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

Introduction

This Supplementary file contains additional details about:

S1 List of events

The list of events compiled in our CID dataset is described in Table S1.

Table S1: List of events included in the dataset. Events are sorted by magnitude

Event Reference	Mag.	Lat. ; Lon.	Date (DD/MM/YY)	Time (UTC)	Signal duration (s)	Sat.	Samp
Tohoku Astafyeva et al. (2011, 2013a)	9.1	38.3 ; 142.37	11/03/2011	05:46:23	800	G26 G05	1s, 30s
Sumatra 1 Astafyeva et al. (2014)	8.6	2.35 ; 92.8	11/04/2012	08:38:37	300	G32	15s
Tokachi Heki and Ping (2005)	8.3	41.78 ; 143.90	25/09/2003	19:50:06	440	G13 G24	30s
Illapel Bagiya et al. (2019)	8.3	-31.57; -71.61	16/09/2015	22:54:32	600	G25,G12 G24	15s, 30s
Sumatra 2 Astafyeva et al. (2014)	8.2	0.90 ; 92.31	11/04/2012	10:43:09	300	G32	15s
Iquique Bagiya et al. (2019)	8.2	-19.61 ; -70.77	01/04/2014	23:46:47	700	G01,G20 G23	15s, 30s
Macquarie Astafyeva et al. (2014)	8.1	-49.91 ; 161.25	23/12/2004	14:59:03	550	G05	30s
Fiorland Astafyeva et al. (2013b)	7.8	-45.75 ; 166.58	15/07/2009	09:22:29	300	G20	30s
Kaikoura Bagiya et al. (2018)	7.8	42.757 ; 173.077	13/11/2016	11:02:56	550	G20 G29	1s, 30s
Sanriku Thomas et al. (2018); Astafyeva and Shults (2019)	7.3	38.44 ; 142.84	09/03/2011	02:45:20	200	G07 G10	1s, 30s
Kii Heki and Ping (2005)	7.2	33.1 ; 136.6	05/09/2004	10:07:07	425	G15	30s
Chuetsu Cahyadi and Heki (2015)	6.6	37.54 ; 138.45	16/07/2007	01:12:22	300	G26	30s

Table S2: Slope parameters for different sampling rates used by the analytical detector presented in Section S2.

Sampling (s)	s_1 (TECU/epoch)	s_2 (TECU/epoch)	s_3 (TECU/epoch)	s_4 (TECU/epoch)
1	0.017	0.027	0.045	0.05
15	0.08	0.125	0.12	-
30	0.11	0.18	-	-

S2 Analytical detectors

To further assess the RF classification performance, we compare the results to two analytical detection methods: 1) a STA/LTA detection method, and 2) a derivative-based threshold method that has recently been developed. Both methods require an extensive hyper-parameter optimization in order to reduce the increase both true and false positive rates over a specific dataset. This further highlights the benefits of using an optimized RF to classify CIDs. We compare the three methods over a waveform dataset consisting of windows of length 2000 s and centered on each true arrival.

The STA/LTA method requires to set four parameters: the STA and LTA time windows and two thresholds to activate and deactivate the detection trigger. The STA window should represent the average duration of expected earthquake signals while the LTA window should capture the average TEC noise amplitude. The STA/LTA method employed here uses a 60 s STA window and a 400 s LTA window. A detection is triggered if the STA/LTA threshold reaches 2.5 while the end of a wavetrain is chosen where the threshold goes below 0.5. This trigger value of 2.5, lower than typically used by ground seismic station, is used to make sure to capture each arrival, i.e., increase the true positive rate, since small earthquakes generally show low signal-to-noise ratio. This choice of parameters is purely empirical and could be improved with a thorough investigation of the STA/LTA accuracy over the whole dataset. However, this is not straightforward since spurious arrivals can show a similar frequency content than CIDs.

The analytical method used for comparison, referred to as "AN", is based on the analysis of TEC rate-of-change. Maletckii and Astafyeva (2021, Determining spatio-temporal characteristics of Coseismic Travelling Ionospheric Disturbances (CTID) in near-real-time, Submitted to Sci. Reports) noticed that, in a majority of cases, the CID are characterized by a rapid and high increase of TEC. To capture the CID arrival we therefore suggest to analyze the rate of TEC change between the two consecutive epochs, between every two and every three epochs:

$$\partial vTEC_1 = |vTEC_i - vTEC_{i+1}|, \quad (1)$$
$$\partial vTEC_2 = |vTEC_i - vTEC_{i+2}|, \quad (2)$$
$$\partial vTEC_3 = |vTEC_i - vTEC_{i+3}|, \quad (3)$$
$$\partial vTEC_4 = |vTEC_i - vTEC_{i+4}|, \quad (4)$$

where the subscript i corresponds to the time step t_i. The vTEC at epoch i is considered as the CID arrival if each slope $\partial vTEC_1, \partial vTEC_2$, and $\partial vTEC_3$ (and $\partial vTEC_4$ for 1s data) are greater that the thresholds shown in Table S2. These threshold values were determined analytically upon data analysis for numerous CID.

We show the confusion matrix for each method in Figures S1abc. Using the heuristic proposed in Section 3.4, RF outperforms the other methods and show a higher positive rate than over the testing dataset (see Figure 2c). This owes to the choice of windows used for testing here that includes windows only up to 1000s from the true arrival. The STA/LTA filter also performs extremely well to detect true arrivals. However this good true positive rate comes at the cost of a large number of false alerts. The analytical method using only local time derivatives shows a large number of false negatives owing to both the choice of heuristic and the variety in true arrival waveform characteristics.

S3 Random forest classification hyper-parameter optimization

Random forests have two main hyper-parameters to optimize during training: 1) maximum tree depth, and 2) number of trees. Increasing both hyper-parameters generally leads to increased accuracy as well as increased execution time and memory cost. Hyper-parameters should therefore be picked to maximize accuracy while minimizing computational cost. In Figure S2 we show the variations of various accuracy metrics for variations in tree depth and number of trees. The largest accuracy gain comes with the increase in maximum tree depth. However, variations in the number of trees do not lead to any substantial improvement across the different metrics. We select 800 trees as it maximizes the accuracy across recall, precision, and accuracy.

S4 List of input features

All input features used to train the RF classifier presented in Section 3 are described in Table table S3.

The distribution of input features over our training and testing datasets is shown in Figure S3.

S5 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 2b 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 S4 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.

Additionally, the significant overlap between distributions in Figure 2b 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 (KL divergence) between the two distributions with respect to the locations of the points in the map. Both PCA and TSNE projections are shown in Figure S5. No obvious clusters can be identified which suggests that only complex nonlinear relationships can help discriminating signals between noise and CID classes. This further highlights the complexity of this classification problem.

