Münchmeyer, J., Bindi, D., Leser, U., Sippl, C., Tilmann, F. (2020): Low uncertainty multifeature magnitude estimation with 3-D corrections and boosting tree regression: application to North Chile. - *Geophysical Journal International*, *220*, 1, pp. 142—159.

DOI: http://doi.org/10.1093/gji/ggz416

Geophysical Journal International

Geophys. J. Int. (2019) **220,** 142–159
Advance Access publication 2019 September 17
GJI Seismology

doi: 10.1093/gji/ggz416

Low uncertainty multifeature magnitude estimation with 3-D corrections and boosting tree regression: application to North Chile

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

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

Accepted 2019 September 16. Received 2019 August 27; in original form 2019 May 24

SUMMARY

Magnitude estimation is a central task in seismology needed for a wide spectrum of applications ranging from seismicity analysis to rapid assessment of earthquakes. However, magnitude estimates at individual stations show significant variability, mostly due to propagation effects, radiation pattern and ambient noise. To obtain reliable and precise magnitude estimates, measurements from multiple stations are therefore usually averaged. This strategy requires good data availability, which is not always given, for example for near real time applications or for small events. We developed a method to achieve precise magnitude estimations even in the presence of only few stations. We achieve this by reducing the variability between single station estimates through a combination of optimization and machine learning techniques on a large catalogue. We evaluate our method on the large scale IPOC catalogue with $>100\,000$ events, covering seismicity in the northern Chile subduction zone between 2007 and 2014. Our aim is to create a method that provides low uncertainty magnitude estimates based on physically meaningful features. Therefore we combine physics based correction functions with boosting tree regression. In a first step, we extract 110 features from each waveform, including displacement, velocity, acceleration and cumulative energy features. We correct those features for source, station and path effects by imposing a linear relation between magnitude and the logarithm of the features. For the correction terms, we define a non-parametric correction function dependent on epicentral distance and event depth and a station specific, adaptive 3-D source and path correction function. In a final step, we use boosting tree regression to further reduce interstation variance by combining multiple features. Compared to a standard, non-parametric, 1-D correction function, our method reduces the standard deviation of single station estimates by up to 57 per cent, of which 17 per cent can be attributed to the improved correction functions, while boosting tree regression gives a further reduction of 40 per cent. We analyse the resulting magnitude estimates regarding their residuals and relation to each other. The definition of a physics-based correction function enables us to inspect the path corrections and compare them to structural features. By analysing feature importance, we show that envelope and *P* wave derived features are key parameters for reducing uncertainties. Nonetheless the variety of features is essential for the effectiveness of the boosting tree regression. To further elucidate the information extractable from a single station trace, we train another boosting tree on the uncorrected features. This regression yields magnitude estimates with uncertainties similar to the single features after correction, but without using the earthquake location as required for applying the correction terms. Finally, we use our results to provide high precision magnitudes and their uncertainties for the IPOC catalogue.

Key words: South America; Earthquake ground motions; Site effects.

[†] Now at: Institute of Geophysics, Czech Academy of Sciences, Prague, Czech Republic

1 INTRODUCTION

Magnitudes are key metrics for assessing the impact of earthquakes, used for example for seismicity analysis or disaster response. These tasks require a high precision of the magnitude scale as well as quantified uncertainties. Current methods encounter uncertainty in magnitude estimates from single stations, which are caused by site, source and path effects and background noise. In this paper, we aim to find and calibrate interpretable magnitude scales with low uncertainties. We focus on reducing the variability of single station estimates. The magnitude scales and calibration functions improve magnitudes both in a fast assessment context when only data for a single station might be available, and, assuming some degree of independence between stations, for more accurate network wide magnitude estimations.

Many different magnitude scales exist for a multitude of use cases. An extensive overview can be found in Bormann (2012). Here, we focus on magnitudes derived from simple waveform features. The first magnitude of this kind was the local magnitude, M_L, as defined by Richter (1935), which was based on the peak horizontal displacement recorded with a particular instrument, the Wood–Anderson seismometer. M_L has the advantage of a simple definition, allowing for fast and robust determination. Uncertainties typically are reduced by averaging measurements from multiple stations. On the downside, M_L and similar magnitude scales require distance-dependent correction functions, which need to be calibrated for each region in order to take the local earth structure into account. M_L was first developed for crustal events in California (Richter 1935). In subduction zones, where both crustal and interface seismicity is present, a correction function based on hypocentral distance only is inaccurate due to the different travel paths and therefore anelastic and geometric focusing effects experienced for crustal and deep events.

For M_L, the amplitude reducing effects of (physical) attenuation and geometric spreading are modeled empirically using a table of attenuation values over distance. In order to reduce cumbersome expressions, we will refer to the combined effect of (physical) attenuation and geometric spreading simply by attenuation as both will generally reduce the recorded amplitudes with the distance traveled by the waves. This approach has been generalized in the context of non-parametric models by Brillinger & Preisler (1984). Savage & Anderson (1995) propose a simple 1-D model with linear interpolation that can be fit to data using quadratic optimization. A similar model, covering the same area as in this study, was applied to strong motion data from 106 events in the Pisagua sequence 2014 by Bindi *et al.* (2014). More complex attenuation functions have been proposed in the context of ground motion prediction. Dawood & Rodriguez-Marek (2013) propose a 2-D attenuation function, by modelling the attenuation per grid cell in a regular grid. They optimize the model parameters using 7242 measurements from 117 aftershocks of the M_w 9.0 Tohoku earthquake in 2011.

Reducing magnitude uncertainties is a major challenge for rapid assessment of earthquakes. Zollo *et al.* (2006) propose a magnitude estimation based on early *P*- and *S*-wave peak displacement. Lancieri & Zollo (2008) extend the work to a Bayesian approach giving uncertainty estimates. Multiple studies (e.g. Festa *et al.* 2008; Picozzi *et al.* 2018; Spallarossa *et al.* 2019) propose to incorporate the early radiated energy, obtained as the squared velocity integrated over time. As the quantitative results in these studies are obtained in the context of early warning, they can not directly be compared to the results of this study. Nonetheless the methods share the idea of choosing appropriate features to minimize single station uncertainties.

The classical method for local magnitudes follows a two step procedure. It consists of a feature extraction and the application of a hypocentral distance correction. To reduce uncertainties we extend this procedure. In the first step, we define 110 physically motivated features that can be easily derived from the single station waveform. In the second step, we model the attenuation using a 2-D grid function, together with a station specific, adaptive 3-D source correction function to account for the complex subduction zone geometry. One set of calibration functions is derived for each feature. Finally, we add a third step where we combine the single station features using boosting tree regression to obtain more precise estimates. All steps are enabled by the large number of events present in the IPOC catalogue for Northern Chile by Sippl *et al.* (2018).

Using a multistepped approach based on hand-crafted, physics-inspired features instead of an end-to-end machine learning approach offers multiple benefits. First, the resulting scales are more interpretable. This includes analysis of the correction functions, comparison of the scales to each other and interpretation of the key features for the boosting tree regression. Secondly, true magnitudes are not known or even necessarily well defined, as most magnitude scales, except e.g. M_W and M_E, are defined through measured features rather than through independent physical source properties. Therefore our approach uses bootstrapping by first creating high quality single feature network-average magnitudes using extended correction functions, and then applying boosting tree regression with these magnitudes as labels and the corrected measurements as features. Following our analysis we extend the IPOC catalogue with well-calibrated magnitude values and their uncertainties.

2 METHODS

2.1 Earthquake catalogue and stations

Our analysis is based on the earthquake catalogue of Sippl *et al.* (2018). The catalogue covers the region of the northern Chile forearc and contains 101 601 events. The events were extracted from 8 yr of continuous seismic data between 2007 and 2014 using automatic event detection and phase picking routines. The magnitudes range from <2 up to 7.7 and the estimated magnitude of completeness for M_L is ∼2.8 (Sippl *et al.* 2018). All event hypocentres were double-difference relocated, based on picked arrival times. The catalogue is based on data from the IPOC network (CX, GFZ German Research Centre For Geosciences & Institut Des Sciences De L'Univers-Centre National De La Recherche CNRS-INSU 2006). Additional seismic data were obtained from the GEOFON (GE, GEOFON Data Centre 1993), CSN (C, C1, Universidad de Chile 2013), WestFissure (8F, Wigger *et al.* 2016), Iquique (IQ, Cesca *et al.* 2009) and Minas (5E, Asch *et al.* 2011) networks. A full map showing the detected events and the stations used can be found in Fig. 1.

Sippl *et al.* (2018) use this catalogue to analyse the double seismic zone of the northern Chile forearc. The catalogue events are classified into upper plate, plate interface, upper plane and lower plane, based on their location. In addition the authors identify an intermediate depth cluster, which is used as a separate class. The catalogue features some events belonging to none of the classes mentioned, mostly events at the border of the study area. We removed these events from our analysis as they are expected to have higher location uncertainties, resulting in a total number of 96 185 events

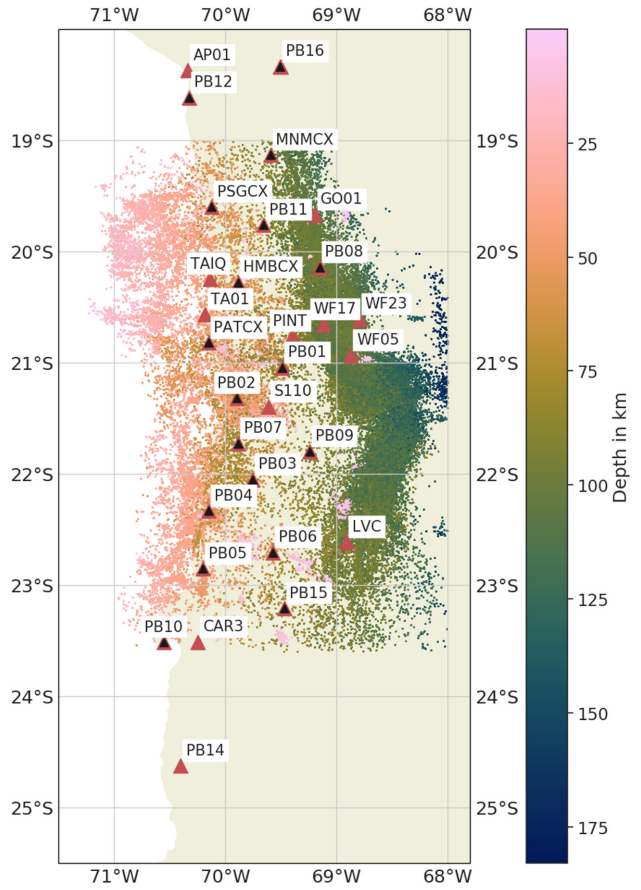

Figure 1. Event distribution (from Sippl *et al.* 2018) and broadband station location. Stations with strong motion data are denoted by an additional black triangle. The sharp boundaries on the north, east and south side of the study area are due to the event selection criteria in the original catalogue.

included in this study. For further information on the classification, catalogue and study region we refer to Sippl *et al.* (2018).

We use the catalogue to evaluate our method, as it is both consistent and challenging, while offering a large amount of data. Consistency is achieved by the low temporal variability in the station coverage, a consistent tool chain and double difference relocated hypocentres. It is a prerequisite for the consistent and low uncertainty calibration of magnitude scales. The challenge arises from the wide range of magnitudes and the different types of seismicity present in the subduction zone.

For our analysis we mostly use the same seismic stations as Sippl *et al.* (2018), but also incorporate data from strong motion stations. A list of all 31 stations can be found in the Table F1.

In total we analyse ∼1 100 000 *P* picks and ∼650 000 *S* picks from the catalogue. A further 450 000 *S* picks were predicted, using the 1-D velocity model of Graeber & Asch (1999).

Fig. 2 shows the distribution of measurements across distance and depth. Nearly all measurements were taken at distances below 400 km and depths shallower than 150 km, while few additional measurements exist up to 500 km distance and 200 km depth. We observe multiple peaks in the depth distribution, with two smaller peaks around 5 and 30 km and one large peak around 110 km. These are caused by the different types of seismicity present, namely crustal events and events in the intermediate depth cluster. For further details on the catalogue and seismicity in Northern Chile we refer to the original publication of the catalogue by Sippl *et al.* (2018).

