6-2020

Neural Network Models for Nuclear Treaty Monitoring: Enhancing the Seismic Signal Pipeline with Deep Temporal Convolution

Joshua T. Dickey

NEURAL NETWORK MODELS FOR NUCLEAR TREATY MONITORING: ENHANCING THE SEISMIC SIGNAL PIPELINE WITH DEEP TEMPORAL CONVOLUTION

DISSERTATION

Joshua T. Dickey, DAFC

AFIT-ENG-DS-20-J-004

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-DS-20-J-004

NEURAL NETWORK MODELS FOR NUCLEAR TREATY MONITORING: ENHANCING THE SEISMIC SIGNAL PIPELINE WITH DEEP TEMPORAL CONVOLUTION

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Degree of Doctor of Philosophy

Joshua T. Dickey, B.S., M.S.

DAFC

July 2020

AFIT-ENG-DS-20-J-004

NEURAL NETWORK MODELS FOR NUCLEAR TREATY MONITORING: ENHANCING THE SEISMIC SIGNAL PIPELINE WITH DEEP TEMPORAL CONVOLUTION

Joshua T. Dickey, B.S., M.S.
DAFC

Committee Membership:

Brett J. Borghetti, PhD
Chairman

James C. Petrosky, PhD
Member

Richard K. Martin, PhD
Member

William N. Junek, PhD
Member

Andrew S. Keys, PhD
Dean's Representative

ADEDJI B. BADIRU, PhD
Dean, Graduate School of Engineering and Management

AFIT-ENG-DS-20-J-004

Abstract

The Comprehensive Nuclear-Test-Ban Treaty dictates a de-facto moratorium on the testing of nuclear weapons. To enforce this treaty, a strict verification regime was established consisting primarily of a global geophysical sensor network and a sophisticated data processing center. Because the majority of covert nuclear tests are conducted underground, preliminary verification often involves the processing of seismic signals. This dissertation begins with a brief consideration of the current seismic signal processing pipeline for treaty monitoring, and then proceeds to detail three research studies, utilizing deep neural network architectures to address four prominent tasks in the pipeline: signal detection, event association, event localization and source discrimination.

Study 1 focuses on the signal detection task. The detection of seismic events at regional and teleseismic distances is critical to nuclear treaty monitoring. Traditionally, detecting regional and teleseismic events has required the use of an expensive multi-instrument seismic array; however in this study, we present DeepPick, a novel seismic detection algorithm capable of array-like detection performance from a single-trace. We achieve this performance by training a deep temporal convolutional neural network detector against the arrival times in an array-beam catalog and the single-trace waveforms taken from the vertical channel of the center element of the array. The training data consists of all arrivals in the International Seismological Centre Database for seven seismic arrays over a five year window from 1 Jan 2010 to 1 Jan 2015, yielding a total training set of 608,362 detections. The test set consists of the same seven arrays over a one year window from 1 Jan 2015 to 1 Jan 2016. We report our results by training the algorithm on six of the arrays and testing it on the seventh,

so as to demonstrate the generalization of the technique to new stations. Detection performance against this test set is outstanding. Fixing a type-I error (false positive) rate of 0.1%, the algorithm achieves an overall recall (true positive rate) of 57.8% on the 141,095 array beam picks in the test set, yielding 81,524 correct detections. This represents a 40% increase in performance over state-of-the-art kurtosis-based detectors, and is more than twice the 37,572 detections made by a state-of-the-art frequency-band detector over the same period. Furthermore, DeepPick provides a 4 dB improvement in detector sensitivity over all other current methods tested, with a run-time that is an order of magnitude faster. These results demonstrate the potential of our algorithm to significantly enhance the effectiveness of the global treaty monitoring network.

Study 2 focuses jointly on both event association and source discrimination, utilizing a learned similarity measure to extract source-specific features from three-component seismograms. Similarity search is a popular technique for seismic signal processing, with template matching, matched filters and subspace detectors being utilized for a wide variety of tasks, including both signal detection and source discrimination. Traditionally, these techniques rely on the cross-correlation function as the basis for measuring similarity. Unfortunately, seismogram correlation is dominated by path effects, essentially requiring a distinct waveform template along each path of interest. To address this limitation, we propose a novel measure of seismogram similarity that is explicitly invariant to path. Using Earthscope's USArray experiment, a path-rich dataset of 207,291 regional seismograms across 8,452 unique events is constructed, and then employed via the batch-hard triplet loss function, to train a deep convolutional neural network which maps raw seismograms to a low dimensional embedding space, where nearness on the space corresponds to nearness of source function, regardless of path or recording instrumentation. This path-agnostic embedding space

forms a new representation for seismograms, characterized by robust, source-specific features, which we show to be useful for performing both pairwise event association as well as template-based source discrimination with a single template.

Study 3 focuses on event localization and backazimuth prediction. Single-station location estimates are traditionally limited to array stations, where beamforming provides high-confidence backazimuth prediction. Three-component stations, on the other hand, rely on polarization analysis for backazimuth prediction, which suffers from both high error and low confidence. In this study, we present BAZNet, a deep neural-network-based backazimuth predictor for three-component stations. For existing stations with ample historical training data, the technique achieves an overall median absolute error of around 14°, a modest improvement over polarization. More importantly, each estimate is accompanied by a robust certainty measure, allowing the selection of only high-confidence predictions to be passed on to the associator. Using this certainty measure, roughly 60% of all predictions can be selected, with an accuracy on par with beamforming. This represents a seven-fold improvement over the 8% of signals similarly selectable via polarization. To demonstrate BAZNet, we use 10 years of waveform data from 561,154 cataloged arrivals across 9 stations selected from the global IMS Network.

Seismic signal processing is critical to the verification of the Comprehensive Nuclear Test Ban Treaty, facilitating the detection and identification of covert nuclear tests in near-real time. The three studies in this dissertation provide substantial enhancements to this processing pipeline. Study 1 details a new methodology for the detection and arrival time estimation of regional and teleseismic signals, effecting a 4 dB increase in detector sensitivity over the latest operational methods. Study 2 details a novel representation space for seismograms, with applications both as a complimentary validation measure for event association and as a one-shot classifier

for template-based source discrimination. Finally, Study 3 details a new method for predicting backazimuth angle, providing a seven-fold increase in usable picks over traditional polarization analysis.

Acknowledgements

This dissertation would not have been possible without the sponsorship of the Air Force Technical Applications Center (AFTAC) Systems Development Directorate, and special thanks goes to Director David Merker, SES for his support and endorsement. Thanks also goes to Eli Baker of the Air Force Research Lab, as well as Dr. Gregory Wagner, Jorge Ramon-Nieves, Chris Barbour, and Vince Gillan of AFTAC, for their patience and guidance during the seismic analysis process. The dataset used in this dissertation was only made available by the considerable efforts of Alan Poffenburger and Judy Wheeler of the United States National Data Center. Major credit goes to my research committee: Dr. Richard Martin for his seemingly limitless signal-processing expertise; Dr. James Petrosky for his insightful and timely feedback; and especially Dr. Bill Junek for his unwavering mentor-ship, encouragement and guidance. Finally, I would like to thank my Research Advisor, Dr. Brett Borghetti, for his countless hours of support and direction over the last three years. His influence and instruction are responsible not only for the content of this dissertation, but also for the ongoing content and caliber of my career as a research scientist and engineer.

Joshua T. Dickey

Table of Contents

Page

List of Figures

List of Tables

List of Abbreviations

Abbreviation		Page
AFTAC	Air Force Technical Applications Center	viii
CTBT	Comprehensive Nuclear-Test-Ban Treaty	1
CTBTO	Comprehensive Nuclear-Test-Ban Treaty Organization	1
IMS	International Monitoring System	2
IDC	International Data Centre	2
3C	Three-Component	6
SNR	Signal-to-Noise Ratio	6
STA/LTA	Short-Term Average, Long-Term Average	7
AIC	Akaike Information Criterion	7
TDOA	Time Delay of Arrival	11
TCN	Temporal Convolutional Neural Networks	12
CNN	Convolutional Neural Networks	13
CF	Characteristic Function	25
HOS	Higher-Order Statistics	26
TA	USArray Transportable Array	68
CONUS	Continental US	69
ROC	Receiver Operating Characteristic	72
AUC	Area Under Curve	72
t-SNE	T-Distributed Stochastic Neighbor Embedding	75
USGS	United States Geological Survey	81
IRIS	Incorporated Research Institutions for Seismology	82
ISC	International Seismological Centre	82

NEURAL NETWORK MODELS FOR NUCLEAR TREATY MONITORING: ENHANCING THE SEISMIC SIGNAL PIPELINE WITH DEEP TEMPORAL CONVOLUTION

I. Introduction

1.1 Nuclear Treaty Monitoring

In 1945, the US unleashed the most potent explosive attack in the history of mankind, dropping two 20-kiloton nuclear bombs on Japan, resulting in the instant annihilation of sixty thousand people, and a final death toll of nearly 150,000 [12]. Fortunately, the cost of developing such weapons is commensurate with their power, requiring billions of dollars in research and extensive testing in order to obtain. Unfortunately, many countries have been willing to pay this price, and in the years from 1945 through 1996, more that 2000 nuclear tests were conducted, primarily by the US, Russia, France, the UK and China, resulting in the establishment of five recognized nuclear powers, and a near half-century long cold war between the US and Russia [30].

The Comprehensive Nuclear Test Ban Treaty.

Throughout the cold war, many attempts were made to halt the seemingly rapid proliferation of nuclear weapons, most notably by the negotiation of a comprehensive ban on the testing of such weapons. Finally, in 1996, the Comprehensive Nuclear-Test-Ban Treaty (CTBT) was opened for signature and the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) was established,

creating a de-facto moratorium on the testing of nuclear weapons. Of course, enforcing such a treaty requires a strict verification regime: a system for promptly detecting any and all nuclear tests. The verification regime established by the CTBTO consists primarily of a global geophysical sensor network, the International Monitoring System (IMS), and a sophisticated data processing center, the International Data Centre (IDC). Since its inception, the CTBTO has used this verification regime to detect and identify 8 illicit nuclear tests by 3 countries: India (1998), Pakistan (1998) and North Korea (2006, 2009, 2013, 2016, 2016 and 2017) [73].

The International Monitoring System Network.

Because the majority of illicit nuclear tests are conducted underground, one of the primary verification technologies employed by the IMS is a global network of seismometers which monitor shockwaves in the earth. The network includes 50 primary and 120 auxiliary seismic stations, some of which are lone three-channel instruments, and others which are regional arrays of seismometers. The locations of the primary sensors are presented in Fig. 1. Additionally, the IMS Network also includes 11 hydroacoustic stations, 60 infrasound stations and 80 radionuclide stations [27].

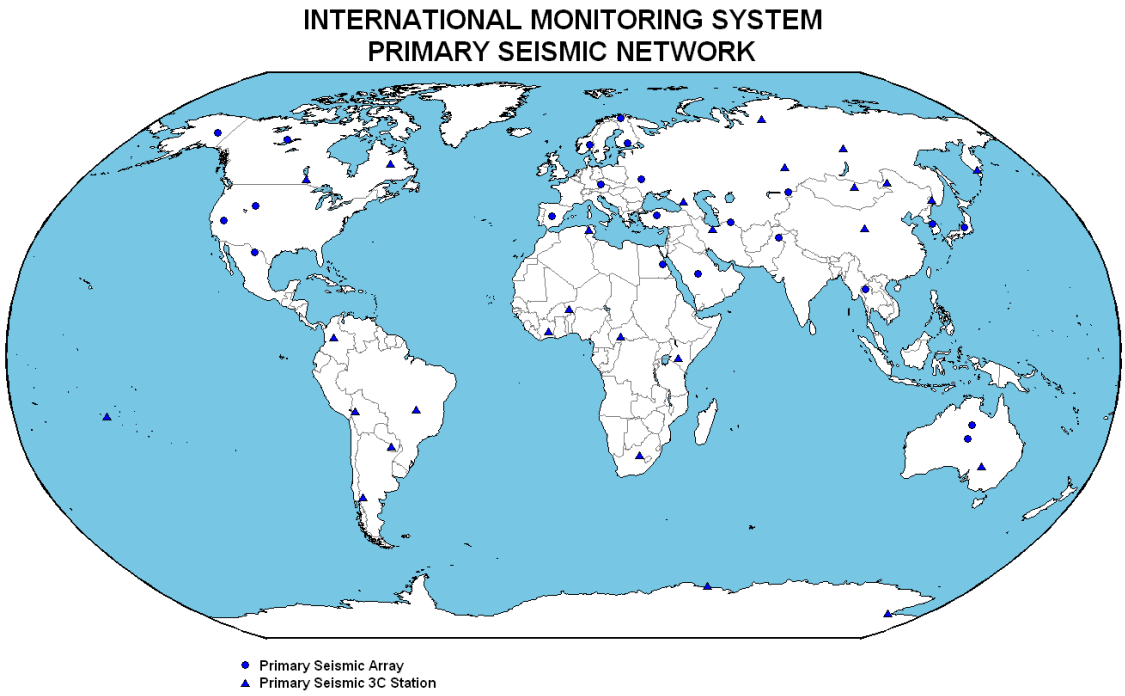

Figure 1. Global Map detailing the IMS Primary Seismic Sensor Network [8].

The International Data Centre.

In total, the IMS consists of more than 300 sensor stations around the globe, many of which report continuous data streams back to the CTBTO's Vienna headquarters in near real-time. Processing and storing this data is accomplished by the IDC, which stores the incoming data, processes it and ultimately reports any verified nuclear tests within 2 hours of occurrence. A basic outline of the IDC is shown in Fig. 2 [27].

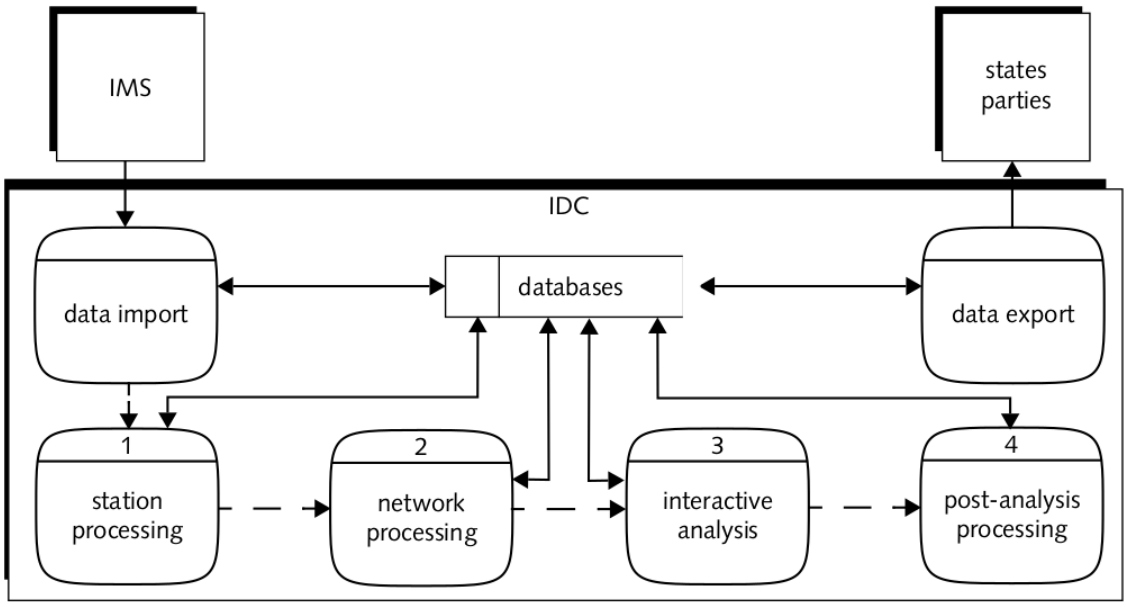

Figure 2. Diagram detailing the operation of the IDC [8].

1.2 Treaty Monitoring Pipeline for Seismic Signals

The IDC receives continuous waveforms from more than 170 seismic stations world-wide, and this data is used to build global seismic events in near-real time. To do so efficiently, the waveforms are broken up into windows (usually 10 minutes in length), and each window is then processed in two steps: Station Processing and Network Processing, where the Station Processing step considers each station individually and the Network Processing step looks at all stations in aggregate [8].

A block diagram of the Station Processing step is detailed in Fig. 3, and can be described briefly as follows: first, a signal detection algorithm identifies any arriving seismic waves; next, each arrival is processed, and features like arrival time, amplitude, period, azimuth and slowness are extracted; finally, these features are used to identify the phase type for each arrival, and the arrivals are grouped together into single-station events with location and magnitude estimates [8].

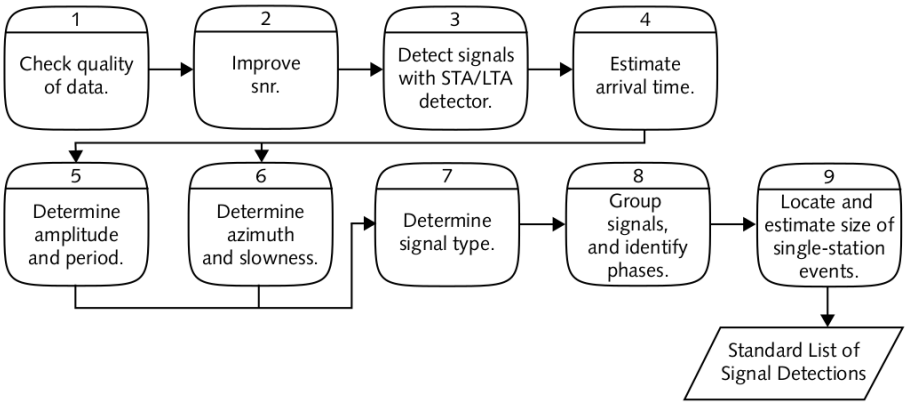

Figure 3. Diagram detailing the operation of the IDC with regards to automated seismic signal processing at the individual station [8].

For Network Processing, the single-station events are associated collectively and global seismic events are built, using a maximum likelihood estimator across a global search grid. This is an iterative process with multiple levels of conflict resolution, as seen in Fig. 4 [8].

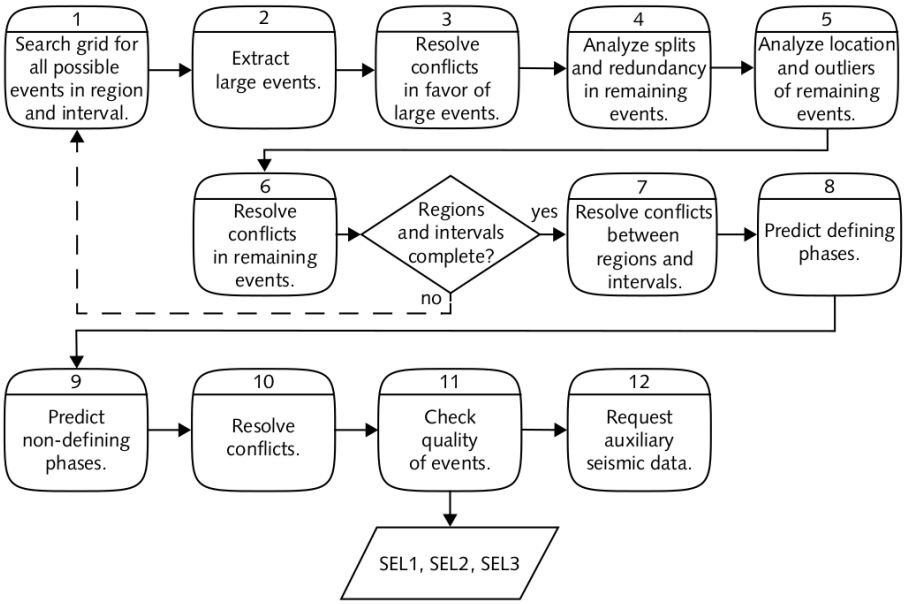

Figure 4. Diagram detailing the operation of the IDC with regards to automated seismic signal processing for the global seismic network [8].

In general, the automatic seismic signal processing pipeline at the IDC can be reduced to the following seven tasks:

1. Signal Detection
2. Arrival Time Estimation
3. Amplitude and Period Estimation
4. Azimuth and Slowness Estimation
5. Phase Classification and Grouping
6. Location and Magnitude Estimation
7. Source Discrimination [1]

The next section briefly considers the current implementation of each of these tasks at the IDC.

Signal Detection.

The first step in processing seismic signals is to detect the arriving seismic waves at each individual station. This step is complicated slightly at the IDC, by the fact that there are two different types of stations: individual stations and array stations. The individual stations usually include a single three-component (3C) seismometer, while the array stations usually include both a 3C seismometer and an array of single-component seismometers. To enhance the signal to noise ratio (SNR), individual stations employ bandpass filters that are manually tuned for each station [6]. To enhance the SNR at array stations, an additional spatial filter is applied, using beam-forming [83]. After SNR enhancements are complete, detections are made using the

[1] Technically, source discrimination is not a part of the signal processing pipeline at the IDC, as this task is designated solely to the individual state parties. However, because it is such a critical task in treaty monitoring, it is considered in this dissertation for completion.

ratio of the short-term average energy to the long-term average energy (STA/LTA), as shown in Fig. 5. A detection is declared when the STA/LTA for a channel or beam exceeds the detection threshold set for that channel [34].

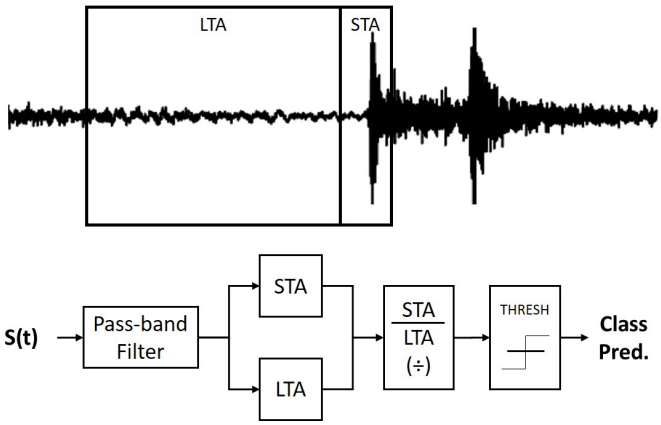

Figure 5. Top: Example seismic waveform, annotated to show the STA and LTA windows. Bottom: Diagram detailing the operation of the STA\LTA algorithm.

Arrival Time Estimation.

The first signal characteristic computed at the IDC is the precise onset time of the arriving seismic wave. This is often referred to as the arrival time. At the IDC, the arrival time is automatically estimated via the Akaike Information Criterion (AIC), a technique which is described in detail in [97] and illustrated in Fig. 6.

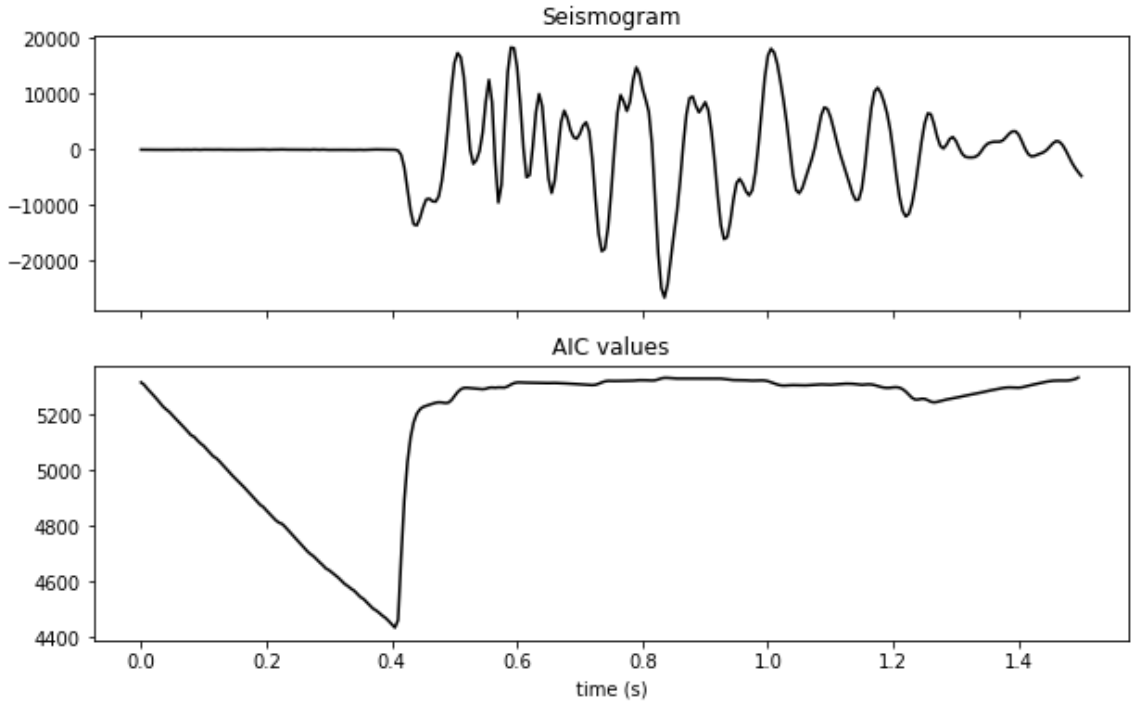

Figure 6. Example seismic waveform, annotated to show the value of the Akaike Information Criterion.

Amplitude and Period Estimation.

The amplitude is measured as half the maximum peak-to-trough amplitude difference in a small window of time taken from the vertical channel or beam used to make the detection. The window starts 0.5 seconds prior to the picked arrival time and ends 5.5 seconds after the arrival time. The period is measured as twice the time between the peak and trough used to calculate the amplitude. The calculations for amplitude and period are shown in Fig. 7.

