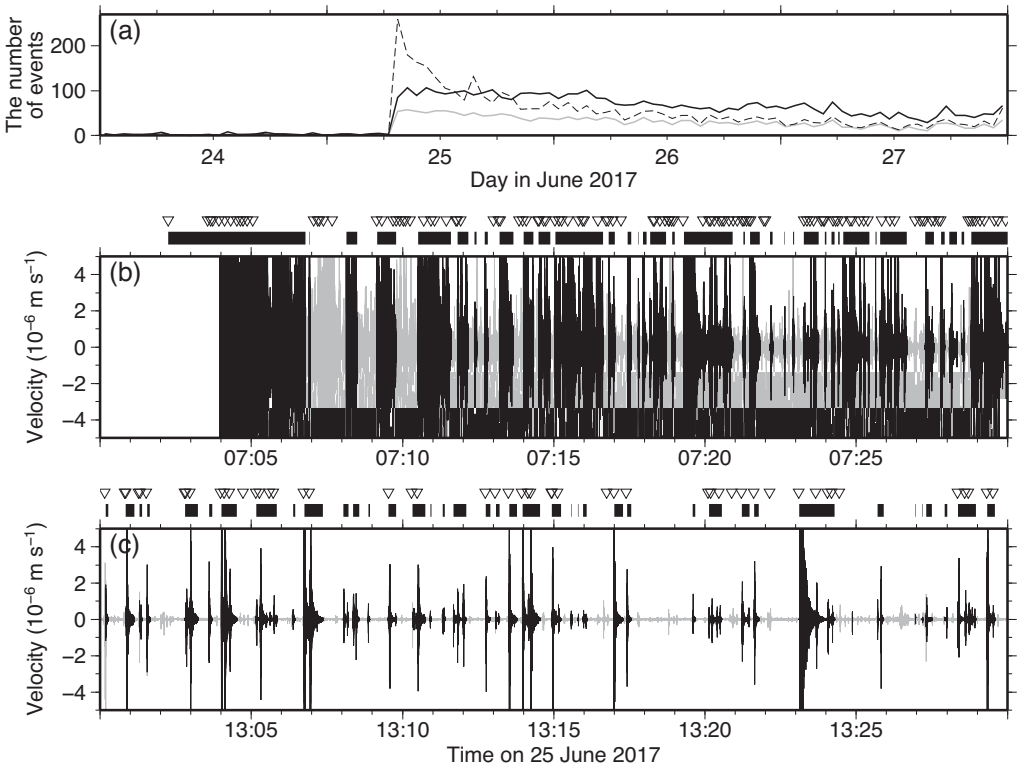

Figure 15. (a) The hourly numbers of events detected by the proposed method (the black solid line) and in the routine catalogue (the dashed line). The grey line represents the number of events included in both of them. (b) Waveforms at NU.KMD in 7:00–7:30 and (c) in 13:00–13:30 on 2017 June 25. Black portions of the waveforms and black bars at the top represent time sections identified as true events by the proposed method. Triangles at the top are origin times of earthquakes in the routine catalogue.

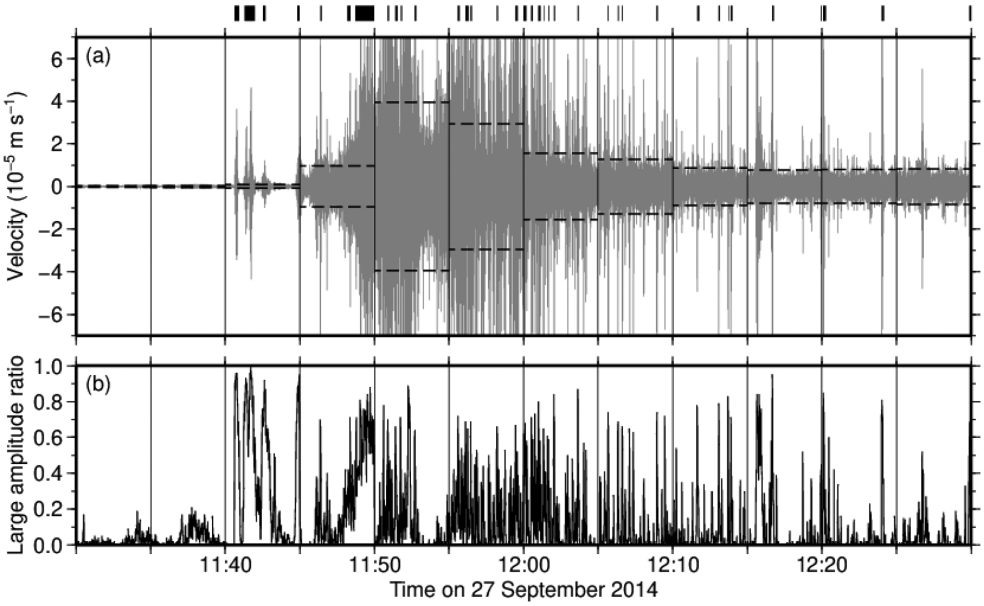

Figure 16. (a) Waveforms (grey), the estimated background noise level for every 5 min (black), and (b) large amplitude ratios around the onset of an eruption at 11:52:30 on 2014 September 27. Data from station V.ONTA are shown, which was nearest to the summit at that time. Bars on the top represent the detected candidate events.

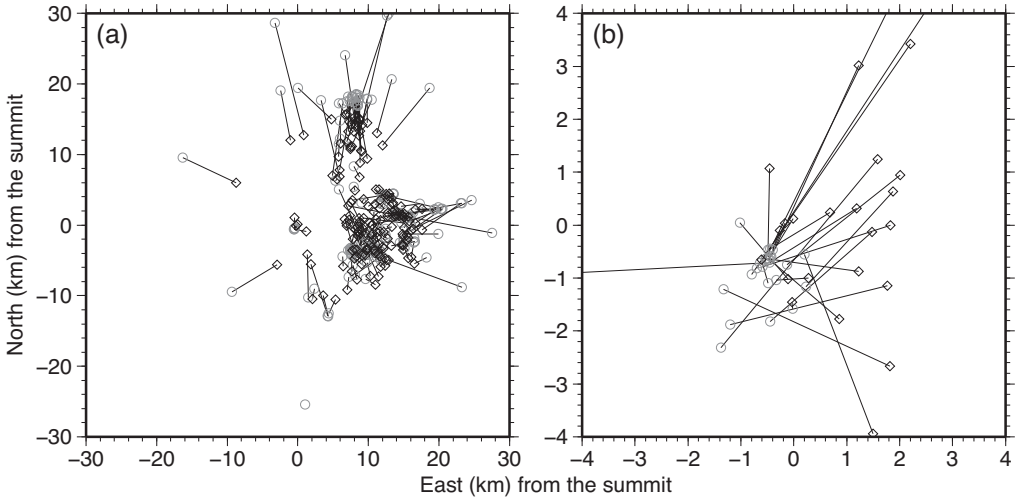

Figure 17. Comparison of the epicentres (circles) and weight centres of large amplitude probabilities (diamonds) for VT and local tectonic earthquakes in (a) routine and (b) summit catalogues.

the little dependence of the SVM model to station layouts, it is natural to expect that the SVM model could identify true and false events at different places. This potential transferability will be examined in the future.

