Münchmeyer, J., Bindi, D., Leser, U., Tilmann, F. (2021): Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network. - Geophysical Journal International, 226, 2, 1086-1104.

https://doi.org/10.1093/gji/ggab139

Geophysical Journal International

Geophys. J. Int. (2021) **226,** 1086–1104
Advance Access publication 2021 April 12
GJI Seismology

doi: 10.1093/gji/ggab139

Earthquake magnitude and location estimation from real time seismic waveforms with a transformer network

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

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

Accepted 2021 March 25. Received 2021 March 10; in original form 2021 January 6

SUMMARY

Precise real time estimates of earthquake magnitude and location are essential for early warning and rapid response. While recently multiple deep learning approaches for fast assessment of earthquakes have been proposed, they usually rely on either seismic records from a single station or from a fixed set of seismic stations. Here we introduce a new model for real-time magnitude and location estimation using the attention based transformer networks. Our approach incorporates waveforms from a dynamically varying set of stations and outperforms deep learning baselines in both magnitude and location estimation performance. Furthermore, it outperforms a classical magnitude estimation algorithm considerably and shows promising performance in comparison to a classical localization algorithm. Our model is applicable to real-time prediction and provides realistic uncertainty estimates based on probabilistic inference. In this work, we furthermore conduct a comprehensive study of the requirements on training data, the training procedures and the typical failure modes. Using three diverse and large scale data sets, we conduct targeted experiments and a qualitative error analysis. Our analysis gives several key insights. First, we can precisely pinpoint the effect of large training data; for example, a four times larger training set reduces average errors for both magnitude and location prediction by more than half, and reduces the required time for real time assessment by a factor of four. Secondly, the basic model systematically underestimates large magnitude events. This issue can be mitigated, and in some cases completely resolved, by incorporating events from other regions into the training through transfer learning. Thirdly, location estimation is highly precise in areas with sufficient training data, but is strongly degraded for events outside the training distribution, sometimes producing massive outliers. Our analysis suggests that these characteristics are not only present for our model, but for most deep learning models for fast assessment published so far. They result from the black box modeling and their mitigation will likely require imposing physics derived constraints on the neural network. These characteristics need to be taken into consideration for practical applications.

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

1 INTRODUCTION

Recently, multiple studies investigated deep learning on raw seismic waveforms for the fast assessment of earthquake parameters, such as magnitude (e.g. Lomax *et al.* 2019; Mousavi & Beroza 2020; van den Ende & Ampuero 2020), location (e.g. Kriegerowski *et al.* 2019; Mousavi & Beroza 2019; van den Ende & Ampuero 2020) and peak ground acceleration (e.g. Jozinović *et al.* 2020). Deep learning is well suited for these tasks, as it does not rely on manually selected features, but can learn to extract relevant information from the raw input data. This property allows the models to use the full information contained in the waveforms of an event. However, the models published so far use fixed time windows and can not be applied to data of varying length without retraining. Similarly, except the model by van den Ende & Ampuero (2020), all models process either waveforms from only a single seismic station or rely on a fixed set of seismic stations defined at training time. The model by van den Ende & Ampuero (2020) enables the use of a variable station set, but combines measurements from multiple stations using a simple pooling mechanism. While it has not been studied so far in

a seismological context, it has been shown in the general domain that set pooling architectures are in practice limited in the complexity of functions they can model (Lee *et al.* 2019).

Here we introduce a new model for magnitude and location estimation based on the architecture recently introduced for the transformer earthquake alerting model (TEAM, Münchmeyer *et al.* 2021), a deep learning based earthquake early warning model. While TEAM estimated PGA at target locations, our model estimates magnitude and hypocentral location of the event. We call our adaptation TEAM-LM, TEAM for location and magnitude estimation. We use TEAM as a basis due to its flexible multistation approach and its ability to process incoming data effectively in real-time, issuing updated estimates as additional data become available. Similar to TEAM, TEAM-LM uses mixture density networks to provided probability distributions rather than merely point estimates as predictions. For magnitude estimation, our model outperforms two state of the art baselines, one using deep learning (van den Ende & Ampuero 2020) and one classical approach (Kuyuk & Allen 2013). For location estimation, our model outperforms a deep learning baseline (van den Ende & Ampuero 2020) and shows promising performance in comparison to a classical localization algorithm.

We note a further deficiency of previous studies for deep learning in seismology. Many of these pioneering studies focused their analysis on the average performance of the proposed models. Therefore, little is known about the conditions under which these models fail, the impact of training data characteristics, the possibility of sharing knowledge across world regions, and of specific training strategies. All of these are of particular interest when considering practical application of the models.

To address these issues and provide guidance for practitioners, we perform a comprehensive evaluation of TEAM-LM on three large and diverse data sets: a regional broad-band data set from Northern Chile, a strong motion data set from Japan and another strong motion data set from Italy. These data sets differ in their seismotectonic environment (North Chile and Japan: subduction zones; Italy: dominated by both convergent and divergent continental deformation), their spatial extent (North Chile: regional scale; Italy and Japan: national catalogues), and the instrument type (North Chile: broadband, Italy and Japan: strong motion). All three data sets contain hundreds of thousands of waveforms. North Chile is characterized by a relatively sparse station distribution, but a large number of events and a low magnitude of completeness. There are far more stations in the Italy and Japan data sets, but a smaller number of earthquakes. This selection of diverse data sets allows for a comprehensive analysis, giving insights for different use cases. Our targeted experiments show that the characteristics are rooted in the principle structure used by TEAM-LM, that is the black box approach of learning a very flexible model from data, without imposing any physical constraints. As this black box approach is common to all current fast assessment models using deep learning, they can be transferred to these models. This finding is further backed by comparison to the results from previous studies.

2 DATA AND METHODS

2.1 Data sets

For this study we use three data sets (Table 1, Fig. 1): one from Northern Chile, one from Italy and one from Japan. The Chile data set is based on the catalogue by Sippl *et al.* (2018) with the magnitude values from Münchmeyer *et al.* (2020b). While there were minor changes in the seismic network configuration during the time covered by the catalogue, the station set used in the construction of this catalogue had been selected to provide a high degree of stability of the locations accuracy throughout the observational period (Sippl *et al.* 2018). Similarly, the magnitude scale has been carefully calibrated to achieve a high degree of consistency in spite of significant variations of attenuation (Münchmeyer *et al.* 2020b). This data set therefore contains the highest quality labels among the data sets in this study. For the Chile data set, we use broad-band seismogramms from the fixed set of 24 stations used for the creation of the original catalogue and magnitude scale. Although the Chile data set has the smallest number of stations of the three data sets, it comprises three to four times as many waveforms as the other two due to the large number of events.

The data sets for Italy and Japan are more focused on early warning, containing fewer events and only strong motion waveforms. They are based on catalogues from the INGV (ISIDe Working Group 2007) and the NIED KiKNet (National Research Institute For Earth Science And Disaster Resilience 2019), respectively. The data sets each encompass a larger area than the Chile data set and include waveforms from significantly more stations. In contrast to the Chile data sets, the station coverage differs strongly between different events, as only stations recording the event are considered. In particular, KiKNet stations do not record continuous waveforms, but operate in trigger mode, only saving waveforms if an event triggered at the station. For Japan each station comprises two sensors, one at the surface and one borehole sensor. Therefore for Japan we have six component recordings (three surface, three borehole) available instead of the three component recordings for Italy and Chile. A full list of seismic networks used in this study can be found in the appendix (Table A1).

For each data set we use the magnitude scale provided in the catalogue. For the Chile catalogue, this is M_A, a peak displacement based scale, but without the Wood-Anderson response and therefore saturation-free for large events (Münchmeyer *et al.* 2020b; Deichmann 2018). For Japan M_{JMA} is used. M_{JMA} combines different magnitude scales, but similarly to M_A primarily uses horizontal peak displacement (Doi 2014). For Italy the catalogue provides different magnitude types approximately dependent on the size of the event: M_L (>90 per cent of the events), M_W (<10 per cent) and m_b (<1 per cent). We note that while the primary magnitude scales for all data sets are peak-displacement based, the precision of the magnitudes vary, with the highest precision for Chile. This might lead to slightly worse magnitude estimation performance for Italy and Japan.

For all data sets the data were not subselected based on the type of seismicity but only based on the location (for Chile and Italy) or depending if they triggered (Japan). This guarantees that, even though we made use of a catalogue to assemble our training data, the resulting data sets are suitable for training and assessing methods geared at real-time applications without any prior knowledge about the earthquakes. We focus on earthquake characterization and do not discuss event detection or separation from noise; we refer the interested reader to Perol *et al.* (2018) and Mousavi *et al.* (2019).

We split each data set into training, development and test set. For Chile and Japan we apply a simple chronological split with approximate ratios of 60:10:30 between training, development and test set, with the most recent events in the test set. As the last 30 per cent of the Italy data set consist of less interesting events for early warning, we instead use all events from 2016 as test set and the remaining events as training and development sets. We reserve all of 2016 for testing, as it contains a long seismic sequence in central

Table 1. Overview of the data sets. 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. Magnitudes are given with two digit precision for Chile, as the precision of the underlying catalogue is higher than for Italy and Japan. The Italy data set uses different magnitudes for different events, which are M_L (>90 per cent of the events), M_W (<10 per cent) and m_b (<1 per cent). For depth and distance minimum, median and maximum are stated. Distance refers to the epicentral distance between stations and events. Note that the count of traces refers to the number of waveform-triplets (for three components, or group of six waveforms for the Japanese stations). The sensor types are broadband (BB) and strong motion (SM).

	Chile	Italy	Japan
Years	2007–2014	2008–2019	1997–2018
Training	01/2007–08/2011	01/2008–12/2015 & 01/2017–12/2019	01/1997–03/2012
Test	08/2012–12/2014	01/2016–12/2016	08/2013–12/2018
Magnitudes	1.21–8.27	2.7–6.5	2.7–9.0
Magnitude scale	M_A	M_L, M_W, m_b	M_{JMA}
Depth [km]	0–102–183	0–10–617	0–19–682
Distance [km]	0.1–180–640	0.1–180–630	0.2–120–3190
Events	96 133	7055	13 512
Unique stations	24	1,080	697
Traces	1 605 983	494 183	372 661
Traces per event	16.7	70.3	27.6
Sensor type	BB	SM	SM & SM-borehole
Catalogue source	Münchmeyer *et al.* (2020b)	INGV	NIED

Italy with two main shocks in August ($M_W = 6.5$) and October ($M_W = 6.0$). Notably, the largest event in the test set is significantly larger than the largest event in the training set ($M_W = 6.1$ L'Aquila event in 2007), representing a challenging test case. For Italy, we assign the remaining events to training and development set randomly with a 6:1 ratio.

2.2 The TEAM for magnitude and location

We build a model for real time earthquake magnitude and location estimation based on the core ideas of the TEAM, as published in Münchmeyer *et al.* (2021). TEAM is an end-to-end peak ground acceleration (PGA) model calculating probabilistic PGA estimates based on incoming waveforms from a flexible set of stations. It uses the transformer network method (Vaswani *et al.* 2017), an attention based neural network which was developed in the context of natural language processing (NLP), at the core of its algorithm. Here, we adapt TEAM to calculate real time probabilistic estimates of event magnitude and hypocentral location. As our model closely follows the architecture and key ideas of TEAM, we use the name TEAM-LM to refer to the location and magnitude estimation model.

Similar to TEAM, TEAM-LM consists of three major components (Fig. 2): a feature extraction, which generates features from raw waveforms at single stations, a feature combination, which aggregates features across multiple stations, and an output estimation. Here, we briefly discuss the core ideas of the TEAM architecture and training and put a further focus on the necessary changes for magnitude and location estimation. For a more detailed account of TEAM and TEAM-LM we refer to Münchmeyer *et al.* (2021), Tables S1–S3 and the published implementation.

The input to TEAM consists of three component seismogramms from multiple stations and their locations. TEAM aligns all seismogramms to start and end at the same times t_0 and t_1. We choose t_0 to be 5 s before the first P arrival at any station. This allows the model to understand the noise conditions at all stations. We limit t_1 to be at latest $t_0 + 30$ s. In a real-time scenario t_1 is the current time, that is the available amount of waveforms, and we use the same approach to imitate real-time waveforms in training and evaluation. The waveforms are padded with zeros to a length of 30 s to achieve constant length input to the feature extraction.

TEAM uses a CNN architecture for feature extraction, which is applied separately at each station. The architecture consists of several convolution and pooling layers, followed by a multilayer perceptron (Table S1). To avoid scaling issues, each input waveform is normalized through division by its peak amplitude. As the amplitude is expected to be a key predictor for the event magnitude, we provide the logarithm of the peak amplitude as a further input to the multilayer perceptron inside the feature extraction network. We ensure that this transformation does not introduce a knowledge leak by calculating the peak amplitude only based on the waveforms until t_1. The full feature extraction returns one vector for each station, representing the measurements at the station.