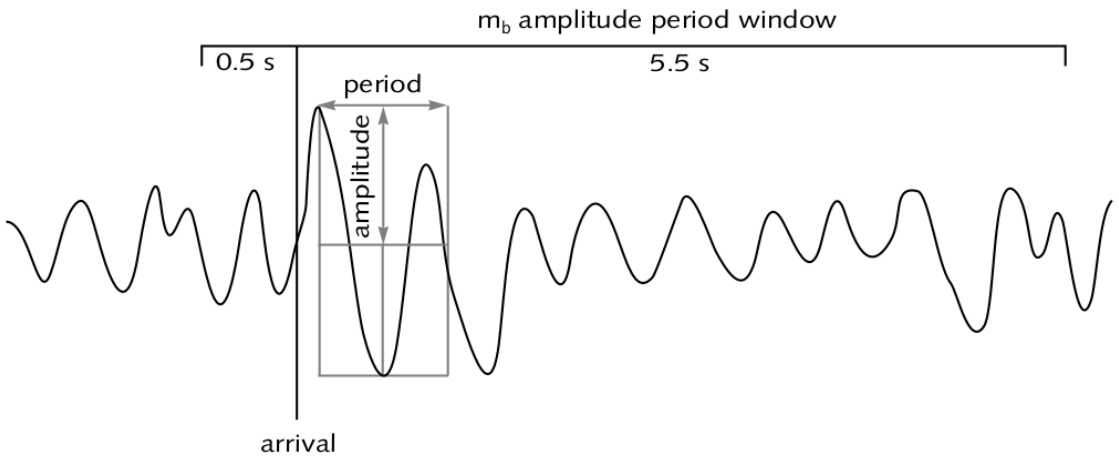

Figure 7. Example waveform detailing the process used at the IDC to measure the amplitude and period of a detected seismic arrival [8].

Backazimuth and Slowness Estimation.

The azimuth and slowness are calculated differently depending on whether the station consists of a single instrument or an array. For single-instrument stations, the traditional method of backazimuth prediction is to analyze the polarization of the three orthogonal components of motion: North-South, East-West and Vertical. This technique is often referred to as polarization analysis, and the algorithm is based on Principle Component Analysis of the filtered and windowed seismograms [57], [69]. In brief, the technique uses an eigendecomposition of the three-component covariance matrix across a window of data to identify the principle directions of both rectilinear and elliptical polarization [39]. Several advancements of this technique have been proposed, most notably the inclusion of variable time windows, which provides a small improvement in performance [77]. For array stations, the azimuth and slowness are calculated using frequency-wavenumber analysis in conjunction with the array beamforming [47]. An example array layout is detailed in Fig. 8, along with a demonstration of the beamforming technique.

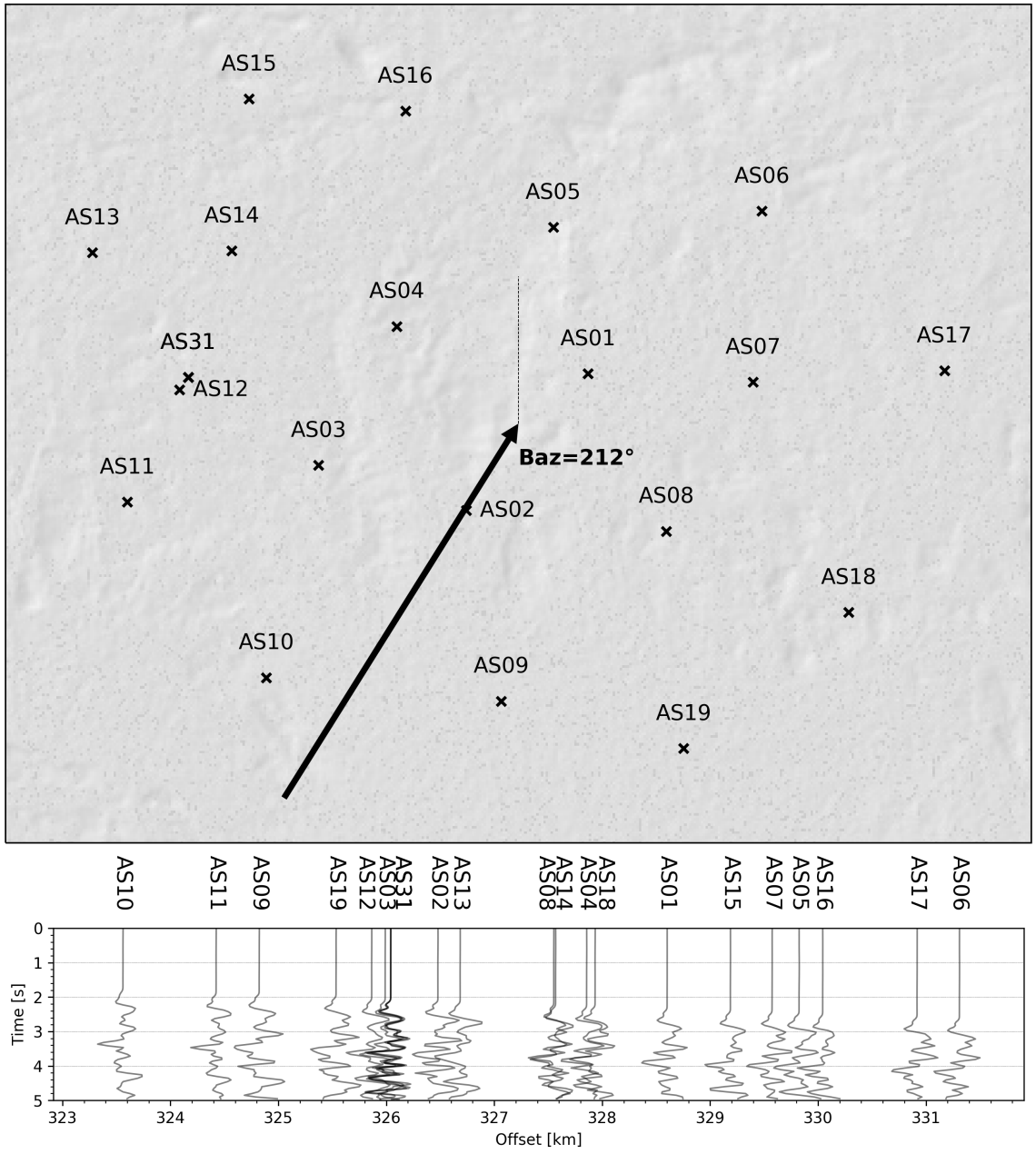

Figure 8. Top: Layout of the 20 element Alice Springs Seismic Array, ASAR, located in central Australia, with an aperture of just under 10 km. The arrow illustrates an incoming seismic wave with a backazimuth of $212°$. Bottom: Seismic waveforms from the corresponding seismic event, stacked in order of distance to epicenter. Beamforming uses the geometry of the array, along with the time-delay of arriving signals, to estimate the backazimuth angle with great precision.

Phase Classification and Grouping.

Each seismic event can produce several seismic wave phases, including primary (P) waves, which travel deep through the earth's core, and secondary (S) waves, which travel along the earth's crust. Often, two or more phases can be detected at a single seismic station, and the process of classifying the phase type of each arrival is accomplished at the IDC via a neural network [109], [95]. Once the phases have been classified, the next step is to group the phases that share a common event, and this is accomplished by a process of Bayesian inference [60].

Event Location and Magnitude Estimation.

Event location is a fundamental task in seismology, and it is a crucial step in Nuclear Treaty Monitoring. Traditionally, the location of a seismic event is estimated by performing a time delay of arrival calculation (TDOA), based on the arrival of various wave phases across a network of seismic sensors [102]. These computations are complicated by the fact that the seismic waves are travelling through a non-homogeneous medium and the equations must be modified to account for earth velocity models, as well as the the backazimuth and slowness estimates at each sensor [17].

Source Discrimination.

Once a seismic event has been detected and located, the final task in the seismic pipeline is to identify the source type of the event. Of particular interest to the Nuclear Treaty Monitoring community is the discrimination between explosions and earthquakes. Traditional discriminants for this task rely on physics-based expert features, such as P to S wave ratios and polarity of first motion [100]. These techniques are reliable and well-understood, but they also require significant tuning, and perform poorly under noisy conditions. Because nuclear tests are often performed at known

test sites, template matching techniques, such as correlation and subspace detectors, are frequently utilized to identify subsequent tests [14]. Finally, recent research has shown some promise in using machine learning to directly learn valid discriminants for nuclear explosions themselves [5].

1.3 Research Overview

The acquisition and processing of raw time-series seismic signals at the IDC has a significant impact on the global security of our world, detecting and identifying covert nuclear tests in near-real time. Most of the underlying algorithms have been in place for more than 20 years, and are time-tested. Unfortunately, many require frequent manual tuning and time-intensive analyst review [80], and others are insufficient for detecting the weakest events [92]. This work proposes several enhancements to the traditional seismic signal pipeline, both in terms of reducing analyst burden and improving detector sensitivity, and these enhancements are unified by a common reliance on the Temporal Convolutional Network, or TCN, a state-of-the-art neural network architecture ideally suited to extracting the long-period features predominant in the regional and teleseismic signals used for Nuclear Treaty Monitoring.

The work is composed of three separate research studies. Study 1 focuses on signal detection, employing a TCN architecture directly against the raw real-time data streams and effecting a 4 dB increase in detector sensitivity over the latest operational methods. Study 2 focuses on both event association and source discrimination, utilizing a TCN-based triplet network to extract source-specific features from three-component seismograms, and providing both a complementary validation measure for event association and a one-shot classifier for template-based source discrimination. Finally, Study 3 focuses on event localization, and employs a TCN architecture against three-component seismograms in order to confidently predict backazimuth

angle and provide a three-fold increase in usable picks over traditional polarization analysis.

Temporal Convolutional Networks.

Deep Convolutional Neural Networks (CNN) are rapidly revolutionizing the science of signal processing, from computer vision to speech recognition, and they are poised to do the same for seismic signal processing as well. CNNs have already been employed in almost every branch of seismological research, from earthquake detection to earthquake early warning systems, ground-motion prediction, seismic tomography, and even earthquake geodesy [62].

CNNs employ many layers of learned digital filters, which are combined with non-linear activations. This structure allows CNNs to accomplish a wide range of signal processing tasks in a way that is actually quite similar to the traditional analyst-driven methods, except for the fact that the empirical search for the optimal filters is performed by a computer in much less time and at a much larger scale. The key to learning an optimal transformation is simply obtaining a sufficient quality and quantity of labeled training data. Due to the vast quantity of labeled seismic data available, seismology is poised to take advantage of the power of CNNs.

While much work has already been done to integrate CNNs into seismic signal processing, these efforts have largely been limited to processing the signals from local seismicity, and little effort has yet been made to extend these techniques to the regional and teleseismic signals commonly encountered in Treaty Monitoring. This is because traditional CNNs are not well-suited to process the long-period features found in regional and teleseismic signals [28]. Fortunately, this limitation can be overcome by the TCN.

TCNs are deep convolutional architectures characterized by three primary fea-

tures:

- Causal convolutions
- Dilated convolutions
- Residual connections

These features combine to allow the neural network to quickly learn long-period features that are critical for teleseismic signal processing and that would be impossible to learn with a traditional CNN architecture. In short, causal convolutions allow the model to make predictions on continuous streaming time-series waveforms, dilated convolutions enable a wide receptive field for long-period feature extraction, and residual connections allow the model to have high-capacity and stable training [7]. Additionally, there are synergies that come from utilizing these three architectural features in concert. Particularly, as you increase the dilation rate on successive layers, this is roughly equivalent to increasingly decimating the input to each layer, but because of the skip connections, each layer's input also includes the inputs to all previous layers. This is an elegant way of capturing long-period features in the data while stabilizing backprop. Compared to RNNs, you get both improved performance [7] and a more meaningful analogue to traditional seismology [28].

The following three studies employ TCN architectures for use in Treaty Monitoring Seismology.

Study 1 - Signal Detection.

The detection of weak seismic events at regional (>200 km) and teleseismic distances (>2000 km) is critical to Nuclear Treaty Monitoring. Traditionally, detecting these weak regional and teleseismic events has required the use of an expensive multi-instrument seismic array, which uses a tuned network of interconnected seismometers

to accomplish an efficient spatial filtering technique called beamforming. This technique is extremely effective, however it is quite expensive to implement due to the additional sensors and processing required.

This study proposes a novel seismic detection algorithm capable of array-like detection performance from a single-trace, demonstrating a 4 dB increase in single-trace detector sensitivity over state-of-the-art techniques including the kurtosis [79] and frequency-band [70] pickers recently implemented for operational use by the Oklahoma Geological Survey. Building on several recent efforts which apply the power of convolutional neural networks to the detection of *local events* [81], [86], [68], this study applies similar techniques to the detection of *regional and teleseismic events*, events previously only detectable using a seismic array. Specifically, this study tackles the following research objective: using the analyst reviewed catalog of events from an array-beam as the data source, and fixing a type-I error rate of 0.1%, create a transportable single-trace detection algorithm with improved recall over existing detectors.

To tackle this objective, we present DeepPick, a single-trace automatic detection algorithm capable of detecting up to 80% of the events in an array-beam catalog. The algorithm is based on a deep Temporal Convolutional Neural Network (TCN), and it is trained against more than five billion raw seismic samples containing 608,362 labeled seismic arrivals from seven array-beam catalogs in the International Monitoring System (IMS) network: TXAR, PDAR, ILAR, BURAR, ABKAR, MKAR and ASAR located in Lajitas Texas, Pinedale Wyoming, Eielson Alaska, Bucovina Romania, Akbulak Kazakhstan, Makanchi Kazakhstan and Alice Springs Australia, respectively. Performance is reported by training the algorithm against five years of data from six of the arrays, and testing it against a full year of data from the seventh, remaining array. All seven arrays are tested in this manner, resulting in

recall rates ranging between 50% and 80% against the individual array beam catalogs, and an overall average of 56% against the combined catalog. This represents a marked improvement over the performance of existing algorithms for the same task. For example, deploying both a state-of-the-art adaptation of the STA/LTA detector (FBPicker) [70] and a kurtosis-based detector (KTPicker) [79] against the same test set, yields only 27% and 42% detection rates respectively. Additionally, the DeepPick algorithm is highly computationally efficient, demonstrating an order of magnitude reduction in computation time over both of the other algorithms.

While there have been several recent efforts to employ convolutional neural networks for seismic detection, our effort here differs in three significant ways. First, our detector was trained and tested using a higher-fidelity reference catalog with an 8 dB improvement in sensitivity over traditional catalogs, which is accomplished by utilizing an array-beam catalog as a reference. Second, whereas previous efforts treat detection as a binary classification problem (thus requiring a secondary algorithm for arrival time picking), our algorithm follows the traditional seismic detection approach of first creating a characteristic function. This effort follows very much in-line with traditional methodologies, but with significant quantitative improvements. Third, our detector is the first to focus on teleseismic detection, a task which depends upon recognizing long-period features, and which is accomplished using a Temporal Convolutional Neural Network (TCN) with a wide receptive field. As such, we present three major contributions to the literature:

- A unique training technique for single-trace detection algorithms, which utilizes array-beam catalogs as a high-fidelity reference
- A novel training objective, *exponential sequence tagging*, which trains the TCN to transform single-trace waveforms into an ideal characteristic function with weighted exponential peaks at predicted arrival times

- DeepPick: a single-trace detection algorithm capable of achieving array-level detection performance

Study 2 - Event Association.

Event Association is a critical step in Nuclear Treaty Monitoring. Traditionally, association is accomplished via a move-out curve predicated on the arriving waveforms in time and space. This work presents a viable complementary validation tool for existing associators using a novel pairwise seismic source similarity measure.

Traditionally, seismic similarity measures have been based on waveform correlation. Case-based discrimination [32], template matching [36], waveform correlation [43], subspace detection [44] and similarity search [113] are all similarity-based algorithms which have been proposed over the last several decades, and deployed against a wide range of seismic signal processing tasks, such as discriminating mining blasts, screening swarm events, identifying aftershock sequences, and even detecting general seismic signals. While these algorithms have different tasks ranging from discrimination to detection, they all share a common measure of similarity: cross-correlation. Unfortunately, these techniques cannot be used for waveform association, as the correlation coefficient between two seismograms is dominated by path effects [94].

This work presents a new measure for seismogram similarity that bypasses correlation entirely, and that is designed to be both path-invariant and source-specific. To be precise, the design goal is to create a measure of seismogram similarity that enables the identification of seismograms sharing a common source event, regardless of the path of travel. While such a measure was previously computationally intractable, it is possible with the careful application of deep convolutional neural networks (CNNs). In 2019, researchers at the Los Alamos National Laboratory pub-

lished a method using a CNN to predict the pairwise association of seismic phase arrivals, for 6 second windows, across a local group of 6 stations in northern Chile, reporting an accuracy of over 80% [72]. Building on these results, we construct a source-dominant, path-invariant measure for seismogram similarity which operates on 180 second windows and is generalized across more than 1,000 sensors across North America. This is accomplished by utilizing a state-of-the-art machine learning technique from the field of facial recognition, called a Triplet Network, which not only indicates pairwise association between seismograms, but actually maps the seismograms to low-dimensional vectors, called embeddings, such that the embedding space distance between seismograms sharing a common source event are minimized, regardless of path, while remaining distinct from any other events. In this way, the embedding function becomes a rich feature extraction technique for source-specific and path-invariant features.

The triplet network architecture accepts three observations - two which are similar and one which is different from the others. Training a triplet network to learn seismic source similarity requires source-similar seismogram triples: two of the three waveforms are associated with a common source event and the third waveform is not. For this task, it is preferable to have a training set containing seismograms recorded from a densely-spaced sensor network, so that the neural network can experience seismogram recordings across numerous paths for the same event. The 400 three-channel broadband sensors of the USArray experiment provided an ideal dataset of seismograms; data from this array is used for training and testing. The triplet network is trained against 7 years of data (2007 - 2013), validated against a single year of data (2014), and tested against the final two years of data (2015-2016). Additionally, a subset of 51 recording stations and a small region of event locations were held out from the algorithm during training, to allow a proper evaluation of the generalizability

of the technique.

The value of this path-invariant measure is demonstrated as a complementary validation tool for pairwise event association. The pairwise event association task of determining whether or not two waveforms depict the same event achieves a binary accuracy of 80%. This accuracy is achieved using only the waveform characteristics, without information on times or recording locations, and the technique has strong potential to augment existing methods of event association [72].

Study 3 - Backazimuth Prediction.

Backazimuth prediction is a critical step in the seismic signal processing pipeline, feeding the downstream processes that associate events and build location estimates. Typically, there are two methods of predicting backazimuth, depending on the type of station. If the station consists of an array of instruments, the backazimuth can be predicted by examination of the time-delay of arrival across the array. This process is called beamforming, and produces angle estimates that can be quite accurate. If the station consists of a single three-component (3C) instrument with North-South, East-West and Vertical components, the backazimuth is traditionally predicted by calculating the polarization of the arriving wavefront. This process produces much less accurate results.

This study proposes BAZNet, a machine-learning-based alternative to polarization analysis that not only produces more accurate backazimuth estimates, but also produces actionable certainty measures for each estimate, allowing downstream algorithms to only use the best estimates available. The BAZNet model directly predicts the backazimuth from raw 3C waveform data, utilizing a deep temporal convolutional neural network architecture [7] to extract meaningful features from the seismograms. It is important to note that the model is trained on a per-station basis against 10

years of historical data; the technique does not generalize across stations, and must be retrained for each station where it will be employed. However, because of the large number of available 3C stations with with extensive analyst-reviewed catalogs, and because of the outstanding certainty measure produced in conjunction with each estimate, BAZNet is able to produce backazimuth estimates for 3C stations with accuracy rivaling a beamformed array.

BAZNet presents three major contributions:

- A novel neural network architecture for the efficient prediction of backazimuth, directly from the raw waveforms with no feature engineering required
- An improvement in accuracy over the traditional polarization analysis
- A robust certainty measure coupled with each backazimuth estimate, allowing a means of preventing bad estimates from corrupting downstream algorithms for event association and location.

II. Study 1 - Improving Regional and Teleseismic Detection for single-trace waveforms using a Deep Temporal Convolutional Neural Network trained with an Array-Beam catalog [28]

2.1 Abstract

The detection of seismic events at regional and teleseismic distances is critical to Nuclear Treaty Monitoring. Traditionally, detecting regional and teleseismic events has required the use of an expensive multi-instrument seismic array; however in this work, we present DeepPick, a novel seismic detection algorithm capable of array-like detection performance from a single-trace. We achieve this performance by training a deep temporal convolutional neural network detector against the arrival times in an array-beam catalog and the single-trace waveforms taken from the vertical channel of the center element of the array. The training data consists of all arrivals in the International Seismological Centre Database for seven seismic arrays over a five year window from 1 Jan 2010 to 1 Jan 2015, yielding a total training set of 608,362 detections. The test set consists of the same seven arrays over a one year window from 1 Jan 2015 to 1 Jan 2016. We report our results by training the algorithm on six of the arrays and testing it on the seventh, so as to demonstrate the generalization of the technique to new stations. Detection performance against this test set is outstanding. Fixing a type-I error (false positive) rate of 0.1%, the algorithm achieves an overall recall (true positive rate) of 57.8% on the 141,095 array beam picks in the test set, yielding 81,524 correct detections. This represents a 40% increase in performance over state-of-the-art kurtosis-based detectors, and is more than twice the 37,572 detections made by a state-of-the-art STA/LTA detector over the same period. Furthermore, DeepPick provides a 4 dB improvement in detector sensitivity over all other current methods tested, with a run-time that is an order of magnitude faster. These results

demonstrate the potential of our algorithm to significantly enhance the effectiveness of the global treaty monitoring network.

2.2 Introduction

Adherence to the Comprehensive Nuclear Test-Ban-Treaty is currently verified by the detection, location and identification of seismic events, often at regional (>200 km) and teleseismic distances (>2000 km). Automated seismic detection is the critical first step in this process, and it is imperative that the events be detected by multiple stations, as this increases the overall accuracy of the final location estimate. As such, maintaining a large network of highly-sensitive seismic detectors is key to the treaty monitoring community [83, 4].

Traditionally, sensitive teleseismic detection has required the use of a multi-instrument seismic array, a strategy which dates back to the Geneva Conference of Experts in 1958 [99]. The sensitivity is achieved through beamforming [105], a spatial filtering technique that relies on a tuned network of interconnected seismometers which form a single station. This technique is extremely effective, however it is quite expensive to implement due to the additional sensors and processing required, and unfortunately, beamforming is inapplicable to single-instrument stations. As such, the vast majority of seismic stations around the globe are simply unable to detect weak regional and teleseismic events.

In this work, we create an automatic detector with array-like performance from a single trace, capable of detecting these signals which were previously too weak to detect with a single sensor. Building on several recent efforts which apply the power of convolutional neural networks to the detection of *local events* [81], [86], [68], we apply similar techniques to the detection of *regional and teleseismic events*, events previously only detectable using a seismic array. Specifically, we tackle the following

research objective: Using the analyst reviewed catalog of events from an array-beam as the reference, and fixing a type-I error rate of 0.1%, create a transportable single-trace detection algorithm with improved recall over existing detectors.

To tackle this objective, we present DeepPick, a single-trace automatic detection algorithm capable of detecting up to 80% of the events in an array-beam catalog. The algorithm is based on a deep Temporal Convolutional Neural Network (TCN), and it is trained against more than five billion raw seismic samples containing 608,362 labeled seismic arrivals from seven array-beam catalogs in the International Monitoring System (IMS) network: TXAR, PDAR, ILAR, BURAR, ABKAR, MKAR and ASAR located in Lajitas Texas, Pinedale Wyoming, Eielson Alaska, Bucovina Romania, Akbulak Kazakhstan, Makanchi Kazakhstan and Alice Springs Australia, respectively. Performance is reported by training the algorithm against five years of data from six of the arrays, and testing it against a full year of data from the seventh, remaining array. All seven arrays are tested in this manner, resulting in recall rates ranging between 50% and 80% against the individual array beam catalogs, and an overall average of 56% against the combined catalog. This represents a marked improvement over the performance of existing algorithms for the same task. For example, we deploy both a modern adaptation of the STA/LTA detector (FBPicker) [70] and a kurtosis-based detector (KTPicker) [79] against the same test set, achieving only 27% and 42% detection rates respectively. Additionally, our algorithm is highly computationally efficient, demonstrating an order of magnitude reduction in computation time over both of the other algorithms.

While there have been several recent efforts to employ convolutional neural networks for seismic detection, our effort here differs in three significant ways. First, our detector was trained and tested using a higher-fidelity reference catalog with an 8 dB improvement in sensitivity over traditional catalogs, which we accomplished by

utilizing an array-beam catalog as a reference. Second, whereas previous efforts treat detection as a binary classification problem (thus requiring a secondary algorithm for arrival time picking), our algorithm follows the traditional seismic detection approach of first creating a characteristic function. As such, we show that our effort follows very much in-line with traditional methodologies, but with significant quantitative improvements. Third, our detector is the first to focus on teleseismic detection, a task which depends upon recognizing long-period features, and which we accomplish using a Temporal Convolutional Neural Network (TCN) with a wide receptive field. As such, we present three major contributions to the literature:

- A unique training technique for single-trace detection algorithms, which utilizes array-beam catalogs as a high-fidelity reference
- A novel training objective, *exponential sequence tagging*, which trains the TCN to transform single-trace waveforms into an ideal characteristic function with weighted exponential peaks at predicted arrival times
- DeepPick: a single-trace detection algorithm capable of achieving array-level detection performance

In the remainder of this work, we provide context for and explain these contributions by first reviewing the related literature, then outlining our methodology, and finally detailing and discussing our results.

