Münchmeyer, J., Bindi, D., Leser, U., Tilmann, F. (2021): The transformer earthquake alerting model: A new versatile approach to earthquake early warning. - Geophysical Journal International, 225, 1, 646-656.

https://doi.org/10.1093/gji/ggaa609

Geophysical Journal International

Geophys. J. Int. (2021) **225,** 646–656
Advance Access publication 2020 December 24
GJI Seismology

doi: 10.1093/gji/ggaa609

The transformer earthquake alerting model: a new versatile approach to earthquake early warning

Jannes Münchmeyer [1,2] Dino Bindi [1] Ulf Leser [2] and Frederik Tilmann [1,3]

[1] *Helmholtzzentrum Potsdam, Deutsches GeoForschungsZentrum GFZ,* 14473 *Potsdam, Germany. E-mail: munchmej@gfz-potsdam.de*
[2] *Institut für Informatik, Humboldt-Universität Berlin,* 10117 *Berlin, Germany*
[3] *Insitut für geologische Wissenschaften, Freie Universität Berlin,* 14195 *Berlin, Germany*

Accepted 2020 December 22. Received 2020 November 27; in original form 2020 October 12

SUMMARY

Earthquakes are major hazards to humans, buildings and infrastructure. Early warning methods aim to provide advance note of incoming strong shaking to enable preventive action and mitigate seismic risk. Their 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. Current approaches suffer from apparent aleatoric uncertainties due to simplified modelling or short warning times. Here we propose a novel early warning method, the deep-learning based transformer earthquake alerting model (TEAM), to mitigate these limitations. TEAM analyses raw, strong motion waveforms of an arbitrary number of stations at arbitrary locations in real-time, making it easily adaptable to changing seismic networks and warning targets. We evaluate TEAM on two regions with high seismic hazard, Japan and Italy, that are complementary in their seismicity. On both data sets TEAM 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.

Key words: Neural networks, fuzzy logic; Probability distributions; Earthquake early warning.

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 (Allen *et al.* 2009; Allen & Melgar 2019). 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 (Chung *et al.* 2019) or FINDER (Böse *et al.* 2018), estimate the source properties of an event, that is, 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 apparent aleatoric uncertainty due to simplified assumptions in the source estimation and in the GMPE (Kodera *et al.* 2018). Propagation based methods, like PLUM (Kodera *et al.* 2018), 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 (Meier *et al.* 2020).

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, for example, estimation of magnitude (Lomax *et al.* 2019; Mousavi & Beroza 2020), location (Kriegerowski *et al.* 2019; Mousavi & Beroza 2019) or peak ground acceleration (PGA, Jozinović *et al.* 2020). 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, the existing methods are restricted in terms of their input stations, as they use either a single seismic station as input (Lomax *et al.* 2019; Mousavi & Beroza 2020) or a fixed set of seismic stations, that needs to be defined at training time (Kriegerowski *et al.* 2019; Jozinović *et al.* 2020; Otake *et al.* 2020). While single station approaches miss out on a considerable amount of information obtainable from combining waveforms from different sources, fixed stations approaches have limited adaptability to changing networks. The latter is of particular concern as for large, dense networks the stations of interest, that is, the stations closest to an event, will change on a per-event basis. Finally, existing methods systematically underestimate the strongest shaking and the highest magnitudes, as these are rare and therefore underrepresented in the training data [figs 6, 8 in Jozinović *et al.* (2020), figs 3, 4 in Mousavi & Beroza (2020)]. 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 data sets from highly seismically active regions with dense seismic networks, namely Japan (13 512 events, years 1997–2018, Fig. 1) and Italy (7055 events, years 2008–2019, Fig. 2). 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 data sets into training, development and test sets with ratios of 60:10:30. We employ an event-wise split, that is, all records for a particular event will be assigned to the same subset. We do not explicitly split station-wise but due to temporary deployments there are a few stations in the test set which have no records in the training set (Fig. 2). 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 data set 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 $M_\mathrm{w} = 6.2$ and $M_\mathrm{w} = 6.5$ Norcia events (Dolce & Di Bucci 2018). Especially the latter event is notably larger than any in the training set ($M_\mathrm{w} = 6.1$ L'Aquila event in 2007), thereby challenging the extrapolation capabilities of TEAM.

Both data sets 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 pre-processing are further described in the supplement text S1.

2.2 TEAM

The early warning workflow with TEAM encompasses three separate steps (Fig. 3): 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 for each target locations where the predicted exceedance probability p for fixed PGA thresholds surpasses a predefined probability α.

TEAM conducts end-to-end PGA estimation: its input are raw waveforms, its output predicted PGA probability densities. There are no intermediate representations in TEAM that warrant an immediate geophysical interpretation. The PGA assessment can be subdivided into three components: feature extraction, feature combination, and density estimation (Fig. S1). Input to TEAM are three, respectively six (three surface, three borehole), component waveforms at 100 Hz sampling rate from multiple stations and the corresponding station coordinates. Furthermore, 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.

TEAM extracts features from input waveforms using a convolutional neural network (CNN). The feature extraction is applied separately to each station, but is identical for all 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 through the alignment of the waveforms and zero-padding: we align all input waveforms in time, that is, 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, allowing the model to understand the noise characteristics. For t_1 we use the current time, that is, the amount of available waveforms. We obtain constant length input, by padding all waveforms after t_1 with zeros up to a total length of 30 s. The feature extraction is described in more detail in supplementary text S2.

TEAM combines the feature vectors and maps them to representations at the targets using a transformer (Vaswani *et al.* 2017). Transformers are attention-based neural networks for combining information from a flexible number of input vectors in a learnable way. To encode the location of the recording stations as well as of the prediction targets, we use sinusoidal vector representations. For input stations, we add these representations component-wise to the feature vectors, for target stations we directly use them as inputs to the transformer. This architecture, processing a varying number of inputs, together with the explicitly encoded locations, allows TEAM to handle dynamically varying sets of stations and targets. The transformer returns one vector for each target representing predictions at this target. Details on the feature combinations can be found in supplementary text S3.

From each of the vectors returned by the transformer, TEAM calculates the PGA predictions at one target. Similar to the feature extraction, the PGA prediction network is applied separately to each target, but is identical for all targets. TEAM uses mixture density networks (Bishop 1994) returning Gaussian mixtures, to computes PGA densities. Gaussian mixtures allow TEAM to predict more complex distributions and better capture realistic uncertainties than a point estimate or a single Gaussian. The full specifications for the final PGA estimation are provided in supplementary text S4.

TEAM is trained end-to-end using a negative log-likelihood loss. To increase the flexibility of TEAM and allow for real-time processing, we use training data augmentation. We randomly select the stations used as inputs and targets in each training iteration. In addition, again in each training iteration, we randomly replace all waveforms after a time t with zeros, matching the input representation of real time data, to train TEAM for real-time application. These data augmentations as well as the complete training procedure are further described in supplementary text S5.

