IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI

S. MOSTAFA MOUSAVI[1], YIXIAO SHENG[1], WEIQIANG ZHU[1], and GREGORY C. BEROZA[1]
[1]Geophysics Department,Stanford University, 397 Panama Mall, Stanford, 94305-2215, CA, United States (e-mail: mmousavi@stanford.edu)

ABSTRACT Seismology is a data rich and data-driven science. Application of machine learning for gaining new insights from seismic data is a rapidly evolving sub-field of seismology. The availability of a large amount of seismic data and computational resources, together with the development of advanced techniques can foster more robust models and algorithms to process and analyze seismic signals. Known examples or labeled data sets, are the essential requisite for building supervised models. Seismology has labeled data, but the reliability of those labels is highly variable, and the lack of high-quality labeled data sets to serve as ground truth as well as the lack of standard benchmarks are obstacles to more rapid progress. In this paper we present a high-quality, large-scale, and global data set of local earthquake and non-earthquake signals recorded by seismic instruments. The data set in its current state contains two categories: (1) local earthquake waveforms (recorded at "local" distances within 350 km of earthquakes) and (2) seismic noise waveforms that are free of earthquake signals. Together these data comprise ~ 1.2 million time series or more than 19,000 hours of seismic signal recordings. Constructing such a large-scale database with reliable labels is a challenging task. Here, we present the properties of the data set, describe the data collection, quality control procedures, and processing steps we undertook to insure accurate labeling, and discuss potential applications. We hope that the scale and accuracy of STEAD presents new and unparalleled opportunities to researchers in the seismological community and beyond.

I. NOMENCLATURE

Benchmark, data set, earthquake, seismic signal, machine learning, AI

II. INTRODUCTION

Earthquakes are sudden movements across faults that release elastic energy stored in rocks and radiate seismic waves that travel throughout Earth. Every day there are about fifty earthquakes worldwide that are strong enough (magnitude > 2.5) to be felt locally, and every few days an earthquake occurs that is capable of damaging structures [1]. In addition, a multitude of smaller earthquakes (magnitude < 2.5) are happening (Fig. 1) that are too weak to be felt, but that are readily recorded by modern instruments. These small earthquakes provide valuable information about earthquake processes [2] .

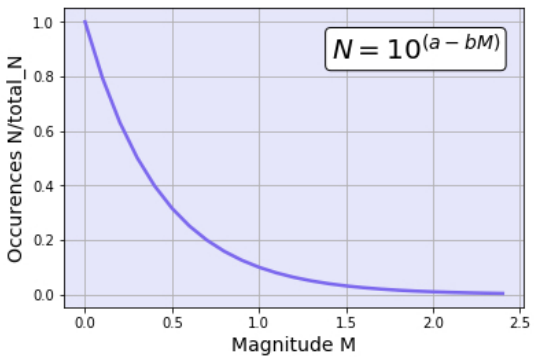


FIGURE 1. Gutenberg-Richter law [3] for b =1. N is the number of earthquakes having a magnitude M, a and b are constant. For the typical case of b=1, the number of earthquakes increases by a factor of 10 for each single unit decrease in M.

The seismic waves generated by earthquakes are recorded in the form of seismograms, which are records of ground motion at a particular place as a function of time. To characterize the vector components of ground motion, earthquakes are usually recorded by three-component instruments (seismographs) equipped with one vertical and two orthogonal horizontal sensors (Fig. 2). Several seismic wave

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

arrivals, called phases, are observable on seismograms. P and S phases are the two fundamental types of seismic phases observable on earthquake seismograms. In P or compressional waves, material moves back and forth in the direction in which the wave propagates, while in S or shear waves, material moves at right angles to the propagation direction. P waves travel faster than S waves, such that the first arriving pulse labeled "P" is a P wave that followed a direct path from the earthquake to the seismic station (Fig. 2). An earthquake begins to rupture at a hypocenter (or focus), which is defined by a position on the surface of the earth (epicenter) and a depth below this point . The hypocenter of an earthquake is found from the arrival times of seismic waves recorded on seismometers at different sites.

The size of an earthquake at its source is measured from the amplitude (or sometimes the duration) of the motion recorded on seismograms, and is expressed in terms of magnitude. Magnitude is a logarithmic measure. At the same distance from the earthquake, the amplitude of the seismic waves from which the magnitude is determined are 10 times as large during a magnitude 5 earthquake as during a magnitude 4 earthquake. The total amount of energy released by an average earthquake, depending on magnitude type, increases by a factor of approximately 32 for each unit increase in magnitude.

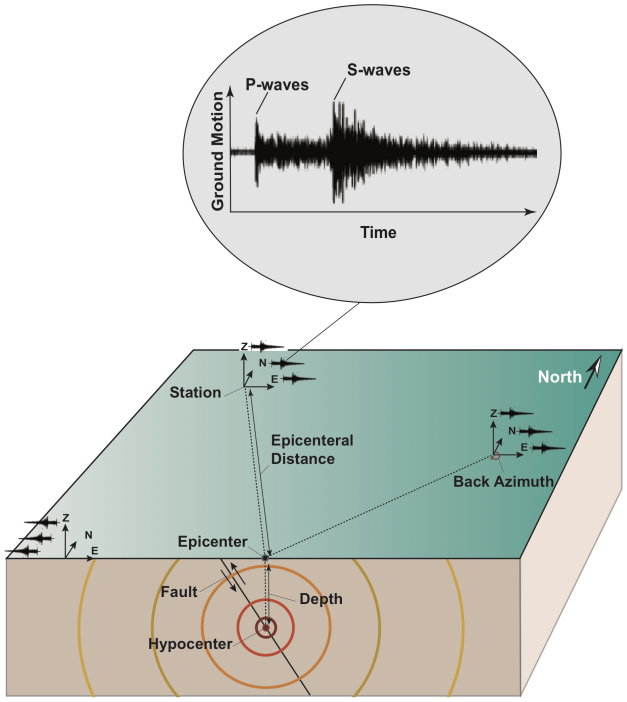


FIGURE 2. A schematic showing propagation of seismic waves and recording of the ground motion from them by seismic stations (receivers). E, N, and Z represent east, north, and vertical components of each instrument recording ground motions. An annotated earthquake waveform is presented in the zoomed window above.

Earthquakes are not the only sources that generate seismic waves. Many other sources such as explosions, landslides, oceanic waves, planes, helicopters, trains, wind, thunderstorms, traffic, and people, generate ground motions that are recorded by thousands of seismic instruments that are continuously operated by seismic monitoring networks around the world. Hence, there is an enormous amount of seismic data generated every day, and much of that ground motion is due to sources other than earthquakes, which we refer to as "non-earthquake" signals.

Seismology is a data-rich and data-driven science, and the rate of data acquisition is accelerating as seismic sensors get steadily less costly. The massive and rapidly growing amount of data highlights the need for more effective tools for the efficient processing and extraction of as much useful information as possible to enable scientist to realize the full potential to gain new insights into earthquake processes from them. Seismologists use only a portion of the recorded data to understand the physics of earthquakes and learn about Earth's deep interior, where direct observations are impossible. Most seismic data sets have not been fully analyzed and important discoveries can result from reanalysis of data sets using new data analysis tools.

Machine learning (ML) techniques have been shown to be powerful tools for processing (e.g. [4]–[6]) and exploring (e.g. [7], [8]) seismic data. The success of these ML-based methods in achieving state-of-the-art performance is mainly due to availability of large-scale and accurately labeled training data sets. Although, hundreds of terabytes of archived seismic waveform data and tens of millions of human picked parameters are available, a large and high-quality-labeled benchmark data set for seismic waveforms does not yet exist. This is attributable to several technical issues regarding reliable synchronization of metadata and waveform data and a lack of comprehensive and efficient quality control mechanisms.

Preparing a training set is one of the most time-consuming steps in making supervised models. Both the quantity and the quality of the training set are crucial to the performance of a model. Without a standard benchmark (e.g. ImageNet [9]), it is difficult to compare the performance of different approaches and to identify, adopt, and improve on best practices [10]. As an example, for the multiple deep-learning-based phase picking models that have been developed recently, each used a different data set for training and demonstration of its performance. In the absence of a standard benchmark, authors set their own criteria for evaluating performance. This inhibits progress because it makes it difficult to determine the relative performance, as well as the advantages and weaknesses, of each method.

Here we introduce STEAD, the first high-quality large-scale global data set of earthquake and non-earthquake signals recorded by seismic instruments. Benchmark data sets such as STEAD can accelerate progress in applying machine learning to problems in seismology. It facilitates training, validation, and performance comparisons, and the adoption of best practices. Moreover, this data set could have applications beyond seismology. The database is pub-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

licly available through https://github.com/smousavi05/STEAD. In the following sections, we first present the properties of the database. Then we discuss pre- and post-processing during the construction of the data set. In the last section we address some potential applications of the data set.