2.3 Related Work

Automatic seismic detection algorithms are a key component of any modern seismic network, and here we review the literature pertaining to this important field. Our review begins with a discussion of the traditional detection algorithms, then investigates teleseismic detection in particular. Finally, this section provides a detailed

examination of the nascent field of convolutional neural network-based detectors, while emphasizing the gaps in the research we intend to address in our own work.

Traditional Seismic Detection.

Traditional algorithms for seismic signal detection usually share a simple, common framework: A comparison is made between the current value of the seismic signal (or some function of it) and a predicted value, and a detection is declared whenever this comparison exceeds some factor. From this simple concept has arisen a vast number of algorithms, which vary primarily based upon their choice of the function to which the detection is applied. This function is often referred to as the characteristic function (CF) of the algorithm [6].

By far the most common traditional technique for seismic signal detection is the short-term average, long-term average (STA/LTA) detector, first described by Freiberger [34]. In its simplest form, this technique employs a bandpass filter to compute the characteristic function, with the predicted value equal to the long-term average and the current value equal to the short-term average. The current and predicted values are then compared via a ratio which is then subjected to some static threshold, as detailed in Fig. 9. Numerous adaptations and enhancements to this STA/LTA detector have been proposed, most notably by Allen [2] and Baer [6], who increased detection efficiency by employing novel characteristic functions based on a combination of the signal and its time derivatives. More recently, modern iterations of the STA/LTA algorithm have employed multiple characteristic functions across multiple frequency bands with great success [70].

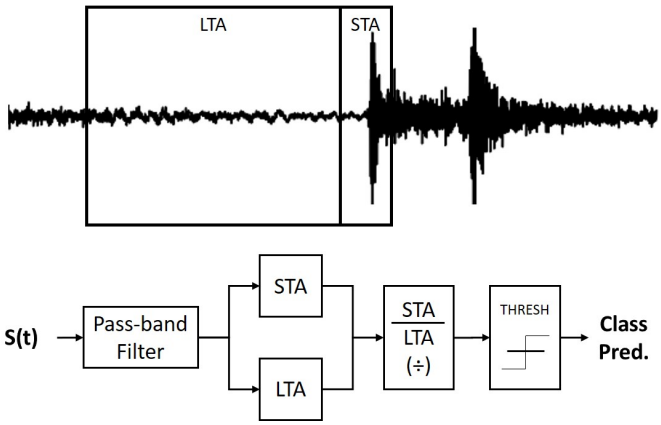

Figure 9. Top: Example seismic waveform, annotated to show the STA and LTA windows. Bottom: Diagram detailing the operation of the STA\LTA algorithm.

Higher-ordered statistics.

Unfortunately, the STA/LTA family of algorithms have an inherent difficulty identifying events that emerge from a noisy pass-band [90]. Fortunately, unlike random noise, seismic signals have higher-order statistics (such as skewness) which are non-zero [35]. This means that the signal and noise energies can be well-separated using characteristic functions based on these higher-order statistics (HOS), which serve as the basis for another common subset of seismic signal detectors, the HOS-based detectors described in [65, 90, 115]. These algorithms can provide excellent performance, but tend to be more computationally expensive.

Other more exotic characteristic functions that have enjoyed success include variations of the Walsh transform [38] and the wavelet transform [3]. Furthermore, there are families of algorithms used to determine the precise arrival time after a detection has been made. These are commonly referred to as autoregressive methods, which employ various techniques, the most common of which was proposed by Sleeman [97] and utilizes the Akaike Information Criterion.

Teleseismic Detection.

Having examined seismic signal detection in general, we now turn our attention specifically to the literature concerning the detection of regional and teleseismic signals. Such signals can be particularly challenging to detect, as their signal strength is often significantly attenuated by the longer path of travel. To address this, one of the most successful techniques for regional and teleseismic signal detection is Beamforming [83], [88], introduced in [21]. Beamforming gains its effectiveness by linearly combining signals from multiple sensors according to the estimated arrival direction, also known as the backazimuth, allowing it to pick out signals beneath the noise floor of a single sensor [105]. Unfortunately, beamforming requires an interconnected array of seismometers, spread out across a large geographical area measuring tens or even hundreds of kilometers. An example array layout, along with a demonstration of the beamforming technique is detailed in Fig. 10.

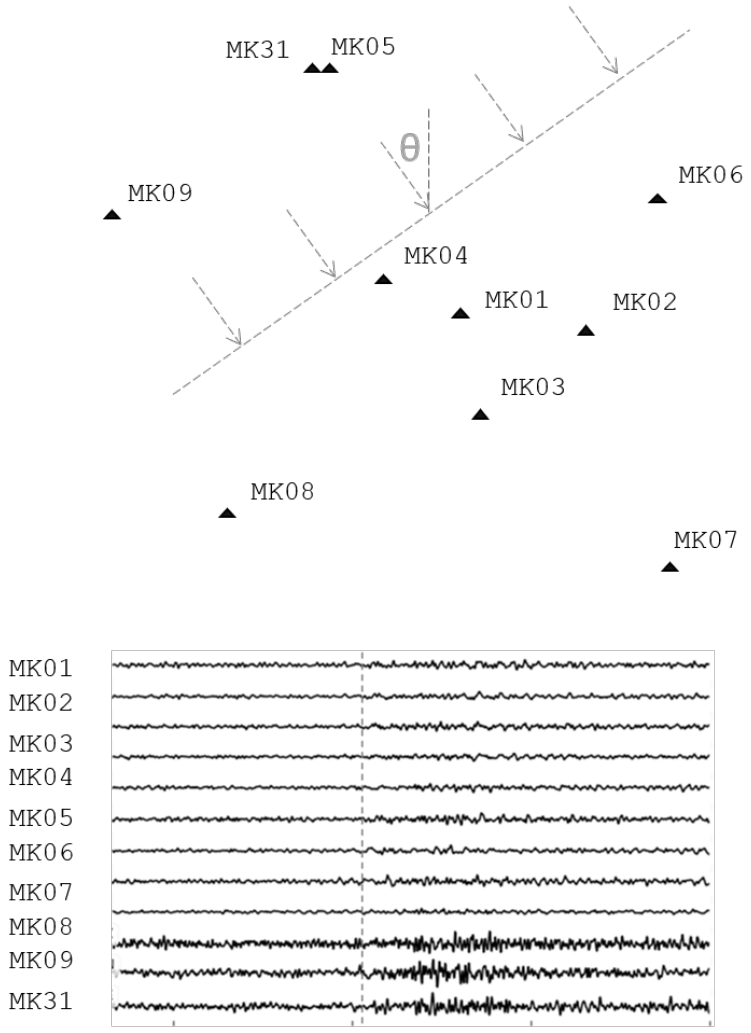

Figure 10. Top: Layout of the 10-element Makanchi Seismic Array, MKAR, in eastern Kazakhstan. The dashed lines illustrate an incoming teleseismic wave with calculated backazimuth, θ. Bottom: Seismic waveforms from an arriving teleseismic event. Beamforming aligns these waveforms via the backazimuth and wavefront velocity, and then linearly combines them to yield a higher SNR, improving the detection threshold significantly.

Another outstanding technique for the detection of weak teleseismic events is the phase-matched filter [104] popularized by [50] and [106] in the early 1990s. These pattern matching techniques are a type of Empirical Signal Detector, that work by comparing incoming seismic waveforms to canonical examples in the extant seismic

record [55, 56]. They are particularly effective for the detection of highly correlated repeating events, even for very weak magnitudes [37]. Unfortunately, to date, this technique is not generally applicable, as only 18% of all global events possess sufficient similarity to be detected with this technique [31].

In [89], the authors demonstrate the power of a richly-featured machine learning based detector. Training a Support Vector Machine against a series of 30 features in the time-frequency plane, they achieved a recall of 97.7% at a type-I error rate of less than 1.3%, for an overall accuracy of 98.2%. These results compare favorably with STA/LTA. Their work is quite promising, with excellent results, however, the signals investigated were once again limited to strong, local signals; the furthest signals detected had epicenters no more than 5 degrees ($\sim$550 km) from the recording sensor.

Seismic Detection with Convolutional Neural Networks.

Convolutional Neural Networks (CNN) are revolutionizing the science of signal processing from computer vision to speech recognition, and they are poised to do the same for seismic signal processing as well. This begs the question: why are CNNs so effective at signal processing tasks? To answer this, we note that at their core, CNNs are comprised of a set of digital filters which are convolved with the signal, where the optimal filter weights are learned by applying stochastic gradient descent across some objective function. In effect, the CNN can quickly explore a wide range of filters and empirically converge on ones that work well. The real power of CNNs comes from the ability of the network to learn many filters simultaneously, combine their outputs with non-linear activations, and then feed these activations into further layers of learned filters, ultimately allowing the CNN to learn complex non-linear transformations. This structure allows CNNs to accomplish a wide range of signal

processing tasks in a way that is actually quite similar to the traditional analyst-driven methods, except for the fact that the empirical search for the optimal filter weights is performed by a computer in much less time and at a much larger scale. The key to learning the optimal transformation is having sufficient quality and quantity of labeled training data for the objective function. And due to the vast quantity of labeled seismic data available, seismology is poised to take advantage of the power of CNNs.

Several recent efforts have already been made to apply deep CNNs to seismic signal detection. Although this research is still in its infancy, early results have shown great promise.

In [81], the researchers utilize a convolutional neural network architecture to perform detection on local seismic signals, formulating the task as a binary classification problem. Their dataset was obtained from two seismic stations in the Oklahoma Geological Survey, consisting of 10-second windows with binary class labels: positive windows were centered around seismic arrival times obtained from an analyst-reviewed arrival catalog, and negative windows were carefully selected to contain no arrival. Against their hold-out test set, they report 100% recall with a high type-I error rate of 1.4%. These results are outstanding, but the most interesting finding in their research comes from their examination of the false positives detected by their algorithm. By applying a correlation detector to their reported false positives, they determined that a substantial portion of these were actually real detections of very weak events. This means that the algorithm learned to detect events below the detection threshold of the catalog on which it was trained. This work highlights the danger of using conventional catalogs to train such a sensitive detector. Additionally, two major limitations exist in this work. First, because of the extreme care taken to produce 'clean' noise windows in the test set, their reported type-I error rate is

not realistic for operational use. Second, their algorithm is applicable only to local events; the short time windows used (10 seconds) limit the algorithm's potential to detect the longer-period (100 second) features characteristic of teleseismic signals.

In [87], the researchers also utilize a deep CNN to perform seismic signal detection on local events. Their dataset consisted of 4.5 million 4-second windows of waveform data recorded and classified by the Southern California Seismic Network. Their task was formulated as a classification problem, assigning one of three classes to each window, P-wave (primary phase arrival), S-wave (secondary phase arrival) and noise. This resulted in 1.5 million windows containing a P-wave arrival, 1.5 million windows containing an S-wave arrival and 1.5 million windows including no arrival. Their validation set consisted of a randomly sampled 25% of the overall data, resulting in 1.1 million seismograms evenly split between the three classes. On the validation set, they report a recall of 96% at a type-I error rate of less than 1%. These results are very impressive, and show that the convolutional neural network is capable of achieving state-of-the-art performance on the seismic signal detection task. A limitation of this work is that it is applicable only to local signals; the researchers only report recall for signals originating within 100 km of the recording station.

In [86], the same research team considers arrival time estimation. Here they formulate the task as a regression problem, and consider only 4-second windows of data, centered around an arrival, with up to half a second of deviation in the arrival time from the center of the window. For this task, they report a mean average error of less than 0.02 seconds from the analyst-recorded picks. Once again, these signals are limited to local events.

2.4 Materials and Methods

The research objective is to build a single-trace detection algorithm capable of detecting weak regional and teleseismic signals with array-like performance. We know that such detections are possible using a full seismic array and we have seen the potential for achieving such detections using a deep neural network. Our approach is to employ a deep TCN model, feed it a single-trace input sequence, and train it to produce a characteristic function with distinct peaks centered on arrival times obtained from an array beam catalog. In this section, we explore this approach in detail, first defining our dataset, and then describing our modeling strategy.

Data Collection.

The success of any deep neural network algorithm lies largely in the careful collection and construction of the training data. This subsection presents a dataset suitable for training a deep seismic detection algorithm. In particular, it details two of our major contributions: First, a description of a novel method for obtaining a high-fidelity dataset of single-trace waveforms with labeled arrival times below the noise floor. Second, it presents exponential sequence tagging, the unique sequence-to-sequence modeling schema used to create an ideal characteristic function for picking arrival times. This subsection concludes with the details of training, test and validation datasets.

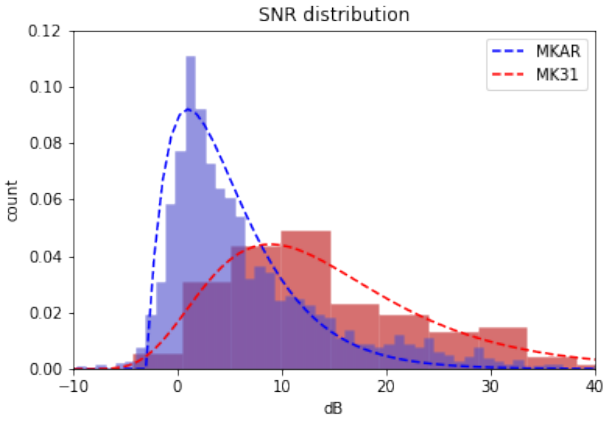

Figure 11. Normalized histograms showing the SNR distributions of detected signals from two seismic arrival catalogs. Both catalogs contain detections for the exact same location, MK31, which is the reference element of the MKAR seismic array. The MK31 catalog is based on a single-trace detection algorithm applied to the MK31 instrument alone, while the MKAR catalog is based on beam-formed picks from the entire 10-instrument array. The mean SNR detected by the array beam is 8 dB lower than that of the single-trace. This lower detection threshold results in nearly an order of magnitude more detections in the MKAR catalog compared to the MK31 catalog.

High Fidelity Arrival Catalog.

At first glance, obtaining a dataset for training a seismic detector would appear to be trivial, as analyst-reviewed arrival catalogs are freely available for millions of seismic events at thousands of seismic sensor elements. Unfortunately, despite the rigorous review process and the extensive cross-referencing, each single-trace arrival catalog only contains picks for signals with sufficient strength to be conventionally detectable from within that trace. This is a significant limitation when the goal is to train a detector *more sensitive* than the conventional one. Fortunately, there are certain sensor elements with accurate cataloged arrival times for regional and teleseismic signals below the noise floor; namely, any sensor element located at the reference point of a seismic array (usually a broadband 3-channel instrument). Using conventional

methods, this 'reference-element' alone is unable to make accurate detections for sub-noise floor events, however the array beam as a whole can make these detections very accurately [88], and the beam arrivals are conveniently aligned to indicate arrivals at this reference element. Thus, by obtaining singe-trace input data from the reference element, and by obtaining labeled arrivals times from the array beam, we can create a labeled single-trace dataset with signals below the noise floor. As an example, Fig. 11 demonstrates the significant 8 dB improvement in detector threshold provided by the Makanchi Array beam in eastern Kazakstan.

For future researchers interested in establishing a similar high-fidelity dataset, we provide here a four step process:

1. **Step 1: Obtain the Array-Beam Catalog** Arrival-time catalogs can be downloaded through a web query of the International Seismological Centre Bulletin (`http://www.isc.ac.uk/iscbulletin/search/arrivals/`), by specifying the desired array station name (i.e. MKAR)

2. **Step 2: Identify the Array-Beam Reference Point** The array-beam reference point coordinates can be found through a web query of the ISC station registry (http://www.isc.ac.uk/registries/search/), by again specifying the desired array station name.

3. **Step 3: Identify the Array-Beam Reference Elements** Available reference elements can then be found by a second web query of the ISC station registry, using the reference point coordinates as the search criteria. For the MKAR array, there are two sensor elements located at the reference point: MK31 and MK32.

4. **Step 4: Obtain Reference Element Waveforms** Raw waveforms can be downloaded from the Incorporated Research Institutions for Seismology (IRIS)

Database using ObsPy.

Idealized Characteristic Function: Exponential Sequence Tagging.

With established high-fidelity sources for both waveforms and arrival times, the next step is to generate input/output pairs for training a seismic detector. Most previous efforts to build an ML-based seismic detector have been formulated as a binary classification task; the input data is partitioned into fixed length windows, each paired with a single Boolean class label: positive class labels are assigned to windows where a signal is present and negative class labels are assigned to windows where signal is absent. This traditional formulation is convenient, as the classes can easily be balanced at training time and it is the common method employed in most recent works in the literature [87, 81, 89]. However, this methodology has three major limitations: First, it is not ideally suited for real-time processing, as the algorithm needs access to a signal window several seconds beyond the signal arrival. Next, it requires a secondary algorithm applied within the detection window, to estimate the precise arrival time [86]. Finally, this methodology is not well suited for the detection of regional and teleseismic signals. Teleseismic signals are characterized by long-period features with frequency components as low as 0.01 Hz [84], and the detection of these features necessitates windows that are several minutes in length. Unfortunately, this resolution is far too coarse for classification, and often covers multiple arrivals in a single window. As such, there are two conflicting requirements for creating binary classification windows in a teleseismic detection dataset:

- Input windows must contain many samples to capture long-period teleseismic features
- Output labels must cover few samples to allow meaningful temporal resolution for the detection windows

To resolve this conflict, we reformulate the task. Instead of performing binary classification on each window, we perform regression on each sample, which is known as sequence-to-sequence modeling [101]. Each training window is labeled with an output sequence of real-valued numbers; each sample in the input sequence is assigned a corresponding value in the output sequence. Coincidentally, this process is nearly identical to the generation of the characteristic function in traditional seismic detector algorithms. The difference is that whereas traditional algorithms specify the transformation in order to produce a characteristic function that has defined arrivals, our algorithm can specify the characteristic function explicitly and let the neural network learn the transformation. As such, we can assign labels corresponding to any idealized characteristic function we desire. But what labels should we assign? A naive formulation is to simply assign a 'one' at each cataloged arrival time and assign a 'zero' everywhere else. This characteristic function would essentially look like a delta function at each cataloged arrival. This formulation is called sequence tagging [82], and it works well for relatively balanced classes [112]. Unfortunately, binary sequence tagging does not work well for teleseismic detection, as it results in an extreme class imbalance of several orders of magnitude, which hinders learning.

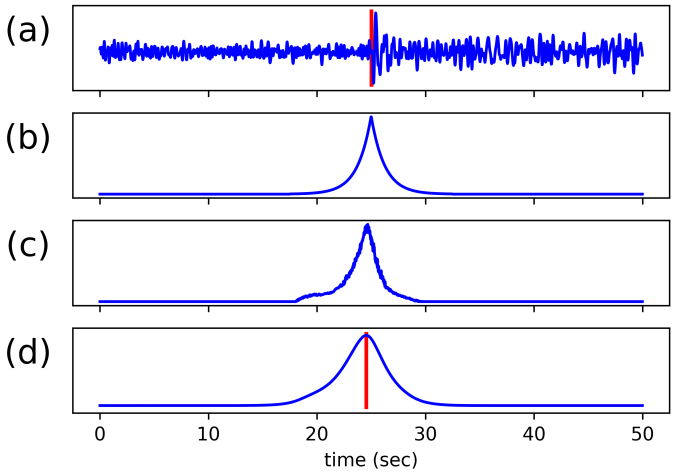

Figure 12. (a): Input Sequence containing an arrival marked in red (b): Labeled output sequence using the exponential function. (c): Predicted output sequence from the model. (d): Cross-correlation between the predicted output sequence and the exponential function. The predicted arrival is marked in red

For this work, we present a novel formulation which we call exponential sequence tagging. This formulation produces a characteristic function that consists of a mirrored-exponential function applied at each cataloged arrival time, as shown in Fig. 12 (b). To be precise, the labels in the output sequence are nominally zero until a cataloged arrival time, at which point they increase and decrease exponentially, according to the mirrored exponential decay function given in Eq. (1), where λ is the decay rate, which is optimized for maximum detection accuracy. This characteristic function is quite similar to that used in the 'suspension bridge' seismic detection algorithm, proposed in [74] and referenced in [114].

$$y(t) = e^{-\lambda|t|} \tag{1}$$

Because each leg of the mirrored exponential decay function is both monotonic and deterministic, the value at each non-zero label can be used to directly infer the precise

arrival time. Because the algorithm learns to match these labels with its output, every non-zero sample in the output is effectively an arrival time estimation. With this in mind, we assign one additional computation to our algorithm at run-time: a cross-correlation of the predicted output sequence with the original exponential decay function. This filters the output and effectively aggregates the arrival time estimates for an even more precise arrival time pick. Because the height of the resulting peak is the correlation between the network model's output and the original exponential, it represents the certainty that the peak is a true arrival and can be used to set the threshold of the detector. Fig. 12 (c) and (d) shows an example of the predicted output, both before and after this cross-correlation is applied, where (d) depicts the final characteristic function.

Training, Validation and Test Sets.

Using this approach to build our training dataset, we obtained a catalog of all local, regional and near-teleseismic arrivals for the seven array beams during a five year period from 1 Jan 2010 to 1 Jan 2015. We generated this catalog through a web query of the ISC Bulletin for seismic arrivals which can be accessed here: `http://www.isc.ac.uk/iscbulletin/search/arrivals/`. The corresponding waveforms were then windowed around each arrival (the windows were 6 minutes in total length, sampled at 40 Hz for a total of 14400 samples per window), and the raw traces were pulled from the IRIS Database, for the vertical channel of the nominal seismometer for each array (PD31_BHZ, TX31_BHZ, IL31_BHZ, MK31_BHZ, ABK31_BHZ, BUR31_BHZ and AS31_BHZ). This was accomplished via a custom Python script based on ObsPy-1.1.0 [13], and yielded a dataset of 608,362 picks out of a total training size of more than five billion samples. The only pre-processing applied to the raw data was a normalization, detrending and bandpass filtering between 0.02 Hz and 10 Hz.

From this training dataset, we selected one month of data from each array (1 Jan 2010 to 1 Feb 2010), as a validation set. This validation set was used to tune the models, with final model selection based on validation set performance.

To build our testing dataset, we also obtained a catalog of all local, regional and near-teleseismic arrivals for the seven array beams, in this case during a one year period from 1 Jan 2015 to 1 Jan 2016. This test includes 141,095 arrivals in the seven array beam catalogs. This test set data was not used to train or tune the models, only to report performance against each array. Additionally, to ensure that our reported performance figures are indicative of the expected performance against novel stations, we actually trained seven separate models, each on a different partition of six arrays and tested against the seventh, such that performance for all seven arrays is reported using a model that did not have access to any training data from that array, demonstrating the generalizability and transportability of our detector.

Modeling.

Now that we have defined our dataset, we present a description of our modeling methodology, detailing the model architecture, hyper-parameter search vectors, and evaluation metrics.

Model Architecture.

Our model architecture is based on the Temporal Convolutional Network (TCN). TCNs are deep convolutional architectures characterized by layered stacks of dilated causal convolutional filters with residual connections [7]. These characteristics offer several distinct advantages for a seismic detection algorithm, which we briefly summarize:

- Residual connections allow the model to have high-capacity and stable training

- Causal convolutions allow the model to make predictions on continuous streaming trace data
- Dilated convolutions allow precise control over the receptive field

The receptive field is of primary importance for time-series modeling, as it explicitly limits the learnable feature periodicity at a given layer. As such, one of our key design parameters was to ensure adequate receptive field for our algorithm. The equation for calculating the receptive field for a given convolutional layer, l, and dilation rate, d is given in Eq. (2):

$$rField(l) = rField(l-1) + [kernelSize - 1] * d(l) \tag{2}$$

Table 1. Layer Parameters for a single stack of our TCN architecture. Descriptions of the columns are as follows: l represents the layer number within the stack, k represents the kernel size (also known as the filter length or number of weights in each learned digital filter), d represents the dilation rate, and Receptive Field represents the number of samples in the input sequence 'seen' by the filters at that layer.

l	k	d	**Receptive Field**
1	16	2	31
2	16	4	91
3	16	16	331
4	16	256	4171

Using Eq. (2), the network is designed to have a receptive field of roughly 100 seconds (or 4,000 samples), allowing it to learn long-period features down to 0.01 Hz. This is accomplished in just 4 layers, as shown in Table 1. Another key design parameter was to ensure that the dilation rate in each layer remained less than the

receptive field in the previous layer to prevent gaps in receptive field coverage. Notice that this constraint is maintained even for the final layer with a dilation rate of 256, as the previous layer had a receptive field of 331. The model architecture is shown in Fig. 13.

