The Transformer Earthquake Alerting Model: A New Versatile Approach to Earthquake Early Warning

Jannes Münchmeyer\textsuperscript{1,2,*}, Dino Bindi\textsuperscript{1}, Ulf Leser\textsuperscript{2}, Frederik Tilmann\textsuperscript{1,3}

\textsuperscript{1} Helmholtzzentrum Potsdam, Deutsches GeoForschungsZentrum GFZ, Potsdam, Germany
\textsuperscript{2} Institut für Informatik, Humboldt-Universität zu Berlin, Berlin, Germany
\textsuperscript{3} Institut für geologische Wissenschaften, Freie Universität Berlin, Berlin, Germany
\textsuperscript{*} To whom correspondence should be addressed: munchmej@gfz-potsdam.de

September 18, 2020

Abstract

Earthquake early warning aims to provide advance notice of incoming strong shaking to enable preventive action and mitigate seismic risk. Its usefulness depends on accuracy, the relation between true, missed and false alerts, and timeliness, the time between a warning and the arrival of strong shaking. Here we propose a novel early warning method, the deep-learning based transformer earthquake alerting model (TEAM). TEAM analyzes raw, strong motion waveforms of an arbitrary number of stations at arbitrary locations in real-time, making it easily adaptable to changing seismic networks. For two regions with high seismic hazard, Japan and Italy, it outperforms existing early warning methods considerably, offering accurate and timely warnings. Using domain adaptation, TEAM even provides reliable alerts for events larger than any in the training data, a property of highest importance as records from very large events are rare in many regions.

1 Introduction

The concept of earthquake early warning has been around for over a century, but the necessary instrumentation and methodologies have only been developed in the last three decades \cite{1, 2}. Early warning systems aim to raise alerts if shaking levels likely to cause damage are going to occur. Existing methods split into two main classes: source estimation based and propagation based. The former, like EPIC \cite{3} or FINDER \cite{4}, estimate the source properties of an event, i.e., its location or fault extent and magnitude, and then use a ground motion prediction equation (GMPE) to infer shaking at target sites. They provide long warning times, but incur a large aleatoric uncertainty due to simplified assumptions in the source estimation and in the GMPE \cite{5}. Propagation based methods, like PLUM \cite{5}, infer the shaking at a given location from measurements at nearby seismic stations. Predictions are more accurate, but warning times are reduced, as warnings require measurements of strong shaking at nearby stations \cite{6}.

Recently, machine learning methods, particularly deep learning methods, have emerged as a tool for fast assessment of earthquakes. Under certain circumstances, they led to improvements in various tasks, e.g., estimation of magnitude \cite{7, 8}, location \cite{9, 10} or peak ground acceleration (PGA) \cite{11}. Nonetheless, no existing method is applicable to early warning because they lack real-time capabilities, instead requiring fixed waveform windows after the P arrival. Furthermore, many of these approaches are designed for single station input, missing out on the potential of jointly using multiple stations across the network. Those methods combining data from multiple stations assume a fixed station set determined at training time, limiting their adaptability to changing networks. Finally, existing methods systematically underestimate the strongest shaking and the highest magnitudes, as these are rare and therefore underrepresented in the training data (Fig. 6, 8 in \cite{11}, Fig. 3, 4 in \cite{8}). However, early warning systems must also be able to provide reliable warnings for earthquakes larger than any previously seen in a region.

Here, we present the transformer earthquake alerting model (TEAM), a deep learning method for early warning, combining the advantages of both classical early warning strategies while avoiding the deficiencies of prior deep
learning approaches. We evaluate TEAM on two data sets from regions with high seismic hazard, namely Japan and Italy. Due to their complementary seismicity, this allows to evaluate the capabilities of TEAM across scenarios. We compare TEAM to two state-of-the-art warning methods, of which one is prototypical for source based warning and one for propagation based warning.

2 Data and Methods

2.1 Data

For our study we use two nation scale datasets from highly seismically active regions with dense seismic networks, namely Japan (13,512 events, years 1997-2018, Figure S1) and Italy (7,055 events, years 2008-2019, Figure S2). Their seismicity is complementary, with predominantly subduction plate interface or Wadati-Benioff zone events for Japan, many of them offshore, and shallow, crustal events for Italy. We split both datasets into training, development and test sets with ratios of 60:10:30. We use the training set for model training, the development set for model selection, and the test set only for the final evaluation. We split the Japan dataset chronologically, yielding the events between August 2013 and December 2018 as test set. For Italy, we test on all events in 2016, as these are of particular interest, encompassing most of the central Italy sequence with the \( M_W = 6.2 \) and \( M_w = 6.5 \) Norcia events [12]. Especially the latter event is notably larger than any in the training set (\( M_w = 6.1 \) L'Aquila event in 2007), thereby challenging the extrapolation capabilities of TEAM.

Both datasets consist of strong motion waveforms. For Japan each station comprises two sensors, one at the surface and one borehole sensor, while for Italy only surface recordings are available. As the instrument response in the frequency band of interest is flat, we do not restitute the waveforms, but only apply a gain correction. This has the advantage that it can trivially be done in real-time. The data and preprocessing are further described in appendix A.

2.2 The transformer earthquake alerting model

The early warning workflow with TEAM encompasses three separate steps (Figure 1): event detection, PGA estimation and thresholding. We do not further consider the event detection task here, as it forms the base of all methods discussed and affects them similarly. The PGA estimation, resulting in PGA probability densities for a given set of target locations, is the heart of TEAM and described in detail below. In the last step, thresholding, TEAM issues warnings at all target locations where the predicted exceedance probability \( p \) for fixed PGA thresholds surpasses a predefined probability \( \alpha \).

The PGA assessment with TEAM is again subdivided into three components: feature extraction, feature combination, and PGA estimation (Figure S3). We present the key ideas for each of these components and their combination in the following. A complete description of the TEAM architecture and training procedure can be found in appendix B. For feature extraction TEAM uses a convolutional neural network (CNN) over the currently recorded waveforms at all input stations. CNNs are well established for feature extraction from seismic waveforms, as they are able to recognize complex features independent of their position in the trace. On the other hand, CNN based feature extraction usually requires a fixed input length, inhibiting real-time processing. We allow real-time processing by zero-padding the waveforms to 30 s length, thereby maintaining the constant input length. We combine the extracted waveform features with sinusoidal vector representations of the station locations, yielding one vector per station. A transformer [13] uses these vectors and the target locations to generate one vector per target location. Transformers are attention-based neural networks for combining information from a flexible number of input vectors in a learnable way. This architecture, processing a varying number of inputs, together with the explicitly encoded locations, allows TEAM to handle varying sets of stations and targets. The vectors resulting from the transformer, each representing predictions at one target, are fed into a mixture density network [14], which computes PGA densities.

To mitigate the systematic underestimation of high PGA values observed in previous machine learning models, TEAM oversamples large events and PGA targets close to the epicenter during training, which reduces the inherent bias in data towards smaller PGAs. When learning from small catalogs or when applied to regions where events substantially larger than all training events can be expected, e.g., because of known locked fault patches or historic records, TEAM additionally can use domain adaptation. To this end the training procedure is modified to include events from other regions, that are similar to the expected events in the target region. While this measure, to some degree, blurs geological properties of the target region, it provides realistic waveforms and PGAs of large events with similar characteristics, leading to models performing better for large events.