S6 Sensitivity of classification accuracy to number of validation points

The heuristic 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 S6, 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.

Table S3: List of attributes.

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
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
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	FT4/FT5
FT8	FT4/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 tim
FT16	Mean distance between the 3rd quartile and the 1st quartile of all FTs as a function of time
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
S11	Energy up to 0.5Nyf Hz
S12	Energy up to 0.75Nyf Hz
S14	Energy up to 1.0Nyf Hz

S7 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 S7, 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%.

S8 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 S8, 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.

S9 Computational cost of detection and association procedures

The distribution of computational costs for a waveform recorded at satellite G07, station 0048 is shown in Figure S9. The time picking step is only present when a detection occurs which explains the jump in computational cost around 7mn after the earthquake. The time evolution of computation costs for the association step over the entire satellite network during Tohoku is shown in Figure S10. We observe a significant increase in computational cost after 9 mn since the earthquake. This higher association cost owes to the increase in detections reported at each combination of satellite/station when the earthquake-induced acoustic wave reaches the ionosphere. Because the association procedure must scan through all possible combinations of detections to lump together arrivals from the same wavefront, the cost increases with the number of detections. The maximum cost for one time step over the whole network is less than 6 s.

S10 Processing of Iquique earthquake

In order to further assess the ability of our model to detect arrivals on new unseen data, we processed waveforms after the 2014 Iquique earthquakes (see Table S1). In Figure S11a, we show the slip distribution of the Iquique earthquake along with, in Figures S11bc, the RF predicted arrivals times and association classes. The earliest arrival times are reported when the ionospheric points are the closest to the regions of maximum slip.

S11 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 S12. We observe that the true arrival is accurately retrieved by our detection model. However, detection probabilities are much lower than for 30s data (see Figure 3a). These lower 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.

S12 Detection of ionospheric signal from volcanic eruptions

Other low-frequency acoustic sources, such as explosions, volcanoes, or meteorites, can generate transient ionospheric perturbations. In particular, volcanic eruptions generate both infrasonic and gravito-acoustic signals in the 1-10 mHz frequency range. While the physics of gravity-wave propagation differs from infrasound propagation Hines (1960), potential similarities between timeseries for such phases when processing incomplete wavetrains could trigger a RF detection. We therefore 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 S13, we observe that the first arrival of the main disturbances associated with the volcanic eruption is accurately captured. However, later arrivals are not detected by our model suggesting that vTEC data from volcanic phases should be included in the training dataset to better discriminate noise vs earthquake vs volcanic activity.

References

E. Astafyeva and K. Shults. Ionospheric gnss imagery of seismic source: Possibilities, difficulties, and challenges. *Journal of Geophysical Research: Space Physics*, 124(1):534–543, 2019. doi: 10.1029/2018JA026107.

E. Astafyeva, P. Lognonné, and L. Rolland. First ionospheric images of the seismic fault slip on the example of the tohoku-oki earthquake. *Geophysical Research Letters*, 38(22), 2011. doi: 10.1029/2011GL049623.

E. Astafyeva, L. Rolland, P. Lognonné, K. Khelfi, and T. Yahagi. Parameters of seismic source as deduced from 1hz ionospheric gps data: Case study of the 2011 tohoku-oki event. *Journal of Geophysical Research: Space Physics*, 118(9):5942–5950, 2013a. doi: 10.1002/jgra.50556.

E. Astafyeva, S. Shalimov, E. Olshanskaya, and P. Lognonné. Ionospheric response to earthquakes of different magnitudes: larger quakes perturb the ionosphere stronger and longer. *Geophys. Res. Letters*, 40:1675–1681, 2013b. doi: 10.1002/grl.50398.

E. Astafyeva, L. M. Rolland, and A. Sladen. Strike-slip earthquakes can also be detected in the ionosphere. *Earth and Planetary Science Letters*, 405:180–193, 2014. ISSN 0012-821X. doi: 10.1016/j.epsl.2014.08.024.

M. S. Bagiya, P. S. Sunil, A. S. Sunil, and D. S. Ramesh. Coseismic contortion and coupled nocturnal ionospheric perturbations during 2016 kaikoura, mw 7.8 new zealand earthquake. *Journal of Geophysical Research: Space Physics*, 123(2):1477–1487, 2018. doi: 10.1002/2017JA024584.

M. S. Bagiya, A. Sunil, L. Rolland, S. Nayak, M. Ponraj, D. Thomas, and D. S. Ramesh. Mapping the impact of non-tectonic forcing mechanisms on gnss measured coseismic ionospheric perturbations. *Scientific reports*, 9(1):1–15, 2019. doi: 10.1038/s41598-019-54354-0.

M. N. Cahyadi and K. Heki. Coseismic ionospheric disturbance of the large strike-slip earthquakes in north sumatra in 2012 mw dependence of the disturbance amplitudes. *Geophysical Journal International*, 200 (1):116–129, 11 2015. doi: 10.1093/gji/ggu343.

K. Heki and J. Ping. Directivity and apparent velocity of the coseismic ionospheric disturbances observed with a dense gps array. *Earth and Planetary Science Letters*, 236(3):845–855, 2005. ISSN 0012-821X. doi: 10.1016/j.epsl.2005.06.010.

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

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.

D. Thomas, M. S. Bagiya, P. S. Sunil, L. Rolland, A. S. Sunil, T. D. Mikesell, S. Nayak, S. Mangalampalli, and D. S. Ramesh. Revelation of early detection of co-seismic ionospheric perturbations in gps-tec from realistic modelling approach: Case study. *Scientific reports*, 8(1):1–10, 2018. doi: 10.1038/s41598-018-30476-9.

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.

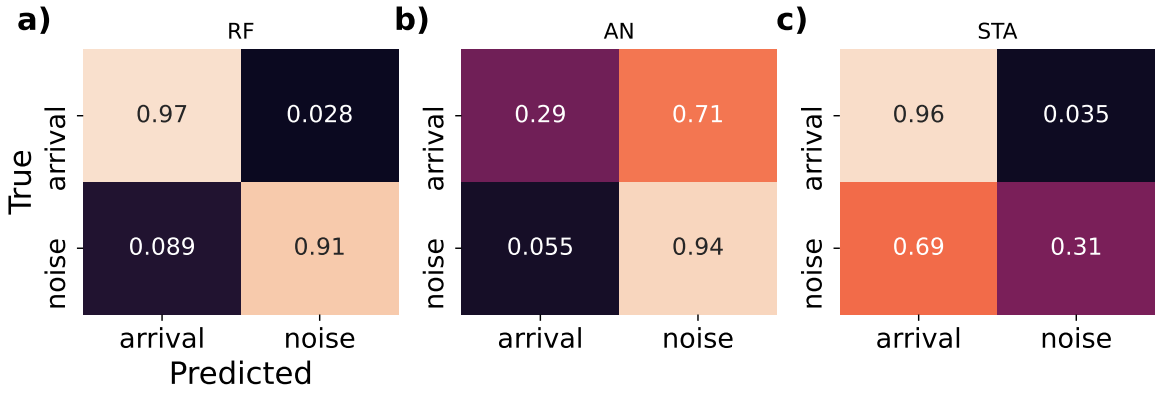

Figure S1: Confusion matrices calculated over a waveform dataset consisting of 2000 s windows centered on each true arrival for (a) the RF classification model, (b) the analytical time-derivative based model, and (c) the STA/LTA filter. The normalization of confusion matrices is identical to what is shown in Figure 2c.

