REPORT
doi:10.26443/seismica.v3i1.1049

Curated Regional Earthquake Waveforms (CREW) Dataset

Albert L. Aguilar Suarez * [1], **Gregory C. Beroza** [1]

[1]Department of Geophysics, Stanford University, Stanford, CA, USA

Author contributions: *Conceptualization*: A.L. Aguilar S., G.C. Beroza. *Software*: A.L. Aguilar S.. *Writing - original draft*: A.L. Aguilar S., G.C. Beroza. *Writing - Review & Editing*: A.L. Aguilar S., G.C. Beroza. *Funding acquisition*: G.C. Beroza.

Abstract We have assembled CREW, the Curated Regional Earthquake Waveforms Dataset, which is a dataset of earthquake arrivals recorded at local and regional distances. CREW was assembled from millions of waveforms with quality control through semi-supervised learning. CREW includes 1.6 million waveforms that have global coverage. Each waveform consists of a 5 minute three component seismogram with labels for both a P and S arrival. CREW provides a high quality labeled waveform data set that can be used to develop and test machine learning models for the analysis of earthquakes recorded at regional distances.

Production Editor:
Carmine Galasso
Handling Editor:
Jack Muir
Copy & Layout Editor:
Théa Ragon

Received:
June 21, 2023
Accepted:
April 20, 2024
Published:
May 24, 2024

1 Introduction

The Deep Learning seismological data landscape is dominated by local recordings. STEAD (Mousavi et al., 2019) contains over 1.2 million three component earthquake waveforms recorded at distances up to 350 km, with 8 percent of the data recorded at more than 110 km. STEAD provides 60 s waveforms from around the world that include both P and S arrival labels. INSTANCE (Michelini et al., 2021) contains over 1.1 million three component earthquake waveforms recorded at distances up to 600 km. INSTANCE provides 120 s waveforms from Italy and its surroundings with at least a P or S arrival. LENDB (Magrini et al., 2020) contains over 600,000 three component earthquake waveforms recorded at distances up to 134 km. LENDB provides 27 s waveforms from around the world with picked P arrivals. The Pacific Northwest AI-ready Seismic Dataset (Ni et al., 2023) contains 190,000 three component waveforms for earthquakes and exotic events. This dataset provides 150 s waveforms. These four datasets also contain noise waveforms. The NEIC dataset (Yeck et al., 2020) contains over 1.3 million earthquake waveforms recorded at distances up to 90 degrees. This dataset provides 60 seconds long waveforms around the phases P, Pn, Pg, Sn, Sg and S, with the majority corresponding to P phases. The MLAAPDE dataset (Cole et al., 2023) contains 5.1 million three component waveforms for earthquakes recorded at distances ranging from local to teleseismic. This dataset provides 120 s waveforms. The GEOFON dataset (Woollam et al., 2022) also covers the local to teleseismic distance range, with nearly 275K labeled arrivals, mostly P waves.

Most seismological deep learning research on earthquake detection and phase picking has used short duration waveforms from small earthquakes at short distances. PhaseNet (Zhu and Beroza, 2019) was trained on 30 second waveforms to predict the timing of P and S wave arrivals in Northern California. Earthquake Transformer (Mousavi et al., 2020) was trained on 60 s waveforms to simultaneously detect earthquakes and pick the arrival times of P and S waves. (Woollam et al., 2019) used 6 s windows for phase picking and (Ross et al., 2018) employed 4 s windows to predict the type of dominant energy in the seismogram (P or S), training on seismograms recorded within 100 km from the epicenter.

Most of the world is sparsely instrumented and many earthquakes are recorded only at distances over 100 km. This is true for the important case of seismicity near most subduction trenches, which are often more than 100 km from the nearest land. At regional distances, which are often taken to be more than 100 km and up to 1,000 km, seismic waves are strongly modified by interaction of the wavefield with the crust and upper mantle. S and P arrivals are also separated by greater times than for shorter distances, such that existing deep learning models may not perform well on these out of distribution data. This provides the motivation for developing CREW. The increase in source-receiver distance comes with mounting complexity in the waveforms due to the accumulation of propagation effects and the decrease in wave amplitudes. Figure 1 schematically compares wave propagation at local distances vs. regional distances.

The waveforms on top are recordings of the 2023 Lake Almanor earthquake in Northern California. This M_W

*Corresponding author: aguilars@stanford.edu

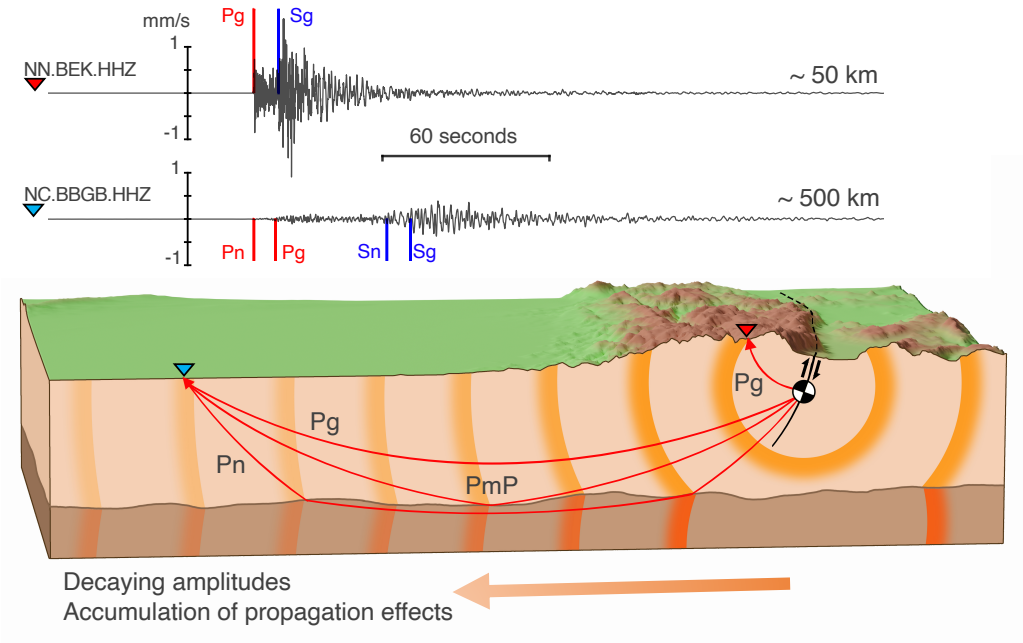

Figure 1 Comparison of local and regional recordings for the 2023 Lake Almanor earthquake in Northern California. The top waveform was recorded at around 50 km while the bottom one was recorded at about 500 km. Note: The focal mechanism, elevation and fault geometry are not related to the real Lake Almanor earthquake setting.

5.5 earthquake was recorded over many instruments at both distance ranges. The top seismogram comes from station BEK from the Nevada Seismological Network at a distance of around 50 km. In this case the arrivals of the crustal phases Pg and Sg are very impulsive and they are around 10 seconds apart. The bottom waveform, from station BBGB of the Northern California Seismic Network recorded the earthquake at a distance close to 500 km from the epicenter, shows that the waves that traveled through the uppermost mantle, Pn and Sn, arrive before the direct crustal arrivals, Pg and Sg. The Pn and Sn arrivals are emergent and more difficult to see. Both seismograms are 5 minutes long and are aligned on the first arrival. The vertical scale of both seismograms is the same, with the top one having a peak ground velocity of 1.81 mm/s while the regional recording has a peak ground velocity of 0.40 mm/s, which is almost a five fold decrease in peak ground velocity.

As indicated in Figure 1, for earthquakes recorded at short local distances, the first arrivals are the direct crustal phases Pg and Sg, which propagated through the crust. As the source to receiver distance increases, earthquake recordings may include the Moho-reflected phases PmP and SmS. Beyond the crossover distance, Pn and Sn will be the first arrivals. These waves travel from the source and propagate through the uppermost mantle before turning to the surface again (Storchak et al., 2003). The crossover distance is a function of earthquake depth and crustal thickness, and ranges from 30 km in thin oceanic crust to 200 km in thick continental crust, since crustal thickness can vary from 6 km to 70 km (Mooney et al., 1998). For reference, for a 30 km thick continental crust and assuming typical seismic velocities for the crust and upper mantle, the crossover distance for a surface source would be ~ 150 km.

For most regional earthquake recordings the first arrival is the Pn phase and for S waves, the first arrival is its analog Sn. As seen in Figure 1, the characteristics of the waveforms are different for the local and the regional recordings. The first arrivals Pn and Sn are known to be emergent, compared to the impulsive nature of Pg and Sg. The decay of coda (its envelope) for local recordings tends to follow a one over time pattern (Sato et al., 2012), with the maximum amplitude very close to the first arrival. In contrast, for the regional recording, the envelope of the P and S codas looks more like a spindle, with the maximum amplitudes not as close to the first arrivals. This change in shape is attributed to scattering, which is strongest in the crust and uppermost mantle (Shearer and Earle, 2004). Even though in Figure 1 the secondary S arrival is labeled as Sg, at longer distances, close to 1000 km the high frequency S wave train has been attenuated and only S waves trapped in the crustal waveguide, known as Lg will be the secondary S wave arrival. Lg phases are complex, and can be blocked by changes in crustal structure (Al-Damegh et al., 2004). Careful attention to the demanding task of precise picking of these regional seismic phases leads to improved earthquake catalogs in zones that are otherwise challenging to monitor (Fuenzalida et al., 2013), which suggests that deep-learning-based methods should be extremely useful at these distances.