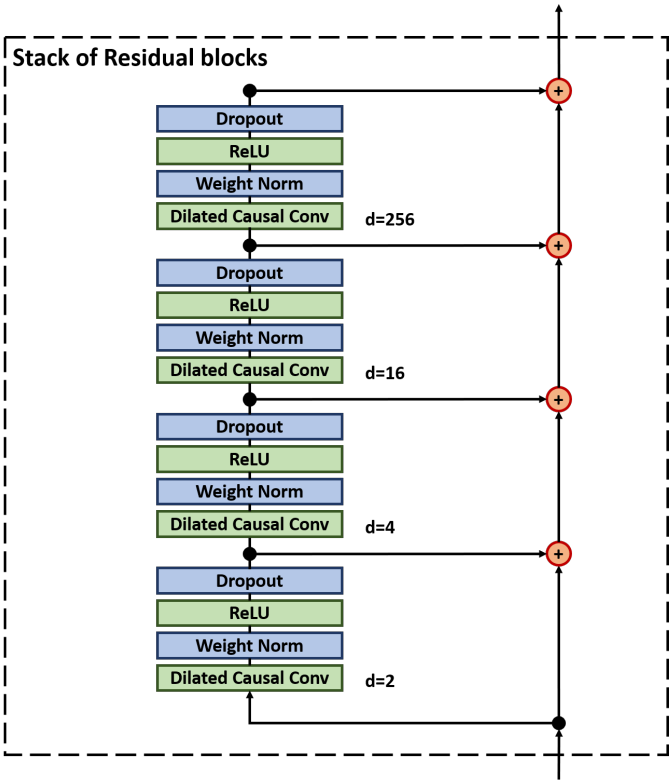

Figure 13. One stack of our chosen TCN architecture. As shown, the stack consists of four separate layers of convolutional filters, which are progressively dilated to provide a wide receptive field. The number of filters in each layer and the overall number of stacks are two hyper-parameters that determine the overall model capacity. As shown, each layer utilizes a Rectified Linear Unit (ReLU) activation function, and employs two forms of regularization: weighted normalization and Dropout.

Hyper-parameter Search Vectors.

Fixing this basic architecture, we engage in a limited hyper-parameter search over two general vectors: the optimal shape for the exponential function, and the optimal

capacity for the neural network.

Optimization over the decay rate of the exponential, λ, was varied across 3 choices, {0.015 Hz, 0.02 Hz, 0.04 Hz}, selected based on visual inspection. Optimization over model capacity was conducted across two parameters, number of stacks and number of filters. Each parameter was varied across 4 choices, {2, 5, 9, 12} and {5, 10, 15, 20} respectively, ranging from a minimal capacity network (2 stacks with 5 filters and only 3,517 parameters) to a high capacity network (12 stacks with 20 filters and 328,681 parameters). Because these two parameters are highly interrelated, the search was conducted exhaustively, for a total of 16 models. The final hyper-parameter selections were based on validation loss curves.

Evaluation Criteria.

The research objective is to determine the maximum achievable recall of our single-trace detection algorithm against the array beam catalogs. Because recall is a classification metric, and because the task is a regression problem, the next step is to define the method for calculating recall.

Each detection window is 4 seconds, identical to the window length used in [86]. The number of Total Positives is the number of labeled arrivals in the dataset, and the number of Total Negatives is the number of windows (length of the dataset in seconds divided by 4) minus the number of Total Positives, which is a conservative estimate. A predicted arrival is any peak in the output sequence with a value above a threshold. A True Positive is any predicted arrival within 2 seconds before or 2 seconds after a labeled arrival, and a False Positive is any predicted arrival not within 2 seconds before or after a labeled arrival. A False Negative is a labeled arrival not within 2 seconds of any predicted arrival, and thus the count of True Negatives is the Total Negatives minus False Negatives. From these definitions, standard equations (3) are

used to calculate recall (true positive rate) and Type-I error (false positive rate):

$$\text{Total Windows} = \frac{\text{Dataset Length}}{\text{Window Length}}$$

$$\text{Total Positives} = \#\text{ of Cataloged Arrivals}$$

$$\text{Total Negatives} = \text{Total Windows - Total Positives} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

$$\text{alpha} = \frac{\text{False Positives}}{\text{Total Negatives}}$$

Using these definitions, and treating the analyst-reviewed array beam catalogs as ground truth, performance is reported in terms of both receiver operating characteristic (ROC) curves and recall. When reporting recall, a type-I error rate of 0.1% is used. It should be noted that this is an order of magnitude lower than the error rate reported in [81], [87] and [89], as it is more appropriate for operational use. Because the primary goal is weak-signal detections, recall is also reported as a function of signal to noise ratio (SNR). SNR is defined as the log ratio between the short-term and long-term average power, as given in Eq. (4), with a short-term window consisting of 5 seconds after the arrival, a long-term window consisting of 40 seconds before the arrival, and a bandpass filter applied from 1.8 to 4.2 Hz.

$$SNR = 10 * \log_{10}\left(\frac{PWR_{STA}}{PWR_{LTA}}\right) \tag{4}$$

Additionally, in order to assess the value of our algorithm over existing single-

trace methods, performance is compared against two common automatic detectors, FBPicker [70] and KTPicker [79]. These detectors are implemented in the PhasePApy [22] package for python, which was developed by the Oklahoma Geological Survey, and has been in operational use there since 2015. The three algorithms are then compared by detector efficiency, arrival time estimation, and overall computation time.

2.5 Results

Two hyper-parameter search vectors were optimized in the model: exponential decay and model capacity. As shown in Table 2, decay rates between 0.015 Hz and 0.040 Hz were explored, and a decay rate of 0.020 Hz yields the highest recall on the validation set.

Table 2. Decay Rate Optimization.

λ (Hz)	Recall ($\alpha = 0.1\%$)	MAE (s)
0.015	62.2%	0.640
0.020	72.1%	0.560
0.040	71.3%	0.476

Fixing the decay rate at 0.020, the overall capacity of the model is varied by increasing both the number of residual stacks, s, and the number of 1D convolutional filters, f. Total training time for each model was approximately 200 hrs on an Nvidia GTX 1080 Ti, and the results of this search indicated that model capacity is optimized with 12 stacks and 15 filters, as increasing capacity beyond this point appears to have marginal value. This yields a final model with 12 residual stacks as shown in Fig. 13, with 15 filters on each 1D convolution, for a total of 185,311 fully convolutional parameters.

Table 3 shows the results of evaluating the final model against the hold out test set. Across the seven arrays, the detector is able to correctly classify 56% of the 141,095 array beam picks, yielding 78,802 correct detections. This is a 35% improvement over the 58,515 detections found by the KTPicker, and more than double the 37,572 detections found by the FBPicker for the same period.

The ROC curves shown in Fig. 14 further illustrate the performance of the algorithm. It should be noted that the type-I error rate of approximately 0.1% represents performance to the left of the elbows of the ROC curves and sub-optimal detector efficiency. For pure academic exercise, a much better choice would be to relaxing the type-I error rate to 1%, as observed in other recent works [87, 81]. This increases the overall recall of DeepPick to 77%. Unfortunately, such a large type-I error rate is not acceptable for operational use, as it represents far too many false positives for a human analyst to deal with. Appendix B presents the performance of the algorithm on several example waveforms, comparing it to FBPicker and KTPicker.

Table 3. Algorithm Efficiency by Station. The efficiency of each algorithm (DeepPick, FBPicker and KTPicker) is shown for each of the seven stations for a full year. The first column contains the total number of events found in the corresponding array beam catalog. The subsequent columns contain the detections (true positives), and recall (true positive rate) and false positive rate for each of the algorithms. The last row of the table gives the overall results of each algorithm against the combined catalog across all seven arrays.

	Catalog	DP Picks			FB Picks			KT Picks		
STA	Events	TP	TPR	FPR	TP	TPR	FPR	TP	TPR	FPR
TXAR	16451	9265	57%	0.1%	2933	18%	0.2%	6040	37%	0.2%
PDAR	12980	6966	54%	0.1%	2118	17%	0.3%	3691	29%	0.1%
ILAR	20769	10269	50%	0.2%	3677	18%	0.5%	6371	31%	0.2%
BURAR	4645	3685	80%	0.1%	1565	34%	0.4%	2679	58%	0.1%
ABKAR	8072	5940	74%	0.2%	4015	50%	0.4%	5951	74%	0.2%
MKAR	40583	24473	61%	0.1%	14118	35%	0.2%	20031	50%	0.1%
ASAR	37595	18204	49%	0.2%	9146	25%	0.5%	13752	37%	0.3%
TOTAL	**141095**	**78802**	**56%**	**0.1%**	**37572**	**27%**	**0.3%**	**58515**	**42%**	**0.2%**

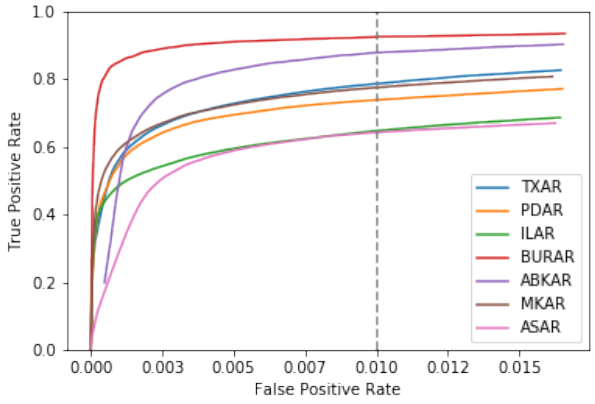

Figure 14. Receiver Operating Characteristic Curves for each of the seven arrays in the hold-out test set. A dashed line is shown in grey, indicating an alpha of 1%.

The objective is to detect weak, distant events. This requires a detector with enough sensitivity to pick out signals near the noise floor. In order to explore the algorithm's performance at this task, its ability to detect signals with very low signal to noise ratio is evaluated. Using the array beam catalog as a baseline, Fig. 15 depicts recall as a function of SNR. This demonstrates that DeepPick maintains a more than 90% recall for signals with an SNR of at least 10 dB for each of the seven arrays in the test set. Signals with an SNR of 10 dB or below are quite difficult to detect from a single trace, as evidenced by the dashed lines in the plot, which represent the detections two other detection algorithms, FBPicker and KTPicker. These graphs indicate that DeepPick maintains at least a 4 dB advantage in sensitivity over both of the other detection algorithms across all seven arrays.

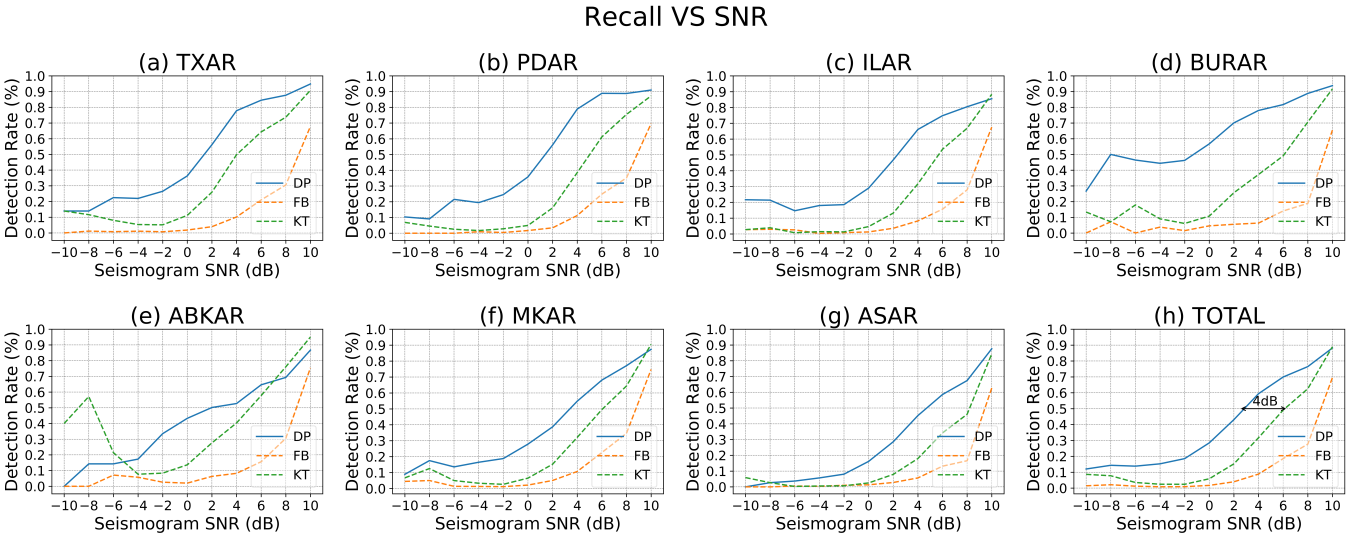

Figure 15. Test-set Recall, reported as a function of SNR, at a fixed type-I error rate of approximately 0.001. Results are compared directly between the three algorithms, DeepPick (DP), FBPicker (FB), and KTPicker (KT). Note that several of the reference catalogs contain fewer arrivals below -8 dB SNR, resulting in some irregularities to the far left of the plots.

Finally, we report the algorithm's performance for the arrival time estimation task

as detailed in Table 4[1]. Here, the algorithm achieves a mean average error of 0.45 seconds from the analyst picked arrival times, with a distribution detailed in Fig. 16. This plot shows that while the most common histogram bin corresponds to an absolute error of less than 0.025 seconds, the weakest signals are frequently missed by more than a second. This error is on par with other automatic detectors as shown in Table 4.

Table 4. Algorithm Precision by Station. Showing the mean average error (in seconds) for the arrival time estimates of each algorithm. The final row shows the average error across all seven arrays.

STA	DP	FB	KT
TXAR	0.447	0.531	0.747
PDAR	0.468	0.487	0.768
ILAR	0.450	0.488	0.690
BURAR	0.477	0.481	0.643
ABKAR	0.384	0.420	0.592
MKAR	0.407	0.443	0.657
ASAR	0.484	0.538	0.692
TOTAL	**0.445**	**0.484**	**0.684**

[1]Arrival time error, Δt, is only reported for true positives ($\Delta t < 2s$).

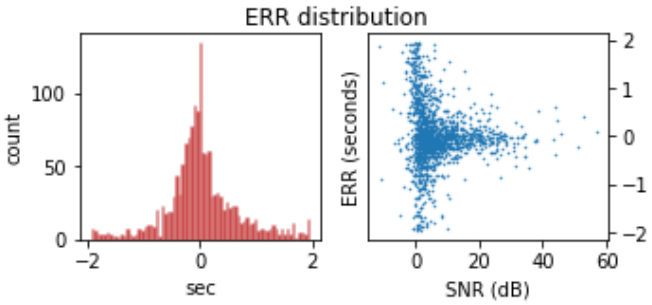

Figure 16. Residual analysis on the errors for the arrival time estimation task. Left: histogram showing the distribution of arrival time errors made by the algorithm against the test set, with a bin width of 0.025 seconds. Right: scatter-plot showing the distribution of errors with respect to SNR.

2.6 Discussion

The results in Table 3 demonstrate that the DeepPick algorithm is capable of achieving a recall of between 50 and 80% against the analyst-reviewed picks from seven array-beam catalogs with a type-I error rate of approximately 0.1%. The low end of this range, 49% recall at ASAR, represents a significant improvement over the performance of existing single trace algorithms (25% and 37% for FB and KT respectively). However, the spread in results is quite large, and suggests the need to examine the underlying cause of this performance variance.

The two stations with the worst performance are ILAR and ASAR. Interestingly, these two stations also utilize a different sensor, the Guralp CMG-3TB, from the other five stations, which all use the Geotech KS54000. This shows the importance of training the algorithm on stations with the same instrument type as the stations for which the algorithm is intended to be deployed against operationally. The two stations with the best results are ABKAR and BURAR. Interestingly, due to higher noise levels at these sites, the array catalogs for these two stations contain relatively fewer events with relatively larger magnitudes. This makes the detection of these

events easier, and the recall rates of 74% and 80% reflect this fact. PDAR, TXAR, and MKAR utilize a common instrumentation, share similar geology and have similar noise levels; as expected, they also share similar recall rates of 54%, 57% and 61% respectively.

The computational efficiency of our algorithm is measured in run-time (seconds) required to build an automatic catalog across a full year of data. Table 5 shows that DeepPick has an order of magnitude increase in computational efficiency over the FBPicker and more than two orders of magnitude increase over the KTPicker. It should be noted that the implementations of FBPicker and KTPicker used here are actual operational implementations used by the Oklahoma Geological Survey. This illustrates the extreme efficiency of the DeepPick algorithm.

Table 5. Algorithm Computational Efficiency by Station. Here we detail the runtime, in seconds, required for each algorithm to process the full year of data at each array.

STA	DP	FB	KT
TXAR	763	22,800	257,800
PDAR	781	18,961	259,243
ILAR	735	19,372	251,210
BURAR	767	22,983	262,368
ABKAR	791	22,913	254,185
MKAR	754	22,838	271,536
ASAR	725	19,059	255,829
AVG	**759**	**21,275**	**258,881**

These results show that the primary determinant of algorithm success lies in the degree of similarity between the training stations and the testing station. As such, when deploying this algorithm for operational use it is important to find suitable

arrays to train on in order to maximize performance. In any case, the algorithm shows decent performance even when trained across different geographical areas and sensor types.

2.7 Conclusion

Weak teleseismic event detection is normally only possible using an array of seismic instruments and sophisticated processing techniques. Even recent works in the literature make little attempt to extend single-trace detection algorithms beyond local events. This is primarily due to the lack of available training data, an issue which we address by mining the seismic catalogs in a unique way, building our catalog for an array beam while taking our event waveforms from a single array element. Using this training data, temporal convolutions and a unique exponential sequence tagging function, we develop a powerful tool for weak signal teleseismic detection. The DeepPick algorithm is able to accurately detect twice the number of events detected by the STA/LTA algorithm commonly used, and does it significantly faster.

The findings in this work represent an important step forward in the field of teleseismic detection, demonstrating that accurate teleseismic event detection is possible from a single seismic instrument. The DeepPick algorithm has the potential to open up thousands of additional automatic detections to single-instrument seismic stations each year, without the need for additional sensors and equipment.

There is still potential for much improvement. In this work, we develop a single-trace detector, applied only to a single channel of data from a three channel instrument; future work could extend our results to include data from all three channels of the instrument. Furthermore, an application of the same technique to an entire array of channels could also prove interesting, and the potential exists to improve our results significantly by incorporating more channels of data. Additionally, the focus of this

work has been primarily centered on producing a detector with increased sensitivity and recall, whereas future work could focus on using similar techniques to produce a detector with an even lower false positive rate.

III. Study 2 - Beyond Correlation: A Path-Invariant Measure for Seismogram Similarity [29]

3.1 Abstract

Similarity search is a popular technique for seismic signal processing, with template matching, matched filters and subspace detectors being utilized for a wide variety of tasks, including both signal detection and source discrimination. Traditionally, these techniques rely on the cross-correlation function as the basis for measuring similarity. Unfortunately, seismogram correlation is dominated by path effects, essentially requiring a distinct waveform template along each path of interest. To address this limitation, we propose a novel measure of seismogram similarity that is explicitly invariant to path. Using Earthscope's USArray experiment, a path-rich dataset of 207,291 regional seismograms across 8,452 unique events is constructed, and then employed via the batch-hard triplet loss function, to train a deep convolutional neural network which maps raw seismograms to a low dimensional embedding space, where nearness on the space corresponds to nearness of source function, regardless of path or recording instrumentation. This path-agnostic embedding space forms a new representation for seismograms, characterized by robust, source-specific features, which we show to be useful for performing both pairwise event association as well as template-based source discrimination with a single template.

3.2 Introduction

Seismograms are time-series records of the earth's motion at a fixed station. This motion results from seismic waves that have often traveled a considerable distance from the source event, and seismograms reflect the combined influence of both the source itself and the propagation path between source location and recording sta-

tion [16]. As illustrated in Fig. 17, two seismograms depicting different events yet sharing a common path can appear similar. This fact has long been recognized by the seismic community [98, 58]. In the earliest days of manual processing and helicorders, analysts were often able to identify mining events from a particular mine, recorded at a particular station, by simply comparing the visual similarity of new seismograms to previously recorded examples [50]. In fact, a common practice was to take two translucent paper seismograms and compare them, by passing the waveforms across one another while holding them up to a light source [94]. Thus began the science of seismogram similarity. Of course, the advent of computer processing ushered in the development of a multitude of techniques to exploit these similarities algorithmically. Case-based discrimination [32], template matching [36], waveform correlation [43], subspace detection [44] and similarity search [113] are all similarity-based algorithms which have been proposed over the last several decades, and deployed against a wide range of seismic signal processing tasks, such as discriminating mining blasts, screening swarm events, identifying aftershock sequences, and even detecting general seismic signals.

While these algorithms have different tasks ranging from discrimination to detection, fundamentally they are all examples of similarity-based classifiers [23], which estimate the class label of a new seismogram based on its similarity to one or more previously labeled templates. Furthermore, these similarity-based classifiers all share a common measure of similarity: cross-correlation. Such methods are generally referred to as correlation detectors [44].

This common reliance on correlation is concerning, because the correlation coefficient of two seismograms is dominated by path effects [94], as demonstrated in Fig. 17. While path-dominant similarity can be desirable, such as when detecting aftershock sequences from a particular fault, or mining blasts from within a small

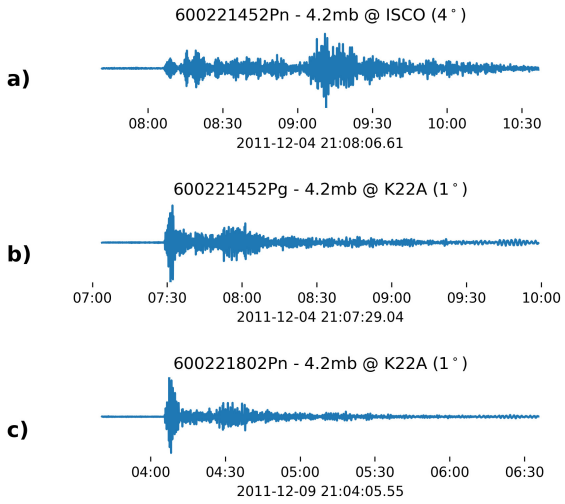

Figure 17. Three seismograms depicting explosions at a coal mine near Thunder Basin, WY. Seismograms a) and b) depict a common source event (600221452), recorded at two separate seismic stations, ISCO and K22A respectively. Seismogram c) depicts a nearby event (600221802), also recorded at K22A. Seismograms a) and b) depict the same event recorded at different stations, while seismograms b) and c) depict different events which share a common path. The correlation between the same-source waveforms a) and b) is only 0.03, and the waveforms visually appear quite different. On the other hand, the visual similarity between the path-similar waveforms b) and c) is obvious, and they are correlated with a coefficient of 0.18. This illustrates the path-dominant similarity inherent to seismogram correlation.

quarry, in general, path-dominant similarity is problematic, as source-similar signals de-correlate with even slight deviations in path [44]. This includes deviations in origin location, such as two explosions occurring at different points in a mining quarry, and deviations in recording location, such as two recordings of the same explosion by separate seismic stations in a regional seismic array. In either case, path differences of even just a quarter wavelength can significantly degrade the correlation of two seismograms [20, 75].

This work presents a new measure for seismogram similarity that bypasses correlation entirely, and is designed to be both path-invariant and source-specific. To be precise, the design goal is to create a measure of seismogram similarity that enables the identification of seismograms sharing a common source event, regardless of the path of travel. While such a measure was previously computationally intractable, it is

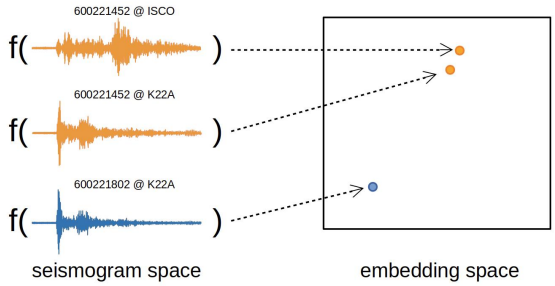

Figure 18. Path-Invariant Embedding Function for Seismograms. The embedding function, $f(\cdot)$, is a non-linear transformation that maps time-series seismograms to low-dimensional embeddings. The mappings should be path-invariant and source-specific, such that regardless of the recording station, all seismograms associated with a particular event are mapped closely in the embedding space, and seismograms not associated with that event have more distant embeddings, as demonstrated in this notional diagram. This embedding function can be learned using a convolutional neural network architecture, trained with seismogram triplets.

possible with the careful application of deep convolutional neural networks (CNNs). In 2019, researchers at the Los Alamos National Laboratory published a method using a CNN to predict the pairwise association of seismic phase arrivals, for 6 second windows, across a local group of 6 stations in northern Chile, reporting an accuracy of over 80% [72]. Building on these results, we construct a source-dominant, path-invariant measure for seismogram similarity which operates on 180 second windows and is generalized across more than 1,000 sensors across North America. We do this by utilizing a state-of-the-art machine learning technique from the field of facial recognition, called a Triplet Network, which not only indicates pairwise association between seismograms, but actually maps the seismograms to low-dimensional vectors, called embeddings, such that the embedding space distance between seismograms sharing a common source event are minimized, regardless of path, while remaining distinct from any other events. This embedding strategy is displayed in Fig. 18. In this way, the embedding function becomes a rich feature extraction technique for source-specific and path-invariant features.

The triplet network architecture accepts three observations - two similar and one

different from the others. Training a triplet network to learn seismic source similarity requires source-similar seismogram triples: two of the three waveforms are associated with a common source event and the third waveform is not. For this task, it is preferable to have a training set containing seismograms recorded from a densely-spaced sensor network, so that the neural network can experience seismograms recordings across numerous paths for the same event. The 400 three-channel broadband sensors of the USArray experiment provided an ideal dataset of seismograms; data from this array is used for training and testing. The triplet network is trained against 13 years of data (2007 - 2013), validated against a single year of data (2014), and tested against the final two years of data (2015-2016). Additionally, a subset of 51 recording stations and a small region of event locations was held out from the algorithm during training, to allow a proper evaluation of the generalizability of the technique. A map detailing the dataset is shown in Fig. 19.