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 epicentre during training, which reduces the inherent bias in data towards smaller

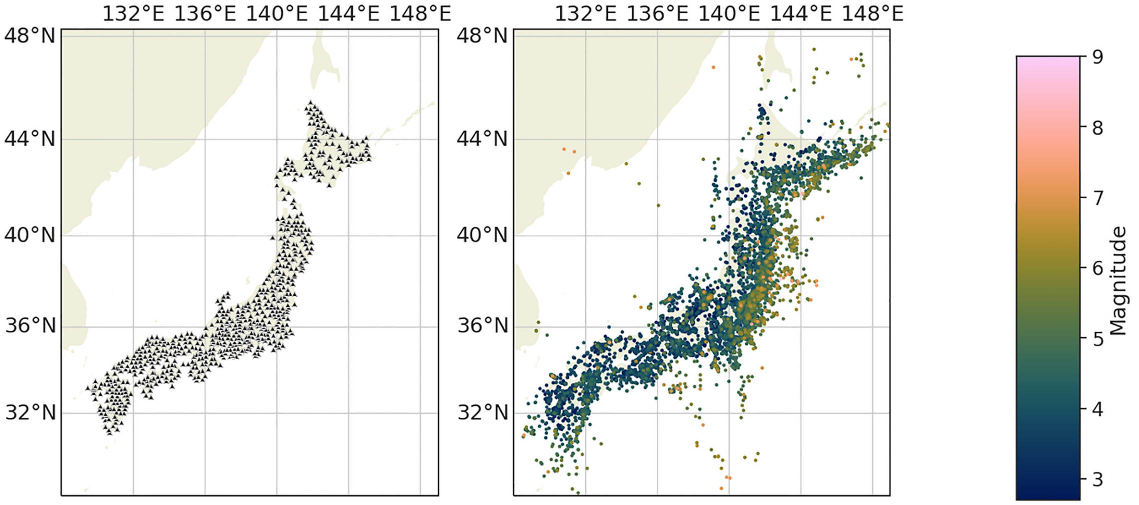


Figure 1. Map of the station (left-hand panel) and event (right-hand panel) distribution in the Japan data set. Stations are shown as black triangles, events as dots. The event colour encodes the event magnitude. There are ∼20 additional events far offshore, which are outside the displayed map region in the catalogue.

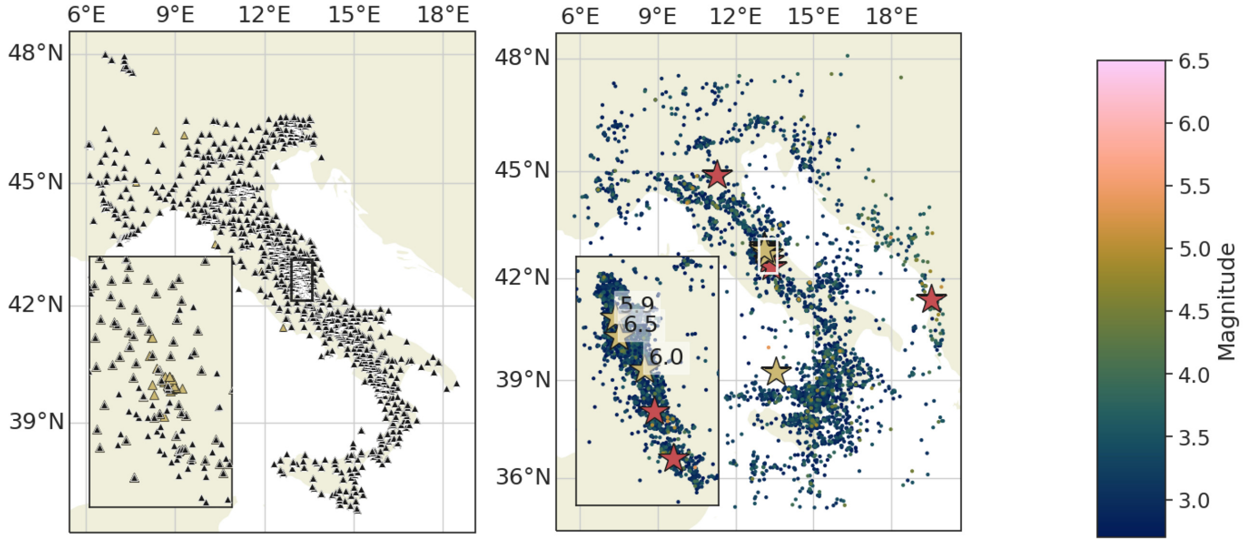


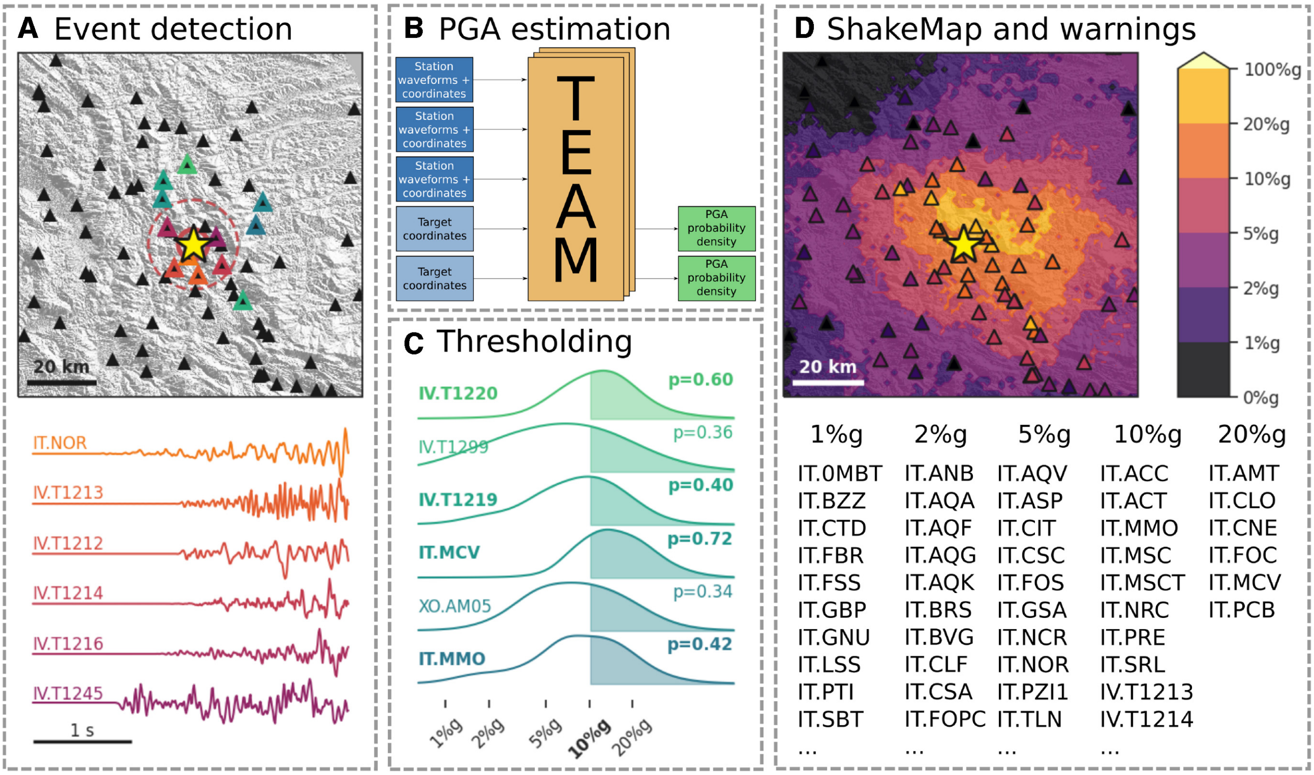
Figure 2. Map of the station (left-hand panel) and event (right-hand panel) distribution in the Italy data set. Stations present in the training set are shown as black triangles, while stations only present in the test set are shown as yellow triangles. Events are shown as dots with the colour encoding 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.