The feature vectors from multiple stations are combined using a transformer network (Vaswani *et al.* 2017). Transformers are attention based neural networks, originally introduced for natural language processing. A transformer takes a set of n vectors as input, and outputs again n vectors which now incorporate the context of each other. The attention mechanism allows the transformer to put special emphasis on inputs that it considers particularly relevant and thereby model complex interstation dependencies. Importantly, the parameters of the transformer are independent of the number of input vectors n, allowing to train and apply a transformer on variable station sets. To give the transformer a notion of the position of the stations, TEAM encodes the latitude, longitude and elevation of the stations using a sinusoidal embedding and adds this embedding to the feature vectors.

TEAM adds the position embeddings of the PGA targets as additional inputs to the transformer. In TEAM-LM, we aim to extract information about the event itself, where we do not know the position in advance. To achieve this, we add an event token, which is a vector with the same dimensionality as the positional embedding of a station location, and which can be thought of as a query vector. This approach is inspired by the so-called sentence tokens in NLP that are used to extract holistic information on a sentence (Devlin

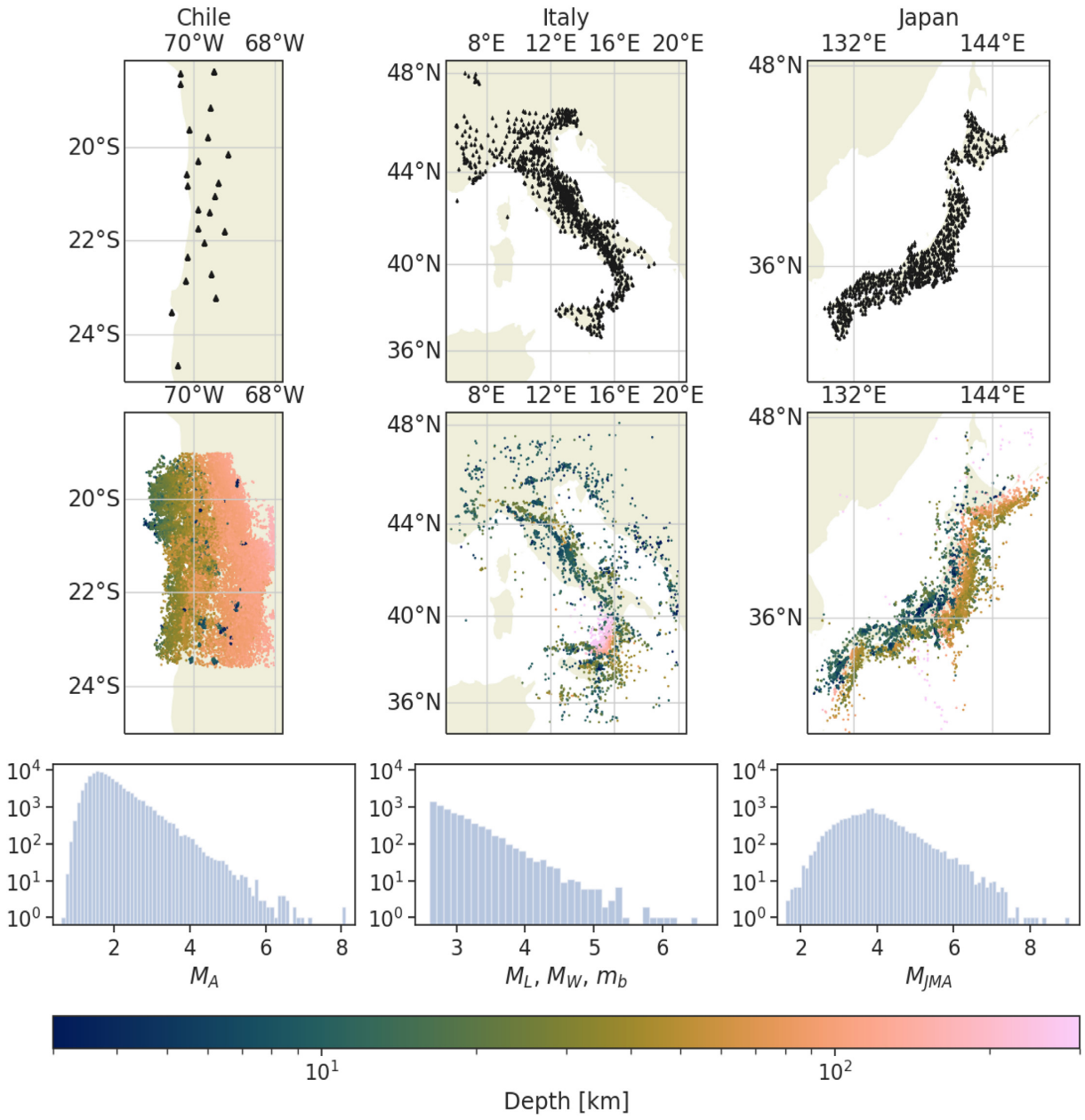

Figure 1. Overview of the data sets. The top row shows the spatial station distribution, the second tow the spatial event distribution. The event depth is encoded using colour. Higher resolution versions of the maps can be found in the supplementary material (Figs S1, S2 and S3). The bottom row shows the distributions of the event magnitudes. The magnitude scales are the peak displacement based M_A, local magnitude M_L, moment magnitude M_W, body wave magnitude m_b and M_{JMA}, a magnitude primarily using peak displacement.

et al. 2018). The elements of this event query vector are learned during the training procedure.

From the transformer output, we only use the output corresponding to the event token, which we term event embedding and which is passed through another multi-layer perceptron predicting the parameters and weights of a mixture of Gaussians (Bishop 1994). We use $N = 5$ Gaussians for magnitude and $N = 15$ Gaussians for location estimation. For computational and stability reasons, we constrain the covariance matrix of the individual Gaussians for location estimation to a diagonal matrix to reduce the output dimensionality. Even though uncertainties in latitude, longitude and depth are known to generally be correlated, this correlation can be modeled with diagonal covariance matrices by using the mixture.

The model is trained end-to-end using a log-likelihood loss with the Adam optimizer (Kingma & Ba 2014). We train separate models for magnitude and for location. As we observed difficulties in the onset of the optimization when starting from a fully random initialization, we pretrain the feature extraction network. To this end we add a mixture density network directly after the feature extraction and train the resulting network to predict magnitudes from single station waveforms. We then discard the mixture density network and use the weights of the feature extraction as initialization for the end-to-end training. We use this pretraining method for both magnitude and localization networks.

Similarly to the training procedure for TEAM we make extensive use of data augmentation during training. First, we randomly

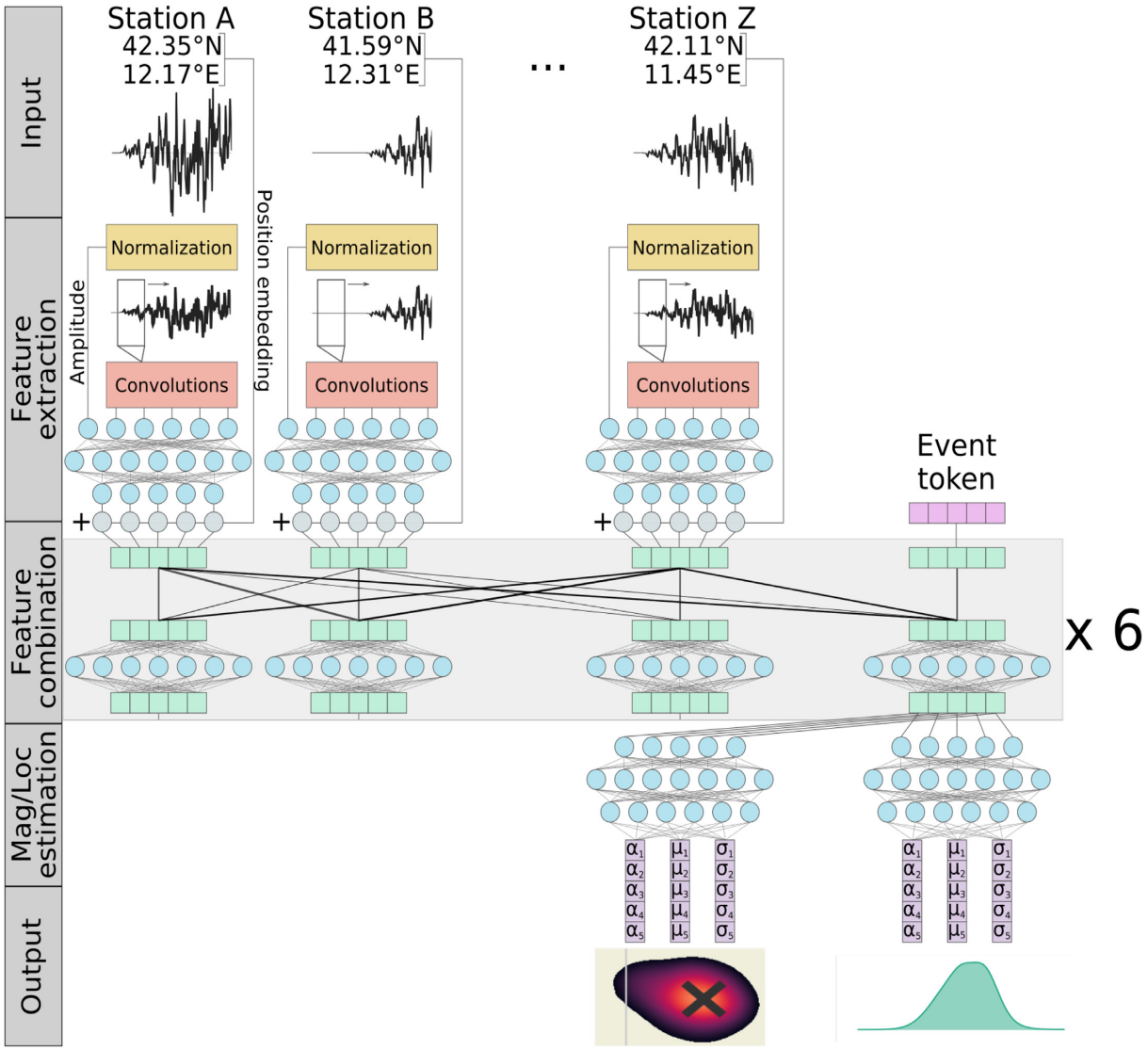

Figure 2. Overview of the adapted transformer earthquake alerting model, showing the input, the feature extraction, the feature combination, the magnitude/location estimation and the output. For simplicity, not all layers are shown, but only their order and combination is visualized schematically. For the exact number of layers and the size of each layer please refer to Tables S1 to S3. Please note that the number of input stations is variable, due to the self-attention mechanism in the feature combination.

select a subset of up to 25 stations from the available station set. We limit the maximum number to 25 for computational reasons. Secondly, we apply temporal blinding, by zeroing waveforms after a random time t_1. This type of augmentation allows TEAM-LM to be applied to real time data. We note that this type of temporal blinding to enable real time predictions would most likely work for the previously published CNN approaches as well. To avoid knowledge leaks for Italy and Japan, we only use stations as inputs that triggered before time t_1 for these data sets. This is not necessary for Chile, as there the maximum number of stations per event is below 25 and waveforms for all events are available for all stations active at that time, irrespective of whether the station actually recorded the event. Thirdly, we oversample large magnitude events, as they are strongly underrepresented in the training data set. We discuss the effect of this augmentation in further detail in the Results section. In contrast to the station selection during training, in evaluation we always use the 25 stations picking first. Again, we only use stations and their waveforms as input once they triggered, thereby ensuring that the station selection does not introduce a knowledge leak.

2.3 Baseline methods

Recently, van den Ende & Ampuero (2020) suggested a deep learning method capable of incorporating waveforms from a flexible set of stations. Their architecture uses a similar CNN based feature extraction as TEAM-LM. In contrast to TEAM-LM, for feature combination it uses maximum pooling to aggregate the feature vectors from all stations instead of a transformer. In addition they do not add predefined position embeddings, but concatenate the feature vector for each station with the location coordinates and apply a multilayer perceptron to get the final feature vectors for each station. The model of van den Ende & Ampuero (2020) is both trained and evaluated on 100 s long waveforms. In its original form it is therefore not suitable for real time processing, although the real

time processing could be added with the same zero-padding approach used for TEAM and TEAM-LM. The detail differences in the CNN structure and the real-time processing capability make a comparison of the exact model of van den Ende & Ampuero (2020) to TEAM-LM difficult.

