Near-real-time detection of co-seismic ionospheric disturbances using machine learning

Quentin Brissaud, Elvira Astafyeva

submitted to *Geophys. J. Int.*

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

1 July 2021

SUMMARY

Tsunamis generated by large earthquake-induced displacements of the ocean floor can lead to tragic consequences for coastal communities. Ionospheric measurements of Co-Seismic Disturbances (CIDs) offer a unique solution to characterize an earthquake's tsunami potential in Near-Real-Time (NRT) since CIDs can be detected within 15min of a seismic event. However, the detection of CIDs rely on human experts which currently prevents the deployment of ionospheric methods in NRT. To address this critical lack of automatic procedure, we train machine-learning models (random forests) over an extensive ionospheric waveform dataset to (1) classify ionospheric waveforms between CIDs and noise, (2) pick arrival times, and (3) associate arrivals across a satellite network in NRT. Our model shows excellent classification and arrival-time picking performances ($\sim 95\%$ recall, average error < 10 s). This model is the first automatic CID detector which paves the way for the NRT imaging of surface displacements from the ionosphere.

Key words: Infrasound – Ionosphere/atmosphere interactions – Tsunami warning – Machine Learning

1 INTRODUCTION

Large seafloor displacements due to earthquakes are known to generate destructive tsunamis. Unfortunately, Near-Real-Time (NRT) mapping of the co-seismic surface displacements to characterize the earthquake tsunami potential

⋆ Correspondence: quentin@norsar.no

is still challenging for conventional methods, especially for earthquakes with $M_w > 8$ (LaBrecque et al. 2019; Wright et al. 2012; Katsumata et al. 2013). NRT corresponds to times within 15-20 minutes after the earthquake onset which is crucial for early-warning application as it gives several tens of minutes for populations to evacuate before the tsunami reaches the coasts.

Recently, several research groups have demonstrated that ionospheric measurements can offer an alternative to seismo-geodetic methods to estimate the tsunami potential of earthquakes. The ionosphere is an electrically charged atmospheric layer that is concentrated around 150-400 km of altitude. This layer is sensitive to the vertically propagating acoustic energy excited by natural hazards (earthquakes, tsunamis, volcanic eruptions) and man-made events (explosions, rocket launches, nuclear tests) (Heki 2006; Rolland et al. 2016; Komjathy et al. 2016; Shults et al. 2016; Astafyeva & Shults 2019). In particular, ionospheric signatures of earthquakes, known as co-seismic ionospheric disturbances (CID), reach ionospheric altitudes in 7-9 minutes after their generation at the surface. CIDs waveform characteristics are correlated to the seismic source properties. For instance, the amplitude of the CID scales almost linearly with the magnitude of an earthquake (Astafyeva et al. 2013c; Cahyadi & Heki 2015; Occhipinti et al. 2018; Heki 2021), or - for submarine earthquakes - with the tsunami wave height or volume of water that was displaced due to an earthquake (Kamogawa et al. 2016; Rakoto et al. 2018; Manta et al. 2020). Additionally, CID arrival times and detection coordinates provide strong constraints on the position of the seismic source, or the origin of tsunami (Afraimovich et al. 2006; Heki et al. 2006; Astafyeva et al. 2009; Tsai et al. 2011; Lee et al. 2018; Bagiya et al. 2020; Inchin et al. 2021; Zedek et al. 2021). Moreover, Astafyeva et al. (2013a,b); Astafyeva & Shults (2019) showed that the distribution of the first-detected CIDs match the position of the maximum displacement on the ground.

Despite the high potential of seismo-ionospheric assessment of natural hazards, the detection and analysis of ionospheric disturbances still rely on human experts. This manual process is problematic for processing large data volume to detect CIDs and estimate seismic source parameters. Only a few studies have focused on the automatization of detection procedures in the ionosphere but only at low frequencies (Efendi & Arikan 2017; Belehaki et al. 2020). Ravanelli et al. (2021) investigated the use of both GNSS ground and ionospheric TEC measurements for NRT tsunami genesis estimation. However, Ravanelli et al. (2021) did not present any detection procedure for CIDs, but only showed TEC variations in NRT scenario. In addition, their TEC processing procedure included the use of 8th order polynomial fit in order to highlight the co-seismic signature. The latter is not possible in our definition of NRT mode, i.e. 15-20 minutes after the earthquake onset time. Therefore, for both future NRT developments, and for processing of large amount of TEC data retrospectively, the community needs methods allowing for rapid automatic detection and recognition of CIDs.

To address the lack of automatic detection method, we build a machine-learning model, called a Random Forest (RF, Breiman (2001)), over an extensive CID waveform dataset from 12 large-magnitude earthquakes, to classify vTEC waveforms between CIDs and noise and pick arrival times in NRT. RFs have been employed for seismic waveform classification and show excellent performances to generalize training datasets (Provost et al. 2017; Li et al. 2018; Wenner et al. 2021). Our model is, to the best of our knowledge, the first automatic classifier and arrival-time

picker of CIDs. In this paper, we first describe the generation of our waveform dataset, our detection procedure, and our machine-learning models. We show classification performance results over our testing dataset and against other analytical detection methods. We finally discuss the future implementation of such method for NRT applications.

2 DATA COLLECTION

The Global Navigation Satellite Systems (GNSS) are nowadays widely used for detection of ionospheric disturbances. GNSS signals transmitted by satellites and captured by ground-based dual-frequency GNSS receivers enable the calculation of the differential slant TEC (sTEC). The technique of sTEC estimation is described in detail in numerous studies (Hofmann-Wellenhof et al. 2008; Afraimovich et al. 2006; Shults et al. 2016). The sTEC is equal to the number of electrons along a line-of-sight (LOS) between a satellite and a receiver. sTEC is measured in TEC units (TECU), with 1 TECU$= 10^{16}$ electrons/m^2. Because the sTEC is affected by the elevation angle of the LOS, we convert sTEC to vertical TEC (vTEC) by using the standard "mapping function" that is a function of the LOS elevation angle and the altitude of ionospheric detection H_{ion}. To construct our database, we collected GNSS-TEC data for 12 earthquakes that occurred between 2003 and 2016 (see Figure 1a). These events produced visible response in the ionospheric vTEC (see Figures 1bcd). For some events, CID were recorded by multiple satellites with sampling rates from 1 to 30 seconds (see Supplementary Table S1). The M6.6 Chuetsu earthquake is the smallest earthquake ever recorded in ionospheric GNSS data.

