GTUNE: An Assembled Global Seismic Dataset of Underground Nuclear Test Blasts

Louisa Barama*[1] , Zhigang Peng[1] , Andrew V. Newman[1], and Jesse Williams[1]

Abstract

From catalogs of available declassified underground nuclear explosions, we compiled a comprehensive seismic waveform and event catalog termed GTUNE (Georgia Tech Underground Nuclear Explosions). Nuclear blast seismic records are sourced from previously prepared published datasets and openly available waveforms from online sources. All seismic traces were assembled into a user-friendly format compatible with most python-based machine learning (ML) packages. The GTUNE dataset includes the raw seismogram time series, event coordinates and origin time, sampling rate, station metadata, channel, epicentral distance, and *P*-wave arrival time from the origin dataset when available and otherwise identified using a tuned automated picker. This is the first openly available comprehensive global underground nuclear blast seismic dataset and consists of 28,123 vertical-component waveforms from 774 nuclear test blasts between 1961 and 2017 recorded between 0 and 90 epicentral degrees. For stations where data are not directly included due to data-sharing restrictions, the mechanisms to acquire and process these data are included. In this article, we describe various steps involved in data collection and quality control to ensure accurate labels, and present summary properties of the catalog and data set. The catalog was initially developed for applications with ML methods but can be used for a wide range of studies such as source physics, earth structure, and event detection methodological development.

Introduction

Problems of data availability, quality, and the incapability of synthetic data to match recorded ground motions at short periods, are problems that have often been the factors limiting progress in seismology using historical data (Kim and Ekström, 1996; Richards and Hellweg, 2020). In addition, the vast majority of nuclear events were recorded during the predigital seismic era (Figs. 1 and 2) in analog form on either paper or microfilms that remain largely underused (Bent et al., 2020). These issues are particularly problematic for research studies focusing on nuclear explosion events (Richards and Hellweg, 2020), as well as monitoring long-term changes in the deep Earth interior (Song and Richards, 1996; Vidale et al., 2000; Wang and Vidale, 2022). Over the past few decades, some seismograms digitized from analog recordings have become available to the seismological community (Richards et al., 1992; Walter et al., 2004; Ishii et al., 2015; Vidale, 2021). However, they are stored in different formats and locations, have restricted access or expired links (Bennett et al., 2010), making it difficult to utilize them in a systematic and efficient way.

With recent advancements in machine learning (ML), modern methods of seismic analysis and digitization of analog seismograms, the potential for using lower quality historical data in combination with high-quality digital data from more recent seismic events for effective analysis and research is proving possible (Del Pezzo et al., 2003; LeCun et al., 2015; Okal, 2015; Maceira et al., 2017; Nakano et al., 2019; Dickey et al., 2020; Richards and Hellweg, 2020). However, historical seismograms, such as the vast majority of underground nuclear test data, can be made more usable to most seismologists and data scientists if they are in formats that enable modern methods of analysis (e.g. Richards and Hellweg, 2020). A large, and more importantly, adequate quality dataset of nuclear blast seismograms can be used for developing more robust global nuclear blast detectors, and also to serve as examples of seismic events with verified ground-truth information including origin times, depths, and locations for Earth structure and source physics studies (e.g., Richards and Hellweg, 2020).

1. School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia, U.S.A., https://orcid.org/0000-0002-0049-0770 (LB); https://orcid.org/0000-0002-0019-9860 (ZP)

*Corresponding author: lbarama@gatech.edu

© Seismological Society of America
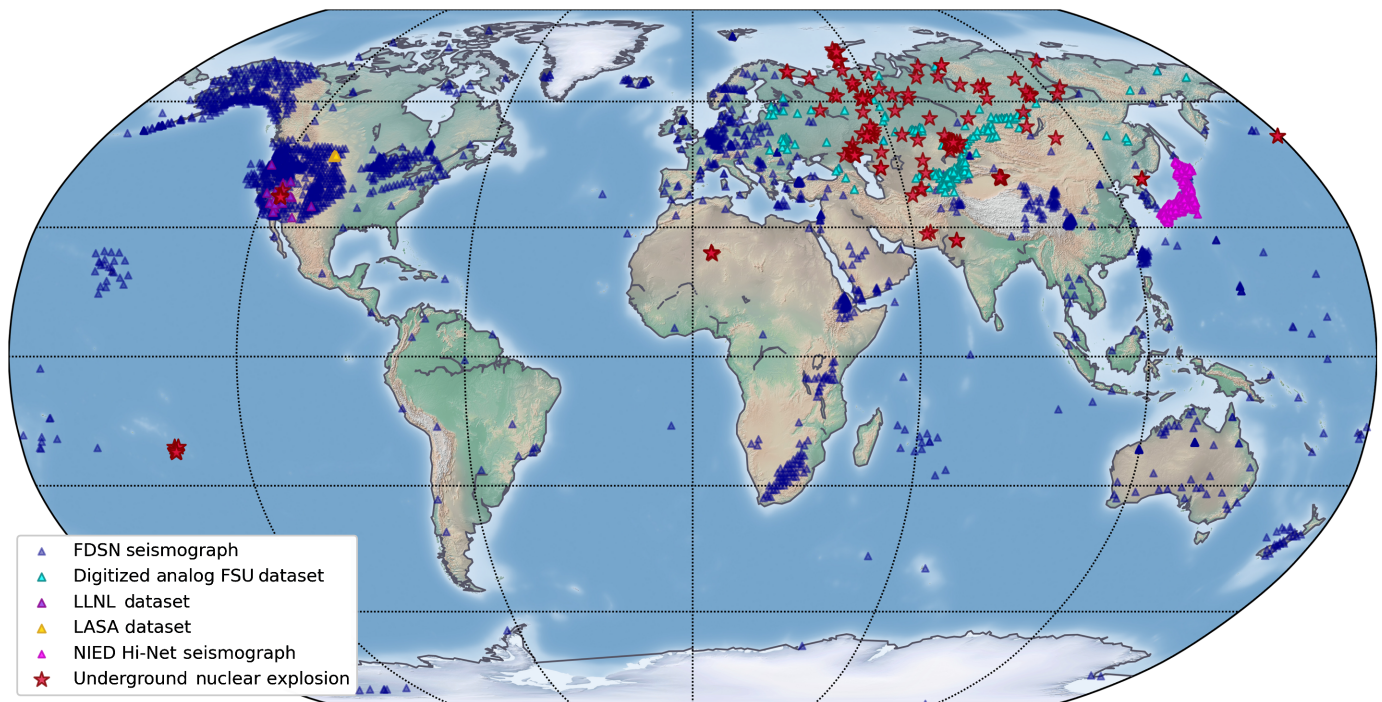
Figure 1. Global map of all underground nuclear explosions (UNE) included in the Georgia Tech Underground Nuclear Explosions (GTUNE) data. Multiple UNEs (red stars) frequently recur at or near the same location. Recording seismic stations (triangles) are also shown, with colors differentiating data sources: International Federation of Digital Seismograph Networks (FDSN) (black), digitized former Soviet Union (FSU) (teal; Richards *et al.*, 1992), Lawrence Livermore National Laboratory (LLNL) (purple; Walter *et al.*, 2004), and National Research Institute for Earth Science and Disaster Resilience (NIED) (pink; NIED, 2019). The color version of this figure is available only in the electronic edition.
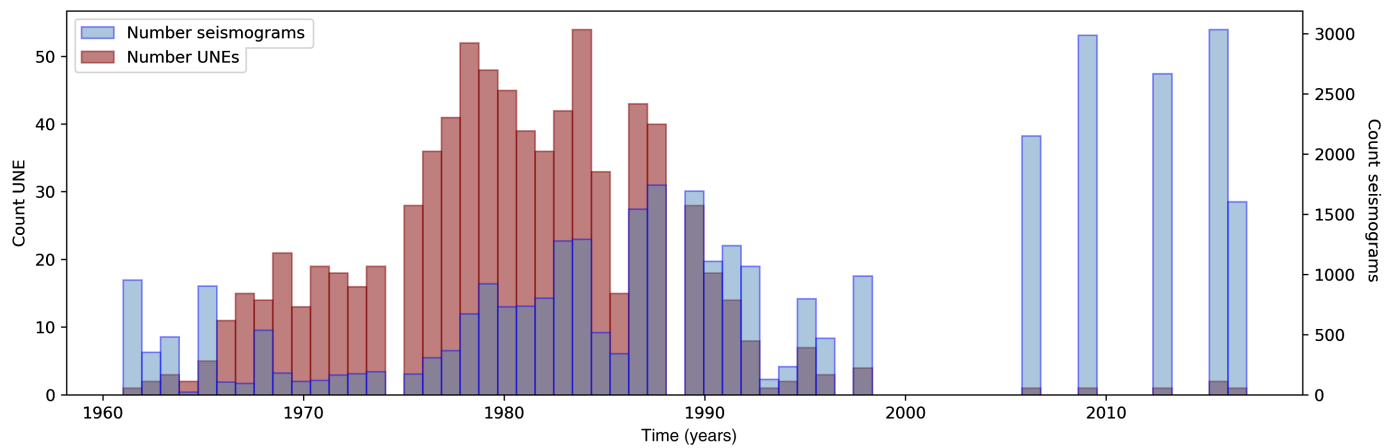
Figure 2. Count of UNE tests performed (red) and count of digital and digitized waveforms (blue) of UNE's in this dataset. The color version of this figure is available only in the electronic edition.