Figure 2 shows the International Seismological Centre station inventory list with the inverted blue triangles. Seismically active parts of Europe, Japan, New Zealand and the United States, have the greatest concentration of instruments. The green contours and areas are those for which there is a minimum of 5 stations within a radius of 3 degrees. These green outlines enclose those highly instrumented regions of the world where "local" earthquake monitoring with direct crustal

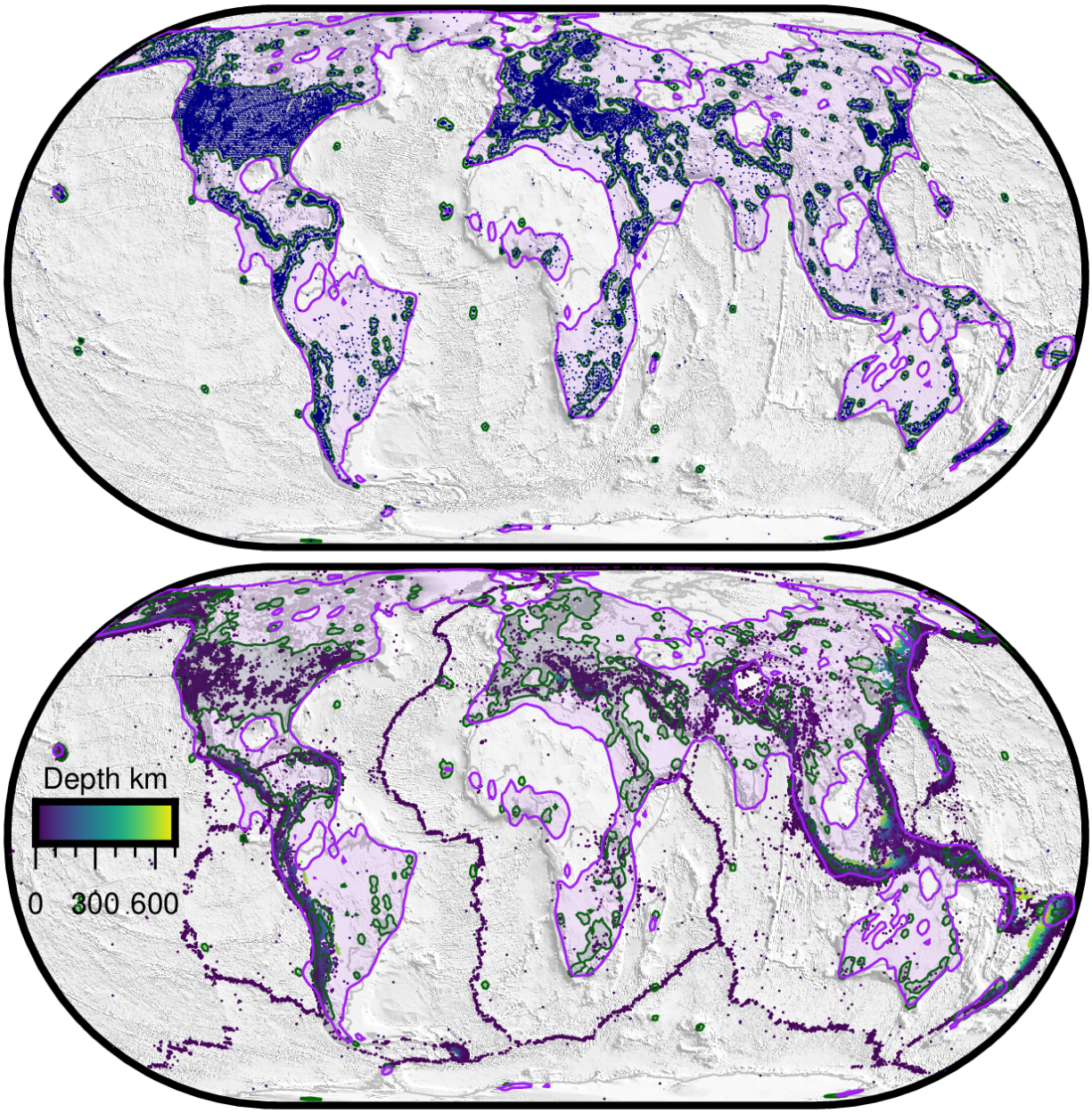

Figure 2 Top. Stations in the ISC inventory list. Bottom. Global seismicity. In both panels the green regions are those for which there are at least 5 stations within a 3 degree radius and the purple regions are those for which there are a minimum of 5 stations within a 10 degree radius and the azimuthal gap for an earthquake within this region would be less than 180 degrees.

phases should be possible. The purple-shaded region indicates where there is a minimum of 5 stations within a 10 degree radius and where the azimuthal gap for an earthquake at each point is less than 180 degrees. These purple regions are those that can be considered suitable for regional monitoring. The area ratio between the green and purple regions is about 10, which indicates that there should be great benefit to more effective regional earthquake monitoring. For the important case of small islands, such as the Azores or Ascension in the South Atlantic, they do not meet the azimuthal gap criteria due to their limited areal footprint. From the point of view of earthquake monitoring, mid-ocean ridges might be considered the least well-monitored seismic zones on Earth. Also, note that this set of stations does not represent current monitoring conditions accurately, since we do not consider information on the lifespan of these stations. For example, the Transportable Array stations across the United States only operated for approximately two years at any particular location, such that much of that area is covered by regional, rather than local, monitoring. That is, monitoring from permanent seismic networks in much of the world is not as effective as this figure suggests.

Adapting the machine learning workflows to regional earthquake monitoring and earthquake catalog building requires adapting the data and the algorithms. The Curated Regional Earthquake Waveforms (CREW) dataset is the first step towards extending deep (i.e., deep learning-based) earthquake catalogs to regional earthquake monitoring, by assembling a high quality benchmark dataset for training deep-learning models.

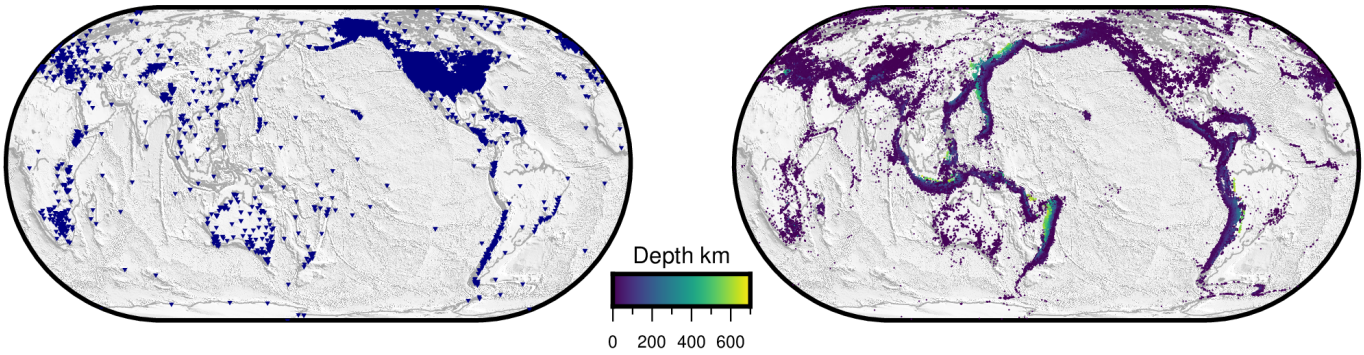

Figure 3 Stations and sources in CREW. Earthquakes are color-coded by depth.

2 Metadata and Data Collection

We queried all datacenters available through Obspy (Beyreuther et al., 2010) to retrieve their earthquake catalogs (see Table 2). We retained only those catalogs that contain information on both P and S arrivals, including phase arrivals P,Pg,Pn and S,Sg,Sn. For instance, catalogs that report only P arrivals were excluded from our workflow. Of the over 30 million metadata entries, we kept only those which for the same station-earthquake pair there were at least one picked arrival of P, Pg, Pn and one picked arrival of S,Sg,Sn on the same trace. That is, we required at least one of the P set and one of the S set to be labeled for each example. The number of traces for which simultaneous P and S information is available is an order of magnitude less than those for which only the first arriving P wave is labeled. Later, we queried all datacenters accesible via Obpsy to retrieve the appropriate waveforms in the distance range of 1 to 20 degrees of source to receiver distance, which is the range where the first arrivals are mainly Pn and Sn. Initially, we retrieved 7 minute waveforms, including 2 minutes before the earliest arriving P phase and 5 minutes after. This included all the instruments for each station, encompassing seismometers and accelerometers, with sampling rates ranging from 20 to 200 Hz. We detrended and resampled these data at 100 Hz. We then cut the waveforms randomly so that the earliest arriving P wave is at least 10 seconds after the start of the seismic trace and the total duration is 5 minutes (300 seconds), zero padding when required to complete the 5 minutes. Then, the waveforms were normalized to the absolute peak amplitude among the three channels. Ultimately, we kept the data from the ISC catalog (Storchak et al., 2013) and waveforms from the IRIS DMC (Trabant et al., 2012). The initial pool of data included nearly 3.3 million waveforms and their corresponding arrivals. The sources and receivers represented in the dataset are shown in Figure 3. The database includes 523,294 unique events recorded at 4,071 unique stations around the world. The distribution of earthquakes is representative of global seismology, spanning all latitudes, longitudes and all depths. In contrast, the coverage of receivers is not uniform, as the places with the highest density of instruments are the USA, Chile, and Europe.

Figure 4 displays five examples in the dataset, with their three-component waveforms, along with the arrivals and their labels, and indicate the instrument type and information on the earthquake location and magnitude, as well as the source to receiver distance. These examples are shown for the presence of 2,3 and 4 picked arrivals. Panels A and B represent the most common cases in CREW, where only the first arriving P wave and the first arriving S wave are labeled. For the example in (A), generic P and S labels are provided, whereas for (B), more specific Pn and Sn labels are provided. Panel (C) depicts a case in which three labels are provided, P, Pn and S, but P and Pn represent the same timestamp, so there are effectively two labeled arrivals. (D) shows the case of three distinct phases labeled, Pn, Sn and Sg. The bottom panel of Figure 4 (E) shows an uncommon example, in which the four phases Pn,Pg, Sn and Sg are all labeled, only a few thousand such examples occur in the dataset because most datacenters do not label arrivals other than the first arrival. Note that this example has been bandpass filtered to enhance the visibility of the arrivals. These rare examples typically come from stable continental regions, where the propagation of regional phases is not blocked by crustal and mantle structure (Gök et al., 2003). Panel C is a case in which there are two differently labeled arrivals, Pn and P that are very close in time, corresponding to the same arrival, but having an almost negligible time difference. In cases like this, we preserved all the available labels, but in subsequent workflows we only employed the earliest of the available P arrivals and the earliest of the S arrivals.