To still compare TEAM-LM to the techniques introduced in this approach, we implemented a model based on the key concepts of van den Ende & Ampuero (2020). As we aim to evaluate the performance differences from the conceptual changes, rather than different hyperparameters, for example the exact size and number of the convolutional layers, we use the same architecture as TEAM-LM for the feature extraction and the mixture density output. Additionally we train the model for real time processing using zero padding. In comparison to TEAM-LM we replace the transformer with a maximum pooling operation and remove the event token.

We evaluate two different representations for the position encoding. In the first, we concatenated the positions to the feature vectors as proposed by van den Ende & Ampuero (2020). In the second, we add the position embeddings element-wise to the feature vectors as for TEAM-LM. In both cases, we run a three-layer perceptron over the combined feature and position vector for each station, before applying the pooling operation.

We use the fast magnitude estimation approach (Kuyuk & Allen 2013) as a classical, that is non-deep-learning, baseline for magnitude. The magnitude is estimated from the horizontal peak displacement in the first seconds of the *P* wave. As this approach needs to know the hypocentral distance, it requires knowledge of the event location. We simply provide the method with the catalogue hypocentre. While this would not be possible in real time, and therefore gives the method an unfair advantage over the deep learning approaches, it allows us to focus on the magnitude estimation capabilities. Furthermore, in particular for Italy and Japan, the high station density usually allows for sufficiently well constrained location estimates at early times. For a full description of this baseline, see supplement section SM 1.

As a classical location baseline we use NonLinLoc (Lomax *et al.* 2000) with the 1-D velocity models from Graeber & Asch (1999) (Chile), Ueno *et al.* (2002) (Japan) and Matrullo *et al.* (2013) (Italy). For the earliest times after the event detection usually only few picks are available. Therefore, we apply two heuristics. Until at least 3/5/5 (Chile/Japan/Italy) picks are available, the epicentre is estimated as the arithmetic mean of the stations with picked arrivals so far, while the depth is set to the median depth in the training data set. Until at least 4/7/7 picks are available, we apply NonLinLoc, but fix the depth to the median depth in the data set. We require higher numbers of picks for Italy and Japan, as the pick quality is lower than in Chile but the station density is higher. This leads to worse early NonLinLoc estimates in Italy and Japan compared to Chile, but improves the performance of the heuristics.

3 RESULTS

3.1 Magnitude estimation performance

We first compare the estimation capabilities of TEAM-LM to the baselines in terms of magnitude (Fig. 3). We evaluate the models at fixed times $t = 0.5$, 1, 2, 4, 8, 16 and 25 s after the first *P* arrival at any station in the network. In addition to presenting selected results here, we provide tables with the results of further experiments in the supplementary material (Tables S5–S15).

TEAM-LM outperforms the classical magnitude baseline consistently. On two data sets, Chile and Italy, the performance of TEAM-LM with only 0.5 s of data is superior to the baseline with 25 s of data. Even on the third data set, Japan, TEAM-LM requires only approximately a quarter of the time to reach the same precision as the classical baseline and achieves significantly higher precision after 25 s. The RMSE for TEAM-LM stabilizes after 16 s for all data sets with final values of 0.08 m.u. for Chile, 0.20 m.u. for Italy and 0.22 m.u. for Japan. The performance differences between TEAM-LM and the classical baseline result from the simplified modelling assumptions for the baseline. While the relationship between early peak displacement and magnitude only holds approximately, TEAM-LM can extract more nuanced features from the waveform. In addition, the relationship for the baseline was originally calibrated for a moment magnitude scale. While all magnitude scales have an approximate 1:1 relationship with moment magnitude, this might introduce further errors.

We further note that the performance of the classical baseline for Italy are consistent with the results reported by Festa *et al.* (2018). They analysed early warning performance in a slightly different setting, looking only at the nine largest events in the 2016 Central Italy sequence. However, they report a RMSE of 0.28 m.u. for the PRESTO system 4 s after the first alert, which matches approximately the 8 s value in our analysis. Similarly, Leyton *et al.* (2018) analyse how fast magnitudes can be estimated in subductions zones, and obtain values of 0.01 ± 0.28 across all events and -0.70 ± 0.30 for the largest events ($M_W > 7.5$) at 30 s after origin time. This matches the observed performance of the classical baseline for Japan. For Chile, our classical baseline performs considerably worse, likely caused by the many small events with bad SNR compared to the event set considered by Leyton *et al.* (2018). However, TEAM-LM still outperforms the performance numbers reported by Leyton *et al.* (2018) by a factor of more than 2.

Improvements for TEAM-LM in comparison to the deep learning baseline variants are much smaller than to the classical approach. Still, for the Japan data set at late times, TEAM-LM offers improvements of up to 27 per cent for magnitude. For the Italy data set, the baseline variants are on par with TEAM-LM. For Chile, only the baseline with position embeddings is on par with TEAM-LM. Notably, for the Italy and Japan data sets, the standard deviation between multiple runs with different random model initialization is considerably higher for the baselines than for TEAM-LM (Fig. 3, error bars). This indicates that the training of TEAM-LM is more stable with regard to model initialization.

The gains of TEAM-LM can be attributed to two differences: the transformer for station aggregation and the position embeddings. In our experiments we ruled out further differences, for example size and structure of the feature extraction CNN, by using identical network architectures for all parts except the feature combination across stations. Regarding the impact of position embeddings, the results do not show a consistent pattern. Gains for Chile seem to be solely caused by the position embeddings; gains for Italy are generally lowest, but again the model with position embeddings performs better; for Japan the concatenation model performs slightly better, although the variance in the predictions makes the differences non-significant. We suspect these different patterns to be caused by the different catalogue and network sizes as well as the station spacing.

We think that gains from using a transformer can be explained with its attention mechanism. The attention allows the transformer to focus on specific stations, for example the stations which have recorded the longest waveforms so far. In contrast, the maximum pooling operation is less flexible. We suspect that the high gains

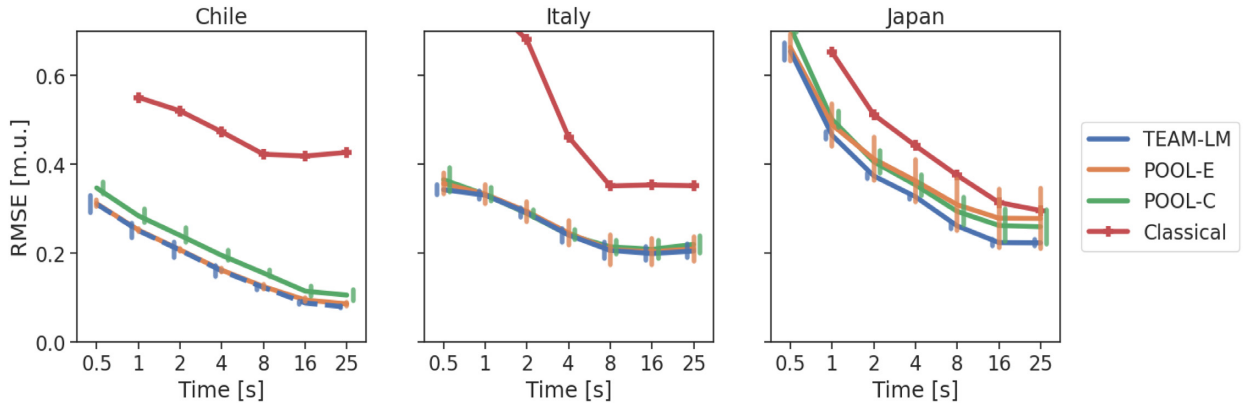

Figure 3. RMSE of the mean magnitude predictions from TEAM-LM, the pooling model with sinusoidal location embeddings (POOL-E), the pooling model with concatenated positions (POOL-C) and the classical baseline method. The time indicates the time since the first *P* arrival at any station, the RMSE is provided in magnitude units [m.u.]. Error bars indicate ±1 standard deviation when training the model with different random initializations. For better visibility error bars are provided with a small x-offset. Standard deviations were obtained from six realizations. Note that the uncertainty of the provided means is by a factor $\sqrt{6}$ smaller than the given standard deviation, due to the number of samples. We provide no standard deviation for the baseline, as it does not depend on a model initialization.

for Japan result from the wide spatial distribution of seismicity and therefore very variable station distribution. While in Italy most events are in Central Italy and in Chile the number of stations are limited, the seismicity in Japan occurs along the whole subduction zone with additional onshore events. This complexity can likely be handled better with the flexibility of the transformer than using a pooling operation. This indicates that the gains from using a transformer compared to pooling with position embeddings are likely modest for small sets of stations, and highest for large heterogeneous networks.

In many use cases, the performance of magnitude estimation algorithms for large magnitude events is of particular importance. In Fig. 4, we compare the RMSE of TEAM-LM and the classical baselines binned by catalogue magnitude into small, medium and large events. For Chile/Italy/Japan we count events as small if their magnitude is below 3.5/3.5/4 and as large if their magnitude is at least 5.5/5/6. We observe a clear dependence on the event magnitude. For all data sets the RMSE for large events is higher than for intermediate sized events, which is again higher than for small events. On the other hand the decrease in RMSE over time is strongest for larger events. This general pattern can also be observed for the classical baseline, even though the difference in RMSE between magnitude buckets is smaller. As both variants of the deep learning baseline show very similar trends to TEAM-LM, we omit them from this discussion.

We discuss two possible causes for these effects: (i) the magnitude distribution in the training set restricts the quality of the model optimization, (ii) inherent characteristics of large events. Cause (i) arise from the Gutenberg-Richter distribution of magnitudes. As large magnitudes are rare, the model has significantly less examples to learn from for large magnitudes than for small ones. This should impact the deep learning models the strongest, due to their high number of parameters. Cause (ii) has a geophysical origin. As large events have longer rupture durations, the information gain from longer waveform recordings is larger for large events. At which point during the rupture the final rupture size can be accurately predicted is a point of open discussion (e.g. Meier *et al.* 2017; Colombelli *et al.* 2020). We probe the likely individual contributions of these causes in the following.

Estimations for large events not only show lower precision, but are also biased (Fig. 5, middle column). For Chile and Italy a clear saturation sets in for large events. Interestingly the saturation starts at different magnitudes, which are around 5.5 for Italy and 6.0 for Chile. For Japan, events up to magnitude 7 are predicted without obvious bias. This saturation behavior is not only visible for TEAM-LM, but has also been observed in prior studies, for example in Mousavi & Beroza (2020, their figs 3, 4). In their work, with a network trained on significantly smaller events, the saturation already set in around magnitude 3. The different saturation thresholds indicate that the primary cause for saturation is not the longer rupture duration of large events or other inherent event properties, as in cause (ii), but is instead likely related to the low number of training examples for large events, rendering it nearly impossible to learn their general characteristics, as in cause (i). This explanation is consistent with the much higher saturation threshold for the Japanese data set, where the training data set contains a comparably large number of large events, encompassing the year 2011 with the Tohoku event and its aftershocks.

As a further check of cause (i), we trained models without upsampling large magnitude events during training, thereby reducing the occurrence of large magnitude events to the natural distribution observed in the catalogue (Fig. 5, left-hand column). While the overall performance stays similar, the performance for large events is degraded on each of the data sets. Large events are on average underestimated even more strongly. We tried different upsampling rates, but were not able to achieve significantly better performance for large events than the configuration of the preferred model presented in the paper. This shows that upsampling yields improvements, but can not solve the issue completely, as it does not introduce actual additional data. On the other hand, the performance gains for large events from upsampling seem to cause no observable performance drop for smaller event. As the magnitude distribution in most regions approximately follows a Gutenberg–Richter law with $b \approx 1$, upsampling rates similar to the ones used in this paper will likely work for other regions as well.

The expected effects of cause (ii), inherent limitations to the predictability of rupture evolutions, can be approximated with physical models. To this end, we look at the model from Trugman *et al.* (2019), which suggests a weak rupture predictability, that is predictability after 50 per cent of the rupture duration. Trugman *et al.* (2019) discuss the saturation of early peak displacement and the effects for magnitude predictions based on peak displacements.