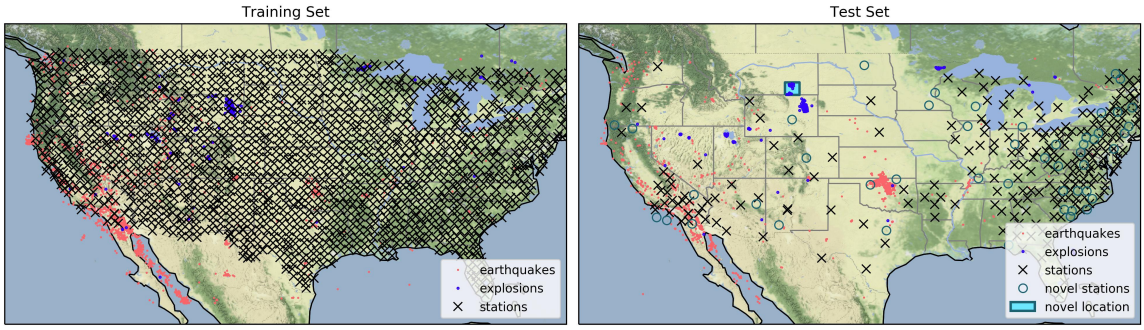

Figure 19. Map showing the geographical location of each recording station and event in the training and testing datasets. The majority of the stations were installed as part of the Earthscope's Transportable USArray, and were in operation from 18 to 24 months before being moved. Additionally, 51 novel stations and a small region of novel event locations are unique to the test set.

The value of this path-invariant measure is demonstrated through performance evaluation on two common seismic tasks: event association and source discrimination. The event association task of determining whether or not two waveforms depict the same event achieves a binary accuracy of 80%. This accuracy is achieved using

only the waveform characteristics, without information on times or recording locations, and the technique has strong potential to augment existing methods of event association [72].

The real promise of the technique, however is for source discrimination. The embedding space is a rich basis for source-specific seismic feature extraction [41]. Our similarity-based explosion discriminator achieves 95.8% accuracy with no explicit training for the source discrimination task; the discriminator simply compares the similarity of unknown waveforms to a single randomly-selected explosion template. This technique is often referred to as one-shot learning [61], and shows promise for discrimination of novel sources when only a few extant templates are available.

In the remainder of this work, these contributions and conclusions are explored in detail, by reviewing the related literature, outlining methodology, and detailing and discussing the results.

3.3 Background

This work merges two relatively disparate fields of science. On the one hand, the application is seismogram similarity, a field with a rich history and considerable previous research. On the other hand, the methodology employs learned similarity, a relatively nascent field that has principally been associated with machine learning image processing applications. This background section is divided into three distinct subsections: seismogram similarity; learned similarity; and learned seismogram similarity. Each subsection contains a brief background and literature review, as well as a discussion of the limitations and gaps in the current research, which this work attempts to fill.

Seismogram Similarity.

A seismogram represents the composition of several effects, including the seismic source itself, the propagation path from the source to the seismometer, the frequency response of the seismometer as well as any ambient noise at the seismometer's location [16]. Because of this diverse composition, estimating and even defining seismogram similarity can be quite challenging.

The traditional measure for seismogram similarity is the cross-correlation function. This measure has been used for detecting and discriminating seismic signals since the late 1980s [32], and such techniques are commonly referred to as correlation detectors [44]. Correlation detectors are exquisitely sensitive, allowing detections near the noise floor for known repeating events in highly confined geographical regions [37]. Unfortunately, this confinement is also a limitation, as seismogram correlation has been shown to decay exponentially with even minor differences in path distance [50]. In fact, early research suggested that correlation-based similarity was limited to signals with hypo-centres separated by no more than a quarter wavelength [33, 75], although later efforts have since shown improvements, allowing the correlation length to be up to two wavelengths [43]. Additionally, researchers have also shown that seismograms quickly decorrelate across small variations in mechanism and source function [49]. These facts limit the applicability of the correlation detector to only the most repetitive sources that are confined to localized geographical regions [44].

To increase the applicability of the correlation detector, there have been numerous adaptations proposed. To address variations in ambient noise, narrow bandpass filters were applied [50]. To address minor variations in mechanism, composite templates were employed, derived from linear combinations of several master templates representing a range of mechanisms [43]. To address path effects, dynamic waveform matching was developed, introducing a non-linearity to the correlation, allowing rel-

ative stretching or squeezing of the template [94]. Subspace detectors attempt to address all of these variations at once, with even more robust composite templates [44]. Recently, efforts focused on a multiplicity of templates and a computationally efficient search across them [113, 116, 9, 11]. These efforts have significantly increased the effectiveness of correlation-based detectors. In fact, for regions with a high sensor density, such as Northern California, it is estimated that more than 90% of events have sufficient similarity to be detected via correlation [107]. However, this figure is highly dependent on both the density of the sensor network and the completeness of the template library [103]. As such, Dodge and Walter estimate that still only 18% of all global events possess sufficient similarity to be detected by these methods [31].

In summary, cross-correlation is a powerful measure for seismogram similarity, especially as a tool for detecting highly-repeating path-specific events. However, cross-correlation is fundamentally limited as a general measure of seismogram similarity, due to its inherent path-dependence. In this study, we address this limitation directly, and propose an alternative measure of seismogram similarity that is invariant to path, instrumentation and ambient noise.

Learned Similarity.

Each of the traditional seismogram similarity measures discussed so far has been fundamentally built around the cross-correlation function. However, it is interesting to note that almost none of those measures performed cross-correlation directly on the raw waveforms. Instead, each measure first applied some pre-processing function to the raw waveforms, either linear (time shifts, bandpass filters, linear combinations) or non-linear (dynamic time warping) prior to performing cross-correlation. We can generally understand these pre-processing functions to be mappings, from raw waveform space to a new *embedding space*. In each case, the mapping function is chosen

such that the cross-correlation of two objects in the embedding space meets some desired similarity objective.

As it turns out, this embedding process used in traditional correlation-based similarity closely mirrors the process accomplished in machine learning-based similarity. For learned similarity, a parameterized embedding function architecture is established, and the parameters are optimized such that the distance between two objects in the space achieves the desired similarity objective. Over the last several years, such learned similarity measures have revolutionized the field of facial recognition in particular and the field of image processing in general, fueling advances in image recognition [108], object tracking [66] and even vision navigation [64]. In the remainder of this section, we review some of the state of the art techniques available for constructing deep learned similarity measures, focusing particularly on the embedding function architecture and similarity objective, in turn.

Embedding Function Architecture.

Many early efforts to create learned similarity spaces utilized a linear architecture, such as the Mahalanobis distance [111, 51, 52]. However, in recent years, much success has been gained by employing non-linear architectures [10], particularly in the form of deep convolutional neural networks (CNNs) [41]. These CNNs were originally developed with 2-dimensional kernels, or filters, which allowed them to closely model the hand-crafted kernels traditionally used in image processing [67]. To adapt these powerful CNN architectures to process time-series waveforms, 1-dimensional CNNs were developed [18], enabling learned similarity spaces for audio waveforms [54].

A more recent advancement to the traditional CNN architecture is the Temporal Convolutional Network (TCN), which is characterized by layered stacks of dilated causal convolutions and residual connections [7], as illustrated in Fig. 20. Such an

architecture is particularly applicable to time-series waveforms with long-period dependencies, and offers several distinct advantages for seismic feature extraction [28], including:

- Residual connections allow the model to have high-capacity and stable training.
- Dilated convolutions allow precise control over the receptive field.

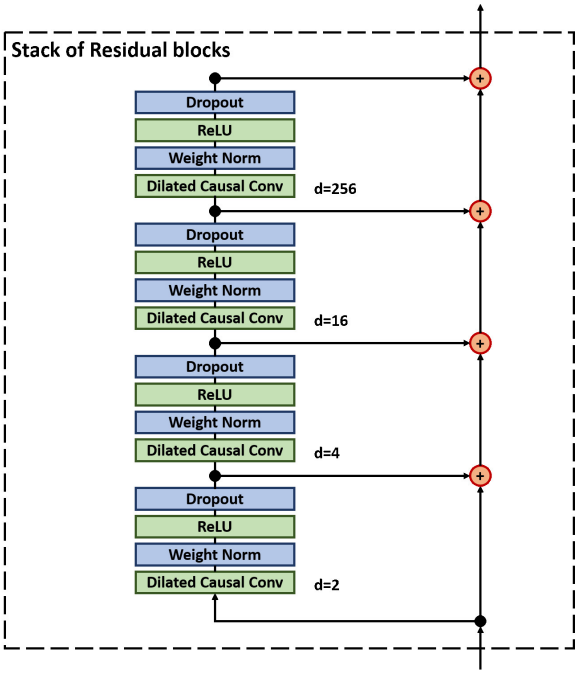

Figure 20. A single stack of 4 dilated residual blocks commonly found in a Deep Temporal Convolutional Neural Network Architecture. In this case, the residual blocks have exponentially increasing dilation rates, increasing from 2 to 256 across the 4 blocks. This rapid dilation provides the network a wide receptive field which is critical for learning long-period features frequently found in time-series waveform data.

The receptive field is of primary importance for time-series modeling, as it explicitly limits the learnable feature periodicity at a given layer. The equation for calculating the receptive field, r, for a given convolutional layer, l, kernel size, k, and dilation rate, d is given in (5):

$$r_l = r_{l-1} + d_l(k-1)$$
$$\text{where } r_0 = 0 \tag{5}$$

In summary, the TCN is ideally suited for the efficient embedding of seismograms. This architecture presents a rich search space for learning an optimal embedding function. However, optimizing this function requires defining a suitable similarity objective, detailed next.

Similarity Objective.

Defining a quantitative similarity objective begins with a qualitative understanding of what similarity means for the given task, which is often referred to as a semantic definition of similarity. Once the semantic definition is established, the next step is to approximate it with an embedding function, such that nearness in the embedding space implies the semantic similarity [26]. This embedding function is learned via back-propagation of loss, $\mathcal{J}$, that reinforces the semantic definition.

One of the simplest semantic definitions of similarity is the concept of a *match*, where a matched pair of objects share a common identity, and an unmatched pair of objects have different identities. For example, in the facial recognition task, a matched pair is defined as two images of the same person and an unmatched pair is defined as two images of distinct persons. The similarity objective is to optimize the parameters of the embedding function such that the embedding space distance between matched pairs is small, while the distance between unmatched pairs is large. This embedding function can be learned directly by a Siamese Neural Network, which takes in a batch of m object pairs, of which half are matched, and half are unmatched. The two objects, $X_A^{(i)}$ and $X_B^{(i)}$, are then embedded via twin copies of the embedding function, $f(\cdot)$, with tied parameter weights w. The parameters of the embedding function

are updated via the contrastive loss function, which penalizes two contrasting cases: matched pairs are penalized for being embedded too far apart and non-matched pairs are penalized for being embedded too close together with respect to some margin, α, as given in Eq. (6) and Eq. (7), respectively [26], where $[\ ]_+$ indicates the ramp function[1].

$$\mathcal{J} = \sum_{i=1}^{m/2} \left[\left\langle f(X_A^{(i)}), f(X_B^{(i)}) \right\rangle \right]_+ \tag{6}$$

$$\mathcal{J} = \sum_{i=1}^{m/2} \left[\alpha - \left\langle f(X_A^{(i)}), f(X_B^{(i)}) \right\rangle \right]_+ \tag{7}$$

This technique works well, however, one drawback is the relatively inefficient use of the embedding space. Matches are too greedy, as the Siamese Network attempts to map all matches to a single point in the space. Meanwhile, non-matches are inefficient, being pushed apart by only a fixed distance [48]. As a result, the Siamese Network is used less frequently in favor of the Triplet Network.

The Triplet Network is similar to the Siamese Network [48], however it is trained on batches of m triples, where each triple is comprised of an anchor object, $X_A^{(i)}$, a positive object, $X_P^{(i)}$, and a negative object, $X_N^{(i)}$. From within each triple, both a matched and non-matched pair can be constructed, however, the triplet loss function computes the relative embedding distance between the matched pair and non-matched pair, and no loss is accrued as long as the matched pair is closer by some margin, α, as given in Eq. (8).

[1]The ramp function simply zeros out all negative values while passing positive values unchanged.

$$\mathcal{J} = \sum_{i=1}^{m} \left[\left\langle f(X_A^{(i)}), f(X_P^{(i)}) \right\rangle - \left\langle f(X_A^{(i)}), f(X_N^{(i)}) \right\rangle + \alpha \right]_+ \tag{8}$$

where $[\ ]_+$ indicates the ramp function.

The Triplet network avoids the greediness of the Siamese network, and makes more efficient use of the embedding space, however it has its own drawbacks. Particularly, it can converge quickly at first, but learning slows rapidly, as the majority of the negative pairs are pushed beyond the margin, failing to train the weights appreciably. This can be solved by sampling hard pairs, semi-hard pairs and several other sampling strategies, all of which rely on iterative processing via forward propagation to determine embedding space distances, selectively sampling based on those distances, and then applying back propagation on the sample [46]. The algorithm used to sample hard pairs is commonly referred to as the *batch hard* loss function, and it requires that each batch be composed by randomly sampling L distinct identities and then randomly sampling K examples of each identity. In this way, the total number of objects in a batch is $L * K$, and each object is double indexed so that object $X_u^{(v)}$ represents the u_{th} example of the v_{th} identity. The triplet loss is calculated using Eq. (8), except that in this case, every object in the batch is treated as an anchor $X_A^{(i)}$, and used to form a new triplet by selecting the *hardest* positive and *hardest* negative samples, $X_P^{(i)}$ and $X_N^{(j)}$ respectively, for that anchor within that batch, as detailed in Eq. (9).

$$\mathcal{J} = \overbrace{\sum_{i=1}^{L} \sum_{A=1}^{K}}^{\text{all anchors}} \left[\overbrace{\max_{\substack{P=1\ldots K \\ P \neq A}} \left\langle f(X_A^{(i)}), f(X_P^{(i)}) \right\rangle}^{\text{hardest positive}} - \overbrace{\min_{\substack{j=1\ldots L \\ N=1\ldots K \\ j \neq i}} \left\langle f(X_A^{(i)}), f(X_N^{(j)}) \right\rangle}^{\text{hardest negative}} + \alpha \right]_+ \tag{9}$$

where $[\ ]_+$ indicates the ramp function.

Deep Seismogram Similarity.

Deep Neural Networks are now being used across many areas of seismological research, from earthquake detection to earthquake early warning systems, ground-motion prediction, seismic tomography, and even earthquake geodesy [62]. However, no effort has yet been made to use deep neural networks to build a seismogram similarity metric. The closest related work was in early 2019, where researchers at Los Alamos National Labs published a paper describing a convolutional neural network for the pairwise association of seismograms depicting a common event, regardless of path [72]. This work shows that path-invariant features do exist within the seismogram record. The seismograms considered in their work had a signal length of 6 seconds, and were restricted to recordings from 6 seismic stations. To process the signals, they used a shallow CNN with 4 layers, the input accepting two seismograms, the output producing a single Boolean. This results in a similar output to a Siamese network, but without tied weights. The lack of tied weights means there is no embedding layer, which prevents their technique from being used for feature extraction. And the small number of stations limits the generalizability and transportability of their algorithm. Finally, the short signal length (6 s) limits each individual seismogram to containing a single phase arrival, thereby limiting the ability of the model to extract long-period features, such as P and S wave energy ratios, which are particularly pertinent to general source discrimination tasks.

3.4 Methodology

We present a novel seismogram similarity measure, based on a learned embedding function, which is both source-dominant and path-invariant. We show that the resultant embedding space is a rich representation space for seismic signals, useful for performing similarity-based classification against two common class dichotomies for

seismograms: common event vs. different events (event association) and earthquake vs. explosion (source discrimination). The remainder of this section describes the embedding function architecture, the similarity objective, the USArray dataset and the evaluation criteria for the two classification tasks.

Embedding Function Architecture.

The goal is to learn a path-invariant embedding function for seismograms, useful for source discrimination at up to regional distances. This is accomplished using a hybrid architecture with two distinct parts: first, a TCN is employed with a receptive field wide enough to capture both P and S wave phases; second, a densely connected output layer, with 32 nodes, is employed to facilitate a rich low-dimensional embedding space.

Using Eq. (5), the TCN is designed to have an overall receptive field of 4171 samples, equivalent to 104 seconds at the given 40 Hz sample rate of the data, allowing it to learn long-period features down to 0.01 Hz, with just four dilated convolutional layers, as shown in Table 6. The TCN architecture consists of two residual stacks, shown in Fig. 20, each with 50 filters and a kernel size (filter length) of 16 samples. Finally, the TCN output is encoded by a densely connected output layer with 32 nodes, and the final output vector is normalized to have unit length. This results in 553,835 trainable parameters, and a network which reduces the three-channel 21,600 dimensional input into just 32 dimensions, for a 99.9% reduction in dimensionality.

Table 6. TCN Layer Parameters

l	k	d	Receptive Field
1	16	2	31
2	16	4	91
3	16	16	331
4	16	256	4171

Similarity Objective.

This embedding function is learned via a Triplet Network with batch-hard loss. Specifically, the batch size was set at 100, with L (the number of distinct source events in a batch) and K (the number of seismogram recordings for each event in a batch) both set equal to 10. In this way, each batch consists of 100 randomly selected seismograms, evenly represented across 10 different source events. These values were selected primarily on the basis of availability, since increasing K beyond 10 would have limited the dataset ($\sim$95% of the events in the USArray dataset were recorded by at least 10 stations), and increasing L beyond 10 would require more memory than the 12 GB available in the Nvidia 1080Ti GPU used for training.

Embedding space distances are computed using the L_2 norm. Because the output of the embedding function is normalized, the embedding space vectors are all constrained to a hypersphere with radius = 1. This ensures a bounded distance between any two embeddings, as chord lengths are always bounded by [0,2] for any unit hypersphere. Because these pairwise distances are bounded, a fixed margin can be used throughout training [96]. In this work, the margin is fixed at $\alpha = 0.2$, which is common [93].

Data Collection.

Learning a path-invariant measure for seismogram similarity requires a training dataset with many recordings of a single seismic event across many disparate paths. This is best accomplished by a dense network of seismometers across a wide region. EarthScope's USArray dataset is ideally suited for this endeavor. In particular, this work utilizes two EarthScope observatories, the Transportable Array and the Reference Array, as the basis for the Training and Test Sets, respectively.

The USArray Transportable Array (TA) consists of 400 temporary seismic in-

struments that were deployed at more than 2,000 temporary station locations across the Continental US between 2007 and 2015 [19]. Each station utilized a broadband 3-channel (North-South, East-West and Vertical) instrument, installed in a post-hole configuration and digitized at 40 Hz. The instruments were generally one of three types, Guralp CMG3T, Quantera STS, or Nanometrics Trillium; the digitizers were primarily Kinemetrics Q330, Q680 or RefTek. In this work, the training and validation datasets are taken from the full array of TA seismograms, minus a random subset of 51 stations and a region of events located near the Rosebud mine in Montana, which were held out for testing. The training and validation sets were distinct in time, covering the periods from 2007-2013 and 2014, respectively. Associated arrival times were obtained by querying the ISC reviewed catalogs for any Continental US (CONUS) events over this period, resulting in 149,036 seismogram recordings of 4,825 distinct seismic events for the training set, and 22,561 seismogram recordings of 1,175 distinct seismic events for the validation set. A map detailing the layout of the training stations is shown in the left plot of Fig. 19.

The USArray Reference Array consists of 120 permanent seismic instruments deployed across the Continental US, utilizing similar equipment to the Transportable Array. In this work, the test set is taken from the full array of TA and REF stations available from 2015 and 2016. Associated arrival times were obtained by querying the ISC reviewed catalog for CONUS events, resulting in a test set with 35,694 seismogram recordings of 2,452 distinct seismic events. All of the events in the test set are mutually exclusive with the training and validation data. Additionally, because of the stations and locations held out during testing, 6,934 of these seismograms were recorded by the 51 novel stations, and 87 seismograms represent events from the novel location near the Rosebud mine in Montana. Performance is evaluated explicitly on these novel data to explore the power and generalization of the technique.

All three datasets, training, validation and test, were limited to events with near CONUS epicenters, as defined by the following limits on latitude and longitude: 25 < LAT < 50, and -125 < LON < -75. This accomplishes two purposes. First, this produces a catalog with more balanced samples of explosions and earthquakes, 207,291 and 26,568 respectively. Second, this restricts the study to regional signals. Regional signals are preferred due to the more manageable window length requirements vs. teleseismic signals. Furthermore, the regional association task is much more interesting than the teleseismic association task, due to the fact that the teleseismic signals recorded by such a dense regional network look much more similar even using traditional seismic similarity. We leave the exploration of this technique against teleseismic signals to future work. For completeness, we also have included histograms of seismogram station-to-event distances as well as event magnitudes for both the test and training sets, shown in Figs. 21 and 22, respectively.

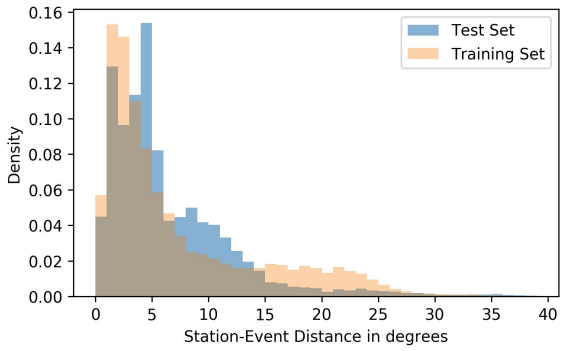

Figure 21. Histogram showing the distributions of station-event distances for all seismograms in the test and training sets. The distributions show that the test and training sets are similar, and that the majority of the seismograms in the combined dataset were recorded within 15 degrees of the epicenter.

For each of the 207,291 seismograms in the combined datasets, a 180-second window is selected which includes the 30 seconds prior to the cataloged arrival time and the 150 seconds subsequent to the arrival. The only pre-processing applied to the raw data was a normalization and de-trending. This window size was chosen so as

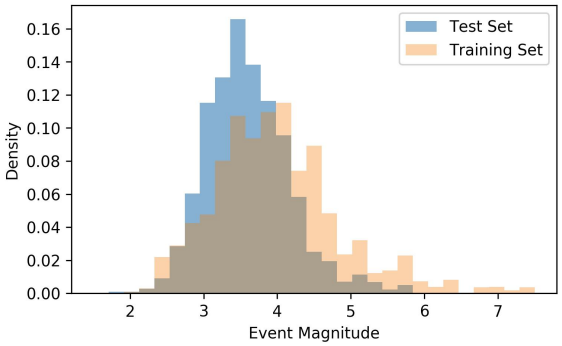

Figure 22. Histogram showing the distributions of event magnitudes for all seismograms in the test and training sets. The distributions show that the test and training sets are similar, and that the majority of the events in the combined dataset have a magnitude between 2 and 5 Mb.

to ensure the presence of both P and S waves within the window. While this long window does present the opportunity for multiple arrivals within a single window, investigation shows that this occurs in only 0.15% of the seismograms in the dataset, and its effects are negligible on the results.

To create the training and validation triples, a generator function randomly selects an anchor, as well as positive (same event, different station) and negative (different) events. Due to multiple site recordings of many of the individual events (on average, each event was recorded by 30 different stations), there are upwards of 300 million possible triples, which makes this a robust training set for learning seismogram similarity.

Evaluation Criteria.

To demonstrate the performance of the similarity measure, it is applied to two tasks: pairwise event association and source discrimination. Evaluation criteria for each of these tasks is shown below.

Event association is the process of correctly associating the arriving seismic phases of a single event across a network, and is a critical step in seismic analysis. The

traditional algorithms used for this task have always been based on travel times and earth velocity models, however our method is similarity-based: we associate the seismograms entirely based on their pairwise similarity in the embedding space, with no external information about arrival times or recording locations. This is a binary classification task: given a pair of seismograms, X_A and X_B, the algorithm must classify the pair as matched or unmatched, where a matched pair is defined as two seismogram recordings of the same event. Classification is accomplished by comparing the similarity-based test statistic, S, against a user defined threshold, τ, as seen in Eq. (10).

$$\begin{aligned} &H_0\text{: UNMATCHED } (X_A \text{ and } X_B \text{ depict distinct events}) \\ &H_A\text{: MATCHED } (X_A \text{ and } X_B \text{ depict a common event}) \\ &S = \frac{1}{\left\langle f(X_A), f(X_B) \right\rangle} \\ &\text{reject } H_0 \text{ if } S \geq \tau \end{aligned} \tag{10}$$

To report performance, a receiver operating characteristic (ROC) curve is built by varying τ across the full range of S, and plotting the rate of false positives against the rate of false negatives for each τ. Additionally, for the threshold τ which maximizes accuracy, area under the ROC curve (AUC), binary classification accuracy, precision and recall are shown. The evaluation is performed across 50,000 random pairs of seismograms drawn from the test set, and compared directly against the results found in [72]. The results are also explored with respect to a subset of novel stations and events that were withheld during training, in order to better understand the abilities and limitations of the technique.

The source discrimination task is also formulated as binary classification, where unlabeled seismograms X are classified as either explosion or earthquake, based on

their embedding space similarities to both the centroid of a set of explosion templates, X_{EXP} and the centroid of a set of earthquake templates, X_{EQK}. This is shown in Eq. (11), where ϵ is machine precision.

$$
\begin{aligned}
&H_0\text{: EARTHQUAKE } (X \text{ depicts an earthquake}) \\
&H_A\text{: EXPLOSION } (X \text{ depicts an explosion}) \\
&S = \frac{\Big\langle f(X), f(X_{EQK}) \Big\rangle}{\Big\langle f(X), f(X_{EXP}) \Big\rangle + \epsilon} \\
&\text{reject } H_0 \text{ if } S \geq \tau
\end{aligned}
\tag{11}
$$

The source discrimination test is performed against the full 35,694 seismograms in the test set. The ROC curve, AUC, accuracy, precision, and recall are presented.