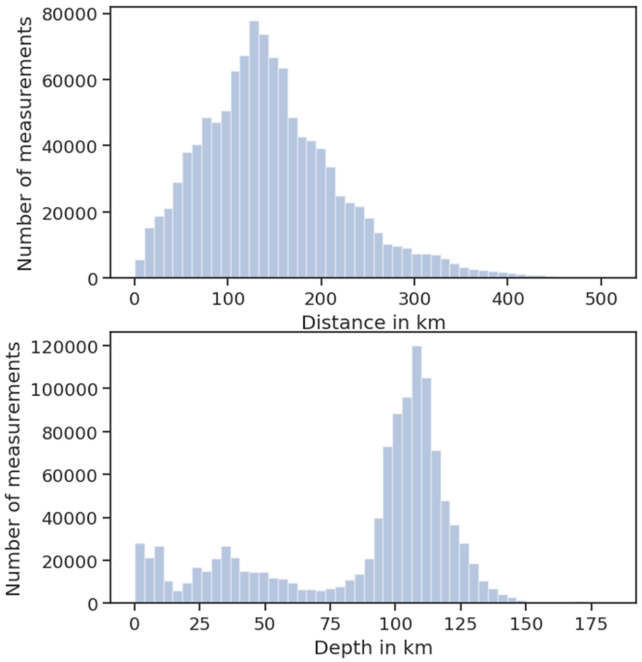

Figure 2. Distribution of the number of measurements binned by epicentral distance and event depth.

2.2 Feature extraction

The feature extraction process encompasses some common preprocessing steps and the actual feature generation. A schematic overview of the workflow for each waveform is shown in Fig. F1. For each event we generate the features for all stations, for which the catalogue contained at least one phase pick.

We generally use broadband records. We assume a clipped trace if its peak value exceeds 80 per cent of the maximum output of the digitizer, as estimated from its bit count. In this case we use the strong motion record instead. A similar procedure is used by Cauzzi *et al.* (2016). If no strong motion data is available, the record is discarded. We also discard traces with gaps.

We remove the instrument response using the inventories provided by GEOFON. We apply a cosine taper in the frequency domain with corner frequency parameters 0.005, 0.01, 30, 35 Hz before the deconvolution. Whenever strong motion data is used, the data is integrated to obtain velocity traces.

As the recorded signal is often below the noise level in the broadband records, we high-pass filter the data to increase the signal-to-noise ratio (SNR), while retaining as much of the low frequency information as is possible. We select the frequency interval f_{low}, f_{high} with the lowest frequencies from a pre-defined set of candidates (Table A1), such that the mean spectral amplitude ratio between the 30 s before the *P* pick and the 30 s after is at least 4. The data is then high-pass filtered with the corner frequency f_{low}. The frequency f_{high} is only used for frequency selection, but is discarded for the filtering of the actual data. This strategy is applied as we observed sufficient SNRs for high frequencies for all events. Therefore we only need to identify the lowest band, where the SNR is still sufficient for the following steps. More details on the applied filtering and the distribution of selected frequencies can be found in Appendix A.

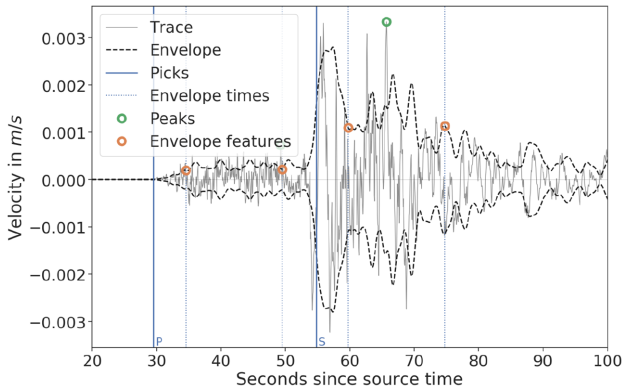

Figure 3. Example trace with extracted features denoted by circles. The trace shows the vertical component from station PB01 for an $M_w = 6.3$ event at a depth of 21 km and an epicentral distance of 193 km.

As a final step of the pre-processing, we detrend the filtered data using the best linear fit in a 300 s window around the event.

The resulting velocity trace is differentiated to obtain the acceleration trace and integrated to obtain the displacement trace for the vertical (Z), radial (R) and transverse component (T). We use the absolute value of all horizontal components (NE) as well as the absolute value of all components (ZNE) as additional traces where we compute the absolute value from the vectorial sum of the single components.

From each trace we export six values, as shown in Fig. 3. We extract the peak values of the *P* and *S* wave. For the *P* wave peak we search the waveform between the *P* pick and the *S* pick minus a safety margin in order to avoid interference from the *S* waves. The safety margin is taken to be 5 per cent of the measured or estimated *S* wave traveltime, but always at least 0.5 s. For the *S* wave peak we restrict the search window to the first 30 s after the *S* pick to minimize the possibility of overlapping events.

In addition, we extract values from the *P* and *S* wave envelopes. We calculate the signal envelope and low-pass filter it at 0.5 Hz. We then export the values at 5 and 20 s after the *P* and *S* picks. The envelope values for the *P* wave are not reported in case the time difference between the *P* and *S* pick is less than the respective lag times. We include the envelope values as we expect them to be less influenced by the radiation pattern as well as distance uncertainties. We chose delays of 5 and 20 s because the 5 s envelope value should be representative of the energy in the direct arrival for moderately sized events but less variable than the peak, while the 20 s value represents a compromise between accessing, for most event-station pairs, the late coda where the wavefield is fully diffusive but still retaining signal levels well above the noise level for practically all events. For further details on the choice of envelope times we refer to Appendix B.

In addition to the features from the displacement, velocity and acceleration traces, we export the energy and the peak value of a simulated Wood–Anderson instrument. We calculate the energy as the integral of the squared velocity trace. We export both the integral over the time between *P* and *S* pick and the integral over the first 30 s after the *S* pick. For the Wood-Anderson instrument we report the peak values from the *P* and *S* waves as before. All resulting feature values are logarithmized with base 10.

We rescale the resulting energy features by a factor of 2/3. Following the analysis by Deichmann (2018b) the factor of 2/3 is the theoretically derived scaling factor between M_L and $\log E$. The different scaling of energy compared to the displacement scale is further discussed in Section 3.

In total we extract 110 features, 22 from each component or combination thereof. Of the 22 features half are from the P wave and half from the complete waveform. The features are energy and the simulated Wood–Anderson peak as well as the peak, 5 s envelope and 20 s envelope values from displacement, velocity and acceleration (see Table 1).

In our data set features might be incomplete due to missing waveform data for single components or because the *P* envelope values are later than the *S* arrival. All features are present in at least 98.8 per cent of the measurements. The only exceptions are the 5 s *P*-wave envelope value with only 97.7 per cent availability and the corresponding envelope value at 20 s with only 21.4 per cent. This lower availability is expected, as the value can only be measured at a significant distance to the event.

2.3 Correction terms and normalization

To correct the measurements for the source–receiver distance, station bias and source conditions, we use a set of non-parametric correction functions. The classical approach of Richter (1935) uses a table of hypocentral distance correction values. We extend this method by using a non-parametric 2-D correction function incorporating source–receiver distance and source depth, as well as by adding a station correction and a station-specific source correction term. The latter will be mostly affected by propagation effects related to 3-D heterogeneity, but could theoretically also incorporate radiation pattern effects, if certain mechanisms are dominant in some area.

Let E be the set of events and S be the set of stations. Let $E_s \subseteq E$ be the subset of E measured at station $s \in S$. For station $s \in S$ and event $e \in E_s$ we model the difference between the measured feature Y_s^e and the corresponding event magnitude M^e through an attenuation function Γ, a station specific source correction term L_s and a station correction B_s. With an error term ε_s^e we obtain:

$$Y_s^e - M^e = \Gamma(r_s^e, d^e) + L_s(p^e) + B_s + \varepsilon_s^e, \tag{1}$$

where r_s^e is the epicentral distance between event and station, d^e the event depth and p^e the hypocentre. We formulate a quadratic minimization problem on the squared error objective function:

$$Obj_\varepsilon = \frac{1}{n} \sum_{s \in S} \sum_{e \in E_s} (\varepsilon_s^e)^2 \,. \tag{2}$$

Here n denotes the number of error terms or equivalently the number of measurements for the feature. We now describe the definitions of the different correction terms, as well as their normalization and regularization. This will also lead to an extension of the objective function for the quadratic optimization.

The attenuation function Γ is defined as a 2-D non-parametric function on a grid of epicentral distances and depth values. We use a grid G with 50 linearly spaced distance values between 20 and 500 km and 20 linearly spaced depth values between 10 and 200 km. Values between the grid points are interpolated bilinearly between the four adjacent values.

We enforce smoothing of the attenuation function by introducing a penalty term derived from the 2nd order finite difference approximation of the Laplacian with a regularization term

$$R_\Gamma = \frac{1}{|G|} \sum_{(r,d) \in G} \lambda_r \left(\frac{\partial^2 \Gamma}{\partial r^2} \right)^2 + \lambda_d \left(\frac{\partial^2 \Gamma}{\partial d^2} \right)^2 \,. \tag{3}$$

For clarity reasons we write the continuous version of the Laplacian here, rather than its finite difference approximation. The factors λ_r and λ_d are model hyperparameters describing the level of smoothing. We use $|G|$ to denote the cardinality of the set G, that is the number of grid points.

To account for source location specific systematic errors, we introduce a source specific correction function L_s for each station s. We randomly sample a set of events $\bar{E}_s \subset E_s$ and assign to each of the events $e \in \bar{E}_s$ a correction term l_s^e. The correction for a single event is defined through the correction terms of the k nearest neighbours:

$$L_s(e) = \frac{1}{k} \sum_{e' \in \mathrm{kNN}(e, \bar{E}_s)} l_s^{e'}. \tag{4}$$

Here $\mathrm{kNN}(e, \bar{E}_s)$ is the set of the k nearest neighbours of e in $\bar{E}_s$. For our experiments we chose $k = 10$ and $\bar{E}_s$ such that $|\bar{E}_s|/|E_s| \approx 0.1$. As distance metric for the determination of the nearest neighbours we use the euclidean distance between the hypocentres, but scale the depth difference with a factor of 3, to account for the high importance of the depth. We use the average over the set of neighbours to obtain a smoothly varying function of position. As the density of events is not uniform over the region, the nearest neighbour based function can represent higher variability in regions with many events, while being smoother in regions, where a high resolution function would not be well constrained. The subsampling $\bar{E}_s$ from E_s is necessary for performance reasons, as each element in $\bar{E}_s$ introduces an additional free parameter. We choose one subset $\bar{E}_s \subseteq E_s$ for each feature and station.

The location correction is normed and regularized by:

$$R_L = \lambda_L \frac{1}{|S|} \sum_{s \in S} \frac{1}{|\bar{E}_s|} \sum_{e \in \bar{E}_s} l_s^{e2} \tag{5}$$

$$\forall s \in S : \textstyle\sum_{e \in E_s} L_s(e) = 0 \tag{6}$$

The factor λ_L is a hyperparameter to adjust the level of regularization.

For each station s we add a station bias B_s to account for site effects. We constrain the biases of all stations to sum up to zero:

$$\sum_{s \in S} B_s = 0. \tag{7}$$

The magnitude scale needs to be calibrated as the system would otherwise be underdetermined. Specifically attenuation with depth can not be extracted from the data, as the depth is only event but not station specific as the distance. The Richter definition resolves attenuation with depth by using hypocentral distance. Due to the separation of depth and distance in our approach, the standard Richter definition of assigning magnitude 3.0 to a 1 mm displacement at a distance of 100 km is not applicable. Therefore we calibrate our scale against M_w, which also includes information on the attenuation in depth direction.