III. PROPERTIES OF THE DATA SET

STEAD includes two main classes of earthquake and non-earthquake signals recorded by seismic instruments. At this stage the earthquake class contains only one category of local-earthquakes with about 1,050,000 three-component seismograms (each 1 minute long) associated with ~ 450,000 earthquakes (Fig. 3) that occurred between January 1984 and August 2018. The earthquakes in the data set were recorded by 2,613 receivers (seismometers) (Fig. 4) worldwide located at local distances (within 350 km of the earthquakes). The non-earthquake class currently contains only one category of seismic noise including ~100,000 samples. Locations of instruments recording noise waveforms are presented in Fig. 5. Most of the seismograms have been recorded since 2000 (Fig. 6) in the United States and Europe where denser station coverage is available.

We provide seismic data as individual NumPy arrays containing three waveforms (each waveform has 6000 samples associated with 60 seconds of ground motion recorded in east-west, north-south, and vertical directions respectively). 35 attributes (labels) for each earthquake and 8 attributes for each noise seismogram are associated with each NumPy array. Noise attributes are mainly limited to the information about the recording instrument (e.g. network code, code, type, and location of the reciever) (Fig. 7). For the earthquake data (Fig. 8), in addition to the station information, we also provide information about the earthquake (e.g. origin time, epicentral location, depth, magnitude, magnitude type, focal mechanism, arrival times of P and S phases, estimated errors, etc), and recorded signal (e.g. measurement of the signal-to-noise ratio for each component, the end of signal's dominant energy (coda-end), and epicentral distance).

The unit of each attribute is included in the attribute's name. The epicenters of earthquakes (source_latitude and source_longitude) are given in units of latitude and longitude in the WGS84 reference frame. The depths (source_depth_km) where the earthquakes begin to rupture, are given in km. Based on the seismic network providing the metadata, this depth may be relative to the WGS84 geoid, mean sea-level, or the average elevation of the seismic stations that provided arrival-time data for the earthquake location.

Earthquake hypocenters and origin times (source_origin_time), when an earthquake began to rupture, have been estimated by seismic networks using earthquake location methods based on observed phase arrival times at multiple stations. The distances between earthquakes (source_distance_km and source_distance_deg) and the recording stations are calculated and provided in two formats of degree (the angle subtended at the center of the earth by the great circle arc between the two points) and kilometers. The distribution of the source_distance_km are given in Fig. 9. Most of the seismograms were recorded within 110 km of the earthquakes. Earthquakes are mainly shallower than 50 km (Fig. 10).

Magnitude is approximately related to the released seismic energy and provides an estimate of the relative size or strength of an earthquake. There are different methods (scales) for measuring the magnitude. The data set contains seismograms associated with a wide range of earthquake sizes from magnitude -0.5 to magnitude 7.9 (Fig. 11), but small earthquakes (magnitudes < 2.5) comprise the majority of the data set. Magnitudes have been reported in 23 different magnitude scales where local (ml) and duration (md) magnitudes are the majority(Fig. 12). This is because of the distance range of the data where these two magnitude scales are the most common scales. Unfortunately, the uncertainties for magnitude estimations have not been reported and only in ~ 24 % of the cases, the name of institute that calculated the magnitude (source_magnitude_author) were reported and have been provided.

source_id is a unique identification number provided by monitoring network that can be used to retrieve the waveforms and metadata (or additional information such as shake maps, etc) from established earthquake data centers.

More than 6200 waveforms contain information about the earthquake focal mechanisms (Fig. 13). These include one or two nodal plane solutions for events at different locations and with different mechanisms.

The category of each seismogram (trace_category) and its name (trace_name) are given in the attributes as well. The trace_name is a unique name containing station, network, recording time, and category code ("EV" for earthquake and "NO" for noise data).

The sample points where P and S phases arrive (p_arrival_sample and s_arrival_sample) are provided while status (p_status and s_status) shows how these arrival times have been determined. There are three types of arrival statuses in the data set (Fig. 14). "Manual" picks are arrival times that are hand-picked by human analysts, "automatic" picks are those measured by automatic algorithms by monitoring networks, and "autopicker" are arrival times determined using our AI-based model in this study. About 70 % of the picks are manually picked arrival times that we expect to have high accuracy. For the "autopicker" picks we use only arrival times with high confidence (high probabilities given by the deep-learning model [4]). As a measure of uncertainties in arrival time picks, a weight (a number between 0 and 1) is provided for most cases. Moreover, we have cross-checked the quality of the "manual" and "automatic" picks using the deep-learning method as discussed in the next section.

The back azimuth angle (back_azimuth_deg) is the direction that seismic waves arrive at the receiver. It is measured clockwise from the local direction of north at the receiver to the great circle arc connecting the receiver

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

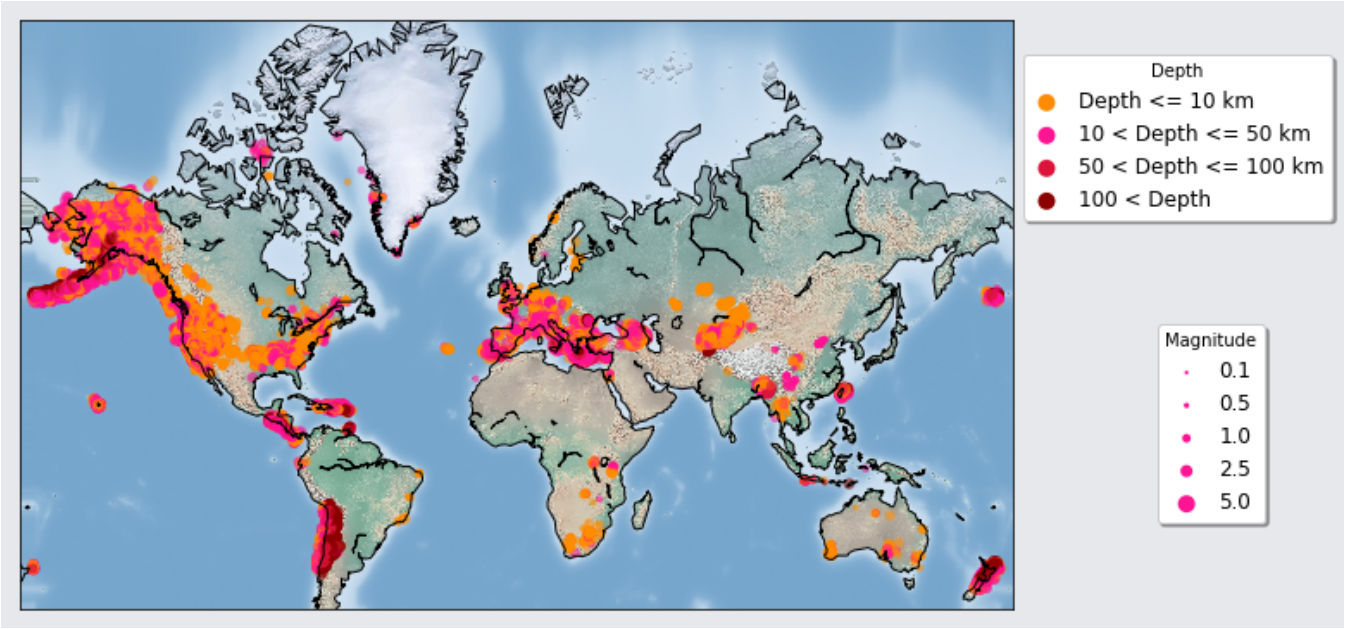


FIGURE 3. Location, size, and depth distributions of recorded earthquakes.

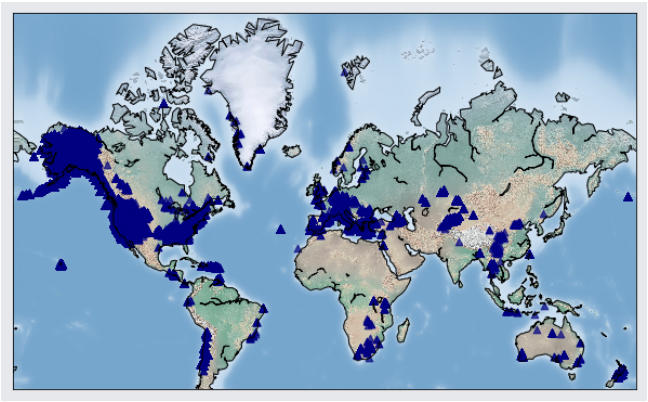


FIGURE 4. Locations of seismic instruments recording earthquakes shown by navy blue triangles.

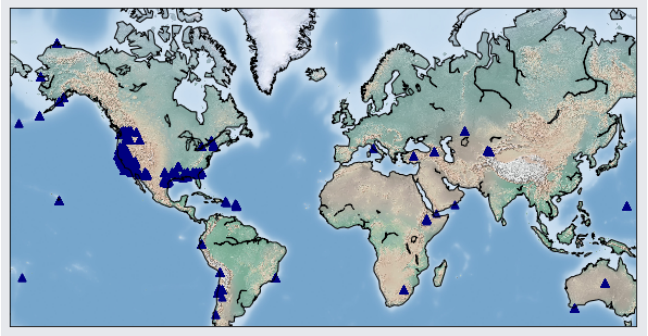


FIGURE 5. Distribution of stations recording seismic noise shown by navy blue triangles.

