Supporting Information for "Near-real-time detection of co-seismic ionospheric disturbances using machine learning"

Quentin Brissaud[1] and Elvira Astafyeva[2]

[1]NORSAR, Kjeller, Norway
[2]Université de Paris, Institut de Physique du Globe de Paris (IPGP), CNRS UMR7154, 35-39 Rue Hélène Brion, 75013 Paris, France

Contents of this file

1. Texts S1 to S8
2. Figures S1 to S10
3. Table S1

Introduction

This Supplementary file contains additional details about:

- Text S1 List of input features
- Text S2 R2 cross correlations of input features and clustering analysis
- Text S3 Sensitivity of classification accuracy to number of validation points
- Text S4 Arrival time picking optimization
- Text S5 Time evolution of detected arrivals
- Text S6 Detection of CIDs at higher sampling rates
- Text S7 Impact of H_{ion} on association classes
- Text S8 Detection of ionospheric signal from volcanic eruptions and Rayleigh waves

S1 List of input features

All input features used to train the RF classifier presented in Section 3 are described in Table table S1.

The distribution of input features over our training and testing datasets is shown in Figure S1.

S2 R2 cross correlations of input features and clustering analysis

The RF model can provide an estimate of the relative feature importance through the calculation of the Gini's impurity during training. Figure 4b shows that the three best features have been extracted from the timeseries in contrast to other signal classification studies Wenner et al. (2021). However, the calculation of feature importance can be biased when considering continuous or high-cardinality categorical variables or when inputs features are co-linear. To assess the input features correlations within our CID dataset, we show in Figure S2 the R2 cross correlations. Co-linearity is present in our input dataset between spectral and time-series features which indicates a potential bias in variable importance results.

Table S1: List of attributes. $Nyf = 0.0165$ Hz is the Nyquist frequency. These attributes are commonly-used in signal-classification studies. We refer the reader to the following references for more details: (Bessason et al., 2007; Curilem et al., 2009; Hammer et al., 2012; Hibert et al., 2014; Provost et al., 2017; Wenner et al., 2021)

Short name	Description
W0	Ratio of the mean over the maximum of the envelope signal
W1	Ratio of the median over the maximum of the envelope signal
W2	Kurtosis of the raw signal (peakness of the signal)
W3	Kurtosis of the envelope
W4	Skewness of the raw signal
W5	Skewness of the envelope
W6	Number of peaks in the autocorrelation function
W7	Energy in the first third part of the autocorrelation function
W8	Energy in the remaining part of the autocorrelation function
W9	W7/W8
W10	Maximum of the envelope signal
W11	Energy of the signal filtered in 0.001-0.005 Hz
W12	Energy of the signal filtered in 0.005-0.015 Hz
W13	Kurtosis of the signal filtered in 0.001-0.005 Hz
W14	Kurtosis of the signal filtered in 0.005-0.015 Hz
S0	Mean of the Fourier transform (FT)
S1	Maximum of the FT
S2	Frequency at the FT maximum
S3	Frequency at the FT centroid
S4	Frequency at the FT 1st quartile
S5	Frequency at the FT 2nd quartile
S6	Median of the normalized FT
S7	Variance of the normalized FT
S8	Number of Fourier transform peaks (> 0.75 FT max.)
S9	Mean of FT peaks (S8)
S10	Gyration radius
S11	Energy up to 0.5Nyf Hz
S12	Energy up to 0.75Nyf Hz
S14	Energy up to 1.0Nyf Hz
FT0	Kurtosis of the maximum of all Fourier transforms (FTs) as a function of time
FT1	Kurtosis of the maximum of all FTs as a function of frequency
FT2	Mean ratio between the maximum and the mean of all FTs
FT3	Mean ratio between the maximum and the median of all FTs
FT4	Number of peaks in the curve showing the temporal evolution of the FTs maximum
FT5	Number of peaks in the curve showing the temporal evolution of the FTs mean
FT6	Number of peaks in the curve showing the temporal evolution of the FTs median
FT7	ratio of FT4 over FT5
FT8	ratio of FT4 over FT6
FT9	Number of peaks in the curve of the temporal evolution of the FTs central frequency
FT10	Number of peaks in the curve of the temporal evolution of the FTs maximum frequency
FT11	FT9/FT10
FT12	Mean distance between the curves of the temporal evolution of the FTs maximum and mean frequency
FT13	Mean distance between the curves of the temporal evolution of the FTs maximum and median frequency
FT14	Mean distance between the 1st quartile and the median of all FTs as a function of time
FT15	Mean distance between the 3rd quartile and the median of all FTs as a function of time
FT16	Mean distance between the 3rd quartile and the 1st quartile of all FTs as a function of time

Additionally, the significant overlap between distributions in Figure 4b motivates the choice of a large number of features to properly discriminate between each class. However, multidimensional clusters in input data are difficult to represent in 2d which prevents human interpretation. Two standard methods to facilitate the visualization of clusters in data are called Principal Component Analysis (PCA), and T-Distributed Stochastic Neighbor Embedding (TSNE, Van der Maaten and Hinton (2008)). PCA consists in project the input features in onto a space with orthogonal, i.e., uncorrelated, vector basis such that the greatest variance of the data comes to lie on the first coordinate. TSNE builds probability distributions over pairs of high-dimensional vectors so that similar data points have a higher probability while dissimilar data points are assigned a lower probability. Then, TSNE constructs a similar probability distribution over the points in the low-dimensional space, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map. No obvious clusters can be identified in the two first components of the PCA (Figure S3a) which indicates that only complex nonlinear relationships can help discriminating signals between noise and CID classes. TNSE non-linear mapping suggests two clusters (Figure S3b). These clusters are particularly visible for events showing strong CID amplitudes such as Sanriku. However, events with low signal-to-noise ratio CIDs, such as Kii, do not show a significant overlap between noise and arrival clusters. This further highlights the complexity of this classification problem.