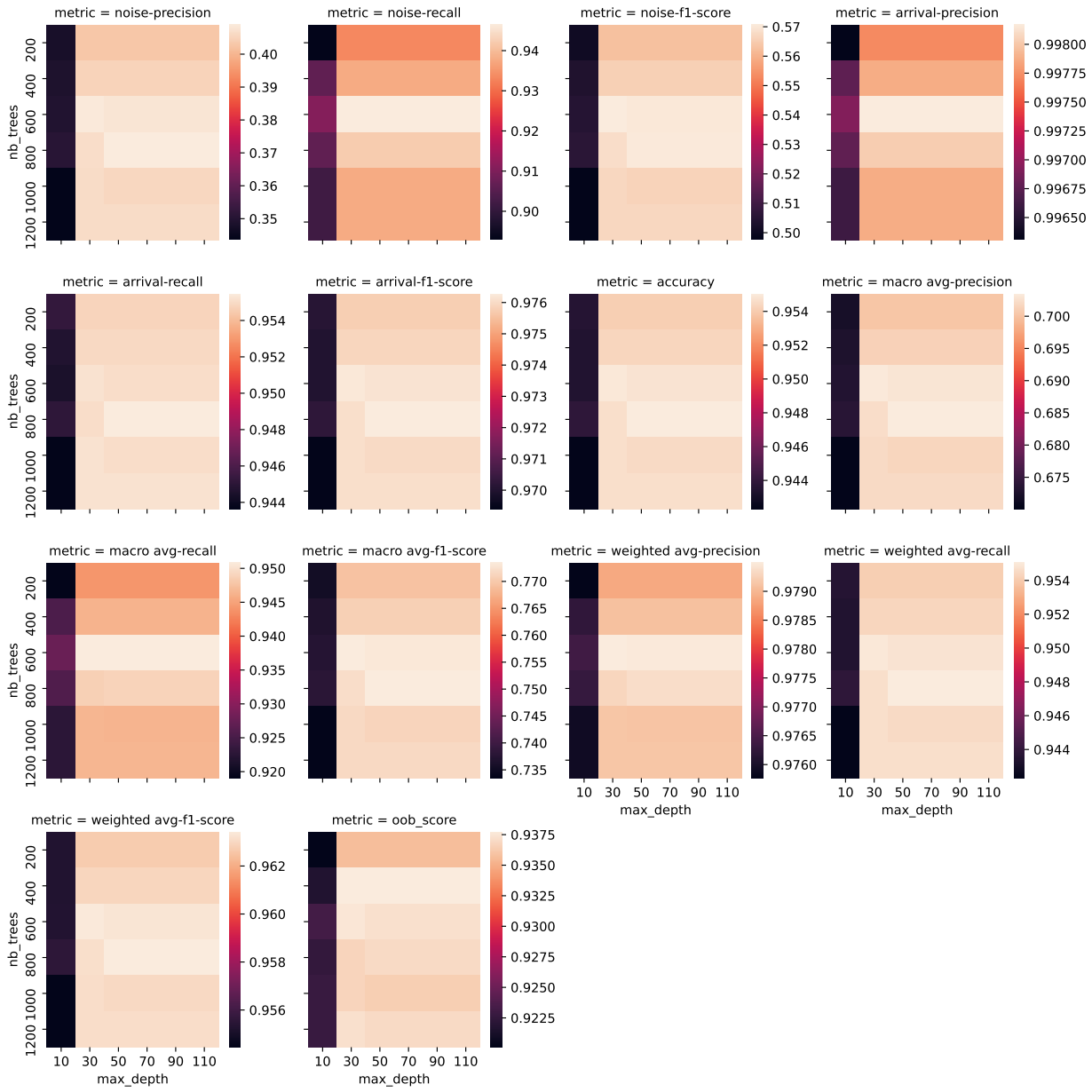

Figure S2: List of metrics to assess the performances of our RF classifier for detection (see Section 3.2) as a function of tree depth and number of trees.