Our Italy dataset is an example of this situation. Accordingly, TEAM applies domain adaptation to this case: It first trains a joint model using data from Japan and from Italy, which is then fine-tuned using the Italy data on its own, except for the addition of a few large, shallow, onshore events from Japan. We chose these events, as for Italy one also expects large, shallow, crustal events due to its tectonic setting and earthquake history.
Figure 1: Schematic view of TEAM’s early warning workflow for the October 2016 Norcia event ($M_w = 6.5$) 2.5 s after the first P wave pick (≈3.5 s after origin time). a. An event is detected through triggering at multiple seismic stations. The waveform colors correspond to the stations highlighted with orange to magenta outlines. The circles indicate the approximate current position of P (dashed) and S (solid) wavefronts. b. TEAM’s input are raw waveforms and station coordinates; it estimates probability densities for the PGA at a target set. A more detailed TEAM overview is given in Figure S3. c. The exceedance probabilities for a fixed set of PGA thresholds are calculated based on the estimated PGA probability densities. If the probability exceeds a threshold $\alpha$, a warning is issued. The figure visualizes a 10%g PGA level with $\alpha = 0.4$, resulting in warnings for the stations highlighted. The colors correspond to the stations with green outlines in a. d. The real-time shake map shows the highest PGA levels for which a warning is issued. Stations are colored according to their current warning level. The table lists all stations for which warnings have already been issued.

2.3 Baselines

We compare TEAM to two state-of-the-art early warning methods, one using source estimation and one propagation based. As source estimation based method we use the estimated point source approach (EPS), which estimates magnitudes from peak displacement during the P-wave onset [15] and then applies a GMPE [16] to predict the PGA. For simplicity, our implementation assumes knowledge of the final catalog epicentre, which is impossible in real-time, leading to overoptimistic results for EPS. As propagation based method we chose PLUM [5], which issues warnings if a station within a radius $r$ of the target exceeds the level of shaking. Additionally we apply the GMPE used in EPS to catalog location and magnitude as an approximate upper accuracy bound for point source algorithms (Catalog-GMPE or C-GMPE). C-CMPE is a theoretical bound that can not be realized in real-time. It can be considered as an estimate of the modeling error for point source approaches. A detailed description of the baseline methods can be found in appendix C.

3 Results

3.1 Alert performance

We compare the alert performance of all methods for PGA thresholds from light (1%g) to very strong (20%g) shaking, regarding precision, the fraction of alerts actually exceeding the PGA threshold, and recall, the fraction of issued alerts among all cases where the PGA threshold was exceeded [17, 18]. As precision and recall trade-off against each other
Figure 2: Warning statistics for the three early-warning models (TEAM, EPS, PLUM) for the Japan and Italy datasets. In addition, statistics are provided for C-GMPE, which can only be evaluated post-event due to its reliance on catalog magnitude and location. **a.** Precision and recall curves across different thresholds $\alpha = 0.05, 0.1, 0.2, \ldots, 0.8, 0.9, 0.95$. As PLUM has no tuning parameter, its performance is shown as a point. Enlarged markers show the configurations yielding the highest F1 scores. Numbers in the corner give the area under the precision recall curve (AUC), a standard measure quantifying the predictive performance across thresholds. **b.** Precision, recall and F1 score at different PGA thresholds using the F1 optimal value $\alpha$. Threshold probabilities $\alpha$ were chosen independently for each method and PGA threshold. **c.** Number of events and traces exceeding each PGA threshold for training and test set. Training set numbers include development events and show the numbers before oversampling is applied. For Italy training and test event curve are overlapping due to similar numbers of events.

depending on $\alpha$, we additionally use the summary metric $F1$ score, the harmonic mean of precision and recall. The evaluation metrics and full setup of the evaluation are defined in appendix D. Additionally, we discuss the calibration of the predicted probabilities from TEAM and EPS in appendix E.

TEAM outperforms both EPS and PLUM for both datasets and all PGA thresholds, indicated by the precision-recall-curves of TEAM lying to the top-right of the baseline curves (Figure 2a). For any baseline method configuration, there is a TEAM configuration surpassing it both in precision and in recall. Improvements are larger for Japan, but still substantial for Italy. Next, we selected $\alpha$ values yielding the highest F1 score separately for each PGA threshold and method. Again, TEAM outperforms both baselines significantly on both datasets, irrespective of the PGA level (Figure 2b). Further performance statistics are available in tables S1 and S2.

All methods degrade with increasing PGA levels, particularly for Japan. This degradation is intrinsic to early warning for high thresholds due to the very low prior probability of strong shaking $[6,18,17]$. Furthermore, shortage of training data with high PGA values results in less well constrained model parameters.
Figure 3: Warning time statistics. **a.** Area under the precision recall curve for different minimum warning times. All alerts with shorter warning times are counted as false negatives. **b.** Warning time histograms for true alerts at F1-optimal $\alpha$, chosen separately for each threshold and method. Warning time dependence on hypocentral distance is shown in Figure S6.

Using domain adaptation techniques, TEAM copes well with the Italy data, even though the largest test event ($M_w = 6.5$) is significantly larger than the largest training event ($M_w = 6.1$), and three further test events have $M_W \geq 5.8$. To assess the impact of this technique, we compared TEAM's results to a model trained without it (Figures S4, S5). While for low PGA thresholds differences are small, at high PGA levels they grow to more than 20 points F1 score. For large events, TEAM strongly outperforms TEAM without domain adaptation even for low PGA thresholds. This shows that domain adaptation does not only allow the model to predict higher PGA values, but also to accurately assess the region of lighter shaking for large events.

### 3.2 Warning times

The previous evaluation considered prediction accuracy irrespective of the warning time, which is the time between issuing of the warning and first exceedance of the level of shaking. Actually, TEAM consistently outperforms both baselines across different required warning times and irrespective of the PGA threshold (Figure 3b). In particular PLUM warning times are short and only competitive at high PGA thresholds, where potential maximum warning times are naturally short due to the proximity between stations with strong shaking and the epicenter [19].

Warning times depend on $\alpha$: a lower $\alpha$ value naturally leads to longer warning times but also to more false positive warnings. At F1-optimal thresholds $\alpha$, EPS and TEAM have similar warning time distributions (Figure 3b, Table S3), but lowering $\alpha$ leads to stronger increases in warning times for TEAM. For instance, at 10%g, lowering $\alpha$ from 0.5 to 0.2 increases average warning times of TEAM by 2.3 s/1.2 s (Japan/Italy), but only by 1.1 s/0.1 s for EPS. Short times as measured here are critical in real applications: First, they reduce the time available for counter measures. Second, real warning times will be shorter than reported here due to telemetry and compute delays. However, compute delays
Figure 4: Scenario analysis of the 22nd November 2016 $M_J = 7.4$ Fukushima earthquake, the largest test event located close to shore. Maps show the warning levels for each method (top three rows) at different times (shown in the corner, $t = 0$ s corresponds to P arrival at closest station). Dots represent stations and are colored according to the PGA recorded during the full event, i.e., the prediction target. The bottom row shows (left to right), the catalog based GMPE predictions, the warning time distributions, and the true positives (TP), false negatives (FN) and false positives (FP) for each method, both at a 2%g PGA threshold. EPS and GMPE shake map predictions do not include station terms, but they are included for the bottom row histograms.

for TEAM are very mild: analysing the Norcia event (25 input stations, 246 target sites) for one time step took only 0.15 s on a standard workstation using non-optimized code.
4 Discussion

We analyze differences between the methods using one example event from each dataset (Japan: Figure 4, Italy: Figure S7). All methods underestimate the shaking in the first seconds (left column Figures 4, S7). However, TEAM is the quickest to detect the correct extent of the shaking. Additionally, it estimates even fine-grained regional shaking details in real-time (middle and right columns). In contrast, shake maps for EPS remain overly simplified due to the assumptions inherent to GMPEs (right column and bottom left panel). PLUM produces very good PGA estimates, but exhibits the worst warning times.