PGAs. When learning from small catalogues or when applied to regions where events substantially larger than all training events can be expected, for example, 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 large events from other regions that are similar to the expected events in the target region. While records from those events will differ in certain aspects, for example, site responses or the exact propagation patterns, other aspects, for example, the average extent of strong shaking or the duration of events of a certain size, will mostly be independent of the region in question. The domain adaptation aims to enable the model to transfer the region immanent aspects of large events, at the cost of a certain blurring of the specific regional aspects of the target region. TEAM aims to mitigate the blurring of regional aspects by the choice of training procedure.

Our Italy data set 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. As we use events from Italy in both training steps and in particular in the second step the overwhelming number of events are from Italy, we expect that this scheme only results in a small degradation in the modelling of the regional specifics of the Italy region.

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



1%g	2%g	5%g	10%g	20%g
IT.0MBT	IT.ANB	IT.AQV	IT.ACC	IT.AMT
IT.BZZ	IT.AQA	IT.ASP	IT.ACT	IT.CLO
IT.CTD	IT.AQF	IT.CIT	IT.MMO	IT.CNE
IT.FBR	IT.AQG	IT.CSC	IT.MSC	IT.FOC
IT.FSS	IT.AQK	IT.FOS	IT.MSCT	IT.MCV
IT.GBP	IT.BRS	IT.GSA	IT.NRC	IT.PCB
IT.GNU	IT.BVG	IT.NCR	IT.PRE	
IT.LSS	IT.CLF	IT.NOR	IT.SRL	
IT.PTI	IT.CSA	IT.PZI1	IV.T1213	
IT.SBT	IT.FOPC	IT.TLN	IV.T1214	
...	...	...	...	

Figure 3. Schematic view of TEAM's early warning workflow for the October 2016 Norcia event ($M_\mathrm{w} = 6.5$) 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 colours correspond to the stations highlighted with orange to magenta outlines. The circles indicate the approximate current position of *P* (dashed) and S (solid) wave fronts. (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 Fig. S1. (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 α, a warning is issued. The figure visualizes a 10 per cent g PGA level with $\alpha = 0.4$, resulting in warnings for the stations highlighted. The colours 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 coloured according to their current warning level. The table lists all stations for which warnings have already been issued.

estimation based method we use the estimated point source approach (EPS), which estimates magnitudes from peak displacement during the *P*-wave onset (Kuyuk & Allen 2013) and then applies a GMPE (Cua & Heaton 2009) to predict the PGA. For simplicity, our implementation assumes knowledge of the final catalogue epicentre, which is impossible in real-time, leading to overoptimistic results for EPS. As propagation based method we chose an adaptation of PLUM (Kodera *et al.* 2018), which issues warnings if a station within a radius *r* of the target exceeds the level of shaking. In contrast to the original PLUM, which operates on the Japanese seismic intensity scale, I_JMA (Shabestari & Yamazaki 2001), our adaptation applies the concept of PLUM to PGA, thereby making it comparable to the other approaches. Whereas I_JMA is also a measure of the strongest acceleration and is thus strongly correlated with PGA, it considers a narrower frequency band and imposes a minimum duration of strong shaking. As such, although the performance might vary slightly for our PLUM-like approach compared to the original PLUM, it still exhibits its key features, in particular the effects of the localized warning strategy. Additionally we apply the GMPE used in EPS to catalogue location and magnitude as an approximate upper accuracy bound for point source algorithms (Catalogue-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 supplementary text S6.

3 RESULTS

3.1 Alert performance

We compare the alert performance of all methods for PGA thresholds from light (1 per cent g) to very strong (20 per cent 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 (Meier 2017; Minson *et al.* 2019). Precision and recall trade-off against each other depending on α. While the PGA predictions of TEAM, EPS and the C-GMPE are probabilistic, the thresholding transforms the predictions into alerts or non-alerts. The probability distribution describes the uncertainty of the models, for example, for the GMPE the apparent aleatoric uncertainty from aspects not accounted for, which makes false and missed alerts inevitable. The threshold value controls the trade-off between both types of errors, and its appropriate value will depend on user needs, specifically the costs associated with false and missed alerts. Therefore, to analyse the performance of the models across different user requirements, we look at the precision recall curves for different thresholds α. In addition to precision and recall, we use two summary metrics: *F1 score*, the harmonic mean of precision and recall, and *AUC*, the area under the precision recall curve. The evaluation metrics and full setup of the evaluation are defined in detail in supplement text S7.