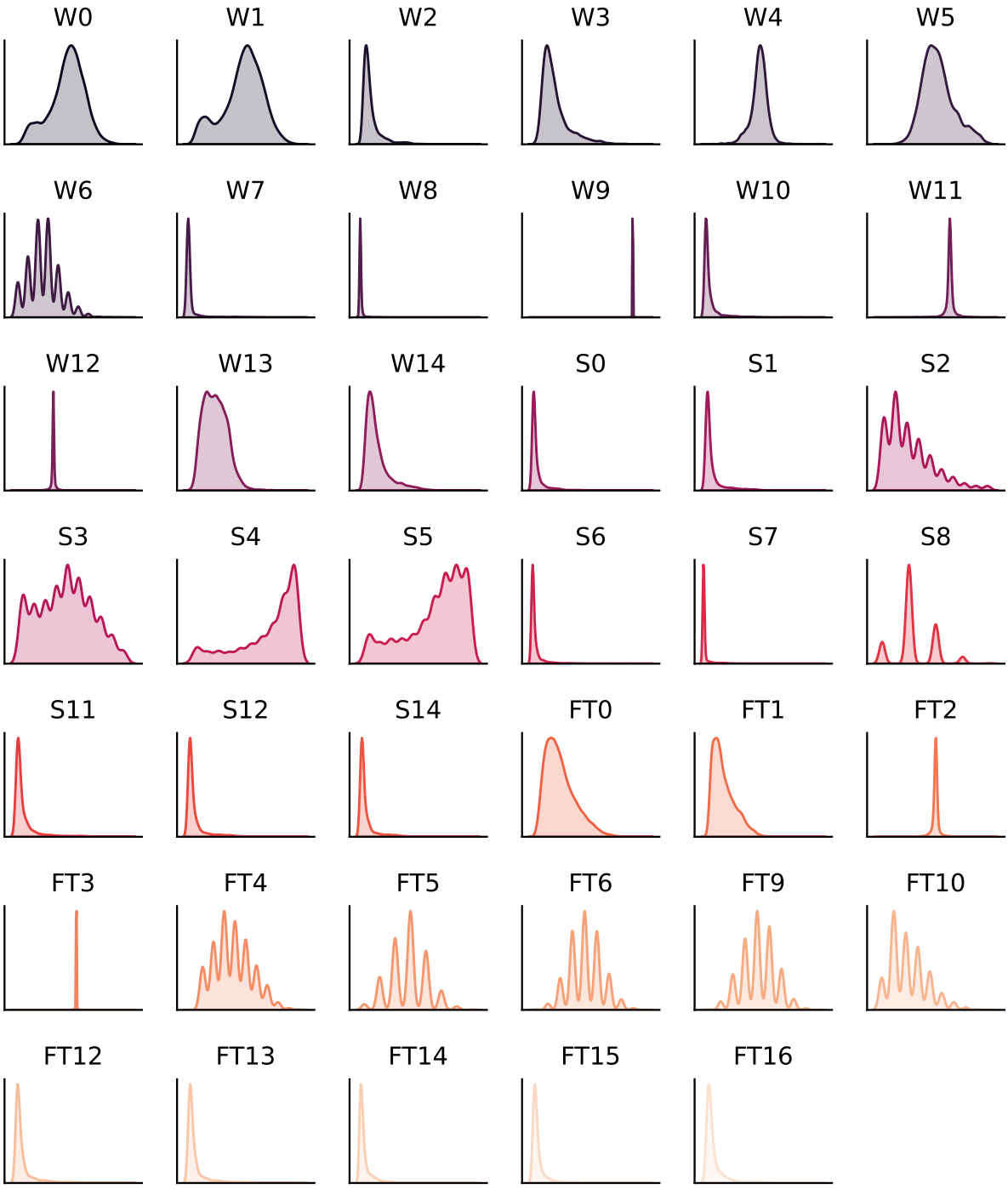

Figure S3: Distribution of each input features over our training and testing datasets. The short name of feature for each plot is shown above the plot. The description of each feature is given in Table table S3

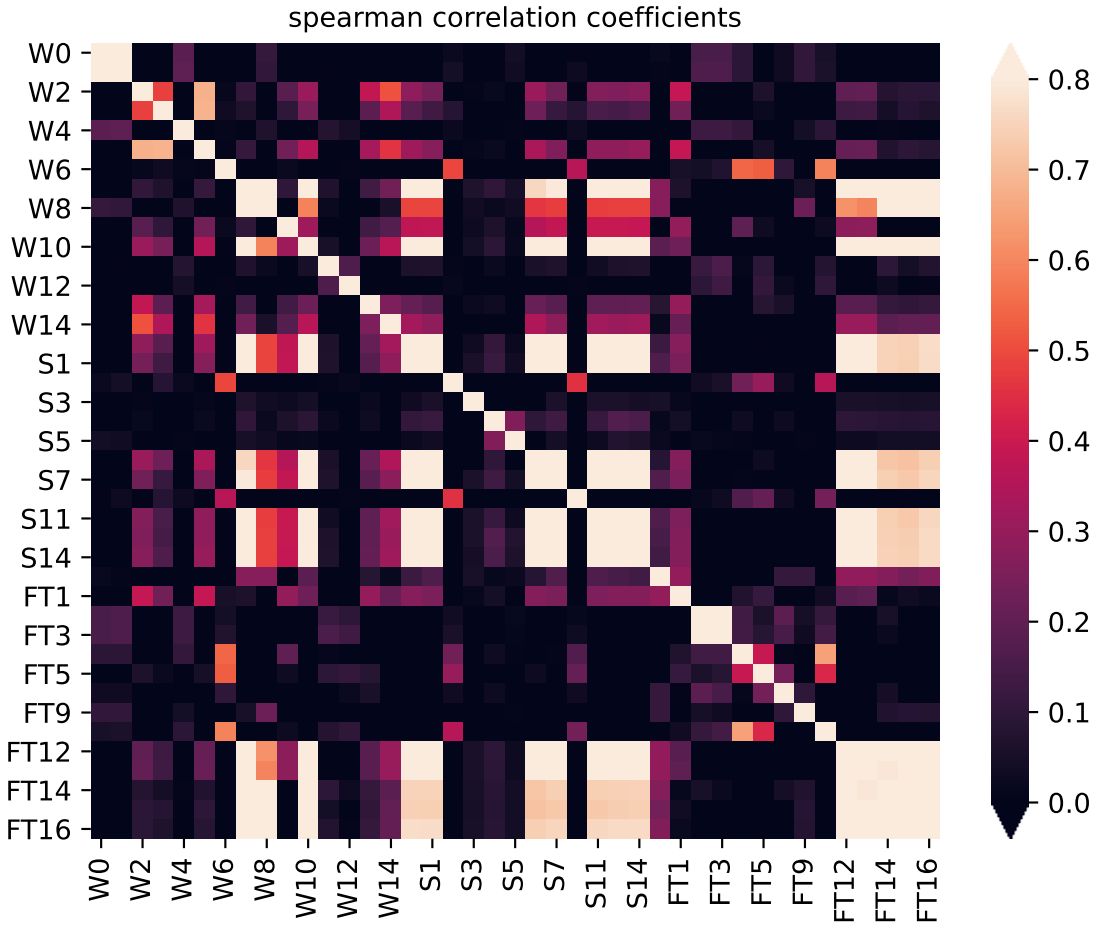

Figure S4: Spearman's correlation coefficients between each feature used for training. A description of each feature is given in Table table S3.