3 From Big Data to Good Data

In several fields employing machine learning, performance gains from dataset cleaning and refinement have been shown to surpass those from model architecture improvements (Northcutt et al., 2021a,b). Moreover, if data quality is high, effective training of deep neural networks requires fewer data (Motamedi et al., 2021). This has caused a shift in attention from the quantity of data to the quality of the data, as Data-Centric AI has gained traction (Zha et al., 2023) and led to data-centric initiatives (https://cleanlab.ai/) and competitions (Ng et al., 2021) (https://https-deeplearning-ai.github.io/data-centric-comp/) for more controlled benchmark datasets. (Northcutt et al., 2021a) documented the prevalence of faulty examples for ten of the most used machine learning benchmark datasets including image, text and

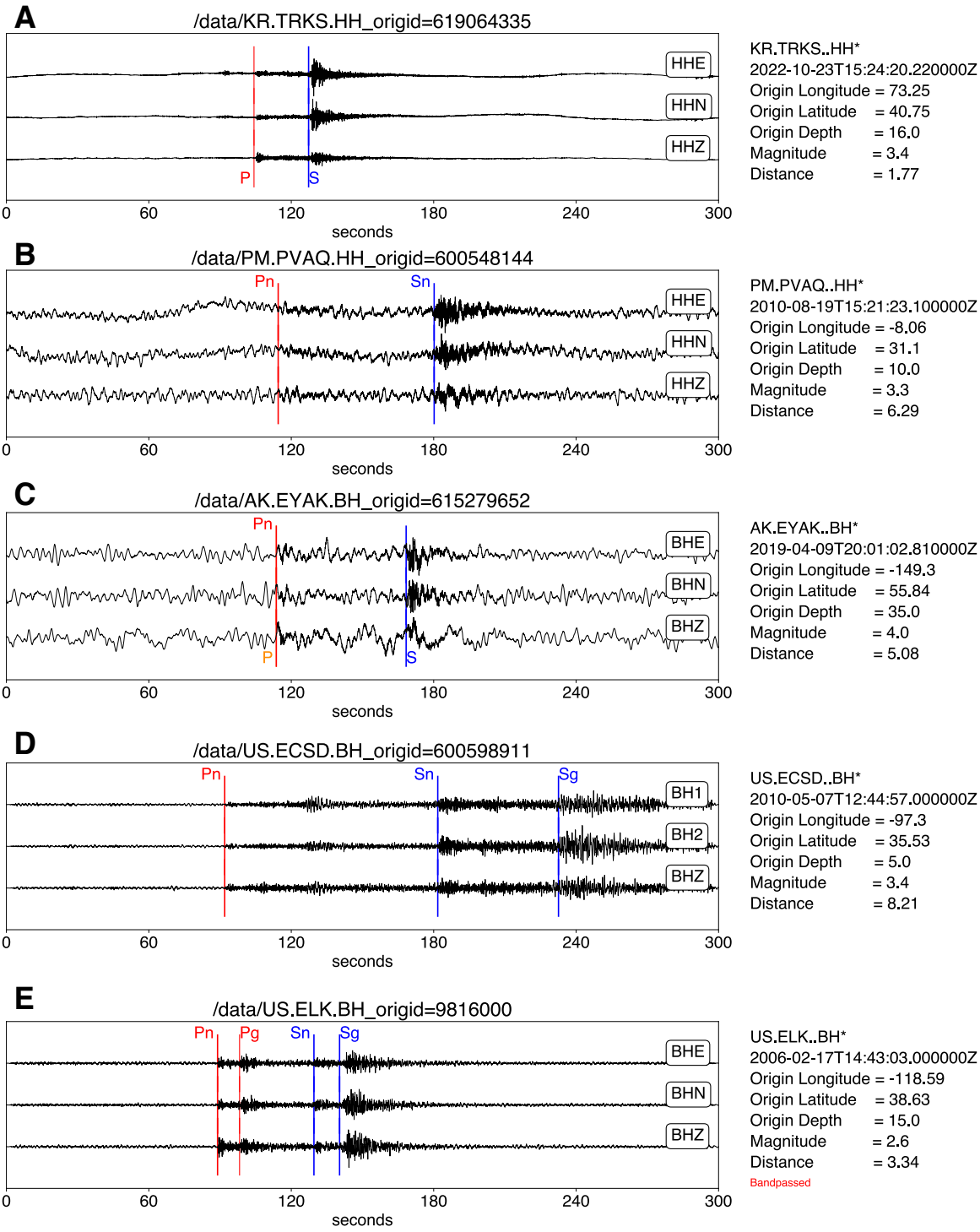

Figure 4 Examples from CREW with 2 (A,B), 3 (B,C), and 4 (E) labeled arrivals. Depth is in km, and distance in degrees. Example in panel (E) has been bandpassed filtered between 1 and 10 Hz to facilitate visualization of the arrivals.

audio, MNIST, CIFAR, and ImageNet among others (https://labelerrors.com/). These examples contain either faulty data or defective or incomplete labels, that end up affecting model selection and performance. Seismological data can contain errors such as inaccurate picked phase arrivals, and seismometer data is prone to corrupted transmission or storage. Strategies to mitigate the effects of bad labels and bad data have been devised (Cordeiro and Carneiro, 2020; Northcutt et al., 2021b), and it remains an active research field. This motivated our shift in approach, from iterating over a fixed dataset and optimizing for model parameters, to fixing the architecture and iteratively improving the dataset by identifying faulty examples, outliers, and edge cases, and/or by synthesizing new examples. CREW was built to have both big and good data.

Upon inspection of the initial dataset, it was clear that there were many faulty examples of various types. We manually checked a random sample of 10,000 examples and classified them into the following categories: (1) good examples, which have seemingly accurate arrival time labels on clean seismograms and (2) bad examples, which have inaccurate arrival time labels, corrupted seismograms, or other problems. Another cate-

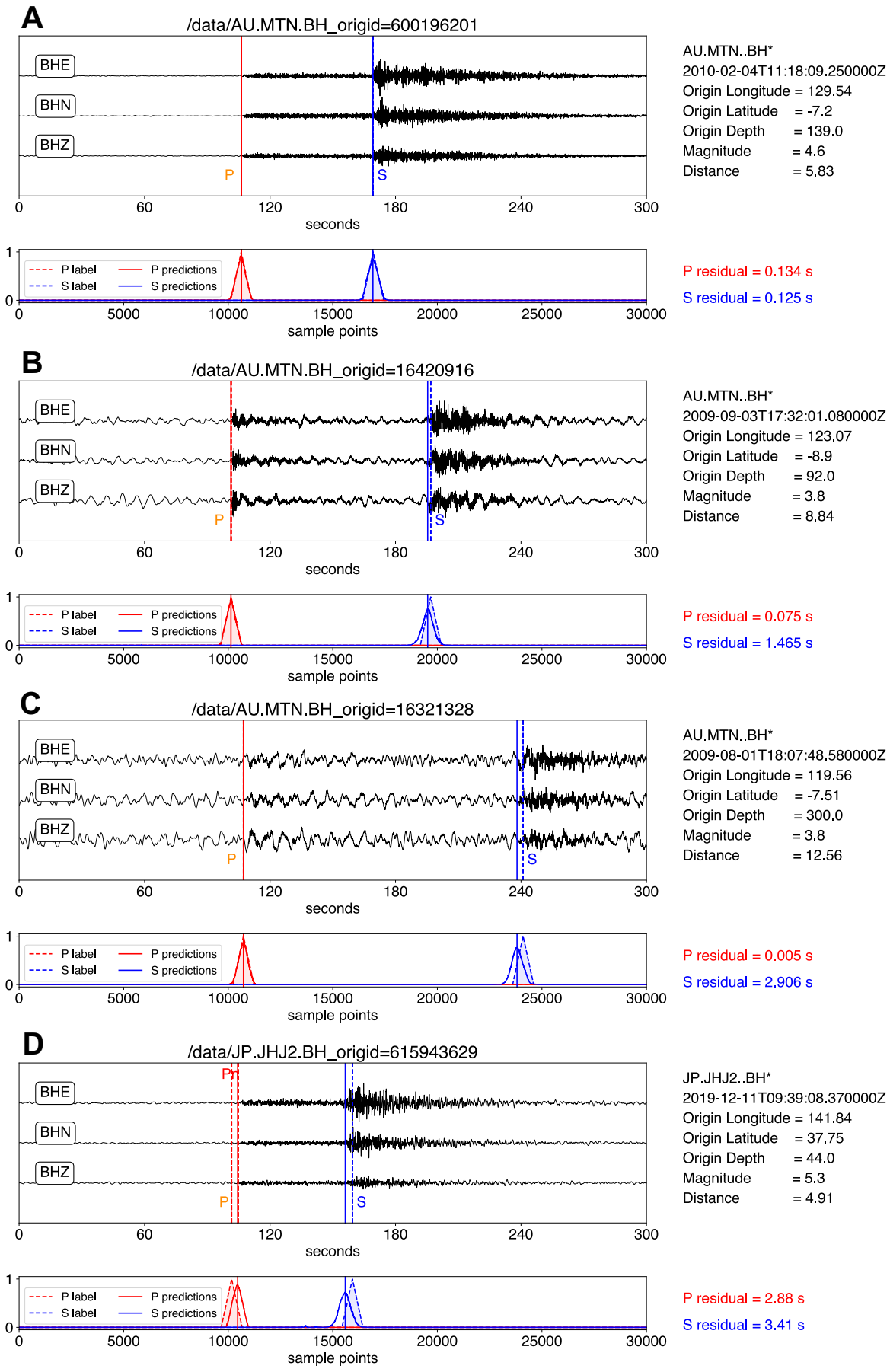

Figure 5 Comparison of labels and predictions, dotted lines are the labels in red for P waves and blue for S waves. The solid lines are the predictions and the inferred picks. The time difference between the two are displayed in the bottom right. Examples in A and B were kept in CREW, while examples in C and D were removed.