We obtain a total of 155 M_w values from the Global CMT Project (Dziewonski *et al.* 1981; Ekström *et al.* 2012). As we do not expect a linear scaling between M_w and our magnitude scales for the full range of magnitudes covered by Global CMT, we only used the 114 events with magnitudes between 5.0 and 6.0 in the calibration. For incorporating the information into our model, let E_{M_w} denote the events for which a moment tensor solution is available. We then define an objective by:

$$Obj_{M_w} = \lambda_{M_w} \frac{1}{|E_{M_w}|} \sum_{e \in E_{M_w}} (M^e - M_w^e)^2. \tag{8}$$

The factor λ_{M_w} controls the trade-off between fitting to M_w and smoothness of the correction functions. For our analysis we use $\lambda_{M_w} = 0.1$.

We use a weak connection between M^e and M_w^e instead of setting $M^e = M_\mathrm{w}^e$ for multiple reasons. First, we do not expect the features to correspond 1:1 with M_w as they might depend on other parameters than the seismic moment, for example the stress drop, which influences the high frequency content in particular. We investigate this scaling in more detail in Section 3.1. Secondly, we only have values for M_w for a small subset of the data set available. In conclusion, enforcing equality to M_w might introduce perturbations into the correction functions. The weak connection resolves the underdetermination of our system, while minimizing the perturbation on the correction functions.

All correction functions and bias terms are optimized concurrently using quadratic optimization on the full objective:

$$\min(Obj_\varepsilon + Obj_{M_w} + R_\Gamma + R_L). \tag{9}$$

It consists of the primary objective, the calibration against M_w and the regularization terms for Γ and L. It is additionally constrained by the relations (6) and (7). The free parameters are the event magnitudes M^e, the values of grid G of the correction function Γ, the correction terms $\{l_s^e\}_{s \in S, e \in \bar{E}_s}$ and the station biases $\{B_s\}_{s \in S}$.

While the source-path correction term could in principle incorporate the whole attenuation function, we still decided to split the attenuation into the distance-depth, the source-path and the station term for multiple reasons. First, the source-path term is station specific, while the distance-depth term is universal for all stations. This enables a by far better calibration of the attenuation with distance and depth, especially for stations and ranges with only few measurements. Secondly, we can formulate a sensible regularization more easily: the distance-depth correction is forced to be smooth, whereas the source-path correction is damped towards zero to ensure deviations from the generic distance-depth correction are only introduced where clearly required by the data. Thus, the correction functions are easier to interpret, as the station specific and the mean attenuation effects are separated. For details on the interpretation see Section 4.3.

2.4 Multifeature estimation

The methods proposed so far only use each feature separately, but do not leverage combinations of features. As a framework for combining multiple features for a joint magnitude estimate we state a regression problem: Given all features of a *single station* estimate a chosen target *network* magnitude. We use the term *network* magnitude to refer to the average across all *single station* magnitude estimates. We want to emphasize that the key point is estimating *network wide* information from *single station* features. The target magnitude scale can be chosen arbitrarily among the scales derived using the calibration functions, for example the event magnitude from peak displacement on the horizontal components. Due to the limited amount of data, we do not use M_w as a target magnitude.

The regression problem has a canonical baseline, which is the station magnitude estimate from the feature corresponding to the target magnitude. If for example the target magnitude is the peak

displacement magnitude average over all stations, the baseline prediction from a single station would be its peak displacement magnitude estimate. The error level of this baseline is exactly the error from the modeling obtained in the previous step. The task of the regression problem is to estimate network wide information from the combined features of a single station.

We use boosting trees (Friedman 2002) for regression, training one common model for all stations. Boosting trees are a special class of gradient boosting models and use decision trees as the underlying classifiers. They are a rather popular regression technique for non-linear problems. We use a non-linear approach to model complex dependencies between the features. We show by quantitative comparison to linear regression that those complex dependencies are indeed present.

Boosting trees are better suited for our problem than other non-linear approaches like support vector machines or neural networks. Support vector machines suffer from long training times for our problem size and are therefore intractable. Neural networks are harder to train in the presence of missing values, as they represent smooth functions. We tried multiple imputation techniques to alleviate this problem, but were not able to achieve the performance level of boosting trees using neural networks. Boosting trees can handle this problem by learning a default action for splits at missing data points [for details see Chen & Guestrin (2016), algorithm 3].

An additional upside of boosting trees is their interpretability regarding feature importance. We can analyse the information gain through splits at specific features to get a view of their internal workings. This is in strong contrast to neural networks, where such an interpretation is not simple.

As boosting trees rely on decision trees, their value range is discrete. While this poses a theoretical limitation, the number of values inside the range is high enough that the discretization is barely observable. The residuals in the regression predictions are still by far higher than those added by the discretization. This effect is only causing higher approximation errors for events with high magnitudes, as their number in the training set is limited.

2.5 Evaluation

We split our data into a training, a development and a test set with the ratios 60:10:30. All measurements for one event are guaranteed to be in the same split. The sets contain ∼670 000, ∼110 000 and ∼330 000 measurements and ∼58 000, ∼9 600 and ∼29 000 events. We split randomly between the events, but keep the splits fixed for all evaluation steps and across all features. All models are trained only on the training set. This includes the correction functions as well as the boosting tree for feature combination. We use the development set for hyperparameter selection and report the scores on the test set. An overview of hyperparameter values can be found in Tables B1 and B2. We discuss the choice of hyperparameters and give advice for the adaptation to other data sets in Appendix B.

To evaluate the uncertainty of our models we are using the root mean square error (RMSE) between station magnitude and event magnitude. Using the definitions from Section 2.3 we define the RMSE as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{s\in S}\sum_{e\in E}\varepsilon_s^{e2}}. \tag{10}$$

To compare the uncertainty of different scales with each other we need to normalize the scales. This is necessary to ensure a fair comparison, as the different scales show different slopes. We normalize by dividing the RMSE by the difference between the 25th and 75th percentile of the predicted magnitudes. We rescale all magnitudes by multiplying by the 25th and 75th interquantile distance of the Wood–Anderson magnitude on the horizontal components. Thereby we obtain RMSE values that approximately resemble local magnitude units. We chose these quantile values as all scales show a relatively linear dependency with each other between those values. We only use the scaling to compare the scales with each other. For our experiments on the combination of multiple features we use the plain RMSE, as we only compare the uncertainty between scales with the same value range.

For the multifeature regression, we always report the RMSE between the predictions from the single station and the network magnitude from the target feature. We do not optimize for the mean of all multifeature predictions, as this would trivially converge against a constant.

Our feature extraction is based on Obspy (Beyreuther *et al.* 2010). The extraction is parallelized event wise and conducted on a compute cluster. As no dependencies between events exist, parallelization can easily be scaled to clusters of arbitrary size. To optimize our models we used the Gurobi optimizer (Gurobi Optimization LLC 2018) using a free academic license. Optimization took ∼3 hr per model using 64 threads on four Intel Xeon E7-4870 CPUs and required ∼120 GB of main memory. All boosting tree experiments were conducted using XGBoost (Chen & Guestrin 2016) on four Intel Xeon E7-4870 CPUs. Each training process took less than 30 min.

3 RESULTS

We report the average RMSE for all extracted features and components in Table 1. We can see that differences in RMSE depend more on the feature and less on the component. Nonetheless we see differences between the components as well. For the peaks of the *P* wave the average normalized RMSE over all feature classes (i.e. displacement, velocity and acceleration) is 0.216 on the *Z* component and 0.239 on the *T* component. In between are the ZNE, R and NE components (in this order). This matches the characteristics of the *P* wave as a longitudinal wave, which is expected to have smaller amplitudes on the transverse than on the vertical and radial components. The effect is also observable for the envelope values, although the combinations of components tend to perform similarly or even better in this case.

For the peak amplitude measurements of the full waveform all components achieve nearly identical RMSE values. For the envelope values, the differences are more pronounced, especially for the 20 s envelope values. The best scoring component at 20 s is ZNE with 0.162 normalized RMSE and the worst Z with 0.183 normalized RMSE. We suspect that taking the absolute value of all components effectively reduces noise and thereby improves envelope performance.

We see major differences regarding the normalized RMSE between the different feature classes. For the peaks of the full trace the lowest average RMSE across all components occurs for velocity (0.163), followed by acceleration (0.172), Wood–Anderson (0.197) and displacement (0.198). Energy (0.134) achieves a better score than all peak values.

Envelope values behave differently for different features. For displacement, the envelope derived scales show considerably higher RMSE on most components. In contrast the best 20 s velocity

Table 1. Normalized RMSE (in local magnitude units) for all analysed features and components on the test set. The second column indicates whether the features are extracted from the full wave (*P*+*S*) or the *P* wave (*P*) only. The third column indicates if peak or envelope values are used. The best combination of peak or envelope, wave and component for each feature class is highlighted in bold. The columns denotes the components, where NE is the absolute value of all horizontal components and ZNE the absolute value of all components. The rightmost column indicates the average normalization factor applied. The norm factor does not vary significantly between different components of the same feature and is therefore only given as average across the components. We note that measurement for the 20 s envelope value on the *P* wave are only possible for the ∼ 30 per cent of the event-station pairs with sufficiently large distances to achieve at least 20 s separation between *P* and *S* arrivals. They are thus skewed towards larger magnitudes.

			Z	R	T	NE	ZNE	∅ Norm factor
Displacement	Full	Peak	0.191	0.195	0.195	0.194	**0.190**	1.01
		Env 5*s*	0.241	0.243	0.243	0.232	0.225	1.01
		Env 20*s*	0.279	0.266	0.264	0.248	0.241	1.28
	P	Peak	0.270	0.285	0.302	0.297	0.290	1.39
		Env 5*s*	0.322	0.328	0.347	0.330	0.320	1.47
		Env 20*s*	0.263	0.259	0.295	0.254	0.252	1.48
Velocity	Full	Peak	0.147	0.164	0.162	0.163	0.155	0.94
		Env 5*s*	0.172	0.199	0.201	0.195	0.184	0.91
		Env 20*s*	0.132	0.138	0.138	0.129	**0.120**	1.02
	P	Peak	0.183	0.191	0.194	0.194	0.191	1.03
		Env 5*s*	0.141	0.162	0.168	0.158	0.144	1.06
		Env 20*s*	0.143	0.149	0.155	0.144	0.138	1.13
Acceleration	Full	Peak	0.160	0.171	0.170	0.172	0.165	0.98
		Env 5*s*	0.169	0.193	0.196	0.190	0.179	0.94
		Env 20*s*	0.128	0.132	0.133	0.125	**0.117**	1.03
	P	Peak	0.181	0.187	0.187	0.189	0.187	1.01
		Env 5*s*	0.137	0.146	0.150	0.142	0.132	1.01
		Env 20*s*	0.119	0.124	0.125	0.120	0.116	1.02
Wood-Anderson	Full	Peak	0.195	0.195	0.197	0.193	**0.188**	1.02
	P	Peak	0.292	0.308	0.332	0.320	0.301	1.58
Energy	Full		0.124	0.134	0.134	0.132	**0.122**	0.75
	P		0.144	0.160	0.165	0.160	0.147	0.81

and acceleration envelope values have a 23 per cent and 29 per cent lower RMSE than the respective best peak scales. Scales derived from features on the *P* wave show a higher normalized RMSE in all cases. The increase is up to 69 per cent for the Wood–Anderson instrument compared to the full wave.

The lowest normalized RMSE value over all are the 20 s envelope values of acceleration and velocity on the ZNE component (0.120 and 0.117). The best peak derived feature is velocity on the Z component with 0.147. The best combination of peak or envelope, wave and component for each feature is highlighted in Table 1.

3.1 Relations between the scales

We now compare the scales obtained from the peak values of the ZNE components for the different feature classes and energy (Fig. 4). As reference scale we use peak displacement, as this scale shows no saturation effects. We denote the scale by M_A as proposed by Deichmann (2018a). The scatter visible in the plot reflects both systematic effects of earthquake physics and the uncertainties of both M_A and the other scale under considerations.

As all scales are bound to M_w between 5.0 and 6.0, they match between those values. They deviate outside this range. The Wood–Anderson based magnitude scales 1:1 with the displacement magnitude for magnitudes below 6.0 and slowly saturates above. Unsurprisingly, it shows the lowest variance in comparison with the displacement scale. The velocity magnitude scales 1:1 with displacement for small magnitudes and increases more slowly $M_A >$ 5.0. The acceleration shows a similar behavior as the velocity, but nearly completely saturates above $M_A > 6.0$. The saturation effects for velocity and acceleration have previously been observed by Katsumata (2001) and are due to the shifted frequency spectra. Interestingly, the variance of the acceleration magnitude is highest among the scales, suggesting a high variability of peak acceleration compared to peak displacement.