S3 Sensitivity of classification accuracy to number of validation points

The heuristic model presented in Section 3.4 relies on a single parameter to confirm a detection: the number of consecutive time steps with a detection probability $> 50\%$, referred to as N_d. To determine the optimal value of N_d we varied this parameter between 2 and 5, and computed the true and false positive and negative rates over our true-arrival dataset, i.e., 2000 s waveforms centered on each true arrival. In Figure S4, we observe that the variations in N_d (Nb points trigger) do not affect significantly the true and false positive rates. Because we observe a slight decrease in False positive rate with an increase in N_d, we select $N_d = 3$ as a trade-off between false alerts and time delay to confirm a detection.

S4 Arrival time picking optimization

The arrival time picking procedure is based on a RF model. This model takes vTEC time derivatives as an input and gives a time shift from the window central time as an output. The RF will therefore be sensitive to the window size, as larger windows increase the number of inputs and tend to complicate the picking procedure while small time windows lack data points to regularize the time picking problem. Additionally, the range of window overlap with the true wavetrain used for training plays a significant role on the RF performances. Using small overlaps will train the machine to pick arrivals on incomplete waveforms and therefore makes the problem more difficult. However this will enable the machine to more efficiently pick arrival times over the first detection time windows of a given wavetrain. We show in Figure S5, the variations in time picking accuracy with window size and overlap (called deviation). As a trade-off between errors and the ability of our RF model to pick arrival times over incomplete waveforms, we choose a window size similar to the RF classifier (see Section 3.2) and an overlap of 30%.

S5 Time evolution of detected arrivals

A requirement for NRT applications is to obtain alerts within 20mn after the event. Therefore, our detection and association procedure should trigger a valid alert as soon as possible in addition to providing accurate arrival times. In Figure S6, we show the evolution of the distribution of arrival times with time since the event for the earthquake Tohoku. We observe that after 12mn, we already observe a specific trend in arrival-time values highlighting that the acoustic energy is propagating from East to West. After 15mn, almost all hand-picked arrival times have been correctly determined by our model.

S6 Detection of CIDs at higher sampling rates

A machine learning model trained with data sampled at 30s might learn patterns that are invariant with frequency. To assess how our classification model performs on 1s data, we extracted features in each time window without downsampling waveforms. In addition, we used a 1s time shift between two consecutive time windows. Detection probability and picked arrival times are shown in Figure S7. Detection probabilities are always over 50%, i.e., RF classified the whole timeseries as a CID with the use of a detection threshold at 50%. Yet, we observe a significant increase in detection probability around 2.85 UT, from 60% to 95%, that matches the arrival of the CID. Jumps in detection probabilities indicates that using a larger detection threshold, such as $\geq 70\%$ instead of $\geq 50\%$, could enable the processing of higher sampling-rate data with of our algorithm. These larger probabilities owe to the additional noise introduced by higher frequencies when extracting input features. The higher-frequency spectral content can lead to substantial variations in certain input features. For example, energy peaks at higher frequencies, that would normally be smoothed out at lower frequencies, can drastically alter the envelope kurtosis and skewness, which are critical parameters for discrimination between noise and arrival windows. Nonetheless, the ability of our model to recover the true arrival time is extremely promising for near-real-time applications.

S7 Impact of $H_{\mathbf{ion}}$ on association classes

The position of ionospheric detection points is dependent on the altitude of detection H_{ion}, which could impact the association classes. To assess the sensitivity of the association classes on H_{ion}, we changed the altitude of the ionospheric points for the Tohoku event from 180 to 250 km. The location of the center of the main association class (light purple in Figure S8c) tends to shift towards the South-East with the increase in H_{ion}. While the location of the ionospheric points changes with H_{ion}, the true arrival times (Figure S8a) are still correctly associated in the same class (light purple in Figure S8c).

S8 Detection of ionospheric signal from volcanic eruptions and Rayleigh waves

Other low-frequency acoustic sources, such as volcanoes or surface Rayleigh waves can generate transient ionospheric perturbations. In particular, volcanic eruptions generate both infrasonic and gravito-acoustic signals in the 0.1-10 mHz frequency range known as Co-Volcanic Ionospheric Disturbances (CVID). While gravity waves show a much lower frequency content Hines (1960), near-epicentral CVID can show short-period signals with significant energy below 5 minutes Shestakov et al. (2021). We therefore first assessed the sensitivity of our RF model to travelling volcanic-induced ionospheric propagation using the example of the Calbuco volcanic eruption on April 22, 2015 Shults et al. (2016). In figure S9, we observe that the entire volcanic-induced gravito-acoustic wavetrain is classified as CID. This can be explained by the similarity of CIDs and CVIDs in the feature space due to significant energy at high frequencies corresponding to infrasound signals mixed with the graviy wavefield.