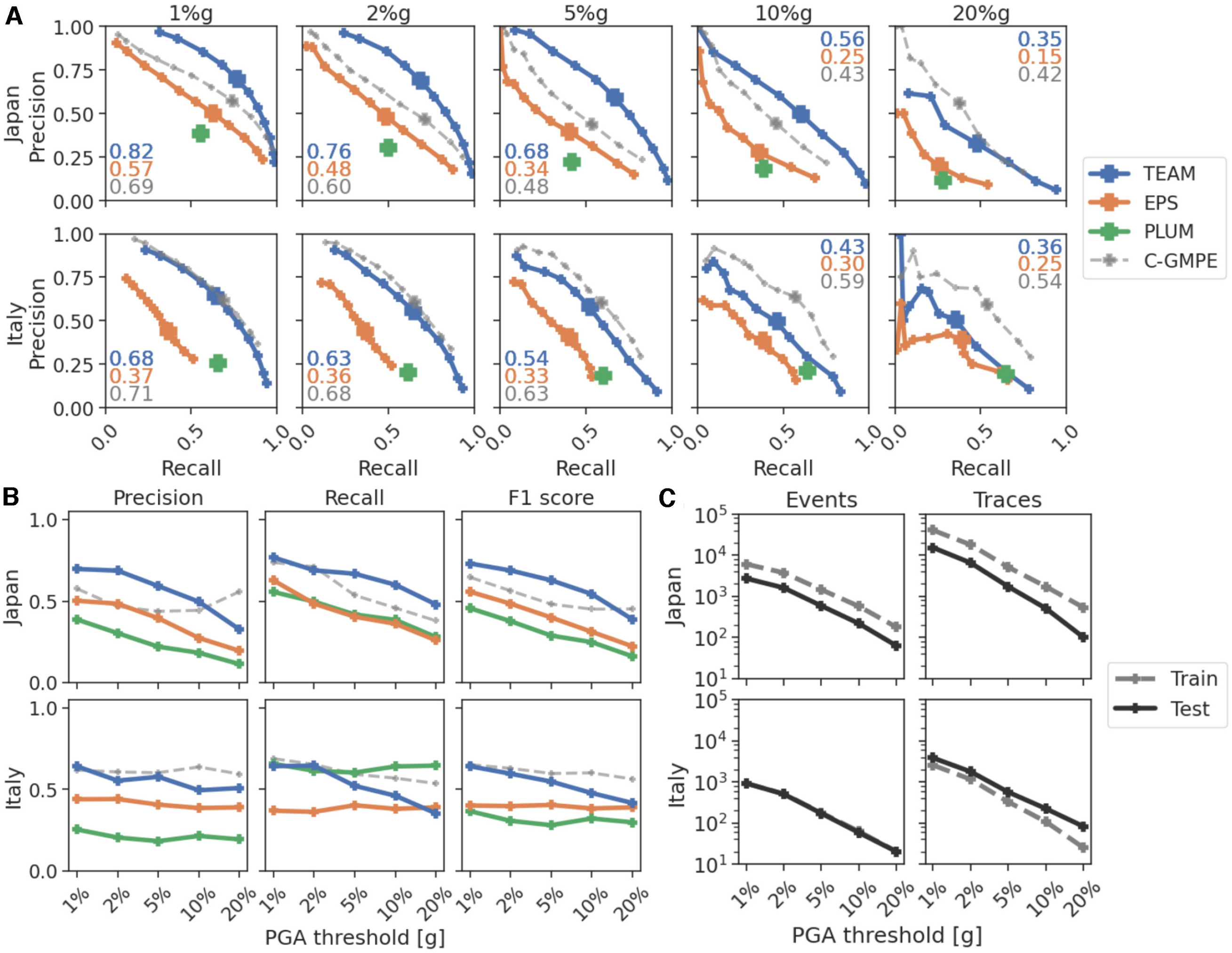


Figure 4. Warning statistics for the three early-warning models (TEAM, EPS, PLUM) for the Japan and Italy data sets. In addition, statistics are provided for C-GMPE, which can only be evaluated post-event due to its reliance on catalogue magnitude and location. (a) Precision and recall curves across different thresholds $\alpha = 0.05, 0.1, 0.2, \ldots, 0.8, 0.9, 0.95$. As the PLUM-like approach 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 α. Threshold probabilities α 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.

TEAM outperforms both EPS and the PLUM-like approach for both data sets and all PGA thresholds, indicated by the precision-recall-curves of TEAM lying to the top-right of the baseline curves (Fig. 4a). 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. To compare the performance at fixed α, we selected α values yielding the highest F1 score separately for each PGA threshold and method. Again, TEAM outperforms both baselines on both data sets, irrespective of the PGA level (Fig. 4b). Performance statistics in numerical form 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 (Meier 2017; Minson *et al.* 2019; Meier *et al.* 2020). Furthermore, shortage of training data with high PGA values results in less well constrained model parameters.

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 (Figs S2 and S3). While for low PGA thresholds differences are small, at high PGA levels they grow to more than 20 points F1 score. Interestingly, 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. Domain adaptation therefore helps TEAM to remain accurate even for events quite far from the training distribution.

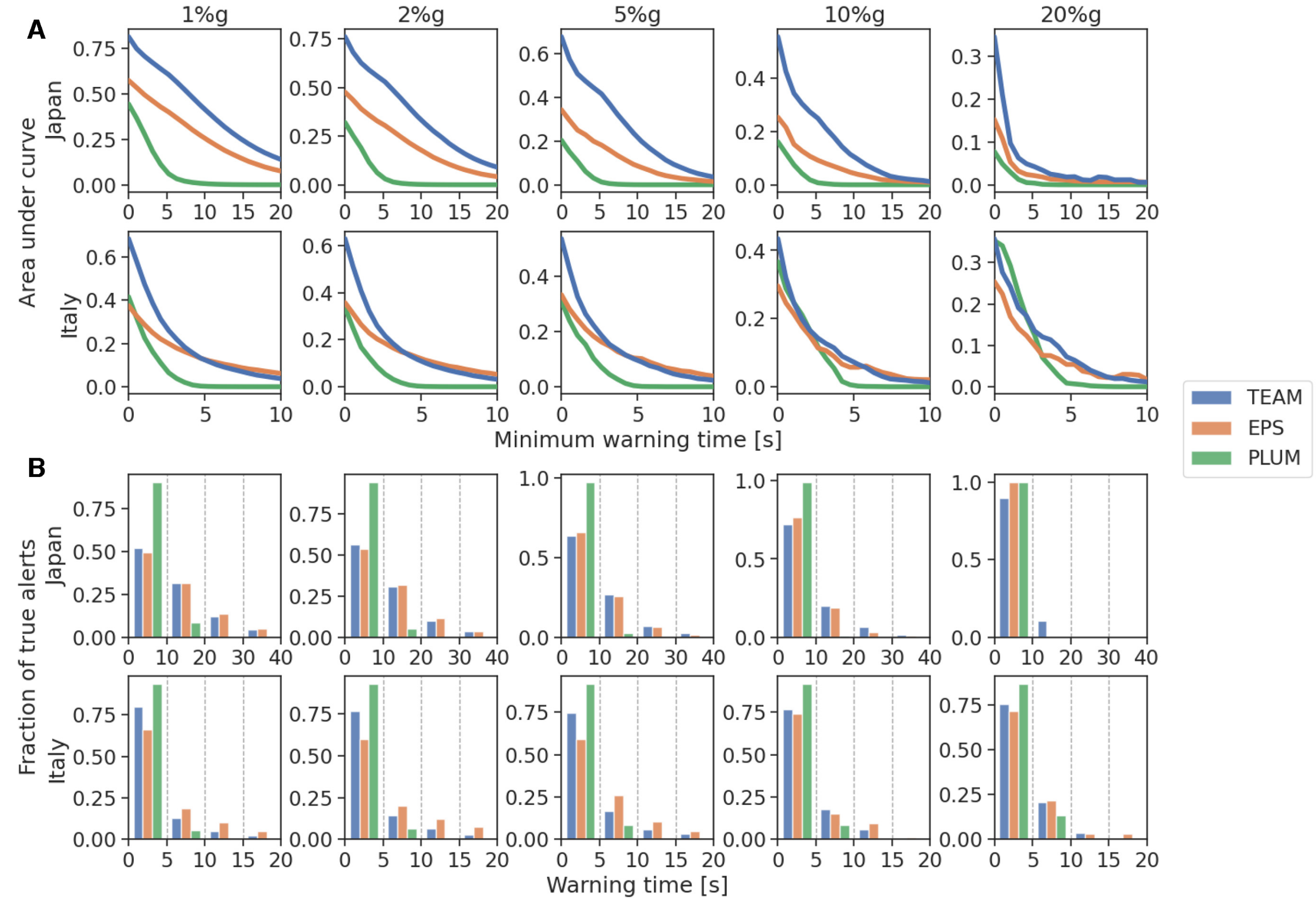


Figure 5. Warning time statistics. (a) Area under the precision recall curve for different minimum warning times. All alerts with shorter warning times are counted as false negatives. (b) Warning time histograms showing the distribution true alerts across distances for the different methods. Please note that the total number of true alerts differs by method and is not shown in this subplot. Therefore the values of different methods cannot be directly compared, but only the differences in the distributions. TEAM and EPS are shown at F1-optimal α, chosen separately for each threshold and method. Warning time dependence on hypocentral distance is shown in Fig. S4.