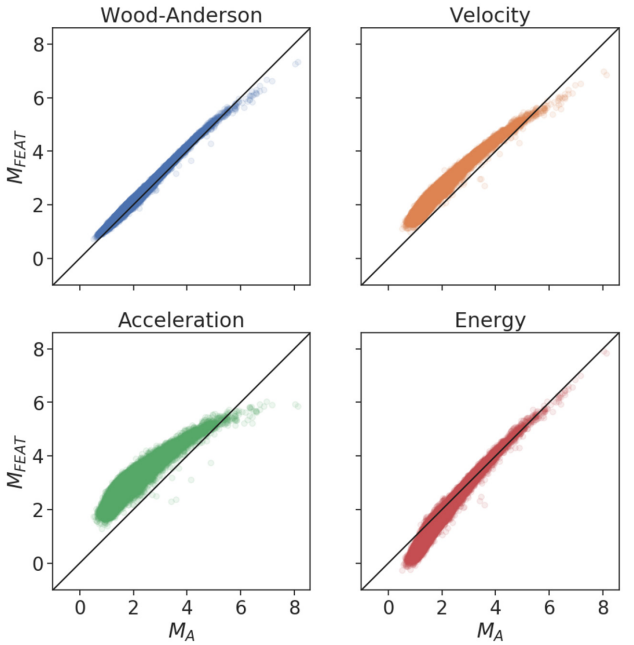

Figure 4. Comparison of the different magnitude scales, all relative to the scale based on displacement (M_A). All scales are based on the ZNE component on the full waveform, using the peak values except the energy scale. The identity line has been added in black for comparison.

The energy magnitude grows slightly stronger than M_A below 4.0 and scales nearly 1:1 above, exhibiting scatter similar to the

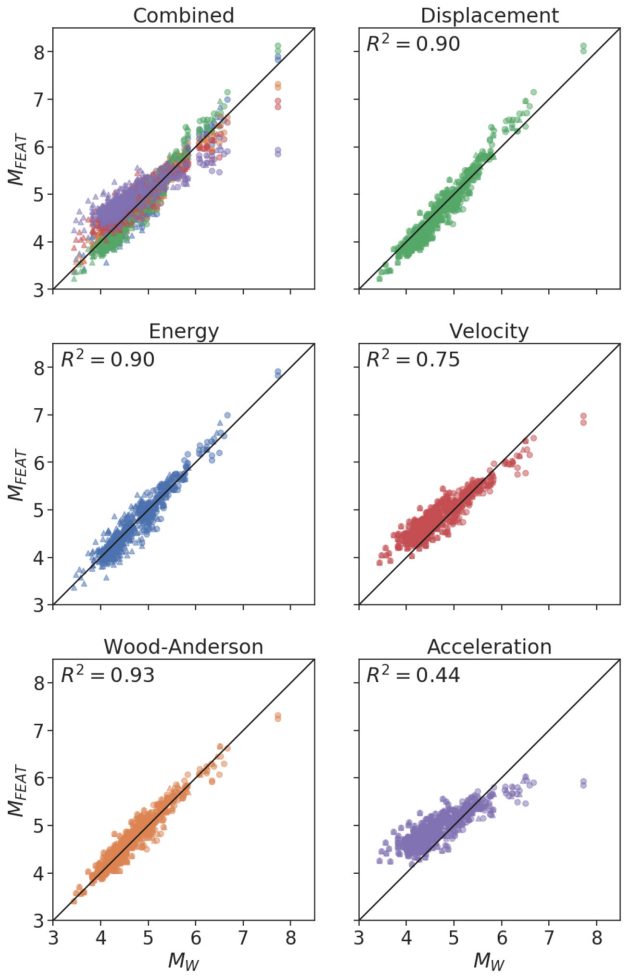

Figure 5. Estimated magnitudes from different features in comparison to M_w from Global CMT and additional solutions. Comparisons to Global CMT are shown as circles, comparisons to our moment tensor solutions as triangles. All magnitudes were determined from the peak values of the ZNE component of the full waveform (except energy). The identity line has been added added in black for comparison. We report R^2 scores as a further orientation.

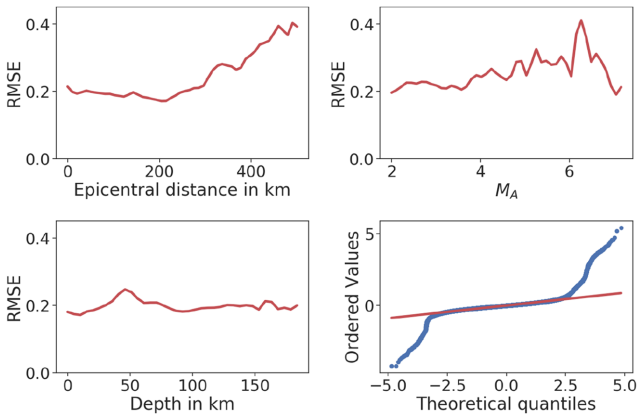

Figure 6. Residual analysis for displacement magnitudes on the NE components. Three plots show the dependency of RMSE on depth, M_A and epicentral distance. M_A refers to the peak displacement magnitude on the horizontal components. All traces represent running square means. The averaging window widths are 20 km (epicentral distance), 10 km (depth) and 0.2 m.u. (M_A). The bottom-right plot shows the distribution of the residuals in comparison to a normal distribution as a Q–Q plot.

velocity magnitude. Below 2.0, the energy magnitude compared to M_L approximately follows a 4:3 scaling. Combined with the factor of 2:3 in the definition of our energy features, this scaling provides empirical evidence for the 2:1 scaling between M_E and M_L derived by Deichmann (2018b). For large magnitudes this scaling only holds true compared to M_L, caused by the Wood–Anderson response, but not for M_A.

We compare the magnitude scales to M_w using 155 moment tensor solutions from the Global CMT project and 507 further solutions we determined using regional moment tensor inversion (see Appendix C). Fig. 5 shows the relation between M_w and the scales generated from different features. Due to the calibration used, all scales match M_w fairly well between 5.0 and 6.0. Strong differences can be seen outside this range, especially for larger events. Saturation effects cause velocity and acceleration magnitudes to both underestimate large events. The saturation effect also causes them to overestimate smaller events, as the saturation already affects the calibration magnitude range *M*5–*M*6. The Wood–Anderson magnitude shows a saturation effect only above $M \sim 6.5$.

The trends of M_A and energy magnitude match M_w approximately over the whole range of magnitudes. The energy magnitude exhibits more scatter, possibly related to varying source properties, for example stress drop, although ambient noise could also affect the measurements, of course. Deichmann (2018a) proposed M_A as a non saturating alternative to M_L. Our empirical results support this proposal.

3.2 Residual distribution

We analyse the the residuals as a function of depth, M_A, hypo- and epicentral distance (Fig. 6). While residual variations do not depend strongly on depth, we observe a near doubling of residuals with distance and presumably a weak increase with M_A. The increase with distance can be easily understood, as the SNR decreases with distance.

Varying residuals could also be caused by implicit frequency dependencies of the correction terms. The increased RMSE for larger magnitude values could be caused by the lower dominant frequency of large events, compared to the small events composing the majority of the training events. Lower frequency waves might encounter less physical attenuation and scattering, therefore experience reduced amplitude decay with larger distances. As site response might be frequency dependent, we expect a weak distance dependence in the station term. On the other hand, this effect is offset by the source-path correction, as it is station specific. While frequency effects can not be accounted for by the linear single-feature model, we expect the boosting tree regression to mitigate them, as it has access to spectral information through the combined use of displacement, velocity and acceleration features.

We observed different RMSE values for different types of seismicity. We used the classification from Sippl *et al.* (2018) to classify events into upper plate, plate interface, upper plane, lower plane and intermediate-depth cluster events. The lowest RMSE for peak displacement on the combined horizontal components occurs for crustal and intermediate depth events, a 0.02 higher RMSE for lower plane and plate interface events and another 0.01 for upper plane events. Results are similar for other features.

Table 2. Test set RMSE for models based on the combinations of multiple features. Separate columns represent different feature sets. A plus sign indicates that further information was added, a minus sign indicates that all features of the respective type were removed. Unadjusted refers to the plain features, on which no correction terms have been applied. Please note that the RMSE values are not normalized, therefore comparisons of absolute values are only valid inside each column, but not between the columns.

Features	Displacement NE	Acceleration *Z*
Single (Baseline)	0.196	0.159
All + Timing	0.103	0.097
All	0.105	0.099
All - Env	0.113	0.108
All - *P* wave	0.110	0.103
All - Env - *P* Wave	0.121	0.112
Only Z component	0.111	0.101
Only velocity	0.122	0.115
Unadjusted + Timing	0.162	0.162
Unadjusted	0.177	0.176
Unadjusted *P* wave	0.203	0.199

The Q–Q plot in Fig. 6 shows that the residual distribution deviates from a normal distribution, as it exhibits heavy tails. Those likely indicate measurement errors, caused by overlapping events, instrument issues or wrong frequency selection, causing low SNR. We observe a general positive skewness of outlier residuals, that is, magnitudes are more likely to be grossly overestimated than underestimated. This holds true for all stations except AP01, LVC, PINT and S100 that exhibit a negative skewness. In the appendix, we give a further analysis of the residuals for each station (Fig. F2), possible time dependency (Fig. F3) and the effect of different SNR thresholds (Appendix D).

3.3 Multifeature magnitude estimation

For the experiments with multifeature magnitude estimation, we use the peak horizontal displacement and the peak vertical acceleration as target scales. We choose horizontal displacement because of its similarity to the standard M_L for smaller magnitudes and no observed saturation effects and we choose vertical acceleration as a challenging benchmark, as it already has a low RMSE.

A single boosting tree predictor is trained on the multifeature sets for all stations simultaneously (Table 2). We achieve the best results for both target scales using the full feature set with additional features measuring temporal information, that is the difference between *P* and *S* pick time and the time at which each feature was extracted relative to the *P* pick. For horizontal displacement we are able to reduce the RMSE by 47 per cent; for vertical acceleration the reduction is 39 per cent. The smaller improvement for acceleration is likely caused by the already lower RMSE of this feature. To elucidate the effect of different features on prediction quality, we removed certain features from the full feature set. The RMSE still improves significantly with respect to the single feature for all tested combinations, although of course the prediction accuracy decreases somewhat (see top part of Table 2). To evaluate the benefit of combining features from velocity, displacement and acceleration, we conducted an experiment solely on velocity features. The resulting RMSE is 16 per cent higher than for the full feature set. The information gain from including features from the displacement, velocity and acceleration traces can be explained with the different frequency bands effectively covered by the features. While acceleration covers the higher frequency ranges, displacement covers mostly the lower frequencies. Hanks & McGuire (1981) discuss the relations between acceleration, velocity and displacement and argue that their values are affected differently by attenuation and that their peaks are expected to occur at different times in the waveform. This gives a further explanation for the information gains from incorporating all three feature classes.

Table 3. RMSE for different subsets of the correction functions. Full refers to the complete correction function, as described in Section 2.3. Distance-depth only contains the Γ term and the station corrections, but not the source corrections. Distance in addition reduces the Γ function to a 1-D function using hypocentral distance.

Corrections	Displacement NE	Acceleration *Z*
Full	0.196	0.159
Distance–depth	0.227	0.202
Distance	0.237	0.221

We additionally trained a boosting tree on the plain features from step one, without applying the correction functions from step two. In addition, we removed all features based on the radial and transverse components, as they can only be obtained if the epicentre location is known. It therefore uses no information about the location or time of the event, but only information gained from the single station. We experiment with both, a feature set with and without temporal information. In particular, the *S*–*P* arrival time difference represents a strong constraint on the hypocentral distance, which controls the dominant term of the correction function. For horizontal displacement both reduced feature sets still clearly outperform the (corrected) single feature baseline. The reduction in RMSE is 17 per cent with timing and 10 per cent without timing (Table 2, bottom part). For acceleration the RMSE is nearly identical with timing and 11 per cent higher without. We conclude that, when properly combined, the uncorrected features are already competitive with the single corrected features. As is natural, boosting tree regression on the corrected features clearly outperforms the uncorrected features.