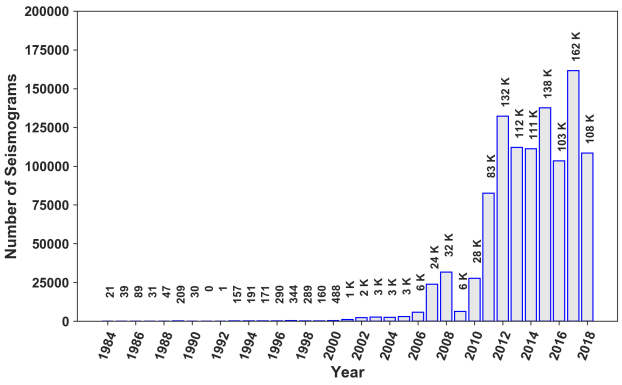


FIGURE 6. Number of earthquake seismograms as a function of time.

and epicenter. The data set contains earthquake signals arriving at receiver from all backazimuths (Fig. 15). P_travel times (p_travel_sec) are given in seconds and are calculated based on the arrival time of the P-wave at a receiver and the earthquake origin time. The coda_end_sample is the sample point where the dominance of scattered energy from an earthquake signal ends and the noise takes over. The network_code is the code for the seismic monitoring network to which the instrument belongs. This code can be use for retrieving either the waveform or metadata directly from the monitoring network. The instruments used for making the data set belong to 144 seismic networks operated at local, regional, and global scales by different national and international agencies. Here, we used data recorded by only 7 types of instruments. Of these, 99.5% are either high-gain broad band or extremely short period (Fig. 16). All seismograms (earthquake and non-earthquake) are three-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

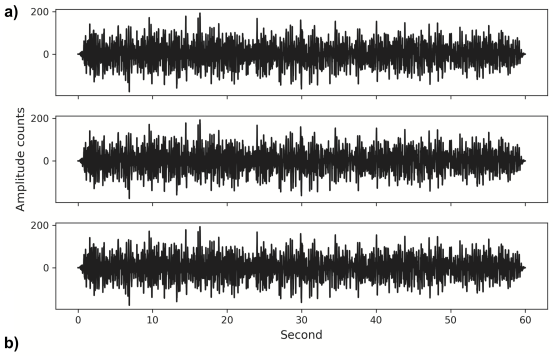


FIGURE 7. Example of noise data. a) time-series ground motions for east-west, north-south, and vertical directions respectively. b) header information (labels) associated with the seismogram.

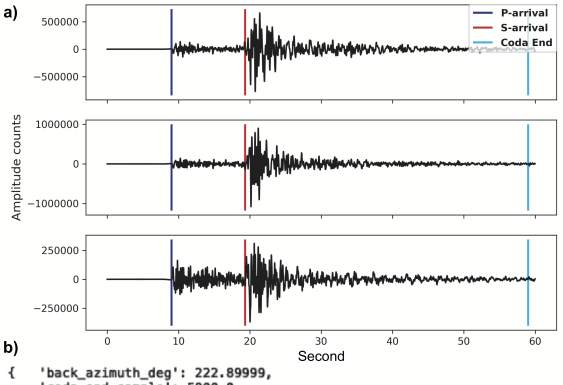


FIGURE 8. A sample earthquake seismogram. a) time-series ground motions for east-west, north-south, and vertical directions respectively. b) header information (labels) associated with the seismogram. The unit of each label is given in the label name.

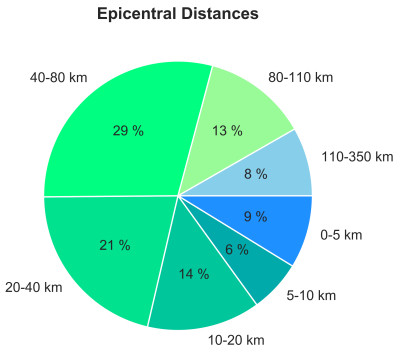


FIGURE 9. Distribution of epicentral distances for earthquake data.

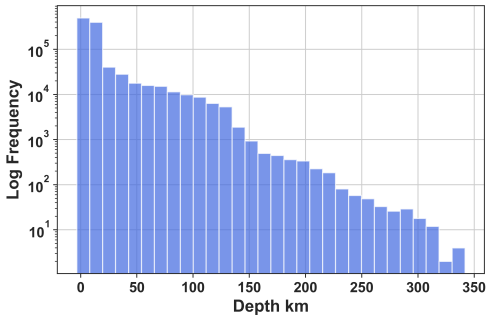


FIGURE 10. Distribution of earthquake depths.

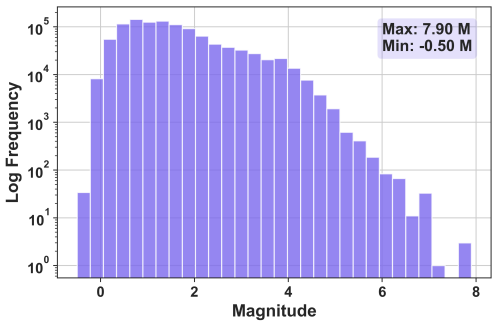


FIGURE 11. Distribution of earthquake magnitudes.

component, resampled to 100 HZ, and have the same 60 second (6000 samples) duration where the time of first sample is given by trace_start_time in UTC. trace_start_time is randomly selected to be between 5 and 10 seconds prior to the P-arrival time. For more details see the following section.

The focal mechanism refers to the direction of slip in an earthquake and the orientation of the fault on which it occurs. These focal mechanisms are computed using a method that attempts to find the best fit to the direction of P-wave first motions observed at each receiver. There is an ambiguity in distinguishing the fault plane, on which slip occurred, from the orthogonal, mathematically equivalent,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

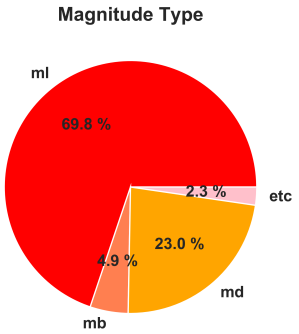


FIGURE 12. Distribution of magnitude scales for earthquake data. ml is the local magnitude, mb, body wave magnitude, and md is the duration magnitude. etc include mw, ms, mwr, mb_lg, mn, mpv, mlg, mwc, mc, mg, mh, mlr, mww, mpva, mbr, mblg, mwb, mlv, h, m, and mdl scales.

auxiliary plane. Hence, the parameters for two nodal planes are provided for those earthquakes that the focal mechanism solutions have been calculated and available through data centers. Each nodal plane is given by 3 values (strike, dip, and rake). Fault strike is the direction of a line created by the intersection of a fault plane and a horizontal surface, 0° to 360°, relative to North. Strike is always defined such that a fault dips to the right side of the trace when moving along the trace in the strike direction. Fault dip is the angle between the fault and a horizontal plane, 0° to 90°. Rake is the direction a hanging wall block moves during rupture, as measured on the plane of the fault. A rake of 90° means that the hanging wall moves up-dip (thrust), 0° means it moves in the strike direction (left-lateral), -90° means it moves in down-dip direction (normal), and 180° means it moves opposite to the strike direction (right-lateral).

IV. CONSTRUCTION OF STEAD

Metadata:

The metadata used in the construction of STEAD mainly consist of the information about the recording stations, recorded earthquakes, and hand-picked parameters, such as arrival times of P and S waves at each station. The metadata was acquired from multiple resources including: 1) the International Seismological Center [11], 2) the National Earthquake Information Center [12], 3) the Northern California Seismic Network [13], 4) the Southern California Seismic Network [14], 5) the Pacific Northwest Seismic Network [15], 6) the New Madrid Seismic Network [16], 7) the Incorporated Research Institutions for Seismology (IRIS) [17], 8) the Advanced National Seismic System Composite Catalog [18], 9) the Global Seismograph Network (GSN) [19] and 10) the broader literature (e.g. [20], [21]). In total, we processed more than 120 million data entries from these resources to extract and re-organize the metadata associated with local waveforms. For the lower magnitude ranges where fewer manual picks were available, we used theoretical arrival times. This information was combined with the earthquake and station information to build a comprehensive relational database. The final database includes more than 4 million phase arrival times of earthquake waveforms recorded by 3-component stations at local stations from around the world between January 1984 and August 2018.

Earthquake Waveforms:

We used the database of metadata to request the associated waveforms from continuous time-series archived at the IRIS data management center [22]–[24]. To ensure that each waveform only includes one earthquake signal (with known parameters) and to prevent inclusion of unknown (non-cataloged) earthquake signals, we used a short, fixed window (1 minute) around the phase arrival times at different stations to request data. Each window contains both P and S waves and begins from 5 to 10 seconds prior to the P arrival and ends at least 5 second after the S arrival. Only 1.5 million waveforms associated with the earthquakes in our database were available on the IRIS archive. We then detrended and removed the mean from all the waveforms, and resampled them at 100 Hz.

In the post-processing step, we checked the quality of existing labels using auxiliary algorithms, added new labels such as P-wave travel time, the end of earthquake signal (coda_end_sample) and computed a measure of the signal-to-noise ratio (snr). We estimated the end of earthquake signal based on the time series envelope, and measured the snr separately for each component as:

$$snr \;=\; 10\;\log_{10}\frac{\|S\|_2^2}{\|N\|_2^2}\;, \qquad (1)$$

where S and N are 95th percentile of amplitudes in a short window after S and prior to the P arrival time respectively. The distribution of the signal-to-noise ratio for earthquake seismograms is presented in Fig. 17. Most of the seismograms have snr between 10 and 40 decibels. The snr can be used to distinguish data with one or two faulty channels (where some of the components are mainly noise but earthquake signal can still be observed on a remaining component) or to select high-quality waveforms for tasks that are sensitive to the waveform quality.