Recently, several global seismic datasets have been compiled by several groups (e.g. Mousavi *et al.*, 2019; Michelini *et al.*, 2021; Yeck *et al.*, 2021), and they have been extensively used by the seismological and other scientific communities to develop ML and other modern methods for seismic processing (e.g., Mousavi *et al.*, 2020). However, all of them are generated by natural earthquakes at local, regional, or teleseismic distances. Here we assembled a comprehensive dataset of all known, declassified, and globally available nuclear blast time-series data primarily for modern seismic data analysis and ML applications.

From catalogs of available declassified underground nuclear explosions (UNE), we compiled a comprehensive seismic waveform and event catalog termed GTUNE (Georgia Tech Underground Nuclear Explosions). The GTUNE dataset includes the raw seismogram time series, event coordinates and origin time, sampling rate, station metadata, channel, epicentral distance, and P-wave arrival time from the origin dataset when available and otherwise identified using a tuned automated picker. In this compiled dataset, both python algorithms to download digital times series from the Incorporated Research Institutions for Seismology (IRIS) International Federation of Digital Seismograph Networks Seismograph Networks (FDSN) webservers and the National Research Institute for Earth Science and Disaster Resilience (NIED) High Sensitivity Seismograph Network (HI-Net) array, as well as combined historical datasets are included. A labeled training dataset that includes one minute long, vertical-component seismograms of earthquake and nuclear blast P waves, and noise is also included. We envision that this dataset can be used to study deep-earth structures such as inner core scattering and rotations (e.g., Wang and Vidale, 2022) as well as event classification and discrimination studies. This is the first globally comprehensive underground nuclear blast time series dataset to be assembled for the use of seismologists and data scientists alike for modern python usage and the purpose of ML applications.

Data
Creating a comprehensive global underground nuclear blast catalog

We initially created a catalog of all UNE performed by the nuclear-weapon states, which include the United States, the former Soviet Union, United Kingdom, France, China, and more recently India, Pakistan, and Democratic People's Republic of Korea (DPRK) from 1945 to present. We compiled a catalog of confirmed declassified nuclear tests from the Sweden Defense Research Establishment and Stockholm International Peace Research Institute (SIPRI) Defense Research Establishment Division of systems and Underwater Technology's report (Bergkvist and Ferm, 2000) and U.S. Department of Energy (DOE), National Nuclear Security Administration, Nevada Field Office "United States Nuclear Tests July 1945 through September 1992" report (DOE/NV-209, 2016). These reports provide accurate event coordinates and detonation times from 1945 to 1998. The SIPRI report is a comprehensive list of all declassified global nuclear test blasts. However, for more precise origin information for the nuclear test blasts the United States is responsible for, we augmented the SIPRI catalog with the more recent and precise DOE report that includes information on all nuclear tests performed or aided by the United States and includes updated origin information. For the most recent international events not covered in either the SIPRI or DOE reports, namely the six DPRK nuclear test blasts, origin information was sourced from the International

Seismological Centre (ISC) Engdahl–van der Hilst–Buland catalog (EHB) Bulletin reports, which uniquely identify these events' origin time, location, and body-wave magnitudes (m_b), but remains a nonauthoritative source for nuclear events.

We initially translated on-screen PDF catalogs to accurate machine-readable digital tables. Although the files all have either partial optical character recognition, or direct font recognition, they are not directly computer readable text files, and require manual evaluation and correction to be converted to simple tabular formats for program readability. We developed algorithms to efficiently read through the UNE catalog, query the data center at the IRIS Data Management Center (DMC), and check and report on timing issues and data availability and to update the SIPRI catalog with origin timing, location, and yield information for each event included in the DOE report. Ultimately, a singular computer-readable catalog of all declassified nuclear events that reports the source information and country responsible for the test was compiled in a text file in the format seen in Table 1. Because above ground (and underwater) tests have relatively poor seismic coupling with the solid earth, we only include underground tests for the final digitized catalog. Underground tests are those detonated beneath the surface of the Earth and have location classification of "shaft" (bottom-drilled vertical holes, with "shaft/G": in a well, and "shaft/LG": in lagoon of an atoll), "tunnel" or "gallery" (horizontal tunnels in mountain or mesa), or "mine" (detonated in a mine) (Bergkvist and Ferm, 2000). Five events (four "surface" and one "crater") are also included that were detonated on the Earth's surface.