3 AUTOMATIC DETECTION AND ASSOCIATION MODELS

We propose a multi-step Random-Forest (RF) based detection procedure to detect and associate CIDs (see Figure 1e): 1) selection of a time window, 2) data preprocessing, 3) waveform features extraction, 4) RF-based classification of inputs features between noise and earthquake classes, 5) RF-based arrival time picking within windows showing a detection probability $> 50\%$, 6) confirmation of the presence of an arrival, and 7) if a detection is confirmed at step 6), we then associate this arrival to previously detected wavetrains. Finally, we shift the time window and start the procedure over at step 1).

3.1 Preprocessing and feature extraction

To extract consistent waveform features in TEC data with different sampling rates, we first downsample all waveforms down to 30s. TEC data may contain short-wavelength signals from transient sources (e.g., volcano, explosions) and long-term trends due to GNSS satellite motion and other long-period TEC changes. Therefore, we simplify the CID detection problem by first taking the time derivative of vTEC waveforms to remove long-wavelength trends. Derivatives are computed using second order central differences in the interior points and second order one-sides (forward or backwards) differences at the boundaries. Once the TEC waveforms have been pre-processed, we extract 39 features calculated from the vTEC timeseries, spectra, and spectrograms (see Supplementary Section S4). These features are

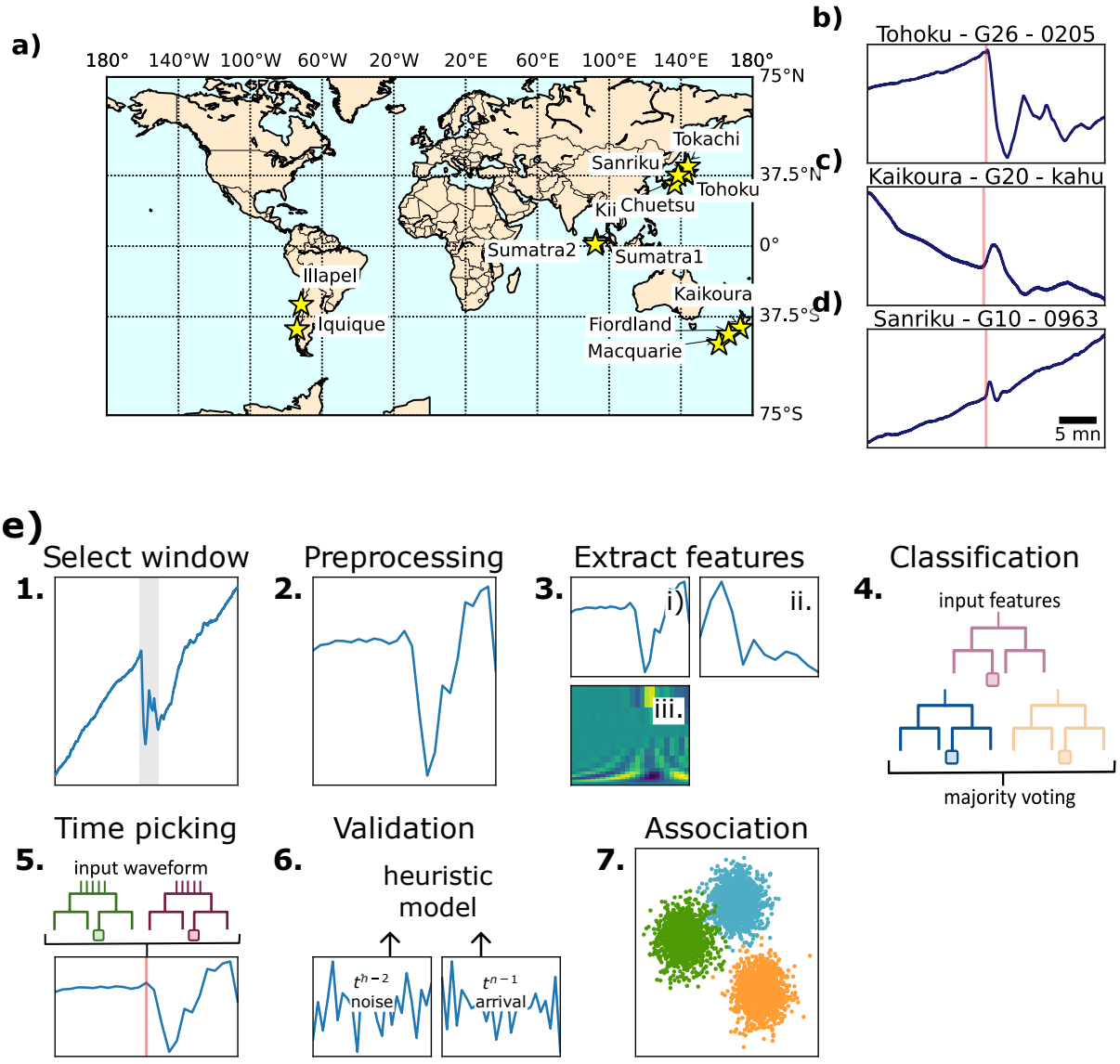

Figure 1. Detection and association procedures proposed in this study. (a) map showing the event included in the training dataset. Details about each event can be found in Table S1. (b,c,d) vTEC waveforms against time for event (b) Tohoku, (c) Kaikoura, and (d) Sanriku. (e) Detection procedure described in Section 3: 1) selection of a time window, 2) preprocessing of the waveform, 3) extraction of waveform features from i) time series, ii) spectrum, and iii) spectrogram, 4) RF classification of input waveform, 5) RF arrival time picking, 6) confirmation of an arrival if RF has classified three consecutive time windows (at times t^{n-2}, t^{n-1}, t^{n}) as arrival, and 7) association of arrivals across different satellites and stations.

commonly used for signal classification tasks (e.g., Hammer et al. (2013); Hibert et al. (2014); Provost et al. (2017); Wenner et al. (2021)).