Errors:

Four types of errors can be included in the waveform data. 1) earthquake characterization errors: these include errors in location, depth, origin time, and magnitude estimates of the earthquakes and can be due to errors in the arrival time picking, inaccurate velocity models, non-robust algorithms, number of recording stations etc. These errors can also affect the calculated epicenter distance, back azimuth, and P travel time. 2) errors in arrival time picks: these are either due to inaccurate theoretical arrival time estimates or human errors in the manual picks. 3) some time series do not contain the expected earthquake signals: this can be due to either inaccurate theoretical arrival time estimation during the preparation of the database or to timing errors between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

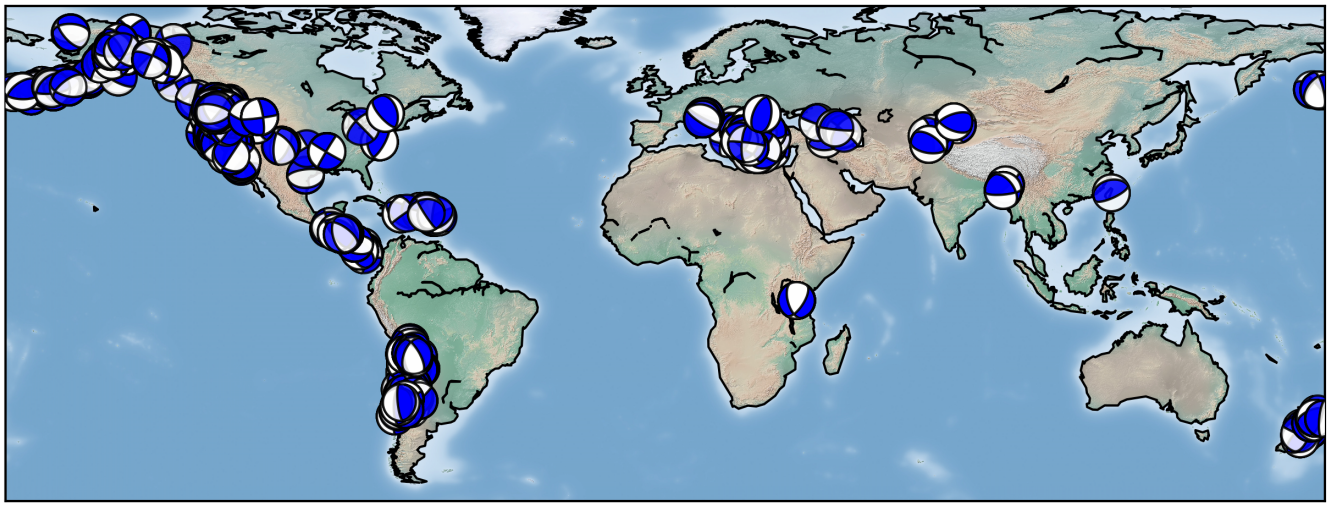


FIGURE 13. Geographical distribution of focal mechanisms shown by beach balls.

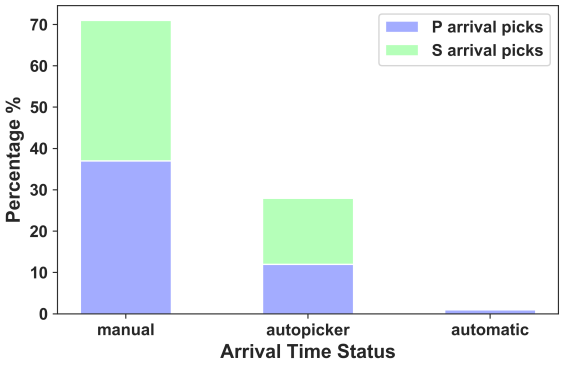


FIGURE 14. Proportions of the status of P-arrival and S-arrival picks. Manual picks are arrival times that were hand-picked by experienced human analysts. Automatic picks are those made by automatic algorithms reported by seismic networks, while autopicker are arrival times that we picked using our AI-based model.

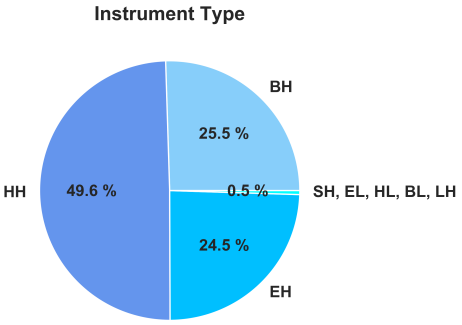


FIGURE 16. Types of seismic instruments used in building the data set. The first letter specifies the general sampling rate and the response band of the instrument where B are broad band, H represnts high broad band, E are extremely short period, and S are short period instruments. The second letter specifies the family to which the sensor belongs where H and L represent high gain and low gain seismometers respectively.

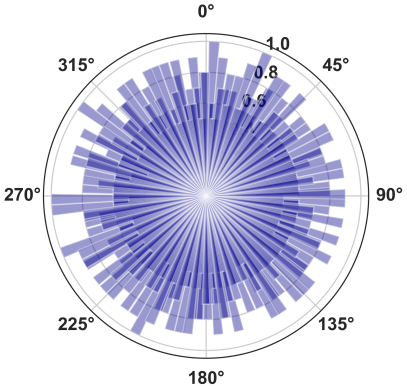


FIGURE 15. Distribution of the back-azimuths at which earthquake signals arrive from at seismic statins.

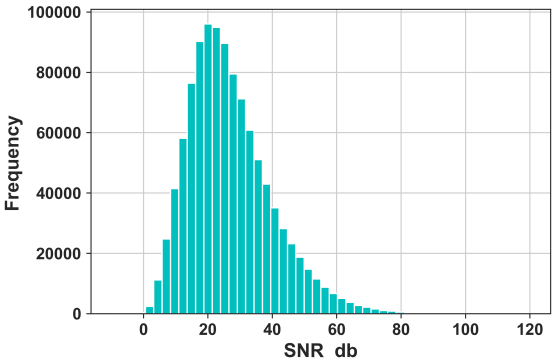


FIGURE 17. Distribution of signal-to-noise ratio (averaged over all components) for earthquake seismograms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

phase catalogs and archived data. 4) some time-series containing multiple uncatalogued earthquakes in addition to the expected earthquakes: this is due to either non-robustness or lack of sensitivity of current detection algorithms used by seismic networks, and leads to an incompleteness in current earthquake catalogs. From our point of view, this would lead to labeling errors to the data set by labeling the waveforms of uncatalogued earthquakes as noise or vice versa.

Unfortunately, the uncertainties in location, depth, and origin time estimates are not uniformly reported for all events by our resources and it is difficult to estimate them; however, we provide five parameters (source_gap_deg, source_error_sec, source_horizontal_uncertainty_km, source_origin source_depth_uncertainty_km)Fig. 18, for earthquakes for which this information were available. This can be used to assess the quality of reported parameters. source_gap_deg Fig. 18c, is the largest azimuthal gap between azimuthally adjacent stations (in degrees). In general, the smaller this number, the more reliable is the calculated horizontal position of the earthquake. Earthquake locations in which the azimuthal gap exceeds 180 degrees typically have large location and depth uncertainties. source_horizontal_uncertainty_kmFig. 18d, defined as the length of the largest projection of the three principal errors on a horizontal plane. The horizontal uncertainty varies from about 100 m horizontally for the best located events to 10s of kilometers for global events. source_depth_uncertainty_km, defined as the largest projection of the three principal errors on a vertical line. source_error_sec, is the RMS of the travel time residuals of the arrivals used for the origin computation.

The source depth is the least-constrained parameter in the earthquake location, and the error bars are generally larger than the variation due to different depth determination methods. Sometimes when depth is poorly constrained by available seismic data, the location program will set the depth at a fixed value. For example, 33 km is often used as a default depth for earthquakes determined to be shallow, but whose depth is not satisfactorily determined by the data, whereas default depths of 5 or 10 km are often used in mid-continental areas and on mid-ocean ridges since earthquakes in these areas are usually shallower than 33 km.