The atmospheric perturbations generated by seismic Rayleigh waves can also propagate to the ionosphere and be observed on TEC data (Rolland et al., 2011). Such signals typically show energy between XXX s and XXX s, similar to epicentral infrasound. Testing our method on a Rayleigh-wave signal observed after the XXX event, we observe that the transient signal is well captured and its arrival time accurately predicted (see Figure S10). This indicates that both epicentral and Rayleigh-wave infrasound can be observed and associated by our detection method.

References

E. Astafyeva, K. Heki, E. Afraimovich, V. Kiryushkin, and S. Shalimov. Two-mode long-distance propagation of coseismic ionosphere disturbances. *J. Geophys. Res.*, 118:A10307, 2009. doi: 10.1029/2008JA013853.

B. Bessason, G. Eiríksson, Ó. Thórarinsson, A. Thórarinsson, and S. Einarsson. Automatic detection of avalanches and debris flows by seismic methods. *Journal of Glaciology*, 53(182):461–472, 2007. doi: 10.3189/002214307783258468.

G. Curilem, J. Vergara, G. Fuentealba, G. Acuña, and M. Chacón. Classification of seismic signals at villarrica volcano (chile) using neural networks and genetic algorithms. *Journal of volcanology and geothermal research*, 180(1):1–8, 2009. doi: 10.1016/j.jvolgeores.2008.12.002.

C. Hammer, M. Beyreuther, and M. Ohrnberger. A Seismic-Event Spotting System for Volcano Fast-Response Systems. *Bulletin of the Seismological Society of America*, 102(3):948–960, 06 2012. ISSN 0037-1106. doi: 10.1785/0120110167.

C. Hibert, A. Mangeney, G. Grandjean, C. Baillard, D. Rivet, N. M. Shapiro, C. Satriano, A. Maggi, P. Boissier, V. Ferrazzini, and W. Crawford. Automated identification, location, and volume estimation of rockfalls at piton de la fournaise volcano. *Journal of Geophysical Research: Earth Surface*, 119(5): 1082–1105, 2014. doi: https://doi.org/10.1002/2013JF002970.

C. O. Hines. Internal atmospheric gravity waves at ionospheric heights. *Canadian Journal of Physics*, 38 (11):1441–1481, 1960.

F. Provost, C. Hibert, and J.-P. Malet. Automatic classification of endogenous landslide seismicity using the random forest supervised classifier. *Geophysical Research Letters*, 44(1):113–120, 2017. doi: 10.1002/2016GL070709.

L. M. Rolland, P. Lognonné, and H. Munekane. Detection and modeling of Rayleigh wave induced patterns in the ionosphere. *J. Geophys. Res. Space Phys.*, 116(A5), 2011. doi: 10.1111/j.1651-2227.2005.tb01817.x.

N. Shestakov, A. Orlyakovskiy, N. Perevalova, N. Titkov, D. Chebrov, M. Ohzono, and H. Takahashi. Investigation of ionospheric response to june 2009 sarychev peak volcano eruption. *Remote Sensing*, 13 (4):638, 2021. doi: 10.3390/rs13040638.

K. Shults, E. Astafyeva, and S. Adourian. Ionospheric detection and localization of volcano eruptions on the example of the april 2015 calbuco events. *Journal of Geophysical Research: Space Physics*, 121(10): 10,303–10,315, 2016. doi: 10.1002/2016JA023382.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. URL `https://jmlr.csail.mit.edu/papers/volume9/vandermaaten08a/vandermaaten08a.pdf`.

M. Wenner, C. Hibert, A. van Herwijnen, L. Meier, and F. Walter. Near-real-time automated classification of seismic signals of slope failures with continuous random forests. *Natural Hazards and Earth System Sciences*, 21(1):339–361, 2021. doi: 10.5194/nhess-21-339-2021.