Notably, TEAM predictions at later times correspond even better to the measured PGA than C-GMPE estimates, although these are based on the final magnitude (top right and bottom left panels). For the Japan data, this is also visible in Figure 2 showing higher accuracy of TEAM’s prediction compared to C-GMPE for all thresholds except 20%g.

We assume TEAM’s superior performance is rooted in both global and local aspects. Global aspects are the abilities to exploit variations in the waveforms, e.g., frequency content, to model complex event characteristics, such as stress drop, radiation pattern or directivity, and to compare to events in the training set. Local aspects include understanding regional effects, e.g., frequency dependent site responses, and the ability to consider shaking at proximal stations. Thereby, combining a global event view with propagation aspects, TEAM can be seen as a hybrid model between source estimation and propagation.

5 Conclusion

In this study we presented the transformer earthquake alerting model (TEAM). TEAM outperforms existing early warning methods in terms of both alert performance and warning time. Using a flexible machine learning model, TEAM is able to extract information about an event from raw waveforms and leverage the information to model the complex dependencies of ground motion. We point out two further aspects that make TEAM appealing to users. First, TEAM can adapt to various user requirements by combining two thresholds, one for shake level and one for the exceedance probability. As TEAM outputs probability density functions over the PGA, these thresholds can easily be adjusted by individual users on the fly, e.g., by setting sliders in an early warning system. Second, deep learning models typically exhibit large performance improvements from larger training datasets due to the high number of model parameters. In our study this reflects in the better performance on the twofold larger Japan dataset. This indicates that TEAM’s performance can be improved just by collecting more comprehensive catalogs, which happens automatically over time.

Acknowledgements

We thank the National Research Institute for Earth Science and Disaster Resilience (NIED) for providing the catalog and waveform data for our Japan dataset. We thank the Istituto Nazionale di Geofisica e Vulcanologia and the Dipartimento della Protezione Civile for providing the catalog and waveform data for our Italy dataset. J. M. acknowledges the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS). We thank Matteo Picozzi for discussions on earthquake early warning that helped improve the study design. Publications of the code and datasets used in this study are under preparation.

References

[1] R. M. Allen, P. Gasparini, O. Kamigaichi, and M. Bose. The Status of Earthquake Early Warning around the World: An Introductory Overview. Seismological Research Letters, 80(5):682–693, September 2009.
[2] Richard M. Allen and Diego Melgar. Earthquake Early Warning: Advances, Scientific Challenges, and Societal Needs. Annual Review of Earth and Planetary Sciences, 47(1):361–388, 2019.
[3] Angela I. Chung, Ivan Henson, and Richard M. Allen. Optimizing Earthquake Early Warning Performance: ElarmS-3. Seismological Research Letters, 90(2A):727–743, March 2019.
[4] M. Böse, D. E. Smith, C. Felizardo, M.-A. Meier, T. H. Heaton, and J. F. Clinton. FinDer v.2: Improved real-time ground-motion predictions for M2–M9 with seismic finite-source characterization. Geophysical Journal International, 212(1):725–742, January 2018.
[5] Yuki Kodera, Yasuyuki Yamada, Kazuyuki Hirano, Koji Tamaribuchi, Shimpei Adachi, Naoki Hayashimoto, Masahiko Morimoto, Masaki Nakamura, and Mitsuyuki Hoshiba. The Propagation of Local Undamped Motion (PLUM) Method: A Simple and Robust Seismic Wavefield Estimation Approach for Earthquake Early Warning The Propagation of Local Undamped Motion (PLUM) Method. Bulletin of the Seismological Society of America, 108(2):983–1003, April 2018.
[6] Men-Andrin Meier, Yuki Kodera, Maren Böse, Angela Chung, Mitsuyuki Hoshiba, Elizabeth Cochran, Sarah Minson, Egill Hauksson, and Thomas Heaton. How Often Can Earthquake Early Warning Systems Alert Sites With High-Intensity Ground Motion? *Journal of Geophysical Research: Solid Earth*, 125(2):e2019JB017718, 2020.

[7] Anthony Lomax, Alberto Michelini, and Dario Jozinović. An Investigation of Rapid Earthquake Characterization Using Single-Station Waveforms and a Convolutional Neural Network. *Seismological Research Letters*, 90(2A):517–529, 2019.

[8] S. Mostafa Mousavi and Gregory C. Beroza. A Machine-Learning Approach for Earthquake Magnitude Estimation. *Geophysical Research Letters*, 47(1):e2019GL085976, 2020.

[9] Marius Kriegerowski, Gesa M. Petersen, Hannes Vasyura-Bathke, and Matthias Ohrnberger. A Deep Convolutional Neural Network for Localization of Clustered Earthquakes Based on Multistation Full Waveforms. *Seismological Research Letters*, 90(2A):510–516, 2019.

[10] S. Mostafa Mousavi and Gregory C. Beroza. Bayesian-Deep-Learning Estimation of Earthquake Location from Single-Station Observations. *arXiv:1912.01144 [physics]*, December 2019.

[11] Dario Jozinović, Anthony Lomax, Ivan Štajduhar, and Alberto Michelini. Rapid prediction of earthquake ground shaking intensity using raw waveform data and a convolutional neural network. *Geophysical Journal International*, 222(2):1379–1389, 2020.

[12] Mauro Dolce and Daniela Di Bucci. The 2016–2017 Central Apennines Seismic Sequence: Analogies and Differences with Recent Italian Earthquakes. In Kyriazis Pitilakis, editor, *Recent Advances in Earthquake Engineering in Europe: 16th European Conference on Earthquake Engineering-Thessaloniki 2018*, Geotechnical, Geological and Earthquake Engineering, pages 603–638. Springer International Publishing, Cham, 2018.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[14] Christopher M Bishop. Mixture density networks. Technical report, Aston University, 1994.

[15] Huseyin Serdar Kuyuk and Richard M. Allen. A global approach to provide magnitude estimates for earthquake early warning alerts. *Geophysical Research Letters*, 40(24):6329–6333, 2013.

[16] Georgia Cua and Thomas H. Heaton. Characterizing Average Properties of Southern California Ground Motion Amplitudes and Envelopes. EERL Report, Earthquake Engineering Research Laboratory, Pasadena, CA, 2009.