3.2 Warning times

In application scenarios, a user will usually require a certain warning time, which is the time between issuing of the warning and first exceedance of the level of shaking, as this time is necessary for taking action. As the previous evaluation considered prediction accuracy irrespective of the warning time, we now compare the methods while imposing a certain minimum warning time. Actually, TEAM consistently outperforms both baselines across different required warning times and irrespective of the PGA threshold (Fig. 5a). While the margin for TEAM compared to the baselines is smaller for Italy than for Japan, TEAM shows consistently strong performance across different warning times. In contrast, EPS performs clearly worse at short warning times, the PLUM-based approach at longer warning times. The latter is inherent to the key idea of PLUM and makes the method 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 epicentre (Minson *et al.* 2018). We further note that while the PLUM-like approach shows slightly higher AUC than TEAM for short warning times at 20 per cent g, this is only a hypothetical result. As PLUM does not have a tuning parameter between precision and recall, this performance can actually only be realized for a specific precision/recall threshold, where it performs slightly superior to TEAM (Fig. 4a bottom right-hand panel).

Warning times depend on α: a lower α value naturally leads to longer warning times but also to more false positive warnings. At F1-optimal thresholds α, EPS and TEAM have similar warning time distributions (Fig. 5b, Table S3), but lowering α leads to stronger increases in warning times for TEAM. For instance, at 10 per cent g, lowering α 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. Secondly, real warning times will be shorter than reported here due to telemetry and compute delays. However, compute delays 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

4.1 Calibration of probabilities

Even though TEAM and EPS give probabilistic predictions, it is not clear whether these predictions are well-calibrated, that is, 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

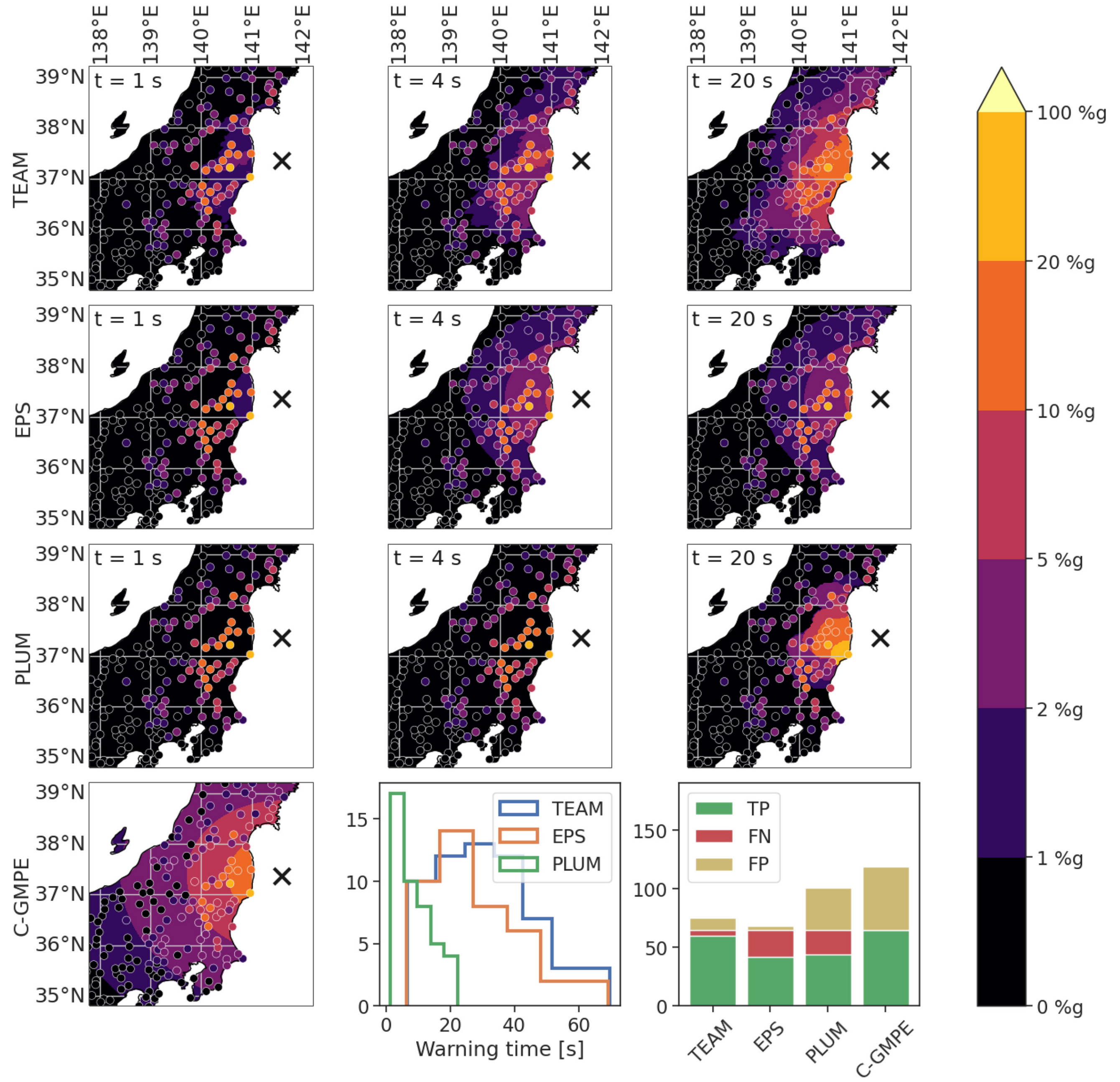


Figure 6. Scenario analysis of the 22 November 2016 $M_{\rm J} = 7.4$ Fukushima earthquake, the largest test event located close to shore. Maps show the warning levels for each method (top three rows) at different times (shown in the corner, $t = 0$ s corresponds to P arrival at closest station). Dots represent stations and are coloured according to the PGA recorded during the full event, that is, the prediction target. The bottom row shows (left- to right-hand panels), the catalogue 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 per cent g PGA threshold. EPS and GMPE shake map predictions do not include station terms, but they are included for the bottom row histograms.

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.

To assess the calibration, we use calibration diagrams (Figs S9 and S10) for Japan and Italy at different times after the first P arrival. These diagrams compare the predicted probabilities to the actually observed fraction of occurrences. In general, both models are well calibrated, with a slightly better calibration for TEAM. Calibration is generally better for Japan, where only EPS is slightly underconfident at earlier times for the highest PGA thresholds. For Italy, EPS is generally slightly overconfident, while TEAM is well calibrated, except for a certain overconfidence at 20 per cent g. We suspect that the worse calibration for the largest events is caused by the domain adaptation strategy, but the better performance in terms of accuracy clearly weighs out this downside of domain adaptation.