3.2 Detection

We selected a RF model (Breiman 2001) to discriminate vTEC signals between earthquakes and noise classes. RFs have excellent generalization abilities, and do not require an extensive hyper-parameter tuning. We used the "Extra-Trees" scikit implementation of the random forest (Pedregosa et al. 2011) which introduces an additional layer of randomness when building decision trees which allow for better generalization of the training dataset (Geurts et al. 2006). The training procedure relies on bootstrap samples to build each tree along with out-of-bag samples to estimate the generalization score. Bootstrapping makes decision trees less sensitive to the choice of training dataset which reduces the probability of overfitting. Additionally, the error computed from out-of-bag samples provides an excellent metric for RF's classification performances.

For each station, CID wavetrains are described by an arrival time and a duration, the latter being uniform across satellites and stations for a given event (see Supplementary Table S1). We consider a time-window to contain a CID if it overlaps the true wavetrain by at least 70% which makes the RF more flexible to detect partial CID waveforms. Similar to Ross et al. (2018), we augment our training dataset by selecting three time-windows over each CID arrival by randomly perturbing the beginning of the time window while still fulfilling the 70% overlap condition. Noise waveforms are selected randomly across all dataset with the condition that it should not overlap any CID wavetrain. The arrival waveform dataset (2110 CIDs and 2110 randomly-picked noise waveforms) is splitted into a 90% training and validation dataset and a 10% testing dataset. The testing dataset is used to calculate confusion matrices and measure the rate of false and true positives which not accessible when bootstrapping samples. Finally, best results were found using 800 decision trees to build our classifier with a maximum tree depth up to 50 (see Supplementary Section S3).

3.3 Arrival-time picking

After the classification step, our detection algorithm needs to accurately select the arrival time in each window with a detection probability $> 50\%$. This time picking procedure remains challenging using threshold-based conditions such as STA/LTA filters (Allen 1982). False positives will degrade the arrival time estimate when using threshold-based methods since signal-to-noise ratio, signal duration and dispersion characteristics vary significantly between events. To overcome this problem, we build an automatic arrival-time picking procedure by using an "ExtraTrees" RF regressor. We train the RF using normalized time-derivative of vTEC amplitudes over windows containing a true arrival as inputs and the signal arrival time as an output. In order to lower the range of arrival-time output values, we use the offset in seconds from the window central time as an output instead of the absolute time. Input waveforms are pre-processed identically in both the detector described in Section 3.2 and in this arrival-time picking procedure.

We select arrival window for waveforms that overlaps the true wavetrain by at least 30%. Note that this overlap is significantly lower than for the detector. This choice aims at training the RF to pick arrival times over the first detection window with incomplete CID waveforms. Similar to Section 3.2, we augment our training dataset by selecting three time-windows over each CID arrival by randomly perturbing the beginning of the time window while still fulfilling

the 30% overlap condition which captures the uncertainty in arrival-time picking. The arrival waveform dataset (2110 CIDs) is splitted into a 90% training and validation dataset and a 10% testing dataset. A sensitivity analysis of the RF accuracy is provided in S7.

3.4 Validation

Owing to the natural variability of the ionosphere, false detections can still be present after the RF classification step. These false detections generally correspond to short-time spikes in RF detection probabilities while true detections show an increase in RF detection probabilities over longer time periods. To further remove false positives, we confirm a detection if 3 consecutive time windows show a detection probability over 50%. Variations of this value between 2 and 5 have a relatively small ($< 1\%$) influence on both recall and precision (see Supplementary Section S6). Short-time decrease in detection probabilities can occur within long CID wavetrains, caused by large earthquakes, compared to the processing time window. To reduce the number of false negatives, we notify the end of an CID wavetrain if 4 consecutive time windows show a detection probability below 50%. Once a detection is confirmed, we compute its arrival time as the 8^{th} decile of the 10 first predicted arrival times across the detections windows. This choice of quantile removes the influence of outliers in predicted arrival times. We do not include predicted arrival times beyond 10 time steps, i.e. 300 s, since these arrivals might correspond to time windows that do not include the true arrival time.

3.5 Association

After the detection is complete for a given combination of station and satellite, we can extract the spatial variations of detected arrival times across a satellite network. However, this step requires to associate arrivals belonging to the same wavefront as false positives can still pollute the detection dataset after step 5. This association step is performed on a set of confirmed arrivals and consists of three steps: 1) for new detections $d_{current}$, give $d_{current}$ an unused association number $s_{current}$, 2) For each detection $d_{current}$ find other confirmed detections d_{accept} across the satellite network within an acceptable time range from the current detection $d_{current}$. By acceptable time range, we consider all arrivals with a time offset from the current detection $t_{offset} < \frac{r_{max}}{c_{min}}$, where $r_{max} = 500$ km is the maximum association range, and $c_{min} = 0.65$ km/s is the minimum horizontal acoustic velocity. r_{max} is chosen as the maximum possible radius of a CID wavefront, and c_{min} corresponds to the minimum acoustic velocity in the lower ionosphere. Finally, 3) for each detection in an acceptable time range d_{accept}, if detection has an association number s_{accept}, change $s_{current}$ to s_{accept}.