[17] Men-Andrin Meier. How “good” are real-time ground motion predictions from Earthquake Early Warning systems? *Journal of Geophysical Research: Solid Earth*, 122(7):5561–5577, 2017.

[18] Sarah E. Minson, Annemarie S. Baltay, Elizabeth S. Cochran, Thomas C. Hanks, Morgan T. Page, Sara K. McBride, Kevin R. Milner, and Men-Andrin Meier. The Limits of Earthquake Early Warning Accuracy and Best Alerting Strategy. *Scientific Reports*, 9(1):2478, February 2019.

[19] Sarah E. Minson, Men-Andrin Meier, Annemarie S. Baltay, Thomas C. Hanks, and Elizabeth S. Cochran. The limits of earthquake early warning: Timeliness of ground motion estimates. *Science Advances*, 4(3):eaao0504, March 2018.

[20] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[21] Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D Sculley, Joshua Dillon, Jie Ren, and Zachary Nado. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pages 13969–13980, 2019.

[22] Kazi R. Karim and Fumio Yamazaki. Correlation of JMA instrumental seismic intensity with strong motion parameters. *Earthquake Engineering & Structural Dynamics*, 31(5):1191–1212, 2002.

[23] Elizabeth S. Cochran, Julian Bunn, Sarah E. Minson, Annemarie S. Baltay, Deborah L. Kilb, Yuki Kodera, and Mitsuyuki Hoshiba. Event Detection Performance of the PLUM Earthquake Early Warning Algorithm in Southern California. *Bulletin of the Seismological Society of America*, 109(4):1524–1541, August 2019.

[24] National Research Institute For Earth Science And Disaster Resilience. Nied k-net, kik-net, 2019.

[25] Istituto Nazionale di Geofisica e Vulcanologia (INGV), Istituto di Geologia Ambientale e Geoingegneria (CNR-IGAG), Istituto per la Dinamica dei Processi Ambientali (CNR-IDPA), Istituto di Metodologie per l’Analisi Ambientale (CNR-IMAA), and Agenzia Nazionale per le nuove tecnologie, l’energia e lo sviluppo economico
sostenibile (ENEA). Centro di microzonazione sismica network, 2016 central italy seismic sequence (centromz), 2018.

[26] Universita della Basilicata. Unibas, 2005.

[27] RESIF - Réseau Sismologique et géodésique Français. Resif-rlbp french broad-band network, resif-rap strong motion network and other seismic stations in metropolitan france, 1995.

[28] University of Genova. Regional seismic network of north western italy. international federation of digital seismograph networks, 1967.

[29] Presidency of Council of Ministers - Civil Protection Department. Italian strong motion network (ran), 1972.

[30] Istituto Nazionale di Geofisica e Vulcanologia (INGV), Italy. Rete sismica nazionale (rsn), 2006.

[31] Dipartimento di Fisica, Università degli studi di Napoli Federico II. Irpinia seismic network (isnet), 2005.

[32] MedNet Project Partner Institutions. Mediterranean very broadband seismographic network (mednet), 1990.

[33] OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale) and University of Trieste. North-east italy broadband network (ni), 2002.

[34] OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale). North-east italy seismic network (nei), 2016.

[35] RESIF - Réseau Sismologique et géodésique Français. Réseau accélérométrique permanent (french ac-celerometrique network) (rap), 1995.

[36] Geological Survey-Provincia Autonoma di Trento. Trentino seismic network, 1981.

[37] Istituto Nazionale di Geofisica e Vulcanologia (INGV). Ingv experiments network, 2008.

[38] EMERSITO Working Group. Seismic network for site effect studies in amatrice area (central italy) (sesaa), 2018.

[39] David J. Wald, Vincent Quitoriano, Thomas H. Heaton, and Hiroo Kanamori. Relationships between Peak Ground Acceleration, Peak Ground Velocity, and Modified Mercalli Intensity in California. *Earthquake Spectra*, 15(3):557–564, August 1999.
A Data and Preprocessing

For our study we use two datasets, one from Japan, one from Italy. The Japan dataset consists of 13,512 events between 1997 and 2018 from the NIED KiK-net catalog [24]. The data was obtained from NIED and consists of triggered strong motion records. Each trace contains 15 s of data before the trigger and has a total length of 120 s. Each station consists of two component strong motion sensors, one at the surface and one borehole sensor. We split the dataset chronologically with ratios of 60:10:30 between training, development and test set. The training set ends in March 2012, the test set begins in August 2013. Events in between are used as development set. We decided to use a chronological split to ensure a scenario most similar to the actual application in an early warning setting.

The Italy dataset consists of 7,055 events between 2008 and 2019 from the INGV catalog. We use data from the 3A [25], BA [26], FR [27], GU [28], IT [29], IV [30], IX [31], MN [32], NI [33], OX [34], RA [35], ST [36], TV [37] and XO [38] networks. We use all events from 2016 as test set and the remaining events as training and development sets. The test set consists of 31% of the events, a similar fraction as in the Japan dataset. We shuffle events between training and development set. While a chronological split would have been the default choice, we decided to use 2016 for testing, as it contains a long seismic sequence in central Italy containing several very large events in August and October. Further details on the statistics of both datasets can be found in Table S4.

Before training we extract, align and preprocess the waveforms and store them in hdf5 format. As alignment requires the first P pick, we need approximate picks for the datasets. For Japan we use the trigger times provided by NIED. Our preprocessing accounts for misassociated triggers. For Italy we use an STA/LTA trigger around the predicted P arrival. While triggering needs to be handled differently in an application scenario, we use this simplified approach as our evaluation metrics depend only very weakly on the precision of the picks.

B TEAM - Architecture and training details

TEAM conducts end-to-end PGA estimation. Input to our model are 3 component waveforms at 100 Hz sampling rate from multiple stations and the corresponding station coordinates. In case of available borehole sensors, we treat the station as having 6 components (3 surface, 3 borehole). In addition the model is provided with a set of output locations, at which the PGA should be predicted. These can be anywhere within the spatial domain of the model and need not be identical with station locations in the training set.

The input waveforms from all stations are aligned in time, i.e., all start at the same time $t_0$ and end at the same time $t_1$. We define $t_0$ to be 5 s before the first P wave arrival at any station. This guarantees that each waveform starts with at least 5 s of noise, allowing the model to understand the current noise conditions at a station. In a real-time scenario $t_1$ is the current time, i.e., limited by the waveforms available up to that point in time. We limit $t_1$ to be at most $t_0 + 30$ s. To obtain a constant length input for TEAM, we pad all waveforms after $t_1$ with zeros up to a total length of 30 s.

The PGA estimation consists of three main components, shown in yellow in Figure S3. The first component is a station-wise feature extraction using a convolutional network (CNN). The second component is a feature combination that combines the information from the different stations based on their location and calculates feature representations for the target locations. The last step is a mixture density estimation calculating probabilistic estimates for the PGA at the target locations. The model has about 13.3 million parameters in total. All components are trained jointly end-to-end.

B.1 Feature extraction network

The feature extraction is conducted separately for each station. Nonetheless the same convolutional neural network (CNN) for feature extraction is applied at all stations, i.e., the same model with the same model weights.

