A Neural Network for Automated Quality Screening of Ground Motion Records from Small Magnitude Earthquakes

Xavier Bellagamba,[a)] **M.EERI, Robin Lee,**[a)] **M.EERI, and Brendon A. Bradley,**[a)] **M.EERI**

The ambitious scopes of recent earthquake ground motion studies are generating a need for more high-quality ground motion records. As the number of deployed sensors is rapidly growing through improved accessibility and cost (e.g., ground motion stations, low-cost accelerometers, smart phones), an exponentially increasing amount of data are being generated. Previously, quality-assured ground motion data sets for engineering applications were generated using both manual and automated quality screening methodologies. More recently, new techniques have emerged that potentially offer both improved classification accuracy and computational expediency. This work presents a machine learning–oriented method to facilitate and accelerate the quality classification of ground motion records from small magnitude earthquakes. Feedforward neural networks are selected for their ability to efficiently recognize patterns and are trained on two New Zealand data sets. An application to physics-based ground motion simulation validation indicates that the proposed approach delivers results that are comparable with manual quality selection. Robust automatic ground motion quality screening allows a significant increase in data set size for development, calibration, and validation of ground motion models. [DOI: 10.1193/122118EQS292M]

INTRODUCTION

The quality of earthquake-induced recorded ground motions has been a topic of discussion for decades, previously with respect to analog instruments, but more recently with digital instruments (Hudson 1979, Douglas 2003, Boore and Bommer 2005, Douglas and Boore 2011). Here, quality refers to how well the instrument recording (comprising signal and noise) represents the actual ground shaking (signal) at a particular location. In this context, the content of the ground motion record attributed to the shaking caused by the earthquake (the desired information) is considered the signal, and any error or undesired disturbance is considered noise. As the quantity of operational strong ground motion recording stations are continuously increasing and providing better spatial coverage, the need for automated determination of ground motion quality becomes a necessity. This is further accentuated by the developing technology for low-cost recording instruments, which could realistically lead to a big data explosion (Anthony et al. 2018).

[a)] Civil and Natural Resources Engineering, University of Canterbury, 20 Kirkwood Avenue, Upper Riccarton, Christchurch 8041, New Zealand; Email: xavier.bellagamba@gmail.com (X. B.)

Earthquake Spectra, Volume 35, No. 4, pages 1637–1661, November 2019; © 2019, Earthquake Engineering Research Institute

Ground motions are inherently complex as a result of the underlying nature of earthquake rupture, wave propagation, and local site effects. Each of these effects can influence the amplitude, frequency content, and duration of ground motions. Small magnitude earthquakes (e.g., $M_w \leq 5.0$) with small amplitude shaking, in particular, are more likely to have their quality compromised because of noise that may have comparable amplitudes and can contaminate the record. At low frequencies (below 1 Hz), seismic noise is predominantly due to natural causes, such as ocean waves, variations in atmospheric pressure, and wind. At high frequencies (above 1 Hz), seismic noise is caused by human activities, such as motor vehicles, industrial work and machinery, and electrical currents as well as natural sources, such as rivers (Okada and Suto 2003). While conventional signal-processing techniques, such as baseline correction and filtering, are able to correct records for some types of noise, often, ground motions are affected in a way that no amount of processing can reliably recover the signal. These low quality recordings are detrimental for ground motion studies and are generally discarded to maintain robust inferences.

The most commonly used metric for determining ground motion quality is the signal-to-noise ratio (SNR; Ancheta et al. 2014, Dawood et al. 2016, Kishida et al. 2017, Rennolet et al. 2018). Fundamentally, SNR is a measure of the strength of a signal relative to background noise. This requires identification of the signal and noise, which can be difficult to isolate. To simplify the calculation, it is common to take the noise as the pre-event duration of the ground motion record or to use a generic white noise corresponding with expected amplitudes and frequency content (e.g., Cauzzi and Clinton 2013). The signal can also be isolated but is often taken as the entire ground motion record for simplicity (Boore and Bommer 2005, Dawood et al. 2016). There are several mathematical definitions for SNR. The most simple definition is to take the peak ground acceleration (PGA) and divide it by the peak absolute acceleration of the noise. A more complex definition, from an energy perspective, compares the Arias intensity (AI) of the signal with that of the noise. Boore and Bommer (2005) also suggests other means to check the quality of ground motions, including examining the shape of the Fourier amplitude spectra (FAS) and integrated velocities and displacements, to see if they are realistic. Additionally, Dawood et al. (2016) also enforced limits on final displacements and velocities. However, many of these tests are inherently manual and hence time-consuming when conducted for a large number of ground motions.

In recent years, there have been large ground motion data sets created using various methods of quality screening and processing (Ancheta et al. 2014, Cauzzi et al. 2016, Dawood et al. 2016, Van Houte et al. 2016). While the ground motion processing steps are broadly similar, comprising baseline corrections, instrument corrections (if necessary), and filtering at low and high frequencies, the tests for whether a ground motion is of acceptable quality varied. The Pacific Earthquake Engineering Research Center record-processing methodology (Ancheta et al. 2013), for example, employed in the development of the NGA-West2 database (Ancheta et al. 2014) is one of the most commonly adopted methodologies. However, a significant drawback of this methodology is the need for manual, user-determined high-pass and low-pass filter frequencies, subsequent visual inspection of corrected displacements, and review of Fourier spectra, which can be excessively time-consuming. Dawood et al. (2016) developed a step-by-step automated protocol to systematically process ground motion records, which alleviated the need for manual intervention, by iterating processing

parameters (e.g., high-pass corner frequency) until the record met some predefined criteria or was otherwise discarded.

Despite numerous challenges (nonuniform geospatial and temporal recordings, amorphous boundaries of studied objects or highly multidimensional problems), machine learning applications in geosciences are gaining momentum (Karpatne et al. 2018). Such methods are now used to detect hurricanes (e.g., Racah et al. 2017) or monitor volcanic activity (e.g., Titos et al. 2018). In seismology, machine learning techniques are applied to a relatively broad range of problems, such as early detection (e.g., Kong et al. 2016, Li et al. 2018), phase picking (e.g., Ross et al. 2018, 2019), or earthquake location detection (e.g., Perol et al. 2018). More examples of applications are given in Kong et al. (2018). However, to the knowledge of the authors, little effort has so far been dedicated to optimize the expediency of ground motion record quality screening.

This work presents a machine learning–based method that automatically classifies ground motion records based on their underlying characteristics. The training data set comprises both pre-existing and newly classified data. The former has been previously used to validate physics-based ground motion modeling in the Canterbury, New Zealand, region, whereas the latter comes from the Wellington, New Zealand, region and has been classified to increase the data set diversity. The proposed classifier is based on a feedforward neural network (described in the *Feedforward Neural Network For Quality Screening* section) based on its recognized ability to efficiently detect, extract, and recombine patterns hidden in the data (Hastie et al. 2008, p. 352). This paper presents the details of the utilized data set, the assumptions and methodology to build the neural network classifier, its validation and effects on the distribution of the selected ground motion records from both an intensity measure and a geospatial perspective, and finally its application to physics-based ground motion model validation.

GROUND MOTION DATA SET

DATA SET CHARACTERISTICS

To develop a neural network for ground motion quality classification, a large data set is required for model training and validation purposes. This study considers ground motion recordings from small magnitude earthquakes ($3.5 \leq M_w \leq 5.0$) that have occurred in the Canterbury and Wellington regions of New Zealand between 2003 and 2018, as shown in Figure 1. The two regions are geographically separated by approximately 300 km, resulting in diverse source, path, and site characteristics. The considered Earthquakes generally occur on active shallow crustal faults (i.e., centroid depth less than 20 km), and shallow site conditions vary from softer sedimentary basin deposits (gravel and marine fines) to harder rock (Begg and Johnston 2000, Forsyth et al. 2008).