Additionally, the performance of this similarity-based discriminator is directly compared to that of two state-of-the-art methods: the SVM-based discriminator proposed in [63] and the SRSpec-CNN discriminator adapted from the work of [76]. In particular, the SVM and CNN implementations both utilize the full 149,036 training waveforms from the training set. The SVM uses 36 features, composed of nine frequency bins ([1-3 Hz], [2-5 Hz], [4-7 Hz], [6-9 Hz], [8-11 Hz], [10-13 Hz], [12-15 Hz], [14-17 Hz], [16-19 Hz]) and four time divisions (P, P coda, S and S coda), with the S-P time differences based on the iasp91 velocity model. SRSpec-CNN uses 64x64 spectrogram images extracted from 180s normalized seismogram windows, with frequency bins between 2-10 Hz.

3.5 Results

Pairwise Event Association.

To demonstrate that event association using this technique is possible, a special test set is created by sampling 50,000 pairs of seismograms from the test set, in-

cluding 25,000 pairs of seismograms that originate from common events, and 25,000 pairs of seismograms that originate from different events. Plotting histograms of the embedding space distances for each pair, as shown in Fig. 23, demonstrates that the distribution for matched-pair distances are considerably lower than the unmatched-pair distances.

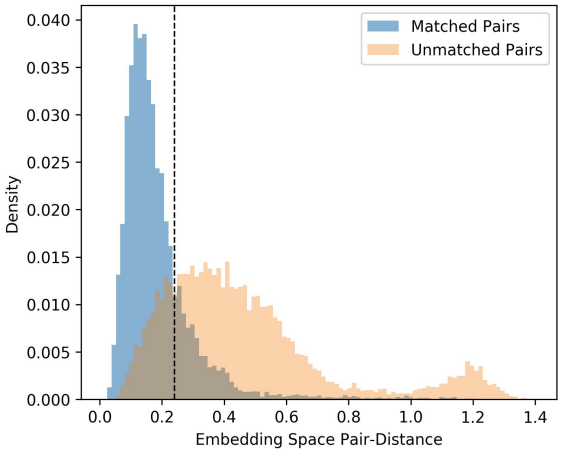

Figure 23. Histograms of matched and unmatched pair distances for the test set. The matched-pair distribution includes embedding space distances for 25,000 pairs of seismograms, where the two embeddings come from the same event. The unmatched-pair distribution includes embedding space distances for 25,000 pairs of seismograms, where the two embeddings come from different events. A cutoff threshold of 0.24 was used to obtain maximum classification accuracy, and is annotated by the dashed line. For this threshold, the area of overlap between the two density plots represents the total classification error, which is ∼20%.

We then apply the similarity-based association classifier defined in Eq. (10). The ROC curve for the task has an AUC of 86.8% as shown in Fig. 24. The overall accuracy is 80.0% with a precision and recall of 80.2% and 79.6%, respectively, and these results are nearly identical to the 80% accuracy reported in [72], extended across a much larger network of stations. Performance is also investigated with respect to the distance between recording stations. As noted previously, correlation-based seismogram similarity is known to decay exponentially with an increase in the distance between recording stations [50]. Our path-invariant measure is also negatively affected by increasing this distance, but the decay is only linear. This is clearly demonstrated

in Table 7.

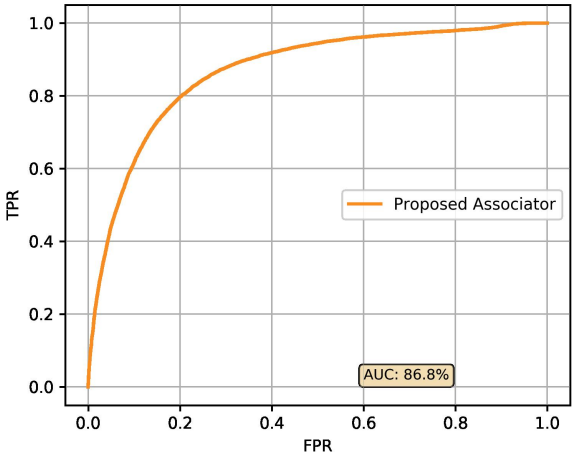

Figure 24. Receiver Operating Characteristic Curve for the Event Association task. The overall area under the curve is 86.8%.

Table 7. Waveform Association Performance vs. Inter-station Distance

Distance (km)	Precision	Recall	Accuracy
0000-0250 km	0.864	0.783	0.830
0250-0500 km	0.852	0.791	0.827
0500-0750 km	0.802	0.766	0.789
0750-1000 km	0.805	0.789	0.799
1000-1250 km	0.778	0.840	0.800
1250-1500 km	0.785	0.811	0.794
1500-1750 km	0.744	0.866	0.784
1750-2000 km	0.731	0.863	0.773
2000-2250 km	0.732	0.794	0.751
2250-2500 km	0.741	0.826	0.769

To further investigate the ability of the embedding space to facilitate event association, Fig. 25 displays 120 seismogram embeddings in 2-dimensions using t-Distributed Stochastic Neighbor Embedding (t-SNE) [71]. The figure clearly demonstrates a clustering of embeddings of common events. However, there are obviously other clusters present as well, shown by the dashed lines in the plot. As it turns out, these other clusters can be quite useful, and are explored further in the discussion of the source discrimination task.

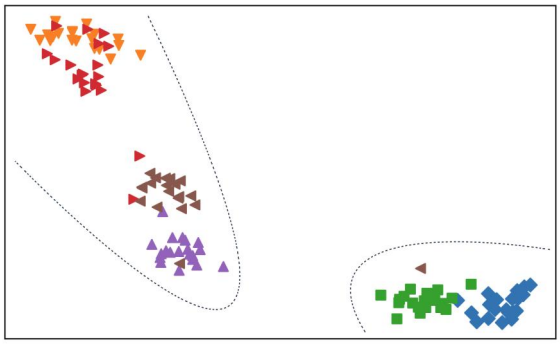

Figure 25. t-SNE Embeddings for Waveform Association. Six unique seismic events were randomly selected from the dataset, along with 20 seismograms for each event, recorded at various stations. These 120 seismograms were then mapped to the 32-dimensional embedding space via the trained neural network. Finally, the 32-dimensional embedding space was visualized here in two dimensions using t-SNE, with each unique event assigned a unique marker. The clustering of same-event embeddings is the result of shared feature commonalities between seismograms of that event. It is interesting to note that there appears to be some aggregate clustering as well, indicated by the dashed lines. This aggregate clustering is the result of feature commonalities shared across seismograms of multiple events. These inter-event commonalities are explored further in the analysis of results for the source discrimination task.

The ability of the embedding space to associate regional events across hundreds of stations with 80% accuracy based entirely on waveform similarity is surprising, and begs the question: is the neural network really extracting generalized path-invariant features, or is it merely 'memorizing' all the training paths exactly, in a way that appears to support conclusions that are unwarranted. To answer this question, we investigate the ability of the embedding space to associate waveforms from novel stations and locations as detailed in Tables 8, 9 and 10. Here, we find that although the performance does drop for such events, the drop is relatively minor. For instance, accuracy only drops from 80% to 79% when considering novel stations, which demonstrates that the neural network has indeed learned to extract features that are invariant to recording location, even novel ones. The accuracy drop is slightly more significant when considering novel event locations, decreasing from 80% to 76% for pairs where at least one event originated near the held-out Rosebud mine. This is understandable, as withholding training events from a certain source location obviously

impairs the ability of the neural network to extract features unique to such events at test time.

Table 8. Association Performance for Novel Stations

Novel STA?	COUNT	ERROR	ACCURACY
No	32221	6358	0.80
Yes	17779	3712	0.79

Table 9. Association Performance for Novel Source Location

Novel LOC?	COUNT	ERROR	ACCURACY
No	49792	10020	0.80
Yes	208	50	0.76

Table 10. Association Performance for Novel Station and Location

Novel STA&LOC?	COUNT	ERROR	ACCURACY
No	49908	10038	0.80
Yes	92	32	0.65

Source Discrimination.

To further demonstrate the power of our embedding space, we consider its utility to facilitate template-based source discrimination. The results here are particularly interesting, as the neural network was not explicitly trained in this task: although the neural network was exposed to many examples of earthquakes and explosions during training (207,291 and 26,568 respectively), the network had no access to these source labels. However, the network did have access to event labels, and was thus trained to extract features with source-specificity and path-invariance. Unsurprisingly, these source-specific features are well-suited for source discrimination. In Fig. 26, the embedding space is visualized using t-SNE, and labeled by source type, demonstrating a significant separation between the two source classes in the embedding space, with no pre-processing or training.

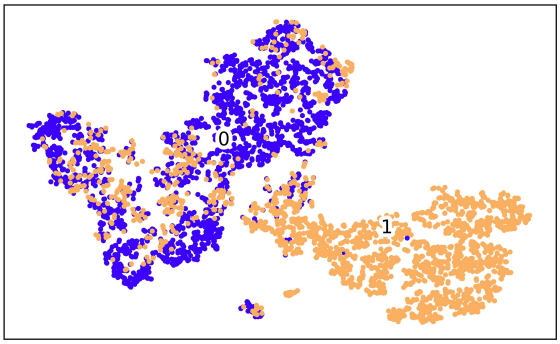

Figure 26. Two hundred embeddings are shown, visualized in 2D using t-SNE, and labeled according to source function. The light-colored dots represent explosions and the darker dots represent earthquakes; the cluster centroids are annotated by 1 and 0, respectively. The 2D clustering of embeddings demonstrates the inherent association between embeddings with a common source function.

Template-based discrimination performance is demonstrated with three different quantities of randomly-selected exemplar templates: 1, 3 and 10, as shown in Fig. 27. The discriminator achieves a mean AUC of 82.8% for just a single template. This is known as one-shot learning, and enables the creation of a viable classification algorithm with only a single training example. The variance on this AUC is a bit high; however with three templates, this method achieves an AUC of 86.7% with low variance. Choosing the threshold so as to maximize accuracy, the algorithm is then evaluated for accuracy, precision and recall, which are recorded at 95.8%, 73.4% and 73.6% respectively, which exceeds the performance of the SVM discriminator, but falls just short of the 96.4%, 78.1% and 77.2% performance achieved by the SRSpec-CNN classifier applied to the same dataset, as detailed in Fig. 28. This minimal performance gap between SRSpec-CNN and our template-based discriminator is surprising, given that SRSpec-CNN is a state-of-the-art fully-supervised method with well-engineered features while our template-based discriminator utilizes semi-supervised learning, with access to just a single template.

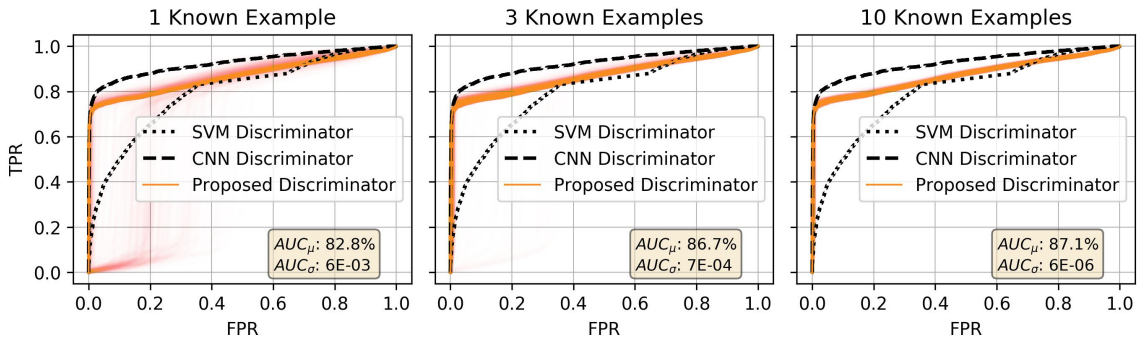

Figure 27. Receiver Operating Characteristic Curves for the Source Discrimination task identifying all explosions. Three plots are shown, demonstrating performance across various numbers of templates (1, 3 and 10). Because the templates are chosen randomly, we have performed 1,000 trials for each plot, with the results of each trial plotted as a separate curve. Performance converges nicely for only 3 templates. The dashed and dotted black lines show the performance of two alternative discriminators applied to the same dataset.

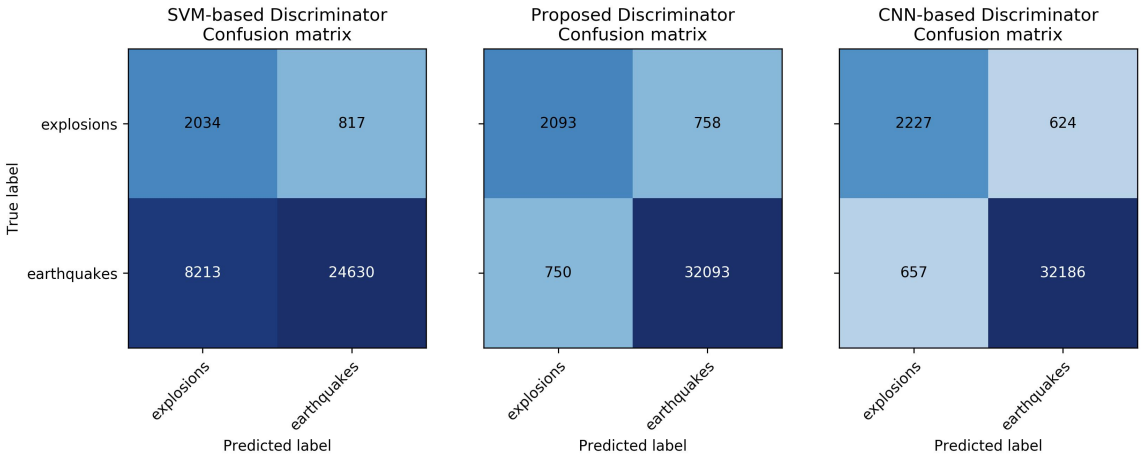

Figure 28. Source Discrimination Confusion Matrix. Three matrices are shown, demonstrating performance of three source discrimination techniques against the test set. Our proposed Similarity-based discriminator utilizes a signal explosion template, whereas the SVM and CNN-based discriminators utilize a large training set with 10,000 labeled earthquakes and 10,000 labeled explosions.

Computation Time.

Optimizing the Neural Network during training requires considerable computation time: approximately 30 hours on the aforementioned Nvidia 1080Ti. However the model only needs to be trained once; after training, deployment is quite fast at runtime, requiring only 1.8 milliseconds to transform a single 180 second window of 3-channel waveform data onto the embedding space. This represents a four-fold improvement over the 7.6 milliseconds required to take the same waveform and extract the spectrogram features used in traditional source discrimination. Runtimes for the Validation and Test sets are shown in Table 11.

Table 11. Comparison of Runtimes against the Validation and Test Sets

	Val Set (22,561 samps)	**Test Set (35,694 samps)**	**Runtime per samp**
NN Embeddings	41 s	66 s	1.8 ms
Spectral Features	171 s	271 s	7.6 ms

3.6 Conclusion

To date, almost all seismogram similarity measures have been based on the cross-correlation function, constraining them to relatively path-dominant similarity, and limiting their use to repetitive and geographically localized signals. In this work, we have presented a path-invariant measure for seismogram similarity, based on a deep triplet network architecture. We have demonstrated the effectiveness of this measure for both pairwise event association and template-based source discrimination.

For the pairwise association task, our similarity measure is able to achieve an accuracy of 80%, without any knowledge of recording time or phase type, across a large and diverse regional network. This is a significant advancement on the work done by McBrearty [72], both in terms of providing increased generalization and extended path distances. While pairwise-similarity is certainly a weaker evidence for association

than a standard moveout curve, it does present a viable complementary validation tool, which could be used to augment existing methods of automatic association. For instance, given an event list from an automatic associator, each event can be scored based on its embedding-space distance from the cluster centroid, and dissimilar events can simply be rejected or flagged for further analyst review based on the desired type-I error rate. Future work could involve constructing a more robust framework for this task, using additional layers of machine learning.

The results for template-based source discrimination are also quite promising. The 95.8% classification accuracy achieved for explosion discrimination is impressive in its own right. However it is astounding considering that the discrimination is based on a single template waveform. This result is not only useful for identifying explosions, but also holds considerable promise for other discrimination tasks. In fact, as with most semi-supervised techniques, the real potential of this similarity-based classifier lies in its application to less well-studied and less robustly labeled classes. For instance, while the United States Geological Survey (USGS) CONUS catalog used in this work includes painstakingly labeled explosions, such labels are simply not available for many other regions. Similarly, there are numerous other source types of interest (volcanoes, ice quakes, rock bursts, tremors, ripple-fire blasts, etc.) for which labels may be scarce or unavailable. As such, our method holds considerable potential for training future discriminators on less well-studied source functions, especially when training examples are limited and fully-supervised methods are unavailable.

In conclusion, we believe that the findings in this work represent an important step forward in the field of seismogram similarity, demonstrating that such similarity measures do not need to be constrained to the path-dominant correlation-based detectors traditionally implemented. However, there is still much work to be done, especially in the application of this method across more diverse datasets, including

global networks and teleseismic signals.

3.7 Data and Resources

The raw seismograms used in this study were collected as part of Earth Scope's USArray experiment [19], and can be accessed via the Incorporated Research Institutions for Seismology (IRIS) Database using ObsPy [13].

Arrival-time catalogs for each station were downloaded through a web query of the International Seismological Centre (ISC) Bulletin for seismic arrivals:

`http://www.isc.ac.uk/iscbulletin/search/arrivals/` (last accessed February 2019).

The Neural Network Architecture was implemented in Keras [25], using the keras-tcn python package written by Philippe Rémy:

`https://github.com/philipperemy/keras-tcn` (last accessed February 2019).

The batch-hard algorithm was implemented in Tensorflow [1], and adapted from the work of Olivier Moindrot, which can be found at:

`https://omoindrot.github.io/triplet-loss` (last accessed February 2019).

A repository containing the code and trained models described in this manuscript has been made available on github, and can be found at:

`https://github.com/joshuadickey/seis-sim` (last accessed September 2019).

IV. Study 3 - BazNet: A Deep Neural Network for Confident Three-component Backazimuth Prediction

This research was submitted to the Journal Pure and Applied Geophysics on 13 November 2019, as an invited paper for the upcoming special issue entitled "Nuclear Explosion Monitoring and Verification: Scientific and Technological Advances." It is currently under review.

4.1 Abstract

As the Treaty Monitoring Community seeks to lower detection thresholds across its sparse sensor network, single-station location estimates and accurate backazimuth predictions become increasingly important. Accurate backazimuth predictions are traditionally limited to array stations, where beamforming provides high-confidence backazimuth prediction that can be reliably passed on to the associator. Three-component stations, on the other hand, rely on polarization analysis for backazimuth prediction, which suffers from both high error and low confidence. As such, very few three-component backazimuth predictions are passed on to the association algorithm. This study presents BazNet, a deep neural-network that takes in a three-component seismogram and produces both a backazimuth prediction and corresponding certainty measure. For existing stations with ample historical training data, the technique achieves an overall median absolute deviation of around 14°, a modest improvement over the 15° achieved by polarization. More importantly, each estimate is accompanied by a robust certainty measure, allowing the selection of high-confidence predictions to be passed on to the associator. Using the BazNet certainty measure, roughly 60% of all three-component predictions can be selected with a median absolute deviation of just 6°, which is on par with the predictions from a full beamformed seismic array. This represents a 7-fold improvement over the 8% of signals similarly selectable

via polarization analysis. BazNet performance is demonstrated against 10 years of waveform data from 561,154 cataloged arrivals across nine stations selected from the global IMS Network: STKA, CPUP, VNDA, LPAZ, AAK, BOSA, ULM, BATI, INK.

4.2 Introduction

Backazimuth prediction is a critical step in the seismic signal processing pipeline, feeding the downstream processes that associate events and build location estimates. Typically, there are two methods of predicting backazimuth, depending on the type of station. If the station consists of an array of instruments, the backazimuth can be predicted by examination of the time-delay of arrival across the array. This process is called beamforming, and produces backazimuth predictions that can be quite accurate. If the station consists of a single three-component (3C) instrument with North-South, East-West and Vertical components, the backazimuth is traditionally predicted by calculating the polarization of the arriving wavefront. This process produces much less accurate results. Fig. 29 demonstrates the performance advantage achieved by beamforming.

As an alternative to polarization analysis, this work presents BazNet, a deep convolutional neural network architecture that operates directly on 3C seismograms and not only produces more accurate backazimuth predictions, but also produces a robust certainty measure, allowing downstream association algorithms to only use the best estimates available. The model is trained on a per-station basis, utilizing 10 years of analyst-reviewed event locations to calculate the true backazimuths for training. The technique does not generalize across stations, and must be retrained for each station where it will be employed. However, because of the large number of available 3C stations with extensive analyst-reviewed catalogs, and because of the outstanding certainty measure produced in conjunction with each estimate, BazNet

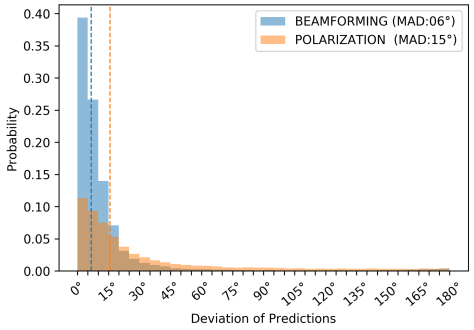

Figure 29. This diagram demonstrates the median absolute deviation for backazimuth prediction using two distinct methodologies: beamforming and polarization analysis. The blue distribution is drawn from 1,116,452 backazimuth predictions made using beamforming at four IMS array stations: CMAR, MKAR, ILAR and ASAR. The orange distribution is drawn from 561,154 predictions made using polarization analysis at nine IMS 3C stations: STKA, CPUP, VNDA, LPAZ, AAK, BOSA, ULM, BATI and INK. The median absolute deviation for both beamforming and polarization are annotated with dashes lines, 6° and 15° respectively. The y-axis of the plot has been normalized for each distribution to allow comparison. This figure clearly illustrates the significant performance advantage enjoyed by beamforming.

is able to produce backazimuth estimates for 3C stations with accuracy rivaling a beamformed array.

BazNet presents three major contributions:

- A novel NN architecture for the efficient prediction of backazimuth, directly from the raw waveforms with no feature engineering required.
- An improvement in accuracy over the traditional polarization analysis.
- A robust certainty measure coupled with each backazimuth estimate, allowing a means of selecting only the best estimates to pass on to downstream algorithms for event association and location.

4.3 Background

This background section is presented in four parts. The first section surveys backazimuth prediction; the second examines backazimuth certainty; the third provides

an overview of convolutional neural network design for time-series signals; and the final section explores the challenges of angle prediction in a machine learning context.

Backazimuth Prediction.

The backazimuth angle is defined as the great circle bearing from the recording station to the event epicenter, measured clockwise from north [15]. Fig. 30 illustrates the backazimuth for an event epicenter located in the south Pacific and a recording station in London, England.

Figure 30. This diagram demonstrates the azimuth (Az) and backazimuth (Baz) angles for the given Event-Station pair. The latitude and longitude coordinates for the station, in radians, are given by (ϕ_s, λ_s), while the event coordinates are given by (ϕ_s, λ_s).

Assuming a spherical earth, simple trigonometry can be used to calculate the backazimuth, as demonstrated in Eq. (12). This algorithm generally works well enough in practice and is useful for illustration here, although it should be noted that newer seismic processing packages utilize a more complex algorithm with better handling of the ellipsoidal shape of the earth [59].

$$
\begin{aligned}
y =& \sin(\lambda_e - \lambda_s)\cos(\phi_e) \\
x =& \sin(\phi_e)\cos(\phi_s) - \sin(\phi_s)\cos(\phi_e)\cos(\lambda_e - \lambda_s) \\
Baz =& atan2(y, x)
\end{aligned} \tag{12}
$$

Currently, the most accurate algorithm for backazimuth prediction is beamforming [83], [105], [88]. Beamforming gains its effectiveness by linearly combining signals from the multiple sensors of a seismic array; unfortunately such arrays are quite expensive, requiring multiple instruments spread out across a large geographical area measuring tens or even hundreds of kilometers. An example array layout is detailed in Fig. 31, along with a demonstration of the beamforming technique.

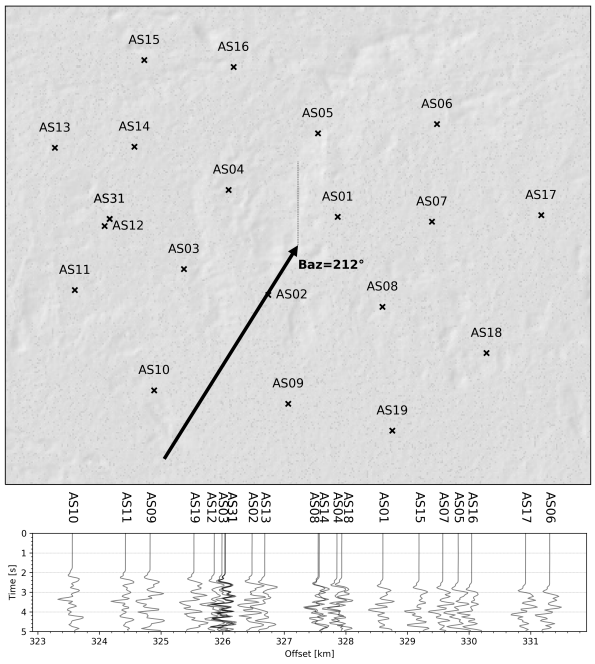

Figure 31. Top: Layout of the 20 element Alice Springs Seismic Array, ASAR, located in central Australia, with an aperture of just under 10 km. The arrow illustrates an incoming seismic wave with a backazimuth of 212°. Bottom: Seismic waveforms from the corresponding seismic event, stacked in order of distance to epicenter. Beamforming uses the geometry of the array, along with the time-delay of arriving signals, to estimate the backazimuth angle with great precision.