There are three tasks needed to be addressed in the future. One is that the proposed method cannot detect an event that occurred before the end of the previous candidate event. This is because the entire network was used for a candidate event detection. This task may be solved by introducing an algorithm to divide stations into subnetworks. The second task is to classify the detected events at the summit region into VT events, LP events and tremors. This classification is not possible using amplitude alone but needs dominant frequencies and durations of the events. Previous studies (e.g. Valentine & Woodhouse 2010; Malfante *et al.* 2018) have shown a usefulness of these spectral features for the classification. Because this task could be done separately from the present work, we do not proceed to the classification in this study. The third task is to enable detection of events for which high-frequency (> 1 Hz) energies are absent. To do this, not only the frequency band but also the window length to compute large amplitude ratios and probabilities would need to be modified for that purpose.

6 CONCLUSIONS

We developed a new algorithm to automatically detect volcano seismic events from continuous waveform records. In the algorithm, candidate events are first detected based on signal-to-noise ratios, and then classified into true and false events using supervised machine learning. The

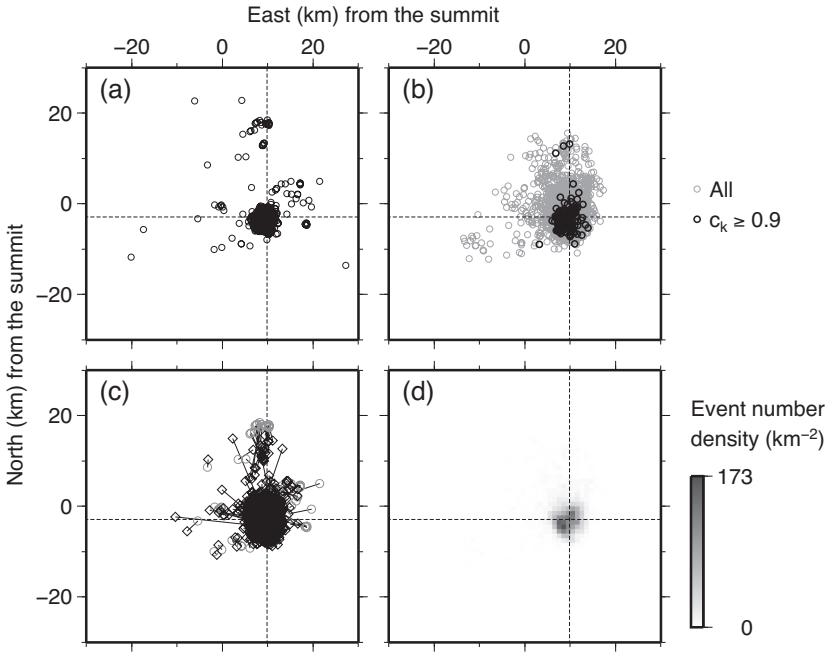

Figure 18. Comparison of (a) the epicentres of earthquakes in the routine catalogue and (b) the weight centres of large amplitude probabilities for all the detected events (grey) and the events with $c_k \geq 0.9$ (black) in 2017 June 24–27. (c) Comparisons of the epicentres (circles) and weight centres (diamonds). (d) Number density of the weight centres of detected events. Dashed lines represent the epicentre of an M5.6 earthquake on 2017 June 25.

data from Mt. Ontake, Japan, during the first 10 d of a dense observation trial in the summit region (2017 November 1–10) are used for the examination of our algorithm.

As the input data to the machine learning, we used the ratio of the number of time samples with amplitudes greater than the background noise level at 1 s intervals (the large amplitude ratio) at every station. The background noise level at each station was estimated by a threshold amplitude, below which the cumulative amplitude distribution of the data is well fitted by Gaussian noise. This estimate was more stable than an LTA in case of intense signals in the data. Candidate events were detected as the times when large amplitude ratios at five or more stations exceeded 0.3 simultaneously.

We used a two-step approach in machine learning. In the first step, the large amplitude ratios at station sites were fit by a neural network model to investigate a continuous spatial distribution of the probability that each location on the ground had an amplitude greater than the background noise level (a large amplitude probability). In the second step, we extracted eight features from the large amplitude probability at 1 s intervals. We used the first 3 s of each candidate event, resulting in 24 features. A relationship between the features and their classification as true or false events was investigated by a SVM using manual classifications as teaching data. The optimal SVM model for the Mt. Ontake data showed a classification accuracy of more than 97 per cent.

ACKNOWLEDGEMENTS

We thank Toshiko Terakawa for collaborating on the seismic monitoring at Nagoya University. The hypocentres in the routine catalogue in the study period were located by Junko Sumida and Eri Hibino. Takahiro Kunitomo, Kazushi Tanoue, Hiroshi Ichihara, Takashi Okuda and Keisokugiken Co. participated in installation of the summit trial seismic network. We used the continuous seismic records from stations operated by the Japan Meteorological Agency, the Nagano and Gifu Prefectures, and the NIED. We used a digital elevation model from the Geospatial Information Authority of Japan. We used the libSVM library (Chang & Lin 2011) for the SVM analysis. Comments by Andrew Valentine, Phillip Dawson and an anonymous reviewer helped to improve the manuscript. This work was supported by JSPS KAKENHI grant number 19K04016.

AUTHOR CONTRIBUTIONS

YM conducted the analyses and drafted the manuscript. YY, SH and YM designed the summit seismic observation trial. SH, TI and YM installed the summit stations used in the study period. YY and TI provided comments to improve the manuscript. Numerical data of the study results are available upon request to the corresponding author.

REFERENCES

Allen, R., 1982. Automatic phase pickers: their present use and future prospects, *Bull. seism. Soc. Am.,* **72**(6B), S225–S242.

Bergen, K.J. & Beroza, G.C., 2018. Detecting earthquakes over a seismic network using single-station similarity measures, *Geophys. J. Int.,* **213**(3), 1984–1998.

Brenguier, F., Shapiro, N.M., Campillo, M., Ferrazzini, V., Duputel, Z., Coutant, O. & Nercessian, A., 2008. Towards forecasting volcanic eruptions using seismic noise, *Nat. Geosci.,* **1,** 126–130.

Chang, C.-C. & Lin, C.-J., 2011. LIBSVM: a library for support vector machines, *ACM Trans. Intell. Syst. Technol.,* **2**(3), 27, doi:10.1145/1961189.1961199.

Chouet, B.A. & Matoza, R.S., 2013. A multi-decadal view of seismic methods for detecting precursors of magma movement and eruption, *J. Volc. Geotherm. Res.,* **252,** 108–175.

Chouet, B. *et al.*, 2003. Source mechanisms of explosions at Stromboli Volcano, Italy, determined from moment-tensor inversions of very-long-period data, *J. geophys. Res.,* **108**(B1), 2019, doi:10.1029/2002JB001919.