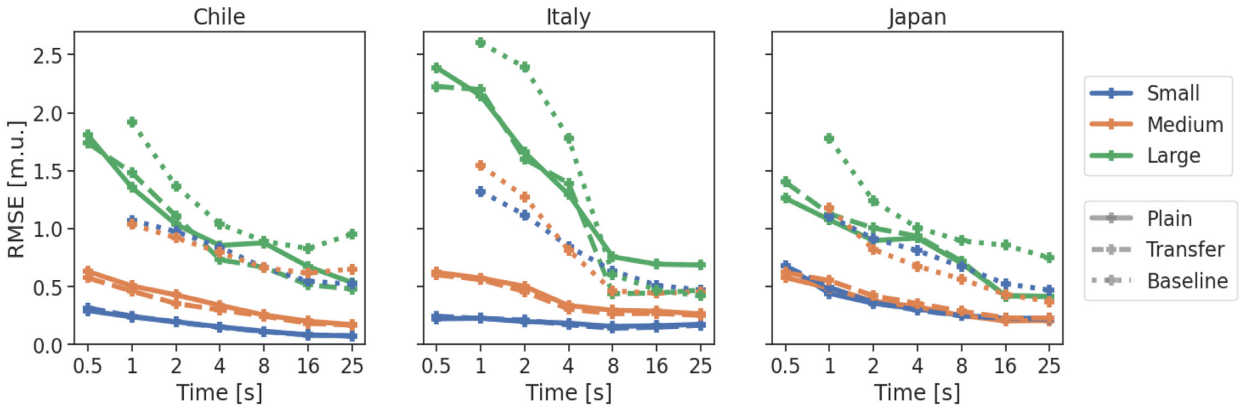

Figure 4. RMSE comparison of the TEAM-LM mean magnitude predictions for different magnitude buckets. Linestyles indicate the model type: trained only on the target data (solid line), using transfer learning (dashed), classical baseline (dotted). For Chile/Italy/Japan we count events as small if their magnitude is below 3.5/3.5/4 and as large if their magnitude is at least 5.5/5/6. The time indicates the time since the first P arrival at any station, the RMSE is provided in magnitude units [m.u.].

Following their model, we would expect magnitude saturation at approximately magnitude 5.7 after 1 s; 6.4 after 2 s; 7.0 after 4 s; 7.4 after 8 s. Comparing these results to Fig. 5, the saturation for Chile and Italy clearly occurs below these thresholds, and even for Japan the saturation is slightly below the modeled threshold. As we assumed a model with only weak rupture predictability, this makes it unlikely that the observed saturation is caused by limitations of rupture predictability. This implies that our result does not allow any inference on rupture predictability, as the possible effects of rupture predictability are masked by the data sparsity effects.

3.2 Location estimation performance

We evaluate the epicentral error distributions in terms of the 50th, 90th, 95th and 99th error percentiles (Fig. 6). In terms of the median epicentral error, TEAM-LM outperforms all baselines in all cases, except for the classical baseline at late times in Italy. For all data sets, TEAM-LM shows a clear decrease in median epicentral error over time. The decrease is strongest for Chile, going from 19 km at 0.5 s to 2 km at 25 s. For Italy the decrease is from 7 to 2 km, for Japan from 22 to 14 km. For all data sets the error distributions are heavy tailed. While for Chile even the errors at high quantiles decrease considerably over time, these quantiles stay nearly constant for Italy and Japan.

Similar to the difficulties for large magnitudes, the characteristics of the location estimation point to insufficient training data as source of errors. The Chile data set covers the smallest region and has by far the lowest magnitude of completeness, leading to the highest event density. Consequently the location estimation performance is best and outliers are very rare. For the Italy and Japan data sets, significantly more events occurred in regions with only few training events, causing strong outliers. The errors for the Japanese data set are highest, presumably related to the large number of offshore events with consequently poor azimuthal coverage.

We expect a further difference from the number of unique stations. While for a small number of unique stations, as in the Chile data set, the network can mostly learn to identify the stations using their position embeddings, it might be unable to do so for a larger number of stations with fewer training examples per station. Therefore the task is significantly more complicated for Italy and Japan, where the concept of station locations has to be learned simultaneously to the localization task. This holds true even though we encode the station locations using continuously varying position embeddings. Furthermore, whereas for moderate and large events waveforms from all stations of the Chilean network will show the earthquake and can contribute information, the limitation to 25 stations of the current TEAM-LM implementation does not allow a full exploitation of the information contained in the hundreds of recordings of larger events in the Japanese and Italian data sets. This will matter in particular for out-of-network events, where the wavefront curvature and thus event distance can only be estimated properly by considering stations with later arrivals.

Looking at the classical baseline, we see that it performs considerably worse than TEAM-LM in the Chile data set in all location quantiles, better than TEAM-LM in all but the highest quantiles at late times in the Italy data set, and worse than TEAM-LM at late times in the Japan data set. This strongly different behavior can largely be explained with the pick quality and the station density in the different data sets. While the Chile data set contains high quality automatic picks, obtained using the MPX picker (Aldersons 2004), the Italy data set uses a simple STA/LTA and the Japan data set uses triggers from KiKNet. This reduces location quality for Italy and Japan, in particular in the case of a low number of picks available for location. On the other hand, the very good median performance of the classical approach for Italy can be explained from the very high station density, giving a strong prior on the location. An epicentral error of around 2 km after 8 s is furthermore consistent with the results from Festa *et al.* (2018). Considering the reduction in error due to the high station density in Italy, we note that the wide station spacing in Chile likely caused higher location errors than would be achievable with a denser seismic network designed for early warning.

In addition to the pick quality, the assumption of a 1-D velocity model for NonLinLoc introduces a systematic error into the localization, in particular for the subduction regions in Japan and Chile where the 3-D structure deviates considerably from the 1-D model. Because of these limitations the classical baseline could be improved by using more proficient pickers or fine-tuned velocity models. Nonetheless, in particular the results from Chile, where the classical baseline has access to high quality *P*-picks, suggest that TEAM-LM can, given sufficient training data, outperform classical real-time localization algorithms.

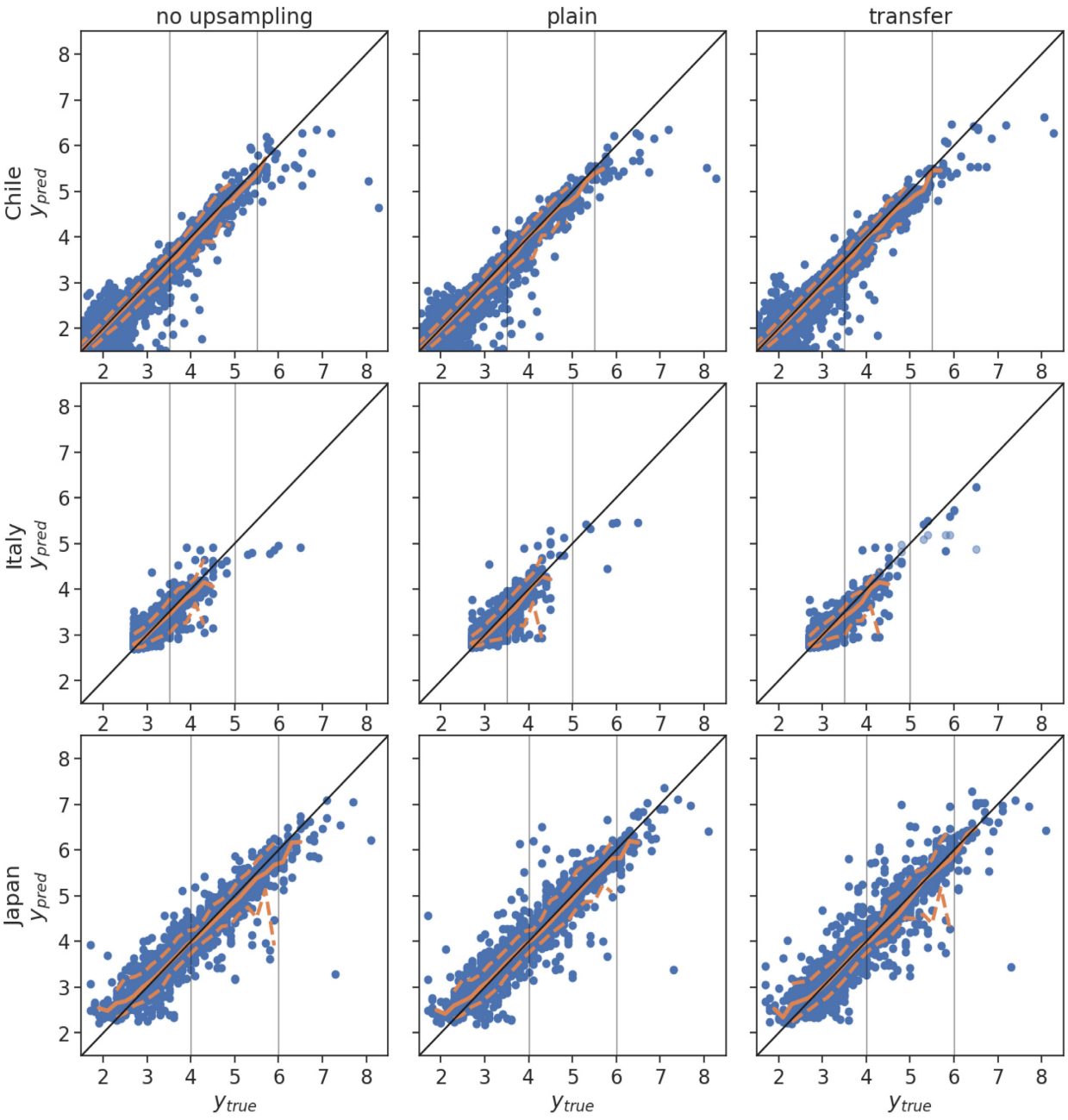

Figure 5. True and predicted magnitudes without upsampling or transfer learning (left-hand column), with upsampling but without transfer learning (middle column) and with upsampling and transfer learning (right-hand column). All plots show predictions after 8 s. In the transfer column for Chile and Japan we show results after fine-tuning on the target data set; for Italy we show results from the model without fine-tuning as this model performed better. For the largest events in Italy ($M > 4.5$) we additionally show the results after fine-tuning with pale blue dots. We suspect the degraded performance in the fine tuned model results from the fact, that the largest training event ($M_W = 6.1$) is considerably smaller than the largest test event ($M_W = 6.5$). Vertical lines indicate the borders between small, medium and large events as defined in Fig. 4. The orange lines show the running 5th, 50th and 95th percentile in 0.2 m.u. buckets. Percentile lines are only shown if sufficiently many data points are available. The very strong outlier for Japan (true ∼7.3, predicted ∼3.3) is an event far offshore (>2000 km).

For magnitude estimation no consistent performance differences between the baseline approach with position embeddings and the approach with concatenated coordinates, as originally proposed by van den Ende & Ampuero (2020), are visible. In contrast, for location estimation, the approach with embeddings consistently outperforms the approach with concatenated coordinates. The absolute performance gains between the baseline with concatenation and the baseline with embeddings is even higher than the gains from adding the transformer to the embedding model. We speculate that the positional embeddings might show better performance because they explicity encode information on how to interpolate between locations at different scales, enabling an improved exploitation of the information from stations with few or no training examples. This is more important for location estimation, where an explicit notion

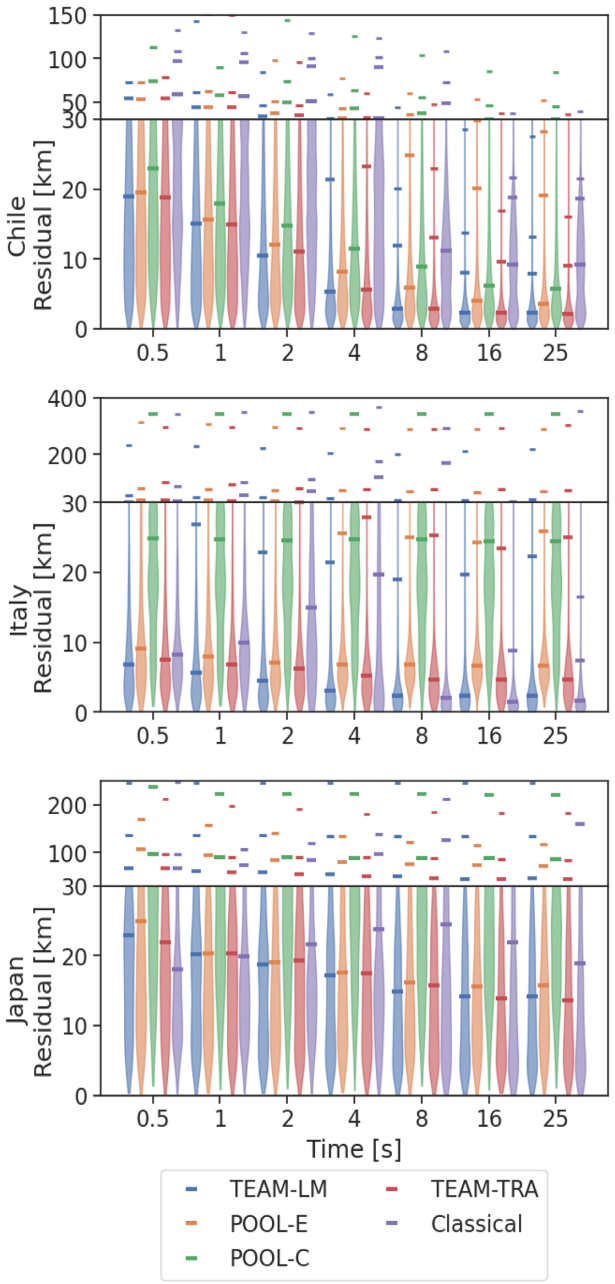

Figure 6. Violin plots and error quantiles of the distributions of the epicentral errors for TEAM-LM, the pooling baseline with position embeddings (POOL-E), the pooling baseline with concatenated position (POOL-C), TEAM-LM with transfer learning (TEAM-TRA) and a classical baseline. Vertical lines mark the 50th, 90th, 95th and 99th error percentiles, with smaller markers indicating higher quantiles. The time indicates the time since the first *P* arrival at any station. We compute errors based on the mean location predictions. A similar plot for hypocentral errors is available in the supplementary material (Fig. S4).

of relative position is required. In contrast, magnitude estimation can use further information, like frequency content, which is less position dependent.