The Canterbury data set is based on the ground motions considered in a hybrid broadband ground motion simulation validation study conducted by Lee et al. (2019), with 3,989 records from 148 earthquakes across 43 stations. This subset of data has been geographically limited for the application that it was initially intended for and is used as-is (this limitation can be alleviated in the future, as more data can be assimilated into the model developed in this study). Source-to-site distances do not exceed 100 km as a result of this geographic limitation. The Wellington data set has been specifically developed for this study, with

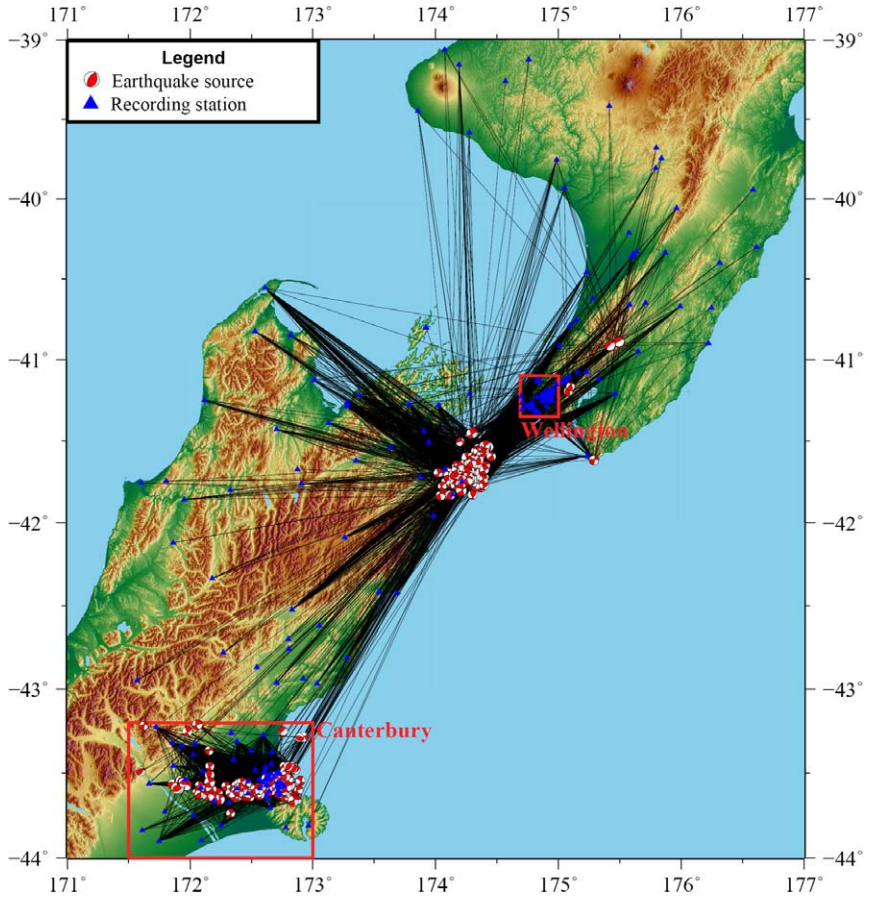

Figure 1. In this study, 327 earthquake sources with 8,467 ground motions recorded across 195 stations are considered. Schematic ray paths of observed ground motions are also shown as black lines. The Canterbury and Wellington regions are explicitly highlighted.

4,478 records from 179 earthquakes across 142 stations. This subset of data has no enforced geographic limitation. Collectively, this amounts to a data set of 8,467 ground motion records. Instruments used to record the ground motions generally have flat response between roughly 0.1 to 50 Hz or 0.1 to 200 Hz (Patterson et al. 2007, Van Houte et al. 2016). Sample rates of instruments at broadband stations are mostly 50 Hz, while instruments at strong motion stations are usually 200 Hz.

MANUAL GROUND MOTION QUALITY SCREENING

In determining the quality score of a ground motion record, several key pieces of information are considered from both raw and processed records. Here, raw records refer to version 1 unprocessed records obtained from the GeoNet file transfer protocol (GeoNet 2018). Two variations of processed records are considered to investigate the ground motions at

different frequencies. Both variants are baseline corrected (by removing the mean of the record and the remaining linear trend) but are bandpass filtered at different frequencies (with fourth-order Butterworth filters). One set of processed motions is filtered with a high-pass frequency of 0.08 Hz (to retain Fourier amplitudes at 0.1 Hz) and a low-pass frequency of 20 or 50 Hz for records with sample rates of 50 or 200 Hz, respectively, and is referred to as broadband processed. The other set of processed motions is filtered with high-pass and low-pass frequencies of 0.08 and 1 Hz, respectively, and is referred to as low frequency processed.

The quality of the ground motions was manually determined primarily by visual examination of time series and FAS plots of the raw and processed records. The raw records were examined to determine how the time series and FAS looked as they were recorded (allowing for determination of SNR), the broadband-processed motions provided insight on how the records appear in engineering applications, and the low frequency–processed motions emphasized the characteristics of the low-frequency motion (which have small amplitude relative to higher frequencies). Several criteria were considered when visually examining the ground motion time series. A compact summary is presented here, while a comprehensive guideline is included in the online Appendix A.

1. Comparison of the acceleration amplitudes of the pre-event noise with that of the earthquake ground motion signal.
2. Inspection of the acceleration amplitudes at the end of the record to determine if the ground motion has adequately finished or terminated early in the coda.
3. Comparison of record FAS with noise FAS, effectively measuring frequency-dependent SNR.
4. Inspection of the shape of the record FAS, with emphasis on the sloped/decaying low-frequency branch.
5. Inspection of the velocity time series obtained via integration of the filtered acceleration.

Based on the adopted criteria, each three-component set of ground motions for one source-site pair (two horizontal components and one vertical component) are scored with discrete values of 0, 0.25, 0.5, 0.75, or 1, where 0 represents the lowest quality and 1 represents the highest quality. Scores of 0, 0.25, and 0.5 are considered low quality and scores of 0.75 and 1 are considered high quality. Records that have misleading or wrong metadata, distinguished multiple wave trains corresponding to multiple earthquakes, or electronic malfunctions, or those that were triggered late such that the P-wave arrival was not recorded, were manually removed if encountered. It is important to note that the manual classification performed here has an engineering-oriented application mindset, so ground motion features that are considered important may differ for other applications.

Table 1 provides the results of the manual quality classification for the Canterbury and Wellington regions. The data set across both regions is relatively balanced between high-quality ground motions and low-quality ground motions. The Wellington data set also has a larger proportion of 0.25–0.75 scores, as the ground motions were more difficult to definitively classify as high or low quality. Figure 2 presents the distribution of high and low-quality ground motions, showing relatively wide coverage across all magnitudes

Table 1. Distributions of manually determined quality scores for the Canterbury, Wellington, and combined data sets

Quality score	Canterbury	Wellington	Combined
1	1,729	1,313	3,042
0.75	431	916	1,347
0.5	11	88	99
0.25	67	365	432
0	1,751	1,796	3,547

and source-to-site distance. The histograms (which are partially transparent with high and low-quality bars overlapping) show that there is a higher proportion of high-quality ground motions occurring from relatively larger-magnitude earthquakes and at shorter source-to-site distances and vice versa for low-quality ground motions. As can be expected, this indicates that, for the same magnitude, signal quality tends to deteriorate with distance to the site.

GROUND MOTION QUALITY METRICS

To develop a neural network for ground motion quality classification, 20 ground motion quality metrics were defined and calculated related to the amplitude, frequency content, and duration to separate high and low-quality ground motions. The considered metrics are presented in Table 2 with comprehensive details regarding the calculation of the metrics detailed in the online Appendix B.

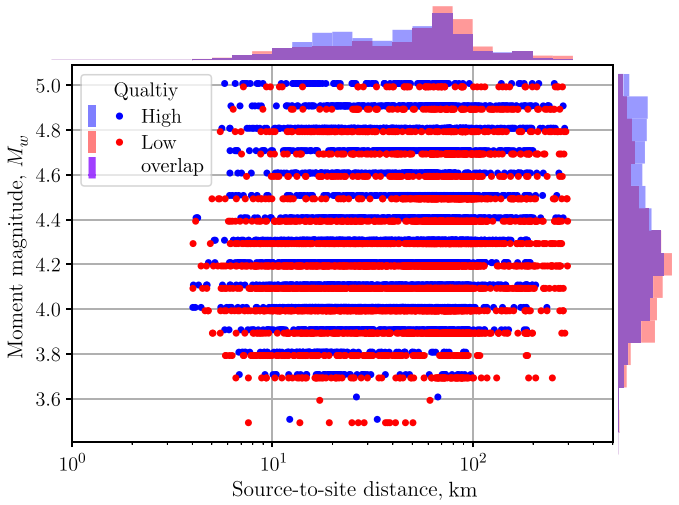

Figure 2. Earthquake ground motion magnitude and source-to-site distance distributions for manually classified high and low-quality ground motions.

Table 2. Ground motion quality metrics used to quantitatively characterize ground motion records and their associated broader categories

Category	Detailed metric
Time domain acceleration amplitude ratios	A: Peak noise to PGA ratio
	B: Average tail ratio
	C: Maximum tail ratio
	D: Average tail noise ratio
	E: Maximum tail noise ratio
	F: Average head ratio
Ground motion durations and duration ratios	G: 10%–20% bracketed duration ratio
	H: 5%–75% significant duration
	I: 5%–95% significant duration
Fourier amplitude ratios	J: Low frequency (below 0.1 Hz) pre-event FAS to maximum signal FAS ratio
	K: Low frequency (below 0.1 Hz) entire signal FAS to maximum signal FAS ratio
	L: Fourier amplitude ratio
SNR	M: Minimum SNR
	N: Maximum SNR
	O: Average SNR
	P: Average SNR between 0.1–0.5 Hz
	Q: Average SNR between 0.5–1.0 Hz
	R: Average SNR between 1.0–2.0 Hz
	S: Average SNR between 2.0–5.0 Hz
	T: Average SNR between 5.0–10.0 Hz

Note: Further details are provided in the online Appendices A and B.