Curilem, G., Vergara, J., Fuentealba, G., Acuña, G. & Chacón, M., 2009. Classification of seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms, *J. Volc. Geotherm. Res.,* **180**(1), 1–8.

Endo, E.T. & Murray, T., 1991. Real-time seismic amplitude measurement (RSAM): a volcano monitoring and prediction tool, *Bull. Volcanol.,* **53**(7), 533–545.

Goodfellow, I., Bengio, Y. & Courville, A., 2016. *Deep Learning,* pp. 164–223, MIT Press, Cambridge, MA.

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

Horikawa, S. *et al.*, 2017., Development of portable seismic observation and telemetry equipment, *Abst. Volcanol. Soc. Jpn.,* P036 (in Japanese), doi:10.18940/vsj.2017.0_156.

Hsu, C.-W., Chang, C.-C. & Lin, C.-J., 2003. *A Practical Guide to Support Vector Classification, Technical report,* Department of Computer Science, National Taiwan University, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Killion, K., Kumar, R., Taylor, C.J. & Morra, G., 2018. Seismology and volcanology: exploration of volcanoes, long-periods, and machines - predicting volcano eruption using signature seismic data, *SMU Data Sci. Rev.,* **1**(1), article 11, https://scholar.smu.edu/datasciencereview/vol1/iss1/11.

Kunitomo, T., 2014. An improvement in the precision of measuring seismic traveltime changes with the use of the Hi-net data, *J. Seism. Soc. Jpn.,* **66**(4), 97–112.

Langer, H., Falsaperla, S., Powell, T. & Thompson, G., 2006. Automatic classification and a-posteriori analysis of seismic event identification at Soufrière Hills volcano, Montserrat, *J. Volc. Geotherm. Res.,* **153**(1–2), 1–10.

Langer, H., Falsaperla, S. & Thompson, G., 2003. Application of artificial neural networks for the classification of the seismic transients at Soufrière Hills volcano, Montserrat, *Geophys. Res. Lett.,* **30**(21), 2090, doi:10.1029/2003GL018082.

Li, Z., Meier, M.-A., Hauksson, E., Zhan, Z. & Andrews, J., 2018. Machine learning seismic wave discrimination: application to earthquake early warning, *Geophys. Res. Lett.,* **45**(10), 4773–4779.

Lyons, J.J. & Waite, G.P., 2011. Dynamics of explosive volcanism at Fuego volcano imaged with very long period seismicity, *J. geophys. Res.,* **116**(B9), B09303, doi:10.1029/2011JB008521.

Maeda, Y., Kato, A., Terakawa, T., Yamanaka, Y., Horikawa, S., Matsuhiro, K. & Okuda, T., 2015. Source mechanism of a VLP event immediately before the 2014 eruption of Mt. Ontake, Japan, *Earth Planets Space,* **67**(1), 187, doi:10.1186/s40623-015-0358-0.

Maeda, Y., Takeo, M. & Kazahaya, R., 2019. Comparison of high- and low-frequency signal sources for very-long-period seismic events at Asama volcano, Japan, *Geophys. J. Int.,* **217**(1), 389–404.

Maggi, A., Ferrazzini, V., Hibert, C., Beauducel, F., Boissier, P. & Amemoutou, A., 2017. Implementation of a multistation approach for automated event classification at Piton de la Fournaise Volcano, *Seismol. Res. Lett.,* **88**(3), 878–891.

Malfante, M., Dalla Mura, M., Mars, J.I., Metaxian, J.-P., Macedo, O. & Inza, A., 2018. Automatic classification of volcano seismic signatures, *J. geophys. Res. Solid Earth,* **123**(12), 10645–10658.

Nakamichi, H., Kumagai, H., Nakano, M., Okubo, M., Kimata, F., Ito, Y. & Obara, K., 2009. Source mechanism of a very-long-period event at Mt Ontake, central Japan: response of a hydrothermal system to magma intrusion beneath the summit, *J. Volc. Geotherm. Res.,* **187**(3–4), 167–177.

Nakano, M., Sugiyama, D., Hori, T., Kuwatani, T. & Tsuboi, S., 2019. Discrimination of seismic signals from earthquakes and tectonic tremor by applying a convolutional neural network to running spectral images, *Seismol. Res. Lett.,* **90**(2A), 530–538.

Oikawa, T. *et al.*, 2016. Reconstruction of the 2014 eruption sequence of Ontake Volcano from recorded images and interviews, *Earth Planets Space,* **68**(1), 79, doi:10.1186/s40623-016-0458-5.

Perol, T., Gharbi, M. & Denolle, M., 2018. Convolutional neural network for earthquake detection and location, *Sci. Adv.,* **4**(2), e1700578, doi:10.1126/sciadv.1700578.

Qu, S., Guan, Z., Verschuur, E. & Chen, Y., 2019. Automatic high-resolution micro seismic event detection via supervised machine learning, *Geophys. J. Int.,* **218**(3), 2106–2121.

Reynen, A. & Audet, P., 2017. Supervised machine learning on a network scale: application to seismic event classification and detection, *Geophys. J. Int.,* **210**(3), 1394–1409.

Scarpetta, S., Giudicepietro, F., Ezin, E.C., Petrosino, S., Del Pezzo, E., Martini, M. & Marinaro, M., 2005. Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks, *Bull. seism. Soc. Am.,* **95**(1), 185–196.

Valentine, A.P. & Woodhouse, J.H., 2010. Approaches to automated data selection for global seismic tomography, *Geophys. J. Int.,* **182**(2), 1001–1012.

Yamanaka, Y., Maeda, Y., Terakawa, T. & Horikawa, S., 2018. Seismic activity of Mt. Ontake volcano by multipoint observation test data at summit, *Abst. Jpn. Geosci. Uni.,* SVC41–19, https://confit.atlas.jp/guide/event/jpgu2018/subject/SVC41-19/advanced.

Yoon, C.E., O'Reilly, O., Bergen, K.J. & Beroza, G.C., 2015. Earthquake detection through computationally efficient similarity search, *Sci. Adv.,* **1**(11), e1501057, doi:10.1126/sciadv.1501057.

Zeiler, M.D., 2012. ADADELTA: an adaptive learning rate method, preprint (arXiv:1212.5701v1).

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

SupportingInformation.pdf

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

APPENDIX A. PARAMETER TUNING.

In the main text, only the results from the finally adopted parameters are shown. In this appendix, we describe more details of the parameter choice.