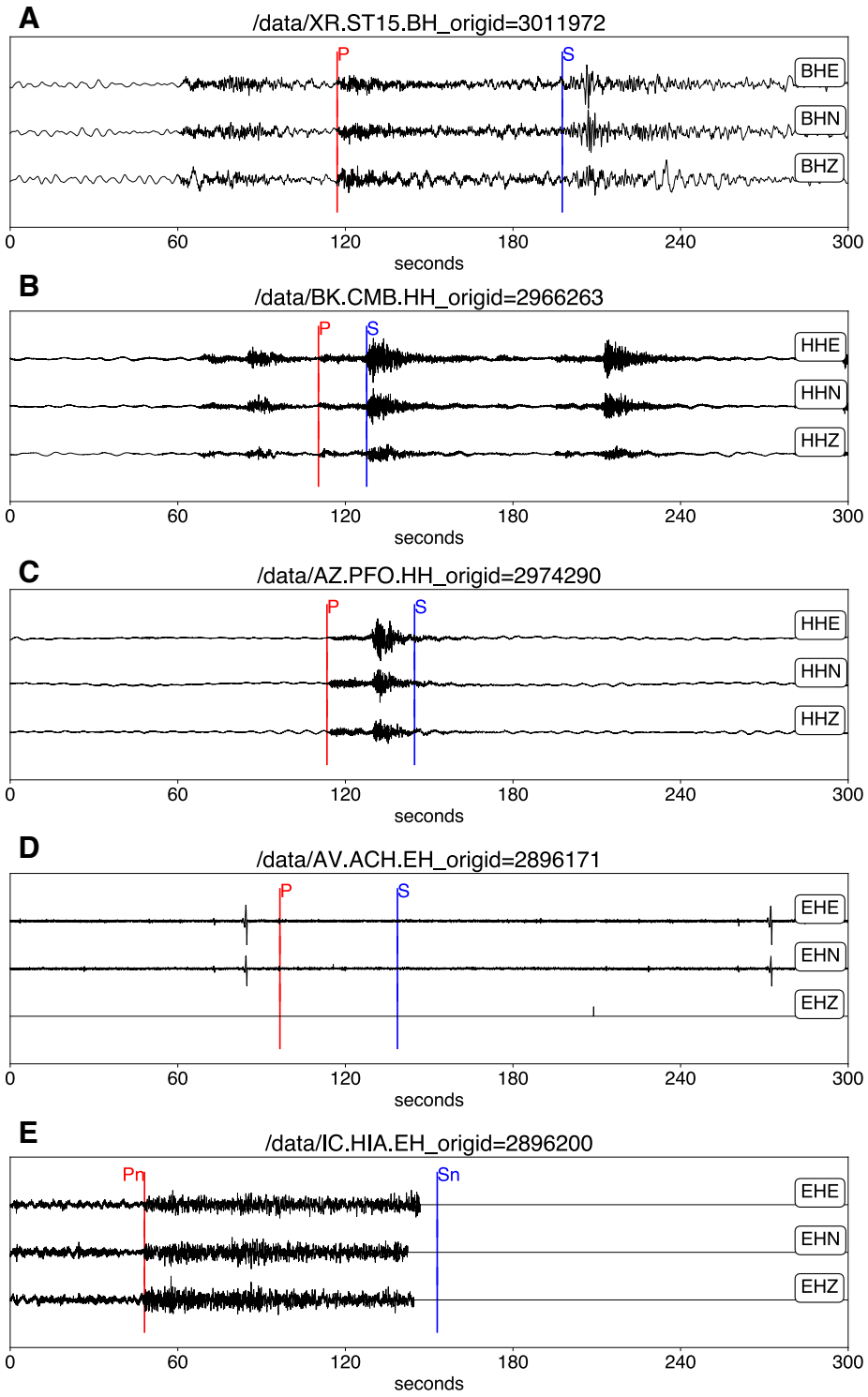

Figure 6 Examples rejected from CREW. (A) uncataloged earthquake. (B) multiple uncataloged earthquakes. (C) accurate P arrival next to an inaccurate S arrival. (D) no earthquake signal visible. (E) accurate Pn arrival but data is incomplete for the Sn arrival.

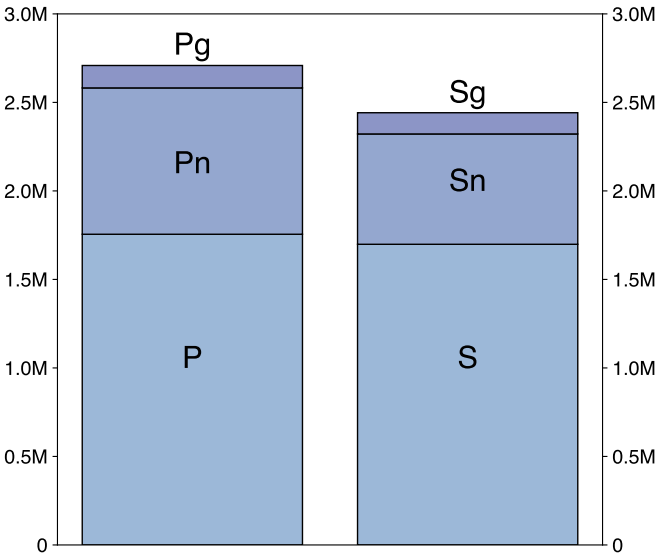

Figure 7 Number of picks in each category.

gory, (3) multiplets, accounted for the class where there are multiple earthquakes, but only one is labeled. From this sorting scheme 72% of the data was deemed as good (category 1) and the remaining 28% was flagged (categories 2 and 3) because it was deemed to contain inaccurate training labels, or corrupted data.

To automate the screening of faulty examples, we trained a convolutional neural network, a U-Net with skip connections, based on the architecture of PhaseNet (Zhu and Beroza, 2019). Our model has several extra layers to process input waveforms that are 10 times longer than those for the original PhaseNet, such that it produces representations in the deepest layers of similar size. The CNN was trained to learn triangular labels that are centered on the pick positions. These labels have a half width of 5 seconds, or 500 sample points on each side, for a total duration of 10 seconds, with a linear increase from 0 to 1 and then a linear decrease from 1 to 0. The labels are the same for both P and S waves, and there is a third channel for noise, which is equal to one minus the label of P and minus the label of S. These triangular labels were made using the earliest arrival among the available arrivals. For instance for the example displayed in Figure 4 the labels used were Pn and Sn. There are multiple examples in the dataset for which multiple arrivals are reported, but in some cases they are very close in time and hard to distinguish. For instance, in Figure 4 panel C there labels for both P and Pn and they are almost overlapping. We leave at the discretion of the user the use of these labels but note that the pruning procedure described here used the earliest arriving among P,Pn, and Pg and the earliest arriving among S,Sn, and Sg. Future research will address working with secondary arrivals.

The training data consists of a mix of data in its raw form, augmented data, and synthetic noise. The details of the architecture, training and deployment of these models and other auxiliary models will be presented in a forthcoming paper. The augmented versions of the data consisted of a superposition of multiple copies of the same example waveforms with a time delay. Depending on the S minus P time, we added a random choice between two or three copies, and with the appropriate labels, those were added to the example. We did this to train the model to work for the frequently encountered scenario where more than one earthquake occurs during a 5-minute window.

Once our phase picker was trained, we applied it to the dataset to remove examples with faulty labels. The criteria used was that the time difference between the dataset labels and the inferred phase picks was under 2 seconds. A large time difference between the label and the prediction was an indication of mistimed arrivals. Figure 5, shows data that passed this criteria and that did not pass it. The delay between labels and predictions are indicated in the bottom right, with red and blue for P and S waves. panel A shows very good alignment of the triangular labels and the model predictions, such that they appear totally superimposed. The predicted arrival times differ by only a tenth of a second, which is an example of what we consider good quality data and labels. Panel B shows good agreement in the P wave, but a delay of nearly a second and a half for the S wave, which is still considered sufficiently good data. Panels C and D show data that was rejected from the dataset because either the P or S prediction differs by more than 2 seconds from the dataset labels. For C, it is the S label that seems to be inaccurate, whereas for D, both the P and S labels are inaccurate, being evident for the S wave, but it requires zooming in to see the P label mislocation.

Figure 6 shows a variety of examples that were flagged as faulty by the described workflow. From top to bottom A through E. (A) Two earthquakes in one window, but only one of them has picked arrivals. (B) At least three uncataloged earthquakes, while only one is labeled. A and B represent the most common way in which the labels are inaccurate. (C) Correct P arrival but faulty S arrival. Without the need for hardcoding a travel time sanity check, our model flagged this type of error. (D) No visible earthquake signal in the waveforms. (E) Data gaps or outages, in this case there is a seemingly accurate Pn arrival, but there is a data gap before the Sn arrival.

The quality-controlled dataset contains 1,599,323 examples (nearly 50% of the initial data pool, nearly 1.1 TB), each a three component waveform sampled at 100 Hz with at least one of P,Pg,Pn and at least one of S,Sg,Sn arrivals. The total number of arrivals is 3,589,986. The proportions of these arrivals are displayed in Figure 7. For the P family there are 1,871,317 arrivals: 1,225,778 for generic P, 564,373 for Pn and 81,166 for Pg. For the S family, there are 1,718,669 arrivals: 1,192,110 for generic S, 446,880 for Sn, and 79,679 for Sg. The relatively low number of Pg and Sg phases is a consequence of excluding data in the 0 to 1 degree distance range. There are multiple existing data sets for those distances as described above.

The resulting dataset consists of 523,294 earthquakes that are globally distributed. The magnitude ranges from 0 to 7.1, with very few earthquakes at either extreme. Figure 8 (left) shows the magnitude-frequency distribution of the earthquakes in CREW as solid bars.

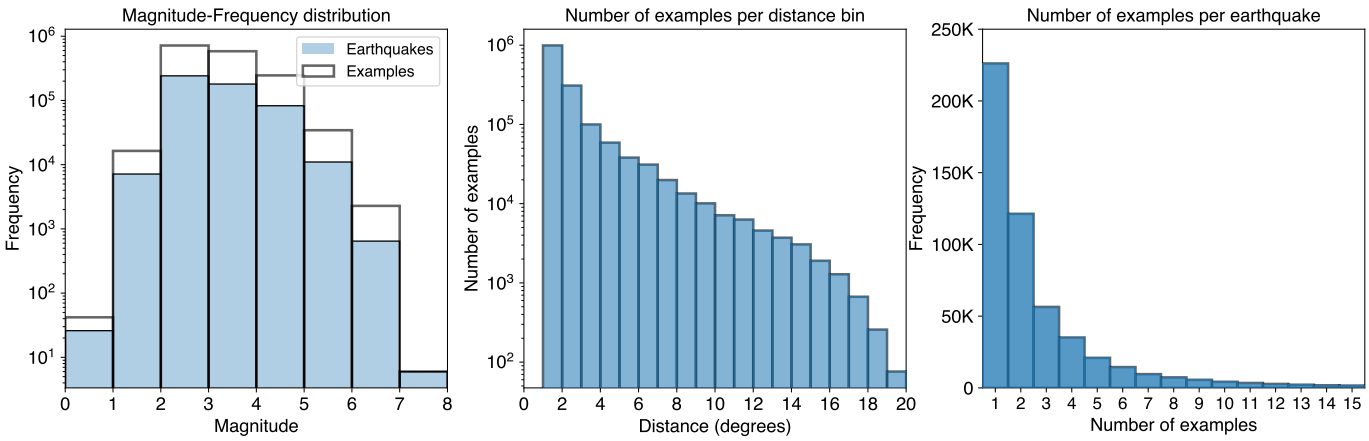

Figure 8 Left. Magnitude-Frequency distribution, the solid bars display the distribution of unique earthquakes in the dataset. The empty bars display the frequency distribution if each example is treated as a separate magnitude. Middle. Distribution of source to receiver distances, which span 1 to 20 degrees. Right. Number of waveforms of examples in the dataset per each unique origin ID, for most earthquakes there is only one or two observations, whereas having more than 5 is rather scarce in the CREW.