4 RESULTS

The performance of the classification procedure presented in Section 3 is sensitive to the window size used for training. In Figure 2a, we show recall and precision for both classes vs the choice of window size. Precision indicates the

proportion of true detections relative to all detections (true positives plus false positives). Recall corresponds to the ratio of correct detections over all detections that should have been made (true positives plus false negatives). We observe that there is a clear improvement in both precision and recall (up to $\sim 94\%$) with an increase in window size over the testing dataset up to 720 s. This owes to the higher number of incomplete CID wavetrain for smaller windows than larger ones. For larger time windows > 720 s, precision and recall values plateau as the predictive power of some input features computed over large time windows diminishes. We selected a time window of 720 s which gives excellent classification results while facilitating the arrival time picking procedure by decreasing the range of possible values compared to larger time windows. Timeseries inputs shown in Figure 2b seem to be the most important features as determined by our RF. However, The overlap between input distributions motivates the choice of a large number of features to classify waveforms (see Supplementary Section S5).

The detection model's confusion matrix and ROC curve are shown in Figures 2c and d. Note that recall for both classes is different than in Figure 2a since these numbers correspond to an average across multiple overlap thresholds. The recall is high for a wide range of probability thresholds indicating that the RF rarely labels true arrivals as noise. This value decreases rapidly for probability thresholds $> 50\%$ corresponding to a stricter classification. However, with larger thresholds, the fall-out, i.e., the number of false alerts will also decrease. This highlights that the threshold can be adapted to specific applications depending on the objective. For early warning applications, the number of missed alert should be low and lower thresholds could therefore be used. In contrast, when building arrival-time catalog to invert for source parameters, precision is key and false alerts should be avoided, which necessitates larger thresholds. Additionally, results indicate that RF outperforms the other analytical methods, including STA/LTA filters, in terms of both true and false positive rates (see Supplementary Section S2).

Detection results for a waveform recorded during the 2011 Sanriku earthquake in Figure 2a show that both predicted (vertical grey line) and true (red vertical line in top panel) arrival times overlap, as the absolute error is low (< 3 s). Note that the predicted arrival time does not match the beginning of the sudden increase in detection probability since this predicted time corresponds to the output of the RF-based arrival time picker, which selects the best arrival time to use in each time window. We observe that the duration of this wavetrain (~ 450 s) is much larger than the true wavetrain (~ 200 s), owing to the large time windows employed in our detection model. Outside of the detected wavetrain, detection probabilities generally remain low ($< 25\%$) in accordance to the high true negative rate in Figure 2c. For NRT applications, the computational time is an important constraint. Numrical tests show that the whole detection process between steps 1) to 6) in Section 3 takes less than 600 ms to run: step 2 and 3 < 50 ms, step 4 < 200 ms,step 5 < 200 ms, and step 6 < 170 ms. This result suggests that this detection method could be implemented for near real-time applications at a higher sampling rate up to 1 Hz.

In addition to the classification of individual waveform snippets, accurate arrival times are crucial for near real-time applications. We assess our model's arrival-time picking accuracy by computing the error between predicted and true arrival times. Arrival-time errors for each event in our CID dataset in Figure 3b indicate that most arrivals ($\sim 95\%$) are captured with an absolute error $< 60s$, i.e., less than two time steps, and a large proportion of arrivals