A.1 Computer environment used to measure computation times

Computation time was an important factor to decide parameters. In particular, the first step of machine learning (Section 3.3) needs to be repeated for all detection target time periods (Fig. 4), and thus, must be able to catch up with the progress of time. For example, the analysis of 1-hr data needs to be completed within 1 hr. Below, we show computation times measured by a Dell Precision T1700 workstation equipped with Intel(R) Xeon(R) CPU E3–1240 v3 (3.40) processors (4C/8T) and 16 GB memory. The computation codes were developed in C and compiled by the GNU GCC compiler with the maximum optimization (-O3) option. Parallel computation was not implemented for measurements of computation times. Most of the computation times are attributed to estimations of background noise levels (Section 3.2.2) and large amplitude probabilities (Section 3.3). The filtering (Section 3.2.1), computation of large amplitude ratios given the background noise levels (Section 3.2.3), candidate event detection (Section 3.2.4) and classification to true and false events given the optimal SVM model have negligible contributions to the total computation time.

A.2 Window length for background noise level estimation

The background noise level is estimated for every 5-min window at each station (i.e. $\bar{T} = 5$ min; Section 3.2.2). This choice is based on the computation time. The length $\bar{T}$ should be as great as possible to avoid overestimating the background noise level, which may occur when the entire time window is occupied by a long-lasting tremor (Fig. 16). However, the computation time is proportional to $\bar{T}^2$, meaning that if we double the window length, then the computation time would be four times greater. In our computer environment (Appendix A.1), a calculation of the background noise level for a 5-min window took 3 s for each trace, that is, 2 min for all 39 traces. This means that the calculation of the background noise level for $\bar{T} = 10$ min would take 8 min, which may be too tight. In case of $\bar{T} = 20$ min, the computation time would be 32 min, meaning that the process cannot catch up with the progress of time.

A.3 Criteria for detection of candidate events

A candidate event is defined as a continuous time period during which the large amplitude ratios (b_{ik}; eq. 14) of at least I^{th} stations exceed a threshold level (b^{th}) simultaneously, where we used $I^{\mathrm{th}} = 5$ and $b^{\mathrm{th}} = 0.3$ for our final choice (Section 3.2.4). We adopted these values based on the success detection ratios of known earthquakes. We searched the optimal I^{th} from 2 to 6 at an interval of 1 and b^{th} from 0.25 to 0.75 at an interval of 0.05. For each value, we calculated the success detection ratios of known earthquakes in the routine and summit catalogues (Section 2.2) and manually identified LP events (Section 2.3). Each earthquake in the routine and summit catalogues is regarded as detected if one of the candidate events starts from 1 s before to 5 s after the origin time of the earthquake, taking into account the start time temporal resolution of the candidate event (1 s) and the traveltime from the hypocentre to nearby stations. Each LP event is regarded as detected if a candidate event starts within $\pm$ 2 s of a manually determined onset time of the event at summit stations considering a relatively large uncertainty in the time. The manually identified tremors (Section 2.3) are not used for this evaluation because of poor onset time resolution.

Fig. A1 shows the success detection ratios of these events for all the combinations of Ith and b^{th}. The ratio was maximal in case of $I^{\mathrm{th}} = 5$ and $b^{\mathrm{th}} = 0.3$. Increasing I^{th} and b^{th} results in an increase in the events in the catalogues that do not meet the detection criterion (white bars in Fig. A1). Decreasing I^{th} and b^{th} causes local noise to be detected as candidate events, which obscures some events in the catalogues (grey bars in Fig. A1).

A.4 Numbers of intermediate layers and neurons in the neural network model: tests from ideal data

To perform a neural network analysis (Section 3.3), the numbers of intermediate layers (M) and neurons (Q_m) need to be set. The number of independent unknowns ($W_{qp}^{(m)}$) is given by

$$\sum_{m=0}^{M-1} (Q_m + 1)\, Q_{m+1} + (Q_M + 1)\,, \tag{A1}$$

which must be less than that of teaching data. In our case, the number of teaching data is the number of stations, which is at maximum 39 and may become smaller when some of the data are defective. Thus, the number of unknowns should be at most ~30, requiring relatively small values of M and Q_m. However, an overly simplified neural network model cannot express expected amplitude patterns. For example, a model with $M = 1$ and $Q_1 = 2$ can never express a closed region of large $P_k(x, y)$ regardless of the data given. To determine proper choices of M and Q_m, we conducted numerical tests using an ideal data set, which consists of a sufficiently large number of noise-free data points of a known pattern. Using the ideal data, we can focus on examining whether a neural network model is too simple.