Of the comprehensive list of 2050 declassified nuclear tests, 802 were detonated underground, and 774 of those had at least one minute of waveform data available sampled at a minimum of 20 sps (see Table 2). There are 107 nuclear test blasts that involved more than one detonation for that location and time, differing temporally by hundredths of seconds to minutes. About 13 of those 107 had a blast detonation time difference greater than 1 s. The blasts were labeled as a single event if the detonations were within 30 s of each other, and the origin time of the first event was used for the catalog. If the difference of detonation times was 10 min or greater, each blast is considered a unique event, and not a source with multiple blasts. Following the format of the SIPRI catalog, the GTUNE catalog notes the number of blasts for each event. Event names in the catalog are defined as a 12-digit serial number comprised of the test year, month, day, hour, and minute: "YYYYMMDDHHMM." Figure 1 shows locations of underground nuclear blasts and recording stations included in this dataset.

UNE waveforms
Using the GTUNE catalog, all available digital waveforms and response files were downloaded from the FDSN webservers, which incorporates data from a global archive of national and international networks (see Data and Resources). For

TABLE 1
Georgia Tech Underground Nuclear Explosions (GTUNE) Catalog Header Columns

GTUNE evid	GTUNE identification number
Date (GMT)	GMT year month and day
Origin time (GMT)	Hour minute second and tenth of second
Berg ID number	Bergkvist Catalog identification number
Country	Country responsible for explosion
Region	Name of test site and/or geographical region
Source	Explosion reporting source
Lat	Approximate latitude
Long	Approximate longitude
m_b	Body-wave magnitude reported by source
M_s	Surface-wave magnitude reported by source
Depth (km)	Depth in kilometers
Yield lower	Lower range of yield estimate
Yield upper	Upper range of yield estimate
Purpose	Purpose of blast
Type	Method of deployment
Name	Detonation name
Number blasts/F?	Number blasts in test and/if footnote in Bergkvist cat
FSU	Blast included in FSU dataset
LLNL	Blast included in LLNL dataset
LASA	Blast included in LASA dataset

FSU, former Soviet Union; LASA, large-aperture seismic arrays; LLNL, Lawrence Livermore National Laboratory.

further completeness (particularly for the DPRK blasts), we added the Japan NIED Hi-Net (National Research Institute for Earth Science and Disaster Resilience [NIED], 2019) data that are not available through the FDSN webservices. The Python algorithm to download and reformat these seismograms into the scheme developed for other waveforms in this dataset is included in the repository. An interested researcher must first acquire individual account credentials from the NIED for access.

Our digital data fetching algorithm acquires all seismic data from stations up to 90° distance from source epicenters, includes 5 min of background signal prior to, and 30 min of data after the initial *P* arrival, from all possible station channels and location codes. To include stations that are recorded in triggered modes (i.e., not continuously), we allow for data that have gaps or overlap, and have at least 5% of the requested data length. Although it is useful to filter out stations that are part of different networks but are at the same physical station, no station distance criteria was set to ensure all available data
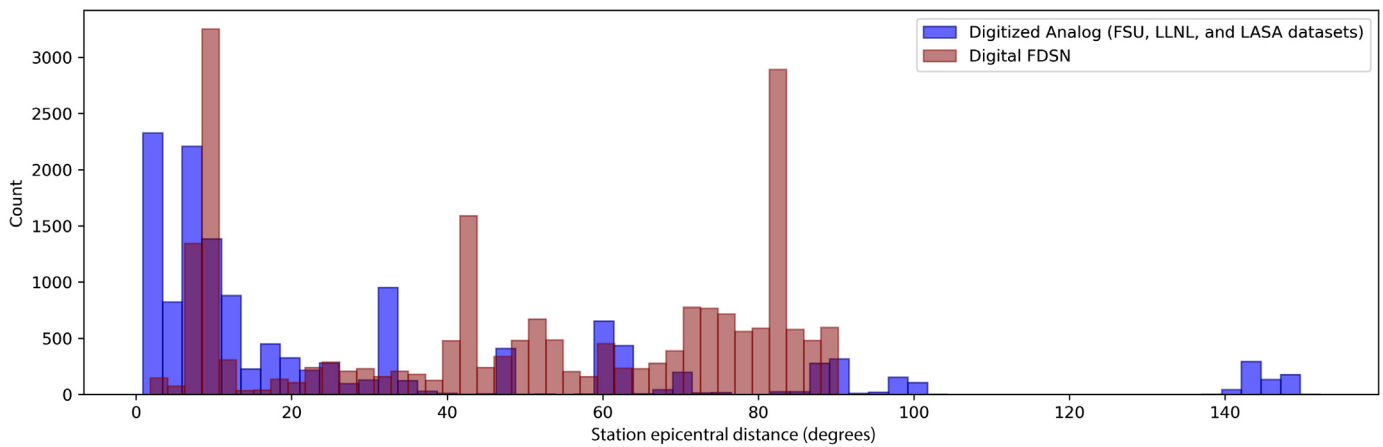
possible was collected. Our algorithm identifies which stations each data center offers, then acquires miniSEED combined waveform files and associated station XML metadata information using the ObsPy package for Python (Beyreuther *et al.*, 2010). The parameters for station distance-degrees, seismogram length, and sensor channel are editable for the user's flexibility in the provided script for digital data acquisition (IRIS query.py). For the prepared GTUNE dataset, all seismograms are cut to one-minute-long windows with the initial *P*-phase arrival occurring at 10 s.

Waveforms are available through the FDSN webservices for 538 UNE's from the year 1973 to 2017. Prepared historical datasets enable us to fill in data missing from earlier tests (1961 to 1998) that were largely originally recorded on tapes and later digitized. These include the Lawrence Livermore National Laboratory (LLNL) western United States seismic dataset (Walter *et al.*, 2004), the digitized analog seismic records from the former Soviet Union (FSU) (Richards *et al.*, 2015), and the large-aperture seismic arrays (LASA) in Montana (Capon, 1970; Vidale *et al.*, 2000; Vidale, 2021). The LASA dataset includes 2755 waveforms from 10 UNE's from 1969 to 1974. The LLNL dataset includes 2950 seismograms from 73 UNE's from 1968 to 1992 (Walter *et al.*, 2004). 5045 digitized analog waveforms from 498 UNE's from 1961 to 1999 are included from the FSU dataset (Richards *et al.*, 2015). Complete uncut waveforms and metadata from all nuclear explosion tests (underground and surface) sourced from the LLNL, LASA and FSU prepared datasets are also included in the GTUNE repository. Tables 2 and 3 summarize the waveforms and data sources that make up the GTUNE repository and Figures 2 and 3 detail the temporal distribution of data compared to total number of events and the distribution of station distance.