The outlined bars show the distribution if each example is treated as a different earthquake, that is counting the same earthquake multiple times. Figure 8 (middle) shows the distance distribution of the examples. Each bar represents a 1 degree distance bin. Most of the data is at the closest distances (distance < 4 degrees). The number of recordings decays dramatically with distance, due to the combined effects of amplitude decay and recording limitations, with only larger earthquakes visible at greater distances.

Table 1 summarizes the metadata attributes in CREW. These can be separated into three main categories, station information, earthquake origin information and arrivals information. CREW is stored in hdf5 format, the examples are stored in a group called **data**, where each individual example is named a combination of the station id and the event origin ID. Examples of these names are shown at the top of each panel in Figure 4. For the arrivals, the timestamps are available as well as the sample position corresponding to the location of the arrivals in the waveforms. In Figures 4 and 5 part of the metadata is displayed in the right panels. The names of the variables are mostly in Seisbench format (Woollam et al., 2022), except for the channels list.

The right panel of Figure 8 shows how many examples there are that correspond to a unique earthquake. The most common scenario is that only one recording per earthquake made it through the quality control. Nearly 230,000 earthquakes have only one example, i.e., at least two phase arrivals. On the other hand, 1,251,900 examples correspond to an earthquake with at least 6 phase readings, i.e. a seismic source for which there are at least 3 examples in the dataset. This aspect should be useful for machine learning implementations that perform seismic phase picking incorporating information from multiple stations, e.g. (Feng et al., 2022), or for models that perform earthquake arrival association, e.g. (McBrearty and Beroza, 2023). The plot is clipped at 15, but the earthquake that has the most examples associated with it has 121, which means over 242 picked arrivals. CREW contains more examples with both P and S arrival information than other datasets that cover the same distance range.

We reviewed the data and metadata in CREW, which is global in coverage, containing waveforms from earthquakes from all longitudes, latitudes and depths. CREW includes events up to magnitude 7. Moreover it provides data and labels useful at the single station level as well as the network level, with the majority of the data corresponding to earthquakes with at least 6 arrivals, which should be enough to produce a location.

4 Conclusions and Future Directions

We introduce CREW as a large, high-quality labeled data set for simultaneous regional seismic P and S phase waveforms recorded on seismometers around the world. CREW is the first benchmark data set that focuses on regional phases, rather than phases from local earthquake recordings or teleseismic recordings. Monitoring using regional phases is essential for large parts of the Earth where local monitoring is logistically impractical or is not a high priority due to relatively low seismic hazard. It should also prove useful for the important case of nuclear test ban treaty monitoring. We hope that its availability will enable progress in machine learning for regional earthquake monitoring and structural imaging.

Most machine learning research on seismology has focused on supervised learning (Mousavi and Beroza, 2023), especially for earthquake monitoring, and CREW contributes to this paradigm by curating data with the best available labels for regional first and secondary arrivals. The combination of algorithmic advances and data advances will contribute to multiscale earthquake monitoring.

Future research directions include working on secondary arrivals, such as reflected e.g. PmP, PP or converted waves e.g. SP, PS, that even though not often used for earthquake location, are nevertheless very sensitive to Earth structure and provide insight into the deep interior of the planet. These secondary arrivals are also

Station information	Event Information	Arrivals Information
network_station_code	source_id	{P,Pg,Pn}_arrival_time
station_code	source_origin_time	{P,Pg,Pn}_arrival_sample
channels	source_latitude_deg	{S,Sg,Sn}_arrival_time
station_latitude_deg	source_longitude_deg	{S,Sg,Sn}_arrival_sample
station_longitude_deg	source_depth_km	trace_start_time
station_elevation_m	source_magnitude	
	path_ep_distance_deg	

Table 1 Metadata attributes in CREW. Most ot these attributes are in seisbench convention.

a challenge for machine learning due to the relative scarcity of labeled examples. For these phases, architectures that rely less heavily on labeled data, such as semi-supervised and self-supervised learning that can learn from incomplete labels or partial data might prove successful (Assran et al., 2023). Also, future implementations that aim to characterize the full wavefield by picking all types of seismic phases present should provide improved capabilities for both monitoring and studies of the deep Earth.

Data and Code Availability

CREW is hosted in Stanford University DataFarm: https://redivis.com/datasets/1z6w-e1w70hpmt (https://doi.org/10.57761/60b3-cv76). All codes used to generate and process the dataset are available at https://github.com/albertleonardo/CREW, as well as example data and notebooks. CREW will be made accessible via SeisBench Woollam et al. (2022).

Competing Interests

The authors declare no competing interests. This is a ChatGPT free manuscipt.

Acknowledgements

This work was supported by AFRL under contracts FA9453-19-C-0073 and FA9453-21-9-0054. The authors thank two anonymous reviewers and editors Jack Muir and Théa Ragon for their help in improving this manuscript.

References

Al-Damegh, K., Sandvol, E., Al-Lazki, A., and Barazangi, M. Regional seismic wave propagation (Lg and Sn) and Pn attenuation in the Arabian Plate and surrounding regions. *Geophysical Journal International*, 157(2):775–795, 2004. doi: 10.1111/j.1365-246X.2004.02246.x.

Alaska Earthquake Center, Univ. of Alaska Fairbanks. Alaska Geophysical Network, 1987. doi: 10.7914/SN/AK.

Alaska Volcano Observatory/USGS. Alaska Volcano Observatory, 1988. doi: 10.7914/SN/AV.

Alberta Geological Survey / Alberta Energy Regulator. Regional Alberta Observatory for Earthquake Studies Network, 2013. doi: 10.7914/SN/RV.

Albuquerque Seismological Laboratory (ASL)/USGS. US Geological Survey Networks, 1980. doi: 10.7914/SN/GS.

Albuquerque Seismological Laboratory (ASL)/USGS. China Digital Seismograph Network, 1986. doi: 10.7914/SN/CD.

Albuquerque Seismological Laboratory (ASL)/USGS. United States National Seismic Network, 1990. doi: 10.7914/SN/US.

Albuquerque Seismological Laboratory (ASL)/USGS. New China Digital Seismograph Network, 1992. doi: 10.7914/SN/IC.

Albuquerque Seismological Laboratory (ASL)/USGS. Global Telemetered Seismograph Network (USAF/USGS), 1993. doi: 10.7914/SN/GT.

Albuquerque Seismological Laboratory (ASL)/USGS. New England Seismic Network, 1994. doi: 10.7914/SN/NE.

Albuquerque Seismological Laboratory (ASL)/USGS. Intermountain West Seismic Network, 2003. doi: 10.7914/SN/IW.

Albuquerque Seismological Laboratory/USGS. Central and Eastern US Network, 2013. doi: 10.7914/SN/N4.

Albuquerque Seismological Laboratory/USGS. Global Seismograph Network (GSN - IRIS/USGS), 2014. doi: 10.7914/SN/IU.

Andy Nyblade. Africa Array- Uganda/Tanzania, 2007. doi: 10.7914/SN/ZP_2007.

Andy Nyblade. AfricaArray - Namibia, 2015. doi: 10.7914/SN/8A_2015.

Anne Meltzer. RAMP Virginia, 2011. doi: 10.7914/SN/YC_2011.

Anne Meltzer and Susan Beck. 2016 Pedernales Earthquake Aftershock Deployment Ecuador, 2016. doi: 10.7914/SN/8G_2016.

Anne Sheehan, Francis Wu, and Roger Bilham. Himalayan Seismotectonics, Nepal and Tibet, 2001. doi: 10.7914/SN/YL_2001.

Aristotle University of Thessaloniki. Aristotle University of Thessaloniki Seismological Network, 1981. doi: 10.7914/SN/HT.

Arizona Geological Survey. Arizona Broadband Seismic Network, 2007. doi: 10.7914/SN/AE.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., Lecun, Y., Ballas, N., and Fair, M. A. I. Joint-Embedding Predictive Architecture. *arXiv preprint arXiv:2301.08243v1*, 2023.

Australian National University (ANU, Australia). Australian Seismometers in Schools, 2011. doi: 10.7914/SN/S1.

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. ObsPy: A python toolbox for seismology. *Seismological Research Letters*, 81(3):530–533, 2010. doi: 10.1785/gssrl.81.3.530.

Botswana Geoscience Institute. Botswana Seismological Network, 2001. doi: 10.7914/SN/BX.

British Geological Survey. Great Britain Seismograph Network, 1970. doi: 10.7914/AV8J-NC83.

Bureau of Economic Geology, The University of Texas at Austin. Texas Seismological Network, 2016. doi: 10.7914/SN/TX.

California Institute of Technology and United States Geological

Survey Pasadena. Southern California Seismic Network, 1926. doi: 10.7914/SN/CI.

Cascades Volcano Observatory/USGS. Cascade Chain Volcano Monitoring, 2001. doi: 10.7914/SN/CC.

Central Asian Institute for Applied Geosciences. Central Asian Seismic Network of CAIAG, 2008. doi: 10.7914/SN/KC.

Centro de Geociencias, UNAM. Permanent Seismic Network of Queretaro State, Mexico, 2003. doi: 10.7914/SN/MG.

Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Ensenada. Red Sísmica del Noroeste de México, 1980. doi: 10.7914/SN/BC.

Charles University in Prague (Czech), Institute of Geonics, Institute of Geophysics, Academy of Sciences of the Czech Republic, Institute of Physics of the Earth Masaryk University (Czech), and Institute of Rock Structure and Mechanics. Czech Regional Seismic Network, 1973. doi: 10.7914/SN/CZ.

Cindy Ebinger. AFAR07, 2007. doi: 10.7914/SN/ZE_2007.

Cole, H. M., Yeck, W. L., and Benz, H. M. MLAAPDE: A Machine Learning Dataset for Determining Global Earthquake Source Parameters. *Seismological Research Letters*, 20(20):1–11, 2023. doi: 10.1785/0220230021.