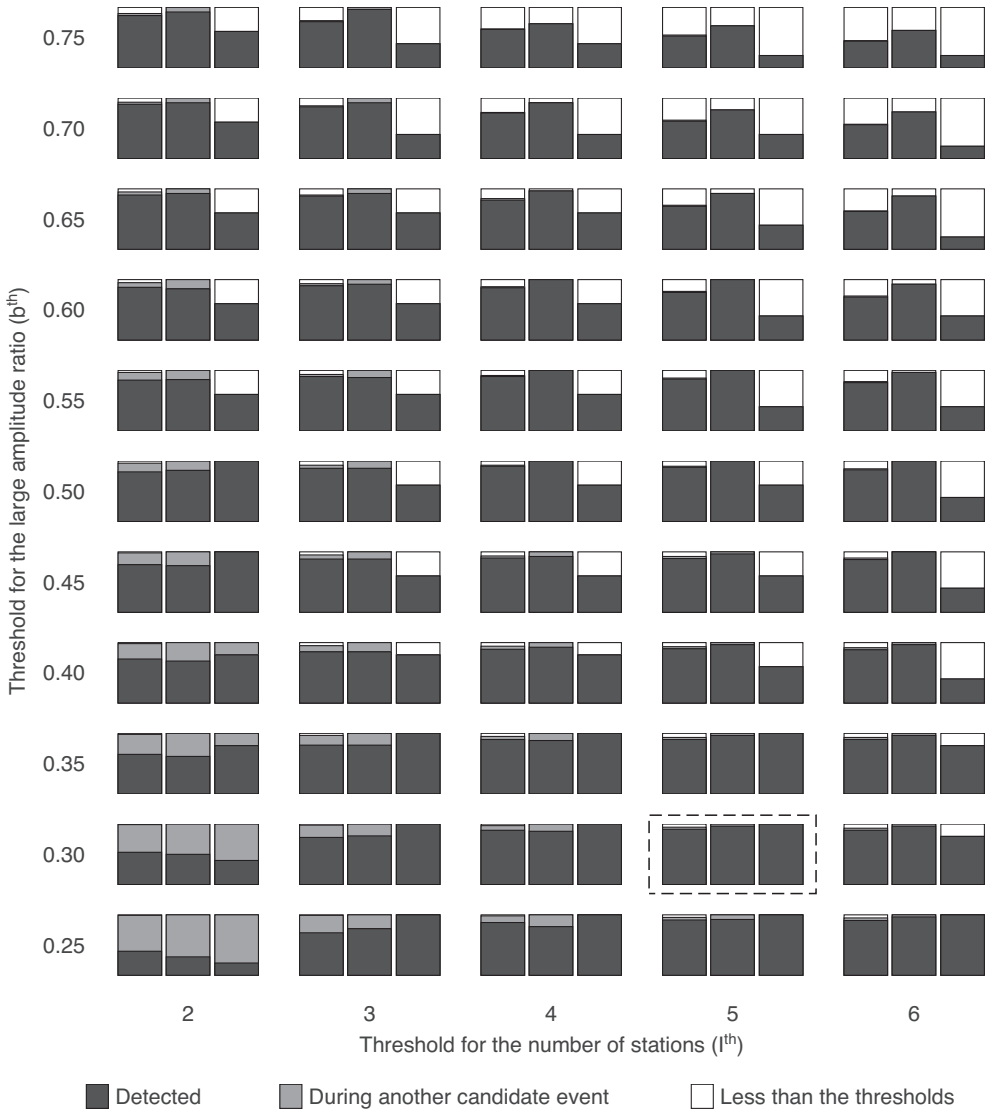

Figure A1. The ratios of the numbers of known earthquakes detected as candidate events, plotted against I^{th} and b^{th}. Results for earthquakes in the routine catalogue (left-hand bars), the summit catalogue (middle bars) and the list of manually detected LP events (right-hand bars) are shown. Black, grey and white bars represent the earthquakes detected as candidate events, not detected because they were hidden by other candidate events, and not detected because the detection criteria (I^{th} and b^{th}) were not met, respectively. The dashed square represents the best choices of Ith and b^{th}.

We prepared four ideal teaching data sets, each composed of 2000 samples randomly distributed in a $[-1,\ 1] \times [-1,\ 1]$ range of a 2-D plane (A2A). Each sample has an attribute of either 0 or 1 depending on location. Attribute 1 represents a large amplitude, whereas 0 represents a small one. In the first data set (Fig. A2aA), the attribute of each sample is 1 if the data point is in a circle of radius 0.7 centred on (0, 0). It simulates a situation where seismic amplitudes are recorded by 2000 stations and large amplitudes are observed by stations within the circle. In the second data set, the attribute boundary is given by an ellipsoid (Fig. A2bA). In the third data set, a doughnut pattern is used for the attribute boundary, which simulates a situation where seismic amplitudes are observed after wave propagation to some distance (Fig. A2cA). In the fourth data set, the attribute boundary is elongated in four directions, which approximates the 4-quadrant pattern of seismic amplitudes (Fig. A2dA). For each data set, we randomly divided the 2000 samples into 1000 for training and the remaining 1000 for testing.

We tried all the combinations of $0 \leq M \leq 3$ and $2 \leq Q_m \leq 7$ restricting the number of free parameters ($W_{qp}^{(m)}$) less than or equal to 30. We used a sigmoid function (eq. 20) for f_{M+1}, whereas for f_m with $m \leq M$, we examined the sigmoid, ReLU (eq. 21), and tanh (eq. 22) functions. For each combination of M, Q_m and f_m, we investigated $W_{qp}^{(m)}$ by minimizing the cross-entropy error for the training data. We investigated them iteratively, starting from 1000 sets of random initial values for which we examined both a uniform distribution within $[-1,\ 1]$ and a Gaussian distribution with an average of 0 and a standard deviation of 1. From each initial model, we updated $W_{qp}^{(m)}$ by 5000 iterations. The $W_{qp}^{(m)}$ values that minimized the cross-entropy error for the training data during these iterations were adopted as the final model. We evaluated the goodness of the final model by cross-entropy error for the test data. We repeated this evaluation for each combination of M, Q_m and f_m to determine the optimal choices of each variable.

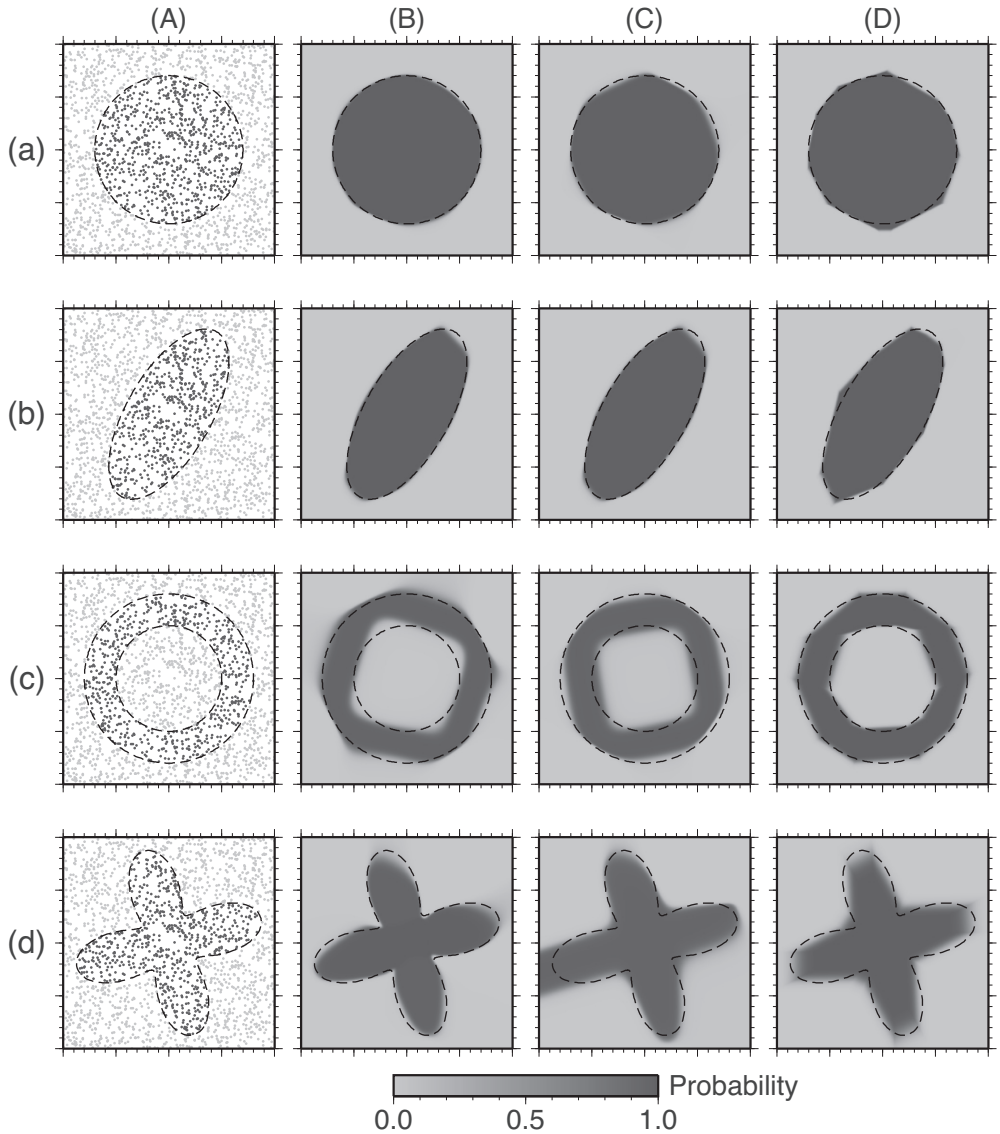

Figure A2. Tests using ideal data. (A) The data on a map, where dark and light grey dots represent the teaching data in groups 1 and 0 (virtual stations with large and small amplitudes), respectively. The group boundaries are shown by the dashed lines. (B)–(D) The estimated probability distributions for group 1 using configurations (1)–(3) in Table A2, respectively.