To work with the FSU data (Richards *et al.*, 2015), an algorithm was built to compile all necessary information including event origin time, trace start time, *P*-arrival time with respect to the origin time, sampling rate, source to station distance, station and nuclear blast coordinates, and the path to the waveform file from the existing catalog's format that follows CSS 3.0 (Center for Seismic Studies v3.0) schema tables (wfdisc, site, sitechan, assoc, and origin) of Anderson *et al.* (1990). We applied a similar method to convert the LLNL dataset into ObsPy readable formats using LLNL database client, an ObsPy client for the LLNL DB database (Alvizuri and Tape, 2018), and then converted these to the format of our training labels. Although not all of this data has a high signal-to-noise ratio (SNR), there are still many usable waveforms from this set that will be useful for training data or seismic studies.

Earthquake and seismic noise
An earthquake and seismic noise training dataset was developed (Fig. 4; Table 3) using the data fetching algorithms used for UNEs and included in the repository. Our earthquake dataset includes waveforms from a global catalog of earthquakes

Figure 3. Distribution of station and blast epicentral distance in the GTUNE underground dataset. Blue denotes station distribution for digitized analog seismograms from the large-aperture seismic arrays (LASA), LLNL, and FSU prepared datasets. Red denotes station distribution for digital data from the FDSN webservers. The color version of this figure is available only in the electronic edition.

from 1 January 2000 to 1 January 2020, with depths shallower than 50 km, and magnitudes ranging from 4.5 to 6.5 to reflect typical UNE blast magnitudes. In addition, we removed events that occur within 30 min of each other to ensure the *P* arrival is not mixed with phases from other events, and used only waveforms with an SNR greater than 2. Using ObsPy's MassDownloader module, all available vertical seismograms from all data centers that implement the FDSN webservices were downloaded. A total of 5-min-long 127,472 vertical-component waveforms from the 25,524 shallow earthquakes were downloaded and included in the dataset.

To build the noise dataset free of any earthquake seismic phases, we used stations from the relatively seismically quiet Central and Eastern United States (CEUS). This region likely represents the highest density of suitable and globally available seismic sensors in a region that is both sufficiently seismically quiet and well cataloged for regional events that do occur. The primary difficulty with choosing a global distribution of stations for noise labels is that many more sparsely distributed sites

(e.g., island sites near volcanism or subduction zones) will likely have substantial small nearby earthquake activity that is un or undercataloged. As such, it becomes a logistical nightmare to ensure that the developed noise catalog is free from seismic sources. Nevertheless, we included some Python algorithms that users can use to build their own "earthquake free" noise dataset. In addition, they can import noise datasets from other publicly available seismic labels such as STanford EArthquake Dataset (STEAD) (Mousavi *et al.*, 2019) or the Italian seismic dataset for machine learning INSTANCE (Michelini *et al.*, 2021).

Using an earthquake catalog that includes all events within a 60° radius of CEUS, we excluded any earthquake that has an origin time occurring within 25 min of another event to ensure that our noise dataset does not include any phases from known local or regional seismic events. The catalog for regional earthquakes was downloaded from the ISC Bulletin (International Seismological Centre, 2000–2099), and includes earthquakes of all magnitudes from 1 June 2011 to 1 June 2016, in the rectangular region bounded by latitudes from 15° to 55° in the northern hemisphere, and longitudes from −130° to −52° in the western hemisphere. The traces within our noise dataset

TABLE 2
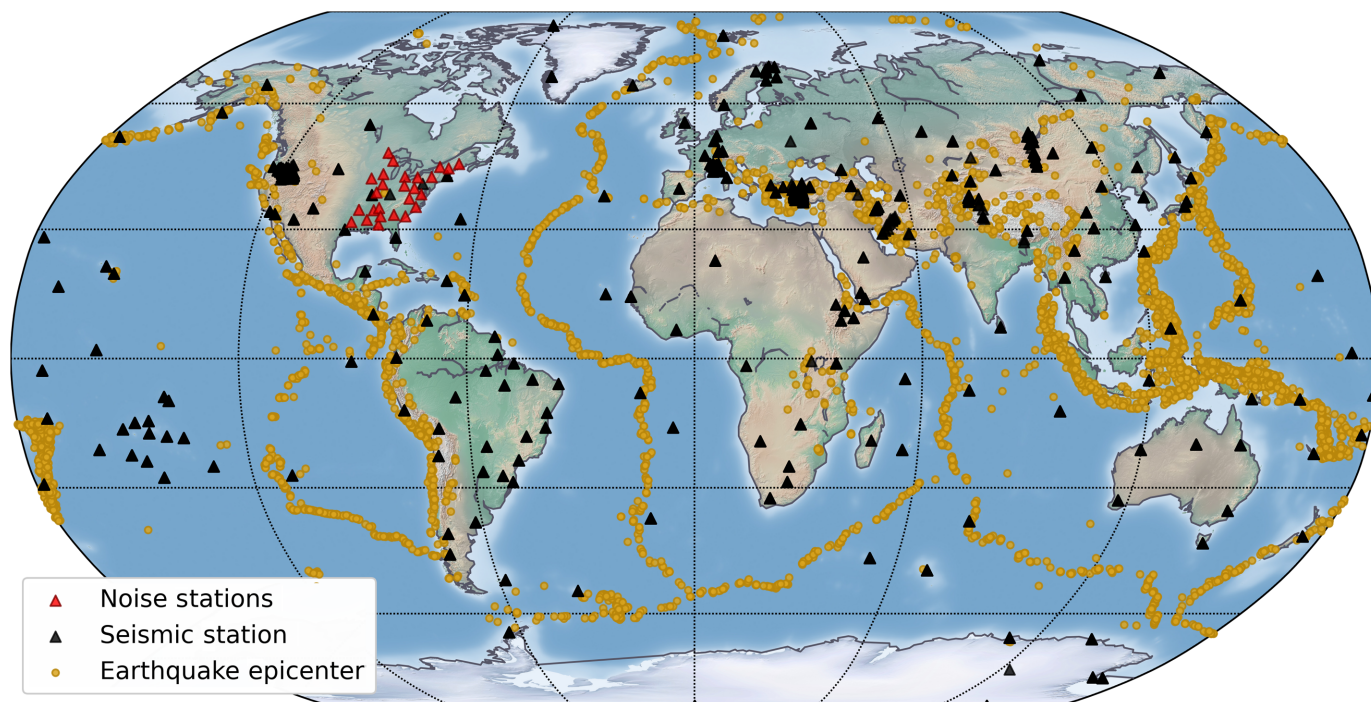**Total Underground Nuclear Explosions and Associated Waveforms in the Prepared Dataset**

Data Source	Number of Explosions	Number of Waveforms
FDSN query on catalog	538	12,733
FSU digitized seismic records	498*	5045
LLNL dataset	73	2,950
LASA	10	2,755
NIED HI-Net data	6	4,640
Total GTUNE data	774	28,123

FDSN, International Federation of Digital Seismograph Networks; HI-Net, High Sensitivity Seismograph Network; NIED, National Research Institute for Earth Science and Disaster Resilience.