3.3 Transfer learning

A common strategy for mitigating data sparsity is the injection of additional information from related data sets through transfer learning (Pan & Yang 2009), in our use case waveforms from other source regions. This way the model is supposed to be taught the properties of earthquakes that are consistent across regions, for example attenuation due to geometric spreading or the magnitude dependence of source spectra. Note that a similar knowledge transfer implicitly is part of the classical baseline, as it was calibrated using records from multiple regions.

Here, we conduct a transfer learning experiment inspired by the transfer learning used for TEAM. We first train a model jointly on all data sets and then fine-tune it to each of the target data sets. This way, the model has more training examples, which is of special relevance for the rare large events, but still is adapted specifically to the target data set. As the Japan and Italy data sets contain acceleration traces, while the Chile data set contains velocity traces, we first integrate the Japan and Italy waveforms to obtain velocity traces. This does not have a significant impact on the model performance, as visible in the full results tables (Tables S5–S8).

Transfer learning reduces the saturation for large magnitudes (Fig. 5, right-hand column). For Italy the saturation is even completely eliminated. For Chile, while the largest magnitudes are still underestimated, we see a clearly lower level of underestimation than without transfer learning. Results for Japan for the largest events show nearly no difference, which is expected as the Japan data set contains the majority of large events and therefore does not gain significant additional large training examples using transfer learning. The positive impact of transfer learning is also reflected in the lower RMSE for large and intermediate events for Italy and Chile (Fig. 4). These results do not only offer a way of mitigating saturation for large events, but also represent further evidence for data sparsity as the reason for the underestimation.

We tried the same transfer learning scheme for mitigating mislocations (Fig. 6). For this experiment we shifted the coordinates of stations and events such that the data sets spatially overlap. We note that this shifting is not expected to have any influence on the single data set performance, as the relative locations of events and stations within a data set stay unchanged and nowhere the model uses absolute locations. The transfer learning approach is reasonable, as mislocations might result from data sparsity, similarly to the underestimation of large magnitudes. However, none of the models shows significantly better performance than the preferred models, and in some instances performance even degrades. We conducted additional experiments where shifts were applied separately for each event, but observed even worse performance.

We hypothesize that this behaviour indicates that the TEAM-LM localization does not primarily rely on traveltime analysis, but rather uses some form of fingerprinting of earthquakes. These fingerprints could be specific scattering patterns for certain source regions and receivers. Note that similar fingerprints are exploited in the traditional template matching approaches (e.g. Shelly *et al.* 2007). While the traveltime analysis should be mostly invariant to shifts and therefore be transferable between data sets, the fingerprinting is not invariant to shifts. This would also explain why the transfer learning, where all training samples were already in the pretraining data set and therefore their fingerprints could be extracted, outperforms the shifting of single events, where fingerprints do not relate to earthquake locations. Similar fingerprinting is presumably also used by other deep learning methods for location estimation, for example by Kriegerowski *et al.* (2019) or Perol *et al.* (2018), however further experiments would be required to prove this hypothesis.

4 DISCUSSION

4.1 Multitask learning

Another common method to improve the quality of machine learning systems in face of data sparsity is multitask learning (Ruder 2017), that is having a network with multiple outputs for different objectives and training it simultaneously on all objectives. This approach has previously been used for seismic source characterization (Lomax *et al.* 2019), but without an empirical analysis on the specific effects of multitask learning.

We perform an experiment, in which we train TEAM-LM to predict magnitude and location concurrently. The feature extraction and the transformer parts are shared and only the final MLPs and the mixture density networks are specific to the task. This method is known as hard parameter sharing. The intuition is that the individual tasks share some similarity, for example in our case the correct estimation of the magnitude likely requires an assessment of the attenuation and geometric spreading of the waves and therefore some understanding of the source location. This similarity is then expected to drive the model towards learning a solution for the problem that is more general, rather than specific to the training data. The reduced number of free parameters implied by hard parameter sharing is also expected to improve the generality of the derived model, if the remaining degrees of freedom are still sufficient to extract the relevant information from the training data for each subtask.

Unfortunately, we actually experience a moderate degradation of performance for either location or magnitude in any data set (Tables S5–S11) when following a multitask learning strategy. The RMSE of the mean epicentre estimate increases by at least one third for all times and data sets, and the RMSE for magnitude stays nearly unchanged for the Chile and Japan data sets, but increases by ∼20 per cent for the Italy data set. Our results therefore exhibit a case of negative transfer.

While it is generally not known, under which circumstances multitask learning shows positive or negative influence (Ruder 2017), a negative transfer usually seems to be caused by insufficiently related tasks. In our case we suspect that while the tasks are related in a sense of the underlying physics, the training data set is large enough that similarities relevant for both tasks can be learned already from a single objective. At the same time, the particularities of the two objectives can be learned less well. Furthermore, we earlier discussed that both magnitude and location might not actually use traveltime or attenuation based approaches, but rather frequency characteristics for magnitude and a fingerprinting scheme for location. These approaches would be less transferable between the two tasks. We conclude that hard parameter sharing does not improve magnitude and location estimation. Future work is required to see if other multitask learning schemes can be applied beneficially.

4.2 Location outlier analysis

As all location error distributions are heavy tailed, we visually inspect the largest deviations between predicted and catalogue locations to understand the behavior of the localization mechanism of TEAM-LM. We base this analysis on the Chile data set (Fig. 7), as it has generally the best location estimation performance, but observations are similar for the other data sets (Figs S5 and S6).

Nearly all mislocated events are outside the seismic network and location predictions are generally biased towards the network. This matches the expected errors for traditional localization algorithms. In contrast to traditional algorithms, events are not only predicted to be closer to the network, but they are also predicted as lying in regions with a higher event density in the training set (Fig. 7, inset). This suggests that not enough similar events were included in the training data set. Similarly, Kriegerowski *et al.* (2019) observed a clustering tendency when predicting the location of swarm earthquakes with deep learning.

We investigated two subgroups of mislocated events: the Iquique sequence, consisting of the Iquique main shock, foreshocks and aftershocks, and mine blasts. The Iquique sequence is visible in the north-western part of the study area. All events are predicted approximately 0.5° too far east. The area is both outside the seismic network and has no events in the training set. This systematic mislocation may pose a serious threat in applications, such as early warning, when confronted with a major change in the seismicity pattern, as is common in the wake of major earthquakes or during sudden swarm activity, which are also periods of heightened seismic hazard.

For mine blasts, we see one mine in the northeast and one in the southwest (marked by red circles in Fig. 7). While all events are located close by, the location are both systematically mispredicted in the direction of the network and exhibit scatter. Mine-blasts show a generally lower location quality in the test set. While they make up only ∼1.8 per cent of the test set, they make up 8 per cent of the top 500 mislocated events. This is surprising as they occur not only in the test set, but also in similar quantities in the training set. We therefore suspect that the difficulties are caused by the strongly different waveforms of mine blasts compared to earthquakes. One waveform of each a mine blast and an earthquake, recorded at similar distances are shown as inset in Fig. 7. While for the earthquake both a *P* and *S* wave are visible, the S wave can not be identified for the mine blast. In addition, the mine blast exhibits a strong surface wave, which is not visible for the earthquake. The algorithm therefore can not use the same features as for earthquakes to constrain the distance to a mine blast event.

4.3 The impact of data set size and composition

Our analysis so far showed the importance of the amount of training data. To quantify the impact of data availability on magnitude and location estimation, we trained models only using fractions of the training and validation data (Fig. 8). We use the Chile data set for this analysis, as it contains by far the most events. We subsample the events by only using each kth event in chronological order, with $k = 2, 4, 8, 16, 32, 64$. This strategy approximately maintains the magnitude and location distribution of the full set. We point out, that TEAM-LM only uses information of the event under consideration and does not take the events before or afterwards into account. Therefore, the 'gaps' between events introduced by the subsampling do not negatively influence TEAM-LM.

For all times after the first *P* arrival, we see a clear increase in the magnitude-RMSE for a reduction in the number of training samples. While the impact of reducing the data set by half is relatively small, using only a quarter of the data already leads to a twofold increase in RMSE at late times. Even more relevant in an early warning context, a fourfold smaller data sets results in an approximately fourfold increase in the time needed to reach the same precision as with the full data. This relationship seems to hold approximately across all subsampled data sets: reducing the data set k fold increases the time to reach a certain precision by a factor of k.

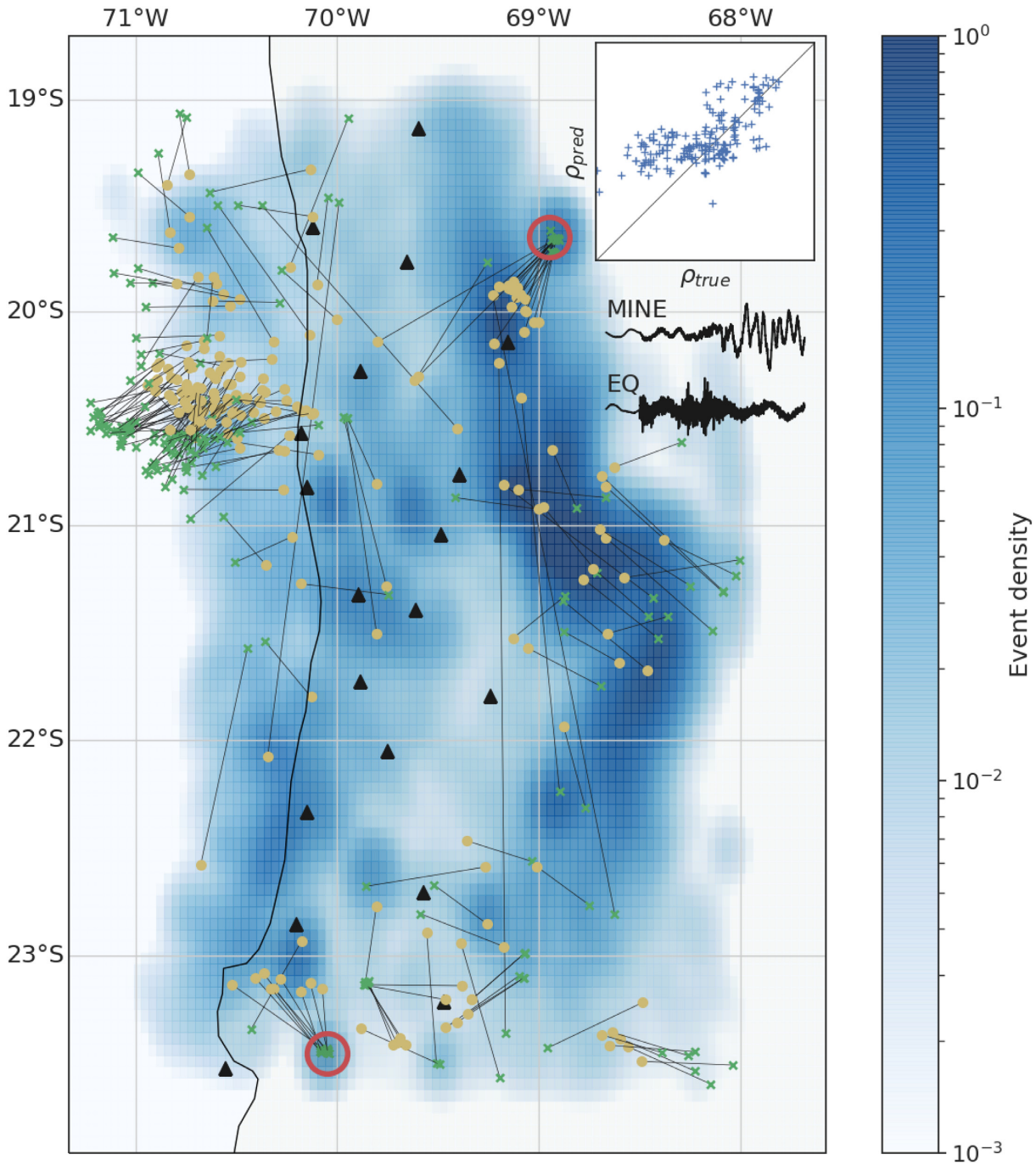

Figure 7. The 200 events with the highest location errors in the Chile data set overlayed on top of the spatial event density in the training data set. The location estimations use 16 s of data. Each event is denoted by a yellow dot for the estimated location, a green cross for the true location and a line connecting both. Stations are shown by black triangles. The event density is calculated using a Gaussian kernel density estimation and does not take into account the event depth. The inset shows the event density at the true event location in comparison to the event density at the predicted event location for the 200 events. Red circles mark locations of mine blast events. The inset waveforms show one example of a waveform from a mineblast (top) and an example waveform of an earthquake (bottom, 26 km depth) of similar magnitude ($M_A = 2.5$) at similar distance (60 km) on the transverse component. Similar plots for Italy and Japan can be found in the supplementary material (Figs S5 and S6).