Table A1 summarizes the best neural network model configurations for the four ideal data sets (Fig. A2A). For all data, a model with $M = 2$ and $Q_2 = 2$ starting from Gaussian random values yielded the best performance. However, optimal choices for Q_1 and f_m depended on the data sets; $Q_1 = 5$ and $f_m = \tanh$ (configuration 1) for circular and 4-quadrant geometries, $Q_1 = 4$ and $f_m = \tanh$ (configuration 2) for an ellipsoid, and $Q_1 = 5$ and $f_m = \mathrm{ReLU}$ (configuration 3) for a doughnut (Table A1). Thus, the best choice is not uniquely determined. We consider configuration 1 to be better than 2 because the former yielded substantially smaller cross-entropy errors than the latter in circular and 4-quadrant cases, whereas the cross-entropy error differences were small in the other two cases (Table A2). Configuration 1 seems to be better than 3 because the latter yielded a smaller cross-entropy error only in one case (doughnut) among the four data sets (Table A2). As our analysis region of the real data is relatively narrow, the doughnut shape is not expected to occur frequently. Moreover, the angular geometries of $P_k(x, y)$ caused by the ReLU function (Fig. A2D) are unnatural. Therefore, we selected configuration 1 as the optimal choice. Using it, various teaching data geometries were well reproduced (Fig. A2B).

A.5 Dummy station intervals, the numbers of initial models and iterations: tests from typical real data

We applied the neural network model to the real data, with configuration 1 (Table A2), which was shown to be best by the tests using ideal data (Appendix A.4). We first used 5000 iterations from 1000 initial models (full estimate) to avoid falling to a local minimum solution. For this stage, we used a limited number of time frames, which exhibited typical spatial amplitude patterns (Fig. 8).

Table A1. Best configurations for the first machine-learning step obtained by numerical tests with four ideal data sets (Fig. A2A). The combination of initial values, f_m, M and Q_m that minimized test data cross-entropy error within each data set is shown. The last three rows represent the results for the test data.

Data set	Circle	Ellipsoid	Doughnut	4-quadrants
Initial values for $W_{qp}^{(m)}$	Gaussian	Gaussian	Gaussian	Gaussian
Function f_m ($m \leq M$)	tanh	tanh	ReLU	tanh
M	2	2	2	2
Q_1	5	4	5	5
Q_2	2	2	2	2
Cross-entropy error (E_k)	0.0155	0.0146	0.0681	0.0914
Misdetections	4	4	18	5
Missed detections	2	2	14	33

Table A2. Results of numerical tests with four ideal data sets (Fig. A2A) using the three candidate configurations given byTable A1; (1) $f_m = \tanh$, $M = 2$, $Q_1 = 5$ and $Q_2 = 2$; (2) $f_m = \tanh$, $M = 2$, $Q_1 = 4$ and $Q_2 = 2$ and (3) $f_m = \mathrm{ReLU}$, $M = 2$, $Q_1 = 5$ and $Q_2 = 2$ with Gaussian initial values.

Dataset	Circle	Ellipsoid	Doughnut	4-quadrants
Configuration (1)				
Cross-entropy error (E_k)	0.0155	0.0158	0.1461	0.0914
Misdetections	4	4	33	5
Missed detections	2	2	28	33
Configuration (2)				
Cross-entropy error (E_k)	0.0363	0.0146	0.1347	0.1447
Misdetections	1	4	23	33
Missed detections	12	2	36	18
Configuration (3)				
Cross-entropy error (E_k)	0.0381	0.0328	0.0681	0.1231
Misdetections	9	6	18	14
Missed detections	6	4	14	41

In Fig. A3(a), we compared the large amplitude ratio data given at station sites (circles) and a continuous distribution of large amplitude probabilities estimated from the data (the background gray scale) for the time frame of a typical LP event (Fig. 8a). The match between them was not good, with a region of high probabilities extending toward the eastern, southern and northwestern domain edges shown by grey background shades. We obtained more stable estimates of the large amplitude probabilities by adding dummy stations with zero values along the domain edges (Fig. A3b). The narrower the dummy station intervals, the more stable the results (Figs A3c–e). However, the computation time increases by increasing the number of dummy stations. Because we found no significant difference between the results from dummy station intervals of 2 km (Fig. A3d) and 1 km (Fig. A3e), we use the 2 km interval below.

The full estimate (5000 iterations from 1000 initial models) took approximately 900 s for the analysis of 1 s data. To save computation time, we reduced the numbers of initial models and iterations. Fig. A4(a) illustrates how the cross-entropy errors of the 1000 individual models of an LP event (Fig. 8a) improved during the iterations. They stagnate at several values (around 0.29, 0.25, 0.16 and 0.14) before reaching the final value around 0.10, suggesting that at least four local minima are present. These are indeed local minima, as indicated by large amplitude probabilities (Figs A5a–d), which are completely different from that for the best model (Fig. A3d). The probability for a model with a cross-entropy error around 0.10 (Fig. A5e) was similar to that for the best model. Fig. A4(b) shows cross-entropy errors of the 1000 models after the 5000 iterations sorted in ascending order. Approximately 90 per cent of the models resulted in almost the same cross-entropy errors as that of the best model. This ratio was 85 per cent after 2000 iterations (Fig. A4c) and 80 per cent after 1000 iterations (Fig. A4d). We made similar evaluations for the typical events in Figs 8(b)–(e). In the worst case, the ratio was 40 per cent after 1000 iterations. The probability that 1000 iterations from 20 initial models (brief estimate) would fall into a local minimum is thus $(1 - 0.4)^{20} = 3.7 \times 10^{-5}$ (0.0037 per cent). In Fig. A6, we compare the full and brief estimates of large amplitude probabilities for the seven typical events in Fig. 8. Both estimates showed similar patterns in case of earthquakes (Figs A6a–e). Some differences occurred in case of noise; nevertheless, true and false events can be distinguished by the results from the brief estimate (Figs A6f and g). We therefore conclude that the brief estimate using 1000 iterations from 20 initial models is satisfactory.