Colorado Geological Survey. Colorado Geological Survey Seismic Network, 2016. doi: 10.7914/SN/C0.

Cordeiro, F. R. and Carneiro, G. A Survey on Deep Learning with Noisy Labels: How to train your model when you cannot trust on the annotations? *Proceedings - 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2020*, pages 9–16, 2020. doi: 10.1109/SIBGRAPI51738.2020.00010.

Cynthia Ebinger. Magadi-Natron Magmatic Rifting Studies, 2013. doi: 10.7914/SN/XJ_2013.

DANA . Dense Array for North Anatolia, 2012. doi: 10.7914/SN/YH_2012.

Douglas Christensen, Roger Hansen, and Geoff Abers. Broadband Experiment Across the Alaska Range, 1999. doi: 10.7914/SN/XE_1999.

Douglas Wiens. Southwest Pacific Seismic Experiment, 1993. doi: 10.7914/SN/XB_1993.

Douglas Wiens. Mantle serpentinization and water cycling through the Mariana Trench and Forearc, 2012. doi: 10.7914/SN/XF_2012.

Dublin Institute for Advanced Studies. Irish National Seismic Network, 1993. doi: 10.7914/SN/EI.

Federal Institute for Geosciences and Natural Resources. German Regional Seismic Network (GRSN), 1976. doi: 10.25928/MBX6-HR74.

Feng, T., Mohanna, S., and Meng, L. EdgePhase: A Deep Learning Model for Multi-Station Seismic Phase Picking. *Geochemistry, Geophysics, Geosystems*, 23(11), 2022. doi: 10.1029/2022GC010453.

Frank Vernon. ANZA Regional Network, 1982. doi: 10.7914/SN/AZ.

Frank Vernon. Broadband Seismic Characterization of the Arabian Shield, 1995. doi: 10.7914/SN/XI_1995.

Fuenzalida, A., Schurr, B., Lancieri, M., Sobiesiak, M., and Madariaga, R. High-resolution relocation and mechanism of aftershocks of the 2007 Tocopilla (Chile) earthquake. *Geophysical Journal International*, 194(2):1216–1228, 2013. doi: 10.1093/gji/ggt163.

Geoffrey A. Abers and Karen M. Fischer. Broadband Tomography Under Costa Rica and Nicaragua, 2003. doi: 10.7914/SN/YO_2003.

GEOFON Data Centre. GEOFON Seismic Network, 1993. doi: 10.14470/TR560404.

Geological Survey of Canada. Public Safety Geoscience Program Canadian Research Network, 2013. doi: 10.7914/SN/PQ.

Geophysical Survey of the National Academy of sciences of Tajikistan. Tajikistan National Seismic Network, 2009. doi: 10.7914/SN/TJ.

Geoscience Australia. Australian National Seismograph Network Data Collection, 2021. doi: 10.26186/144675.

GNS Science. GeoNet Aotearoa New Zealand Seismic Digital Waveform Dataset, 2021. doi: 10.21420/G19Y-9D40.

Gök, R., Sandvol, E., Türkelli, N., Seber, D., and Barazangi, M. Sn attenuation in the Anatolian and Iranian plateau and surrounding regions. *Geophysical Research Letters*, 30(24):3–6, 2003. doi: 10.1029/2003GL018020.

Idaho National Laboratory. INL Seismic Monitoring Program, 1972. doi: 10.7914/SN/IE.

Incorporated Research Institutions For Seismology. SINGLE STATION, 1970. doi: 10.7914/SN/SS.

Institut Cartogràfic i Geològic de Catalunya. Catalan Seismic Network, 1984. doi: 10.7914/SN/CA.

Institut De Physique Du Globe De Paris (IPGP). GNSS, seismic broadband and strong motion permanent networks in West Indies, 2008. doi: 10.18715/ANTILLES.WI.

Institut de physique du globe de Paris (IPGP) and École et Observatoire des Sciences de la Terre de Strasbourg (EOST). GEOSCOPE, French Global Network of broad band seismic stations, 1982. doi: 10.18715/GEOSCOPE.G.

Institute of Earth Sciences, Academia Sinica, Taiwan. Broadband Array in Taiwan for Seismology, 1996. doi: 10.7914/SN/TW.

Institute of Geophysics China Earthquake Administration (IGPCEA). China National Seismic Network, Data Management Centre of China National Seismic Network at Institute of Geophysics, CEA, 2000. doi: 10.7914/SN/CB.

Institute of GeoSciences (IGEO), Polytechnic University of Tirana (PUT). Albanian Seismological Network, 2002. doi: 10.7914/SN/AC.

Instituto Dom Luiz - Faculdade de Ciências da Universidade de Lisboa. Instituto Dom Luiz (IDL) - Faculdade de Ciencias Universidade de Lisboa, 2003. doi: 10.7914/SN/LX.

Instituto Nacional de Sismologia, Vulcanologia, Meteorologia e Hidrologia (INSIVUMEH). Red Sismologica Nacional, 1976. doi: 10.7914/SN/GI.

Instituto Nicaraguense de Estudios Territoriales (INETER). Nicaraguan Seismic Network, 1975. doi: 10.7914/SN/NU.

Instituto Politecnico Loyola. Observatorio Sismológico Politécnico Loyola, 2012. doi: 10.7914/SN/LO.

Instituto Português do Mar e da Atmosfera, I.P. Portuguese National Seismic Network, 2006. doi: 10.7914/SN/PM.

IRIS Transportable Array. USArray Transportable Array, 2003. doi: 10.7914/SN/TA.

Istituto Nazionale di Oceanografia e di Geofisica Sperimentale. Antarctic Seismographic Argentinean Italian Network - ASAIN, 1992. doi: 10.7914/SN/AI.

Istituto Nazionale di Oceanografia e di Geofisica Sperimentale. International OGS Network, 2014. doi: 10.7914/SN/IO.

Istituto Nazionale di Oceanografia e di Geofisica Sperimentale - OGS. North-East Italy Seismic Network, 2016. doi: 10.7914/SN/OX.

Jay Pulliam. Greater Antilles Seismic Program, 2013. doi: 10.7914/SN/ZC_2013.

Jim Gaherty, Cindy Ebinger, Andy Nyblade, and Donna Shilling-

ton. Study of Extension and maGmatism in Malawi aNd Tanzania, 2013. doi: 10.7914/SN/YQ_2013.

Jim Ni, Tom Hearn, and John Nabelek. INDEPTH-III, 1997. doi: 10.7914/SN/XR_1997.

Kandilli Observatory And Earthquake Research Institute, Boğaziçi University. Kandilli Observatory And Earthquake Research Institute (KOERI), 1971. doi: 10.7914/SN/KO.

Kate Miller. Bhutan Pilot Experiment, 2002. doi: 10.7914/SN/XA_2002.

Kentucky Geological Survey/Univ. of Kentucky. Kentucky Seismic and Strong Motion Network, 1982. doi: 10.7914/SN/KY.

KNDC/Institute of Geophysical Research (Kazakhstan). Kazakhstan Network, 1994. doi: 10.7914/SN/KZ.

KNMI. Netherlands Seismic and Acoustic Network, 1993. doi: 10.21944/E970FD34-23B9-3411-B366-E4F72877D2C5.

Kyrgyz Institute of Seismology, IVTAN/KIS and University of California, San Diego. Kyrgyz Seismic Telemetry Network, 1991. doi: 10.7914/SN/KN.

Kyrgyz Institute of Seismology, KIS . Kyrgyz Digital Network, 2007. doi: 10.7914/SN/KR.

Lamont Doherty Earth Observatory (LDEO), Columbia University. Lamont-Doherty Cooperative Seismographic Network, 1970. doi: 10.7914/SN/LD.

LTD Seismological Experience and Methodology Expedition of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan. Seismic network of the Seismological Experience and Methodology Expedition CS MES RK, 2003. doi: 10.7914/SN/QZ.

Magrini, F., Jozinović, D., Cammarano, F., Michelini, A., and Boschi, L. Local earthquakes detection: A benchmark dataset of 3-component seismograms built on a global scale. *Artificial Intelligence in Geosciences*, 1(July):1–10, 2020. doi: 10.1016/j.aiig.2020.04.001.

McBrearty, I. W. and Beroza, G. C. Earthquake Phase Association with Graph Neural Networks. *Bulletin of the Seismological Society of America*, 113(2):524–547, 2023. doi: 10.1785/0120220182.

MedNet Project Partner Institutions. Mediterranean Very Broadband Seismographic Network (MedNet), 1990. doi: 10.13127/SD/FBBBTDTD6Q.

Michael West. The Colima Deep Seismic Experiment: Imaging the Magmatic Root of Colima Volcano, 2006. doi: 10.7914/SN/ZA_2006.

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V. INSTANCE-the Italian seismic dataset for machine learning. *Earth System Science Data*, 13(12):5509–5544, 2021. doi: 10.5194/essd-13-5509-2021.

Montana Bureau of Mines and Geology/Montana Tech (MBMG, MT USA). Montana Regional Seismic Network, 1982. doi: 10.7914/SN/MB.

Mooney, W. D., Laske, G., and Masters, T. G. CRUST 5.1: A global crustal model at 5° × 5°. *Journal of Geophysical Research: Solid Earth*, 103(1):727–747, 1998. doi: 10.1029/97jb02122.

Motamedi, M., Sakharnykh, N., and Kaldewey, T. A Data-Centric Approach for Training Deep Neural Networks with Less Data. *35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021. doi: 10.48550/arXiv.2110.03613.

Mousavi, S. M. and Beroza, G. C. Machine Learning in Earthquake Seismology. *Annual Review of Earth and Planetary Sciences*, 51 (1), 2023. doi: 10.1146/annurev-earth-071822-100323.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C. STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI. *IEEE Access*, 7:179464–179476, 2019. doi: 10.1109/ACCESS.2019.2947848.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., and Beroza, G. C. Earthquake transformer—an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, 11(1):1–12, 2020. doi: 10.1038/s41467-020-17591-w.

Mozambique Rift Tomography, George Helffrich, and Fonseca, J F B D. Mozambique Rift Tomography, 2011. doi: 10.7914/SN/6H_2011.

National Centre for Seismological Research (CENAIS Cuba). Servicio Sismologico Nacional de Cuba, 1998. doi: 10.7914/SN/CW.