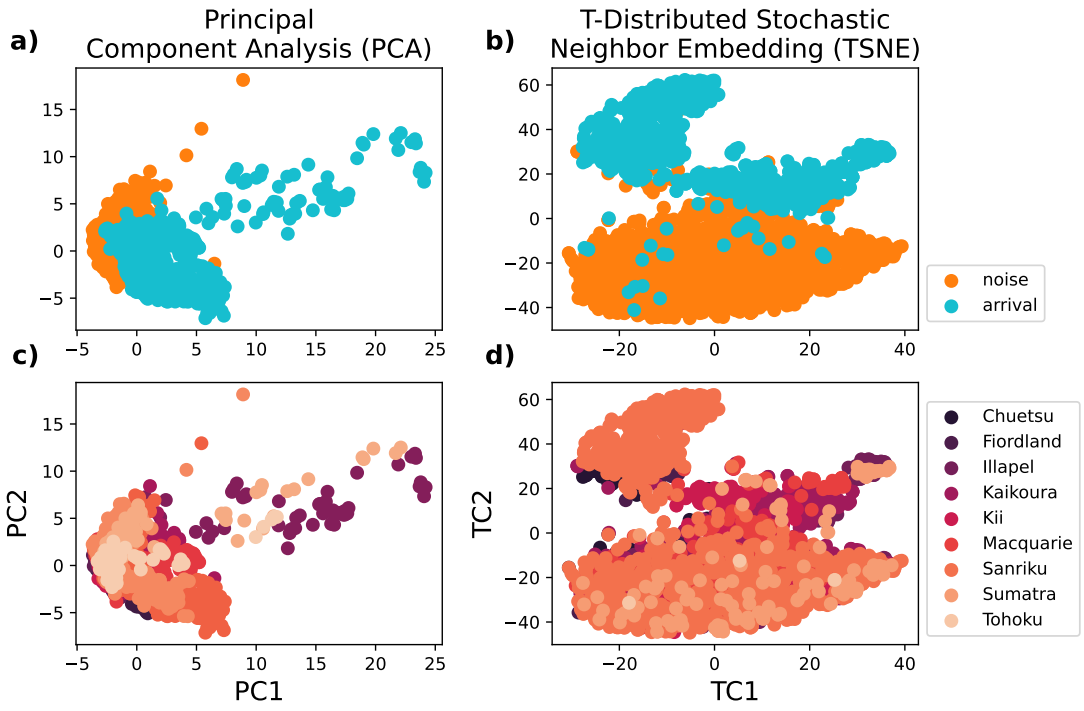

Figure S5: First versus second component of (a,c) a Principal Component Analysis (PCA) and (b,d) a T-Distributed Stochastic Neighbor Embedding (TNSE, Van der Maaten and Hinton (2008)). Points are colorcoded with (a,b) the detection class, and (c,d) the event name for the arrival class.

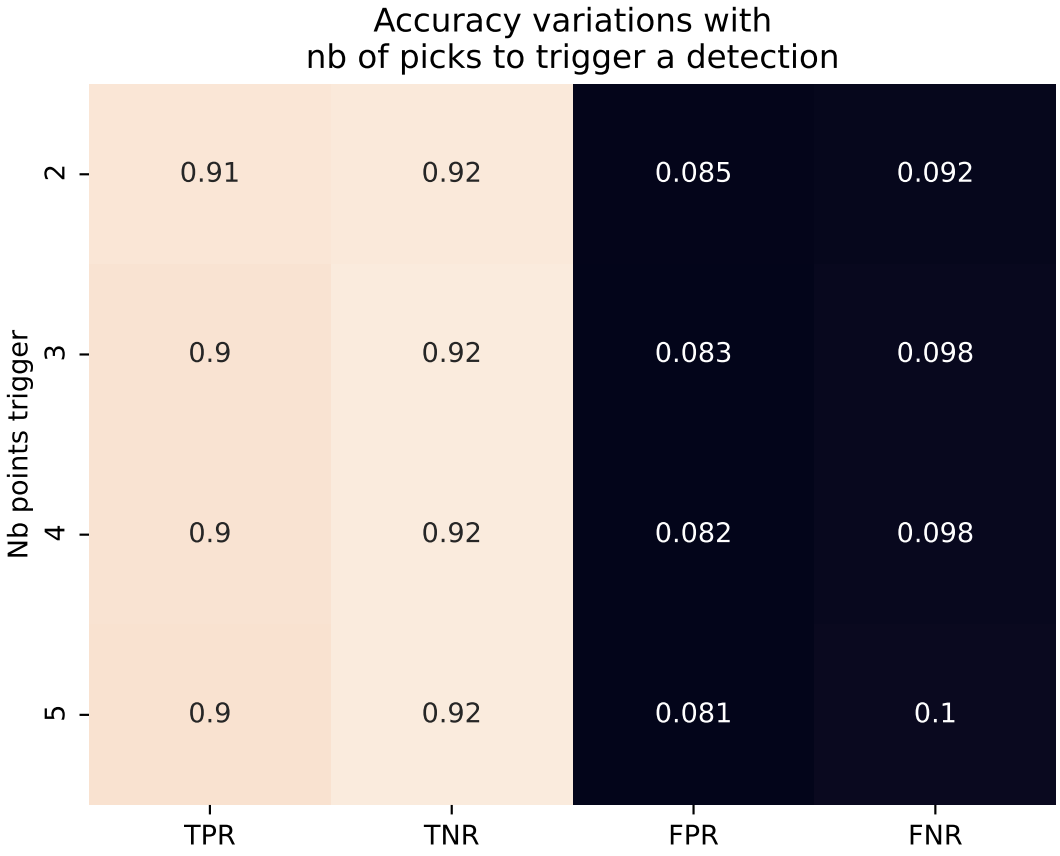

Figure S6: True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) with the choice of number of time steps for validation in the heuristic presented in Section 3.4