With reference to Table 2, the various time domain acceleration amplitude ratios serve different purposes. The peak noise-to-PGA ratio (*A*) is a simple SNR based on time domain acceleration amplitudes. The tail ratio metrics are used to determine if the acceleration amplitudes at the termination of the record are small enough relative to the PGA (*B* and *C*) and peak noise (*D* and *E*) to consider that significant earthquake-induced ground motions have ceased. The maximum head ratio (*F*) is used to ensure that records do not begin with relatively large acceleration amplitudes. The 10%–20% bracketed duration ratio (*G*) is a metric that describes the relative shape of the acceleration time series, while significant durations (*H* and *I*) help identify ground motions that may be too long for a small magnitude rupture. The Fourier amplitude ratios (*J*, *K*, and *L*) are used to define the behavior of the FAS at low frequencies. Lastly, the SNR metrics are used to quantify the strength of the recorded earthquake-induced ground motion to the recorded noise. In order to classify a bidirectional horizontal ground motion recording (rather than individual classifications for each horizontal component), the values for each metric are geometric means of the two horizontal components of the raw (unprocessed) records.

The analyst determination of ground motion classification requires a heuristic examination of numerous ground motion features (such as those represented by the metrics in Table 2) and their joint influence on an overall quality assignment for the intended use of the ground motion record. As such, the interaction between the numerical values of each metric can be viewed as complex and nonlinear; thus the problem is well suited to solution via machine learning.

To highlight the difference between a high and low-quality ground motion records, which would be scored 1 and 0, respectively, an example of each is provided in Figure 3. Figure 3a and 3b presents the raw acceleration time series of one horizontal component of a high-quality record and a low-quality record, respectively. Annotated on the plots are the maximum absolute acceleration amplitudes of the noise/pre-event trace (PN), entire record (PGA), and tail end (PT), and the dashed line denotes the P-wave arrival. Based on this comparison, it is evident that the PN and PT are small relative to the PGA for the high-quality record compared with the low-quality record. Large PN/PGA values suggest that the noise is significant relative to the earthquake signal, while large PT/PGA values suggest that the ground motion records may have been terminated early in the coda. Figure 3c and 3d present Husid plots corresponding to Figure 3a and 3b, respectively. Prior to the arrival of seismic waves, the high-quality record has practically zero buildup in normalized cumulative AI compared with the low-quality record, which has roughly 3% at the P-wave arrival. Figure 3e and 3f presents the FAS of the pre-event trace and entire record with corresponding smoothed spectra. The FAS of the high-quality signal has a shape that is broadly consistent with seismic theory, with slopes at low and high frequencies that are linear in log space and plateau at moderate frequencies, while the low-quality signal has an irregular shape. Figure 3g and 3h presents the SNR of the records, in which the high-quality record has large SNR at frequencies of interest (0.1–20 Hz), while the low-quality record has low SNR (<1) at low frequencies.

FEEDFORWARD NEURAL NETWORK FOR QUALITY SCREENING

The quality screening of ground motion records is a multivariate and subjective classification problem. Furthermore, the potentially large data set size makes it an ideal target for machine learning–based approaches. As discussed subsequently, selected quality metrics show nonlinear interactions that complicate the record classification. The idea is therefore to train an algorithm that searches for specific patterns in the data that qualify high-quality ground motion records. Neural networks have been proven particularly efficient in pattern recognition (Hastie et al. 2008, p. 352). They also have better predictive capabilities than most other classification techniques, with the notable exception of support-vector machines (Hastie et al. 2008, table 10.1), in which fitting can be excessively computationally expensive (Abdiansah and Wardoyo 2015). However, neural networks remain less interpretable than other machine learning algorithms, like tree-based methods (Benítez et al. 1997; Hastie et al. 2008, p. 352; Heinert 2008, table 10.1). Weights between the input and the first hidden layer can nevertheless provide some intuitive insights on the importance of each utilized metric. Given the problem at hand, neural networks are selected to assess the quality of ground motion records. This section briefly introduces neural networks, details the composition of the training and testing data sets, provides and justifies the preprocessing of the data, and describes the key elements of the neural network training.

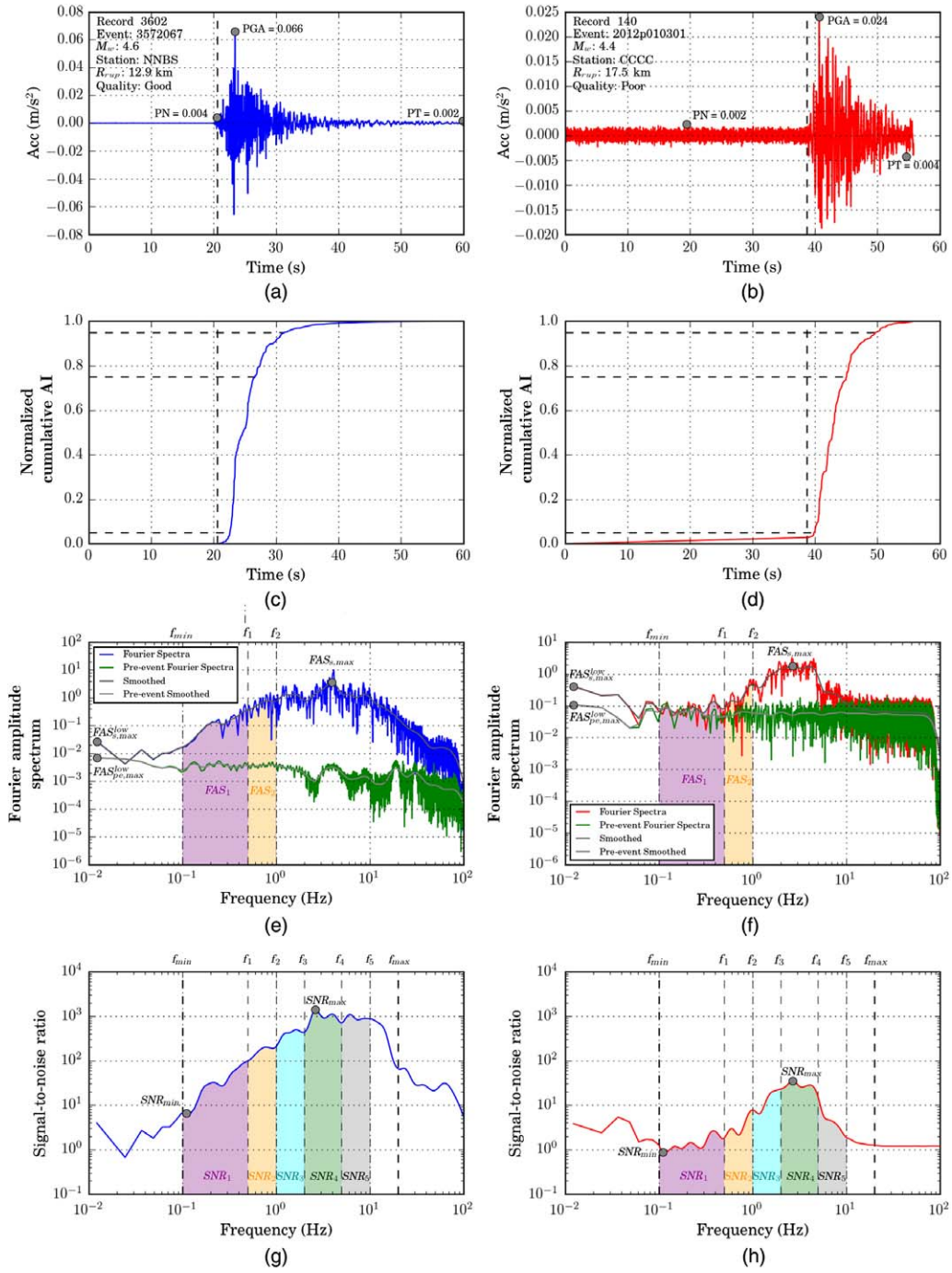

Figure 3. Comparison between a high (a, c, e, g) and low (b, d, f, h) quality ground motion record: (a, b) raw ground motion time series; (c, d) Husid plots; (e, f) FAS; and (g, h) SNR.

INTRODUCTION TO FEEDFORWARD NEURAL NETWORKS

A feedforward neural network is a supervised classification technique using repeated, activated, weighted linear combinations of a given input to predict its class (in the present case, low or high-quality ground motion records). A very simple feedforward neural network architecture is provided in Figure 4. More details about neural network architecture, training methodologies, and validation techniques than the herein introduction can be found in Haykin (1994), Hastie et al. (2008), and Goodfellow et al. (2016). In the machine learning context, supervised means that the algorithm learns to recognize, or fit, labeled training data (i.e., being provided the correct result of each data point). Feedforward denotes the type of neural network being used in this study. Feedforward neural networks do not perform any convolution on the input data, and do not possess any feedback loops as, for example, recurrent neural networks. A feedforward neural network is a stack of multiple layers composed of neurons. A neuron is a series of two operations. The first operation is the weighted sum of the neuron input and its internal bias, and the second one is the activation of this weighted sum by a nonlinear function, like the hyperbolic tangent, sigmoid, or rectifier functions. Figure 4b shows the anatomy of a neuron.

A typical feedforward neural network, like the one proposed in Figure 4a, is composed of three main parts: (1) an input layer, (2) N hidden layers, and (3) an output layer. The input layer represents the vector of input variables for the problem, in this case, the ground motion record quality metrics. Hidden layers create the nonlinear hyperplane that separates the data into different classes by recombining and activating the output from previous layers. The output layer delivers the model prediction, which is further used to estimate its loss (or residual).