Estimated uncertainties for most of the arrival time picks are given in terms of weights. To replace the theoretical arrival times with more accurate picks and to double check the quality of manual and automatic picks, we used PhaseNet [4], a deep-leaning based phase picker. To identify traces with no earthquake or with more than one earthquake, we used CRED [6], a deep-learning-based model that detect earthquakes signals based on their time-frequency characteristics. With the help of these algorithms, we found during pos-tprocessing that many of the traces that should have lacked earthquake signals, contained uncatalogued-earthquake signals, or suffered from inaccurate arrival time picks. Examples of problematic data with incorrect labels identified by post-processing are shown in Fig. 19. This processing to remove problematic waveforms reduced the

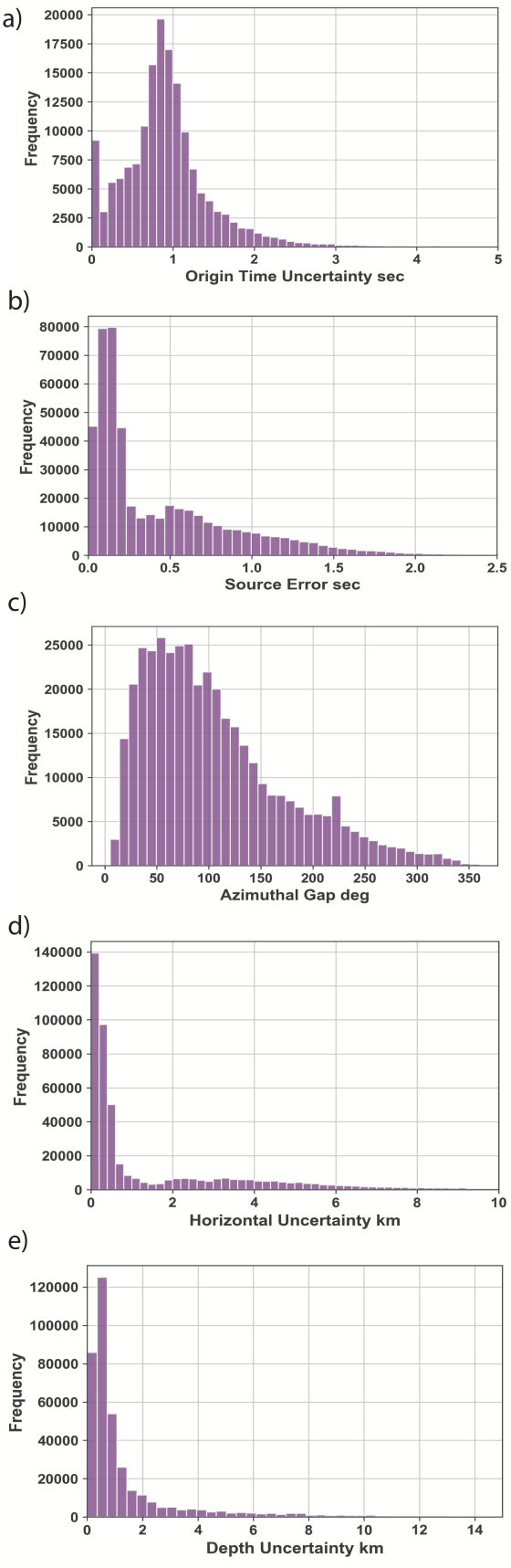


FIGURE 18. Uncertainties in the earthquake characterization.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

size of the original waveform data set by ~ 8 %. To estimate the remaining errors, we visually inspected 116,000 waveforms, randomly selected from the data set after the post-processing. Based on that sample, the remaining waveform data with error types of 2, 3, and 4 combined, make up less than 1% of the data set.

Noise Waveforms:

We randomly selected one-minute noise waveforms from the time periods between the cataloged earthquakes. After performing the same pre-processing (detrending, band-pass filtering, and resampling), we performed post-processing consisting of de-signaling followed by double checking using the generalized earthquake detector, CRED [6] to ensure that the noise traces do not contain earthquake signal (even hidden within the background noise). The de-signaling algorithm used here is a combination of the methods introduced in [25] and [26] that identifies the anomalous spectral features associated with earthquake signals (based on statistical considerations) in a continuous wavelet domain.

V. STEAD APPLICATIONS

Developing more robust models for processing seismic signals and characterizing earthquakes is a direct application of STEAD. Previous studies showed that deep-learning approaches can outperform traditional algorithms in these tasks. Existence of a large-scale data set with highly accurate labels like STEAD can facilitate development of more robust deep-learning models.

Denoising, detection, phase picking, and classification/discrimination are common processes performed on seismic signals. Denoising refers to suppressing the noise level and is traditionally done using simple band-pass filtering [27]. Earthquake signals generally have simpler waveforms compared to signals such as speech or audio; however, denoising of seismic signals can be more challenging due to the existence of strong coherent, non-stationary, and non-Gaussian noise [28]. Seismic denoising is particularly important because it can improve the snr and as a result improve subsequent processing such as detection [29] and phase picking. Examples of applications of machine learning methods for denoising seismic signals include both supervised [30] and unsupervised [31]–[34] methods. Recorded seismic noise and earthquake signals characterized by their snr and the beginning/end of the signals make the data set well-suited for building denoising models. Moreover, the data set can be used for developing decomposition models for separating overlapping signals (either two earthquakes, or earthquake and non-earthquake signals), which is another common and closely related problem in observational seismology.

Earthquake detection is one of the first data processing steps and remains a challenging problem in earthquake seismology. A good detection algorithm should: have few false positives (does not detect non-earthquake signals as earthquakes), few false negatives (does not miss small or weak earthquake signals), generalize well (is not limited to

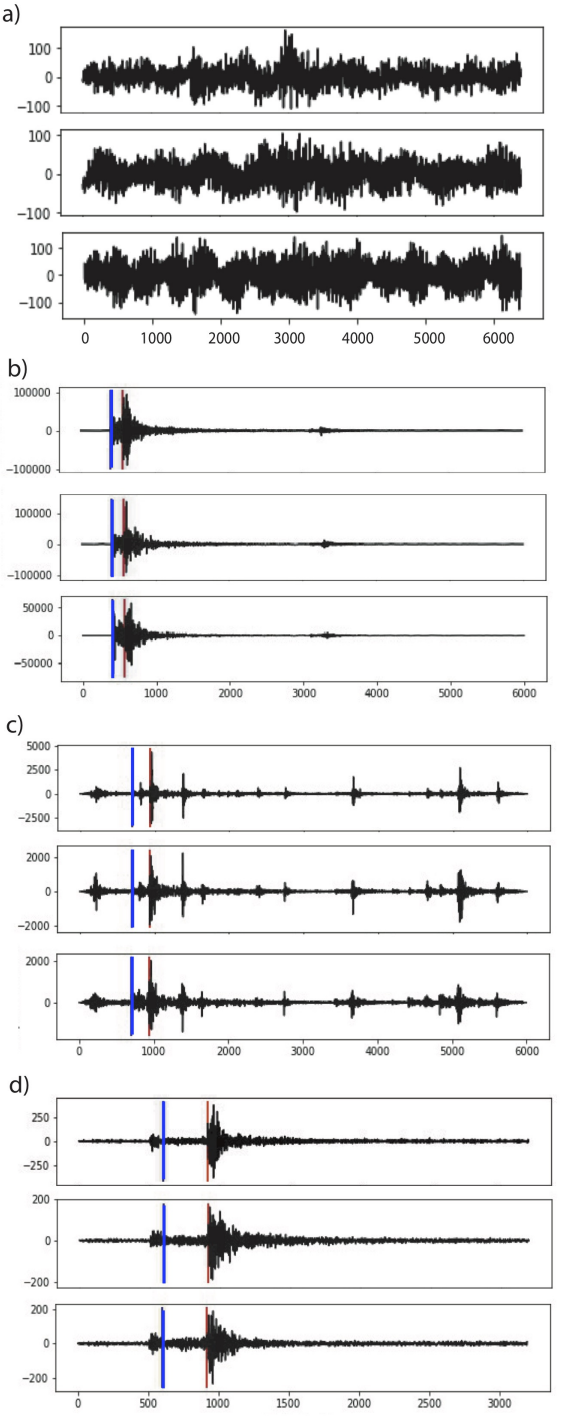


FIGURE 19. Examples of problematic seismograms detected by AI-based models during post-processing. a) is a seismogram that does not contain any earthquake signal. b) and c) are seismograms that in addition to the expected earthquake (with annotated picks) contain signals from uncataloged earthquakes. d) is an example of seismogram where the manual P-arrival pick is incorrect. P and S arrival times are marked by vertical blue and red lines respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2947848, IEEE Access

a specific shape, range, or setting of earthquakes), be insensitive to background noise, and be efficient for processing large data volumes. Characteristic-function-based (e.g. [35]) and similarity-search based (e.g. [36]–[38]) are the two main categories of algorithms commonly used for detection. In the characteristic-function based method a simple transformation is typically used to construct a function (e.g. STA/LTA) that highlights abrupt changes in the continuous data and makes it easier to distinguish earthquake signals. The advantages are that these methods are fast and generalize well-meaning that they can detect non-repeated earthquakes with non-similar waveforms. This generalization tends to also be the weakness of these methods because they inherently can not make a distinction between an earthquake signal and a non-earthquake pulse. Moreover, they are sensitive to background noise. On the other hand the similarity-search based methods look for repeated events with strictly similar waveforms. So they are more robust and generally result in much lower false positive rates; however, they are limited to repeated events and this can come with much higher computational cost. Neural networks have been used for earthquake signal detection (e.g. [39]–[46]). These methods can combine the advantages of characteristic-function and similarity-search based methods. In this approach a machine is trained to learn general characteristics of an earthquake signal by being exposed to many examples of earthquake and non-earthquake signals. Once the machine learns this general model, its application is fast since the detection is done in just one round. Previous studies showed that supervised learning can be a powerful tool for earthquake signal detection, however, there is still ample room for improvement and the development of more general and robust models. The global distribution of data, wide magnitude range, high accuracy of labels, and the end of earthquake signal as well as its beginning, positions STEAD to serve as an ideal data set for building more robust and comprehensive detection models.