*The FSU prepared data in the repository includes 789 nuclear blasts and 8236 waveforms from the FSU dataset, however only but 498 of these are underground blasts that fit the GTUNE criteria.

TABLE 3
Summary of Earthquake and Noise Labeled Waveforms

Data Source	Data Type	Number of Waveforms
FDSN query on catalog	Earthquakes	127,472
	Noise	20,833

Figure 4. Global map of earthquake locations (yellow circles) and seismic stations (black triangles) included in the dataset, occurring between 1 January 2000 and 1 January 2020, ranging in magnitude between 2 and 6.5. Noise stations (red triangles) across the eastern United States between 1 June 2011 and 1 June 2016, avoiding periods where known earthquakes, both regionally and globally, may have contaminated signal (stations are red triangles). The color version of this figure is available only in the electronic edition.

likely include earthquake phases from regional or distant earthquakes that are not listed in the ISC bulletin. However, because there are earthquakes occurring around the world almost at all times, seismic phases from distant sources are likely within most seismic data. These distant phases have lower frequencies and should not interfere with the higher frequency *P*-wave phases that many seismic studies focus on. We did not remove time windows of predicted arrivals from teleseismic seismic events during this time period.

P-phase arrivals

Because waveforms of explosions often have much less discernible *S*-wave arrival as compared to comparably energetic earthquake sources (Murphy *et al.*, 2009; Walter *et al.*, 2018), and many older recordings were only capturing vertical signals, we only included *P*-wave phase arrival determinations in our prepared UNE dataset. All *P*-phase arrivals were used from the associated meta data of the source dataset, when available or sourced from the ISC-EHB Bulletin catalog (Engdahl *et al.*, 1998). The source of the *P* phase for each waveform is included in the dataframes of the prepared data files (see Table 4). For the digital FDSN acquired waveforms or when *P* phases are not available in the prepared dataset, we apply the Baer Picker method (pk Baer in ObsPy; Baer and Kradolfer 1987; Beyreuther *et al.*, 2010). This algorithm determines the approximate onset times using an automatic detection built to detect both teleseismic and local *P*-wave arrivals, first motion, and relative reliability of the pick while ignoring noise and transient events. The Baer picker method works well with the included signals. Multiple-component data should be added later, the method could use all components for phase determination.

For more accurate *P*-wave determinations and to decrease processing time, we narrowed the allowable window for the algorithm to run, including 5 s before and 15 s after (to allow for delays in the real phase arrival due to heterogeneities in Earth) the theoretically predicted *P*-wave arrival (using the IASPEI-91 velocity model; Engdahl *et al.*, 1998; Crotwell *et al.*, 1999; Snoke, 2009). An example of Baer *P*-wave picks on regional to teleseismic raw seismograms from the 1992 China Lop Nor nuclear test is shown in Figure 5. Our adaptation of the *P*-wave picker algorithm reports the result in an output list that includes the labels for the nuclear explosion *P* wave to be converted to an array (Python NumPy). This list includes for each waveform, the explosion origin timestamp in seconds (UTC epoch time), the timestamp of beginning of waveform trace (same format as origin timestamp), the calculated *P*-wave arrival time in seconds with respect to the origin time, the predicted *P*-wave arrival in seconds with respect to the origin time, sampling rate of waveform trace, distance from explosion to seismic station in degrees, station latitude, station longitude, explosion latitude, explosion longitude, the path to the waveform data file. Phase arrivals

TABLE 4
Prepared Datasets Dataframe Column Format

Evid	GTUNE identification number
Origin time	GMT epoch time
Trace start time	GMT epoch time
P arrival	GMT epoch time
Phase	Phase name
Source	Source of phase pick
Sampling rate	Samples per second
Station distance	Station epicentral distance (degrees)
Net	Network
Chn	Channel
Station	Station name
Stla	Station latitude
Stlo	Station longitude
Evla	Event latitude
Evlo	Event longitude
Mag	Magnitude
Magtype	Magnitude type
Waveform	Time series (NumPy array)

provided in the LLNL, LASA, and FSU datasets, were utilized when available. The source of the *P* phases are included as metadata in the prepared datasets (see Table 4).

Summary and Catalog Use Case

All the prepared datasets are in the format of a compiled table (Pandas' dataframe saved in hdf5 and pickle formats) in which each row contains the event information, station metadata, *P*-arrival time, and time-series array of the raw seismogram. Table 4 details the format of the prepared data sets.

For all UNE waveforms sourced from the LLNL, LASA, or FSU datasets, the seismogram is uncut (full length of the available seismogram). For all data sourced from FDSN webservers all seismograms are cut to a one-minute trace, with the *P* wave set at 10 s, (see Fig. 6 for cut window examples). No preprocessing of the waveforms was done. Because instrument response is not available for all UNEs in the dataset, instrument response was not removed from waveforms.

All the available metadata and catalog origin information for the GTUNE blasts and earthquakes are available within the repository as text files (extensions: .txt). All scripts are written to run in Python 3.8 (extension: .py). All waveform data in the format of time-series arrays within the dataframe table that are saved as a flattened or serialized pickle (a Python format designed for backward compatibility) or hdf5 (binary file format for storage of large scientific data sets) files, extensions: .pkl and .hdf5, respectively. Because of space limitations, we include all Python query algorithms for the user to download all available earthquake data from FDSN webservers. Table 4 details the format for each individual column of the prepared earthquake and UNE data.

The most substantial limitation in developing a rich catalog from the data available is that most nuclear tests occurred before the substantial global growth of digital seismic recordings beginning in the 1980s (Fig. 2), and that the majority of these historical seismograms are recorded on more regional networks. Therefore, the final dataset is highly unbalanced in the number of earlier analog (fewer waves for many events) versus modern digital waveforms (many waves from relatively few events). In addition, the combination of prepared and historical datasets with the FDSN and Hi-Net available data gives a large spread in epicentral distance degrees, with a slight bias to regional distances less than 20° (see Fig. 5).