While beamforming is an incredibly accurate backazimuth prediction technique for

seismic arrays, the vast majority of seismic stations consist of only a single 3C sensor, making beamforming an impossibility. For these stations, the traditional method of backazimuth prediction is to analyze the polarization of the three orthogonal components of motion: North-South, East-West and Vertical. This technique is often referred to as polarization analysis, and the algorithm is based on an eigen-analysis of the filtered and windowed seismograms [57], [69]. In brief, the technique uses an eigen-decomposition of the three-component covariance matrix across a window of data to identify the principle directions of both rectilinear and elliptical polarization [39]. Several advancements of this technique have been proposed, most notably the inclusion of variable time windows, which provides a small improvement in performance [77]. Despite these advancements, the backazimuth predictions produced by polarization analysis are quite inaccurate, especially when compared to the predictions produced by beamforming [45] as shown in Fig. 29.

Recently, several attempts have been made to apply machine learning techniques to backazimuth prediction. In [78], researchers applied Support Vector Machines to estimate backazimuth for large earthquakes in Columbia. These efforts utilized feature vectors derived directly from the polarization algorithm, and showed good success. In this work, we build off of these efforts by bypassing the polarization features and learning directly from the raw waveforms using a convolutional neural network.

Backazimuth Certainty.

Backazimuth prediction is an intermediary step on the way to event association, with potentially dozens of backazimuth predictions available from various stations to feed the downstream associator. Because of the high number of 3C stations available, and because of the relatively high error rate for the 3C backazimuth predictions, it is critical that there be some statistical measure of certainty for each prediction, al-

lowing the good predictions to be utilized, while preventing the bad predictions from corrupting the associator. The simplest measure for thresholding 3C backazimuth predictions is the signal to noise ratio (SNR). In particular, an SNR threshold of 10 dB has proven useful for thresholding predictions based on polarization, as described in [45]. Unfortunately, for most stations, this discards more than 80% of the potential predictions, greatly negating the potential for using 3C stations to boost the performance of the downstream event association algorithms.

To address this, an angular measure of uncertainty was developed specifically for backazimuth predictions, called *delaz*, described in [8] and detailed in Eq. (13).

$$delaz = 2 \arcsin\left(\frac{delslo}{2slow}\right)\frac{180}{\pi} \tag{13}$$

The *delaz* measure of uncertainty, in turn, relies heavily on the uncertainty of the slowness estimate, *delslo*, which is calculated differently for array and 3C stations, as shown in Eq. (14). For array stations, *delslo* varies primarily based on $fstat$, which is a measure of the beam's spectral coherence, and f_c which is the center frequency of the processing band. For 3C stations, *delslo* varies primarily based on *rect*, which is the measured linearity of particle motion. Finally, for both array and 3C stations, *delslo* also depends on the estimated measurement error, *dk*, and the estimated modeling error, *ds*, which are both stored in site-specific lookup tables based on historical data. When certain values of *delslo* are exceeded, the *delaz* measure is given a null value, which in our dataset is set to be 180°.

$$\begin{aligned} delslo_{(AR)} &= \sqrt{ds^2 + dk^2\frac{f_c^2}{fstat}} \\ delslo_{(3C)} &= \sqrt{ds^2 + dk^2\frac{(1-rect)}{2}} \end{aligned} \tag{14}$$

For array stations, the *delaz* statistic is highly correlated with actual prediction error. Unfortunately, for 3C stations, not only are the predictions less accurate, but the *delaz* statistic is also much more loosely correlated with actual error. This is shown clearly in Fig. 32, where boxplots of backazimuth error are plotted for each decile of *delaz*. In particular, it should be noted that the 90% confidence interval for non-null beamformed predictions never extends beyond 45°, even for the least-certain decile. The boxplots for polarization certainty are much wider than those for beamforming, with 90% confidence intervals extending beyond 45° for all but the first decile. Additionally, many more null-values are assigned to the *delaz* for polarization predictions than for beamformed predictions, with null-values filling the last four deciles vs. the last decile, respectively.

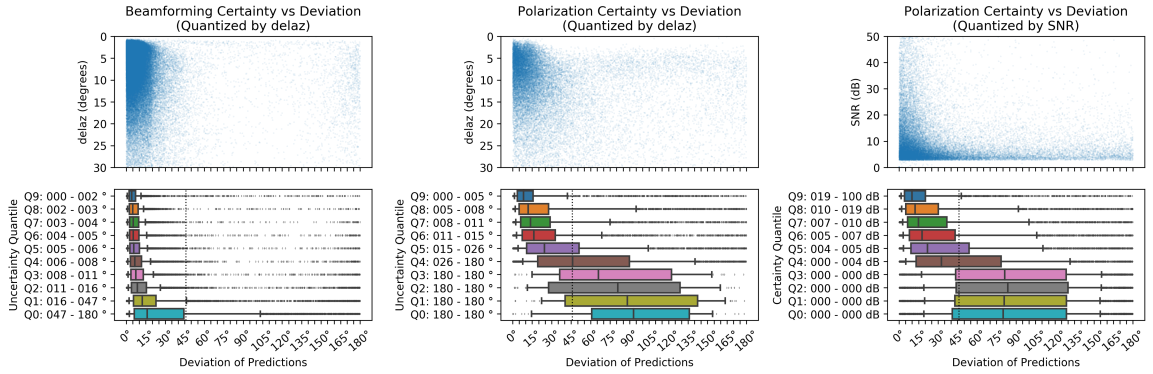

Figure 32. A demonstration of *delaz* and SNR as certainty measures for backazimuth prediction. The top row of plots are scatter-plots of certainty vs. error. The bottom row of plots are box-plots of certainty vs. error, quantized into ten evenly-sized deciles, such that Q9 shows the error distribution of predictions in the most certain decile (top 10%), while Q0 shows the error distribution of predictions in the least certain decile (bottom 10%). Predictions with invalid *delaz* are assigned a null-value of 180°. Similarly, signals where SNR is unavailable are assigned a null-value of 0 dB. This accounts for the large number of predictions assigned to either $delaz = 180°$ or $SNR = 0$dB. Finally, each boxplot is annotated with a vertical dashed line at 45°, aligned with the 90% confidence interval of the least-certain non-null decile for beamforming. Using this as a performance threshold, it can be seen that only 10% of polarization predictions meet this criteria.

Obviously, the backazimuth predictions from a 3C station will never reach the same level of performance as the predictions from a multi-instrument beamformed

array, due to the rich additional information available to the beamformed predictor. However, due to the high number of 3C stations available, great potential does exist for utilizing the backazimuth predictions from 3C stations in the downstream associator, provided that the predictions are accompanied by a reliable certainty measure. Unfortunately this criteria is currently not met by either the *delaz* statistic, or the signal to noise ratio. As a result, defining a robust certainty measure for 3C backazimuth predictions is an important and open task in seismology, particularly for the Treaty Monitoring community. This is an issue which we attempt to address directly in this work.

Convolutional Neural Networks.

Convolutional Neural Networks (CNNs) are revolutionizing the science of signal processing, from computer vision to speech recognition, and they are poised to do the same for seismic signal processing as well [85]. CNNs have already been employed in almost every branch of seismological research, from earthquake detection to earthquake early warning systems, ground-motion prediction, seismic tomography, and even earthquake geodesy [62]. Fundamentally, a CNN is composed of a set of digital filters, called kernels, which are identical in form to the digital filters commonly employed in traditional seismology, with filter coefficients that are convolved across the signal, per usual. There are two differences, however, which enhance the power of CNNs over traditional digital filters. First, the CNN filter coefficients are actually trainable parameters, which are empirically optimized against a large-scale training dataset. Second, the kernels are applied in layers, with the output of each layer undergoing an activation prior to entering the next layer. Critically, these activation functions are non-linear (such as the hyperbolic tangent function), allowing the CNN model to learn a wide range of complex non-linear processes directly from

the data [67].

In addition to using the data to learn the kernel parameters, developing a CNN also requires specifying several architectural parameters for the model, often referred to as hyper-parameters. Some typical hyper-parameters include the number of layers, the number of digital filters in each layer (kernel depth, f), and the number of coefficients in each digital filter (kernel size, k). These hyper-parameters are fixed during training, but can be varied between training runs, and optimized by comparing model performance against the validation set.

When designing a CNN architecture for time-series data, like seismograms, an important consideration is the receptive field of the model, which describes the number of input samples that can be 'seen' by each sample in the output. This is of critical importance, as it limits the feature periodicity learnable by the model. For instance, for a single-layer CNN architecture processing a 40 Hz signal, if the kernel size is 10, then the receptive field is also 10, and the model can only extract features with a periodicity of 0.25 seconds or less. In practice, designing CNNs to process long-period signals can be quite complex, requiring either many layers, or very long kernels to obtain the desired receptive field. As such, it is common to use a specialized CNN architecture known as the Temporal Convolutional Neural Network, or TCN, to process long-period signals [7]. The TCN provides a large receptive field primarily by using dilated convolution, which simply spreads out the kernel coefficients across the signal, allowing a smaller kernel to see a longer window in time. The dilated convolution equation is given in Eq. (15), where F is the time series signal, G is the kernel, and d is the dilation rate. It should be noted that this equation represents a generalized form of convolution, equivalent to standard convolution when the dilation rate is equal to 1.

$$(F *_d G)[n] = \sum_{m=-\infty}^{\infty} F[n - d \cdot m]G[m] \tag{15}$$

The TCN makes aggressive use of dilated convolution, by increasing the dilation rate at each successive layer of the network, providing a rapid expansion of the receptive field through the network. The recurrence equation for calculating the receptive field, rf_l, for a given convolutional layer, l, is given in Eq. (16), where k is the kernel size, and d_l is the dilation rate for that layer. In addition to dilated convolution, the TCN also makes use of residual connections, which simply add the output of each layer to the output of all subsequent layers, allowing the network to easily learn the identity function for any given layer, which stabilizes training [7].

$$\begin{aligned} rf_l &= rf_{l-1} + d_l(k-1) \\ &\text{where } r_0 = 0 \end{aligned} \tag{16}$$

In summary, the TCN is ideally suited for processing seismograms [28], particularly for the regional and teleseismic signals used for backazimuth angle prediction. This is exactly the research objective this work seeks to address. To this end, the next section briefly explores the general task of angle prediction in machine learning.

Angle Prediction & Circular Statistics.

Any attempt at angle prediction requires that some consideration be given to circular statistics [53]. For example, subtraction is an invalid distance measure for angles, as the linear difference between 345° and 15° is 330°, whereas the circular difference is 30°. Two valid circular distance measures are given in Eq. (17) and Eq. (18).

$$\begin{aligned} d_1(\theta_1, \theta_2) &= min\big(\theta_1 - \theta_2, 360^\circ - (\theta_1 - \theta_2)\big) \\ &= 180^\circ - \big|180^\circ - |\theta_1 - \theta_2|\big| \end{aligned} \tag{17}$$

$$d_2(\theta_1, \theta_2) = 1 - cos(\theta_1 - \theta_2) \tag{18}$$

Similarly, mean squared error is inappropriate as an angular loss function. Instead, it is common to either implement a custom circular loss function based on Eq. (17) or Eq. (18), or to instead pre-transform the angles to and from the unit circle, using Eq. (19) and Eq. (20), so that mean squared error can be used effectively.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \tag{19}$$

$$\theta = atan2(x, y) \tag{20}$$

Another solution is to discretize the angle-space into N classes, transforming the usual regression-based angle prediction into a classification task [24]. For neural networks, this is accomplished by adding a final fully-connected layer with N nodes, applying the softmax activation function, replacing the typical mean-squared error loss with categorical cross-entropy, and replacing the typical real-valued training angles with N-dimensional one-hot[1] encoded training vectors, t. This approach has a significant advantage, in that the model outputs are no longer just scalar estimates of the angle, but are instead vector estimates of the class probabilities, providing a built-in certainty measure for each prediction [40]. The softmax function and categorical cross-entropy function are given in Eq. (21) and Eq. (22), respectively, where s is

[1] A one-hot encoding is an N-dimensional binary vector representation of an integer value between zero and N-1. The vector has a single non-zero entry, located in the column corresponding to the integer value being encoded.

the N-dimensional vector output of the final fully-connected layer prior to activation, and t is the N-dimensional one-hot encoded training vector.

$$f(s)_n = \frac{e^{s_n}}{\sum_{n'}^{N} e^{s_{n'}}} \tag{21}$$

$$CE = -\sum_{n}^{N} t_n log\big(f(s)_n\big) \tag{22}$$

Unfortunately, this approach has a significant drawback, in that it introduces discontinuities to the unit circle at each class boundary. Because these discontinuities do not occur naturally in the data, examples from the dataset that happen to lie arbitrarily close to either side of a class boundary are basically indistinguishable from each other, and this artificially increases the miss-classification rate of the model near each boundary. Furthermore, the overall number of these boundary-induced miss-classifications will increase with the number of boundaries, thereby limiting both the number of classes and the angular resolution attainable by any model utilizing this standard discretization scheme [110].

To mitigate these effects, researchers in [42] adopt an M-N discretization scheme for classifying angles, using M separate classifiers, with N classes each. By keeping N small, each classifier has relatively few class boundaries, reducing the number of boundary-induced miss-classifications. However, by employing M of these classifiers, uniformly shifted around the unit circle, a high resolution, r, can be achieved, as given in Eq. (23). In effect, M-N discretization avoids the problems associated with arbitrary class boundaries, by reducing their number and then shifting them around the unit circle, such that all examples lie sufficiently far from any of the N class boundaries on the majority of the M classifiers.

$$r = \frac{360^\circ}{N \times M} \tag{23}$$

M-N discretization encodes an angle, θ, by the MxN matrix, B, with rows indexed by $m \in [0, 1, \ldots, M-1]$ and columns indexed by $n \in [0, 1, \ldots, N-1]$. Each row represents the class probabilities of a distinct N-class classifier with an initial class boundary shifted from the origin by $m \cdot r^\circ$. When encoding deterministic angles, the matrix rows are one-hot encodings, with non-zero entries corresponding to the class assignment for each row. When encoding angle predictions, the matrix contains real-valued class probabilities, such that each row sums to one. Mathematically, the encoding is defined in Eq. (24), where the class assignment for each classifier is made by taking the difference between the angle, θ, and the initial class boundary, $m \cdot r$, normalized by the class width, $360^\circ/N$, and rectified by the floor and modulo operators.

$$B_{m,n} = \begin{cases} 1, & \text{if } n = \left\lfloor \frac{\theta - m \cdot r}{360^\circ / N} \right\rfloor \mod N \\ 0, & \text{otherwise} \end{cases} \tag{24}$$

An estimate of the original angle, $\hat{\theta}$, can then be recovered by decoding the matrix. For deterministic encodings, there are exactly $M \cdot N$ possible permutations of B, each encoding an angle-space of r°. It is helpful to parameterize these permutations sequentially around the unit circle, using the indices $v \in [0, 1, \ldots, N-1]$ and $u \in [0, 1, \ldots, M-1]$, representing class and shift respectively, such that the central angle of each permutation is given by the product of the class and class width $(v \cdot 360^\circ/N)$ plus the product of the shift and shift width $(u \cdot r)$ plus half again the shift width $(r/2)$. Decoding is then accomplished according to Eq. (25), where the indicated class and shift indices for a given matrix B are found by taking the argmax of the

sum of the one-hot class probabilities corresponding to each permutation.

$$\hat{\theta} = (v \cdot 360°/N) + (u \cdot r) + r/2, \text{ where}$$
$$u, v = \underset{\substack{u\in[0,1,\ldots,M-1]\\ v\in[0,1,\ldots,N-1]}}{\operatorname{argmax}} \sum_{m=0}^{M-1} B_{m,(v-(m>u)) \mod N} \tag{25}$$

For clarity, a detailed example is explored for the case where $M = 3$ and $N = 4$. To illustrate the encoding scheme, Fig. 33 shows the class boundaries for each of the 3 classifiers, shifted around the unit circle, as well as the encoding matrix B for each 3-4 discretization. To illustrate the decoding scheme, Table 12 shows the central (predicted) angle and corresponding one-hot elements for each permutation of B, indexed by u and v.

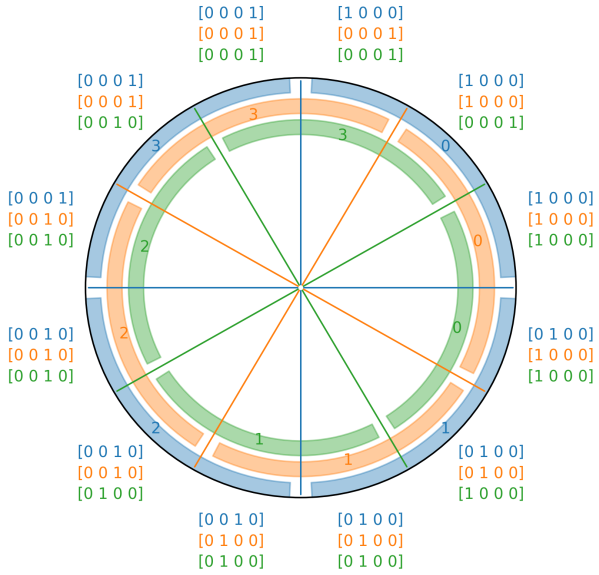

Figure 33. An example M-N discretization where $M = 3$ and $N = 4$. This discretization employs three distinct classifiers, annotated in the figure by three distinct colors: blue, orange and green. Each classifier is composed of four classes, labeled in the figure as 0, 1, 2 and 3. Because these classifiers are shifted evenly around the unit circle, this effectively creates 4*3=12 discrete regions, each encoded by a distinct one-hot permutation of the MxN matrix B.

To implement this M-N classification scheme in a neural network, a final layer

Table 12. Decoding Scheme for M-N Discretization where $M = 3$ and $N = 4$

v	u	central angle	one-hot elements
0	0	15°	$B_{(0,0)}$, $B_{(1,3)}$, $B_{(2,3)}$
0	1	45°	$B_{(0,0)}$, $B_{(1,0)}$, $B_{(2,3)}$
0	2	75°	$B_{(0,0)}$, $B_{(1,0)}$, $B_{(2,0)}$
1	0	105°	$B_{(0,1)}$, $B_{(1,0)}$, $B_{(2,0)}$
1	1	135°	$B_{(0,1)}$, $B_{(1,1)}$, $B_{(2,0)}$
1	2	165°	$B_{(0,1)}$, $B_{(1,1)}$, $B_{(2,1)}$
2	0	195°	$B_{(0,2)}$, $B_{(1,1)}$, $B_{(2,1)}$
2	1	225°	$B_{(0,2)}$, $B_{(1,2)}$, $B_{(2,1)}$
2	2	255°	$B_{(0,2)}$, $B_{(1,2)}$, $B_{(2,2)}$
3	0	285°	$B_{(0,3)}$, $B_{(1,2)}$, $B_{(2,2)}$
3	1	315°	$B_{(0,3)}$, $B_{(1,3)}$, $B_{(2,2)}$
3	2	345°	$B_{(0,3)}$, $B_{(1,3)}$, $B_{(2,3)}$

must be added to the network consisting of M fully-connected N-node outputs in parallel. Each of these outputs must be activated with a softmax function, and each must be trained with a separate catagorical cross-entropy loss function against a separate training vector, corresponding to one row of the matrix B.

4.4 Methodology

Our stated task is to build an accurate 3C backazimuth predictor, along with a robust certainty measure, allowing for the selection of high-confidence predictions to be passed on to downstream event association algorithms. This section details the dataset, neural network architecture, and evaluation metric used to accomplish this task.

Data Description.

The BazNet model takes in 3C waveforms, sampled at 40 Hz, and windowed to include 3 seconds prior and 17 seconds after the cataloged arrival time. The model is trained on a per-station basis, across ten years of cataloged waveform data from the

International Data Center (IDC), spanning from 2009 to 2018. The stations consist of nine three-component seismic stations from the International Monitoring System (IMS) Network. The nine stations include diversity in both geographic location and seismic region, and the station locations and event origins are displayed in Fig. 34. This dataset includes training, validation and testing sets according to three distinct time windows, 2009-2015, 2016-2017 and 2018, respectively. The overall catalog includes 561,154 arrivals and a detailed per-site breakdown of these arrivals can be found in Table 13, along with the true backazimuth angle distributions in Fig. 35.

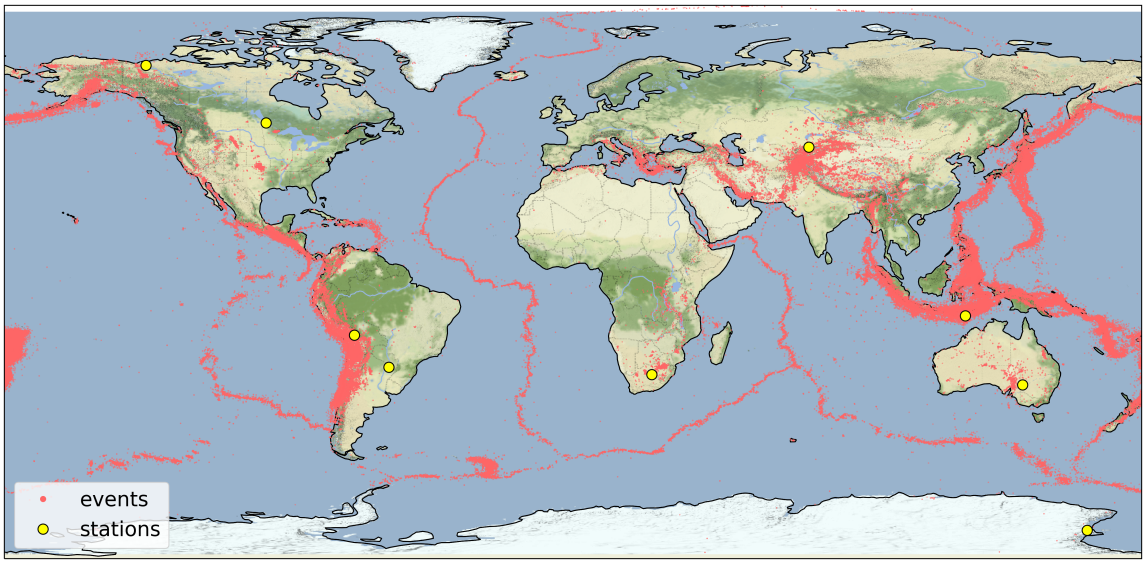

Figure 34. Map showing the geographical location of each recording station and event in the combined training, validation and testing datasets.

Model Architecture.

The BazNet architecture consists of three structural components: model input, feature extraction and model output; each structure is discussed in turn below.

The BazNet model input consists of windowed three-component seismic waveforms. Based on a survey of the time windows typically used for polarization analysis [45], [77], [39], [57], [69], as well as empirical testing against the validation set, 20

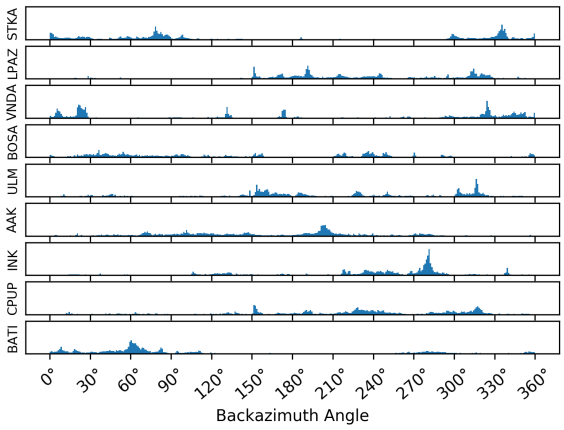

Figure 35. Histogram showing the distribution of backazimuth angles for each of the nine stations in the dataset.

Table 13. Cataloged Arrivals Across the Nine Stations.

STATION	TRAINING	VALIDATION	TESTING
STKA	94,530	24,953	12,599
LPAZ	61,247	14,904	6,444
VNDA	41,225	12,345	4,418
BOSA	34,478	9,758	5,321
ULM	36,046	8,664	5,572
AAK	31,830	11,113	5,229
INK	35,751	10,141	4,274
CPUP	35,548	10,423	5,414
BATI	26,663	9,237	3,027
TOTAL	397,318	111,538	52,298

second windows were selected. For the 40 Hz data used, this results in input windows that are 800 samples long and three samples deep.

BazNet feature extraction utilizes convolutional layers and ReLU[2] activations, which is standard for processing structured data like waveforms or images. In particular the BazNet architecture is designed to have a receptive field matching the full window length of at least 800 samples. This is accomplished by employing the TCN architecture described in [7], with a a dilation scheme of $d = [2, 4, 8, 16, 32]$ and a kernel size of $k = 15$. This results in a receptive field of 869 samples, as calculated by Eq. (16) and detailed in Table 14. The filter depth, or number of filters in each layer, was varied from as low as 4 to as high as 100, and an optimal value of $f = 45$ was selected. Following the standard TCN architecture, causal padding is used during convolution, meaning each sample in the output time-series depends only on prior samples from the input time-series. This allows the output time-series to be truncated just after the final convolutional layer, discarding all but the last sample from each of the 45 filters, as described in [7].