Once an earthquake signal is detected, the arrival times of P and S waves need to be picked to locate the source. In addition to low false positive and false negative rates, pick accuracy is a crucial factor for obtaining reliable locations. Only 1 millisecond of error in determining P-wave arrivals can lead to ~ 7 m errors in estimated location [47]. While traditional algorithms for phase picking have a statistical basis [48]–[52], machine learning approaches use a variety of techniques (e.g. [4], [53]–[58]) to identify and pick different phases. The scale and reliability of picks in STEAD can foster building more accurate phase pickers. The random time lag between the beginning of each earthquake seismogram and first arrival reduces the data preparation process for this purpose.

Classification/discrimination of seismic signals is another problem in observational seismology where STEAD could be useful through transfer learning. Some examples include classification of volcanic signals (e.g. [59]–[62]), the discrimination of explosions from natural earthquakes (e.g. [63]–[67]), discrimination of quarry blasts from microearthquakes (e.g. [68], [69]), discrimination of seismic signals from earthquake and tectonic tremor (e.g. [70]), and discrimination of local from teleseismic earthquakes (e.g. [71]).

Direct earthquake characterization is yet another line of research where STEAD can be useful. Rapid estimation of the back-azimuth (e.g. [72]–[74]), magnitude, distance, and depth have applications for earthquake early warning systems. This is where the limited data used in previous efforts at applying machine learning techniques (e.g. [75], [76]) may have been problematic. A large, accurately labeled data set like STEAD could help overcome these limitations. Moreover, STEAD also has potential to be used to directly determine the earthquake locations using machine learning approaches (e.g. [77]–[79]), a challenging problem that has not yet been fully solved. This data set might be used for building ground-motion prediction models. These models are one of the most important elements used for seismic hazard assessments [80], [81]. Ground-motion prediction models are used to estimate the strong motion given a hypothetical earthquake source. Linear regression analysis is commonly used for developing ground-motion prediction equations [82], [83]. However, ML has shown to be a powerful tool for developing such models [84]–[87].

In addition to these, similarity of seismic signals to other time series data such as audio (see [88]–[91]) suggests a potential for using STEAD beyond seismological applications. Denoising, detection, and classification are common problems for audio and acoustic signals as well (e.g. [92]–[94]). Despite some differences, the existence of millions of human-picked labels, and extra information such as known locations of sources and receivers are unique characteristics of STEAD that do not exist in most equivalent audio data sets.

VI. CONCLUSION

Understanding the properties of earthquakes and subsurface processes they express must come through the analysis of recorded signals by near surface sensors. The complex, non-stationary nature of these signals requires powerful and sensitive processing tools to exploit them fully. Machine learning (ML) techniques are powerful tools that can learn the relationships and discover patterns directly from the data. The efficient extraction of as much useful information as possible from the recorded signals and the potential of gaining new insight is a challenge and the focus of an active field of research.

Here we introduce STEAD as the first high-quality large-scale global labeled data set of earthquake and non-earthquake signals recorded by seismic instruments. Benchmark data sets such as STEAD can accelerate progress in applying machine learning to problems in the seismology. It facilitates validation and comparison of competing methods, which promotes adoption of best practices, and accelerates research progress.

Future directions will concentrate on expanding the data set to regional (400 to 2000 km distance) and teleseismic (> 2000 km distance) earthquake seismograms, and include other non-earthquake categories such as seismic waves generated by explosions, volcanoes, landslides, oceanic waves, planes, helicopters, trains, wind, thunderstorms, and traffic.

We hope the high-precision monitoring techniques and models that will be developed with the help of this data set, can ultimately improve our understanding of earthquake processes by sharpening our ability to characterize seismicity.

ACKNOWLEDGMENTS

We thank Tim Ahern, Jerry Carter, and Chad Trabant from IRIS data services and Harley Benz from USGS for their help during compilation of the data set. The authors also thank William Ellsworth for his helpful suggestions. The facilities of IRIS-DS, and specifically the IRIS Data Management Center (http://ds.iris.edu/ds/nodes/dmc/, last accessed August 2018), were used for access to waveform data required in this study. The IRIS-DS is funded through the National Science Foundation (NSF) and specifically the GEO Directorate through the Instrumentation and Facilities Program of the NSF.

REFERENCES

[1] P. M. Shearer, Introduction to seismology. Cambridge University Press, 2009.

[2] E. E. Brodsky, "The importance of studying small earthquakes," Science, vol. 364, no. 6442, pp. 736–737, 2019.

[3] B. Gutenberg and C. F. Richter, "Magnitude and energy of earthquakes," Nature, vol. 176, no. 4486, p. 795, 1955.

[4] W. Zhu and G. C. Beroza, "PhaseNet: a deep-neural-network-based seismic arrival-time picking method," Geophysical Journal International, vol. 216, no. 1, pp. 261–273, 2018.

[5] Z. E. Ross, M.-A. Meier, E. Hauksson, and T. H. Heaton, "Generalized seismic phase detection with deep learning," Bulletin of the Seismological Society of America, vol. 108, no. 5A, pp. 2894–2901, 2018.

[6] S. M. Mousavi, W. Zhu, Y. Sheng, and G. C. Beroza, "CRED: A Deep Residual Network of Convolutional and Recurrent Units for Earthquake Signal Detection," arXiv preprint arXiv:1810.01965, 2018.

[7] B. K. Holtzman, A. Paté, J. Paisley, F. Waldhauser, and D. Repetto, "Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field," Science advances, vol. 4, no. 5, p. eaao2929, 2018.

[8] B. Rouet-Leduc, C. Hulbert, and P. A. Johnson, "Continuous chatter of the Cascadia subduction zone revealed by machine learning," Nature Geoscience, vol. 12, no. 1, p. 75, 2019.

[9] "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision (IJCV), vol. 115, no. 3, pp. 211–252, 2015. [Online]. Available: 10.1007/s11263-015-0816-y

[10] K. J. Bergen, P. A. Johnson, V. Maarten, and G. C. Beroza, "Machine learning for data-driven discovery in solid Earth geoscience," Science, vol. 363, no. 6433, p. eaau0323, 2019.

[11] 2016. [Online]. Available: http://www.isc.ac.uk

[12] "National Earthquake Information Center." [Online]. Available: https://earthquake.usgs.gov/contactus/golden/neic.php

[13] U. S. G. S. M. Park, "USGS Northern California Network. International Federation of Digital Seismograph Networks," Dataset/Seismic Network, vol. 10, no. 7914, 1967.

[14] "Southern California Seismic Network," California Institute of Technology and United States Geological Survey Pasadena, vol. 10, no. 7914, 1926.

[15] "Pacific Northwest Seismic Network. International Federation of Digital Seismograph Networks," Dataset/Seismic Network, vol. 10, no. 7914, 1963.

[16] C. Langston and H. DeShon, "Detection and location of non-volcanic tremor in the New Madrid Seismic Zone," International Federation of Digital Seismograph Networks. Dataset/Seismic Network, vol. 10, no. 7914, 2009.

[17] "Incorporated Research Institutions for Seismology." [Online]. Available: https://www.iris.edu/hq/

[18] "Advanced National Seismic System." [Online]. Available: https://www.sciencebase.gov/catalog/item/52eab950e4b0444d1ce67917

[19] "Global Seismograph Network (GSN -," IRIS/USGS). International Federation of Digital Seismograph Networks. Dataset/Seismic Network, vol. 10, no. 7914, 1988. [Online]. Available: https://doi.org/10.7914/SN/IU

[20] P. O. Ogwari, S. P. Horton, and S. Ausbrooks, "Characteristics of induced/triggered earthquakes during the startup phase of the Guy-Greenbrier earthquake sequence in North-Central Arkansas," Seismological Research Letters, vol. 87, no. 3, pp. 620–630, 2016.

[21] C. E. Yoon, Y. Huang, W. L. Ellsworth, and G. C. Beroza, "Seismicity during the initial stages of the Guy-Greenbrier, Arkansas, earthquake sequence," Journal of Geophysical Research: Solid Earth, vol. 122, no. 11, pp. 9253–9274, 2017.

[22] The facilities of IRIS Data Services, and specifically the IRIS Data Management Center, were used for access to waveforms, related metadata, and/or derived products used in this study. IRIS Data Services are funded through the Seismological Facilities for the Advancement of Geoscience and EarthScope.

[23] L. Krischer, T. Megies, R. Barsch, M. Beyreuther, T. Lecocq, C. Caudron, and J. Wassermann, "ObsPy: A bridge for seismology into the scientific Python ecosystem," Computational Science & Discovery, vol. 8, no. 1, p. 014003, 2015.

[24] M. Beyreuther, R. Barsch, and L. Krischer, "Tobias Megies, Yannik Behr, Joachim Wassermann; ObsPy: A Python Toolbox for Seismology," Seismological Research Letters, vol. 81, no. 3, pp. 530–533. [Online]. Available: https://doi.org/10.1785/gssrl.81.3.530

[25] S. M. Mousavi and C. A. Langston, "Automatic noise-removal/signal-removal based on general cross-validation thresholding in synchrosqueezed domain and its application on earthquake data," Geophysics, vol. 82, no. 4, pp. 211–227, 2017.