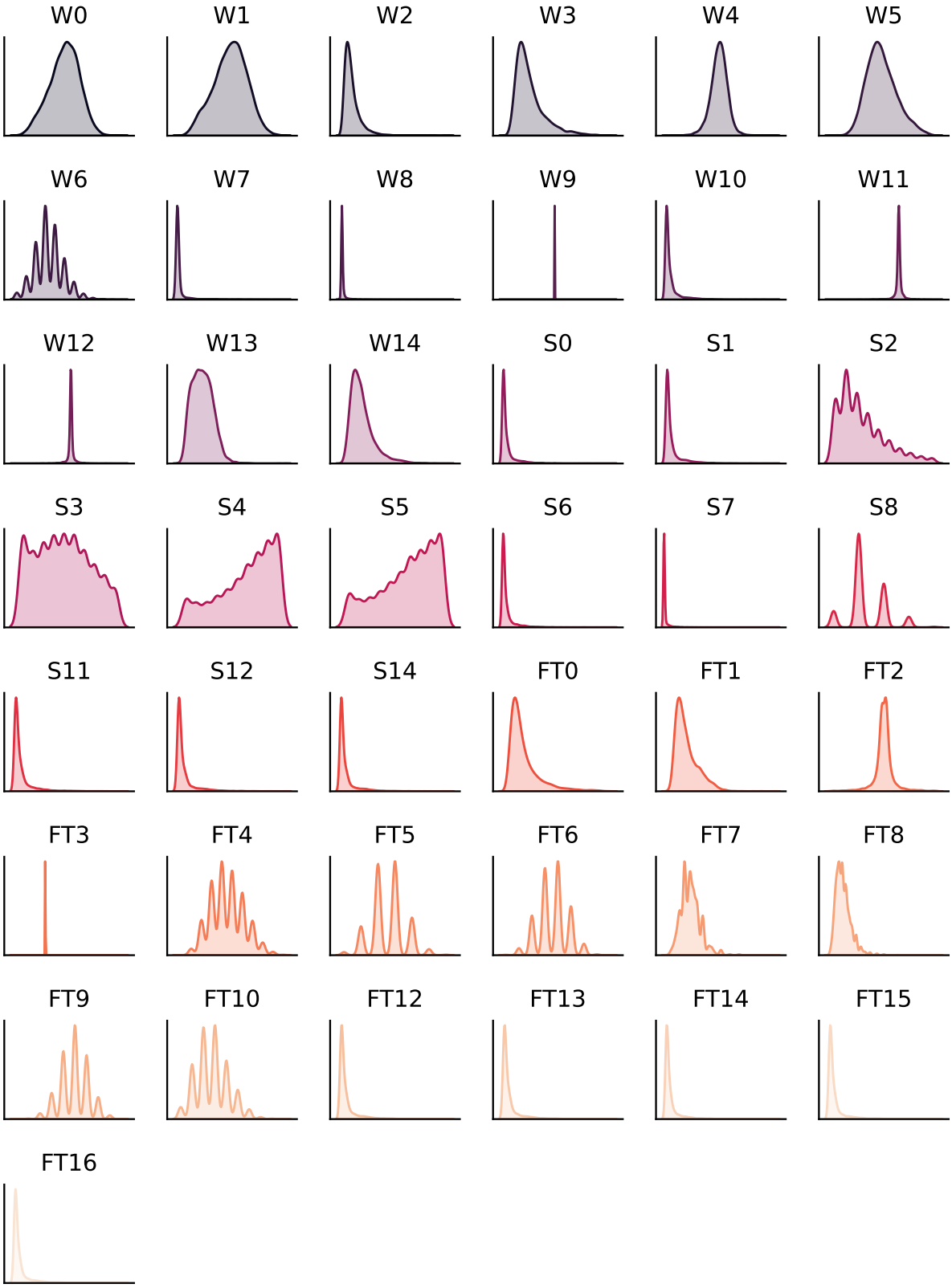

Figure S1: Probability density of each input features over our training and testing datasets. The short name of feature for each plot is shown above the plot. The description of each feature is given in Table table S1

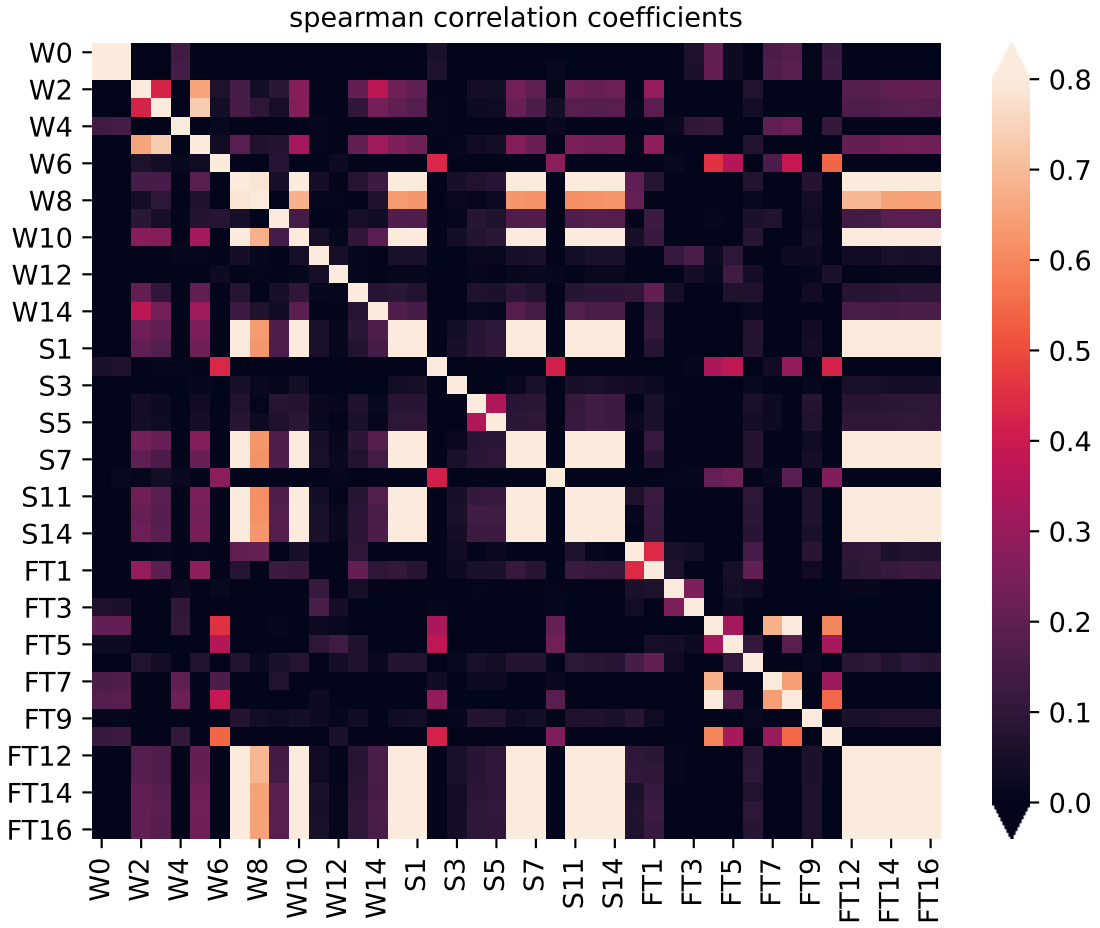

Figure S2: Spearman's correlation coefficients between each feature used for training. A description of each feature is given in Table table S1.