A neural network is trained using stochastic gradient descent (SGD) coupled with the so-called back-propagation method, the former being responsible for randomly picking the training data and evaluating the learning rate–weighted loss gradient, and the latter for updating the model parameters following the chain rule of calculus. By repeating SGD and back propagation iteratively (each iteration referred to as an epoch), the neural network weights and biases are updated, and it learns to correctly classify the data.

In machine learning, it is usual to keep a portion of the data undisclosed to the algorithm during its training phase. This data are referred to as the test data set and is used to evaluate the

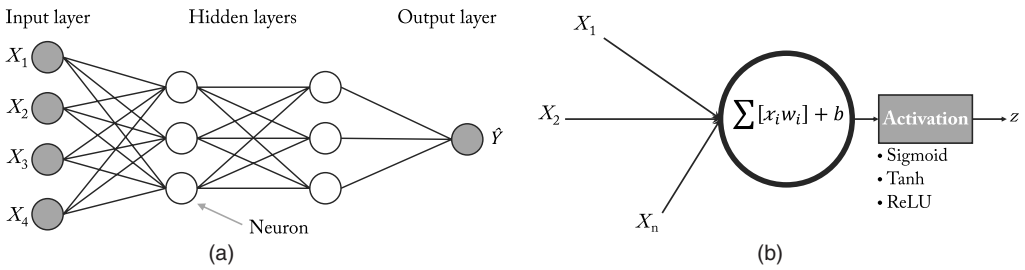

Figure 4. (a) Layout of a two-hidden-layer feedforward neural network and (b) anatomy of a single neuron.

predictive capabilities of the developed model. To avoid overfitting, a portion of the training data set, called the validation data set, is used to monitor the loss during the training phase.

GROUND MOTION RECORD TRAINING AND TESTING DATA SETS

Given the different origins of the considered data sets (Canterbury and Wellington regions), two different neural networks are trained against different data mixes. The first model is trained exclusively using the Canterbury data set and tested against the entire Wellington data set. The second model is trained against a mix of the Wellington and Canterbury data sets and tested against the remainder of the data. The partitioning of the data are summarized in Table 3. Note that as k-fold cross validation is used to select the best model architecture (detailed in the *Neural Network Architecture Selection and Training Method* section), the data in the training data set will also be used for validation. Separating and mixing geographically distinct data sets allows a direct comparison of the effect of ground motion record origin diversity within the training set.

HYPOTHESIS ON GROUND MOTION QUALITY AND DATA PREPROCESSING

To assess ground motion record quality, the assumption is made that ground motion records of high and low quality do not share the same quality metric distribution. The preprocessing of the data used on the neural network input data will show that this assumption cannot be rejected.

As noted by da Silva and Adeodato (2011), neural networks tend to exhibit better performance on decorrelated and amplitude-like data. Hence the data preprocessing consists of a deskewing operation based on a Tukey's ladder of powers (Helsel and Hirsch 2002) and a standardizing and decorrelating procedure using a ZCA whitening transform (Kessy et al. 2018). The transformed input variables are therefore decorrelated and follow the standard Gaussian distribution ($\mathcal{N}(0.0,1.0)$). Because of the mathematical transform, these new

Table 3. Composition of the training and test data sets used to train and test both the single region and the mixed regions neural networks. The composition of each subset is given by a quality score. As described in the *Hypothesis on Ground Motion Quality and Data Preprocessing* section, numbers in parentheses are not utilized

Training configuration	Quality score	Training set	Test set
A: Canterbury only	1.0	1,729	1,313
	0.75	(431)	916
	0.5	(11)	88
	0.25	(67)	365
	0.0	1,751	1,796
B: Canterbury and Wellington	1.0	1,525	1,517
	0.75	(691)	656
	0.5	(41)	58
	0.25	(208)	224
	0.0	1,796	1,751

variables lose their physical meaning (e.g., B' does not represent the transformed average tail ratio). Note that all the transformed variables from the ZCA are kept in order to preserve the full variability of the data set. Details on the original quality metric distributions (cumulative density functions in Figure C1 and correlation matrix in Figure C2) can be found in online Appendix C.

The preprocessing operations are only fitted on the ground motions of highest quality, such that the distribution of ground motion records of lower quality would be different if the initial assumption is proven to be true. Neural networks trained on data split A and B presented in Table 3 have their own preprocessing operation series (i.e., both preprocessings are fitted on their respective training data set). Results of the preprocessing are available in online Appendix D and show that all transformed variables follow the Gaussian distribution in Figures D1 and D2 for the Canterbury and Canterbury-Wellington preprocessings, respectively.

Figure 5a–f presents the joint plots of six pairs of transformed variables. It can be observed in Figure 5e that some important overlaps could exist between variables. Nevertheless, even such extreme cases of transformed variable pairs allow partial separation of low and high-quality ground motion records. In the other presented pairs, the separation of data is more evident. However, there are no observed cases in which the data are linearly separable (i.e., minor overlaps exist between all pairs). P-values developed from Kolmogorov–Smirnov (KS) tests

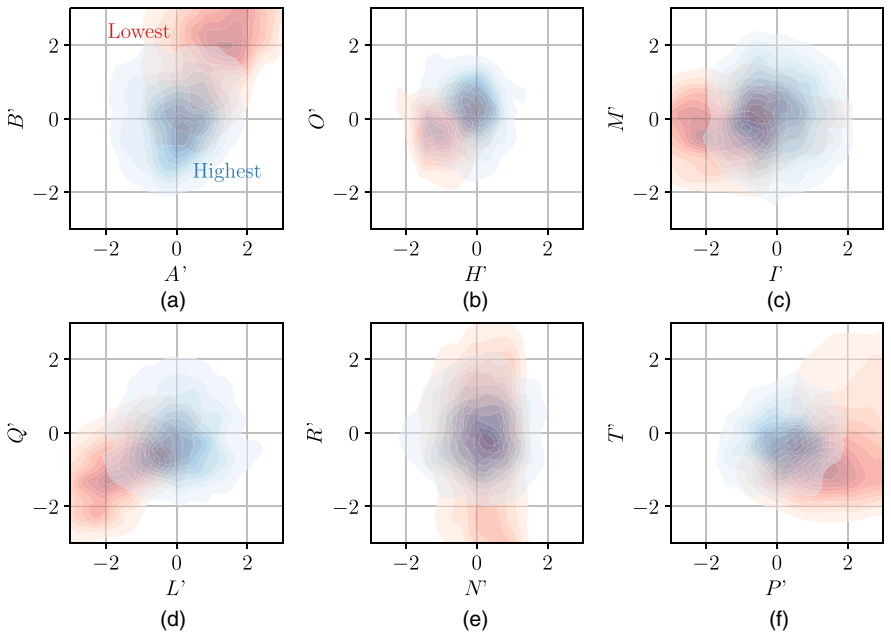

Figure 5. Joint plots of transformed variable pairs: (a) A'-B'; (b) H'-O'; (c) I'-M'; (d) L'-Q'; (e) N'-R'; and (f) P'-T'. Red and blue colours indicate lowest and highest quality ground motion records, respectively.

(Benjamin and Cornell 1970, pp. 466–471) indicate that the distributions of lowest and highest-quality ground motion records show appreciable differences. Details about the KS tests and their results are available in online Appendix D and Figure D3, respectively.

NEURAL NETWORK ARCHITECTURE SELECTION AND TRAINING METHOD

Bergstra and Bengio (2012) have shown that the performance of a neural network is linked to its architecture and parametrization (e.g., number of hidden layers, number of neurons, learning rate, etc.). While the exploration of the entire parameter search space is practically impossible (Bergstra and Bengio 2012, Bengio 2012), combinations of only a few parameters given in Table 4 are tested via grid search (Bengio 2012; Goodfellow et al. 2016, pp. 434–436). In this approach, the number of layers (1 or 2), number of neurons, and size of the mini-batches are tested. As oversized and overtrained neural networks tend to overfit (Baum and Haussler 1989, Geman et al. 1992, Lawrence et al. 1996), the size of the developed neural network remains small to retain good generalization capabilities, and the so-called early-stopping method is used to stop training if validation loss has not improved for ten epochs (Prechelt 1998). When mini-batches are used (mini-batch size $m > 1$), the back-propagation algorithm combines the results from m SGD before updating the weights and biases of the neural network. The Nadam optimization technique (Dozat 2016) is utilized to enhance the SGD algorithm. It automatically adapts the learning rate at each epoch during training based on the previously evaluated gradient loss. For this binary classification problem, the trained neural networks deliver two outputs: a score for its resemblance to the highest quality ground motion records and a score for its resemblance to the lowest quality ground motion records. The output activation function is the sigmoid function, and the loss is characterized by the binary cross entropy (Hastie et al. 2008; equation 2.36) presented in Equation 1:

$$L(\Theta) = \sum_{i=1}^{N} \log p_{g_i,\Theta}(x_i) \tag{1}$$

where $L(\Theta)$ represents the loss given the set of model parameters Θ (e.g., number of layers and neurons, values of each of the neuron weights and biases), N is the number of considered data points (in our case, the ground motion records), x_i corresponds to the model input of data point i, and $p_{g_i,\Theta}(x_i)$ is the probability of x_i belonging to class g_i given Θ and could be rewritten as $P(G = g_i(\Theta) \mid X = x_i)$. It must be noted that the loss does not represent the accuracy of the algorithm and is solely used to quantify the goodness of the model during training. The result of each combination is validated using a fivefold cross-validation scheme, as proposed

Table 4. Neural network parameters and their respective values used in the grid search

Varying neural network hyperparameter	Tested values
Number of neurons in the first hidden layer, n_1	10; 15; 20
Number of neurons in the second hidden layer, n_2	0, 10; 15; 20
Mini-batch size, m	4, 8, 16

by Hastie et al. (2008, pp. 241–249). This framework has been implemented in Keras (Chollet et al. 2015) with a TensorFlow back end (Abadi et al. 2016).