[26] ——, "Hybrid seismic denoising using higher-order statistics and improved wavelet block thresholding," Bulletin of the Seismological Society of America, vol. 106, no. 4, pp. 1380–1393, 2016.

[27] S. M. Mousavi, C. A. Langston, and S. P. Horton, "Automatic microseismic denoising and onset detection using the synchrosqueezed continuous wavelet transform," Geophysics, vol. 81, no. 4, pp. 341–355, 2016.

[28] S. M. Mousavi and C. A. Langston, "Adaptive noise estimation and suppression for improving microseismic event detection," Journal of Applied Geophysics, vol. 132, pp. 116–124, 2016.

[29] S. M. Mousavi and C. Langston, "Fast and novel microseismic detection using time-frequency analysis," in Society of Exploration Geopysics, and others, Ed., 2016, pp. 2632–2636.

[30] W. Zhu, S. M. Mousavi, and G. C. Beroza, "Seismic signal denoising and decomposition using deep neural networks," arXiv preprint arXiv:1811.02695, 2018.

[31] Y. Chen, M. Zhang, M. Bai, and W. Chen, "Improving the signal-to-noise ratio of seismological datasets by unsupervised machine learning," Seismological Research Letters, 2019.

[32] L. Zhu, E. Liu, and J. H. McClellan, "Seismic data denoising through multiscale and sparsity-promoting dictionary learning," Geophysics, vol. 80, no. 6, pp. 45–57, 2015.

[33] Y. Chen, J. Ma, and S. Fomel, "Double-sparsity dictionary for seismic noise attenuation," Geophysics, vol. 81, no. 2, pp. 103–116, 2016.

[34] C. Zhang, M. van der Baan, and T. Chen, "Unsupervised dictionary learning for signal-to-noise ratio enhancement of array data," Seismological Research Letters, vol. 90, no. 2A, pp. 573–580, 2018.

[35] R. V. Allen, "Automatic earthquake recognition and timing from single traces," Bulletin of the Seismological Society of America, vol. 68, no. 5, pp. 1521–1532, 1978.

[36] S. J. Gibbons and F. Ringdal, "The detection of low magnitude seismic events using array-based waveform correlation," Geophysical Journal International, vol. 165, no. 1, pp. 149–166, 2006.

[37] C. E. Yoon, O. O'Reilly, K. J. Bergen, and G. C. Beroza, "Earthquake detection through computationally efficient similarity search," Science advances, vol. 1, no. 11, p. e1501057, 2015.

[38] K. Rong, C. E. Yoon, K. J. Bergen, H. Elezabi, P. Bailis, P. Levis, and G. C. Beroza, "Locality-sensitive hashing for earthquake detection: a case study of scaling data-driven science." vol. 11, 2018, pp. 11–1674.

[39] J. Wang and T.-L. Teng, "Artificial neural network-based seismic detector," Bulletin of the Seismological Society of America, vol. 85, no. 1, pp. 308–319, 1995.

[40] Y. Zhao and K. Takano, "An artificial neural network approach for broadband seismic phase picking," Bulletin of the Seismological Society of America, vol. 89, no. 3, pp. 670–680, 1999.

[41] T. Perol, M. Gharbi, and M. Denolle, "Convolutional neural network for earthquake detection and location," Science Advances, vol. 4, no. 2, p. e1700578, 2018.

[42] Y. Wu, Y. Lin, Z. Zhou, D. C. Bolton, J. Liu, and P. Johnson, "DeepDetect: A cascaded region-based densely connected network for seismic event detection," IEEE Transactions on Geoscience and Remote Sensing, no. 99, pp. 1–14, 2018.

[43] S. M. Mousavi, W. Zhu, Y. Sheng, and G. C. Beroza, "Cred: A deep residual network of convolutional and recurrent units for earthquake signal detection," arXiv preprint arXiv:1810.01965, 2018.

[44] Z. E. Ross, M.-A. Meier, E. Hauksson, and T. H. Heaton, "Generalized seismic phase detection with deep learning," Bulletin of the Seismological Society of America, vol. 108, no. 5A, pp. 2894–2901, 2018.

[45] R. M. Dokht, H. Kao, R. Visser, and B. Smith, "Seismic Event and Phase Detection Using Time-Frequency Representation and Convolutional Neural Networks," Seismological Research Letters, vol. 90, no. 2A, pp. 481–490, 2019.

[46] Z. Zhou, Y. Lin, Z. Zhang, Y. Wu, and P. Johnson, "Earthquake Detection in 1D Time-Series Data with Feature Selection and Dictionary Learning," Seismological Research Letters, vol. 90, no. 2A, pp. 563–572, 2019.

[47] F. Waldhauser and W. L. Ellsworth, "A double-difference earthquake location algorithm: Method and application to the northern Hayward fault, California." Bulletin of the Seismological Society of America, vol. 90, no. 6 (2000, pp. 1353–1368.

[48] T. TAKANAMI and G. KITAGAWA, "A new efficient procedure for the estimation of onset times of seismic waves," Journal of Physics of the Earth, vol. 36, no. 6, pp. 267–290, 1988.

[49] A. Lomax, C. Satriano, and M. Vassallo, "Automatic picker developments and optimization: FilterPicker—A robust, broadband picker for real-time seismic monitoring and earthquake early warning," Seismological Research Letters, vol. 83, no. 3, pp. 531–540, 2012.

[50] E. Kalkan, "An automatic P-phase arrival-time picker." Bulletin of the Seismological Society of America, vol. 106, no. 3, pp. 971–986, 2016.

[51] Z. E. Ross and Y. Ben-Zion, "Automatic picking of direct P, S seismic phases and fault zone head waves," Geophysical Journal International, vol. 199, no. 1, pp. 368–381, 08 2014. [Online]. Available: 10.1093/gji/ggu267

[52] C. Chen and A. A. Holland, "PhasePApy: A robust pure Python package for automatic identification of seismic phases." Seismological Research Letters, vol. 87, no. 6, pp. 1384–1396, 2016.

[53] J. Woollam, A. Rietbrock, A. Bueno, and S. D. Angelis, "Convolutional neural network for seismic phase classification, performance demonstration over a local seismic network," Seismological Research Letters, pp. 16–90, 1 2019.

[54] Y. Chen, "Automatic microseismic event picking via unsupervised machine learning," Geophysical Journal International, vol. 212, no. 1, pp. 88–102, 09 2017. [Online]. Available: 10.1093/gji/ggx420

[55] Z. E. Ross, M. A. Meier, and E. Hauksson, "P wave arrival picking and first-motion polarity determination with deep learning," Journal of Geophysical Research: Solid Earth, vol. 2018, pp. 123–6.

[56] J. Zheng, J. Lu, S. Peng, and T. Jiang, "An automatic microseismic or acoustic emission arrival identification scheme with deep recurrent neural networks." Geophysical Journal International, vol. 212, no. 2 (2017, pp. 1389–1397.

[57] J. Wang, Z. Xiao, C. Liu, D. Zhao, and Z. Yao, "Deep-Learning for Picking Seismic Arrival Times," Journal of Geophysical Research: Solid Earth, 2019.

[58] E. Pardo, C. Garfias, and N. Malpica, "Seismic Phase Picking Using Convolutional Networks," IEEE Transactions on Geoscience and Remote Sensing, 2019.

[59] S. Scarpetta, F. Giudicepietro, E. C. Ezin, E. P. S. Petrosino, M. Martini, and M. Marinaro, "Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks." Bulletin of the Seismological Society of America 95, vol. 1, pp. 185–196, 2005.

[60] C. Hammer, M. Ohrnberger, and D. Fäh, "Classifying seismic waveforms from scratch: a case study in the alpine environment," Geophysical Journal International, vol. 192, no. 1, pp. 425–439, 11 2012. [Online]. Available: 10.1093/gji/ggs036

[61] D. Agliz and A. Atmani, "Seismic signal classification using multi-layer perceptron neural network." International Journal of Computer Applications, vol. 79, p. 15, 2013.

[62] M. Titos, A. Bueno, L. García, and C. Benítez, "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, vol. 11, no. 5 (2018, pp. 1533–1544.

[63] F. U. Dowla, S. R. Taylor, and R. W. Anderson, "Seismic discrimination with artificial neural networks: preliminary results with regional spectral data." Bulletin of the Seismological Society of America, vol. 80, no. 5, pp. 1346–1373, 1990.

[64] Y. Shimshoni and N. Intrator, Classification of seismic signals by integrating ensembles of neural networks., 1998, vol. 5.

[65] Y. V. Fedorenko, E. S. Husebye, and B. O. Ruud, Explosion site recognition; neural net discriminator using single three-component stations., 1999, vol. no.

[66] M. Tarvainen, ""Recognizing explosion sites with a self-organizing network for unsupervised learning." Physics of the earth and planetary interiors," pp. 1–4, 1999.

[67] P. S. Dysart and J. J. Pulli, "Regional seismic event classification at the NORESS array: seismological measurements and the use of trained neural networks." Bulletin of the Seismological Society of America, vol. 80, no. 6, pp. 1910–1933, 1990.