Although the prepared labeled dataset includes a combination of both originally digital and analog seismograms, traces that have poor SNR and relatively small training dataset compared to big data or ML studies, we found it to be robust for training a Convolutional Neural Network (CNN) classifier for automatic identification of nuclear blasts, from earthquakes and background noise, in continuous traces from stations at regional and teleseismic distances (Barama *et al.*, 2020). The trained CNN was highly capable of classifying signals curated in the format of GTUNE datasets, that is, window size and positioned event arrival. A more detailed characterization of this discriminator method is forthcoming, and the associated refined dataset and training algorithms will be similarly released. We expect that other seismologists and data scientists will find this comprehensive and labeled dataset to equally be fruitful for advanced computational methods, including ML.

Data and Resources

The GTUNE data repository is available on Zenodo (www.zenodo.org), the open-access repository for data, Zenodo, doi: 10.5281/zenodo.7026463. See Figure 7 for a schematic of the repository directory. Seismic waveform data and station metadata from digital seismometers used in this study are available from the Incorporated Research Institutions for Seismology Data Management Center (IRIS-DMC; www.fdsn.org/webservices) and National Research Institute for Earth Science and Disaster Resilience (NIED) High Sensitivity Seismograph Network (Hi-Net) array (NIED, 2019). All digitized analog waveforms come from the assembled data sets of Walter *et al.* (2004), Richards *et al.* (2015), and Vidale (2021). A complete list of data centers can be found at www.fdsn.org/webservices/datacenters/. All websites were last accessed in August 2022.

Declaration of Competing Interests

The authors acknowledge that there are no conflicts of interest recorded.
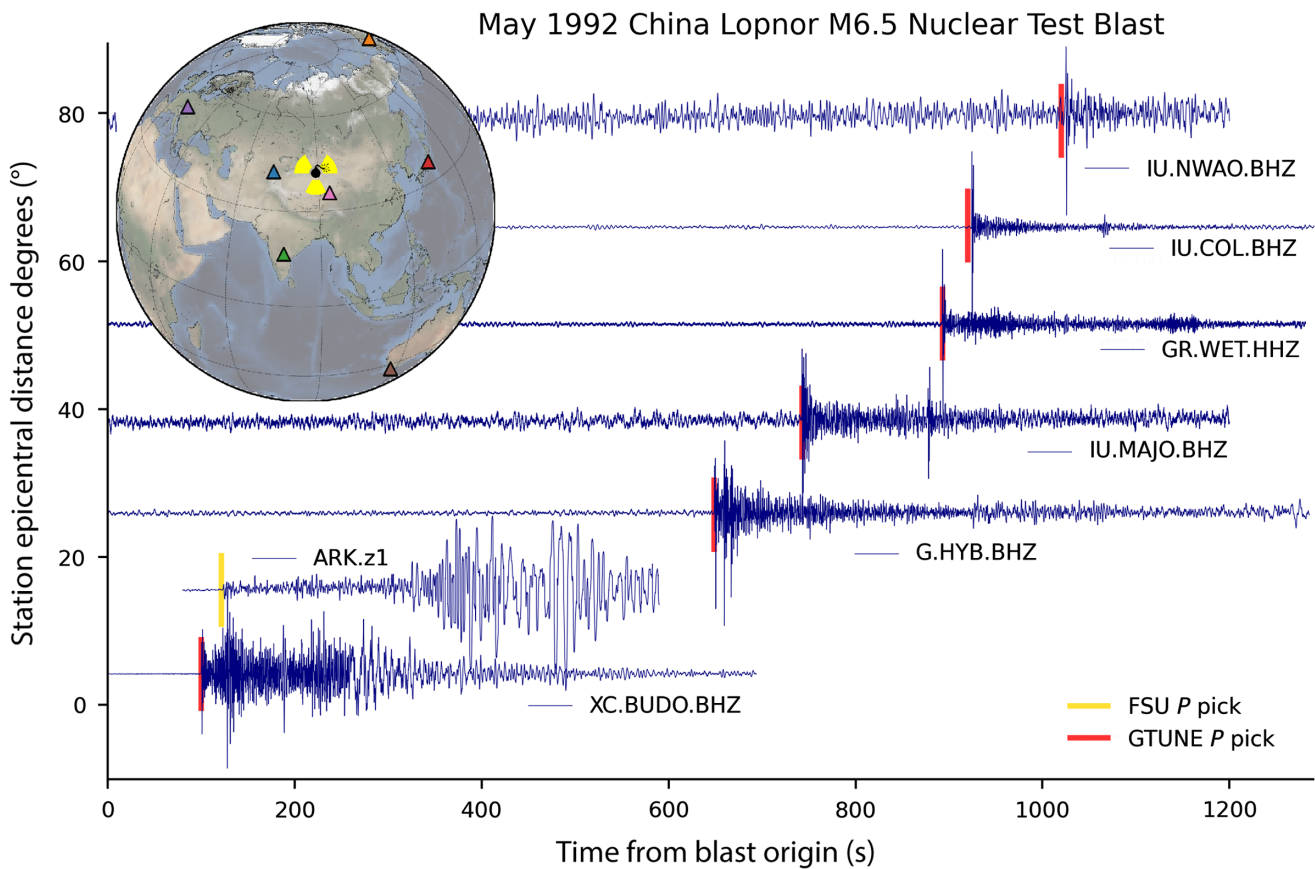
Figure 5. Selected seismograms from the May 1992 Chinese Lop Nor nuclear test blast. Waveforms shown are from the GTUNE FDSN data and the s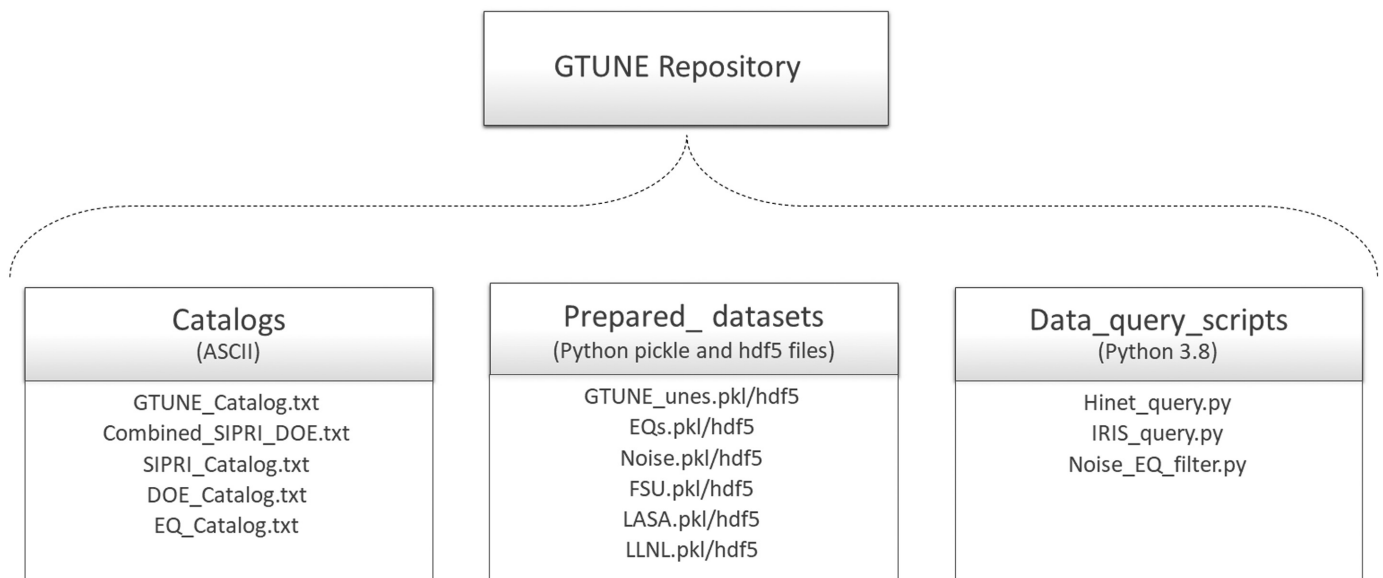ingle waveform available from the FSU dataset for this event. The red and yellow lines indicate *P*-arrival picks from the FSU dataset and GTUNE Baer algorithm picks, respectively. The color version of this figure is available only in the electronic edition.
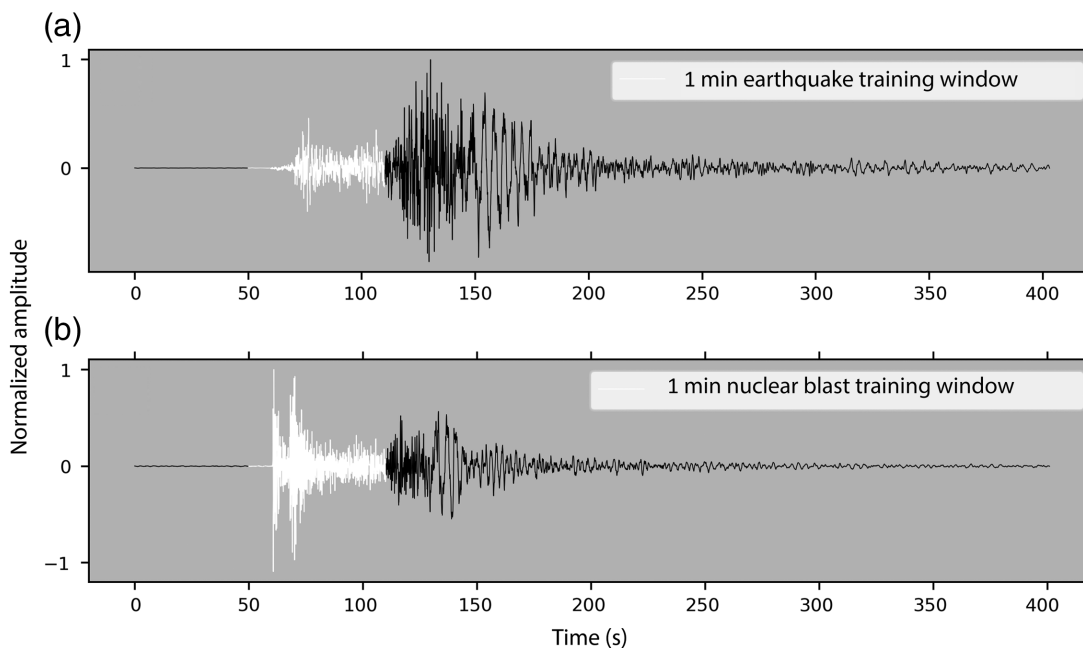
Figure 6. Example of one-minute windows of raw seismic data included in the compiled dataset, including 10 s before, and 50 s after the *P* arrival. Comparison of two vertical seismograms from station IC.MDJ of (a) a 2018 magnitude 5.2 earthquake (distance = 3.57°) and (b) the 2016 magnitude 5.1 nuclear test (distance = 3.34°) from the Democratic People's Republic of Korea (DPRK).

```
                    GTUNE Repository
```