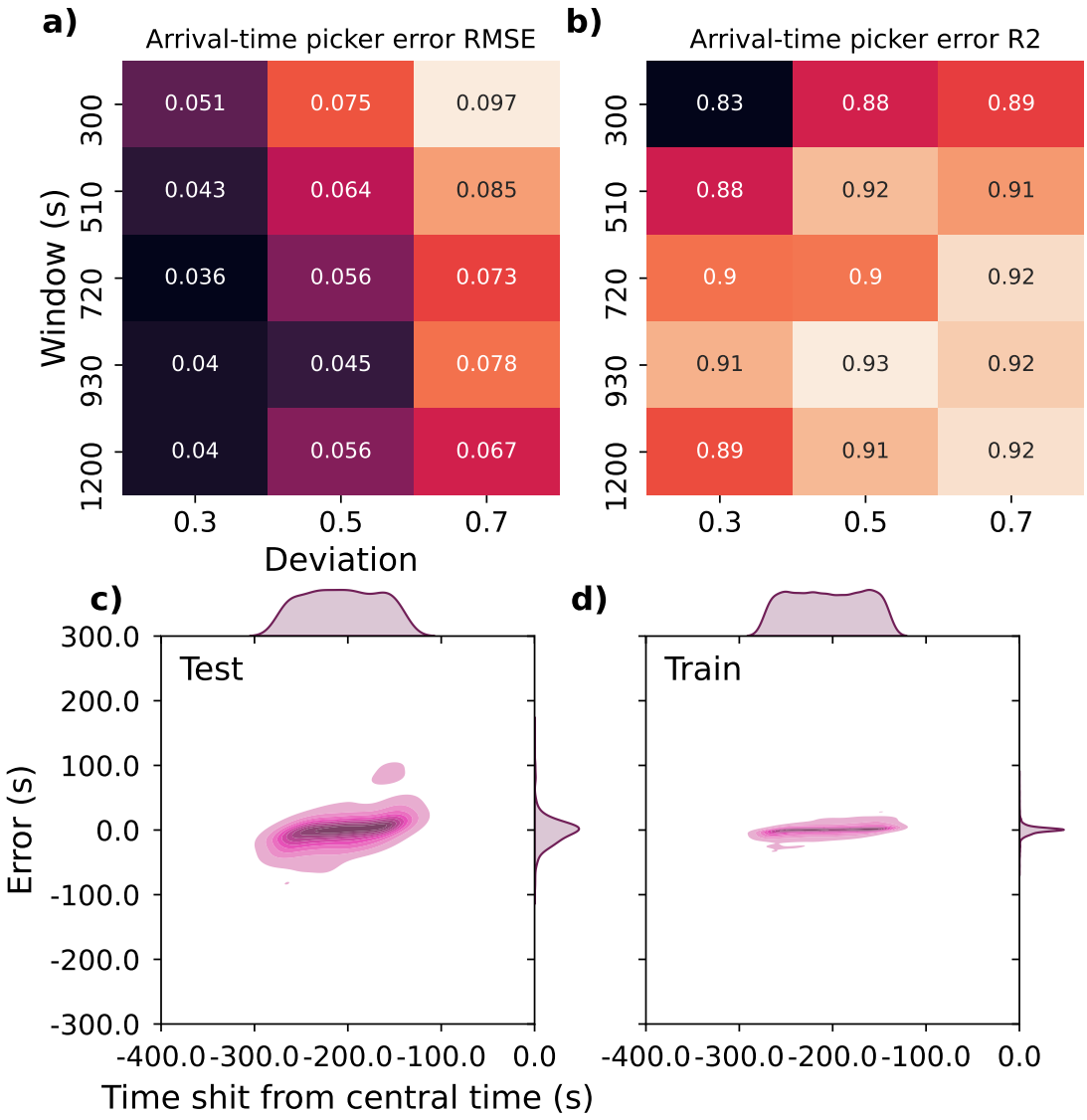

Figure S7: Performance of RF arrival time picker. (a) Root Mean Square Error (RMSE) vs minimum true-wavetrain overlap (deviation) and window size (s). The minimum true-wavetrain overlap corresponds to the minimum fraction of the wavetrain that has to be included in a window to be considered for training. (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.

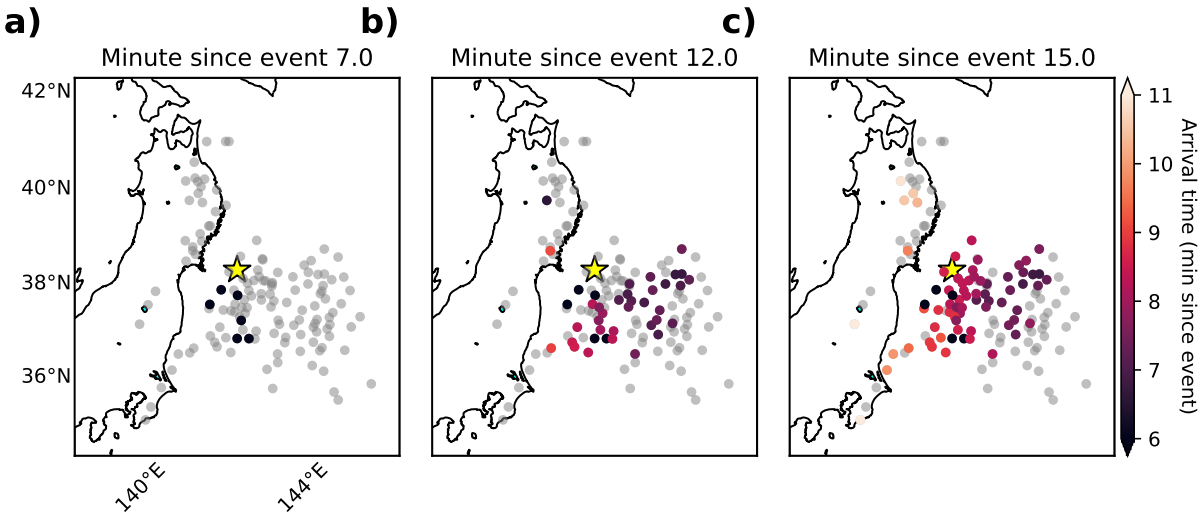

Figure S8: Distribution of detected arrival times after (a) 7 minutes, (b) 12 minutes, and (c) 15 minutes since the event. We used $H_{ion} = 180$ km to determine the location of the ionospheric points. The colorcode corresponds to the predicted arrival time at each satellite/station. Grey dots correspond to the location of satellites/stations where there is no detection yet but where there will eventually be a detection.