As amplitudes of seismic waveforms can span several orders of magnitude, the first layer of the network normalizes the traces by dividing through their peak. All components of one station are normalized jointly, such that the amplitude ratio between the components stays unaltered. As the peak amplitude of the trace is a key predictor, we logarithmize the value and concatenate it to the feature vector after passing through all the convolutional layers, prior to the fully connected layers.

We apply a set of convolutional and max-pooling layers to the waveforms. We use convolutional layers as this allows the model to extract translation invariant features and as convolutional kernels can be interpreted as modeling frequency features. We concatenate the output of the convolutions and the logarithm of the peak amplitude. This vector is fed into a multi-layer perceptron to generate the final features vector for the station. All layers use ReLu activations. A detailed overview of the number and specifications of the layers in the feature extraction model can be found in Table S5.
B.2 Feature combination network

The feature extraction provides one feature vector per input station representing the waveforms. As additional input the model is provided with the location of the stations, represented by latitude, longitude and elevation. The targets for the PGA estimation are specified by the latitude, longitude and elevation as well.

We use a transformer network \[13\] for the feature combination. Given a set of \( n \) input vectors, a transformer produces \( n \) output vectors capturing combined information from all the vectors in a learnable way. We use transformers for two main reasons. First, they are permutation equivariant, i.e., changing the order of input or output stations does not have any impact on the output. This is essential, as there exists no natural ordering on the input stations. Second, they can handle variable input sizes, as the number of parameters of transformers is independent of the number of input vectors. This property allows application of the model to different sets of stations and a flexible number of target locations.

To incorporate the locations of the stations we use predefined position embeddings. As proposed in \[13\], we use pairs of sinusoidal functions, \( \sin(\frac{2\pi}{L} x) \) and \( \cos(\frac{2\pi}{L} x) \), with different wavelength \( \lambda_i \). We use 200 dimensions for latitude and longitude, respectively, and the remaining 100 dimensions for elevation. We anticipate two advantages of sinusoidal embeddings for representing the station position. First, keeping the position embeddings fixed instead of learnable reduces the parameters and therefore likely provides better representations for stations with only few input measurements or sites not contained in the training set. Second, sinusoidal embeddings guarantee that shifts can be represented by linear transformations, independent of the location it applies to. As the attention mechanism in transformers is built on linear projections and dot products, this should allow for more efficient attention scores at least in the first transformer layers. As proposed in the original transformer paper \[13\], the position embeddings are added element-wise to the feature vectors to form the input of the transformer. We calculate position embeddings of the target locations in the same way.

As in our model input and output size of the transformer are identical, we only use the transformer encoder stack \[13\] with six encoder layers. Inputs are the feature vectors with position embeddings from all input stations and the position embeddings of the output locations. By applying masking to the attention we ensure that no attention weight is put on the vectors corresponding to the output locations. This guarantees that each target only affects its own PGA value and not any other PGA values. As the self-attention mechanism of the transformer has quadratic computational complexity in the number of inputs, we restrict the maximum number of input stations to 25 (see training details for the selection procedure). Further details on the hyperparameters can be found in Table 6. The transformer returns one output vector for each input vector. We discard the vectors corresponding to the input stations and only keep the vectors corresponding to the targets.

B.3 Mixture density output

Similar to the feature extraction, the output calculation is conducted separately for each target, while sharing the same model and weights between all targets. We use a mixture density network to predict probability densities for the PGA \[14\]. We model the probability as a mixture of \( m = 5 \) Gaussian random variables. Using a mixture of Gaussians instead of a single Gaussian allows the model to predict more complex distribution, like non-Gaussian distributions, e.g., asymmetric distributions. The functional form of the Gaussian mixture is \[ \sum_{i=1}^{m} \alpha_i \varphi_{\mu_i, \sigma_i}(x) \]. We write \( \varphi_{\mu_i, \sigma_i} \) for the density of a standard normal with mean \( \mu_i \) and standard deviation \( \sigma_i \). The values \( \alpha_i \) are non-negative weights for the different Gaussians with the property \( \sum_{i=1}^{m} \alpha_i = 1 \). The mixture density network uses a multi-layer perceptron to predict the parameters \( \alpha_i, \mu_i \) and \( \sigma_i \). The hidden dimensions are 150, 100, 50, 30, 10. The activation function is ReLu for the hidden layers, linear for the \( \mu \) and \( \sigma \) outputs and softmax for the \( \alpha \) output.

B.4 Training details

We train the model end-to-end using negative log-likelihood as loss function. To increase the amount of training data and to train the model on shorter segments of data we apply various forms of data augmentation. Each data augmentation is calculated separately each time a particular waveform sample is shown, such that the effective training samples vary.

First, if our dataset contains more stations for an event than the maximum number of 25 allowed by the model, we subsample. We introduce a bias to the subsampling to favor stations closer to the event. We use up to twenty targets for PGA prediction. Similarly to the input station, we subsample if more targets are available and bias the subsampling to stations close to the event. This bias ensures that targets with higher PGA values are shown more often during training.

Second, we apply station blinding, meaning we zero out a set of stations in terms of both waveforms and coordinates. The number of stations to blind is uniformly distributed between zero and the total number of stations available minus one. In combination with the first point this guarantees that the model also learns to predict PGA values at sites where no waveform inputs are available.

11
Third, we apply temporal blinding. We uniformly select a time $t$ that is between 1 s before the first P pick and 25 s after. All waveforms are set to zero after time $t$. The model therefore only uses data available at time $t$. Even though we never apply TEAM to times before the first P pick, we include these in the training process to ensure TEAM learns a sensible prior distribution. We observed that this leads to better early predictions. As information about the triggering station distribution would introduce a knowledge leak if available from the beginning, we zero out all waveforms and coordinates from stations that did not trigger until time $t$.

Fourth, we oversample large magnitude events. As large magnitude events are rare, we artificially increase their number in the training set. An event with magnitude $M \geq M_0$ is used $\lambda M - M_0$ times in each training epoch with $\lambda = 1.5$ and $M_0 = 5$ for Japan and $M_0 = 4$ for Italy. This event-based oversampling implicitly increases the number of high PGA values in the training set, too.

We apply all data augmentation on the training and the development set, to ensure that the development set properly represents the task we are interested in. As this introduces stochasticity into the development set metrics, we evaluate the development set three times after each epoch and average the result. In contrast, at test time we do not apply any data augmentation, except temporal blinding for modelling real-time application. If more than 25 stations are available for a test set event, we select the 25 station with the earliest arrivals for evaluation.

We train our model using the Adam optimizer. We emphasize that the model is only trained on predicting the PGA probability density and does not use any information on the PGA thresholds used for evaluation. We start with a learning rate of $10^{-3}$ and decrease the learning rate by a factor of 3 after 5 epochs without a decrease in validation loss. For the final evaluation we use the model from the epoch with lowest loss on the development set. We apply gradient clipping with a value of 1.0. We use a batch size of 64. We train the model for 100 epochs.

To improve the calibration of the predicted probability densities we use ensembles [21]. We use an ensemble size of 10 models and average the predicted probability densities. We weight each ensemble member identically. To increase the entropy between the ensembles, we also modify the position encodings between the ensemble members by rotating the latitude and longitude values of stations and targets. The rotations for the 10 ensemble members are $0^\circ, 5^\circ, \ldots, 40^\circ, 45^\circ$.