TRAINING RESULTS AND MODEL ASSESSMENT

The performance of the proposed model is critical for ground motion studies, such as validation of physics-based ground motion models. Hence in addition to the performance of the neural network, the distribution of the selected ground motion records, both from an intensity measure and a geospatial viewpoint, are also important. This section briefly summarizes the results of the fivefold cross validation, provides the training history of the retained models, discusses the predictive performance on the test data sets, and analyzes the potential introduction of biases from an intensity measure and a geospatial point of view for both neural network configurations.

CROSS-VALIDATION RESULTS AND TRAINING HISTORY

Table 5 presents the selected parameter combination as well as the mean and standard deviation of the fivefold cross-validation error for the two models trained on the different subsets of data presented in Table 3.

Final models are then trained against a newly drawn subset representing 80% of the training sets, which are validated against the remaining 20% to avoid overfitting, and their architectures are based on the aforementioned parameters. According to the observed validation errors, both models are expected to exhibit similar predictive performances on their respective test data sets. In online Appendix E, Figure E1 presents the training of both final models.

MODEL PERFORMANCE ASSESSMENT

The predictive performance of the neural networks is studied against their respective test data set. As the objective is to retain a relatively high number of high-quality ground motion records while discarding as many low-quality ground motion records as possible, several acceptance threshold values are tested on the high-quality ground motion record score: 0.5, 0.6, 0.7, 0.8, 0.9, and 0.95. The acceptance threshold value is used to determine whether an analyzed record is of low or high quality. For example, if the acceptance threshold is fixed to 0.7 and a record has a score of 0.82, it will be flagged as a high-quality record and will therefore be retained. Conversely, if the acceptance threshold is now fixed to 0.9, the same

Table 5. Selected neural network parameter values (see Table 4 for definitions) and their respective 5-fold validation mean and standard deviation loss

Training set	Parameter combination	Mean loss	Standard deviation loss
A: Canterbury	$n_1 = 15$ $n_2 = 20$ $m = 16$	0.0859	0.0108
B: Canterbury and Wellington	$n_1 = 15$ $n_2 = 15$ $m = 8$	0.0725	0.0081

record will be flagged as a low-quality record and will be discarded. Traditionally, performance of binary predictors is given in terms of accuracy and represented by receiver operation characteristics (ROC) curves (Fawcett 2006) and confusion matrices. However, in the present case, not only is the classifier's accuracy of interest but also its relative performance to each class (i.e., lowest, low, average, high, and highest quality). Therefore, the performance results are presented as bar plots in Figure 6 in terms of absolute number and portion of selected ground motion records for each class. Note that confusion matrices and ROC curves are available in Figures E2 and E3 of online Appendix E, respectively.

Results presented in Figure 6 show that, as expected, both models perform very similarly for acceptance thresholds up to 0.8. Above this particular value, the model trained exclusively on the Canterbury data tends to discard more high-quality ground motion records than the model trained on the Canterbury and Wellington data. However, the Canterbury-only trained model also seems to marginally reduce the proportion of average-to-lowest–quality ground motion records compared with the Canterbury-Wellington model. Note that the reduction of this value can be observed for both models with an increasing acceptance threshold.

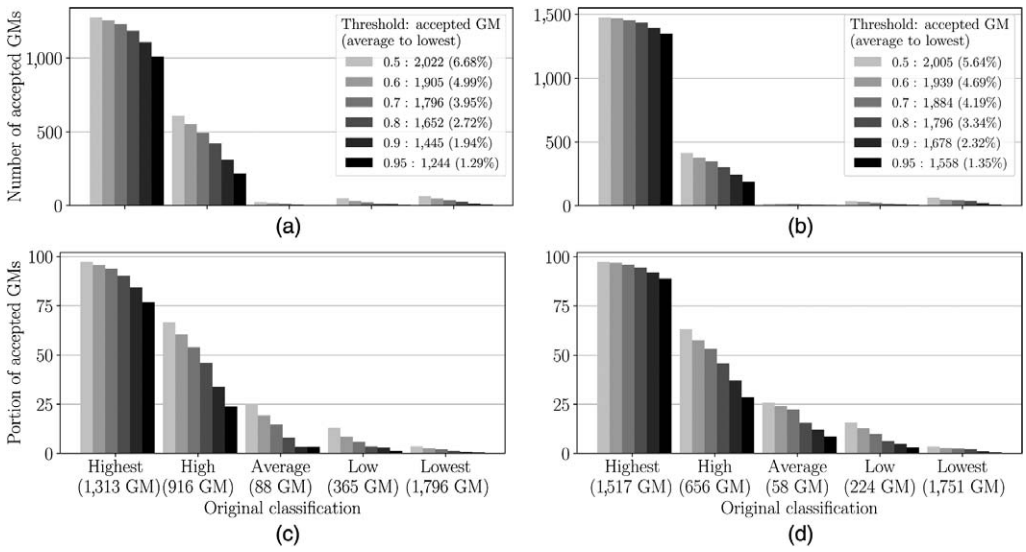

Figure 6. Effect on the number and proportion of selected ground motion records given their quality by applying different acceptance thresholds on the high quality predicting score for the (a,c) Canterbury-trained neural networks and (b,d) Canterbury-Wellington–trained neural networks. The number of ground motions in each category from both test data sets is given in brackets under the *x*-axis of (c) and (d) for the Canterbury trained and Canterbury-Wellington–trained models, respectively. For each tested threshold, the total number of selected ground motions and the proportion of average-to-lowest quality ground motion records (between brackets) is given in the legends of (a) and (b) for the Canterbury trained and Canterbury-Wellington–trained models, respectively.

Tables F1 and G1 in the online Appendices summarize results for manually removed records and records from two large earthquake magnitudes, respectively. In both cases, the proposed model tends to accurately predict the quality of these recorded ground motions. In addition, online Appendix G provides implementation strategies to apply the proposed model in a near-real-time fashion, giving workarounds to alleviate potential signal clipping-related problems and the possible absence of a pre-event noise window.

INTENSITY MEASURE COMPARISON BETWEEN MANUALLY AND AUTOMATICALLY SELECTED GROUND MOTION RECORDS

Figure 7 shows the distribution of the discarded high-quality records for PGA, peak ground velocity, spectral acceleration at 3.0 s (SA(3.0 s)), 5%–95% significant duration (D_{s595}), and AI. For conciseness, only results using an acceptance threshold of 0.5 are