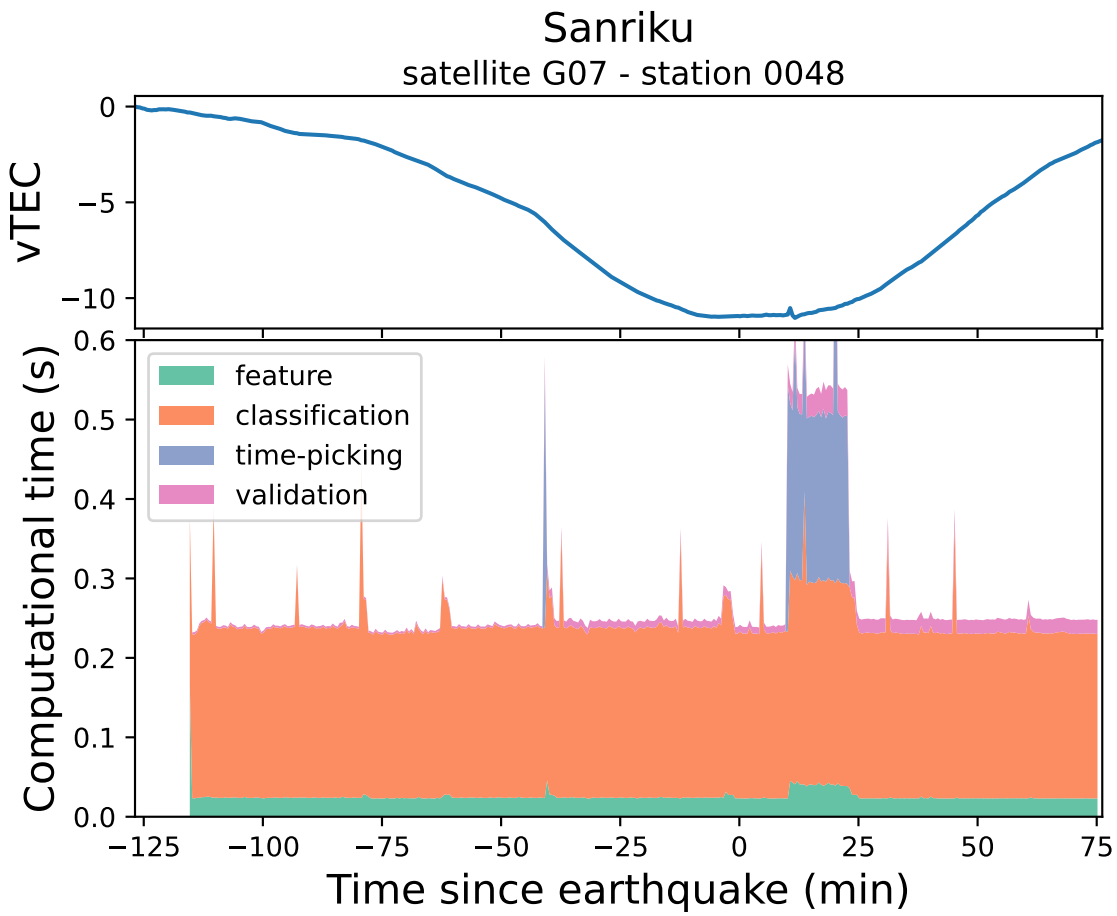

Figure S9: Distribution of detection procedure's computational cost for one station/satellite. Top, vTEC timeseries for satellite G07 and station 0048. Bottom, stack of computational time (s) for earthquake Sanriku, satellite G07 and station 0048, for various detection operations: feature extraction (green, see Section 3.1), RF classification (orange, see Section 3.2), RF arrival-time picking (blue, 3.3), and heuristic validation (pink, 3.4).

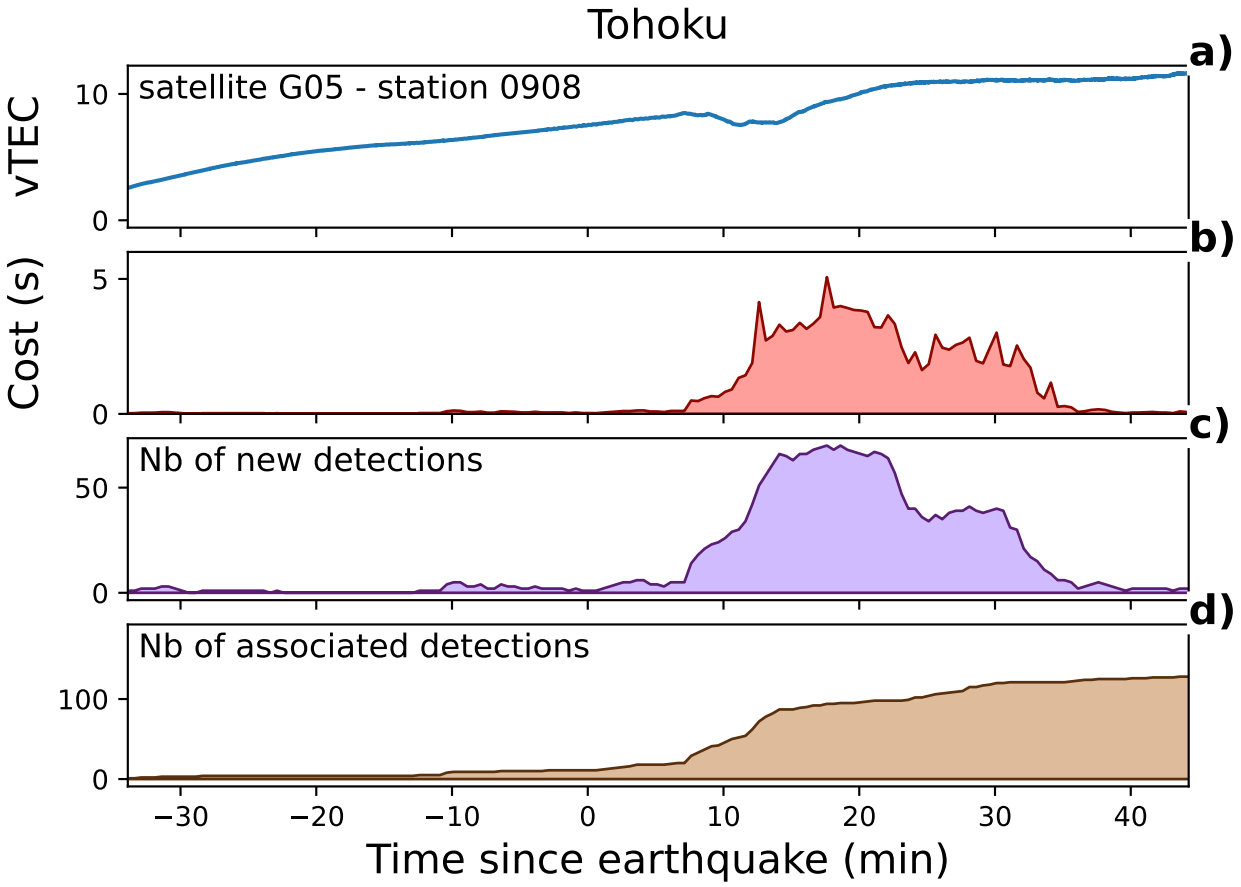

Figure S10: Time evolution of the association procedure's computational cost for earthquake Tohoku. (a) vTEC timeseries for satellite G05 and station 0908. (b) Computational cost (s) at each time iteration of the association procedure. (c) number of new detections per time iteration. (d) number of associated detections up to current time iteration.