For the Italy model we use domain adaptation by modifying the training procedure. We first train a model jointly on the Italy and Japan data set, according to the configuration described above. We use the resulting model weights as initialization for the Italy model. For this training we reduce the number of PGA targets to 4, leading to a higher fraction of high PGA values in the training data, and the learning rate to $10^{-5}$. In addition, we train jointly on an auxiliary data set, comprised of 77 events from Japan. The events were chosen to be shallow, crustal and onshore, having a magnitude between 5.0 and 7.2. We shift the coordinates of the stations to lie in Italy. We use 85% of the auxiliary events in the training set and 15% in the development set.

We implemented the model using Tensorflow. We trained each model on one GeForce RTX 2080 Ti or Tesla V100. Training of a single model takes approximately 5 h for the Japan dataset, 10 h for the joint model and 1 h for the Italy data set. We benchmarked the inference performance of TEAM on a common workstation with GPU acceleration (Intel i7-7700, Nvidia Quadro P2000). Running TEAM with ensembling at a single timestep took 0.15 s for all 246 PGA targets of the Norcia event. As our implementation is not optimized for run time, we expect an optimized implementation to yield multifold lower run times, enabling a real-time application of TEAM with high update rate and low compute latency.

C Baseline methods

We compare TEAM to two baseline methods, EPS and PLUM. We do not compare to any deep learning baseline, because we are not aware of any published deep learning method for early warning that can actually be applied in real-time. For the EPS method we use a GMPE based on the functional form by Cua and Heaton [16] and add a quadratic magnitude term as proposed by Meier [17]. We make further minor adjustments to accommodate the wider range of magnitudes in our datasets. The functional form of the GMPE is:

$$\log(pga) = a_1 M + a_2 \max(M - M_0, 0)^2 + b(R_d + C(M)) + d \log(R_d + C(M)) + e + \delta_S + \mathcal{N}(0, \sigma^2)$$

$$C(M) := c_1 \exp(c_2 \max(0, M - 5))(\arctan(M - 5) + \pi/2)$$

$$R_d := \sqrt{R^2 + H_d^2}$$

We write $M$ for magnitude, $R$ for epicentral distance, $\delta_S$ for the station bias, and $e$ for an error term. We use $m/s^2$ as unit for PGA and $km$ as unit for all length measurements. We use a pseudo-depth $H_d$, depending on the event depth and the dataset. This allows to model the stronger attenuation with distance for shallow events. For Italy we set
We analyze the performance of the early warning algorithms using PGA thresholds of 1\%g, 2\%g, 5\%g, 10\%g and 20\%g to account for a few very deep events. We set $M_0 = 4$ for Italy and $M_0 = 6$ for Japan. We fix $c_1 = 1.48$ and $c_2 = 1.11$, as proposed by Cua and Heaton [16], and optimize the other parameters using linear regression. We perform the optimization iteratively to obtain station bias terms, using the union of training and development set. To avoid noise samples in calibration we only use stations for which $R_d < (M - 3) \times 50$ km for Japan and $R_d < (M - 3) \times 200$ km for Italy. The calibrated GMPEs have residual values $\sigma$ of 0.29 for Italy and 0.33 for Japan, matching the value of $\sim 0.3$ proposed as the approximate current optimum for GMPEs [13]. Residual plots can be found in Figure [S].

We note that our GMPE model is using a point source assumption, which is incorrect for larger events. We chose this simplification, as it is common in source based early warning and makes the GMPE performance an upper bound for any method relying on magnitude and location estimate. While there are early warning methods based on extended fault models [4], they perform equally well as point source approaches for all but the largest events [6]. As lower thresholds are dominated by smaller events, for which the point source approximation is valid, the inferior performance of the GMPE compared to TEAM is not an artifact of the point source assumption, but probably related to its inability to account for systematic propagation effects caused by regional structure, and variability of the earthquake source (focal mechanism, stress drop) not captured by the magnitude and location.

For magnitude estimation we use the peak displacement based method proposed by Kuyuk and Allen [15]. We bandpass filter the signal between 0.5 Hz and 3 Hz and discard traces with insufficient signal to noise ratio. We extract peak displacement from the horizontal components in the first 6 s of the P wave. We stop the time window at latest at the S onset. We use the relationship

$$M = c_1 \log(PD) + c_2 \log(R) + c_3 + \mathcal{N}(0, \sigma^2)$$

(4)

to estimate magnitudes from peak displacement. We use $c_1 = 1.23$, $c_2 = 1.38$, $c_3 = 5.69$ (Italy) / $c_3 = 5.89$ (Japan) and $\sigma = 0.31$. These are the values from [15], except for $c_3$ which needed to be adjusted as we do not use moment magnitude. We combine the predictions in probability space assuming independence between the predictions from different stations. We weight stations based on the length of the P wave window recorded so far. We use the mean value of the single-station magnitude estimates for PGA estimation. For both the application of the GMPE and the magnitude estimation we use the catalog hypocenters. As the quality of real-time location estimates will be worse, this leads to inflated performance measures for EPS.

As second baseline, we adapted the PLUM algorithm [5]. While the original paper applies PLUM to seismic intensities, we apply it to PGA values. This adaptation is possible, as approximate linear and especially monotonic relations exist between intensity and PGA [22]. The PGA prediction $pga^s_{ts}$ at a station $s$ at time $t$ is the maximum of all observed PGA values $pga^s_{ts'}$ at stations $s'$ within a radius $r$ of $s$. Therefore a warning for a certain threshold for a station is issued once the threshold has been exceeded at any station within the radius $r$. Due to different station densities in Italy and Japan we used different values for $r$. For Italy we used $r = 15$ km for Japan we used $r = 30$ km. Following the findings of Cochran et al [23], we do not use site correction terms in our implementation of PLUM as they only have minor impact on the performance.

D Evaluation metrics

We analyze the performance of the early warning algorithms using PGA thresholds of 1\%g, 2\%g, 5\%g, 10\%g and 20\%g, approximately matching Modified Mercalli Intensity (MMI) III (light) to VII (very strong) [39]. We calculate PGA from the absolute value of the two horizontal components. For the PGA values for the Japanese data, we use the surface stations and not the borehole stations.

A warning at a site should be issued if anytime during the event the PGA threshold is exceeded at the site. We consider a warning correct (true positive, TP), if a warning for a certain threshold was issued and the threshold was actually exceeded later during the event. Missed warnings (false negative, FN) are all cases, where the PGA threshold was exceeded, but no warning was issued or the warning was issued after the PGA threshold was first exceeded. We consider a warning false (false positive, FP), if a warning was issued, but the threshold was not exceeded. All remaining cases are true negatives (TN).

As the number of true negatives depends strongly on the inclusion criteria of the catalog, we use metrics independent of the true negatives. As summary statistics we use precision, $TP/(TP+FP)$, measuring the fraction of correct warnings among all warnings, and recall, $TP/(TP+FN)$, measuring the fraction of possible correct warnings that was issued. We use the $F1$ score $= 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall})$ as a combined statistic. Any analysis using a fixed $\alpha$ uses the
value maximizing the F1 score, which is specific to each method and PGA threshold. For an analysis independent of the threshold $\alpha$ we use the area under the precision recall curve (AUC). We use values $\alpha = 0.05, 0.1, 0.2, \ldots, 0.8, 0.9, 0.95$ and add additional points at $(0, 1)$ and $(1, 0)$ to the precision recall curve to approximate the AUC. For comparison of PLUM using AUC in Figure 3, we introduce an artificial precision recall line for PLUM with a slope of $-1$ going through the observed precision and recall values.