[68] M. Musil and A. Plešinger, "Discrimination between local microearthquakes and quarry blasts by multi-layer perceptrons and Kohonen maps." Bulletin of the Seismological Society of America, vol. 86, no. 4 (1996, pp. 1077–1090.

[69] A. Ursino, H. Langer, L. Scarfì, G. D. Grazia, and S. Gresta, "Discrimination of quarry blasts from tectonic microearthquakes in the Hyblean Plateau (Southeastern Sicily)." Annals of Geophysics, vol. 44, no. 4, 2001.

[70] M. Nakano, D. Sugiyama, T. Hori, T. Kuwatani, and S. Tsuboi, "Discrimination of seismic signals from earthquakes and tectonic tremor by applying a convolutional neural network to running spectral images." Seismological Research Letters, vol. 90, no. 2, pp. 530–538, 2019.

[71] S. M. Mousavi, W. Zhu, W. Ellsworth, and G. Beroza, "Unsupervised Clustering of Seismic Signals Using Deep Convolutional Autoencoders," IEEE Geoscience and Remote Sensing Letters, 2019. [Online]. Available: 10.1109/LGRS.2019.2909218

[72] A. S. Eisermann, A. Ziv, and G. H. Wust-Bloch, "Real-time back azimuth for earthquake early warning." Bulletin of the Seismological Society of America, vol. 105, no. 4 (2015, pp. 2274–2285.

[73] J. Hu, H. Zhang, and H. Yu, "Accurate determination of P-wave backazimuth and slowness parameters by sparsity-constrained seismic array analysis," Geophysical Journal International, vol. 216, no. 1, pp. 1–18, 09 2018. [Online]. Available: 10.1093/gji/ggy390

[74] S. Noda, S. Yamamoto, S. Sato, N. Iwata, M. Korenaga, and K. Ashiya, ""Improvement of back-azimuth estimation in real-time by using a single station record." Earth, planets and space 64," pp. 305–308, 2012.

[75] A. Lomax, A. Michelini, and D. Jozinović, "An investigation of rapid earthquake characterization using single-station waveforms and a convolutional neural network." Seismological Research Letters, vol. 90, no. 2, pp. 517–529, 2019.

[76] S. M. Mousavi, C. A. Langston, S. P. Horton, and B. Samei, "Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression," Geophysical Journal International, vol. 207, no. 1, pp. 29–46, 2016. [Online]. Available: 10.1093/gji/ggw258

[77] M. Kriegerowski, G. M. Petersen, H. Vasyura-Bathke, and M. Ohrnberger, "A deep convolutional neural network for localization of clustered earthquakes based on multistation full waveforms." Seismological Research Letters, vol. 90, no. 2, pp. 510–516, 2018.

[78] X. Zhang, J. Zhang, C. Yuan, S. Liu, Z. Chen, and and, "Weiping Li. "Locating earthquakes with a network of seismic stations via a deep learning method." arXiv preprint arXiv:1808.09603," 2018.

[79] D. T. Trugman and P. M. Shearer, "GrowClust: A hierarchical clustering algorithm for relative earthquake relocation, with application to the Spanish Springs and Sheldon, Nevada, earthquake sequences," Seismological Research Letters, vol. 88, no. 2A, pp. 379–391, 2017.

[80] S. M. Mousavi, G. C. Beroza, and S. M. Hoover, "Variabilities in probabilistic seismic hazard maps for natural and induced seismicity in the central and eastern United States," The Leading Edge, vol. 37, no. 2, pp. 141–1. [Online]. Available: https://doi.org/10.1190/tle37020141a1.1

[81] S. M. Mousavi and G. C. Beroza, "Evaluating the 2016 One-Year Seismic Hazard Model for the Central and Eastern United States Using Instrumental Ground-Motion Data," Seismological Research Letters, vol. 89, no. 3, pp. 1185–1196. [Online]. Available: https://doi.org/10.1785/0220170226

[82] Y. Bozorgnia, N. A. Abrahamson, L. A. Atik, T. D. Ancheta, G. M. Atkinson, J. W. Baker, A. Baltay, D. M. Boore, K. W. Campbell, B. S.-J. Chiou et al., "NGA-West2 research project," Earthquake Spectra, vol. 30, no. 3, pp. 973–987, 2014.

[83] B. Derras, P. Y. Bard, and F. Cotton, "Towards fully data driven ground-motion prediction models for Europe," Bulletin of earthquake engineering, vol. 12, no. 1, pp. 495–516, 2014.

[84] D. T. Trugman and P. M. Shearer, "Strong correlation between stress drop and peak ground acceleration for recent M 1-4 earthquakes in the San Francisco Bay area," Bulletin of the Seismological Society of America, vol. 108, no. 2, pp. 929–945, 2018.

[85] B. Derras, P.-Y. Bard, F. Cotton, and A. Bekkouche, "Adapting the neural network approach to PGA prediction: An example based on the KiK-net data," Bulletin of the Seismological Society of America, vol. 102, no. 4, pp. 1446–1461, 2012.

[86] A. Alimoradi and J. L. Beck, "Machine-learning methods for earthquake ground motion analysis and simulation," Journal of Engineering Mechanics, vol. 141, no. 4, p. 04014147, 2014.

[87] A. H. Alavi and A. H. Gandomi, "Prediction of principal ground-motion parameters using a hybrid method coupling artificial neural networks and simulated annealing," Computers & Structures, vol. 89, no. 23-24, pp. 2176–2194, 2011.

[88] "Listening to Earthquakes." [Online]. Available: https://earthquake.usgs.gov/learn/topics/listen/

[89] "Listening to Earthquakes – from Inside the Earth." [Online]. Available: https://blogs.ei.columbia.edu/2016/09/23/listening-to-earthquakes-from-inside-the-earth/

[90] "seismicsoundlab." [Online]. Available: https://vimeo.com/seismicsoundlab

[91] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep Learning for Audio Signal Processing," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206–219, 2019.

[92] "DCASE2017 Challenge." [Online]. Available: http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/

[93] "VoxCeleb." [Online]. Available: http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

[94] "audioset." [Online]. Available: https://research.google.com/audioset/

S. MOSTAFA MOUSAVI received the M.S. degree in Risk Engineering from University of Tehran, Tehran, Iran in 2010 and the M.S. and Ph.D. degrees in geophysics from University of Memphis, TN, USA in 2017.

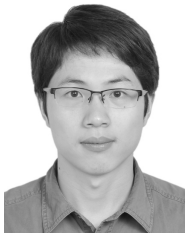
He is currently a Postdoctoral Research Fellow at Stanford University, CA, USA. He is the author of more than 20 journal, and 3 conference papers. His research interests include machine learning, signal processing, statistics, and observational seismology.

Dr. Mousavi is a fellow of National Elite Foundation of Iran and a recipient of Jeannne X. Kasperson award by the American Association of Geographers and SEP/SEG award by ExxonMobil in 2014.

YIXIAO SHENG received his B.S. degree in geophysics from the University of Science and Technology of China, Hefei, Anhui Province, in 2014. He is currently pursuing the PhD degree in Geophysics at Stanford University, CA, USA.

He has been a research assistant at Stanford since 2014. His research interest focuses on extracting useful information from passive seismic signals, including both ambient seismic noise and earthquake ground motion. One of his recent projects involves using graph analytics and geostatistical tools to understand the spatial variability of earthquake ground motion.

Mr. Sheng is a student member of American Geophysical Union, Seismological Society of America, and Southern California Earthquake Center since 2014.



WEIQIANG ZHU received the B.S. (2013) and the M.S. (2016) in Geophysics from Peking University, Beijing, China. He is currently pursuing the Ph.D. in Geophysics at Stanford University, CA, USA.

Since 2016, he is a Research Assistant at Geophysics department, Stanford University. His current research is the application of deep learning in seismology and physics-based earthquake simulation. He is interested in developing new methods based on deep neural networks to solve the challenging problems in seismology, e.g. earthquake monitoring, early warning, ground motion prediction, etc.

Mr. Zhu is a member of American Geophysical Union (AGU), Seismological Society of America (SSA) and Southern California Earthquake Center (SCEC) since 2016.

GREGORY C. BEROZA received his BS degree in Earth Science from UC Santa Cruz in 1983, and his Ph.D. from MIT in geophysics from MIT in 1989.

After a year as a Postdoc at MIT, he joined the Stanford Geophysics Department in 1990 where he now holds the Wayne Loel Professorship. He is the author of over 170 publications. His research interest is in analyzing seismograms to understand how earthquakes work and to quantify the hazards they pose. He is particularly interested in earthquake source processes for shallow earthquakes, intermediate-depth earthquakes, induced earthquakes, and slow earthquakes.

Prof. Beroza received a Presidential Young Investigator award from NSF in 1991, is a fellow of the American Geophysical Union since 2008, has been named to the Brinson, RIT, IRIS/SSA, and Lawson lecturerships, and received the Beno Gutenberg medal from the European Geosciences Union in 2014 for outstanding contributions to seismology. Since 2007 he has been Deputy-Director, then Co-Director of the Southern California Earthquake Center.