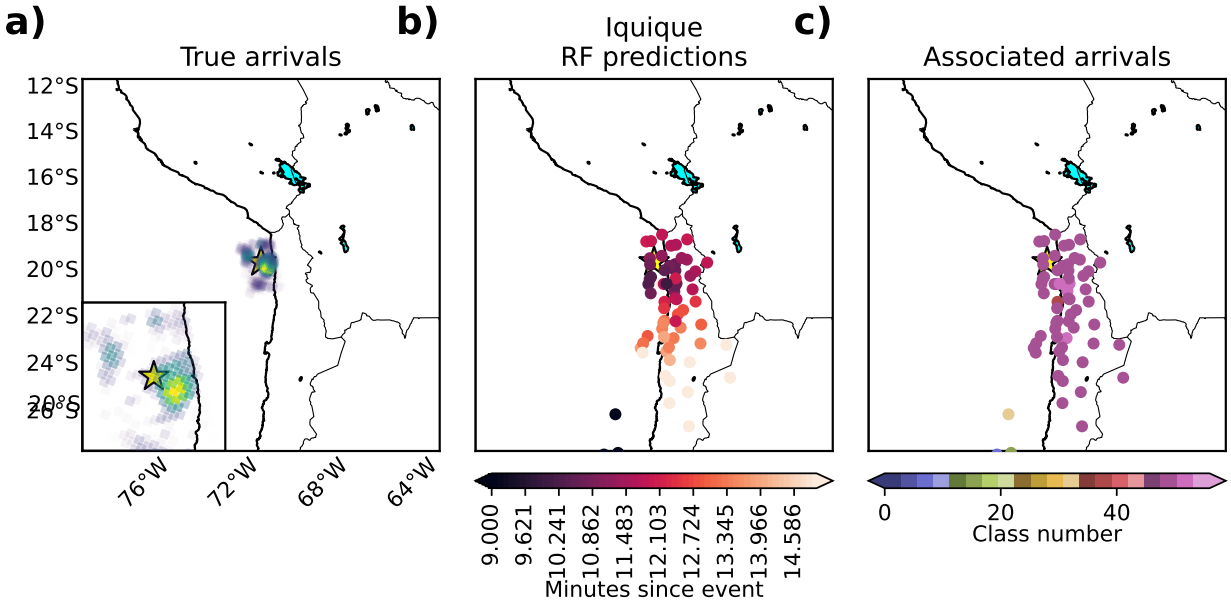

Figure S11: Ionospheric arrival-time maps computed 16 minutes after the Iquique earthquake. (a) map showing the epicenter location (yellow star), and the maximum fault slip (in m) as green to yellow patches, (b) RF-based arrival-time predictions, and (c) association classes determined from the RF-predicted time using the method presented in Section 3.4.

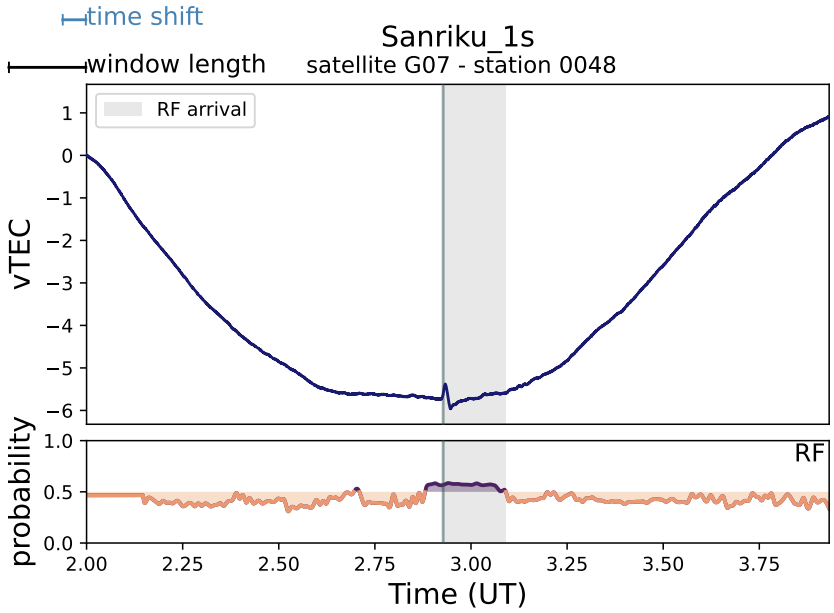

Figure S12: Performance assessment of RF detection and arrival-time picking at a higher sampling rate of 1s. 4-h vTEC waveform for the Sanriku event, satellite G07, station 0048 along with detection probabilities predicted by our RF detection model. The true arrival is shown as a red vertical line while the RF-predicted arrival time as a dark grey vertical line. The wavetrain detected by the RF and heuristic models (see steps 4 and 6 in Section 3) is highlighted with a grey background.

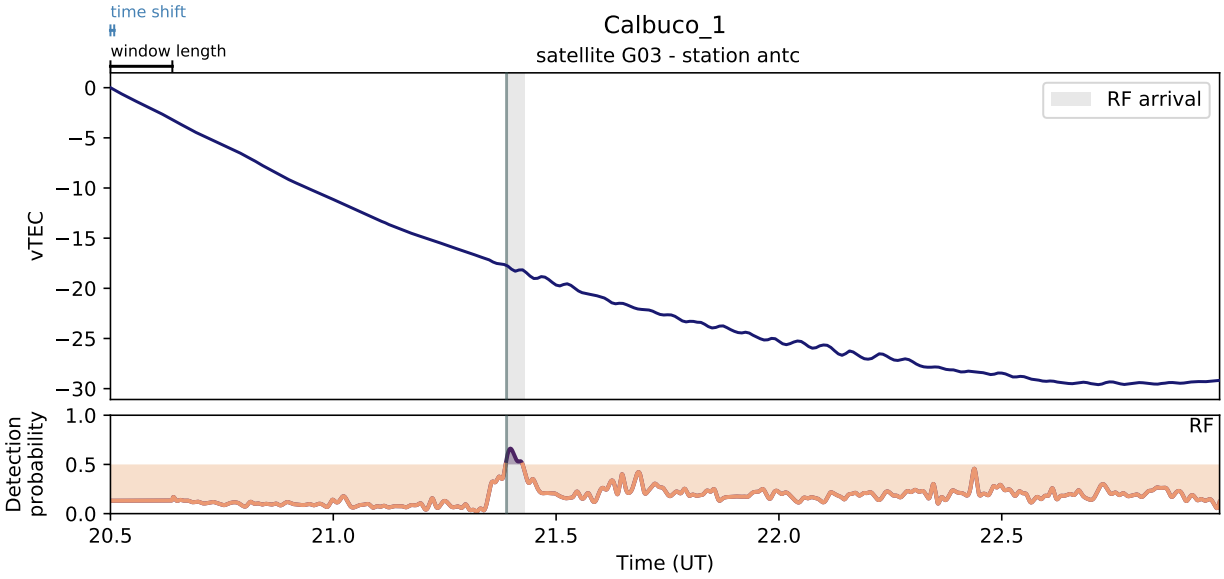

Figure S13: vTEC waveform for the Calbuco eruption, satellite G03, station antc along with detection probabilities predicted by our detection procedure (see Section 3) using a window size $w = 720$ s. Volcano-associated ionospheric perturbations are present between 21.3 and 22.5UT. The RF-predicted arrival time as a dark grey vertical line. The detected wavetrain using the RF is highlighted with a grey background.