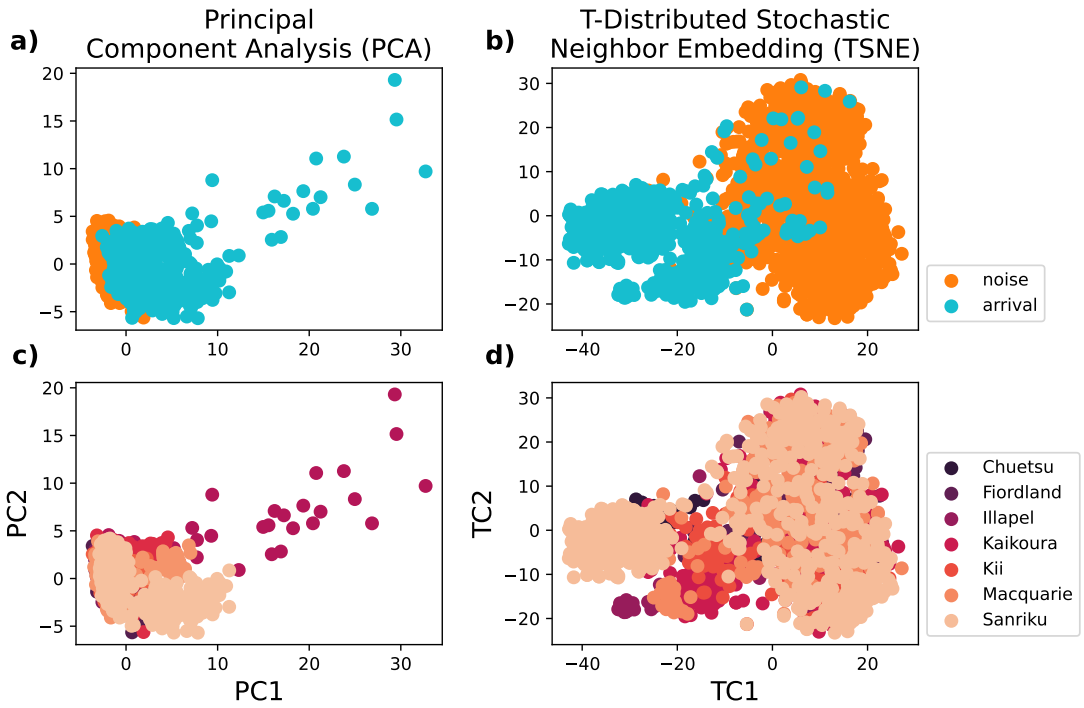

Figure S3: First versus second component of (a,c) a Principal Component Analysis (PCA) and (b,d) a T-Distributed Stochastic Neighbor Embedding (TNSE, Van der Maaten and Hinton (2008)). Points are colorcoded with (a,b) the detection class, and (c,d) the event name for the arrival class.

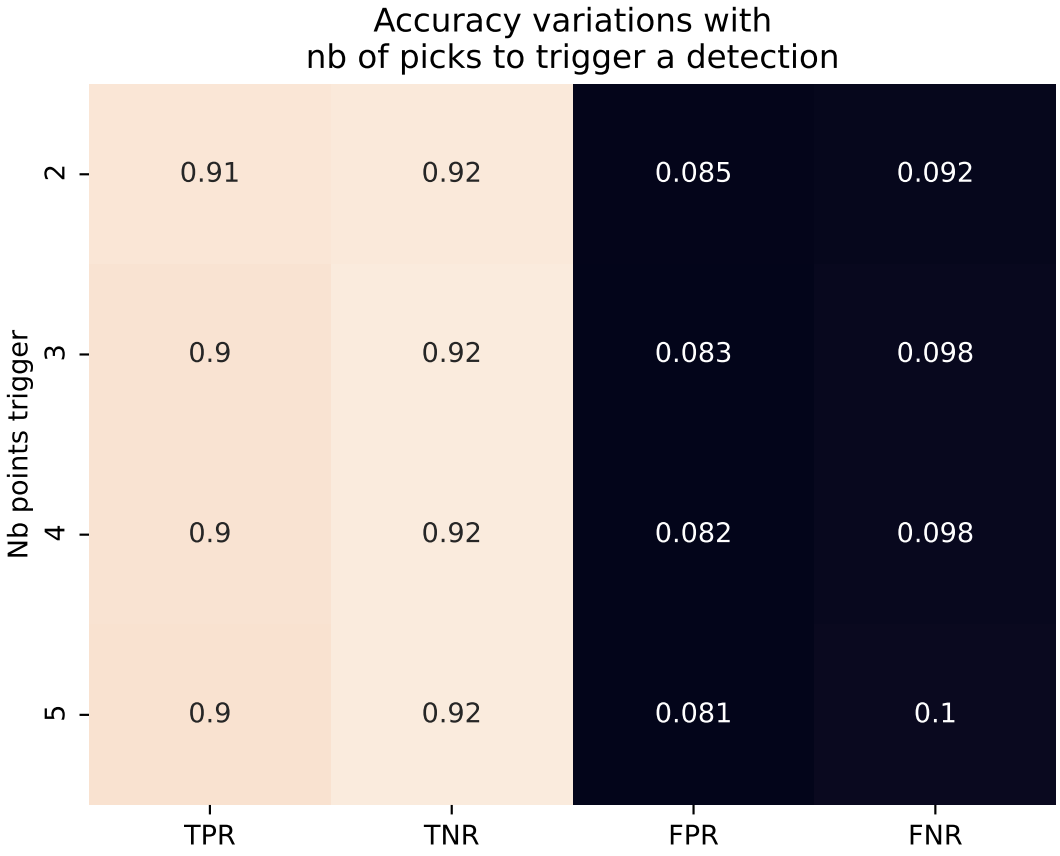

Figure S4: True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) with the choice of number of time steps for validation in the heuristic model presented in Section 3.4