To use our method in an early warning context, the system needs to deliver its estimate rapidly. Therefore we run an additional experiment using only the uncorrected data from the *P* wave. This information is available at the time of the S arrival. For the displacement magnitude, the RMSE is only 4 per cent worse than the single feature after applying corrections. For acceleration the RMSE is 25 per cent higher.

We want to emphasize that the complete feature set can be made available only 30 s after the *S* arrival. All *P* wave features are already available at the moment of the *S* arrival. While this is interesting for fast magnitude estimates, its applicability to early warning is limited. This is caused by the catalogue consisting mostly of small, non-hazardous events and the relatively far source station distance of up to 500 km. Applicability to early warning would need to be assessed on an appropriate catalogue.

4 DISCUSSION

4.1 Influence of different correction functions

We conduct an ablation study to quantify the impact of different correction terms on the residuals. We compare the full model to a model without source correction and a model without source correction and only a 1-D hypocentral distance correction. Similarly to Section 3.3, we conduct the analysis for horizontal displacement and vertical acceleration. The results are shown in Table 3.

Both features incur an improvement from both the 2-D correction as well as the source correction. The improvement from the 2-D source correction is around 4.2 per cent for displacement and 8.6 per cent for acceleration. The effect of the source correction is by far greater, with an additional improvement of 14 per cent for displacement and 21 per cent for acceleration. The combined improvement is 17 per cent for displacement and 28 per cent for acceleration with respect to a classic distance-only dependent correction function.

We next analyse the spatial distribution of residuals with and without source correction. Fig. 7 shows the residuals for station *PB*01. Without source correction they show clear spatial dependencies, while with source correction there is no pattern visible. Without source correction, there are strong azimuthal dependencies. This suggests that the residuals are dominated by path effects, which are similar across a wide distance range with the same azimuth, while being less affected by the properties of the physical source such as radiation pattern.

Changing the correction terms alters the resulting magnitude calibration. While removing the source correction only has a minor impact, switching from a 2-D to a 1-D correction, introduces a depth dependent offset between the scales. Unlike for distance- and source corrections, the depth is an inherent property of the event not improved by averaging. Therefore, the magnitude calibration is performed essentially for each depth level. For depth levels without events in the training set, strictly no magnitude could be determined except by interpolation or extrapolation. This also implies that the calibration of the 2-D correction could be more fragile, requiring careful testing of the performance with the test and validation sets (see also the Section 4.2).

Fig. 8 compares the distance and depth correction functions obtained with and without the source correction. The correction function without the source correction is significantly rougher, suggesting that the depth and distance correction function derived without a source correction term represents a biased estimate influenced by the particularities of the event distribution. Concurrent optimization of source correction and distance and depth correction therefore does not only yield a good source correction, but also improves the smoothness of the distance and depth correction. This suggests that it captures the actual average attenuation in the study region rather than the specifics of the data set.

4.2 Stability of the correction functions

We estimate the level of overfitting in our model by comparing the RMSE on the training and test sets. A high level of overfitting suggests that the model is not well constrained and poses an issue to interpretability. On average the RMSE on the test set is 2.7 per cent larger than on the train set. This increase is fairly constant across the different features, varying from 1.0 to 3.9 per cent. The only exception are 20 s *P* wave envelope values, probably because there are far less measurements. Their RMSE on the test set is on average 8.3 per cent higher. To assess the significance of the increases in RMSE, we evaluated the uncertainty of the RMSE values under the assumption that errors are uncorrelated and identically distributed. In this case the standard deviation of the RMSE is simply the RMSE divided by the square root of the number of samples, which comes out at around 0.1 per cent of the RMSE. Therefore all differences in the RMSE discussed here are significant. While these results show that the source correction functions are slightly underdetermined, the ablation study in Section 4.1 shows that this does not negatively impact their predictive performance.

We investigate how well the parameters of the correction functions are constrained by the given measurements. We therefore partition the set of events randomly into ten equal-sized, disjoint subsets and calibrate a model for each of those. Due to the source correction the changed numbers of events and measurements also change the number of model parameters. To ensure that the model differences are not dominated by changing the subset of events used for calibration with M_w, we always add the events with M_w to the subsets. We analyse models for peak horizontal displacement.

Results show that the station bias terms are robust. For stations with more than 2000 measurements in the complete data set, standard deviation between the ten sets is below 0.01. For stations with few measurements (<2000), we observe standard deviations up to 0.036. We emphasize that these deviations apply between the sliced sets containing less than 200 measurements each for these stations. On the full set this implies that we expect uncertainty of the station biases to be below 0.01 for all stations. Biases, uncertainties and number of measurements for each station are shown in Fig. F4.

The distance and depth correction is also very robust, albeit with a higher level of uncertainty than the station biases (see Fig. F5). At distances below 250 km and depths shallower than 100 km the standard deviation is still always below 0.05. Much higher standard deviations occur at large distances and depths, as data are very sparse there. Standard deviations of more than 0.1 solely occur for distances above 400 km and depths below 175 km. We want to emphasize that uncertainties on the final model are likely to be even smaller by a factor around 3, as it has been trained on ten times the data.

To assess the stability of the source correction, we evaluate the standard deviation between the ten subsets for 100 000 randomly chosen measurements. The mean standard deviation is 0.027. The 90th percentile is 0.039. The parameter uncertainties in the model are clearly below the random effects in the measurements.

In order to analyse the stability of the boosting tree scales we split the data set event-wise into three equal sized partitions A, B and C. We train one boosting tree on A and another one on B, both using the full feature set including timing. We compare the predictions of the boosting trees on C and also compare them to the non-boosting predictions on C. The target scale is again the horizontal peak displacement of the full wave. The event magnitude, averaged across all stations, differs between the non-boosting predictions and the boosting trees by 0.062 (0.063 for tree B) in quadratic mean. The two boosting tree scales only differ by 0.015 in this measure, even though they are trained on completely disjoint sets. The significantly smaller difference between the boosting scales suggests that the boosting trees are actually reducing estimation errors on the event magnitude. This does not hold true for the largest events (>6.0), as only relatively few of these events occurred in the observational period. We experience higher differences between boosting and non-boosting scales for those events, which are likely caused by sparse training data. Boosting tree scales should therefore not be used for the largest events.

4.3 Analysis of the correction functions

To analyse the different correction functions, we first need to emphasize the interconnections between them. Without regularization, the full distance correction could be incorporated into the source

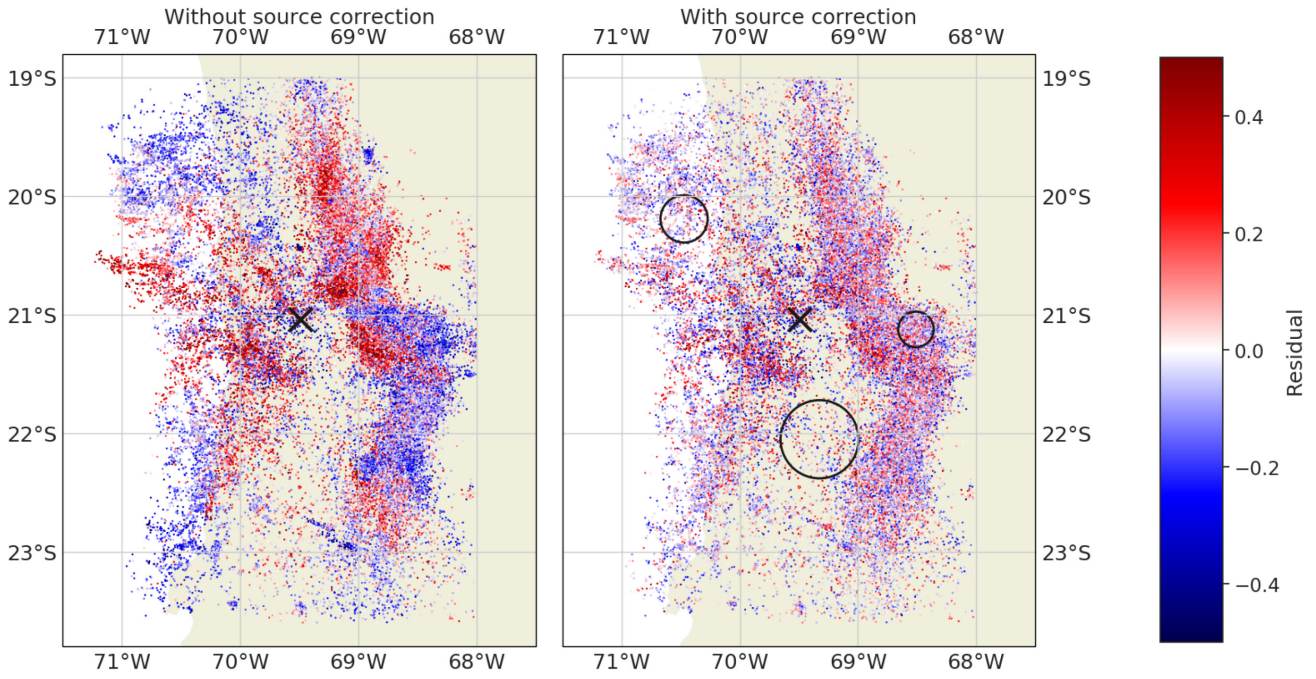

Figure 7. Spatial distribution of residuals at station PB01 for displacement magnitude on the horizontal components with and without source correction. The location of the station is denoted by a cross. While the source correction uses the 3-D location, we reduced the picture to 2-D for simplicity. The circles indicate the distance to the 10th nearest neighbour with a correction term to visualize the adaptive window size.

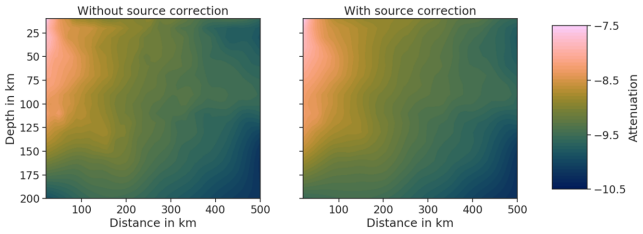

Figure 8. Comparison of the correction functions for displacement on the horizontal components with and without source correction terms.

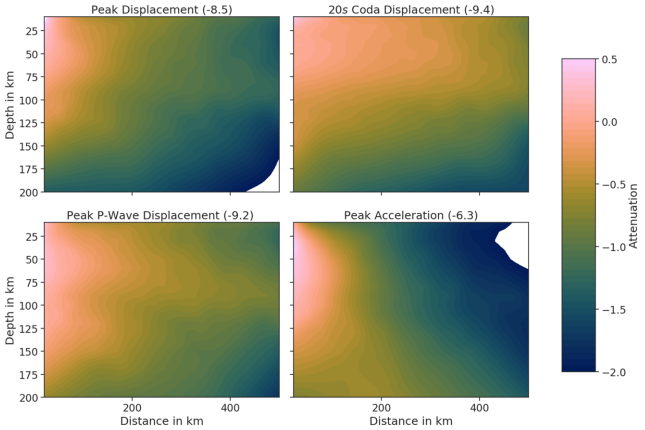

Figure 9. Distance and depth correction functions for selected features on the Z component. All corrections are shifted to 0 at a distance of 50 km and a depth of 30 km, to better visualize the relative differences. The shifts are denoted in brackets behind the title.

correction function, only requiring an offset to calibrate the correction function. We chose our regularization constants in a way that penalizes putting distance corrections into the source correction function. Nonetheless both interact and separation of the effects is not fully possible. In addition our stations and events are not uniformly distributed. Therefore, the distance correction function, being a mean across all stations and events, incorporates effects from the average paths, which do not necessarily reflect the average ground structure.