The processing time needed to perform the brief estimate for each 1 s of data was approximately 3.6 s. We need to investigate the large amplitude probabilities for the first 3 s of each candidate event. The maximum hourly occurrence rate of the candidate events was 62 per hour in the study period (Fig. 13a). This rate was 42 around the eruption in 2014 (Fig. 13b) and 118 around the M5.6 earthquake in 2017 (Fig. 13c). In the case of 118 candidate events/hour, or approximately 10 candidate events/5 min, a total of $3.6 \times 3 \times 10 = 108$ s is needed to estimate the large amplitude probabilities for the first 3 s of all the candidate events in a 5 min interval. The sum of this processing time and

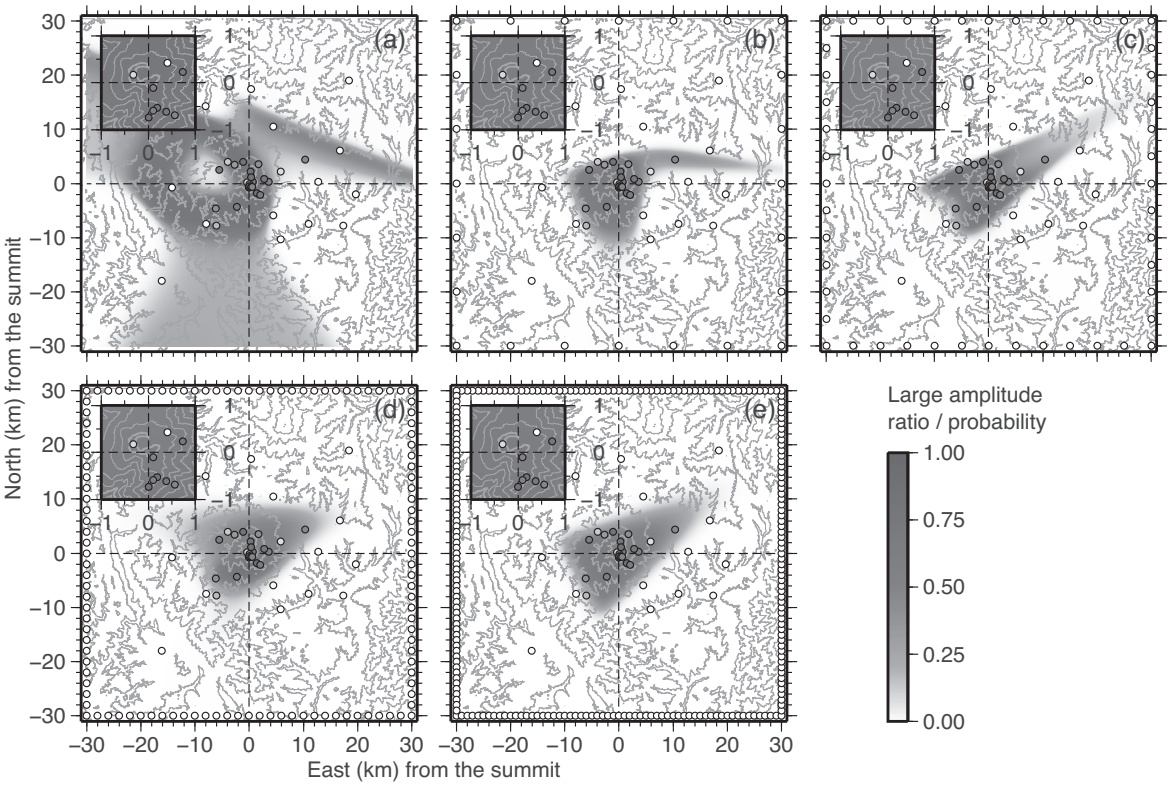

Figure A3. Comparison of the large amplitude ratios at station sites used as the teaching data (circles) and the large amplitude probabilities investigated by a neural network model (the background grey scale) for the data at 19:03:46 on 2017 November 1 (Fig. 8a). The plotting formats are same as those for Fig. 9. No dummy data were added in (a), whereas in (b)–(e), dummy stations with zero values were added on the edges at 10, 5, 2 and 1 km intervals. The main point of this figure is that 2 km intervals are satisfactory for the dummy stations.

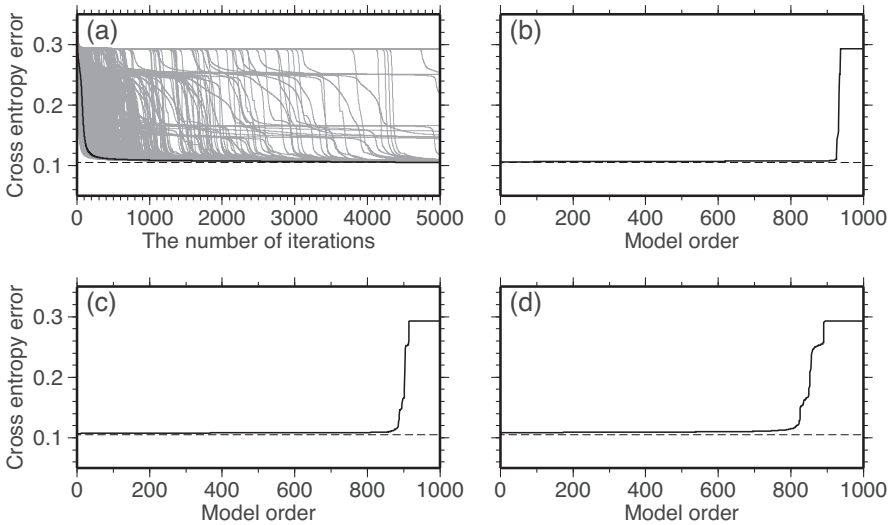

Figure A4. (a) Histories of cross-entropy errors during 5000 iterations starting from 1000 initial models (grey lines). The black line represents the model that finally reached the minimum cross-entropy error shown by the dashed line. (b)–(d) The cross-entropy errors from the 1000 models obtained after 5000, 2000 and 1000 iterations, sorted in ascending order. Results for the data at 19:03:46 on 2017 November 1 (Fig. 8a) are shown.

the 2 min needed to investigate the background noise level (Appendix A.2) does not exceed 5 min, suggesting that the analyses of continuous data can catch up with the progress of time. In principle, at most 50 candidate events can occur in 5 min if 1-s candidate events and 5-s rests are repeated in turn. In this case, the total processing time for the 5-min data is $3.6 \times 3 \times 50 + 120 = 660$ s (11 min), suggesting that the analysis can catch up with the progress of time if computations for different time periods are performed simultaneously by a processor with three cores.