We make three further observations from comparing the predictions to the true values (Fig. S7). First, for nearly all models the RMSE changes only marginally between 16 and 25 s, but the RMSE of this plateau increases significantly with a decreasing number of training events. Secondly, the lower the amount of training data, the lower is the saturation threshold above which all events are strongly underestimated. In addition, for 1/32 and 1/64 of the full data set, an 'inverse saturation' effect is noticeable for the smallest magnitudes. Thirdly, while for the full data set and the largest subsets all large events are estimated at approximately the saturation threshold, if at most one quarter of the training data is used, the largest events even fall significantly below the saturation threshold. For the mod-

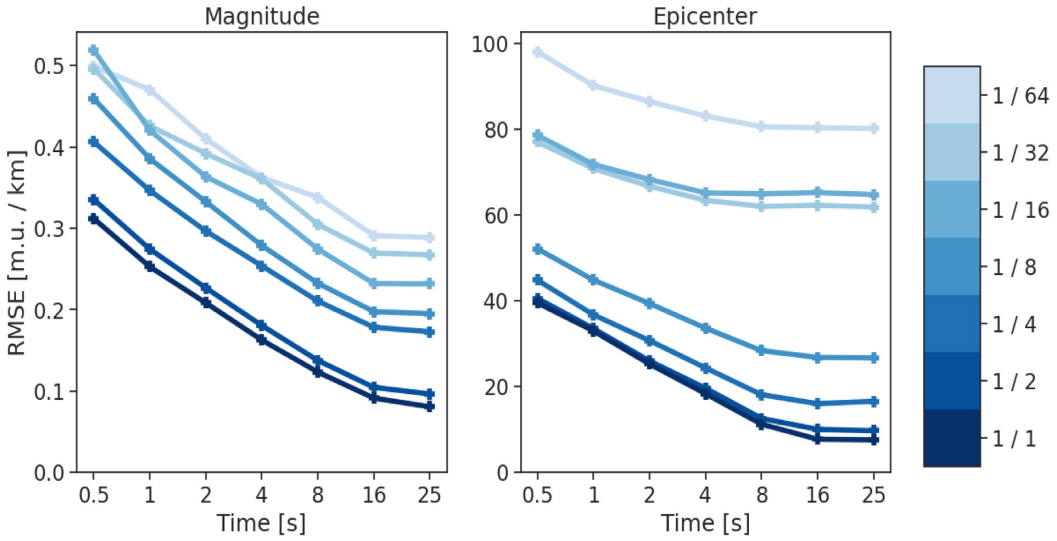

Figure 8. RMSE for magnitude and epicentral location at different times for models trained on differently sized subsets of the training set in Chile. The line colour encodes the fraction of the training and validation set used in training. All models were evaluated on the full Chilean test set. We note that the variance of the curves with fewer data is higher, due to the increased stochasticity from model training and initialization.

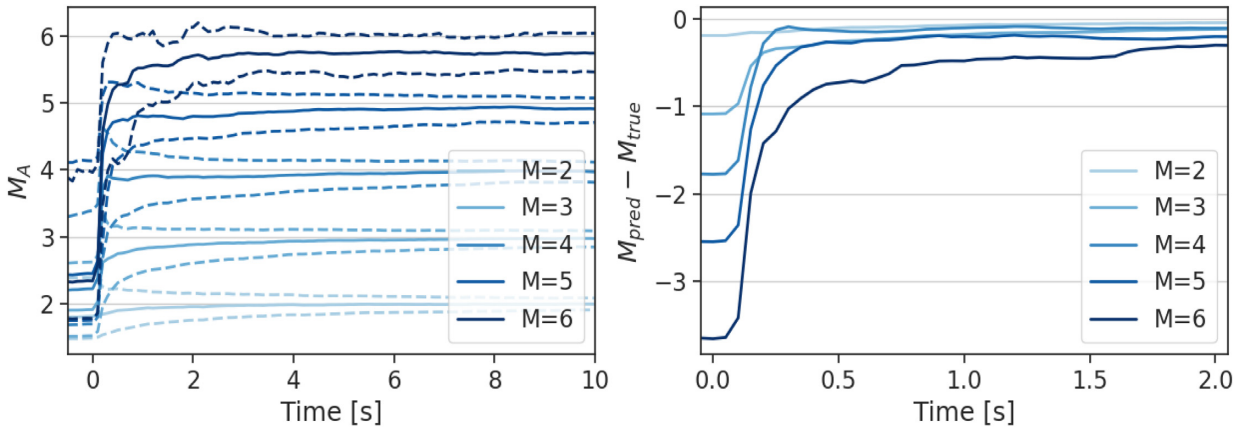

Figure 9. Magnitude predictions and uncertainties in the Chile data set as a function of time since the first *P* arrival. Solid lines indicate median predictions, while dashed lines (left-hand panel only) show 20th and 80th quantiles of the prediction. The left-hand panel shows the predictions, while the right-hand panel shows the differences between the predicted and true magnitude. The right-hand panel is focused on a shorter time frame to show the early prediction development in more detail. In both plots, each colour represents a different magnitude bucket. For each magnitude bucket, we sampled 1000 events around this magnitude and combined their predictions. If less than 1000 events were available within ±0.5 m.u. of the bucket centre, we use all events within this range. We only use events from the test set. To ensure that the actual uncertainty distribution is visualized, rather than the distribution of magnitudes around the bucket centre, each prediction is shifted by the magnitude difference between bucket centre and catalogue magnitude.

els trained on the smallest subsets (1/8 to 1/64), the higher the true magnitude the lower the predicted magnitude becomes. We assume that the larger the event is, the further away from the training distribution it is and therefore it is estimated approximately at the most dense region of the training label distribution. These observations support the hypothesis that underestimations of large magnitudes for the full data set are caused primarily by insufficient training data.

While the RMSE for epicentre estimation shows a similar behavior as the RMSE for magnitude, there are subtle differences. If the amount of training data is halved, the performance only degrades mildly and only at later times. However, the performance degradation is much more severe than for magnitude if only a quarter or less of the training data are available. This demonstrates that location estimation with high accuracy requires catalogues with a high event density.

The strong degradation further suggests insights into the inner working of TEAM-LM. Classically, localization should be a task where interpolation leads to good results, i.e., the traveltimes for an event in the middle of two others should be approximately the average between the traveltimes for the other events. Following this argument, if the network would be able to use interpolation, it should not suffer such significant degradation when faced with fewer data. This provides further evidence that the network does not actually learn some form of triangulation, but only an elaborate fingerprinting scheme, backing the finding from the qualitative analysis of location errors.

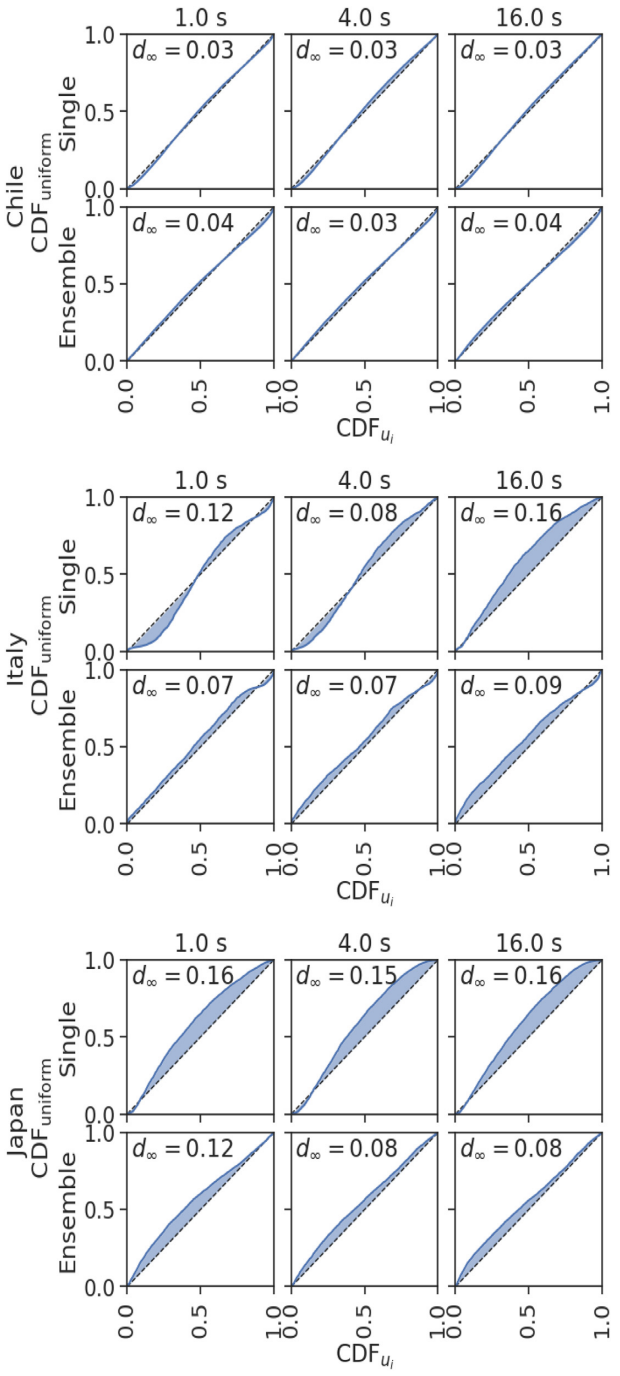

Figure 10. P-P plots of the CDFs of the empirical quantile of the magnitude predictions compared to the expected uniform distribution. The P-P plot shows $(\mathrm{CDF}_{u_i}(z), \mathrm{CDF}_{\mathrm{uniform}}(z))$ for $z \in [0, 1]$. The expected uniform distribution is shown as the diagonal line, the misfit is indicated as shaded area. The value in the upper corner provides d_∞, the maximum distance between the diagonal and the observed CDF. d_∞ can be interpreted as the test statistic for a Kolmogorov–Smirnov test. Curves consistently above the diagonal indicate a bias to underestimation, and below the diagonal to overestimation. Sigmoidal curves indicate overconfidence, mirrored sigmoids indicate underconfidence. See supplementary section SM 2 for a further discussion of the plotting methodology and its connection to the Kolmogorov–Smirnov test.

4.4 Training TEAM-LM on large events only

Often, large events are of the greatest concerns, and as discussed, generally showed poorer performance because they are not well represented in the training data. It therefore appears plausible that a model optimized for large events might perform better than a model trained on both large and small events. In order to test this hypothesis, we used an extreme version of the upscaling strategy by training a set of models only on large events, which might avoid tuning the model to seemingly irrelevant small events. In fact, these models perform significantly worse than the models trained on the full data set, even for the large events (Tables S5–S11). Therefore even if the events of interest are only the large ones, training on more complete catalogues is still beneficial, presumably by giving the network more comprehensive information on the regional propagation characteristics and possibly site effects.

4.5 Interpretation of predicted uncertainties

So far we only analysed the mean predictions of TEAM-LM. As for many application scenarios, for example early warning, quantified uncertainties are required, TEAM-LM outputs not only these mean predictions, but a probability density. Fig. 9 shows the development of magnitude uncertainties for events from different magnitude classes in the Chile data set. The left-hand panel shows the absolute predictions, while the right-hand panel shows the difference between prediction and true magnitude and focuses on the first 2 s. As we average over multiple events, each set of lines can be seen as a prototype event of a certain magnitude.