We compare the absolute values of the distance and depth correction functions between the different displacement, velocity and acceleration features. Due to different units, absolute values are not comparable between displacement, velocity and acceleration. Absolute differences in the correction function represent differences in the magnitude of the signal. For peaks from the full wave the signal level is similar for the *R* and *T* component, but around 0.15 orders of magnitude smaller on the *Z* component. For the *P* wave the signals are strongest on the *Z* and *R* component and about 0.1 orders of magnitude smaller on the *T* component. The envelope derived values are on average 0.4 orders of magnitude smaller after 5 s and 0.5 after 20 s.

Fig. 9 shows the comparison of four selected normalized correction functions. We focus on different features rather than components, as we observed no major differences between the different components. Comparing the peak displacement of the complete waveform with the peak *P* wave displacement, we see that the peak displacement shows a stronger attenuation with both distance and depth. The peak acceleration shows the strongest decay with distance, while being only weakly dependent on depth at near offsets. For far offsets (>∼250 km), deeper events actually are less attenuated then shallower events (opposite the pattern for amplitude). This effect could arise from a dominant importance of physical attenuation over geometric spreading.

In contrast, the 20 s envelope displacement amplitudes only shows a relatively weak dependence on distance. This lower attenuation stems probably from the fact that the envelope is made of

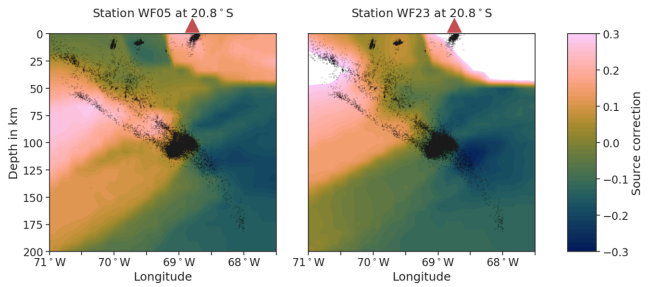

Figure 10. Section through the source correction terms for peak horizontal displacement of the stations WF05 and WF23 at 20.8°S, the latitude in the middle of the two stations. The positions of the stations are marked by red triangles. For orientation, events inside 20.8 ± 0.2°S are shown by black dots. Note that in areas without any seismicity, the correction term will be effectively controlled by the nearest seismicity, even if that is far from the point under consideration. While the source-specific correction term in those areas is likely to be biased, this normally does not matter, because hardly any seismicity occurs in these poorly constrained areas in any case.

multiply scattered waves; the theory of coda normalization predicts that energy will be distributed equally through all space and degrees of freedom after an asymptotically long time after the event (Sato *et al.* 2012). The remaining decay with distance stems most likely from the fairly short time window of 20 s used. This window is necessary to account for the many small events in the catalogue, for which the envelope values tend to fall below the noise quickly.

Fig. 10 shows sections through the source correction terms of the stations WF05 and WF23 at 20.8°S. The two stations are both located approximately 150 km from the coastline, with a distance of only about 40 km between each other. The source correction terms for the two stations are quite similar, which is consistent with the explanation that the source correction terms indeed capture large-scale path effects. The source correction terms also exhibit tectonic features. The most prominent is the sudden change for shallow earthquakes around 69°W. In addition the source corrections resolve the slab, which can be seen as a diagonal boundary in the corrections, approximately matching the slab determined by Sippl *et al.* (2018).

By comparing sections at different latitudes, we observed that the resolution of structural features becomes worse for sections further away from the station. As the source correction measures both source and path effects, for large distances it is dominated by aggregated path effects. Therefore the resolution of structural features gets worse. In contrast, the similarity between the corrections for nearby stations stays similar, as the paths get even more similar. We inspected many sections through the source correction volumes of different stations. All sections showed the clear change for shallow earthquakes around 69°W and an imprint of the slab geometry. In general, the sections for stations located close to each other were mostly very similar.

4.4 Insights into multifeature estimation

As boosting trees use decision trees as their base classifiers, they inherently lead to a ranking of features regarding their feature importance. Feature importance is derived from the information gain of the splits using this feature. We analyse the feature importance for the two target scales used in Section 3.3 (Table 4).

While we are not able to state the reason for the importance of this features with certainty, we provide some intuition. *P* wave features on the vertical and radial components are least affected by local site conditions. As shown before, the vertical component features from *P* waves have the lowest RMSE values among the *P* wave components. The envelope values are less affected by the radiation pattern as well as uncertainties in the location or correction functions. Both *P* wave features and envelope values have worse signal to noise ratios than features from the full waveform, making them less precise scales using only single features, while the combination of those features can likely be used to better separate signal from noise. We attribute the dominance of velocity features to two factors. In contrast to acceleration features, velocity features show less saturation, as discussed earlier. In addition, as our underlying data are velocity traces, velocity is not affected by artifacts from integration that occur for displacement.

To verify the presence of complex interactions, we compare the boosting tree to a simple linear regression. We use the full parameter set without timing information and the same target scales. Similar to boosting trees, linear feature combination significantly reduces RMSE. For displacement the RMSE is 0.133 (0.103 for the boosting tree) and for acceleration it is 0.120 (0.097). Although parts of the gain can be achieved with linear regression, a significant part of the improvement from the boosting tree is due to its capability to model non-linear relationships and complex interactions between multiple parameters.

Table 4. Top 10 features in the boosting regression ordered by importance. The columns denote whether the feature is from peak (no annotation) or envelope, if the feature is from the *P* wave, the trace it was exported from, and the component. NE refers to the horizontal components, ZNE to the combination of all components. We abbreviate displacement (DISP), velocity (VEL) and acceleration (ACC).

DISP NE				ACC Z			
	P	VEL	Z		P	VEL	Z
	P	VEL	R		P	VEL	R
5 s	P	DISP	Z	5 s	P	DISP	Z
	P	VEL	T		P	VEL	T
5 s	P	ACC	R	5 s	P	ACC	R
5 s		ACC	T		P	ACC	Z
5 s	P	DISP	R	5 s		ACC	T
	P	ACC	Z	5 s	P	DISP	R
5 s	P	DISP	T	5 s	P	DISP	T
		ACC	Z			VEL	Z

4.5 Magnitudes for the IPOC catalogue

Following our analysis, we provide well calibrated magnitude values for the IPOC catalogue. For each event we provide magnitude estimates from both the Wood–Anderson instrument and the peak displacement on the horizontal components. The former is chosen for its close resemblance of the standard local magnitude M_L, while the second offers a non-saturating alternative, which we refer to as M_A as proposed by Deichmann (2018a).

We additionally report uncertainty values for the magnitude estimates. We derive those uncertainties from the residuals between the stations. The detailed procedure for uncertainty estimation is described in the Appendix E.

We apply multiple steps to further increase the quality of the published scales. After calibrating and applying the correction functions, we remove all outliers. Outliers are defined as measurements with a residual of at least twice the global RMSE. We recalibrate the correction functions on the set without outliers. Due to overfitting,

we can not use a global boosting tree for the full data set. We therefore randomly split the data set event wise into three equal sized sets A, B and C. We train one boosting tree on each pair of these sets and use it to produce predictions on the last set. The analysis in Section 4.2 suggests that these predictions are consistent between the different boosting trees. This is especially the case, as, contrary to Section 4.2, the training sets of the trees are not even disjoint. Following the results from Section 4.2, we use the non-boosting estimates for events with magnitude >6.0. For events with magnitude <5.5 we use the boosting tree scales. We interpolate linearly for events of magnitude between 5.5 and 6.0 to obtain continuously defined scales.

To enable in-depth analysis we provide the full set of extracted features and magnitude predictions on station level in csv format.[1] We additionally provide our code to calibrate correction functions and train boosting tree models.[2] For convenience we also provide the calibrated correction functions for each feature in the data set.

5 CONCLUSION

In this study, we proposed and evaluated a three step procedure to evaluate magnitude scales based on many different features and reduce their uncertainty. In the first step we extracted a multitude of features from the seismic traces of single stations for over 100 000 events. In the second step we calibrated correction functions for each extracted feature. The correction functions consist of station biases, a 2-D non-parametric distance and depth correction and a 3-D source correction. We show that these correction functions reduce RMSE by up to $\sim$ 23 per cent in comparison to a classical 1-D non-parametric correction function. About three quarters of the gain can be attributed to the source correction, while the remaining gain stems from the 2-D distance and depth correction.

In the last step we use a boosting tree to combine multiple features. This further reduces uncertainties by up to $\sim$ 50 per cent. By analysing the feature importance, we show that key features are derived from the velocity on the Z component, especially from the P wave. The analysis also highlights the importance of envelope derived features. In conclusion of our analysis, we provide calibrated magnitude values M_A and peak Wood-Anderson based magnitude values (similar to standard M_L but with a richer calibration function) and their estimated uncertainties for the catalogue of Sippl *et al.* (2018).

The results from multifeature estimation, especially the results from the experiments with uncorrected features, give a hint at the wealth of information contained in a single trace. This information is not incorporated in the standard M_L, using only the peak displacement. It could however be of major interest for reliable magnitude estimates in the context of early warning. Promising tools for information extraction might be convolutional neural networks, which have lately been shown to be beneficial for multiple seismological tasks, including earthquake localization (Kriegerowski *et al.* 2018), phase picking and polarity determination (Ross *et al.* 2018). Lomax *et al.* (2019) use convolutional neural networks for earthquake monitoring, including rough magnitude estimations. Our results suggest that these estimates can be refined significantly.

This study did not focus on frequency dependency, but rather investigated effects on a broad frequency band. We pursued this approach to capture the wide magnitude range present in the catalogue. We acknowledge that attenuation functions are frequency dependent, as shown for example by Dawood & Rodriguez-Marek (2013) for the Japan subduction zone. This possibly is the cause of the increased RMSE values for larger magnitudes in our estimates, which will be based on longer period data less affected by physical attenuation. Incorporating frequency dependency into the model could also open up a perspective of applying the model to ground motion prediction.

While we applied the method to a catalogue of $\sim$100 000 events, our analysis in Section 4.2 suggests that our method can also be applied to significantly smaller data sets. All correction terms are already well defined with 10 000 events and we expect the boosting tree to work as well. For catalogues with more measurements per event, we even expect a by far lower required number of events.

[1] http://doi.org/10.5880/GFZ.2.4.2019.004
[2] https://github.com/yetinam/magnitude-calibration

ACKNOWLEDGEMENTS

Jannes Münchmeyer acknowledges the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS). We thank the two anonymous reviewers for their comments and suggestions, which have helped to improve the quality of the manuscript. We use color scales from Crameri (2018).

REFERENCES

Asch, G., Tilmann, F., Schurr, B. & Ryberg, T., 2011. *Seismic network 5e: Minas project (2011/2013),* 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.

Bindi, D., Schurr, B., Puglia, R., Russo, E., Strollo, A., Cotton, F. & Parolai, S., 2014. A magnitude attenuation function derived for the 2014 Pisagua (Chile) Sequence using strong-motion datashort note, *Bull. seism. Soc. Am.,* **104**(6), 3145–3152.

Bormann, P., 2012. *New Manual of Seismological Observatory Practice (NMSOP-2),* IASPEI, GeoForschungsZentrum.

Brillinger, D.R. & Preisler, H.K., 1984. An exploratory analysis of the Joyner-Boore attenuation data, *Bull. seism. Soc. Am.,* **74**(4), 1441–1450.

Cauzzi, C., Sleeman, R., Clinton, J., Ballesta, J.D., Galanis, O. & Kaestli, P., 2016. Introducing the European rapid raw strong-motion database, *Seismol. Res. Lett.,* **84**(4), 977–986.

Cesca, S., Sobiesiak, M., Tassara, A., Olcay, M., Günther, E., Mikulla, S. & Dahm, T., 2009. *The iquique local network and picarray,* doi:10.14470/vd070092.

Chen, T. & Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16,* pp. 785–794, ACM, New York, NY, USA.

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

Dawood, H.M. & Rodriguez-Marek, A., 2013. A method for including path effects in ground-motion prediction equations: an example using the Mw 9.0 Tohoku earthquake aftershocksmethod for including path effects in GMPEs Using Mw 9.0 Tohoku earthquake aftershocks, *Bull. seism. Soc. Am.,* **103**(2B), 1360–1372.