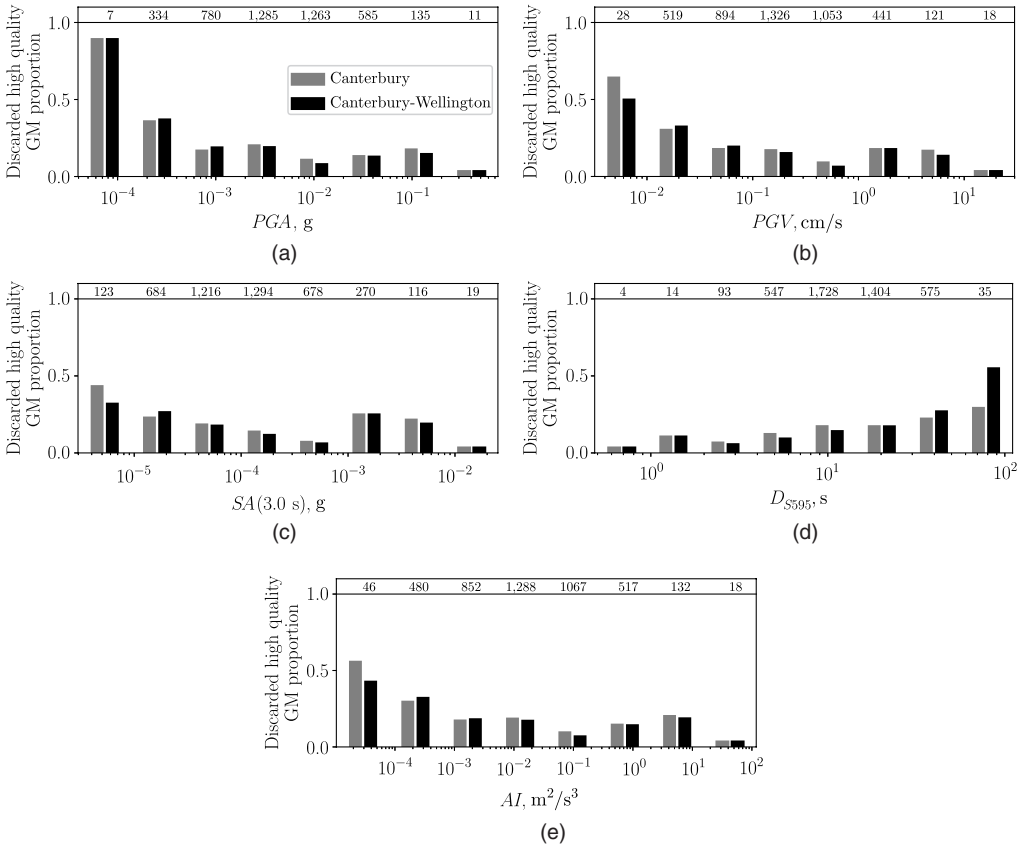

Figure 7. Proportion of discarded high-quality ground motions from the Canterbury trained (in gray) and Canterbury-Wellington–trained (in black) neural networks on the entire data set for (a) PGA, (b) peak ground velocity, (c) SA(3.0 s), (d) D_{S5-95}, and (e) AI using the 0.5 acceptance threshold. The number of manually selected ground motion records are given at the top of each bar pair.

shown here, but the same analysis is carried out for all other acceptance threshold values and presented in Figures H3–H7 of online Appendix H.

It can be observed that, with the exception of D_{s595}, the smaller the intensity measure value, the more likely the record would be discarded. This can be explained by the different SNR metrics used to predict the quality of the records: the smaller the signal, the lower the SNR. As ground motion duration is highly correlated with the distance to the source, the distribution of D_{s595} also implies that the farther away from the source the recording is occurring, the more likely it is to be rejected. However, there is no statistically significant case in which an entire subset is removed by the algorithm, meaning that the same range of intensities will remain present regardless of the selection method (manual or automated). To statistically show that the distributions from manual and automated selection are not significantly different, KS tests are performed and shown in Figure H8 of online Appendix H. Results indicate that for an acceptance threshold below 0.7, intensity measure distributions of the manually and automatically selected ground motion records are similar. This similarity tends to diminish with higher acceptance thresholds.

GEOSPATIAL ASSESSMENT OF AUTOMATIC GROUND MOTION RECORD SELECTION

By investigating the percentage of false predictions occurring at each recording station, it is possible to determine if there is any spatial bias being introduced by automatic screening of ground motions. In other words, there is a significant percentage of ground motions being wrongly discarded or included at a site as a result of false predictions. This exercise is carried out using the Canterbury and Wellington neural network with an acceptance threshold of 0.5 and is focused on the Canterbury and Wellington areas.

Figure 8a and 8b presents maps that detail the false negative prediction percentage (manually classified as high quality but predicted to be low quality) at recording stations in the Canterbury and Wellington areas, respectively. Triangles represent the recording stations and are sized based on how many recordings are observed at that station and colored based on what percentage of these recorded ground motions are false negatives. This means that lighter-shaded triangles imply good model performance, while darker-shaded triangles imply poor model performance. However, if the number of records is small, then the result is less substantial (i.e., when there is only one record at a station, 100% false prediction does not provide a strong statement of poor model performance). Therefore, the only real significant issue is where there is a darker-shaded large triangle. There are no stations that match these criteria but a few that are darker shaded and smaller (e.g., SMTC and ROLC in Canterbury and PIPS and INSS in Wellington). Upon inspection of the records at these stations, it was found that most records were manually scored 0.25–0.75 and were therefore already fringe cases. Additionally, the list of considered ground motion quality metrics do not completely represent the quality of the ground motion, which may contribute to the number of false predictions as well. In terms of applications, discarding high-quality ground motions is generally less detrimental than including low-quality ground motions for strong motion studies.

Figure I1a and I1b in online Appendix I presents maps that detail the false-positive prediction (manually classified as low quality but predicted to be high quality) percentage at recording stations in Canterbury and Wellington, respectively. In both Canterbury and

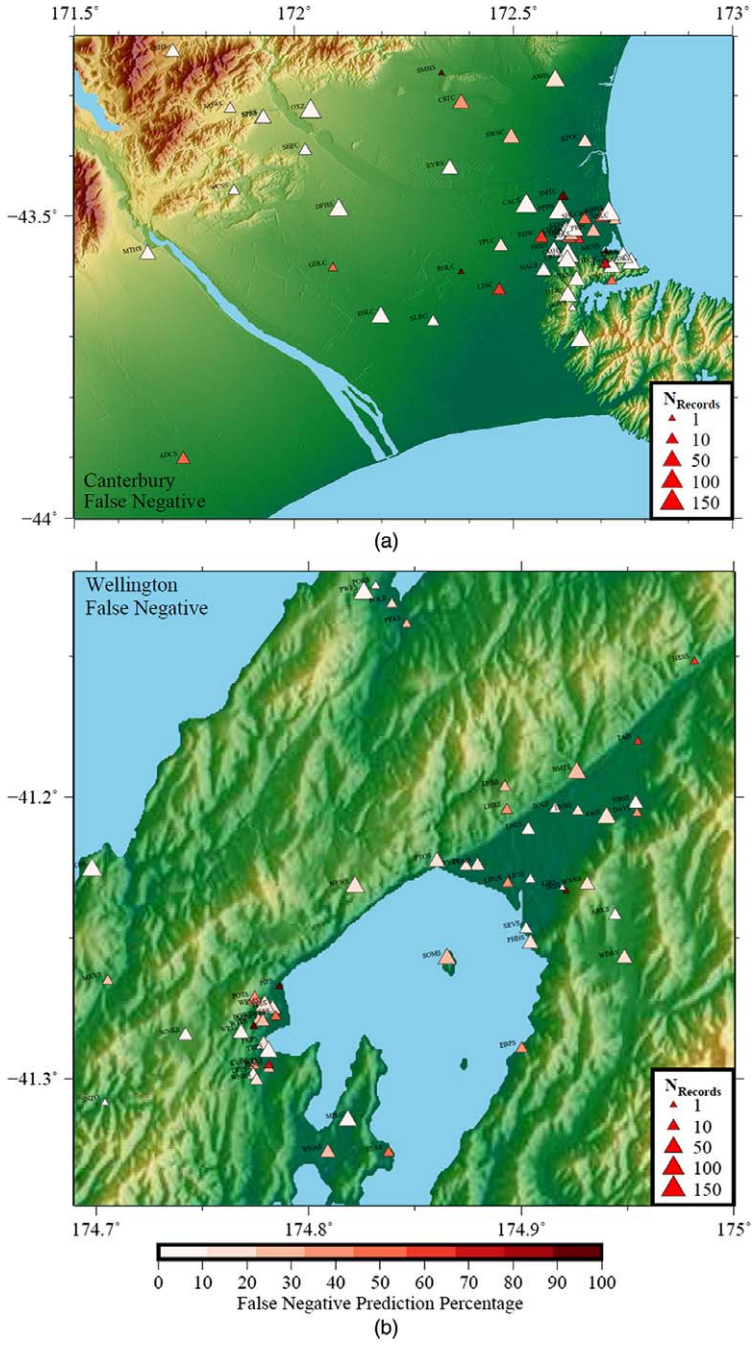

Figure 8. False-negative prediction (manual high quality, predicted low quality) percentage of ground motion recording stations in (a) Canterbury and (b) Wellington (locations shown in Figure 1).

Wellington, there appear to be fewer false-positive predictions than false-negative predictions. None of the stations have both a large quantity of records and high percentage of false-positive predictions. Compared with the false-negative predictions, there appears to be fewer false predictions, which is advantageous, as including low-quality ground motions in strong-motion studies can lead to incorrect inferences. Overall, there does not appear to be any significant geospatial bias being introduced with the Canterbury and Wellington screening model with a 0.5 acceptance threshold.

APPLICATION TO PHYSICS-BASED GROUND MOTION MODEL VALIDATION

To test the developed neural networks, they are used in an example of physics-based ground motion simulation validation. Ground motion simulation validation involves the comparison between simulated and observed ground motions, and hence quantification of the predictive capability of the simulation methodology. For this example, the adopted broadband simulation methodology is the Graves and Pitarka (2010, 2015) hybrid approach. Rather than carry out a comprehensive validation (which is beyond the scope of this paper), an assessment of the model prediction bias and total standard deviation and their sensitivity to the different neural networks and acceptance thresholds is the focus of the discussion that follows.

A previous study by Lee et al. (2019) carried out ground motion simulation validation in the Canterbury region using the subset of Canterbury ground motions in this study (148 earthquake sources with 3,989 ground motion records across 43 strong motion stations). The analysis compared a suite of ground motion intensity measures from simulated and observed ground motions and partitioned the residuals, using mixed-effects regression, into various components of ground motion variability to infer systematic biases inherent in the simulations. In terms of model prediction bias, the ground motion simulations were found to overpredict peak acceleration metrics (e.g., PGA and SA) but underpredict significant durations (e.g., D_{s575} and D_{s595}). This example follows the framework used by Lee et al. (2019) to identify the model bias and associated total standard deviation.