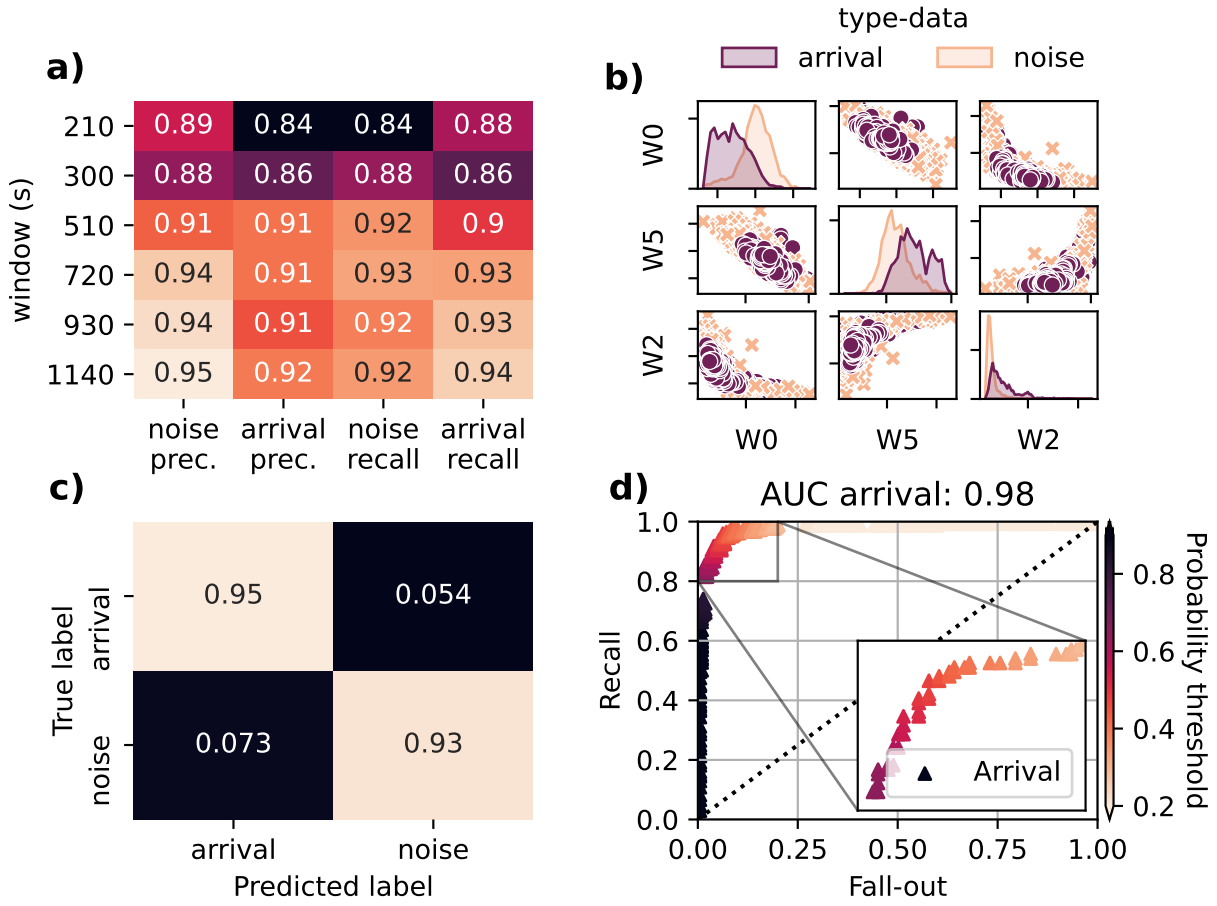

Figure 2. Sensitivity and accuracy of the RF classification step. (a) Precision (prec.) and recall for noise and arrival classes and various window sizes averaged over multiple overlap thresholds: $30\%, 50\%, 70\%$, and 90%. The following formula are used to compute recall and precision for arrival and noise: recall arrival = $\frac{TP}{TP+FN}$, recall noise = $\frac{TN}{TN+FP}$, precision arrival = $\frac{TP}{TP+FP}$, and precision noise = $\frac{TN}{TN+FN}$. TP, TN, FP, and FN correspond to True positive, True Negative, False positive, and False Negative. The correct detection of a CID corresponds to a TP. (b) Distribution of the three best features against each other. In the diagonal, we show univariate histograms for each feature. Best features are determined during training by calculating the Gini's impurity. W0 corresponds to the ratio of the envelope mean over the envelope maximum, W2 is the kurtosis of the timeseries, and W5 is the envelope skewness. (c) Confusion matrix for the detection model with window size $w = 720$ s and an overlap of 70%. The confusion matrix is normalized over each row. (d) Arrival-class ROC curve using the detection model with window size $w = 720$ s. The Area Under Curve (AUC) value is shown above the panel.

($\sim 80\%$) are accurately reproduced with an absolute error < 30 s, which is below the sampling rate in each CID waveform. Some outliers are present for both Illapel and Kaikoura events. Errors for the Kaikoura earthquake owe primarily to the high noise level in the waveforms which leads to large variations in vTEC time derivatives. For Illapel, false positives are lumped together with the true detection windows and degrade the arrival-time picking performance

over 4 time steps. However, we show in Figure 3d that the average arrival-time picking error across the whole dataset decreases significantly as you increase the number of time steps, i.e., the picking time delay.

Once detections are confirmed across multiple satellites/stations, we can associate each detection to the same wavefront using the method presented in Section 3.4. Associated arrival times predicted by our model can then be used to plot ionospheric maps for each event. Comparing Tohoku's ionospheric images in Figures 3f and g, we observe that the spatial distribution of arrival times is accurately reproduced by our detection model. The earliest arrival times match the location of maximum slip at the surface. The slight shift of the first arrivals to the south east owes to our choice of altitude of detection H_{ion} (Astafyeva et al. 2013b). However, some spurious arrivals are present in Figure 3g, with early arrival times west of the fault. These false detections correspond to rapid changes in vTEC occurring more than 20mn before or after the true arrival and classified as earthquake signals by our model. Owing to the large time difference between these spurious arrivals and the true arrivals, these false detections are correctly classified in a different association class. An example of such false detection is visible in the inset plot in Figure 3h as a vertical purple line. The time evolution of the distribution of confirmed arrivals (see Supplementary Section S8) indicates that the entirety of the true arrivals were detected within 15mn since the event. Additionally, we applied the detection and association procedure to unseen and unlabelled data extracted after the Iquique earthquake (see Supplementary Section S10). Arrival times are coherent with the region of maximum slip at the surface.

Figures 3f and g, also show that new detections have been reported by our model, in addition to the ones picked by human analysts, for the largest class corresponding the true CID (green class in Figure 3h). An example of such detected arrival is shown as an inset plot in Figure 3g. A low signal-to-noise ratio pulse is visible after the predicted arrival time (vertical line) at $t = 9.9$ mn after the earthquake, which is consistent with acoustic travel time from the source highlighted by other studies (e.g., Astafyeva et al. (2013b)). Using our model also ensures consistency in the choice of arrival times, in contrast to human analysts who introduce a subjective uncertainty range when determining the true onset. This association procedure is computationally expensive since it must scan through all possible neighbors of each new detection to update association classes. The algorithm's computational cost scales linearly with the number of new detections (see Supplementary Figure S10). The procedure takes around 1 s to process 10 new detections, at a given time, over a network of about 100 satellites/stations.

5 CONCLUSIONS AND DISCUSSION

We introduced a new automatic procedure for detection, arrival-time picking, and association of CIDs. Detection and arrival time picking steps are handled using random forests trained over a CID dataset from 12 earthquake events. These methods show excellent classification results with 95% true positive rate and 93% true negative rate, and arrival-time accuracy with an average error < 20 s using a 120 s time delay. Our model outperforms threshold-based detection methods in terms of both recall and precision. Our analytical classification procedure accurately associates all arrivals corresponding to the same wavefront. Classification results also indicate that low signal-to-noise ratio arrival that were not picked by human analysts could also captured by our RF detection model. Additionally, our model