We define the warning time as the time between the issuance of a warning and the first exceedance of the threshold. We consider a zero latency system and do not impose a minimum warning time. For comparing warning times between methods or different parameter combinations, we only use the subset of station event pairs, where both methods/parameter combinations issued correct warnings.

We evaluate PLUM continuously, i.e., warnings are issued immediately at the exceedance of a threshold. TEAM and EPS are evaluated every 0.1 s, starting 1 s after the first P arrival for EPS and 0.5 s after the first P arrival for TEAM. We use a longer time before the first prediction for EPS as the early results of EPS are unstable. Warnings are not retracted, i.e., even if the model later estimates a shake level below the warning threshold, the warning stay active.

E Model calibration

Even though TEAM and EPS give probabilistic predictions, it is not assured whether these predictions are well-calibrated, i.e., if the predicted confidence values actually correspond to observed probabilities. Calibrated probabilities are essential for threshold selection, as they are required to balance expected costs of taking action versus expected costs of not taking action. We note that while good calibration is a necessary condition for a good model, it is not sufficient, as a model constantly predicting the marginal distribution of the labels would be always perfectly calibrated, yet not very useful.

Figures S9 and S10 show the calibration diagrams for Japan and Italy at different times after the first P arrival. For low PGA thresholds the calibration is near optimal for the Japan models at all times. For 10%g and 20%g EPS is underconfident, while the calibration of TEAM is still good. For these thresholds the calibration improves with time since the first P arrival.

For Italy the calibration diagrams are more diverse. EPS is generally slightly overconfident, except for the low confidence values at low PGA thresholds. TEAM shows good calibration for all PGA thresholds except 20%g, where it is underconfident. We suspect that the worse calibration for the largest events is caused by the domain adaptation strategy. In conclusion, both models are generally well calibrated, with a slightly better calibration for TEAM.
Figure S1: Map of the event and station distribution in the Japan data set. Stations are shown as black triangles, events as dots. The event color encodes the event magnitude. There are $\sim 20$ additional far offshore events outside the displayed map region in the catalog.
Figure S2: Map of the event and station distribution in the Italy data set. Stations present in the training set are shown as black triangles, while stations only present in the test set are shown as yellow triangles. Events are shown as dots with the color encoding the event magnitude. All events with magnitudes above 5.5 are shown as stars. The red stars indicate large training events, while the yellow stars indicate large test events. The inset shows the central Italy region with intense seismicity and high station density in the test set. Moment magnitudes for the largest test events are given in the inset.
Figure S3: Overview of the transformer earthquake alerting model. Colors encode if a box is a neural network, an input or an output. The main components of the model are the feature extraction (a CNN), the feature combination (a transformer) and the PGA estimation at targets (a mixture density network). Further inputs are the target coordinates, that are input just like stations, but with a zero feature vector. Ten instances of this network are trained independently and the results ensemble-averaged.
Figure S4: True positives (TP), false negatives (FN) and false positives (FP) for the events in the Italy test sets causing the largest shaking. The methods are the transformer earthquake alerting model without domain adaptation (TEAM base), the transformer earthquake alerting model (TEAM), the estimated point source algorithm (EPS) and PLUM. In addition, a GMPE with full catalog information is included for reference. Values $\alpha$ were chosen separately for each threshold and method to yield the highest F1 score for the whole test set, but are kept constant across all events. TEAM with domain adaptation outperforms TEAM without domain adaptation consistently across all thresholds. This indicates that the domain adaptation not only allows TEAM to better predict higher levels of shaking, but also to better assess large events in general.
Figure S5: Precision, recall and F1 score at different PGA thresholds for Italy including TEAM without domain adaptation. Threshold values $\alpha$ were chosen independently for each method and PGA threshold to yield the highest F1 score. The methods are the transformer earthquake alerting model without domain adaptation (TEAM Base), the transformer earthquake alerting model (TEAM), the estimated point source (EPS) model and the PLUM model. In addition the graph shows the performance of C-GMPE, a GMPE with full catalog information for reference.

Figure S6: Warning time and hypocentral distance between station and event for each true alert at F1-optimal $\alpha$. The white area corresponds roughly to the range of possible warning times and is bounded by the 90th percentile of the times between first detection of an event (i.e., arrival of P wave at the closest station) and first exceedance of the PGA threshold in recordings at that approximate distance.
Figure S7: Scenario analysis of the 30th October 2016 $M_w = 6.5$ Norcia earthquake, the largest event in the Italy test set. See Fig. 4 in the main paper for further explanations. The bottom row diagrams for this scenario analysis use a 10%g PGA threshold.
Figure S8: Predictions and residuals of the GMPEs derived in this study. All PGA values are given as log units using m/s². Every point refers to one recording. Solid lines indicate running means, dashed lines denote the running standard deviation around the running mean. Orange crosses denote mean and standard deviations for magnitude ranges with insufficient data to infer a continuous line. Window sizes are 0.24 m.u./10 km (Italy) and 0.44 m.u./53 km (Japan). Overall σ is 0.29 for Italy and 0.33 for Japan. The plotted magnitude values have been offset by random values between -0.05 and 0.05 m.u. for increased visibility.
Figure S9: Calibration diagrams for Japan at different times after the first P detection and different PGA thresholds. The confidence is defined as the probability of exceeding the PGA threshold as predicted by the model. Each bar represents the traces with a confidence value inside the limits of the bar. Its height is given by the accuracy, the fraction of traces actually exceeding the threshold among all traces in the bar. For a perfectly calibrated model, the confidence equals the accuracy. This is indicated by the dashed line. We note that accuracy estimations for the high PGA thresholds are strongly impacted by stochasticity due to the small number of samples.
Figure S10: Calibration diagrams for Italy at different times after the first P detection and different PGA thresholds. For a further description see the caption of figure S9.
Table S1: Performance statistics for Japan. Probability thresholds $\alpha$ were chosen to maximize F1 scores and are shown in the last column. The AUC value does not depend on the threshold $\alpha$. PGA indicates the used PGA threshold.