For all magnitude classes the estimation shows a sharp jump at $t = 0$, followed by a slow convergence to the final magnitude estimate. We suspect that the magnitude estimation always converges from below, as due to the Gutenberg–Richter distribution, lower magnitudes are more likely *a priori*. The uncertainties are largest directly after $t = 0$ and subsequently decrease, with the highest uncertainties for the largest events. As we do not use transfer learning in this approach, there is a consistent underestimation of the largest magnitude events, visible from the incorrect median predictions for magnitudes 5 and 6. We note that the predictions for magnitude 4 converge slightly faster than the ones for magnitude 3, while in all other cases the magnitude convergence is faster the smaller the events are. We suspect that this is caused by the accuracy of the magnitude estimation being driven by both the number of available events and by the signal to noise ratio. While magnitude 4 events have significantly less training data than magnitude 3 events, they have a better signal to noise ratio, which could explain their more accurate early predictions.

While the Gaussian mixture model is designed to output uncertainties, it cannot be assumed that the predicted uncertainties are indeed well calibrated, that is, that they actually match the real error distribution. Having well calibrated uncertainties is crucial for downstream tasks that rely on the uncertainties. Neural networks trained with a log-likelihood loss generally tend to be overconfident (Snoek *et al.* 2019; Guo *et al.* 2017), that is underestimate the uncertainties. This overconfidence is probably caused by the strong overparametrization of neural network models. To assess the quality of our uncertainty estimations for magnitude, we use the observation that for a specific event i, the predicted Gaussian mixture implies a cumulative distribution function $F^i_{pred} : \mathbb{R} \to [0, 1]$. Given the observed magnitude y^i_{true}, we can calculate $u_i = F^i_{pred}(y^i_{true})$. If y^i_{true} is

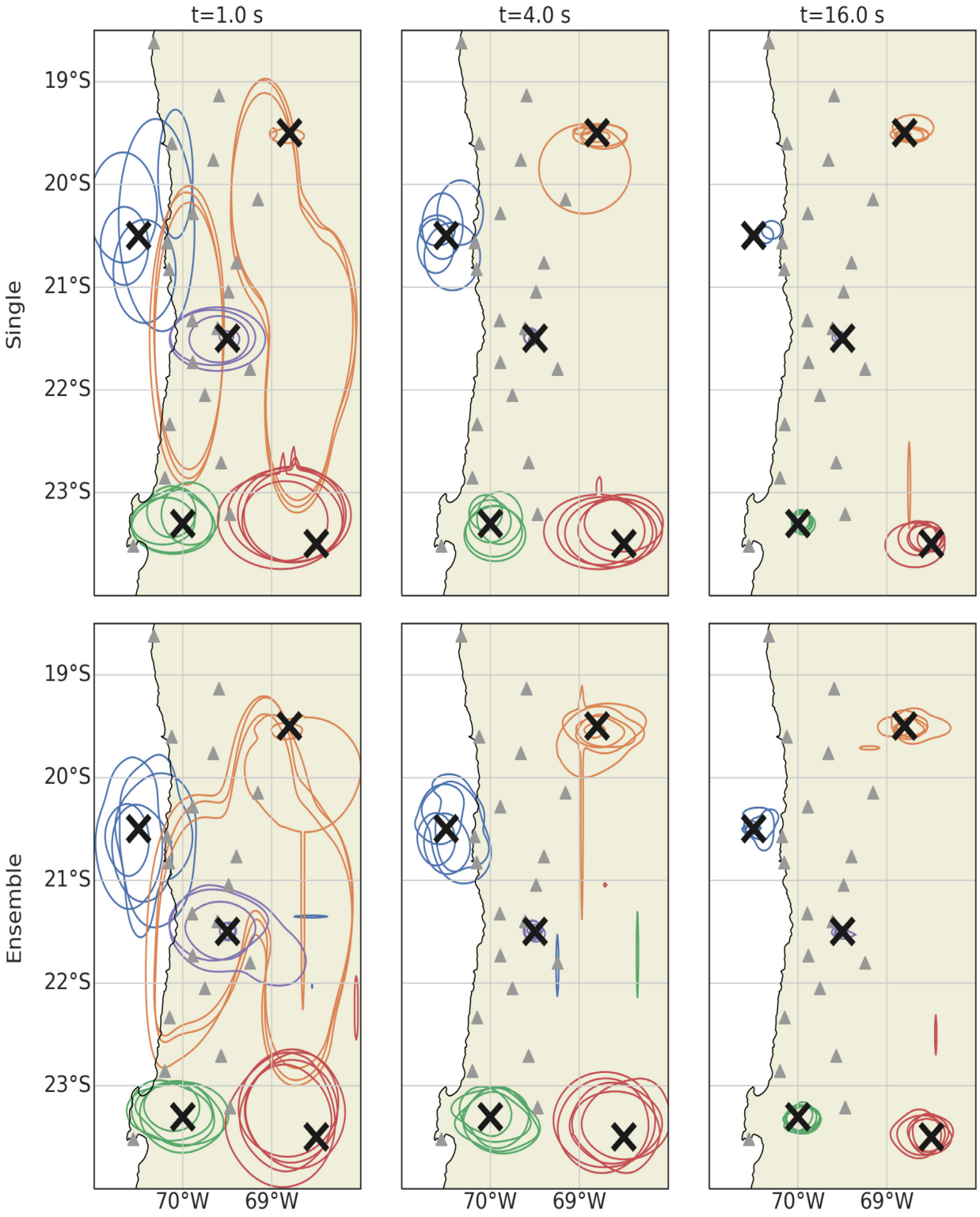

Figure 11. The figure shows 90th per cent confidence areas for sample events around 5 example locations. For each location the 5 closest events are shown. Confidence areas belonging to the same location are visualized using the same colour. Confidence areas were chosen as curves of constant likelihood, such that the probability mass above the likelihood equals 0.9. To visualize the result in 2-D we marginalize out the depth. Triangles denote station locations for orientation. The top row plots show results from a single model, while the bottom row plots show results from an ensemble of 10 models.

indeed distributed according to F^i_{pred}, then u_i needs to be uniformly distributed on [0,1]. We test this based on the u_i of all events in the test set using P-P plots (Fig. 10). Further details on the method can be found in the supplementary material (Section SM 2). Note that good calibration is a necessary but not sufficient condition for a good probabilistic forecast. An example of a perfectly calibrated but mostly useless probabilistic prediction would be the marginal probability of the labels.

Fig. 10 shows the P-P plots of u in comparison to a uniform distribution. For all data sets and all times the model is signficantly miscalibrated, as estimated using Kolmgorov–Smirnov test statistics (Section SM 2). Miscalibration is considerably stronger for Italy and Japan than for Chile. More precisely, the model is always overconfident, that is estimates narrower confidence bands than the actually observed errors. Further, in particular at later times, the model is biased towards underestimating the magnitudes. This is least visible for Chile. We speculate that this is a result of the large training data set for Chile, which ensures that for most events the density of training events in their magnitude range is high.

To mitigate the miscalibration, we trained ensembles (Hansen & Salamon 1990), a classical method to improve calibration. Instead of training a single neural network, a set of n neural networks, in our case $n = 10$, are trained, which all have the same structure, but different initialization and batching in training. The networks therefore represent a sample of size n from the posterior distribution of the model parameters given the training data. For Italy and Japan, this improves calibration considerably (Fig. 10). For Chile, the ensemble model, in contrast to the single model, exhibits underconfidence, that is estimates too broad uncertainty bands.

The maximum distance between the empirical cumulative distribution function of u and a uniformly distributed variable d_∞ is the test statistic of the Kolmogorov–Smirnov test. While d_∞ is reduced by nearly half for some of the ensemble results, the Kolmogorov–Smirnov test indicates, that even the distributions from the ensemble models deviate highly significantly from a uniform distribution ($p \ll 10^{-5}$). A table with d_∞ for all experiments can be found in the supplementary material (Table S8).

To evaluate the location uncertainties qualitatively, we plot confidence ellipses for a set of events in Chile (Fig. 11). Again we compare the predictions from a single model to the predictions of an ensemble. At early times, the uncertainty regions mirror the seismicity around the station with the first arrival, showing that the model correctly learned the prior distribution. Uncertainty ellipses at late times approximately match the expected uncertainty ellipses for classical methods, that is they are small and fairly round for events inside the seismic network, where there is good azimuthal coverage, and larger and elliptical for events outside the network. Location uncertainties are not symmetric around the mean prediction, but show higher likelihood towards the network than further outwards. Location errors for the ensemble model are more smooth than from the single model, but show the same features. The uncertainty ellipses are slightly larger, suggesting that the single model is again overconfident.

In addition to improving calibration, ensembles also lead to slight improvements regarding the accuracy of the mean predictions (Tables S5–S11). Improvements in terms of magnitude RMSE range up to ∼10 per cent, for epicentral location error up to ∼20 per cent. Due to the high computational demand of training ensembles, all other results reported in this paper are calculated without ensembling. We note that in addition to ensembles a variety of methods have been developed to improve calibration or obtain calibrated uncertainties. For a quantitative survey, see for example Snoek *et al.* (2019). One of these methods, Monte Carlo Dropout, has already been used in the context of fast assessment by van den Ende & Ampuero (2020).

5 CONCLUSION

In this study we adapted TEAM to build TEAM-LM, a real time earthquake source characterization model, and used it to study the pitfalls and particularities of deep learning for this task. We showed that TEAM-LM achieves state of the art in magnitude estimation, outperforming both a classical baseline and a deep learning baseline. Given sufficiently large catalogues, magnitude can be assessed with a standard deviation of ∼0.2 magnitude units within 2 s of the first P arrival and a standard deviation of 0.07 m.u. within the first 25 s. For location estimation, TEAM-LM outperforms a state of the art deep learning baseline and compares favorably with a classical baseline.

Our analysis showed that the quality of model predictions depends crucially on the training data. While performance in regions with abundant data is excellent, in regions of data sparsity, prediction quality degrades significantly. For magnitude estimation this effect results in the underestimation of large magnitude events; for location estimation events in regions with few or no training events tend to be mislocated most severely. This results in a heavy tailed error distribution for location estimation. Large deviations in both magnitude and location estimation can have significant impact in application scenarios, for example for early warning where large magnitudes are of the biggest interest.

Following our analysis, we propose a set of best practices for building models for fast earthquake source characterization:

(i)Build a comprehensive evaluation platform. Put a special focus on outliers and rare or large events. Analyse which impact outliers or out of distribution events will have for the proposed application.
(ii)Use very large training catalogues, spanning long time spans and having a low magnitude of completeness. If possible, use transfer learning. We hope the catalogues used in this study can give a starting point for transfer learning.
(iii)Use training data augmentation, especially upsampling of large events, which improves prediction performance in face of label sparsity at virtually no cost.
(iv)If probabilistic estimates are required, use deep ensembles to improve the model calibration.
(v)When using deep learning for location estimation, put special emphasis on monitoring possible distribution shifts between training data and application.

While these points give guidance for training current models they also point to further directions for methodological advances. First, our transfer learning scheme is fairly simple. More refined and targeted schemes could increase the amount of information sharable across data sets considerably. We further expect major improvements from training with simulated data, but are aware that generating realistic, synthetic seismogramms, especially for large events, poses major challenges. Another promising alternative might be to move away from the paradigm of black box modeling, that is training algorithms that are built solely by fitting recorded data. Instead, incorporation of physical knowledge and a move towards physics informed deep learning methods seems promising (Raissi *et al.* 2019). However, physics informed neural networks are still in their infancy and the application to seismic tasks still needs to be developed.

ACKNOWLEDGEMENTS

We thank the National Research Institute for Earth Science and Disaster Resilience 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.

We thank Christian Sippl for providing the P picks for the Chile catalogue. We thank Sebastian Nowozin for insightful discussions on neural network calibration and probabilistic regression. We thank Martijn van den Ende for his comments that helped improve the manuscript. Jannes Münchmeyer acknowledges the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS). We use obspy (Beyreuther *et al.* 2010), tensorflow (Abadi *et al.* 2016) and colour scales from Crameri (2018).

DATA AVAILABILITY

The Italy data set has been published as Münchmeyer *et al.* (2020). The Chile data set has been published as Münchmeyer *et al.* (2021a). An implementation of TEAM-LM and TEAM has been published as Münchmeyer *et al.* (2021b). Download instructions for the Japan data set are available in the code publication.

REFERENCES

Abadi, M., *et al.*, 2016. Tensorflow: a system for large-scale machine learning, in *Proceedings of the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16),* pp. 265–283.

Aldersons, F., 2004. Toward three-dimensional crustal structure of the dead sea region from local earthquake tomography, *PhD thesis*, Senate of Tel-Aviv University, Tel Aviv, Israel.

Asch, G., Tilmann, F., Schurr, B. & Ryberg, T., 2011. Seismic network 5E: MINAS Project (2011/2013), GFZ Data Services, doi:10.14470/ab466166.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y. & Wassermann, J., 2010. Obspy: a python toolbox for seismology, *Seismol. Res. Lett.,* **81**(3), 530–533.