Catalogs (ASCII)	Prepared_ datasets (Python pickle and hdf5 files)	Data_query_scripts (Python 3.8)
GTUNE_Catalog.txt Combined_SIPRI_DOE.txt SIPRI_Catalog.txt DOE_Catalog.txt EQ_Catalog.txt	GTUNE_unes.pkl/hdf5 EQs.pkl/hdf5 Noise.pkl/hdf5 FSU.pkl/hdf5 LASA.pkl/hdf5 LLNL.pkl/hdf5	Hinet_query.py IRIS_query.py Noise_EQ_filter.py

Figure 7. Schematic of GTUNE repository directory structure. All header information is included in the catalog text files. The length of waveforms in the prepared GTUNE.pkl, EQ.pkl, and Noise.pkl datasets are set at 1 min long. The length of the waveforms in the FSU LASA llnl.pkl datasets are uncut, retaining the length of the waveform in the respective source dataset.

Acknowledgments

References

Alvizuri, C., and C. Tape (2018). Full moment tensor analysis of nuclear explosions in North Korea, *Seismol. Res. Lett.* **89,** no. 6, 2139–2151.

Anderson, J., W. Farrell, K. Garcia, J. Given, and H. Swanger (1990). Center for seismic studies version 3 database: Schema reference manual, *CSS Technical Rept. C90-01,* 5–66.

Baer, M., and U. Kradolfer (1987). An automatic phase picker for local and teleseismic events, *Bull. Seismol. Soc. Am.* **77,** no. 4, 1437–1445.

Barama, L., J. Williams, Z. Peng, and A. V. Newman (2020). Nuclear blast discrimination using a convolutional neural network, *AGU Fall Meeting Abstracts,* Virtual Online Conference, December 2020, S053–0009.

Bennett, T. J., V. Oancea, B. W. Barker, Y.-L. Kung, M. Bahavar, B. C. Kohl, J. R. Murphy, and I. K. Bondár (2010). The Nuclear Explosion Database (NEDB): A new database and web site for accessing nuclear explosion source information and waveforms, *Seismol. Res. Lett.* **81,** no. 1, 12–25, doi: 10.1785/gssrl.81.1.12.