4.2 Insights into TEAM

We analyse differences between the methods using one example event from each data set (Japan: Fig. 6, Italy: Fig. S5). All methods underestimate the shaking in the first seconds (left-hand column Figs 6 and S5). However, TEAM is the quickest to detect the correct

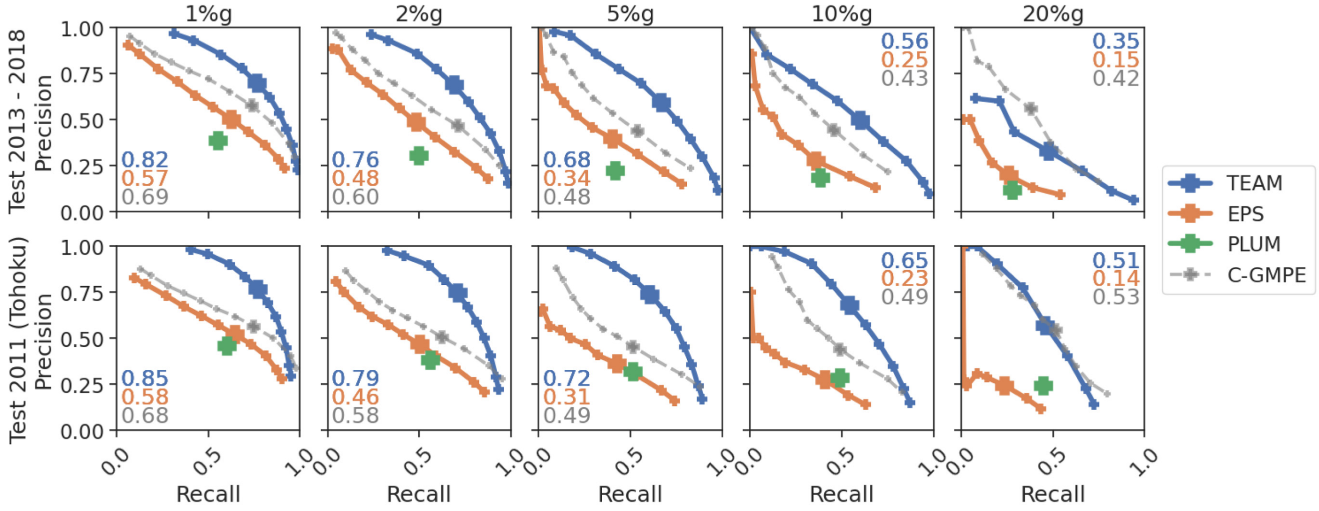


Figure 7. Precision recall curves for the Japanese data set using the chronological split (top panel) and using the events in 2011 as test set (bottom panel). The year 2011 contains the $M_\mathrm{w} = 9.1$ Tohoku event as well as its aftershocks.

extent of the shaking. Additionally, it estimates even fine-grained regional shaking details in real-time (middle and right-hand columns). In contrast, shake maps for EPS remain overly simplified due to the assumptions inherent to GMPEs (right-hand column and bottom left-hand panel). For the Japan example, even late predictions of EPS underestimate the shaking, due to an underestimation of the magnitude. The PLUM-based approach 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-hand panels). For the Japan data, this is not only the case for the example at hand, but also visible in Fig. 4, showing higher accuracy of TEAM's prediction compared to C-GMPE for all thresholds except 20 per cent g on the full Japan data set. 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, for example, 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, for example, frequency dependent site responses, and the ability to consider shaking at proximal stations. We note that for our Italy experiments, the modelling of local aspects resulting from regional characteristics might be slightly degraded by the domain adaptation. However, the first-order propagation effects such as, for example, amplitude decay due to geometric spreading, are similar between regions and therefore not negatively affected by the domain adaptation. In conclusion, combining a global event view with propagation aspects, TEAM can be seen as a hybrid model between source estimation and propagation.

4.3 TEAM performance on the Tohoku sequence

We evaluated TEAM for Japan on a chronological train/dev/test split, as this split ensures the evaluation closest to the actual application scenario. On the other hand, this split put the $M = 9.1$ Tohoku event in March 2011 into the training set. To evaluate the performance for this very large event and its aftershocks, we trained another TEAM instance using the year 2011 as test set and the remainder of the data for training and validation. Fig. 7 shows the precision recall curves for the chronological split and the year 2011 as test set. In general, the performance of all models stays similar when evaluated on the alternative split. A key difference between the curves is, that TEAM, in particular for high PGA thresholds, does not reach similar levels of recall for 2011 as for the chronological split, while achieving higher precision. As we will describe in the next paragraph, this trend probably results from a tendency to underestimate true PGA amplitudes, which will naturally reduce recall and boost precision. Nevertheless, the performance of TEAM as quantified by the AUC actually improves, and significantly so for the highest thresholds. We suspect that this tendency for underestimation is either caused by the higher number of large events in the 2011 test set compared to the chronological split, or by the lower number of high PGA events in the training set without 2011.

Fig. S6 presents a scenario analysis for the Tohoku event. All models underestimate the event considerably, with the strongest underestimation for the EPS method. Even 20 s after the first *P* wave arrival, all methods underestimate both the severity and the extent of shaking. Due to its localized approach, the PLUM-based model achieves the highest number of true warnings, albeit at short warning times and a certain number of false positives, which due to the underestimation are totally absent from TEAM and EPS predictions. The performance of both EPS and TEAM is likely degraded by the slow onset of the Tohoku event as described by Koketsu *et al.* (2011). According to Koketsu *et al.* (2011) the main subevent with a displacement of 36 m only initiated 20 s after the onset of the Tohoku event. Therefore only the first *P* waves for EPS or at most the first 25 s of waveforms for TEAM is most likely insufficient to correctly estimate the size of the Tohoku event.

For Italy, we showed that underestimation for large events can be mitigated using transfer learning. However, the Tohoku event clearly shows the limitations of this strategy, as nearly no training data for events of comparable size are available, even when using events across the globe. Therefore, for the largest events alternative strategies need to be developed, for example, training using simulated data. Furthermore, the 25 s of waveforms used by TEAM in the current implementation may, for a very large event, not capture the largest subevent. While we decided to use only 25 s of event waveforms, as there is only insufficient training data of longer events, this window could be extended when developing training strategies and models for the largest events.

5 CONCLUSION

In this study we presented 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, for example, by setting sliders in an early warning system. Secondly, deep learning models typically exhibit large performance improvements from larger training data sets (Sun *et al.* 2017) due to the high number of model parameters. In our study this reflects in the better performance on the twofold larger Japan data set. This indicates that TEAM's performance can be improved just by collecting more comprehensive catalogues, which happens automatically over time.