National Observatory of Athens, Institute of Geodynamics, Athens. National Observatory of Athens Seismic Network, 1975. doi: 10.7914/SN/HL.

National Seismological Centre. Nepal Kathmandu, 1978. doi: 10.7914/SN/NK.

National Seismological Centre. Centro Nacional de Sismologia, 1998. doi: 10.7914/SN/DR.

Natural Resources Canada. Canadian National Seismograph Network, 1975. doi: 10.7914/SN/CN.

Ng, A., Laird, D., and He, L. Data-Centric AI Competition. *https://https-deeplearning-ai.github.io/data-centric-comp/*, 2021. https://https-deeplearning-ai.github.io/data-centric-comp/.

Ni, Y., Hutko, A. R., Skene, F., Denolle, M. A., Malone, S. D., Bodin, P., Hartog, J. R., and Wright, A. K. Curated Pacific Northwest AI-ready Seismic Dataset. *Seismica*, 2, 2023. doi: 10.26443/seismica.v2i1.368.

NOAA National Oceanic and Atmospheric Administration (USA). National Tsunami Warning Center Alaska Seismic Network, 1967. doi: 10.7914/SN/AT.

Norsar. NORSAR Station Network, 1971. doi: 10.21348/D.NO.0001.

Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv preprint arXiv:2103.14749*, pages 1–16, 2021a. http://arxiv.org/abs/2103.14749.

Northcutt, C. G., Jiang, L., and Chuang, I. L. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021b. doi: 10.1613/JAIR.1.12125.

Northern California Earthquake Data Center. Berkeley Digital Seismic Network (BDSN), 2014. doi: 10.7932/BDSN.

Obsercatorio Vulcanológico y Sismológico de Costa Rica, Universidad Nacional. Observatorio Vulcanológico y Sismológico de Costa Rica, 1984. doi: 10.7914/SN/OV.

Observatório Nacional, Rio de Janeiro, RJ. Rede Sismográfica do Sul e do Sudeste, 2011. doi: 10.7914/SN/ON.

Ocean Networks Canada. NEPTUNE seismic stations, 2009. doi: 10.7914/SN/NV.

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

Ohio Geological Survey. Ohio Seismic Network, 1999. doi: 10.7914/SN/OH.

Oklahoma Geological Survey. Oklahoma Seismic Network, 1978. doi: 10.7914/SN/OK.

Oklahoma Geological Survey. Oklahoma Consolidated Temporary Seismic Networks, 2018. doi: 10.7914/SN/O2.

Pacific Tsunami Warning Center. Pacific Tsunami Warning Seismic System, 1965. doi: 10.7914/SN/PT.

Paul Silver. Anatomy of an Archean Craton, South Africa, A Multi-

disciplinary Experiment across the Kaapvaal Craton, 1997. doi: 10.7914/SN/XA_1997.

Penn State University. AfricaArray, 2004a. doi: 10.7914/SN/AF.

Penn State University. Pennsylvania State Seismic Network, 2004b. doi: 10.7914/SN/PE.

Red Sismica Volcan Baru. ChiriNet, 2000. doi: 10.7914/SN/PA.

Regional Integrated Multi-Hazard Early Warning System (RIMES Thailand). Regional Integrated Multi-Hazard Early Warning System, 2008. doi: 10.7914/SN/RM.

Roger Hansen and Gary Pavlis. Collaborative Research: St. Elias Erosion/Tectonics Project, 2005. doi: 10.7914/SN/XZ_2005.

Ross, Z. E., Meier, M. A., Hauksson, E., and Heaton, T. H. Generalized seismic phase detection with deep learning. *Bulletin of the Seismological Society of America*, 108(5):2894–2901, 2018. doi: 10.1785/0120180080.

Royal Observatory of Belgium. Belgian Seismic Network, 1985. doi: 10.7914/SN/BE.

San Fernando Royal Naval Observatory (ROA), Universidad Complutense De Madrid (UCM), Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum (GFZ), Universidade De Évora (UEVORA, Portugal), and Institute Scientifique Of Rabat (ISRABAT, Morocco). The Western Mediterranean BB seismic Network, 1996. doi: 10.14470/JZ581150.

Sato, H., Fehler, M. C., and Maeda, T. *Seismic Wave Propagation and Scattering in the Heterogeneous Earth : Second Edition*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-23029-5.

Scripps Institution of Oceanography. Global Seismograph Network - IRIS/IDA, 1986. doi: 10.7914/SN/II.

Seismological Laboratory of University of Basrah. Iraqi Seismic Observatory, 2014. doi: 10.7914/SN/MP.

Servicio Geológico Colombiano. Red Sismologica Nacional de Colombia, 1993. doi: 10.7914/SN/CM.

Shearer, P. M. and Earle, P. S. The global short-period wavefield modelled with a Monte Carlo seismic phonon method. *Geophysical Journal International*, 158(3):1103–1117, 2004. doi: 10.1111/j.1365-246X.2004.02378.x.

Stephane Rondenay. Multi-disciplinary Experiments for Dynamic Understanding of Subduction under the Aegean Sea, 2006. doi: 10.7914/SN/XS_2006.

Steve Roecker and Ray Russo. RAMP response for 2010 Chile earthquake, 2010. doi: 10.7914/SN/XY_2010.

Storchak, D. A., Schweitzer, J., and Bormann, P. The IASPEI standard seismic phase list. *Seismological Research Letters*, 74(6): 761–772, 2003. doi: 10.1785/gssrl.74.6.761.

Storchak, D. A., Giacomo, D. D., Bondár, I., Engdahl, E. R., Harris, J., Lee, W. H., Villaseñor, A., and Bormann, P. Public release of the ISC-GEM global instrumental earthquake catalogue (1900-2009). *Seismological Research Letters*, 84(5):810–815, 2013. doi: 10.1785/0220130034.

Susan Beck, Terry Wallace, and George Zandt. Slab Geometry in the Southern Andes, 2000. doi: 10.7914/SN/YC_2000.

Susan Schwartz. Imaging the Seismogenic Zone with Geodesy and Seismology, 1999. doi: 10.7914/SN/XY_1999.

Swiss Seismological Service (SED) At ETH Zurich. National Seismic Networks of Switzerland, 1983. doi: 10.12686/SED/NETWORKS/CH.

Sylvie Leroy, Derek Keir, and Graham Stuart. Young Conjugate Margins Lab in the Gulf of Aden, 2009. doi: 10.7914/SN/XW_2009.

Texas Seismological Network, Alexandros Savvaidis. Texas Seismological Network, 2018. doi: 10.7914/SN/4T_2018.

Trabant, C., Hutko, A. R., Bahavar, M., Karstens, R., Ahern, T., and Aster, R. Data products at the IRIS DMC: Stepping stones for research and other applications. *Seismological Research Letters*, 83(5):846–854, 2012. doi: 10.1785/0220120032.

UC Santa Barbara. UC Santa Barbara Engineering Seismology Network, 1989. doi: 10.7914/SN/SB.

Universidad de Chile. Red Sismologica Nacional, 2012. doi: 10.7914/SN/C1.

Universidad de Costa Rica. Red Sismológica Nacional de Costa Rica (RSN: UCR-ICE), 2016. doi: 10.15517/TC.

Universidad Nacional Autónoma de México (UNAM). Red Sísmica Mexicana, 1970. doi: 10.21766/SSNMX/SN/MX.

University of Nevada, Reno. Nevada Seismic Network, 1971. doi: 10.7914/SN/NN.

University of Oregon. Pacific Northwest Seismic Network - University of Oregon, 1990. doi: 10.7914/SN/UO.

University of Ottawa (uOttawa Canada). Yukon-Northwest Seismic Network, 2013. doi: 10.7914/SN/NY.

University of Puerto Rico. Puerto Rico Seismic Network & Puerto Rico Strong Motion Program, 1986. doi: 10.7914/SN/PR.

University of South Carolina. South Carolina Seismic Network, 1987. doi: 10.7914/SN/CO.

University of Utah. University of Utah Regional Seismic Network, 1962. doi: 10.7914/SN/UU.

University of Utah. Yellowstone National Park Seismograph Network, 1983. doi: 10.7914/SN/WY.

University of Washington. Pacific Northwest Seismic Network - University of Washington, 1963. doi: 10.7914/SN/UW.

USGS Alaska Anchorage. USGS Northern Mariana Islands Network, 2000. doi: 10.7914/SN/MI.

USGS Hawaiian Volcano Observatory (HVO). Hawaiian Volcano Observatory Network, 1956. doi: 10.7914/SN/HV.

Utrecht University (UU Netherlands). NARS, 1983. doi: 10.7914/SN/NR.

Vadim Levin. Retreating-Trench, Extension, and Accretion LTectonics: A Multidisciplinary Study of the Northern Apennines, 2003. doi: 10.7914/SN/YI_2003.

Various Institutions. International Miscellaneous Stations, 1965. doi: 10.7914/VEFQ-VH75.

Wenyuan Fan, Yihe Huang, and S. Shawn Wei. Lake-induced earthquakes in Lake Erie, 2018. doi: 10.7914/SN/7D_2018.

Won Sang Lee and Yongcheol Park. Korea Polar Observation Network, 2013. doi: 10.7914/SN/KP.

Woollam, J., Rietbrock, A., Bueno, A., and De Angelis, S. Convolutional neural network for seismic phase classification, performance demonstration over a local seismic network. *Seismological Research Letters*, 90(2 A):491–502, 2019. doi: 10.1785/0220180312.

Woollam, J., Münchmeyer, J., Tilmann, F., Rietbrock, A., Lange, D., Bornstein, T., Diehl, T., Giunchi, C., Haslinger, F., Jozinović, D., Michelini, A., Saul, J., and Soto, H. SeisBench-A Toolbox for Machine Learning in Seismology. *Seismological Research Letters*, 93(3):1695–1709, 2022. doi: 10.1785/0220210324.

Yeck, W. L., Patton, J. M., Ross, Z. E., Hayes, G. P., Guy, M. R., Nick, B., Shelly, D. R., Benz, H. M., and Earle, P. S. Leveraging Deep Learning in Global 24 / 7 Real-Time Earthquake Monitoring at the National Earthquake Information Center. *Seismological Research Letters*, 2020. doi: 10.1785/0220200178.Supplemental.

ZAMG - Zentralanstalt für Meterologie und Geodynamik. Austrian Seismic Network, 1987. doi: 10.7914/SN/OE.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., and Hu, X. Data-centric AI:

Perspectives and Challenges. *arXiv preprint arXiv:2301.04819*, 2023. http://arxiv.org/abs/2301.04819.