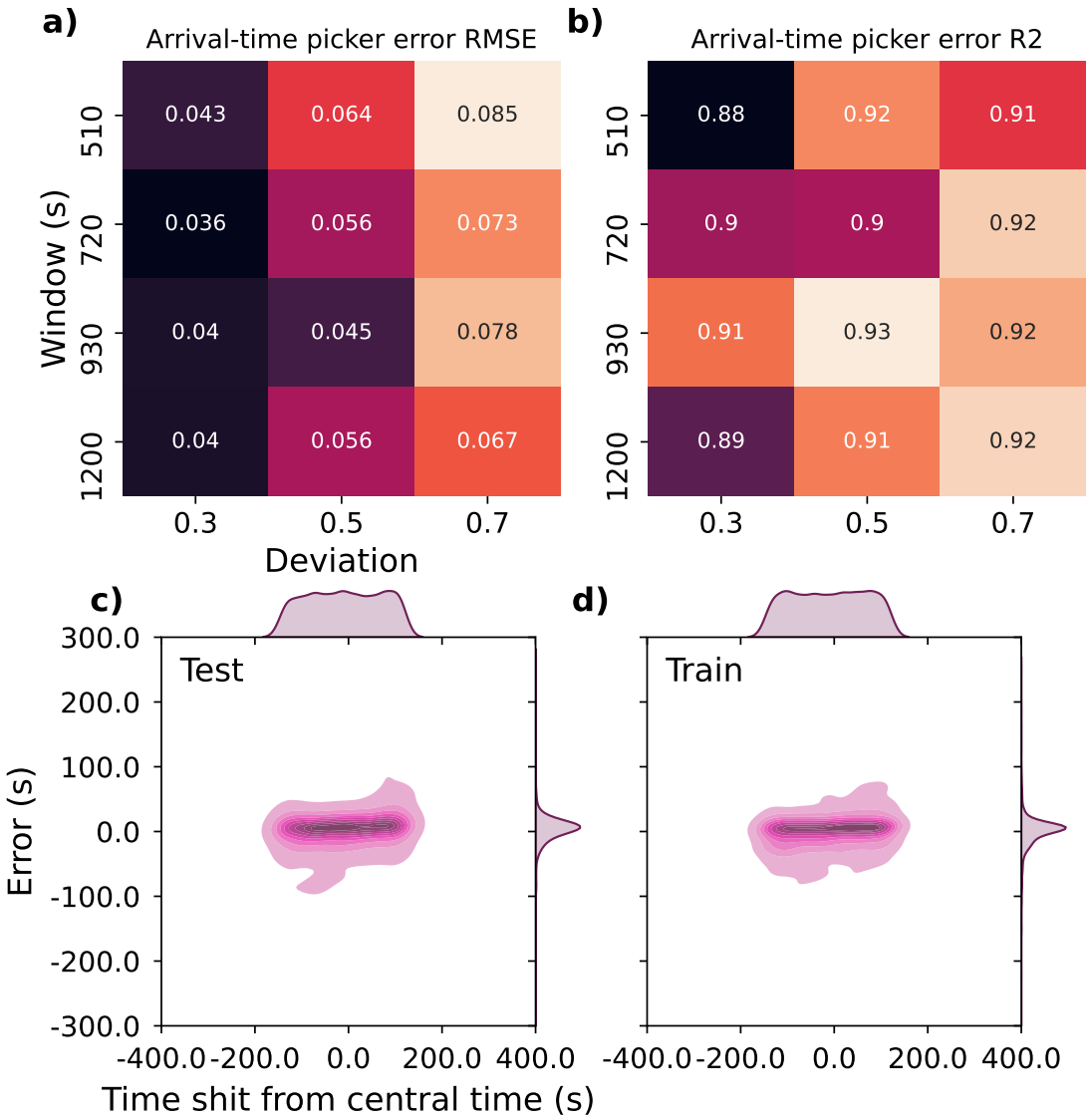

Figure S5: Performance of RF arrival time picker. (a) Root Mean Square Error (RMSE) vs minimum true-wavetrain overlap (deviation) and window size (s). The minimum true-wavetrain overlap corresponds to the minimum fraction of the wavetrain that has to be included in a window to be considered for training. (b) R2 error vs minimum true-wavetrain overlap (deviation) and window size (s). Bottom Distribution of arrival-time picking errors (s) vs true time shift from central time (s) over (c) the testing dataset, and (d) the training dataset.