For this application, the 327 earthquake sources used in developing the Canterbury and Wellington neural network model are simulated with the 8,467 prospective ground motions across 195 stations. As ground motion simulations have finite spatial domains that are chosen to balance scope and computational demands, not all observed ground motions will have corresponding simulated ground motions produced in the simulations. It is important to note that while Lee et al. (2019) sought to validate the suitability of ground motion simulations in Canterbury for use in seismic hazard analysis and engineering applications, the purpose of this example is simply to test the sensitivity of the validation metrics to the various neural networks and acceptance thresholds. Therefore, the relative change in the results is of interest, as opposed to the absolute values of the model prediction biases and total standard deviations, which depend on a multitude of other factors.

The Canterbury only (Cant) and Canterbury and Wellington (CantWell) neural network models are considered with various acceptance thresholds: $Y = [0.5,0.6,0.7,0.8,0.9,0.95]$. A control case with no quality screening is also included along with a case that utilizes the manual screening scores. Figure 9 presents the model bias and total standard deviation

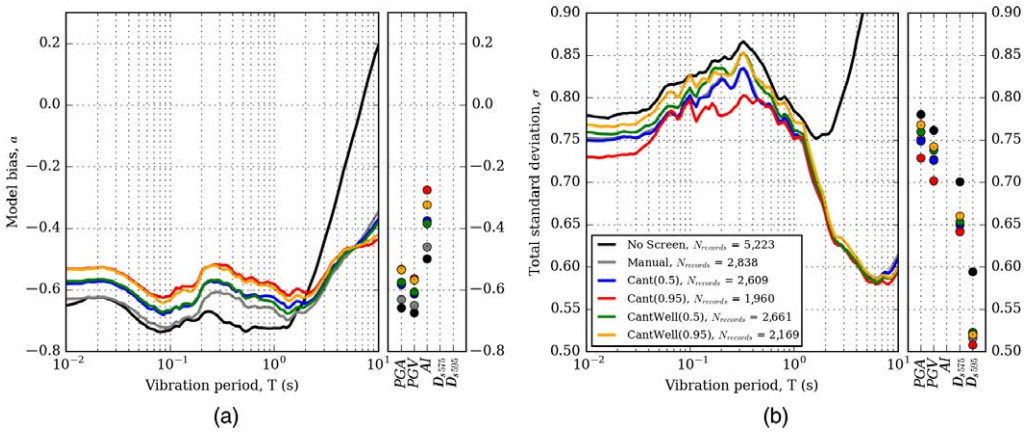

Figure 9. Model prediction (a) bias and (b) total standard deviation from ground motion simulation validation of 327 earthquakes for various intensity measures. Positive bias values indicate underprediction, and negative bias values indicate overprediction.

for the intensity measures considered as well as the number of high-quality ground motions included in the analysis for a few selected scenarios. The scenarios shown capture the extreme cases of screening with acceptance thresholds of $Y = 0.5$ and $Y = 0.95$ for each model. Table J1 in online Appendix J provides additional details regarding numerical values of bias and standard deviation for all cases considered. The bias values for duration are not shown in the plot, as they are too large, but are included in Figure J1. For both Canterbury and Canterbury-Wellington models, an increase in the acceptance threshold expectedly reduces the number of ground motions included in the analysis as more ground motions fail to meet the criteria. Comparing the two neural networks, the Canterbury and Wellington model includes marginally more ground motions than the Canterbury-only model for a given acceptance threshold.

Comparing bias and standard deviation values for the various screening scenarios, there appears to be only small differences between neural network models and acceptance thresholds. The largest difference in bias between screened scenarios for all intensity measures is on the order of 0.1 natural log units, which corresponds to roughly 10% difference. Comparing between the different scenarios, the acceptance threshold appears to be more influential than the choice between the two neural network models. When compared against the control case with no quality screening, the short period SA are similar but the long period SA (e.g., $T \geq 3.0$ s) bias becomes significantly more positive, suggesting that the neural networks remove these observed ground motions that have disproportionately large low-frequency acceleration amplitudes. Standard deviations are also mostly similar between all cases with the exception of the no screening case at long-period SA, which has significantly larger standard deviations, further reinforcing that the screening is removing low-quality ground motions. Overall, using either neural network with any of the considered thresholds appears to improve the observed ground motion data set's quality relative to no screening. As the manual screening also has error, deviations from this case should not be strictly considered as

being less correct. However, there is a trade-off of quantity of ground motions with increasing acceptance threshold. To retain a satisfactory number of observed ground motions, an acceptance threshold of 0.5 or 0.6 seems most appropriate.

CONCLUSIONS

This paper presented a machine learning–based tool that classifies ground motion recordings from small earthquakes based on quantifiable characteristics. The preprocessing of the available data has shown that high and low-quality ground motion records do not share the same quality metric distributions. Using the natural acceptance threshold on the Wellington-Canterbury model yields greater than 95% of correct classification for both lowest and highest quality ground motion records.

Despite the demonstrated accuracy of the fitted classifier, some limitations remain. First, it has been shown that increasing the acceptance threshold (i.e., the minimum score to classify ground motion records as high quality) increases the statistical differences between the manually and automatically selected distributions of intensity measures without removing all low-quality ground motion records. Hence trying to eliminate almost all low-quality records using the proposed method is detrimental, as it drastically reduces the number of available ground motion records and changes their associated intensity measure distributions. Second, the model has so far been tested in two regions only, namely, the Canterbury and Wellington, New Zealand, regions. Higher rejection and acceptance rates for high and low-quality ground motions, respectively, from other regions cannot be excluded. Third, further investigation is required to assess high rejection rates for high-quality ground motion records at particular stations. In some cases, peculiar geological and topographical effects may not be adequately captured by the developed algorithm.

However, as shown in the *Training Results and Model Assessment* and *Application to Physics-Based Ground Motion Model Validation* sections, the global effect of automatic quality screening on the ground motion simulation validation process remains beneficial. First, by using an appropriate acceptance threshold, it has been shown in the *Intensity Measure Comparison Between Manually and Automatically Selected Ground Motion Records* section that the intensity measure distributions from manually and automatically selected ground motion records are not statistically different. Second, the manual time required to scrutinize the observational data has now been reduced to a fraction of what would be required if a manual approach was undertaken.

Further applications may be imagined for the proposed algorithm. By comparing classified data sets from other locations with the results yielded by the algorithm, particular geological and topographical effects can be easily detected as outliers. In other words, if most of the ground motion records considered of high quality are rejected by the algorithm, some particular site effects may distort the signal. Once evaluated and, if needed, recalibrated to other regions, this method can also be used to monitor instruments' average recording quality, helping to maintain a high-quality sensor network.

In the future, more data from other regions should be utilized to further validate or retrain the developed classifier. Furthermore, alternatives to the proposed method may be explored to classify ground motion records. For example, a recurrent neural network, coupled with a

performant phase-picking algorithm, such as Ross et al. (2018), may be designed to assess the ground motion record quality as a full time series as opposed to utilizing summary quality metrics.

RESOURCES

The implementation of the entire framework has been made in Python 3.6 using ObsPy (Krischer et al. 2015) to process ground motions (particularly the ar_pick tool for picking P-wave arrivals), Keras (Chollet et al. 2015) to create the neural network architecture, and TensorFlow (Abadi et al. 2016) to train them. Observed ground motions were originally obtained from the GeoNet file transfer protocol (GeoNet 2018). The ground motion record data set and the final classifiers (Canterbury and Canterbury-Wellington trained models) are available in a GitHub repository (Bellagamba 2019).

ACKNOWLEDGMENTS

This project was supported by QuakeCoRE, a New Zealand Tertiary Education Commission-funded center, and the Royal Society of New Zealand's Rutherford Postdoctoral Fellowship. The authors would like to thank Eric Thompson and one anonymous reviewer for their constructive and thought-provoking comments, which have led to an improved paper, and editor David Wald. This is QuakeCoRE publication number 0368.

APPENDICES

Please refer to the online version of this manuscript to access the supplementary material provided in the Appendices.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X., 2016. Tensorflow: A system for large-scale machine learning, in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, USENIX Association, Berkeley, CA, 265–283.

Abdiansah, A., and Wardoyo, R., 2015. Time complexity analysis of support vector machines (SVM) in LibSVM, *International Journal Computer and Application* **128**, 28–34.

Ancheta, T. D., Darragh, R. B., Stewart, J. P., Seyhan, E., Silva, W. J., Chiou, B. S. J., Wooddell, K. E., Graves, R. W., Kottke, A. R., Boore, D. M., Kishida, T., and Donahue, J. L., 2013. *PEER NGA-West2 Database, PEER Rep. 2013/03*, Pacific Earthquake Engineering Research Center, Berkeley, CA.

Ancheta, T. D., Darragh, R. B., Stewart, J. P., Seyhan, E., Silva, W. J., Chiou, B. S. -J., Wooddell, K. E., Graves, R. W., Kottke, A. R., Boore, D. M., Kishida, T., and Donahue, J. L., 2014. NGA-West2 database, *Earthquake Spectra* **30**, 989–1005.