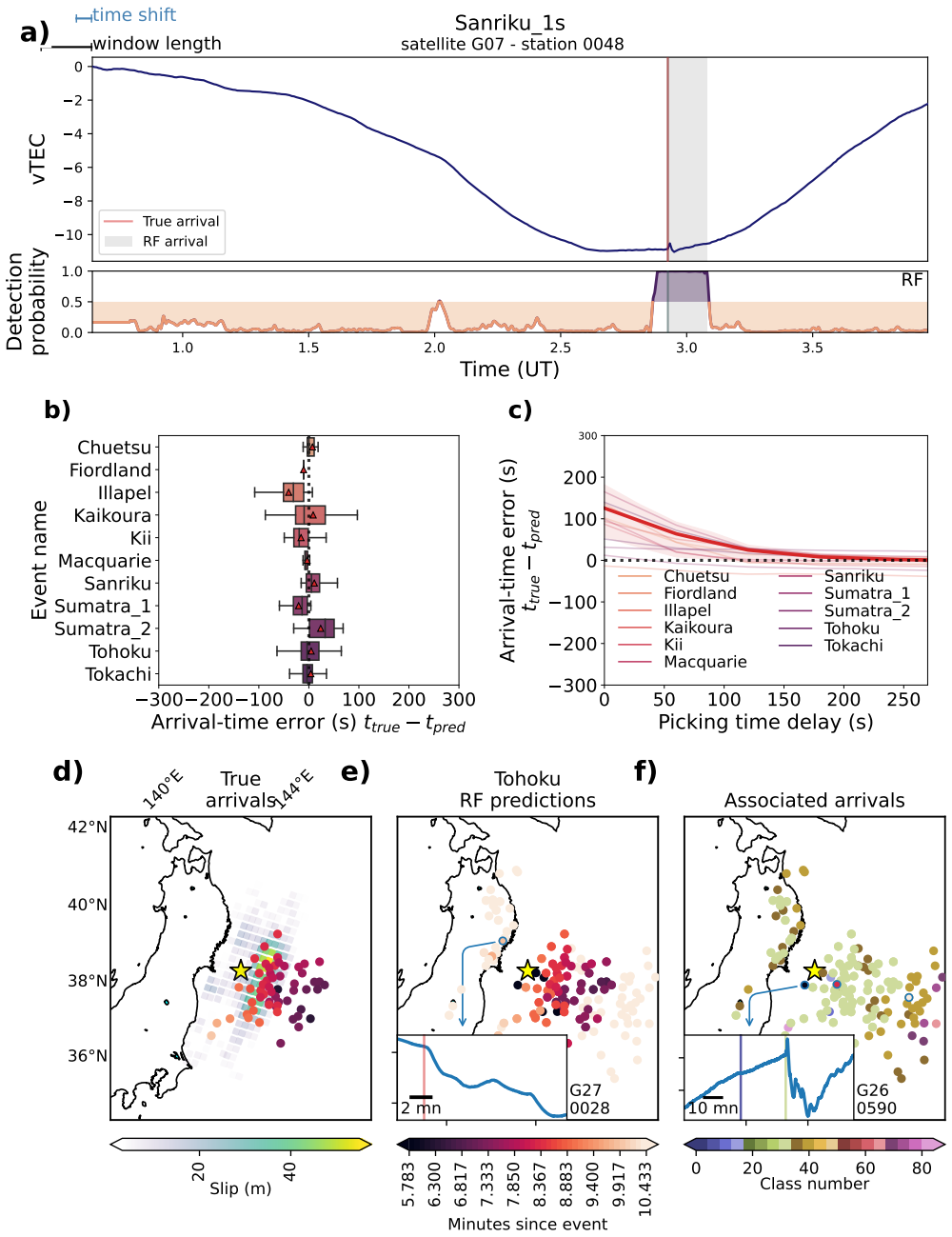

Figure 3. Performance assessment of RF arrival-time picking, and association steps. (a) 4-h vTEC waveform for the Sanriku event, satellite G07, station 0048 along with detection probabilities predicted by our RF detection model. The true arrival is shown as a red vertical line while the RF-predicted arrival time as a dark grey vertical line. The wavetrain detected by the RF and heuristic models (steps 4 and 6 in Section 3) is highlighted with a grey background. (b) box plot of arrival-time picking errors (in s) vs event. (c) Evolution of arrival-time picking error vs time delay since first detected window. The red curve shows the average error across all events. Red shaded background shows the standard deviation from the average across all events. Errors for each events are shown as thin solid lines in the background. Bottom, 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 fault slip (in m) as green to yellow patches, (e) RF-based arrival-time predictions with an inset plot showing a vTEC waveform for satellite G27 and station 0028, i.e., an arrival detected by our model but not reported by human analyst, and (f) association classes determined from predicted arrival times (see Section 3.4), along with an inset plot showing the vTEC data for satellite G26, station 0590. The vertical lines correspond to the arrival times of the two detected arrivals (first arrival is a false detection; the second is the true arrival) for station 0590. Here, the CID coordinates were calculated at the intersection point between the LOS and the ionospheric layer using $H_{ion} = 250$ for higher elevations, and 200 km for lower elevations (Astafyeva et al. 2013b).

seems to be able to detect vTEC variations associated with volcanic explosions (see Supplementary Section S12). However, the waveform's energy is primarily at lower frequencies which explains the inconsistency in volcanic arrival detections. This suggests that a dataset of volcanic-induced vTEC waveforms should be built and used to train an efficient discriminator between noise, earthquake, and volcanic phases.

The performance of our automated procedure is promising for future NRT applications, including the use of CID arrival times for construction of ionospheric images of seismic sources. The first demonstration of seismo-ionospheric imagery was based on retrospective analysis of CID generated by the 2011 Tohoku earthquake (Astafyeva et al. 2013a,b). Here we show that our newly developed method can generate such images in NRT. Note that the position of ionospheric detection points is dependent on the altitude of detection H_{ion}. The latter parameter is not known precisely, but it is presumed to be around the height of ionospheric ionization maximum, i.e. between 150 and 400 km, depending on solar, geomagnetic, seasonal and diurnal conditions. Future studies should account for uncertainties in H_{ion} to obtain accurate source locations. In addition to fault mapping, our method can be used to estimate earthquake magnitude. The latter can be done by removing non-tectonic impacts introduced by the magnetic field configuration at the epicentral area, LOS geometry factor and the background ionization (Bagiya et al. 2019).

We note that our procedure's practical implementation will require an efficient internet connection between the relevant GNSS stations to collect and extract timeseries for classification in NRT. Because the overall computational cost of one time iteration using our method is below 6 s on a single CPU using non-compiled Python codes (see Supplementary Section S9), at least 24 s are available for data acquisition and processing with waveforms sampled at 30 s. The association step is currently the most costly ($\sim 90\%$ of the total cost) but can be run in parallel to the other detection steps. Note that we also explored the feasibility of using our model to detect CIDs at a higher sampling rate by extracting input features without downsampling input data (see Supplementary Section S11). We obtained promising results using a 1s sampling rate, which show that our detection model is able to capture the true arrival time at the cost of a higher false positive likelihood.