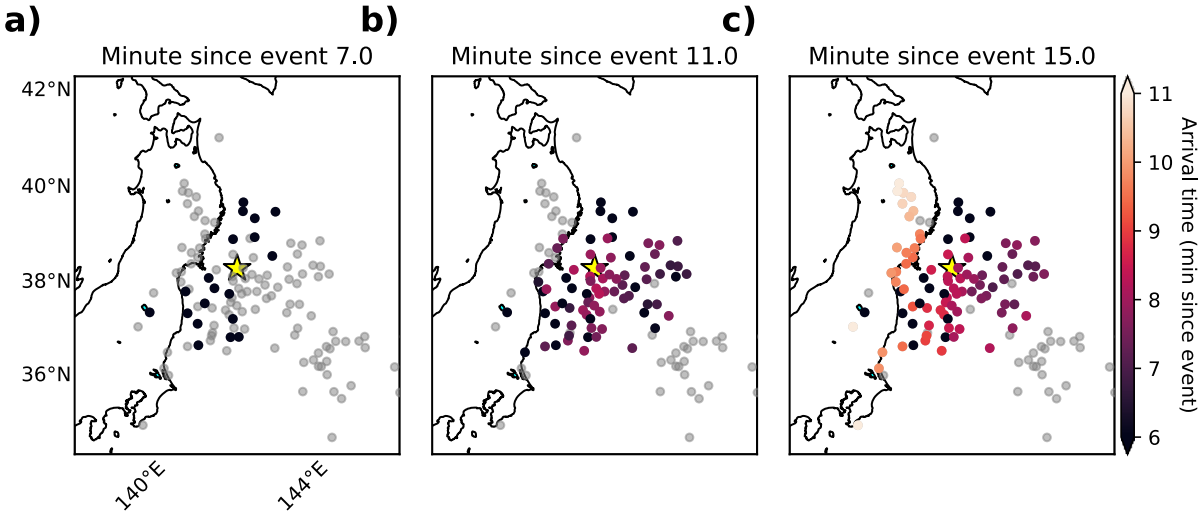

Figure S6: Ionospheric maps after the 2011 Tohoku earthquake generated at various times since the event. (a) to (c) Distribution of detected arrival times after (a) 7 minutes, (b) 11 minutes, and (c) 15 minutes since the event. CID coordinates were calculated at the intersection point between the LOS and the ionospheric layer using $H_{\text{ion}} = 250$ km. The colorcode corresponds to the predicted arrival time at each ionospheric point. Grey dots correspond to the location of ionospheric points where there is no detection yet but with detections after 20 mn.

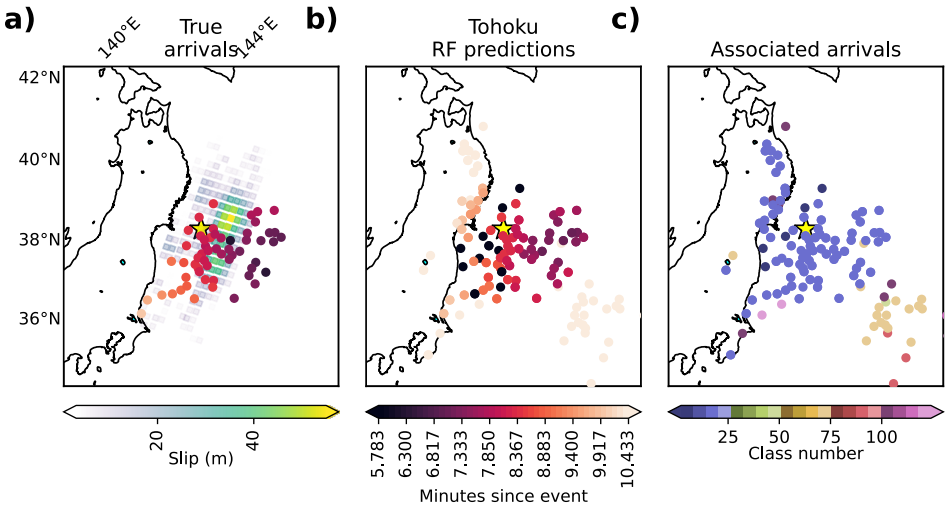

Figure S7: Tohoku's ionospheric arrival-time maps computed 14 minutes after the event for (d) hand-picked arrival times along with the epicenter location (yellow star), and surface projection of the fault slip (in m) as green to yellow patches, (e) RF-based arrival-time predictions, and (f) association classes determined from predicted arrival times.. CID coordinates were calculated at the intersection point between the LOS and the ionospheric layer using $H_{\text{ion}} = 180$ km.

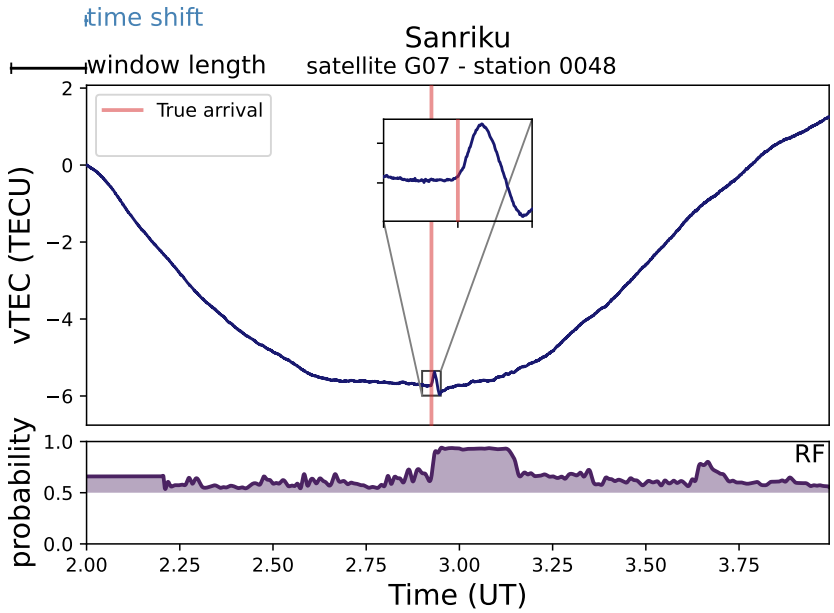

Figure S8: Performance assessment of RF detection and arrival-time picking at a higher sampling rate of 1s. 2-h vTEC waveform for the Sanriku event, satellite G07, station 0048 along with detection probabilities predicted by our RF detection model (bottom). The true arrival is shown as a red vertical line.