ACKNOWLEDGEMENTS

We thank the National Research Institute for Earth Science and Disaster Resilience (NIED) for providing the catalogue and waveform data for our Japan data set. We thank the Istituto Nazionale di Geofisica e Vulcanologia and the Dipartimento della Protezione Civile for providing the catalogue and waveform data for our Italy data set. JM 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. We thank Hiroyuki Goto and an anonymous reviewer for their comments which helped to improve the manuscript.

The source code for TEAM and the baselines are available at https://github.com/yetinam/TEAM. The repository contains the necessary instructions for downloading the Japan data set as well. The Italy data set was published as Münchmeyer *et al.* (2020).

REFERENCES

Allen, R.M. & Melgar, D., 2019. Earthquake early warning: advances, scientific challenges, and societal needs, *Ann. Rev. Earth Planet. Sci.,* **47**(1), 361–388.

Allen, R.M., Gasparini, P., Kamigaichi, O. & Bose, M., 2009. The status of earthquake early warning around the world: an introductory overview, *Seismol. Res. Lett.,* **80**(5), 682–693.

Belkin, M., Hsu, D., Ma, S. & Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off, *Proc. Natl. Acad. Sci.,* **116**(32), 15 849–15 854.

Bishop, C.M., 1994. Mixture density networks, Tech. rep., Aston University.

Böse, M., Smith, D.E., Felizardo, C., Meier, M.-A., Heaton, T.H. & Clinton, J.F., 2018. FinDer v.2: improved real-time ground-motion predictions for m2–M9 with seismic finite-source characterization, *Geophys. J. Int.,* **212**(1), 725–742.

Chung, A.I., Henson, I. & Allen, R.M., 2019. Optimizing earthquake early warning performance: ElarmS-3, *Seismol. Res Lett.,* **90**(2A), 727–743.

Cochran, E.S., Bunn, J., Minson, S.E., Baltay, A.S., Kilb, D.L., Kodera, Y. & Hoshiba, M., 2019. Event detection performance of the plum earthquake early warning algorithm in Southern California, *Bull. seism. Soc. Am.,* **109**(4), 1524–1541.

Cua, G. & Heaton, T.H., 2009. Characterizing average properties of southern California ground motion amplitudes and envelopes, EERL Report, Earthquake Engineering Research Laboratory, Pasadena, CA.

Dipartimento di Fisica, Universitá degli studi di Napoli Federico, II, 2005. Irpinia Seismic Network (ISNet), Istituto Nazionale di Geofisica e Vulcanologia (INGV).

Dolce, M. & Di Bucci, D., 2018. The 2016–2017 central apennines seismic sequence: analogies and differences with recent Italian earthquakes, in *Recent Advances in Earthquake Engineering in Europe: 16th European Conference on Earthquake Engineering-Thessaloniki 2018, Geotechnical, Geological and Earthquake Engineering,* pp. 603–638, ed. Pitilakis, K., Springer International Publishing.

EMERSITO Working Group, 2018. Seismic network for site effect studies in Amatrice Area (Central Italy) (SESAA), Istituto Nazionale di Geofisica e Vulcanologia (INGV). https://doi.org/10.13127/SD/7TXeGdo5X8.

Geological Survey-Provincia Autonoma di Trento, 1981. Trentino seismic network, International Federation of Digital Seismograph Networks, 10.7914/SN/ST.

Istituto Nazionale di Geofisica e Vulcanologia (INGV), 2008. INGV experiments network, Istituto Nazionale di Geofisica e Vulcanologia (INGV).

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), Agenzia Nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile (ENEA), 2018. Rete del Centro di Microzonazione Sismica (CentroMZ), sequenza sismica del 2016 in Italia Centrale, Istituto Nazionale di Geofisica e Vulcanologia (INGV), 10.13127/SD/ku7Xm12Yy9.

Istituto Nazionale di Geofisica e Vulcanologia (INGV), Italy, 2006. Rete sismica nazionale (RSN), Istituto Nazionale di Geofisica e Vulcanologia (INGV), doi.org/10.13127/SD/X0FXnH7QfY.

Jozinović, D., Lomax, A., Štajduhar, I. & Michelini, A., 2020. Rapid prediction of earthquake ground shaking intensity using raw waveform data and a convolutional neural network, *Geophys. J. Int.,* **222**(2), 1379–1389.

Karim, K.R. & Yamazaki, F., 2002. Correlation of JMA instrumental seismic intensity with strong motion parameters, *Earthq. Eng. Struct. Dyn.,* **31**(5), 1191–1212.

Kodera, Y., Yamada, Y., Hirano, K., Tamaribuchi, K., Adachi, S., Hayashimoto, N., Morimoto, M., Nakamura, M. & Hoshiba, M., 2018. The propagation of local undamped motion (PLUM) method: a simple and robust seismic wavefield estimation approach for earthquake early warning, *Bull. seism. Soc. Am.,* **108**(2), 983–1003.

Koketsu, K., Yokota, Y., Nishimura, N., Yagi, Y., Miyazaki, S., Satake, K., Fujii, Y., Miyake, H., Sakai, S., Yamanaka, Y., *et al.*, 2011. A unified source model for the 2011 Tohoku earthquake, *Earth planet. Sci. Lett.,* **310**(3–4), 480–487.

Kriegerowski, M., Petersen, G.M., Vasyura-Bathke, H. & Ohrnberger, M., 2019. A deep convolutional neural network for localization of clustered earthquakes based on multistation full waveforms, *Seismol. Res. Lett.,* **90**(2A), 510–516.

Kuyuk, H.S. & Allen, R.M., 2013. A global approach to provide magnitude estimates for earthquake early warning alerts, *Geophys. Res. Lett.,* **40**(24), 6329–6333.

Lomax, A., Michelini, A. & Jozinović, D., 2019. An investigation of rapid earthquake characterization using single-station waveforms and a convolutional neural network, *Seismol. Res. Lett.,* **90**(2A), 517–529.

MedNet Project Partner Institutions, 1990. Mediterranean Very Broadband Seismographic Network (MedNet), Mediterranean Very Broadband Seismographic Network (MedNet). Istituto Nazionale di Geofisica e Vulcanologia (INGV). https://doi.org/10.13127/SD/fBBBtDtd6q.

Meier, M.-A., 2017. How "good" are real-time ground motion predictions from earthquake early warning systems?, *J. geophys. Res.,* **122**(7), 5561–5577.

Meier, M.-A., Kodera, Y., Böse, M., Chung, A., Hoshiba, M., Cochran, E., Minson, S., Hauksson, E. & Heaton, T., 2020. How often can earthquake early warning systems alert sites with high-intensity ground motion?, *J. geophys. Res.,* **125**(2), e2019JB017718, doi:10.1029/2019JB017718.

Minson, S.E., Meier, M.-A., Baltay, A.S., Hanks, T.C. & Cochran, E.S., 2018. The limits of earthquake early warning: timeliness of ground motion estimates, *Sci. Adv.,* **4**(3), eaaq0504.