Acquiring labeled vTEC data from additional events which will significantly improve the generalization abilities of our RF models. Additionally, the choice of features made in this paper could be further refined to obtain better accuracy (Han & Kim 2019). More accurate RF classifications could also alleviate the need for a validation step presented in Section 3.4. However, RF memory costs increase exponentially with tree depth, and consequently dataset size, $\sim 2^D$, with D the tree depth (Louppe 2014; Solé et al. 2014). The RF classification model is only about 70 mb but will grow considerably larger with new data. With a larger dataset, image segmentation ML techniques such as standard convolutional neural networks (Ross et al. 2018, 2019), transformers (Mousavi et al. 2020) or residual networks (Mousavi et al. 2019) applied on non-engineered inputs such as spectrograms could lead to substantial improvements in accuracy and memory costs for both classification and arrival time picking steps.

Finally, the proposed association algorithm does not incorporate any information about the source nor the atmospheric dynamics. This procedure could be improved by assessing the consistency of arrival time differences across a network of satellites and stations using a range of possible sources, similarly to the methods used for the automated

production of seismic bulletins (Draelos et al. 2015). In contrast to seismic media, atmospheric velocities, i.e., winds, are time-dependent which introduces further complexity when computing theoretical source-receiver arrival times. Fast simulations of acoustic wave propagation up to the ionosphere with realistic atmospheric specifications would greatly improve the classification between true and false arrivals and enable the localization of the largest surface displacements (Bagiya et al. 2019; Inchin et al. 2021; Zedek et al. 2021). Finally, to confirm the detection of an earthquake across a given network and trigger an alert for human analysts, an additional heuristic could be implemented based, for example, on the number of detections per association class.

ACKNOWLEDGMENTS

This work was supported by the French Space Agency (CNES, Project "RealDetect").

DATA AVAILABILITY

GNSS data are available from the following web-services: Japan GNSS Earth Observation System, GEONET (`http://datahouse1.gsi.go.jp/terras/terras_english.html`), GEONET Geological Hazard Information for New Zealand (`https://www.geonet.org.nz`), Scripps Orbit and Permanent Array Center (SOPAC, `http://sopac-old.ucsd.edu/dataBrowser.shtml`), National Seismological Centre, University of Chile (`http://gps.csn.uchile.cl`). Finite-fault data were downloaded from the US Geological Survey website (`https://earthquake.usgs.gov/earthquakes`). RF models, validation, and associations codes will be released upon publication on a FigShare repository.

REFERENCES

Afraimovich, E., Astafyeva, E., & Kiryushkin, V., 2006. Localization of the source of ionospheric disturbance generated during an earthquake, *International Journal of Geomagnetism and Aeronomy*, **6**, GI2002.

Allen, R., 1982. Automatic phase pickers: Their present use and future prospects, *Bulletin of the Seismological Society of America*, **72**(6B), S225–S242.

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

Astafyeva, E., Heki, K., Afraimovich, E., Kiryushkin, V., & Shalimov, S., 2009. Two-mode long-distance propagation of coseismic ionosphere disturbances, *J. Geophys. Res.*, **118**, A10307.

Astafyeva, E., Lognonné, P., & Rolland, L. M., 2013a. First ionosphere images for the seismic slip on the example of the tohoku-oki earthquake, *Geophys. Res. Letters*, **38**, L22104.

Astafyeva, E., Rolland, L. M., Lognonné, P., Khelfi, K., & Yahagi, T., 2013b. Parameters of seismic source as deduced from 1hz ionospheric gps data: case-study of the 2011 tohoku-oki event, *Journal of Geophys. Research*, **118**, 5942–5950.

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

Bagiya, M. S., Sunil, A., Rolland, L., Nayak, S., Ponraj, M., Thomas, D., & Ramesh, D. S., 2019. Mapping the impact of non-tectonic forcing mechanisms on gnss measured coseismic ionospheric perturbations, *Scientific reports*, **9**(1), 1–15.

Bagiya, M. S., Thomas, D., Astafyeva, E., Bletery, Q., Lognonné, P., & Ramesh, D. S., 2020. The ionospheric view of the 2011 tohoku-oki earthquake seismic source: the first 60 seconds of the rupture, *Scientific reports*, **10:5232**.

Belehaki, A., Tsagouri, I., Altadill, D., Blanch, E., Borries, C., Buresova, D., Chum, J., Galkin, I., Juan, J. M., Segarra, A., et al., 2020. An overview of methodologies for real-time detection, characterisation and tracking of traveling ionospheric disturbances developed in the techtide project, *Journal of Space Weather and Space Climate*, **10**, 42.

Breiman, L., 2001. Random forests, *Machine learning*, **45**(1), 5–32.

Cahyadi, M. N. & Heki, K., 2015. 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.

Draelos, T. J., Ballard, S., Young, C. J., & Brogan, R., 2015. A new method for producing automated seismic bulletins: Probabilistic event detection, association, and location, *Bulletin of the Seismological Society of America*, **105**(5), 2453–2467.

Efendi, E. & Arikan, F., 2017. A fast algorithm for automatic detection of ionospheric disturbances: Drot, *Advances in Space Research*, **59**(12), 2923–2933.

Geurts, P., Ernst, D., & Wehenkel, L., 2006. Extremely randomized trees, *Machine learning*, **63**(1), 3–42.

Hammer, C., Ohrnberger, M., & Faeh, D., 2013. Classifying seismic waveforms from scratch: a case study in the alpine environment, *Geophysical Journal International*, **192**(1), 425–439.

Han, S. & Kim, H., 2019. On the optimal size of candidate feature set in random forest, *Applied Sciences*, **9**(5), 898.

Heki, K., 2006. Explosion energy of the 2004 eruption of the asama volcano, central japan, inferred from ionospheric disturbances, *Geophys. Res. Lett.*, **33**, L17101.