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

Cesca, S., Sobiesiak, M., Tassara, A., Olcay, M., Günther, E., Mikulla, S. & Dahm, T., 2009. The Iquique local network and PicArray, (Scientific Technical Report STR - Data; 18/02), 19 p. doi:10.14470/vd070092.

Colombelli, S., Festa, G. & Zollo, A., 2020. Early rupture signals predict the final earthquake size, *Geophys. J. Int.,* **223**(1), 692–706.

Crameri, F., 2018. Geodynamic diagnostics, scientific visualisation and staglab 3.0, *Geosci. Model Dev.,* **11**(6), 2541–2562.

Deichmann, N., 2018. Why does ML scale 1:1 with 0.5logES? *Seismol. Res. Lett.,* **89**(6), 2249–2255.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2019. Bert: pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* **Volume 1 (Long and Short Papers)**

Dipartimento di Fisica, Universitá degli studi di Napoli Federico II, 2005, Istituto Nazionale di Geofisica e Vulcanologia (INGV). Irpinia seismic network (ISNET).

Doi, K., 2014. Seismic network and routine data processing-japan meteorological agency, *Summary Bull. Int. Seismol. Centre,* **47**(7–12), 25–42.

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

Festa, G., *et al.*, 2018. Performance of earthquake early warning systems during the 2016–2017 Mw 5–6.5 Central Italy sequence, *Seismol. Res. Lett.,* **89**(1), 1–12.

GEOFON Data Center, 1993, Deutsches GeoForschungsZentrum GFZ. Geofon seismic network, doi:10.14470/tr560404.

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

Graeber, F.M. & Asch, G., 1999. Three-dimensional models of P wave velocity and P-to-S velocity ratio in the southern central Andes by simultaneous inversion of local earthquake data, *J. geophys. Res.,* **104**(B9), 20 237–20 256.

Guo, C., Pleiss, G., Sun, Y. & Weinberger, K.Q., 2017. On calibration of modern neural networks, *International Conference on Machine Learning,* PMLR, .

Hansen, L.K. & Salamon, P., 1990. Neural network ensembles, *IEEE Trans. Pattern Anal. Mach. Intell.,* **12**(10), 993–1001.

ISIDe Working Group, 2007, Istituto Nazionale di Geofisica e Vulcanologia (INGV). Italian seismological instrumental and parametric database (ISIDe), doi:10.13127/ISIDE.

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. Centro di microzonazione sismica network, 2016 central italy seismic sequence (centromz), doi:10.13127/SD/ku7Xm12Yy9.

Istituto Nazionale di Geofisica e Vulcanologia (INGV), Italy, 2006, Istituto Nazionale di Geofisica e Vulcanologia (INGV). Rete sismica nazionale (rsn), doi: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.

Kingma, D.P. & Ba, J., 2014. Adam: A method for stochastic optimization, *International Conference for Learning Representations*.

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.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S. & Teh, Y.W., 2019. Set transformer: a framework for attention-based permutation-invariant neural networks, in *International Conference on Machine Learning,* pp. 3744–3753, PMLR.

Leyton, F., Ruiz, S., Baez, J.C., Meneses, G. & Madariaga, R., 2018. How fast can we reliably estimate the magnitude of subduction earthquakes? *Geophys. Res. Lett.,* **45**(18), 9633–9641.

Lomax, A., Virieux, J., Volant, P. & Berge-Thierry, C., 2000. Probabilistic earthquake location in 3D and layered models, in *Advances in Seismic Event Location,* pp. 101–134, Springer.

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.

Matrullo, E., De Matteis, R., Satriano, C., Amoroso, O. & Zollo, A., 2013. An improved 1-D seismic velocity model for seismological studies in the Campania–Lucania region (southern Italy), *Geophys. J. Int.,* **195**(1), 460–473.

MedNet Project Partner Institutions, 1990, Istituto Nazionale di Geofisica e Vulcanologia (INGV). Mediterranean very broadband seismographic network (MEDNET), doi:10.13127/SD/fBBBtDtd6q.

Meier, M.-A., Ampuero, J.P. & Heaton, T.H., 2017. The hidden simplicity of subduction megathrust earthquakes, *Science,* **357**(6357), 1277–1281.

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,* **58,** 11, 8211–8224.

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

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

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

Münchmeyer, J., Bindi, D., Leser, U. & Tilmann, F., 2021, **225,** 1, 646–656. The transformer earthquake alerting model: a new versatile approach to earthquake early warning, *Geophys. J. Int.,* doi:10.1093/gji/ggaa609.

Münchmeyer, J., Bindi, D., Sippl, C., Leser, U. & Tilmann, F., 2020b. Low uncertainty multifeature magnitude estimation with 3-D corrections and boosting tree regression: application to North Chile, *Geophys. J. Int.,* **220**(1), 142–159.

Münchmeyer, J., Bindi, D., Leser, U. & Tilmann, F., 2021a. Fast earthquake assessment dataset for chile, doi:10.5880/GFZ.2.4.2021.002, GFZ Data Services

Münchmeyer, J., Bindi, D., Leser, U. & Tilmann, F., 2021b. Team – the transformer earthquake alerting model, doi:10.5880/GFZ.2.4.2021.003, GFZ Data Services.

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

OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale), 2016. North-east Italy seismic network (NEI), doi:10.7914/SN/OX, International Federation of Digital Seismograph Networks.

OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale) and University of Trieste, 2002. North-east Italy broadband network (NI), doi:10.7914/SN/NI, International Federation of Digital Seismograph Networks.

Pan, S.J. & Yang, Q., 2009. A survey on transfer learning, *IEEE Trans. Knowledge Data Eng.,* **22**(10), 1345–1359.

Perol, T., Gharbi, M. & Denolle, M., 2018. Convolutional neural network for earthquake detection and location, *Sci. Adv.,* **4**(2), e1700578, doi:10.1126/sciadv.1700578.

Presidency of Counsil of Ministers - Civil Protection Department, 1972. Italian strong motion network (ran), doi:10.7914/SN/IT, Presidency of Counsil of Ministers - Civil Protection Department.

Raissi, M., Perdikaris, P. & Karniadakis, G.E., 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *J. Comput. Phys.,* **378,** 686–707.

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:10.15778/RESIF.FR.

RESIF - Réseau Sismologique et géodésique Français, 1995b. Réseau accélérométrique permanent (french accelerometrique network) (rap), doi:10.15778/RESIF.RA.

Ruder, S., 2017. An overview of multi-task learning in deep neural networks, preprint arXiv:1706.05098.

Shelly, D.R., Beroza, G.C. & Ide, S., 2007. Non-volcanic tremor and low-frequency earthquake swarms, *Nature,* **446**(7133), 305–307.

Sippl, C., Schurr, B., Asch, G. & Kummerow, J., 2018. Seismicity structure of the Northern Chile forearc from >100,000 double-difference relocated hypocenters, *J. geophys. Res.,* **123**(5), 4063–4087.

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,* pp. 13 969–13 980.

Trugman, D.T., Page, M.T., Minson, S.E. & Cochran, E.S., 2019. Peak Ground Displacement Saturates Exactly When Expected: Implications for Earthquake Early Warning, *J. geophys. Res.,* **124**(5), 4642–4653.

Ueno, H., Hatakeyama, S., Aketagawa, J., Funasaki, J. & Hamada, N., 2002. Improvement of hypocenter determination procedures in the japan meteorological agency, *Quart. J. Seism.,* **65,** 123–134.

Universidad de Chile, 2013. Red sismologica nacional, doi:10.7914/SN/C1, International Federation of Digital Seismograph Networks.

Universita della Basilicata, 2005. Unibas, doi:10.17598/NIED.0004, Italian National Institute of Geophysics and Volcanology (INGV).

University of Genova, 1967. Regional Seismic Network of North Western Italy [Data set]. International Federation of Digital Seismograph Networks, doi:10.7914/SN/GU, International Federation of Digital Seismograph Networks.

van den Ende, M.P.A. & Ampuero, J.-P., 2020. Automated seismic source characterization using deep graph neural networks, *Geophys. Res. Lett.,* **47**(17), e2020GL088690, doi:10.1029/2020GL088690.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I., 2017. Attention is all you need, in *Proceedings of the 31st Conference on Advances in Neural Information Processing Systems,* Long Beach, CA, USA, pp. 5998–6008.

Wigger, P., Salazar, P., Kummerow, J., Bloch, W., Asch, G. & Shapiro, S., 2016. West–fissure- and atacama-fault seismic network (2005/2012), doi:10.14470/3s7550699980, Deutsches GeoForschungsZentrum GFZ.

SUPPORTING INFORMATION

Supplementary data are available at *GJI* online.

Table S1. Architecture of the feature extraction network. The input shape of the waveform data is (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. We want to mention that depending on the existence of borehole data the number of input filters for the first Conv1D varies.

Table S2. Architecture of the transformer network.

Table S3. Architecture of the mixture density network.

Table S4. Experiment names for the results tables

Table S5. Test set RMSE magnitude estimate across all magnitudes. For some experiments we additionally provide standard deviation. The standard deviations were obtained from six runs with different random model initialization. In this case the provided mean value is the mean over six runs. Note that the provided standard deviation denotes the empirical standard deviation of a single run, therefore the uncertainty of the mean expected to be smaller by a factor of $\sqrt{6}$. Due to computational constraints we are only able to provide standard deviations for a selected set of experiments.

Table S6. Test set mean absolute error (MAE) magnitude estimate across all magnitudes

Table S7. Test set R2 score across all magnitudes

Table S8. Test set test statistic d_α for the Kolmogorov–Smirnov test across all magnitudes.

Table S9. Test set RMSE of magnitude estimate for large events

Table S10. Test set MAE of magnitude estimate for large events

Table S11. Test set R2 score of magnitude estimate for large events

Table S12. Test set root squared mean for hypocentral error. We note that only 4 out of 10 models for the Italy location ensemble converged. We used only the converged models for the ensemble evaluation.

Table S13. Test set mean absolute hypocentral error. We note that only 4 out of 10 models for the Italy location ensemble converged. We used only the converged models for the ensemble evaluation.

Table S14. Test set root squared mean for epicentral error. We note that only 4 out of 10 models for the Italy location ensemble converged. We used only the converged models for the ensemble evaluation.

Table S15. Test set mean absolute epicentral error. We note that only 4 out of 10 models for the Italy location ensemble converged. We used only the converged models for the ensemble evaluation.

Figure S1. Event and station distribution for Chile. In the map, events are indicated by dots, stations by triangles. The event depth is encoded using colour.

Figure S2. Event and station distribution for Italy. In the map, events are indicated by dots, stations by triangles. The event depth is encoded using colour.

Figure S3. Event and station distribution for Japan. In the map, events are indicated by dots, stations by triangles. The event depth is

encoded using colour. There are ∼20 additional events far offshore in the catalogue, which are outside the displayed map region.

Figure S4. Distribution of the hypocentral errors for TEAM-LM, the pooling baseline with position embeddings (POOL-E), the pooling baseline with concatenated position (POOL-C), TEAM-LM with transfer learning (TEAM-TRA) and a classical baseline. Vertical lines mark the 50th, 90th, 95th and 99th error percentiles. The time indicates the time since the first *P* arrival at any station. We use the mean predictions.

Figure S5. The 100 events with the highest location error in the Italy data set overlayed on top of the spatial event density in the training data set. The estimations use 16 s of data. Each event is denoted by a dot for the estimated location, a cross for the true location and a line connecting both. Stations are not shown as station coverage is dense. The event density is calculated using a Gaussian kernel density estimation and does not take into account the event depth. The inset shows the event density at the true event location in comparison to the event density at the predicted event location.

Figure S6. The 200 events with the highest location error in the Japan data set overlayed on top of the spatial event density in the training data set. The estimations use 16 s of data. Each event is denoted by a dot for the estimated location, a cross for the true location and a line connecting both. Stations are not shown as station coverage is dense. The event density is calculated using a Gaussian kernel density estimation and does not take into account the event depth. The inset shows the event density at the true event location in comparison to the event density at the predicted event location.

Figure S7. True and predicted magnitudes after 8 s using only parts of the data sets for training. All plots show the Chile data set. The fraction in the corner indicates the amount of training and validation data used for model training. All models were evaluated on the full test data set.

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

Table A1. Seismic networks.

Region	Network	Reference
Chile	GE	GEOFON Data Center (1993)
	C, C1	Universidad de Chile (2013)
	8F	Wigger *et al.* (2016)
	IQ	Cesca *et al.* (2009)
	5E	Asch *et al.* (2011)
Italy	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), Italy (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) and University of Trieste (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)
	XO	EMERSITO Working Group (2018)
Japan	KiK-Net	National Research Institute For Earth Science And Disaster Resilience (2019)