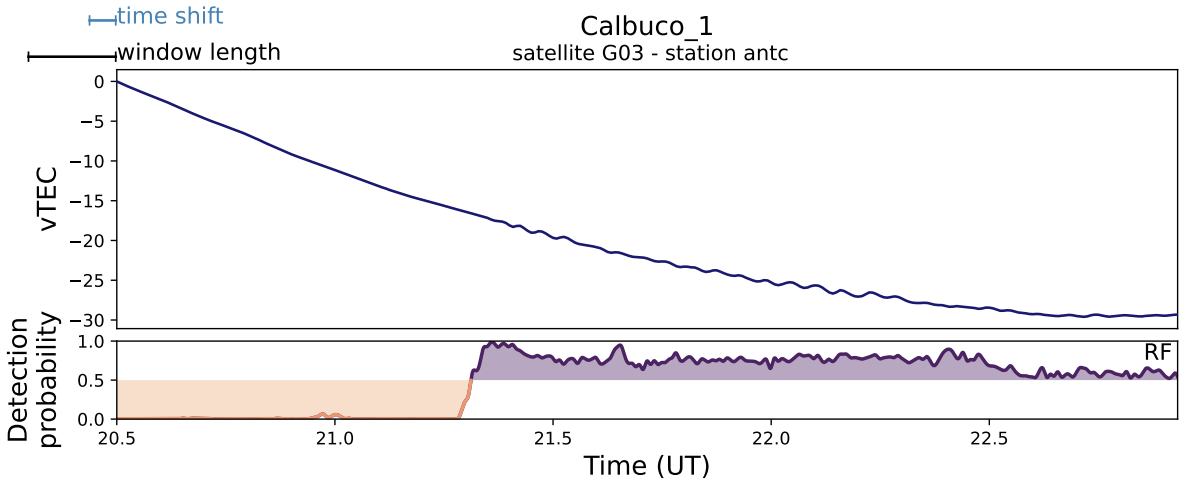

Figure S9: vTEC waveform for the Calbuco eruption, satellite G03, station antc along with detection probabilities predicted by our detection procedure (see Section 3) using a window size $w = 720$ s. Volcano-associated ionospheric perturbations are present between 21.3 and 22.5UT. The RF-predicted arrival time as a dark grey vertical line. The detected wavetrain using the RF is highlighted with a grey background.

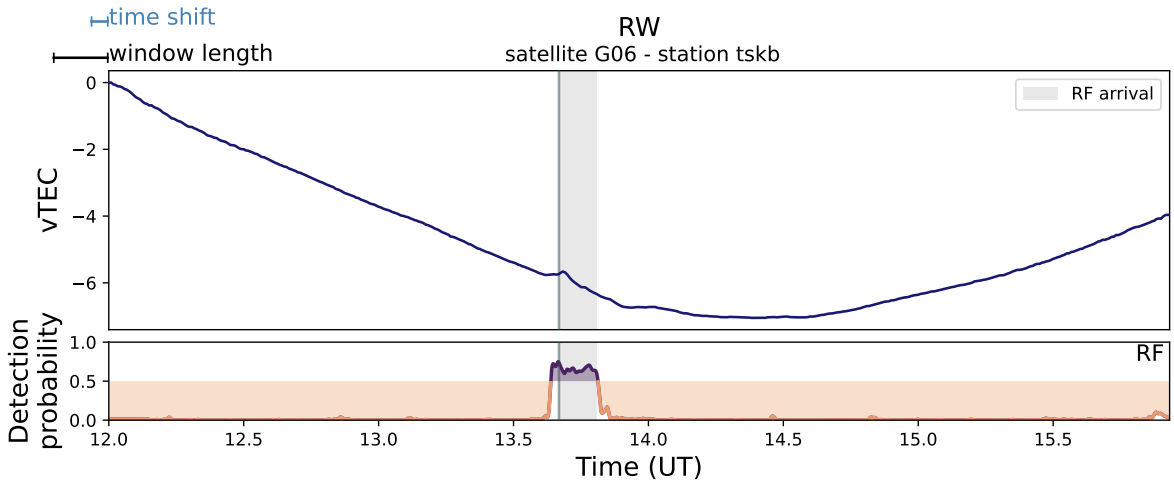

Figure S10: vTEC waveform from seismic Rayleigh waves recorded after the 1994 earthquake in Kuril Islands (Astafyeva et al., 2009), satellite G06, station tskb along with detection probabilities predicted by our detection procedure using a window size $w = 720$ s. Rayleigh-wave-associated ionospheric perturbations are present between 13.6UT and 13.8UT. The RF-predicted arrival time as a dark grey vertical line. The detected wavetrain using the RF is highlighted with a grey background.