Heki, K., 2021. Ionospheric disturbances related to earthquakes in ionospheric dynamics and applications, *Geophys. Monograph, 260, edited by C. Huang, G. Lu, Y. Zhang, and L. J. Paxton*, pp. 511–526.

Heki, K., Otsuka, Y., Choosakul, N., Hemmakorn, N., Komolmis, T., & Maruyama, T., 2006. Detection of ruptures of andaman fault segments in the 2004 great sumatra earthquake with coseismic ionospheric disturbances, *J. Geophys. Res.*, **111**, B09313.

Hibert, C., Mangeney, A., Grandjean, G., Baillard, C., Rivet, D., Shapiro, N. M., Satriano, C., Maggi,

A., Boissier, P., Ferrazzini, V., et al., 2014. Automated identification, location, and volume estimation of rockfalls at piton de la fournaise volcano, *Journal of Geophysical Research: Earth Surface*, **119**(5), 1082–1105.

Hofmann-Wellenhof, B., Lichtenegger, H., & Wasle, E., 2008. *GNSS-Global Navigation Satellite System*, Springer.

Inchin, P., Snively, J., Kaneko, Y., Z., D., M., & Komjathy, A., 2021. Inferring the evolution of a large earthquake from its acoustic impacts on the ionosphere., *AGU Advances*, **2**.

Kamogawa, M., Orihara, Y., Tsurudome, C., Tomida, Y., Kanaya, T., & Ikeda, D., e. a., 2016. A possible space-based tsunami early warning system using observations of the tsunami ionospheric hole, *Scientific Reports*, **6:37989**.

Katsumata, A., Ueno, H., Aoki, S., Yasushiro, Y., & Barrientos, S., 2013. Rapid magnitude determination from peak amplitudes at local stations, *Earth, Planets Space*, **65**, 843–853.

Komjathy, A., Yang, Y., Meng, X., Vekhoglyadova, O., Mannucci, A., & Langley, R., 2016. Review and perspectives: Understanding natural-hazards-generated ionospheric perturbations using gps measurements and coupled modeling, *Radio Science*, **51**, 951–961.

LaBrecque, J., Rundle, J., Bawden, G., Surface, E., & Area, I. F., 2019. Global navigation satellite system enhancement for tsunami early warning systems, *Global Assessment Report on Disaster Risk Reduction*.

Lee, R., Rolland, L., & Mykesell, T., 2018. Seismo-ionospheric observations, modeling and backprojection of the 2016 kaikoura earthquake, *Bulletin of the Seismological Society of America*, **108**(3B), 1794–1806.

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

Louppe, G., 2014. Understanding random forests: From theory to practice, *arXiv preprint arXiv:1407.7502*.

Manta, F., Occhipinti, G., Feng, L., & Hill, E., 2020. Rapid identification of tsunamigenic earthquakes using gnss ionospheric sounding, *Scientific Reports*, **10:11054**.

Mousavi, S. M., Zhu, W., Sheng, Y., & Beroza, G. C., 2019. Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection, *Scientific reports*, **9**(1), 1–14.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C., 2020. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking, *Nature communications*, **11**(1), 1–12.

Occhipinti, G., Aden-Antoniow, F., Bablet, A., Molinie, J.-P., & Farges, T., 2018. Surface waves magnitude estimation from ionospheric signature of rayleigh waves measured by doppler sounder and oth radar, *Scientific Reports*, **8:1555**.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer,

P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E., 2011. Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, **12**, 2825–2830.

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

Rakoto, V., Lognonné, P., Rolland, L., & Coisson, P., 2018. Tsunami wave height estimation from gps-derived ionospheric data, *J. Geophys. Res.*, **123**, 4329–4348.

Ravanelli, M., Occhipinti, G., Savastano, G., Komjathy, A., Shume, E. B., & Crespi, M., 2021. Gnss total variometric approach: first demonstration of a tool for real-time tsunami genesis estimation, *Scientific reports*, **11**(1), 1–12.

Rolland, L. M., Occhipinti, G., Lognonné, P., & Loevenbruck, A., 2016. Ionospheric gravity waves detected offshore hawaii after tsunami, *Geophys. Res. Lett.*, **37**, L17101.

Ross, Z. E., Meier, M.-A., & Hauksson, E., 2018. P wave arrival picking and first-motion polarity determination with deep learning, *Journal of Geophysical Research: Solid Earth*, **123**(6), 5120–5129.

Ross, Z. E., Idini, B., Jia, Z., Stephenson, O. L., Zhong, M., Wang, X., Zhan, Z., Simons, M., Fielding, E. J., Yun, S.-H., et al., 2019. Hierarchical interlocked orthogonal faulting in the 2019 ridgecrest earthquake sequence, *Science*, **366**(6463), 346–351.

Shults, K., Astafyeva, E., & Adourian, S., 2016. 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.

Solé, X., Ramisa, A., & Torras, C., 2014. Evaluation of random forests on large-scale classification problems using a bag-of-visual-words representation, in *CCIA*, pp. 273–276.

Tsai, H.-F., Liu, J.-Y., Lin, C.-H., & Chen, C.-H., 2011. Tracking the epicenter and the tsunami origin with gps ionosphere observation, *Earth, Planets Space*, **63**, 859–862.

Wenner, M., Hibert, C., van Herwijnen, A., Meier, L., & Walter, F., 2021. 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.

Wright, T., Houlie, N., Hildyard, M., & Iwabuchi, T., 2012. Real-time, reliable magnitudes for large earthquakes from 1 hz gps precise point positioning: The 2011 tohoku-oki (japan) earthquake, *Geophys. Res. Lett.*, **38**(L12302).

Zedek, F., Rolland, L. M., Dylan Mikesell, T., Sladen, A., Delouis, B., Twardzik, C., & Coïsson, P., 2021. Locating surface deformation induced by earthquakes using gps, glonass and galileo ionospheric sounding from a single station, *Advances in Space Research*.