Anthony, R. E., Ringler, A. T., Wilson, D. C., and Wolin, E., 2018. Do low-cost seismographs perform well enough for your network? An overview of laboratory tests and field observations of the OSOP Raspberry Shake 4D, *Seismological Research Letters* **90**, 219–228.

Baum, E. B., and Haussler, D., 1989. What size net gives valid generalization? in *Advances in Neural Information Processing Systems* (D. S. Touretzky, ed.), MIT Press, Cambridge, MA, 81–90.

Begg, J. G., and Johnston, M. R., 2000. *Geology of the Wellington Area: Scale 1:250,000*, Institute of Geological and Nuclear Sciences, Lower Hutt, New Zealand.

Bellagamba, X., 2019. GroundMotionRecordClassifier, available at https://github.com/xavierbellagamba/GroundMotionRecordClassifier (last accessed 01 August 2019).

Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures, in *Neural Networks: Tricks of the Trade* (G. Montavon, G. B. Orr, and K. -R. Müller eds.), Springer, Berlin, Germany, 437–478.

Benítez, J. M., Castro, J. L., and Requena, I., 1997. Are artificial neural networks black boxes?, *IEEE Transactions on Neural Networks* **8**, 1156–1164.

Benjamin, J. R., and Cornell, C. A., 1970. Probability, Statistics, and Decision for Civil Engineers, McGraw-Hill, New York, NY, 466–471.

Bergstra, J., and Bengio, Y., 2012. Random search for hyper-parameter optimization, *Journal of Machine Learning Research* **13**, 281–305, available at http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

Boore, D. M., and Bommer, J. J., 2005. Processing of strong-motion accelerograms: Needs, options and consequences, *Soil Dynamics and Earthquake Engineering* **25**, 93–115.

Cauzzi, C., and Clinton, J., 2013. A high-and low-noise model for high-quality strong-motion accelerometer stations, *Earthquake Spectra* **29**, 85–102.

Cauzzi, C., Sleeman, R., Clinton, J., Ballesta, J. D., Galanis, O., and Kästli, P., 2016. Introducing the European rapid raw strong-motion database, *Seismological Research Letters* **87**, 977–986.

Chollet, F. et al., 2015. *Keras*, software, available at https://keras.io (last accessed 01 August 2019).

da Silva, I. B. V., and Adeodato, P. J. L., 2011. PCA and Gaussian noise in MLP neural network training improve generalization in problems with small and unbalanced data sets, in *Proceedings of the 11th International Joint Conference on Neural Networks*, Institute of Electrical and Electronics Engineers, San Jose, CA, 2664–2669.

Dawood, H. M., Rodriguez-Marek, A., Bayless, J., Goulet, C., and Thompson, E., 2016. A flatfile for the KiK-net database processed using an automated protocol, *Earthquake Spectra* **32**, 1281–1302.

Douglas, J., 2003. What is a poor quality strong-motion record?, *Bulletin of Earthquake Engineering* **1**, 141–156.

Douglas, J., and Boore, D. M., 2011. High-frequency filtering of strong-motion records, *Bulletin of Earthquake Engineering* **9**, 395–409.

Dozat, T., 2016. Incorporating nesterov momentum into Adam, in *Proceedings of the 4th International Conference on Learning Representations*, Workshop Track, San Juan, Puerto Rico.

Fawcett, T., 2006. An introduction to ROC analysis, *Pattern Recognition Letters* **27**, 861–874.

Forsyth, P. J., Barrell, D. J. A., and Jongens, R., 2008. *Geology of the Christchurch Area: Scale 1:250,000*, Institute of Geological and Nuclear Sciences, Lower Hutt, New Zealand.

Geman, S., Bienenstock, E., and Doursat, R., 1992. Neural networks and the bias/variance dilemma, *Neural Computation* **4**, 1–58.

GeoNet, 2018. Index of strong/processed/Raw, available at ftp://ftp.geonet.org.nz/strong/processed/Raw (last accessed 01 August 2019).

Goodfellow, I., Bengio, Y., and Courville, A., 2016. *Deep Learning*, MIT Press, Cambridge, MA, available at http://www.deeplearningbook.org (last accessed 01 August 2019).

Graves, R. W., and Pitarka, A., 2010. Broadband ground-motion simulation using a hybrid approach, *Bulletin of the Seismological Society of America* **100**, 2095–2123.

Graves, R., and Pitarka, A., 2015. Refinements to the Graves and Pitarka (2010) broadband ground-motion simulation method, *Seismological Research Letters* **86**, 75–80.

Hastie, T., Tibshirani, R., and Friedman, J., 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* – Second Edition, Vol. 1, Springer, Berlin, Germany, 739 pp.

Haykin, S., 1994. *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, NJ, 768 pp.

Heinert, M., 2008. Artificial neural networks–How to open the black boxes?, in *First International Workshop on Application of Artificial Intelligence in Engineering Geodesy (AIEG 2008)* (A. Reiterer and U. Egly, eds.), Artificial Intelligence in Engineering Geodesy, Vienna, Austria, 42–62.

Helsel, D. R., and Hirsch, R. M., 2002. *Statistical Methods in Water Resources*, U.S. Geological Survey, Reston, VA, 510 pp.

Hudson, D. E., 1979. *Reading and Interpreting Strong Motion Accelerograms*, Vol.1. Earthquake Engineering Research Institute, Berkeley, CA, 112 pp.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V., 2018. Machine learning for the geosciences: Challenges and opportunities, *IEEE Transactions on Knowledge and Data Engineering*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 1544–1554.

Kessy, A., Lewin, A., and Strimmer, K., 2018. Optimal whitening and decorrelation, *The American Statistician* **72**, 309–314.

Kishida, T., Di Giacinto, D., and Iaccarino, G., 2017. Comparison of manual and automated ground motion processing for small-to-moderate-magnitude earthquakes in Japan, *Earthquake Spectra* **33**, 875–894.

Kong, Q., Allen, R. M., Schreier, L., and Kwon, Y. -W., 2016. MyShake: A smartphone seismic network for earthquake early warning and beyond, *Science Advances* **2**, e1501055.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., and Gerstoft, P., 2018. Machine learning in seismology: Turning data into insights, *Seismological Research Letters* **90**, 3–14.

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., and Wassermann, J., 2015. ObsPy: A bridge for seismology into the scientific Python ecosystem, *Computational Science & Discovery* **8**, 014003.

Lawrence, S., Giles, C. L., and Tsoi, A. C., 1996. What Size Neural Network Gives Optimal Generalization? Convergence Properties of Backpropagation, *Tech. Rept.* UMIACS-TR-96-22 and CS-TR-3617, University of Maryland, College Park, MD.

Lee, R. L., Bradley, B. A., Stafford, P., Graves, R., and Rodriguez-Marek, A., 2019. Hybrid broadband ground motion simulation validation of Canterbury, New Zealand, *Earthquake Spectra*, in review.

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

Okada, H., and Suto, K., 2003. *The Microtremor Survey Method*, Society of Exploration Geophysicists, Tulsa, OK, 150 pp.

Patterson, N., Gledhill, K., and Chadwick, M., 2007. *New Zealand National Seismograph Network Report for the Federation of Digital Seismograph Networks Meeting*, *2007*, *Tech. Rept.*, GNS Science, Wellington, New Zealand.

Perol, T., Gharbi, M., and Denolle, M., 2018. Convolutional neural network for earthquake detection and location, *Science Advances* **4**, e1700578.

Prechelt, L., 1998. Early stopping – but when?, in *Neural Networks: Tricks of the Trade* (G. Montavon, G. B. Orr, and K. -R. Müller, eds.), Springer, Berlin, Germany, 55–69.

Racah, E., Beckham, C., Maharaj, T., Kahou, S. E., Prabhat, M., and Pal, C., 2017. ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events, in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), Curran Associates, Inc., Red Hook, NY, 3402–3413.

Rennolet, S. B., Moschetti, M. P., Thompson, E. M., and Yeck, W. L., 2018. A flatfile of ground motion intensity measurements from induced earthquakes in Oklahoma and Kansas, *Earthquake Spectra* **34**, 1–20.

Ross, Z. E., Meier, M. -A., Hauksson, E., and Heaton, T. H., 2018. Generalized seismic phase detection with deep learning, *Bulletin of the Seismological Society of America* **108**, 2894–2901.

Ross, Z. E., Yue, Y., Meier, M. -A., Hauksson, E., and Heaton, T. H., 2019. PhaseLink: A deep learning approach to seismic phase association, *Journal of Geophysical Research: Solid Earth* **124**, 856–869.

Titos, M., Bueno, A., García, L., and Benítez, C., 2018. A deep neural networks approach to automatic recognition systems for volcano-seismic events, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11**, 1533–1544.

Van Houte, C., Bannister, S., Holden, C., Bourguignon, S., and McVerry, G., 2016. The New Zealand strong motion database, *Bulletin of the New Zealand Society for Earthquake Engineering* **50**, 1–20.

(Received 21 December 2018; Accepted 19 April 2019)