Bent, A. L., D. I. Doser, and L. J. Hwang (2020). Preface to the focus section on historical seismograms, *Seismol. Res. Lett.* **91,** no. 3, 1356–1358.

Bergkvist, N.-O., and R. Ferm (2000). Nuclear explosions 1945–1998, *Defence Research Establishment Division of Systems and Underwater Technology SE-172,* Stockholm, Sweden, 90 pp.

Beyreuther, M., R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann (2010). ObsPy: A python toolbox for seismology, *Seismol. Res. Lett.* **81,** no. 3, 530–533.

Capon, J. (1970). Analysis of Rayleigh-wave multipath propagation at LASA, *Bull. Seismol. Soc. Am.* **60,** no. 5, 1701–1731.

Crotwell, H. P., T. J. Owens, and J. Ritsema (1999). The taup toolkit: Flexible seismic travel-time and ray-path utilities, *Seismol. Res. Lett.* **70,** no. 2, 154–160.

Del Pezzo, E., A. Esposito, F. Giudicepietro, M. Marinaro, M. Martini, and S. Scarpetta (2003). Discrimination of earthquakes and underwater explosions using neural networks, *Bull. Seismol. Soc. Am.* **93,** no. 1, 215–223.

Dickey, J., B. Borghetti, W. Junek, and R. Martin (2020). Beyond correlation: A path- invariant measure for seismogram similarity, *Seismol. Res. Lett.* **91,** no. 1, 356–369.

DOE/NV-209 (2016). United States nuclear tests July 1945 through September 1992, *U.S. Department of Energy, National Nuclear Security Administration,* Nevada Field Office.

Engdahl, E. R., R. van der Hilst, and R. Buland (1998). Global teleseismic earthquake relocation with improved travel times and procedures for depth determination, *Bull. Seismol. Soc. Am.* **88,** no. 3, 722–743.

International Seismological Centre (2000–2099). ISC-EHB dataset, doi: 10.31905/PY08W6S3.

Ishii, M., H. Ishii, B. Bernier, and E. Bulat (2015). Efforts to recover and digitize analog seismograms from Harvard-Adam Dziewoński observatory, *Seismol. Res. Lett.* **86,** no. 1, 255–261.

Kim, W.-Y., and G. Ekström (1996). Instrument responses of digital seismographs at Borovoye, Kazakhstan, by inversion of transient calibration pulses, *Bull. Seismol. Soc. Am.* **86,** no. 1A, 191–203.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning, *Nature* **521,** no. 7553, 436–444.

Maceira, M., P. S. Blom, J. K. MacCarthy, O. E. Marcillo, G. G. Euler, M. L. Begnaud, S. R. Ford, M. E. Pasyanos, G. J. Orris, M. P. Foxe, *et al.* (2017). Trends in nuclear explosion monitoring research and development-a physics perspective (tech. rep.), Los Alamos National Laboratory (LANL), Los Alamos, New Mexico.

Michelini, A., S. Cianetti, S. Gaviano, C. Giunchi, D. Jozinović, and V. Lauciani (2021). Instance—The Italian seismic dataset for machine learning, *Earth Syst. Sci. Data* **13,** 5509–5544, doi: 10.5194/essd-13-5509-2021.

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

Mousavi, S. M., Y. Sheng, W. Zhu, and G. C. Beroza (2019). Stanford earthquake dataset (stead): A global data set of seismic signals for AI, *IEEE Access*, doi: 10.1109/ACCESS.2019.2947848.

Murphy, J., B. Barker, D. Sultanov, and O. Kuznetsov (2009). S-wave generation by underground explosions: Implications from observed frequency-dependent source scaling, *Bull. Seismol. Soc. Am.* **99,** no. 2A, 809–829.

Nakano, M., D. Sugiyama, T. Hori, T. Kuwatani, and S. Tsuboi (2019). Discrimination of seismic signals from earthquakes and tectonic tremor by applying a convolutional neural network to running spectral images, *Seismol. Res. Lett.* **90,** no. 2, 530–538.

National Research Institute for Earth Science and Disaster Resilience (NIED) (2019). NIED HI-Net, doi: 10.17598/NIED.0003.

Okal, E. A. (2015). Historical seismograms: Preserving an endangered species, *Geo. Res. J.* **6,** 53–64.

Richards, P. G., and M. Hellweg (2020). Challenges and opportunities in turning large US archives of analog seismograms into a modern usable resource, *Seismol. Res. Lett.* **91,** no. 3, 1531–1541.

Richards, P. G., W.-Y. Kim, and G. Ekström (1992). The Borovoye geophysical observatory, Kazakhstan, feature article, *Eos Trans. AGU* **73,** no. 201, 205–206.

Richards, P. G., W.-Y. Kim, I. N. Sokolova, and N Mikhailova (2015). Digitization of nuclear explosion seismograms from the former Soviet Union, *Geology* doi: 10.21236/ada618975.

Snoke, J. A. (2009). Traveltime tables for iasp91 and ak135, *Seismol. Res. Lett.* **80,** no. 2, 260–262.

Song, X., and P. G. Richards (1996). Seismological evidence for differential rotation of the earth's inner core, *Nature* **382,** no. 6588, 221–224.

Vidale, J. E. (2021). LASA data, available at https://github.com/JohnVidale/LASA_data.git (last accessed December 2021).

Vidale, J. E., D. A. Dodge, and P. S. Earle (2000). Slow differential rotation of the earth's inner core indicated by temporal changes in scattering, *Nature* **405,** no. 6785, 445–448.

Walter, W., D. A. Dodge, G. Ichinose, S. C. Myers, M. E. Pasyanos, and S. R. Ford (2018). Body-wave methods of distinguishing between explosions, collapses, and earthquakes: Application to recent events in North Korea, *Seismol. Res. Lett.* **89,** no. 6, 2131–2138.

Walter, W., K. Smith, J. O'Boyle, T. Hauk, F. Ryall, S. Ruppert, S. Myers, R. Abbot, and D. Dodge (2004). *An Assembled Western United States Dataset for Regional Seismic Analysis*, Lawrence Livermore National Laboratory, Livermore, California.

Wang, W., and J. E. Vidale (2022). An initial map of fine-scale heterogeneity in the Earth's inner core, *Nat. Geosci.* **15,** 240–244, doi: 10.1038/s41561-022-00903-8.

Yeck, W. L., J. M. Patton, Z. E. Ross, G. P. Hayes, M. R. Guy, N. B. Ambruz, D. R. Shelly, H. M. Benz, and P. S. Earle (2021). Leveraging deep learning in global 24/7 real-time earthquake monitoring at the national earthquake information center, *Seismol. Res. Lett.* **92,** 469–480, doi: 10.1785/0220200178.