Minson, S.E., Baltay, A.S., Cochran, E.S., Hanks, T.C., Page, M.T., McBride, S.K., Milner, K.R. & Meier, M.-A., 2019. The limits of earthquake early warning accuracy and best alerting strategy, *Sci. Rep.,* **9**(1), 2478.

Mousavi, S.M. & Beroza, G.C., 2020. Bayesian-deep-learning estimation of earthquake location from single-station observations, *IEEE Transactions on Geoscience and Remote Sensing,* IEEE, **58,** 11, 8211–8224, 10.1109/TGRS.2020.2988770.

Mousavi, S.M. & Beroza, G.C., 2020. A machine-learning approach for earthquake magnitude estimation, *Geophys. Res. Lett.,* **47**(1), e2019GL085976, doi:10.1029/2019GL085976.

Münchmeyer, J., Bindi, D., Leser, U. & Tilmann, F., 2020. Fast earthquake assessment and earthquake early warning dataset for Italy, *GFZ Data Services,* V 1.0., doi: 10.5880/GFZ.2.4.2020.004.

Muthukumar, V., Vodrahalli, K., Subramanian, V. & Sahai, A., 2020. Harmless interpolation of noisy data in regression, *IEEE J. Select. Areas Inform. Theory,* arXiv:1903.09139 [cs.LG]

National Research Institute For Earth Science And Disaster Resilience, 2019. Nied k-net, kik-net, National Research Institute for Earth Science and Disaster Resilience, doi:10.17598/NIED.0004.

OGS (Istituto Nazionale Di Oceanografia E Di Geofisica Sperimentale), 2016. North-East Italy Seismic Network. International Federation of Digital Seismograph Networks.OGS (Istituto Nazionale Di Oceanografia E Di Geofisica Sperimentale), https://doi.org/10.7914/SN/OX.

OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale) and University of Trieste, 2002. North-East Italy Broadband Network. International Federation of Digital Seismograph Networks, https://doi.org/10.7914/SN/NI

Otake, R., Kurima, J., Goto, H. & Sawada, S., 2020. Deep learning model for spatial interpolation of real-time seismic intensity, *Seismol. Res. Lett.,* **91**(6), 3433–3443.

Presidency of Counsil of Ministers - Civil Protection Department, 1972. Italian Strong Motion Network. Presidency of Counsil of Ministers - Civil Protection Department, https://doi.org/10.7914/SN/IT.

RESIF - Réseau Sismologique et géodésique Français, 1995a. RESIF-RLBP French Broad-band network, RESIF-RAP strong motion network and other seismic stations in metropolitan France, doi.org/10.15778/resif.fr.

RESIF - Réseau Sismologique et géodésique Français, 1995b. Réseau accélérométrique permanent (french accelerometrique network) (rap).

Shabestari, K.T. & Yamazaki, F., 2001. A proposal of instrumental seismic intensity scale compatible with mmi evaluated from three-component acceleration records, *Earthq. Spectra,* **17**(4), 711–723.

Snoek, J., *et al.*, 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, in *Advances in Neural Information Processing Systems, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, pp. 13 969–13 980.

Sun, C., Shrivastava, A., Singh, S. & Gupta, A., 2017. Revisiting unreasonable effectiveness of data in deep learning era, in *Proceedings of the IEEE International Conference on Computer Vision,* pp. 843–852.

Universita della Basilicata, 2005. UniBAS, Italian National Institute of Geophysics and Volcanology (INGV).

University of Genova, 1967. Regional seismic network of north western Italy. international federation of digital seismograph networks, International Federation of Digital Seismograph Networks, 10.7914/SN/GU.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I., 2017. Attention is all you need, in *Advances in Neural Information Processing Systems,* pp. 5998–6008.

Wald, D.J., Quitoriano, V., Heaton, T.H. & Kanamori, H., 1999. Relationships between peak ground acceleration, peak ground velocity, and modified Mercalli intensity in California, *Earthq. Spectra,* **15**(3), 557–564.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Figure S1: Overview of the transformer earthquake alerting model, showing the input, the feature extraction, the feature combination, the PGA estimation and the output. For simplicity, not all layers are shown, but only their order and combination is visualized schematically. For the exact number of layers and the size of each layer please refer to tables S5 and S6. Please note that the number of input stations and the number of targets are both variable, due to the self-attention mechanism in the feature combination. Ten instances of this network are trained independently and the results ensemble-averaged.

Figure S2: 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-based approach. In addition, a GMPE with full catalogue information is included for reference. Values α 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 S3: Precision, recall and F1 score at different PGA thresholds for Italy including TEAM without domain adaptation. Threshold values α 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-based model. In addition the graph shows the performance of C-GMPE, a GMPE with full catalogue information for reference.

Figure S4: Warning time and hypocentral distance between station and event for each true alert at F1-optimal α. 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 S5: Scenario analysis of the 30 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 per cent g PGA threshold.

Figure S6: Scenario analysis of the 11 March 2011 $M_w = 9.1$ Tohoku earthquake, the largest event in the Japan data set. See Fig. 4 in the main paper for further explanations. The bottom row diagrams for this scenario analysis use a 2 per cent g PGA threshold.

Figure S7: Training and validation loss curves for the Japan TEAM model and the fine-tuning step of the Italy TEAM model. Each line shows the loss curve for one ensemble member with colours matching between training and validation curves. The models used are determined by the minimum validation loss and are denoted by black crosses. The models were evaluated after the training epoch indicated on the x-axis, that is, the leftmost point of each curve already includes one epoch of training.

Figure S8: Predictions and residuals of the GMPEs derived in this study. All PGA values are given as log units using m s^{-2}. 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 α were chosen to maximize F1 scores and are shown in the last column. The AUC value does not depend on the threshold α. PGA indicates the used PGA threshold.

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

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 α for TEAM and EPS to the optimal value in terms of F1 score.

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 six component traces, while for Italy it refers to three component traces.

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.

Table S6: Architecture of the transformer network. Please note that even though the transformer in TEAM does not apply dropout, we explicitly state this in the table, as transformers commonly use dropout.

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

APPENDIX: DATA SOURCES

We obtained our Japan catalogue and waveforms from NIED and the NIED KiK-net (National Research Institute For Earth Science And Disaster Resilience 2019). For our Italy data set we use the INGV catalogue and waveforms from the 3A (Istituto Nazionale di Geofisica e Vulcanologia (INGV) 2018), BA (Universita della Basilicata 2005), FR (RESIF - Réseau Sismologique et géodésique Français 1995a), GU (University of Genova 1967), IT (Presidency of Counsil of Ministers - Civil Protection Department 1972), IV (Istituto Nazionale di Geofisica e Vulcanologia (INGV) 2006), IX (Dipartimento di Fisica, Universitá degli studi di Napoli Federico II 2005), MN (MedNet Project Partner Institutions 1990), NI (OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale) 2002), OX (OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale) 2016), RA (RESIF - Réseau Sismologique et géodésique Français 1995b), ST (Geological Survey-Provincia Autonoma di Trento 1981), TV (Istituto Nazionale di Geofisica e Vulcanologia (INGV) 2008) and XO (EMERSITO Working Group 2018) networks.