Deichmann, N., 2018a. The relation between ME, ML and Mw in theory and numerical simulations for small to moderate earthquakes, *J. Seismol.,* **22**(6), 1645–1668.

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

Dziewonski, A.M., Chou, T.-A. & Woodhouse, J.H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. geophys. Res.: Solid Earth,* **86**(B4), 2825–2852.

Eaton, M.L., 1992. A group action on covariances with applications to the comparison of linear normal experiments, *Lect. Notes-Monogr. Ser.,* **22,** 76–90.

Ekström, G., Nettles, M. & Dziewoński, A., 2012. The global CMT project 2004–2010: centroid-moment tensors for 13,017 earthquakes, *Phys. Earth planet. Inter.,* **200-201,** 1–9.

Festa, G., Zollo, A. & Lancieri, M., 2008. Earthquake magnitude estimation from early radiated energy, *Geophys. Res. Lett.,* **35**(22), doi:10.1029/2008GL035576.

Friedman, J.H., 2002. Stochastic gradient boosting, *Comput. Stat. Data Anal.,* **38**(4), 367–378.

GEOFON Data Center, 1993. *Geofon seismic network,* doi:10.14470/tr560404.

GFZ German Research Centre For Geosciences, Institut Des Sciences De L'Univers-Centre National De La Recherche CNRS-INSU, 2006. *IPOC seismic network,* doi:10.14470/pk615318.

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.: Solid Earth,* **104**(B9), 20 237–20 256.

Gurobi Optimization LLC, 2018. Gurobi optimizer reference manual, http://www.gurobi.com.

Hanks, T.C. & Kanamori, H., 1979. A moment magnitude scale, *J. geophys. Res.: Solid Earth,* **84**(B5), 2348–2350.

Hanks, T.C. & McGuire, R.K., 1981. The character of high-frequency strong ground motion, *Bull. seism. Soc. Am.,* **71**(6), 2071–2095.

Katsumata, A., 2001. Relationship between displacement and velocity amplitudes of seismic waves from local earthquakes, *Earth planet. Sci. Lett.,* **53,** 347–355.

Kriegerowski, M., Petersen, G.M., Vasyura-Bathke, H. & Ohrnberger, M., 2018. A deep convolutional neural network for localization of clustered earthquakes based on multistation full waveforms, *Seismol. Res. Lett.,* **90,** doi:10.1785/0220180320

Lancieri, M. & Zollo, A., 2008. A Bayesian approach to the real-time estimation of magnitude from the early *P* and *S* wave displacement peaks, *J. geophys. Res.,* **113**(B12), doi:10.1029/2007JB005386.

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,** 517–529.

Nábělek, J. & Xia, G., 1995. Moment-tensor analysis using regional data: application to the 25 March, 1993, Scotts Mills, Oregon, earthquake, *Geophys. Res. Lett.,* **22**(1), 13–16.

Nabelek, J.L., 1984, Determination of earthquake source parameters from inversion of body waves. *PhD thesis*, M. I. T., Dept. of Earth, Atmospheric and Planetary Sciences.

Picozzi, M., Bindi, D., Spallarossa, D., Di Giacomo, D. & Zollo, A., 2018. A rapid response magnitude scale for timely assessment of the high frequency seismic radiation, *Sci. Rep.,* **8**(1), doi:10.1038/s41598-018-26938-9.

Richter, C.F., 1935. An instrumental earthquake magnitude scale, *Bull. seism. Soc. Am.,* **25**(1), 1–32.

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

Sato, H., Fehler, M. & Maeda, T., 2012. *Seismic Wave Propagation and Scattering in the Heterogeneous Earth* , 2nd edn, Springer.

Savage, M.K. & Anderson, J.G., 1995. A local-magnitude scale for the western Great Basin-eastern Sierra Nevada from synthetic Wood-Anderson seismograms, *Bull. seism. Soc. Am.,* **85**(4), 1236–1243.

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.: Solid Earth,* **123**(5), 4063–4087.

Spallarossa, D., Kotha, S.R., Picozzi, M., Barani, S. & Bindi, D., 2019. On-site earthquake early warning: a partially non-ergodic perspective from the site effects point of view, *Geophys. J. Int.,* **216**(2), 919–934.

Universidad de Chile, 2013. *Red sismologica nacional,* doi:10.7914/SN/C1.

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.

Zollo, A., Lancieri, M. & Nielsen, S., 2006. Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records, *Geophys. Res. Lett.,* **33**(23), doi:10.1029/2006GL027795.

APPENDIX A: HIGHPASS FREQUENCY SELECTION

Table A1 shows the candidate intervals for high pass filtering. The last line indicates the fall-back filter, which is used for all events for which the minimum SNR of 4 is not attained with any of the other filters. For velocity (acceleration) the SNR is larger than 2 in 96 per cent (98 per cent) of the waveforms, whereas for displacement this is only true for 70 per cent. Therefore in some cases, particularly for features based on displacement, some of our data might be strongly affected by ambient noise. We nonetheless do not remove these measurements, as the information that the feature is close to noise is still valuable.

The distribution of chosen high pass frequencies by event magnitude is shown in Fig. A1. As expected, for larger events lower frequencies are chosen. Especially for the largest events, only the lowest frequencies are chosen.

Table A1. Intervals for high-pass filtering.

f_{low} [Hz]	f_{high} [Hz]
0.001	0.3
0.1	0.5
0.3	1.0
0.5	1.5
0.75	–

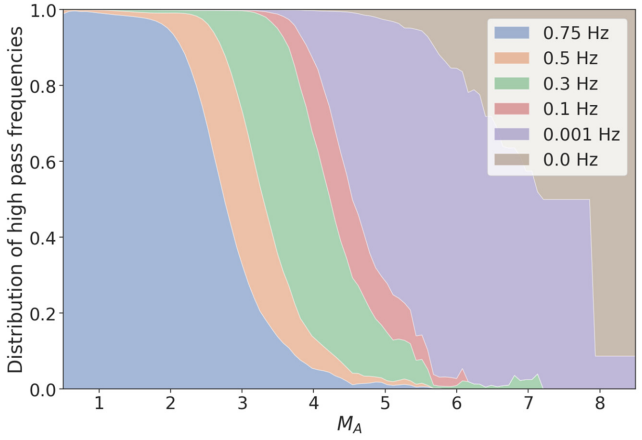

Figure A1. Distribution of applied high pass frequencies by event magnitude. Strong motion records were not high pass filtered and are therefore denoted with a high pass frequency of 0 Hz.

APPENDIX B: CHOICE OF HYPERPARAMETERS AND ENVELOPE TIMES

In this section we give some advice on the selection of hyperparameters and envelope delays. As the experiments, both feature extraction and calibration of the correction functions, are computationally expensive, a grid search for hyperparameter selection is intractable. Hyperparameters therefore need to be tuned by hand. Therefore, we explain the significance of and interaction between the different hyperparameters. For practical applications we suggest to start with the hyperparameters used in this study.

λ_r and λ_d determine the smoothness of the distance-depth correction function. We settled for a higher value of λ_r as we expect a generally lower lateral than vertical variability in ground structure. Both values might need to be increased in the presence of fewer data points and the other way around. λ_d should be increased, if less M_w values are available for the calibration of attenuation with depth. The choice of suitable values can be assisted by plots, as in Fig. 9.

λ_L controls the level of deviation from the distance-depth correction that is caused by the source-path correction. It interacts with the number of neighbours k chosen for averaging and the subsampling rate $|\bar{E}_s|/|E_s|$. In general, a low number of neighbours k or a high subsampling requires a higher λ_L, as the number of free parameters is increased and the parameters are less constrained by the data.

k determines the smoothing of the source-path correction. A higher value will generally cause a smoother function, while a lower value will cause a rougher function. In contrast a higher subsampling rate (at constant k) will cause a rougher function, a lower subsampling rate a smoother function. The choice of subsampling rate will most likely be governed by the available computational capacities. We experienced a superquadratic increase in runtime and memory consumption with the subsampling rate. If the computational capacities turn out to be limiting factors, we recommend slowly increasing the subsampling rate and observing the effect on RMSE.

λ_{M_w} determines the trade-off between the deviation from the prescribed M_w values and the smoothness of the correction functions. A higher value λ_w will lead to a smaller deviation from M_w, but increases the roughness of the correction functions. As the calibration with M_w is mostly required for the calibration of the depth-dependent attenuation, we generally recommend small values for λ_{M_w}.

A good measure for the suitability of hyperparameters is the difference between the RMSE on the training and development sets. In general we recommend a slightly higher RMSE on the training set, indicating some level of overfitting. No overfitting at all suggests that the model is regularized to strongly, while strong deviations between the training and development performance suggest that overfitting negatively impacts performance on the development and test set.

For the envelope delays we chose 5 and 20 s. The 5 s value is intended to capture the early high energy portion of the event, but providing a more stable measurement than the peak. We tried putting the second value as late as possible to approach the diffusive regime and thereby minimize the effects of the radiation pattern and distance uncertainties. As most of our events are small, we can not resort to the classical rule of assuming a diffusive regime after twice the S wave travel time, as this value is below noise level for most measurements. Therefore we need to find a sensible trade-off between diffusiveness and SNR. Whereas we did not carry out systematic testing, we confirmed 20 s as a good choice by comparing the value of the envelope at this time to the noise level 5 s before the P pick, as measured by the envelope value. We found that the noise exceeds the signal in only ~ 3 per cent of cases. In addition we expect the boosting tree to appropriately handle low SNR 20 s envelope values.

The proper choice of envelope delays will usually depend on the data set. In our case we had a favorable data set for long envelope delays, as most IPOC stations are low noise hard rock stations. To choose appropriate values we recommend first to visually inspect the signal envelopes for a subsample of the measurements and second to look at the SNRs for multiple candidate delay times. It is possible to include more than two envelope times. We did not conduct experiments with more than two envelope times, due to computational constraints.

Table B1. Hyperparameters used for the correction functions.

Hyperparameter	Value
G	$\{20\ \mathrm{km} + 9.8\ \mathrm{km} * i \mid i \in \{0, 49\}\}$ $\times$ $\{10\ \mathrm{km} + 10\ \mathrm{km} * i \mid i \in \{0, 19\}\}$
λ_r	$10^3\ \mathrm{km}^4$
λ_d	$10^2\ \mathrm{km}^4$
λ_L	10
λ_{M_w}	10^{-1}
k	10
$\lvert\bar{E}_s\rvert/\lvert E_s\rvert$	10^{-1}

Table B2. Hyperparameters used in the boosting experiments. We use the naming conventions from XGBoost. We only denote parameters that were changed from the defaults for XGBoost version 0.80.

Hyperparameter	Value
Depth	11
Epochs	250
Eta	0.1

APPENDIX C: DETERMINATION OF MOMENT MAGNITUDES FOR MODERATE-SIZE EARTHQUAKES

The global CMT catalogue only covers earthquakes above moment magnitude 5–5.5 reliably. In order to extend our database of events with M_w, additional moment magnitudes were determined with regional moment tensor inversion with the approach of Nabelek (1984) and Nábělek & Xia (1995). We constrained moment tensors to be deviatoric (i.e. no isotropic component), used the period band between 10 s and 35 s and assumed quality factors (inverse attenuation) of 225 for P and 100 for S waves for the calculation of Green's functions. Scalar moments were converted to moment magnitudes using the relation of Hanks & Kanamori (1979). At the utilized long periods, physical attenuation effects only play a minor role.