| PGA [g] | Precision | Recall | F1  | AUC | $\alpha$ |
|---------|-----------|--------|-----|-----|----------|
| TEAM    |           |        |     |     |          |
| 1%      | 0.70      | 0.77   | 0.73| 0.82| 0.60     |
| 2%      | 0.69      | 0.69   | 0.69| 0.76| 0.60     |
| 5%      | 0.59      | 0.67   | 0.65| 0.68| 0.50     |
| 10%     | 0.50      | 0.60   | 0.54| 0.56| 0.40     |
| 20%     | 0.33      | 0.48   | 0.39| 0.35| 0.30     |
| EPS     |           |        |     |     |          |
| 1%      | 0.60      | 0.63   | 0.56| 0.57| 0.40     |
| 2%      | 0.48      | 0.48   | 0.48| 0.48| 0.40     |
| 5%      | 0.40      | 0.40   | 0.40| 0.34| 0.30     |
| 10%     | 0.27      | 0.36   | 0.31| 0.25| 0.20     |
| 20%     | 0.20      | 0.26   | 0.22| 0.15| 0.20     |
| PLUM    |           |        |     |     |          |
| 1%      | 0.39      | 0.56   | 0.46| -   | -        |
| 2%      | 0.30      | 0.50   | 0.38| -   | -        |
| 5%      | 0.22      | 0.42   | 0.29| -   | -        |
| 10%     | 0.18      | 0.39   | 0.25| -   | -        |
| 20%     | 0.11      | 0.28   | 0.16| -   | -        |
| C-GMPE  |           |        |     |     |          |
| 1%      | 0.58      | 0.74   | 0.65| 0.69| 0.30     |
| 2%      | 0.47      | 0.71   | 0.56| 0.60| 0.20     |
| 5%      | 0.44      | 0.54   | 0.48| 0.48| 0.20     |
| 10%     | 0.44      | 0.46   | 0.45| 0.43| 0.20     |
| 20%     | 0.56      | 0.38   | 0.45| 0.42| 0.30     |

Table S2: Performance statistics for Italy. Probability thresholds $\alpha$ were chosen to maximize F1 scores and are shown in the last column. The AUC value does not depend on the threshold $\alpha$. PGA indicates the used PGA threshold.

| PGA [g] | Precision | Recall | F1  | AUC | $\alpha$ |
|---------|-----------|--------|-----|-----|----------|
| TEAM    |           |        |     |     |          |
| 1%      | 0.64      | 0.64   | 0.64| 0.68| 0.60     |
| 2%      | 0.55      | 0.65   | 0.60| 0.63| 0.50     |
| 5%      | 0.58      | 0.52   | 0.55| 0.54| 0.50     |
| 10%     | 0.50      | 0.46   | 0.48| 0.43| 0.40     |
| 20%     | 0.51      | 0.35   | 0.42| 0.36| 0.30     |
| EPS     |           |        |     |     |          |
| 1%      | 0.44      | 0.37   | 0.40| 0.37| 0.30     |
| 2%      | 0.44      | 0.36   | 0.40| 0.36| 0.40     |
| 5%      | 0.41      | 0.40   | 0.40| 0.33| 0.40     |
| 10%     | 0.39      | 0.38   | 0.38| 0.30| 0.40     |
| 20%     | 0.39      | 0.39   | 0.39| 0.25| 0.40     |
| PLUM    |           |        |     |     |          |
| 1%      | 0.25      | 0.66   | 0.37| -   | -        |
| 2%      | 0.21      | 0.61   | 0.31| -   | -        |
| 5%      | 0.18      | 0.60   | 0.28| -   | -        |
| 10%     | 0.22      | 0.64   | 0.32| -   | -        |
| 20%     | 0.19      | 0.65   | 0.30| -   | -        |
| C-GMPE  |           |        |     |     |          |
| 1%      | 0.62      | 0.69   | 0.65| 0.71| 0.30     |
| 2%      | 0.61      | 0.65   | 0.63| 0.68| 0.30     |
| 5%      | 0.60      | 0.59   | 0.60| 0.63| 0.30     |
| 10%     | 0.64      | 0.57   | 0.60| 0.59| 0.30     |
| 20%     | 0.59      | 0.54   | 0.56| 0.54| 0.30     |

Table S3: Relative warning times of the algorithms in seconds. Positive values indicate longer average warning times for the second method, negative values shorter warning times. The difference in average warning times is calculated from all event station pairs, where both methods issued correct warnings. No value is reported if this set is empty. We set $\alpha$ for TEAM and EPS to the optimal value in terms of F1 score.

| PGA [g] | Japan | Italy |
|---------|-------|-------|
| EPS TEAM| 0.39  | 0.18  |
| PLUM TEAM| 8.98 | 1.49  |
| PLUM EPS| 8.53  | 2.95  |
Table S4: Data set statistics for the full data set and the test set. The lower boundary of the magnitude category is the 5th percentile of the magnitude; this limit is chosen as each data set contains a small number of unrepresentative very small events. The upper boundary is the maximum magnitude. The lower part of the table shows how often each PGA threshold was exceeded. An event is counted as exceeding a threshold if at least one station exceeded this threshold during the event. The number of exceedances in the test set for Italy is disproportionally high compared to the number of events in the test set. This is caused by the high seismic activity and the higher station density in 2016. Traces for Japan always refer to 6 component traces, while for Italy it refers to 3 component traces.

|                | Japan       |               | Italy       |               |
|----------------|-------------|---------------|-------------|---------------|
|                | Full        | Test          | Full        | Test          |
| Years          | 1997 - 2018 | 08/2013 - 12/2018 | 2008 - 2019 | 01/2016 - 12/2016 |
| Magnitudes     | 2.7 - 9.0   | 2.7 - 8.1     | 2.7 - 6.5   | 2.7 - 6.5     |
| Events         | 13,512      | 4,054         | 7,055       | 2,123         |
| Unique stations| 697         | 632           | 1,080       | 621           |
| Traces         | 372,661     | 104,573       | 494,183     | 253,454       |
| Avg. traces per event | 27.6 | 25.9 | 70.3 | 119.4 |
| PGA [g]        |             |               |             |               |
| 1%             | 8,761       | 55,618        | 1,841       | 6,379         |
|                | 5,324       | 24,396        | 1,013       | 2,921         |
| 5%             | 2,026       | 6,802         | 348         | 888           |
| 10%            | 782         | 2,223         | 120         | 330           |
| 20%            | 238         | 631           | 40          | 107           |

Table S5: Architecture of the feature extraction network. The input dimensions of the waveform data are (time, channels). FC denotes fully connected layers. As FC layers can be regarded as 0D convolutions, we write the output dimensionality in the filters column. The “Concatenate scale” layer concatenates the log of the peak amplitude to the output of the convolutions. Depending on the existence of borehole data the number of input filters for the first Conv1D layer is 64 instead of 32 in the non-borehole case.

| Layer            | Filters | Kernel size | Stride |
|------------------|---------|-------------|--------|
| Conv2D           | 8       | 5, 1        | 5, 1   |
| Conv2D           | 32      | 16, 3       | 1, 3   |
| Flatten to 1D    | 64      | 16          | 1      |
| Conv1D           | 2       | 2           | 2      |
| MaxPool1D        | 128     | 16          | 1      |
| Conv1D           | 2       | 2           | 2      |
| MaxPool1D        | 32      | 8           | 1      |
| Conv1D           | 2       | 2           | 2      |
| Conv1D           | 32      | 8           | 1      |
| Conv1D           | 16      | 4           | 1      |
| Flatten to 0D    |         |             |        |
| Concatenate scale|         |             |        |
| FC               | 500     |             |        |
| FC               | 500     |             |        |
| FC               | 500     |             |        |

Table S6: Architecture of the transformer network.

| Feature          | Value   |
|------------------|---------|
| # Layers         | 6       |
| Dimension        | 500     |
| Feed forward dimension | 1000   |
| # Heads          | 10      |
| Maximum number of stations | 25     |
| Dropout          | 0       |
| Activation       | GeLu    |