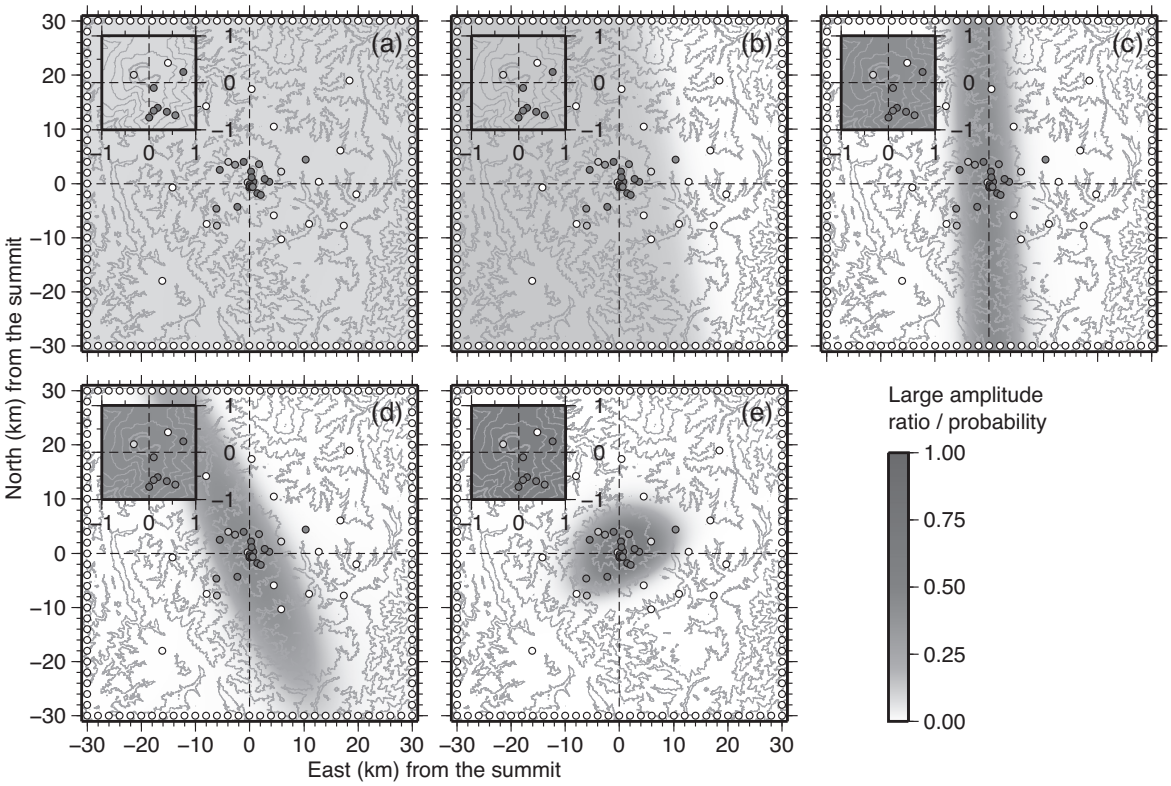

Figure A5. Comparison of large amplitude probabilities of local minimum solutions obtained after 1000 iterations for the data at 19:03:46 on 2017 November 1 (Fig. 8a). The plotting formats are same as those in Fig. 9. (a) The 891th best model with a cross-entropy error $E_k = 0.291$, (b) 890th best model ($E_k = 0.258$), (c) 841th best model ($E_k = 0.164$), (d) 826th best model ($E_k = 0.149$) and (e) 700th best model ($E_k = 0.111$). Note that the cross-entropy error for (e) is close to that of the best solution ($E_k = 0.108$). The main point of this figure is that the local minimum solutions (a–d) are indeed different from the global minimum (Fig. A3d).

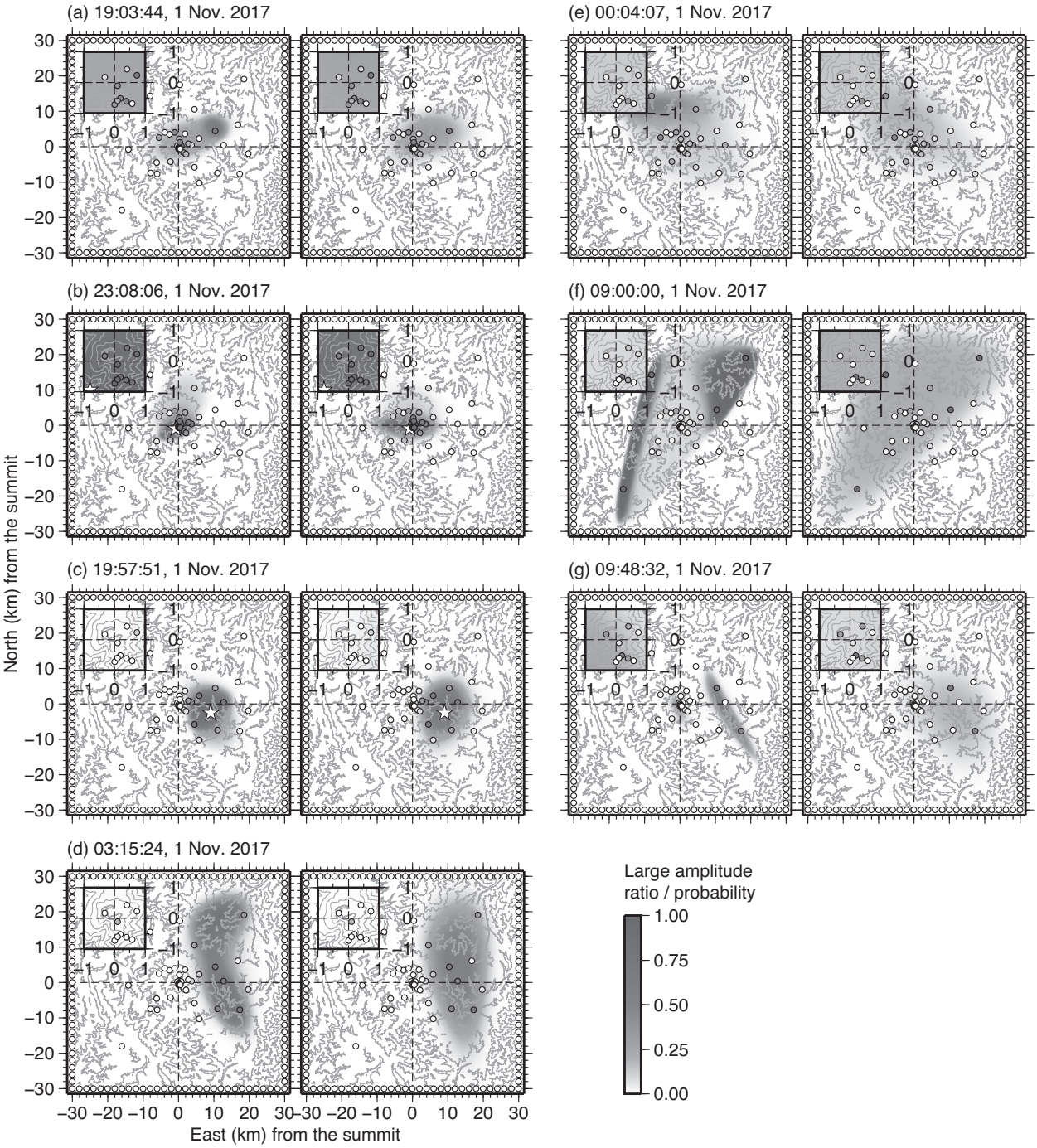

Figure A6. Comparison of the large amplitude ratios at station sites used as the teaching data (circles) and the large amplitude probabilities investigated by a neural network model (the background grey scale) for the typical seven events (Fig. 8). The plotting formats are same as those in Fig. 9. The left- and right-hand figures in each panel represent the results obtained after 5000 iterations from 1000 initial models (full estimate) and 1000 iterations from 20 initial models (brief estimate), respectively. The main point of this figure is that the brief estimate is satisfactory.