Zhu, W. and Beroza, G. C. PhaseNet: A deep-neural-network-based seismic arrival-time picking method. *Geophysical Journal International*, 216(1):261–273, 2019. doi: 10.1093/gji/ggy423.

Table 2 Seismological Networks used in CREW

4T	Texas Seismological Network, Alexandros Savvaidis (2018)
6H	Mozambique Rift Tomography et al. (2011)
7D	Wenyuan Fan et al. (2018)
7F	Central Arkansas Induced Eathquakes 2010-2011
8A	Andy Nyblade (2015)
8G	Anne Meltzer and Susan Beck (2016)
AC	Institute of GeoSciences (IGEO), Polytechnic University of Tirana (PUT) (2002)
AE	Arizona Geological Survey (2007)
AG	Arkansas Seismic Network
AF	Penn State University (2004a)
AI	Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (1992)
AK	Alaska Earthquake Center, Univ. of Alaska Fairbanks (1987)
AR	Northern Arizona Network
AT	NOAA National Oceanic and Atmospheric Administration (USA) (1967)
AU	Geoscience Australia (2021)
AV	Alaska Volcano Observatory/USGS (1988)
AY	Haitian Seismic Network
AZ	Frank Vernon (1982)
BC	Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), Ensenada (1980)
BE	Royal Observatory of Belgium (1985)
BK	Northern California Earthquake Data Center (2014)
BL	Brazilian Lithospheric Seismic Project
BX	Botswana Geoscience Institute (2001)
C	Chilean National Seismic Network
C0	Colorado Geological Survey (2016)
C1	Universidad de Chile (2012)
C8	Canadian Seismic Research Network
CA	Institut Cartogràfic i Geològic de Catalunya (1984)
CB	Institute of Geophysics China Earthquake Administration (IGPCEA) (2000)
CC	Cascades Volcano Observatory/USGS (2001)
CD	Albuquerque Seismological Laboratory (ASL)/USGS (1986)
CH	Swiss Seismological Service (SED) At ETH Zurich (1983)
CI	California Institute of Technology and United States Geological Survey Pasadena (1926)
CK	CAREMON Central Asian Cross-border network
CM	Servicio Geológico Colombiano (1993)
CN	Natural Resources Canada (1975)
CO	University of South Carolina (1987)
CS	Caucusus Array (CS)
CW	National Centre for Seismological Research (CENAIS Cuba) (1998)
CY	Cayman Islands Seismic Network
CZ	Charles University in Prague (Czech) et al. (1973)
DK	Danish Seismological Network
DR	National Seismological Centre (1998)
EC	Ecuador Seismic Network
EI	Dublin Institute for Advanced Studies (1993)
ET	CERI Southern Appalachian Seismic Network

Table 2 continued: Seismological Networks used in CREW

G	Institut de physique du globe de Paris (IPGP) and École et Observatoire des Sciences de la Terre de Strasbourg (EOST) (1982)
GB	British Geological Survey (1970)
GE	GEOFON Data Centre (1993)
GI	Instituto Nacional de Sismologia, Vulcanologia, Meteorologia e Hidrologia (INSIVUMEH) (1976)
GO	National Seismic Network of Georgia
GR	Federal Institute for Geosciences and Natural Resources (1976)
GS	Albuquerque Seismological Laboratory (ASL)/USGS (1980)
GT	Albuquerque Seismological Laboratory (ASL)/USGS (1993)
HK	Hong Kong Seismograph Network
HL	National Observatory of Athens, Institute of Geodynamics, Athens (1975)
HT	Aristotle University of Thessaloniki (1981)
HV	USGS Hawaiian Volcano Observatory (HVO) (1956)
IC	Albuquerque Seismological Laboratory (ASL)/USGS (1992)
IE	Idaho National Laboratory (1972)
II	Scripps Institution of Oceanography (1986)
IM	Various Institutions (1965)
IN	National Seismic Network of India
IO	Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (2014)
IU	Albuquerque Seismological Laboratory/USGS (2014)
IW	Albuquerque Seismological Laboratory (ASL)/USGS (2003)
JP	Japan Meteorological Agency Seismic Network
KC	Central Asian Institute for Applied Geosciences (2008)
KG	Korean Seismic Network - KIGAM
KN	Kyrgyz Institute of Seismology, IVTAN/KIS and University of California, San Diego (1991)
KO	Kandilli Observatory And Earthquake Research Institute, Boğaziçi University (1971)
KP	Won Sang Lee and Yongcheol Park (2013)
KR	Kyrgyz Institute of Seismology, KIS (2007)
KS	Korea National Seismography Network (KNSN-KMA) (KNSN)
KY	Kentucky Geological Survey/Univ. of Kentucky (1982)
KZ	KNDC/Institute of Geophysical Research (Kazakhstan) (1994)
LB	Leo Brady Network (LB)
LD	Lamont Doherty Earth Observatory (LDEO), Columbia University (1970)
LO	Instituto Politecnico Loyola (2012)
LX	Instituto Dom Luiz - Faculdade de Ciências da Universidade de Lisboa (2003)
MB	Montana Bureau of Mines and Geology/Montana Tech (MBMG, MT USA) (1982)
MG	Centro de Geociencias, UNAM (2003)
MI	USGS Alaska Anchorage (2000)
MN	MedNet Project Partner Institutions (1990)
MP	Seismological Laboratory of University of Basrah (2014)
MX	Universidad Nacional Autónoma de México (UNAM) (1970)
MY	Malaysian National Seismic Network
N4	Albuquerque Seismological Laboratory/USGS (2013)
NE	Albuquerque Seismological Laboratory (ASL)/USGS (1994)
NI	OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale) and University of Trieste (2002)
NK	National Seismological Centre (1978)
NL	KNMI (1993)
NM	Cooperative New Madrid Seismic Network
NN	University of Nevada, Reno (1971)
NO	Norsar (1971)
NR	Utrecht University (UU Netherlands) (1983)
NU	Instituto Nicaraguense de Estudios Territoriales (INETER) (1975)
NV	Ocean Networks Canada (2009)
NY	University of Ottawa (uOttawa Canada) (2013)

Table 2 continued: Seismological Networks used in CREW

NZ	GNS Science (2021)
O2	Oklahoma Geological Survey (2018)
OC	Observatorio Sismológico CIGEOBIO CONICET (OSCO)
OE	ZAMG - Zentralanstalt für Meterologie und Geodynamik (1987)
OH	Ohio Geological Survey (1999)
OK	Oklahoma Geological Survey (1978)
ON	Observatório Nacional, Rio de Janeiro, RJ (2011)
OV	Obsercatorio Vulcanológico y Sismológico de Costa Rica, Universidad Nacional (1984)
OX	Istituto Nazionale di Oceanografia e di Geofisica Sperimentale - OGS (2016)
PA	Red Sismica Volcan Baru (2000)
PE	Penn State University (2004b)
PL	Polish Seismological Network
PM	Instituto Português do Mar e da Atmosfera, I.P. (2006)
PO	Portable Observatories for Lithospheric Analysis and Research Investigating Seismicity (POLARIS)
PQ	Geological Survey of Canada (2013)
PR	University of Puerto Rico (1986)
PS	Pacific21 (ERI/STA)
PT	Pacific Tsunami Warning Center (1965)
QZ	LTD Seismological Experience and Methodology Expedition of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (2003)
RM	Regional Integrated Multi-Hazard Early Warning System (RIMES Thailand) (2008)
RV	Alberta Geological Survey / Alberta Energy Regulator (2013)
S1	Australian National University (ANU, Australia) (2011)
SB	UC Santa Barbara (1989)
SC	New Mexico Tech Seismic Network
SE	Southeastern Appalachian Cooperative Seismic Network
SS	Incorporated Research Institutions For Seismology (1970)
SV	Servicio Nacional de Estudios Territoriales (SNET), El Salvador (SNET-BB)
TA	IRIS Transportable Array (2003)
TC	Universidad de Costa Rica (2016)
TJ	Geophysical Survey of the National Academy of sciences of Tajikistan (2009)
TM	Thai Seismic Monitoring Network (TM)
TR	Eastern Caribbean Seismograph Network
TT	Seismic Network of Tunisia
TW	Institute of Earth Sciences, Academia Sinica, Taiwan (1996)
TX	Bureau of Economic Geology, The University of Texas at Austin (2016)
UO	University of Oregon (1990)
US	Albuquerque Seismological Laboratory (ASL)/USGS (1990)
UU	University of Utah (1962)
UW	University of Washington (1963)
WA	West Central Argentina Network
WI	Institut De Physique Du Globe De Paris (IPGP) (2008)
WM	San Fernando Royal Naval Observatory (ROA) et al. (1996)
WU	The Southern Ontario Seismic Network (SOSN)
WY	University of Utah (1983)
XA	Paul Silver (1997), Kate Miller (2002)
XB	Douglas Wiens (1993)
XE	Douglas Christensen et al. (1999)
XF	Douglas Wiens (2012)
XI	Frank Vernon (1995)
XJ	Cynthia Ebinger (2013)
XR	Jim Ni et al. (1997)
XS	Stephane Rondenay (2006)
XW	Sylvie Leroy et al. (2009)

Table 2 continued: Seismological Networks used in CREW

XY	Susan Schwartz (1999),Steve Roecker and Ray Russo (2010)
XZ	Roger Hansen and Gary Pavlis (2005)
YC	Susan Beck et al. (2000),Anne Meltzer (2011)
YG	Carpathian Basins Project Regional Array (CBPRA)
YH	DANA (2012)
YI	Vadim Levin (2003)
YJ	Ethiopia-Afar Geoscientific Lithospheric Experiment (EAGLE)
YK	Coordinated Seismic Experiment in the Azores (COSEA)
YL	Anne Sheehan et al. (2001)
YO	Geoffrey A. Abers and Karen M. Fischer (2003)
YQ	Jim Gaherty et al. (2013)
YV	North East Atlantic Tomography (NEAT)
ZA	Michael West (2006)
ZC	Jay Pulliam (2013)
ZE	Cindy Ebinger (2007)
ZF	Afar Consortium Network (AFAR)
ZP	Andy Nyblade (2007)

The article *Curated Regional Earthquake Waveforms (CREW) Dataset* © 2024 by Albert L. Aguilar Suarez is licensed under CC BY 4.0.