APPENDIX D: EFFECT OF SNR THRESHOLDING

No explicit SNR threshold is imposed but an implicit threshold exists because the data set is assembled based on pre-existing picks, which require a reasonable visibility of at least the P wave. We analysed the impact of imposing an additional threshold on the SNR

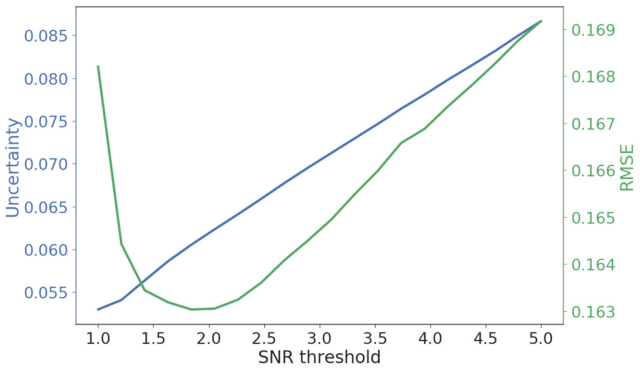

Figure D1. RMSE and resulting uncertainties for the single feature magnitude scale from displacement on the vertical component at different signal-to-noise thresholds.

on the RMSE and the resulting uncertainties (Fig. D1), initally using the vertical displacement magnitude as an example. We obtain the noise level for this analysis as the peak value in the 30 s before the *P* pick, with an additional safety window of 1 s. For each SNR threshold we calculate the RMSE using only measurements with a higher SNR and estimate the uncertainty on the mean. We estimate the uncertainty as the RMSE divided by the square root of the number of stations for each event minus one. As a higher SNR threshold causes a lower number of measurements, the average uncertainty can increase, even if the RMSE falls.

As we see in D1, the RMSE falls for SNRs of up to ∼2 and grows afterwards. The growth can be explained by the fact that measurements with a higher SNR are more often from events with higher magnitudes, which exhibit an increased RMSE in general (Fig. 6). In contrast to the RMSE, the uncertainty does not show any decreasing behavior, but a steady growth due to the decreasing number of measurements. We observe similar behavior for velocity and acceleration. This means that the general quality of our estimates is highest if we do not impose a further SNR threshold. In addition we expect the boosting tree regression to act as denoising, as it is able to combine multiple features representing different frequency spectra.

APPENDIX E: DETERMINATION OF MAGNITUDE UNCERTAINTIES

To obtain magnitude values and uncertainties for each event, we combine the measurements from multiple stations. As the results from multiple stations might not be independent, the stated uncertainty of the magnitude estimate could be erroneous if it is calculated by ignoring possible correlations. Figs E1 and E2 show the correlations between the residuals at pairs of stations and their dependency on interstation distance. Interestingly correlation shows a strong dependence on distance and especially gets negative for distances above ∼100 km. The negative values are partially caused by analysing the residuals with respect to the mean rather than the (unknown) true value. This effect alone causes some apparent negative correlation, but for truly independent errors this would be much smaller than observed.

Determining the optimal estimator and the effective sample size has been discussed by Eaton (1992). Unfortunately, the suggested method uses the inverse of the correlation matrix, which is unstable regarding minor variations of the covariance matrix. This is especially problematic, as we do not have access to the actual correlation matrix, but only to an empirical covariance matrix. In addition we

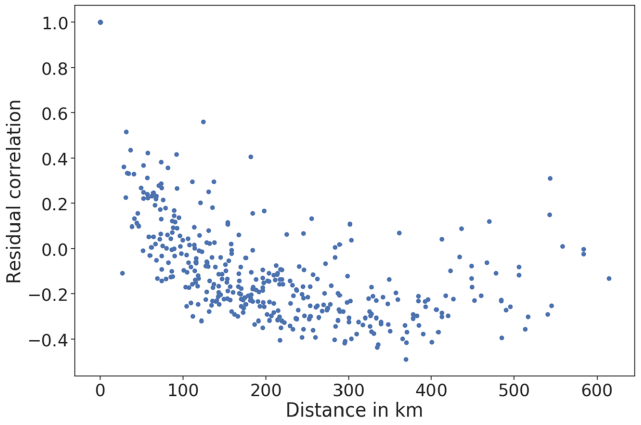

Figure E1. Empirical correlation of the residuals for peak horizontal displacement as a function of interstation distance. Each dot represents a pair of stations. Station pairs with less than 500 common events are discarded.

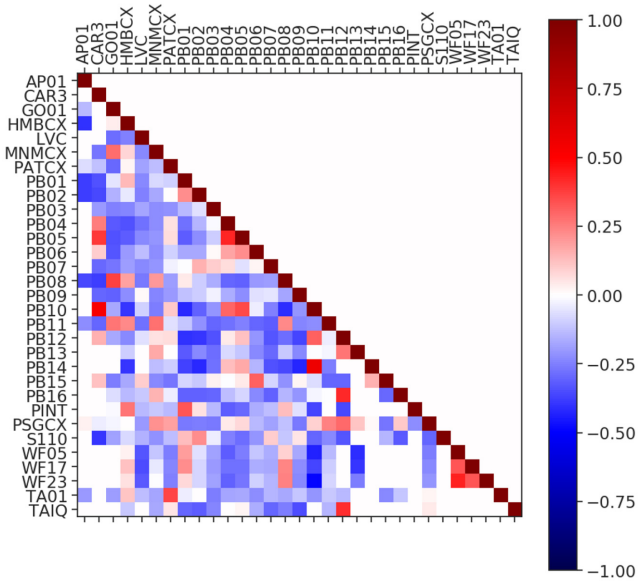

Figure E2. Empirical correlation of the residuals for peak horizontal displacement for station pairs. Station pairs with less than 500 common events are discarded.

are missing some elements of the matrix, for stations with too few events in common. Therefore, the proposed method is not applicable.

Nonetheless we want to present two main results from Eaton (1992). First, a growing correlation does not always reduce effective sample size, but can actually increase it. Secondly, negative correlations in general increase the effective sample size.

Following these observations we adapt a simply *ad hoc* procedure. The mean observed correlation between pairs of stations is close to zero (−0.1). Therefore, we use the mean of all stations as the event magnitude and the standard deviation between the single station estimates divided by the square root of the number of contributing stations minus one as the event magnitude standard deviation. Even though this does not represent the optimal way, following the discussion above we believe to achieve reasonable uncertainty estimates using the method.

APPENDIX F: SUPPLEMENTARY MATERIAL

Table F1. Seismic networks and stations used. Stations including strong motion records are printed in bold. The stations are identical to those used by Sippl *et al.* (2018) except that station PB17 from the CX network was removed because it showed non-documented gain changes over time and for the different components.

Network		Stations
MINAS	(5E)	S110
CSN	(C)	AP01 GO01 TA01
IPOC	(CX)	CAR3 **HMBCX MNMCX PATCX PB01 PB02 PB03 PB04 PB05 PB06 PB07 PB08 PB09 PB10 PB11 PB12 PB13** PB14 **PB15 PB16 PSGCX** TAIQ
GEOFON	(GE)	LVC
Iquique	(IQ)	PINT
WestFissure	(8F)	WF05 WF17 WF23

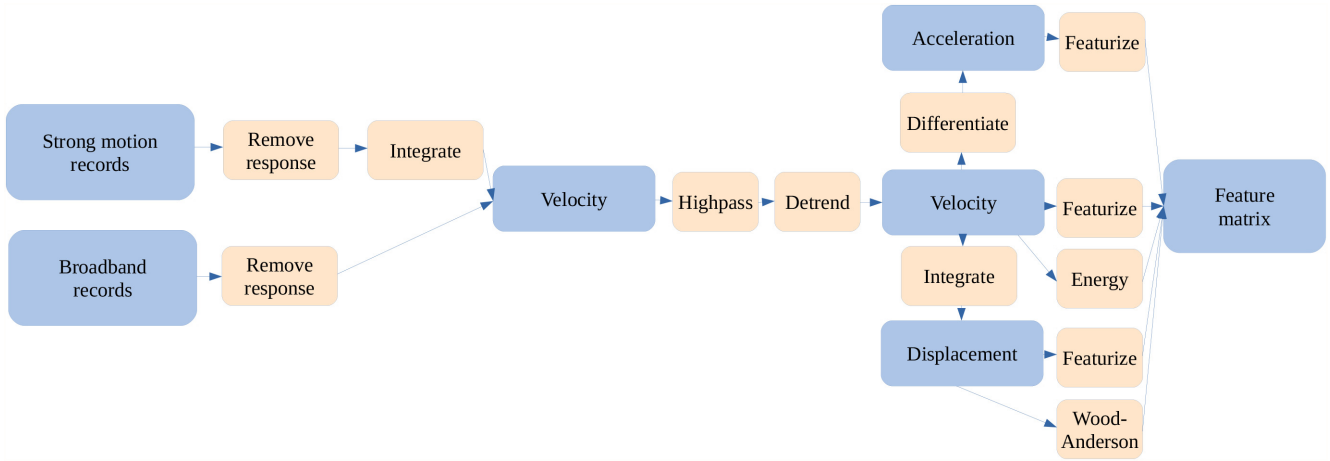

Figure F1. Schematic overview of the pre-processing and feature extraction workflow. The split into the different components is not visualized to keep the figure simple. Featurize refers to the process of extracting the peak and envelope values from the traces.

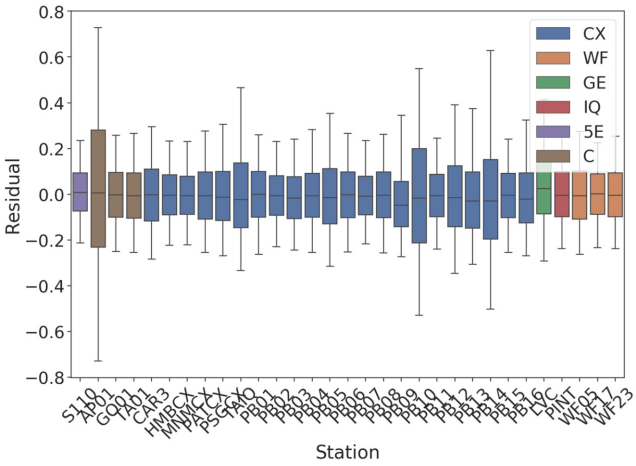

Figure F2. Residual distribution for different stations for displacement on the horizontal component. The middle bar denotes the median, the boxes show the quartile ranges, the whiskers show the 5th and 95th percentiles. Most stations have residuals of similar magnitudes, while a few show significantly higher residuals, e.g. AP01, TAIQ, PB10 and PB15.

Figure F3. Development of residuals for displacement on the horizontal component for station PB01 over time. The lines show running mean and standard deviation over 500 consecutive events. While we observed slight changes in the station bias over time, we were not able to ensure that these changes are not caused by measurement artifacts.

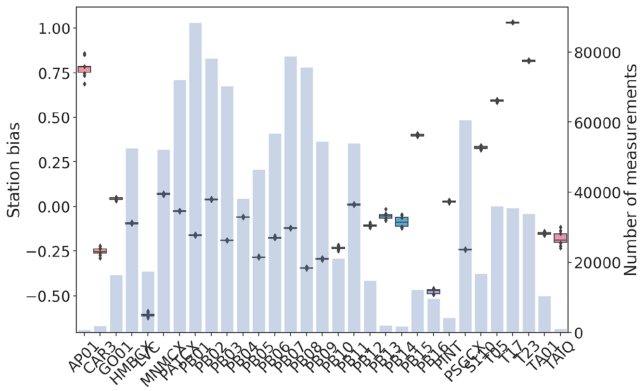

Figure F4. Station bias for peak displacement on the horizontal component. The bias is shown for ten suboptimizations, each containing 10 per cent of the events. Boxes indicate quartiles. The blue bars show the total number of measurments per station.

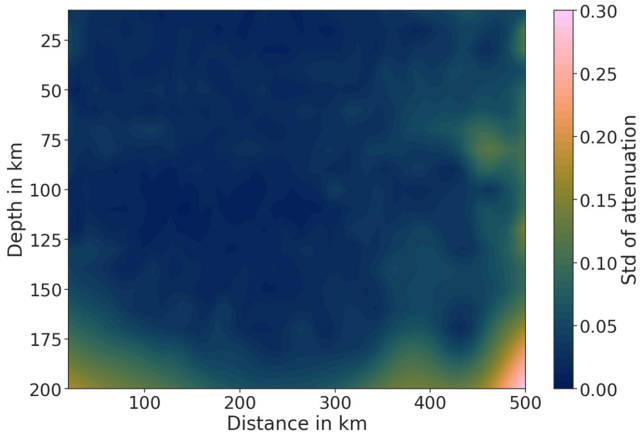

Figure F5. Standard deviation of the distance and depth correction function grid points for peak displacement on the horizontal component. Standard deviation is calculated across the subsets of a 10-fold split of the full data set.