Table 14. TCN Layer Parameters

l	d	k	Receptive Field	Receptive Field
1	2	15	29 samples	0.7 seconds
2	4	15	85 samples	2.1 seconds
3	8	15	197 samples	4.9 seconds
4	16	15	421 samples	10.5 seconds
5	32	15	869 samples	21.7 seconds

The BazNet model output is formulated as a set of class probabilities, in order to obtain the built-in certainty measure described in [40]. Initial efforts focused on standard discretization to classify the backazimuth angles, however the results were largely unsuccessful, due to the limitations described in [110], and the model was unable to

[2]The rectified linear unit (ReLu) is a standard activation function used with convolutional neural networks. Also know as the ramp function, it simply zeros out all negative values while passing positive values unchanged. ReLu activations are commonly employed in CNN design due to the computational advantages of differentiating this function, as the derivative is always one or zero.

achieve accuracy on par with polarization, much less beamforming. To remedy this, the M-N discretization scheme described in [42] was employed. In particular, a grid search was conducted over 28 M-N tuples, with the number of classes and classifiers ranging over $N \in [3, 4, 5, 6]$ and $M \in [4, 6, 8, 10, 12, 14, 16]$, respectively. Comparing model performance for each discretization against the validation set, best empirical values of $M = 8$ and $N = 4$ were selected, as shown in the model performance vs. discretization heatmap in Fig. 36. This 8-4 discretization provides a resolution of 11.3°, a maximum resolution error of 5.6°, and for uniformly distributed angles, an expected mean resolution error of just 2.8°, which is well within the median absolute deviation for beamforming of 6°, which is our goal. To implement this 8-4 classifier in BazNet, a final fully-connected layer was added to the network consisting of eight parallel four-node outputs, each activated by the softmax function. Training was then accomplished using the categorical cross-entropy loss function. A diagram detailing the final model architecture can be seen in Fig. 37.

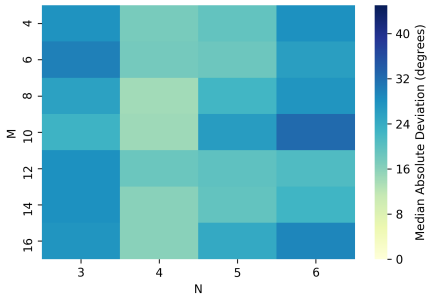

Figure 36. Heatmap showing model performance for various M-N discretization schemes. Model performance is reported by the median absolute deviation of back-azimuths against the validation set, and varies from an optimal performance of 14° for the 8-4 discretization to a dismal performance of 33° for the 10-6 discretization. This large spread in performance illustrates the importance of selecting an optimal discretization scheme.

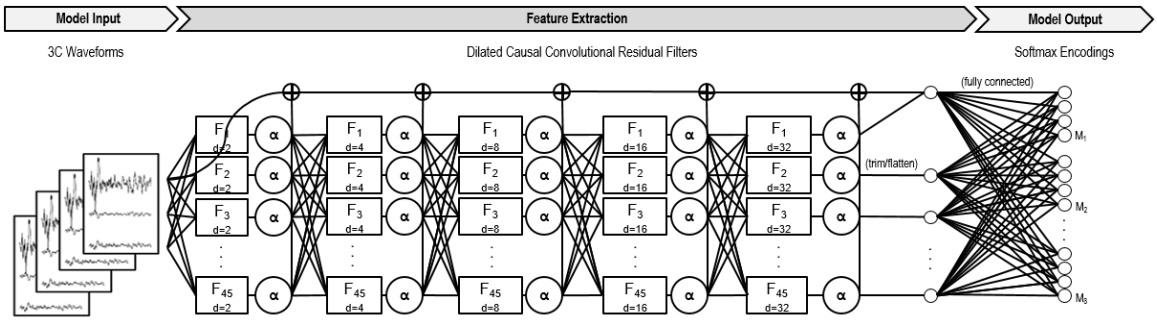

Figure 37. A detailed representation of the BazNet architecture. The model accepts as inputs three-component seismic waveforms. Features are extracted from these waveforms via five dilated convolutional layers, each with 45 filters, progressive dilation rates of 2, 4, 8, 16 and 32, and ReLU activations, α. The final layer consists of eight fully connected dense layers, each with four nodes. All eight dense layers are connected in parallel with softmax activations, producing eight classification outputs, M_1 to M_8, with four classes each.

Evaluation Criteria.

The objective of the BazNet model is to produce accurate backazimuth predictions, along with a robust certainty measure, allowing the retention of a subset of estimates with an error distribution on par with beamforming. To this end, the beamforming error distribution is benchmarked by two statistics: the median absolute deviation and the 90% confidence interval, which capture the central tendency and spread, respectively. To evaluate the BazNet model, it's predictions are thresholded by certainty in order to achieve these statistical distribution benchmarks, and performance is reported as the percentage of predictions retained.

In particular, as shown in Fig. 29, beamforming typically achieves a median absolute deviation of 6°. Enforcing this benchmark, all predictions will be thresholded to achieve a median absolute deviation of no more than 6°, and model performance will be evaluated based on the number of predictions retained. Applying this analysis to the polarization predictions across all nine stations in the test set, using the *delaz* certainty measure to threshold the predictions, a retention-rate of just 8% is achieved. As such, any retention-rate for the BazNet model above the polarization baseline of

8% will be considered a success.

Likewise, as shown in Fig. 32, beamforming typically achieves a 90% confidence interval of less than 45° for all non-null predictions. Enforcing this benchmark, all predictions will be thresholded to achieve a 90% confidence interval of no more than 45°, and model performance will be evaluated based on the number of predictions retained. Applying this analysis to the polarization predictions across all nine stations in the test set, and using the *delaz* certainty measure to threshold the predictions, results in a retention-rate of just 10%. As such, any retention-rate for the BazNet model above the polarization baseline of 10% will be considered a success.

4.5 Results and Discussion

Training and computation time.

Training for each model was accomplished using a single Titan X GPU hardware platform. Early stopping was employed based on validation loss, and the patience was set to seven epochs. The batch size was set at 50 examples, which was approximately the maximum size permitted due to the 12 GB RAM capacity of the Titan X GPUs. Training times varied somewhat due to the stochastic nature of the learning and the variations in length of the training datasets, but the average training time was 128 minutes per station. Finally, computation time is quite fast, taking less than 30 seconds total to process the 53,298 arrivals in the combined test set across the nine stations.

Performance Comparison.

The overall prediction accuracy for BazNet is not much better than that of a finely-tuned polarization algorithm, with the two predictors reporting an overall median absolute deviation of 14° and 15°, respectively, across the combined nine station test

set. On the other hand, the softmax certainty measure provided along with the BazNet predictions is quite successful, correlating strongly with prediction accuracy, as shown in Fig. 38.

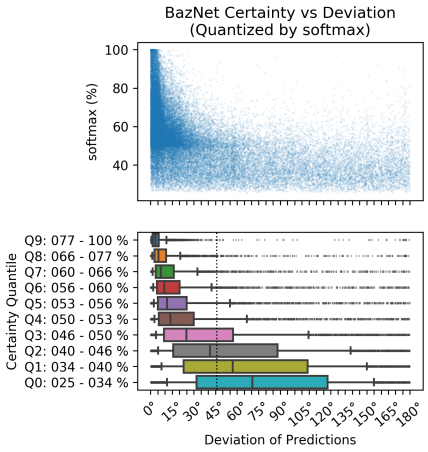

Figure 38. A demonstration of the softmax certainty measure proposed in conjunction with the BazNet backazimuth predictions. The top plot is a scatter-plot of certainty vs. error. The bottom plot is a box-plot of certainty vs. error, quantized into ten evenly-sized deciles, such that Q9 shows the error distribution of predictions in the most certain decile (top 10%), while Q0 shows the error distribution of predictions in the least certain decile (bottom 10%). The prediction errors are tightly aligned with the certainty error, as shown by the quatized boxplots, which widen smoothly as a function of decreasing certainty across all 10 deciles. Finally, the boxplot is annotated with a vertical dashed line at 45°, corresponding to the 90% confidence interval for the least-certain non-null decile for beamforming. Using this as a performance threshold, it can be seen that 40% of polarization predictions meet this criteria.

Comparing Fig. 38 to Fig. 32 allows us to evaluate the BazNet model by the spread of the distribution. For BazNet, there are four deciles (40% of predictions) with 90% confidence intervals extending to less than 45°, equating to a four-fold improvement over the 10% of polarization predictions meeting this same criteria. To evaluate the BazNet model by central tendency, the BazNet predictions are thresholded to achieve a median absolute deviation of 6°, which results in a retained-prediction rate of 59%, as shown in Fig. 39. This is a seven-fold improvement over the baseline of 8% achieved by polarization analysis.

The results are not uniformly distributed across the nine stations, as can be seen

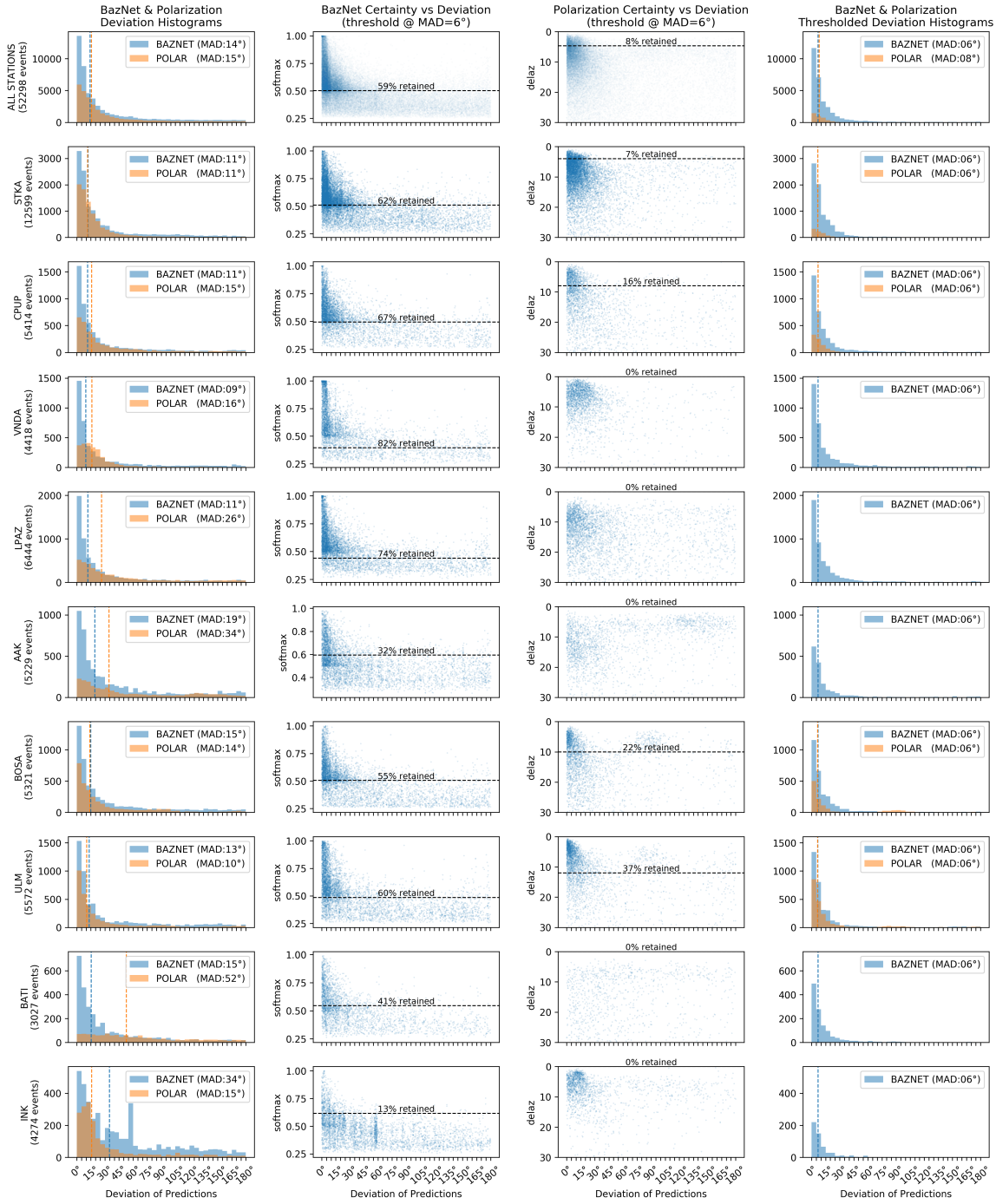

Figure 39. BazNet and Polarization Performance Comparison by Station. The left column displays the overall distribution of errors for each predictor, with the median absolute deviation of each distribution annotated by a vertical dashed line. The left-middle column displays a scatter-plot of error vs. certainty for BazNet, and the right-middle column displays error vs. certainty for polarization. For each scatter-plot, a certainty threshold is selected, such that the retained distribution has a median absolute deviation of $6°$. This threshold is annotated by a horizontal dashed line, along with the percentage of retained predictions. The far right column displays the distribution of errors for the retained predictions. The top row displays the aggregate statistics across all nine stations, while the single-station statistics are broken out in the rows below.

in Fig. 39. For instance, the performance of the predictions for INK are quite poor, and deserve further investigation. Of particular interest is the high percentage of predictions with a deviation of around 60°, seemingly uncorrelated with certainty. Investigating these, it is found that the predictor has converged to a local minimum, with the network always predicting angles of around 287°. To better understand this convergence, the true backazimuth distributions are plotted for each station in Fig. 35. Here, the reason for the convergence to 287° is apparent, as this angle coincides with a large spike in the true backazimuth distributions at INK. In fact, 287° is the central angle encoding the discretization from 281° to 293°, which corresponds to nearly half of the cataloged events. This also explains the concentration of errors around 60°, as the second largest spike in true backazimuths for INK occurs near 343°, resulting in the corresponding spike in errors of 343° - 287° = 56°. This extended analysis is included here to illustrate one of the many dangers of machine learning in general, and of softmax certainty in particular: namely, models that are empirically driven, must be expertly investigated. Perhaps one day, ML models will be able to recognize the signs of convergence to a local minimum, however in the present day, this is left to expert analysis. In this present case, for the backazimuth model at INK, the problem was easily identified by human review, because of the peculiar concentration of errors around 60°. However, the potential exists for this to be a more subtle problem, and this illustrates the value in using a separate validation set and conducting a thorough expert evaluation of the error residuals as shown in this work.

4.6 Conclusion

BazNet shows some promise for outperforming traditional polarization analysis as a viable backazimuth predictor for 3C stations. It does suffer from several limitations, most notably that it requires a large historical catalog of training data and must be

trained separately for each station. As such, it simply cannot be applied to new stations, where polarization is still the best option. However, the certainty measure provided with each prediction is highly correlated with the actual error, and this can be incredibly valuable. As demonstrated, this gives BazNet the potential to provide 7 times the number of backazimuth predictions to downstream algorithms over polarization. More importantly, notice the relationship between BazNet's softmax certainty measure and the confidence intervals, as show in Fig. 38. The relationship is linear for the first 6 deciles. This allows BazNet to report real confidence intervals along with these predictions. This could have a significant impact on event association for global monitoring networks like the IMS.

There is much future work left to be done. In particular, due to the lack of generalization across stations, it is clear that BazNet is not making use of the traditional polarization-type features used by other methods. As such, it would be interesting to combine the results of polarization and BazNet with some type of ensemble predictor. Similarly, it could be possible to build such features directly into BazNet, perhaps by feeding them into the final fully connected layer of the network.

Finally, it would also be interesting to try to learn a new polarization-type feature-space directly. This could be accomplished by employing a semi-supervised learning approach to create polarization-specific embeddings, learning a similarity metric as in [29], but altering the embedding objective to focus on like-angled arrivals. This metric could be trained in general, utilizing all available stations, and then the resulting embedding-space could be used in specific, as a basis for training each station-specific back-azimuth predictor. This technique is akin to transfer learning [91], and could both reduce the required training data and increase the overall performance of the predictor.

Data and Resources

The raw waveforms and analyst-reviewed catalogs described in this manuscript were provided by the Comprehensive Test Ban Treaty Organization through the US National Data Center. This data was made available exclusively to the authors, as employees of the United States Air Force.

The Neural Network Architecture was implemented in Keras [25], using the keras-tcn python package written by Philippe Rémy: `https://github.com/philipperemy/keras-tcn` (last accessed February 2019).

A repository containing the code and trained models described in this manuscript has been made available on github, and can be found at: `https://github.com/joshuadickey/baz-net` (last accessed November 2019).

V. Conclusions and Future Work

Seismic signal processing at the IDC is critical to global security, facilitating the detection and identification of covert nuclear tests in near-real time. This dissertation details three research studies providing substantial enhancements to this pipeline. Study 1 focuses on signal detection, employing a TCN architecture directly against the raw real-time data streams and effecting a 4 dB increase in detector sensitivity over the latest operational methods. Study 2 focuses on both event association and source discrimination, utilizing a TCN-based triplet network to extract source-specific features from three-component seismograms, and providing both a complementary validation measure for event association and a one-shot classifier for template-based source discrimination. Finally, Study 3 focuses on event localization, and employs a TCN architecture against three-component seismograms in order to confidently predict backazimuth angle and provide a seven-fold increase in usable picks over traditional polarization analysis.

5.1 Study 1 - Signal Detection

Study 1 tackles the joint task of signal detection and arrival time estimation, using a deep neural network architecture called DeepPick. The power of DeepPick lies in the training data, which utilizes the arrival catalogs for several regional arrays as labels while using trace waveforms from a single vertical component at the array center. By taking advantage of this training data, temporal convolutions and a unique exponential sequence tagging function, the DeepPick algorithm forms a powerful tool for weak signal teleseismic detection. The DeepPick algorithm is able to accurately detect twice the number of events detected by the STA/LTA algorithm commonly used, and does it significantly faster [Section 2.6].

The findings in this work represent an important step forward in the field of teleseismic detection, demonstrating that accurate teleseismic event detection is possible from a single seismic instrument. The DeepPick algorithm has the potential to open up thousands of additional automatic detections to single-instrument seismic stations each year, without the need for additional sensors and equipment.

There is still potential for much improvement. In this work, develops a single-trace detector, applied only to a single channel of data from a three channel instrument; future work could extend our results to include data from all three channels of the instrument. Furthermore, an application of the same technique to an entire array of channels could also prove interesting, and the potential exists to improve the results significantly by incorporating more channels of data. Additionally, the focus of this work has been primarily centered on producing a detector with increased sensitivity and recall, whereas future work could focus on using similar techniques to produce a detector with an even lower false positive rate.

5.2 Study 2 - Event Association

Study 2 tackles the task of pairwise event association from raw data, utilizing a deep seismic similarity measure. To date, almost all seismogram similarity measures have been based on the cross-correlation function, constraining them to relatively path-dominant similarity, and limiting their use to repetitive and geographically localized signals. In contrast, this study presents a path-invariant measure for seismogram similarity, based on a deep triplet network architecture. The utility of this similarity measure is demonstrated for both pairwise event association and template-based source discrimination.

For the pairwise association task, the similarity measure is able to achieve an accuracy of 80%, without any knowledge of recording time or phase type, across a

large and diverse regional network [Section 3.5]. This is a significant advancement on the work done by McBrearty [72], both in terms of providing increased generalization and extended path distances. And while pairwise-similarity is certainly a weaker evidence for association than a standard moveout curve, it does present a viable complementary validation tool, which could be used to augment existing methods of automatic association. For instance, given an event list from an automatic associator, each event can be scored based on its embedding-space distance from the cluster centroid, and dissimilar events can simply be rejected or flagged for further analyst review based on the desired type-I error rate. Future work could involve constructing a more robust framework for this task, using additional layers of machine learning.

The results for template-based source discrimination are also quite promising. The 95.8% classification accuracy achieved for explosion discrimination is impressive in its own right [Section 3.5]. However it is astounding considering that the discrimination is based on a single template waveform. This result is not only useful for identifying explosions, but also holds considerable promise for other discrimination tasks. In fact, as with most semi-supervised techniques, the real potential of our similarity-based classifier lies in its application to less well-studied and less robustly labeled classes. For instance, while the USGS CONUS catalog used in this work includes painstakingly labeled explosions, such labels are simply not available for many other regions. Similarly, there are numerous other source types of interest (volcanoes, ice quakes, rock bursts, tremors, ripple-fire blasts, etc.) for which labels may be scarce or unavailable. As such, our method holds considerable potential for training future discriminators on less well-studied source functions, especially when training examples are limited and fully-supervised methods are unavailable.

The findings in this work represent an important step forward in the field of seismogram similarity, demonstrating that such similarity measures do not need to

be constrained to the path-dominant correlation-based detectors traditionally implemented. However, there is still much work to be done, especially in the application of this method across more diverse datasets, including global networks and teleseismic signals.

5.3 Study 3 - Backazimuth Prediction

Study 3 tackles the initial task in the process of single-station event location, backazimuth prediction, using a deep neural network architecture called BazNet. BazNet shows promise for overtaking traditional polarization analysis as a viable backazimuth predictor for 3C stations. It does suffer from several limitations, most notably that it requires a large historical catalog of training data and must be trained separately for each station. As such, it simply cannot be applied to new stations, where polarization is still the best option. However, the certainty measure provided with each prediction is highly correlated with the actual error, and this can be incredibly valuable. In fact, it gives a BazNet predictor the potential to provide nearly 3 times the number of backazimuth predictions to downstream algorithms over polarization [Section 4.5]. This could have a significant impact on event association for global monitoring networks like the IMS.

There is much future work left to be done. In particular, due to the lack of generalization across stations, it is clear that BazNet is not making use of the traditional polarization-type features used by other methods. As such, it would be interesting to combine the results of polarization and BazNet with some type of ensemble predictor. Similarly, it could be possible to build such features directly into BazNet, perhaps by feeding them into the final fully connected layer of the network.

Finally, it would also be interesting to try to learn a new polarization-type feature-space directly. This could be accomplished by employing a semi-supervised learning

approach to create polarization-specific embeddings, learning a similarity metric as in [29], but altering the embedding objective to focus on like-angled arrivals. This metric could be trained in general, utilizing all available stations, and then the resulting embedding-space could be used in specific, as a basis for training each station-specific back-azimuth predictor. This technique is akin to transfer learning [91], and could both reduce the required training data and increase the overall performance of the predictor.

Appendix A. DeepPick Comparative Algorithm Settings

In this work, we have utilized the FBPicker algorithm from the PhasePApy python package written by Chen Chen and Austin Holland of the Oklahoma Geological Survey. The FBPicker algorithm is designed to be robust to parameter selection, and the majority of the parameters were left at their default values, however, some tuning was performed. Specifically, t_long and t_ma were set to 5 and 30 respectively, based on established windows for teleseismic signals, and n_sigma was selected empirically to give a type-I error rate of approximately 0.001. Our final parameter selections for the FBPicker are listed in Table 15.

Table 15. FBPicker Parameter Values Used in this Work.

Parameter	Val	Description
`t_long`	5	the time in seconds of moving window to calculate CFn of each bandpass filtered data
`freq_min`	1	the center frequency of first octave filtering band
`cnr`	1	corner order of bandpass filtering
`t_ma`	30	the time in seconds of the moving average window for dynamic threshold
`n_sigma`	6	controls the level of threshold to trigger potential picks
`t_up`	2	the time in seconds not allowed consecutive pick in this duration
`mode`	5	two options: standard deviation (std) or root mean square (rms)
`nr_len`	2	noise ratio filter window length before and after potential picks used to calculate standard deviation
`nr_coeff`	0.05	control threshold level to determine if remove the pick by comparing std or rms on both sides of each potential pick
`pol_len`	10	window length in samples to calculate the standard deviation of waveform before the picks
`pol_coeff`	10	determine if declare first motion as 'Compression' or 'Dilation' by comparing the first local extreme value after pick and standard deviation in previous window
`uncert_len`	30	window length in time to calculate the rms of the CF before the picks, we make it as long as t_ma
`uncert_coeff`	3	control the floating level based on the noise of CF

We have also utilized the KTPicker algorithm from the PhasePApy python package, with the final parameter selections listed in Table 16. The majority of the parameters were left at their default values, however t_win and t_ma were set to 5 and 30 respectively, and n_sigma was selected empirically to give a type-I error rate of approximately 0.001.

Table 16. KTPicker Parameter Values Used in this Work.

Parameter	Val	Description
`t_win`	5	the time in seconds of moving window to calculate kurtosis
`t_ma`	30	the time in seconds of the moving average window for dynamic threshold
`n_sigma`	7	controls the level of threshold to trigger potential picks
`t_up`	2	the time in seconds not allowed consecutive pick in this duration
`nr_len`	2	noise ratio filter window length before and after potential picks used to calculate standard deviation
`nr_coeff`	.05	control threshold level to determine if remove the pick by comparing std or rms on both sides of each potential pick
`pol_len`	10	window length in samples to calculate the standard deviation of waveform before the picks
`pol_coeff`	10	determine if declare first motion as 'Compression' or 'Dilation' by comparing the first local extreme value after pick and standard deviation in previous window
`uncert_len`	30	window length in time to calculate the rms of the CF before the picks, we make it as long as t_ma
`uncert_coeff`	3	control the floating level based on the noise of CF

Appendix B. DeepPick Waveform Examples

In order to more fully represent the capabilities of DeepPick, we now proceed to detail its performance directly against several waveform examples. Specifically, we examine 64 total waveforms from two of our test set arrays, BURAR and ASAR. These two arrays were chosen to represent both the best and worst performing models generated in our work, with recall rates of 80% and 49% respectively. Each waveform is centered around a cataloged arrival time, and labeled with its ISC eventid, phase, magnitude estimate, depth and distance in degrees. Next to each waveform we also present the characteristic functions for each of the three algorithms tested, DeepPick, FBPicker and KTPicker. Finally, each characteristic function is annotated with any predicted arrivals to allows a direct comparison of algorithm performance. We hope that the inclusion of this waveform Appendix will help the reader to better understand the potential limitations of the DeepPick algorithm, as well as its considerable ability to detect very faint signals from a single trace.

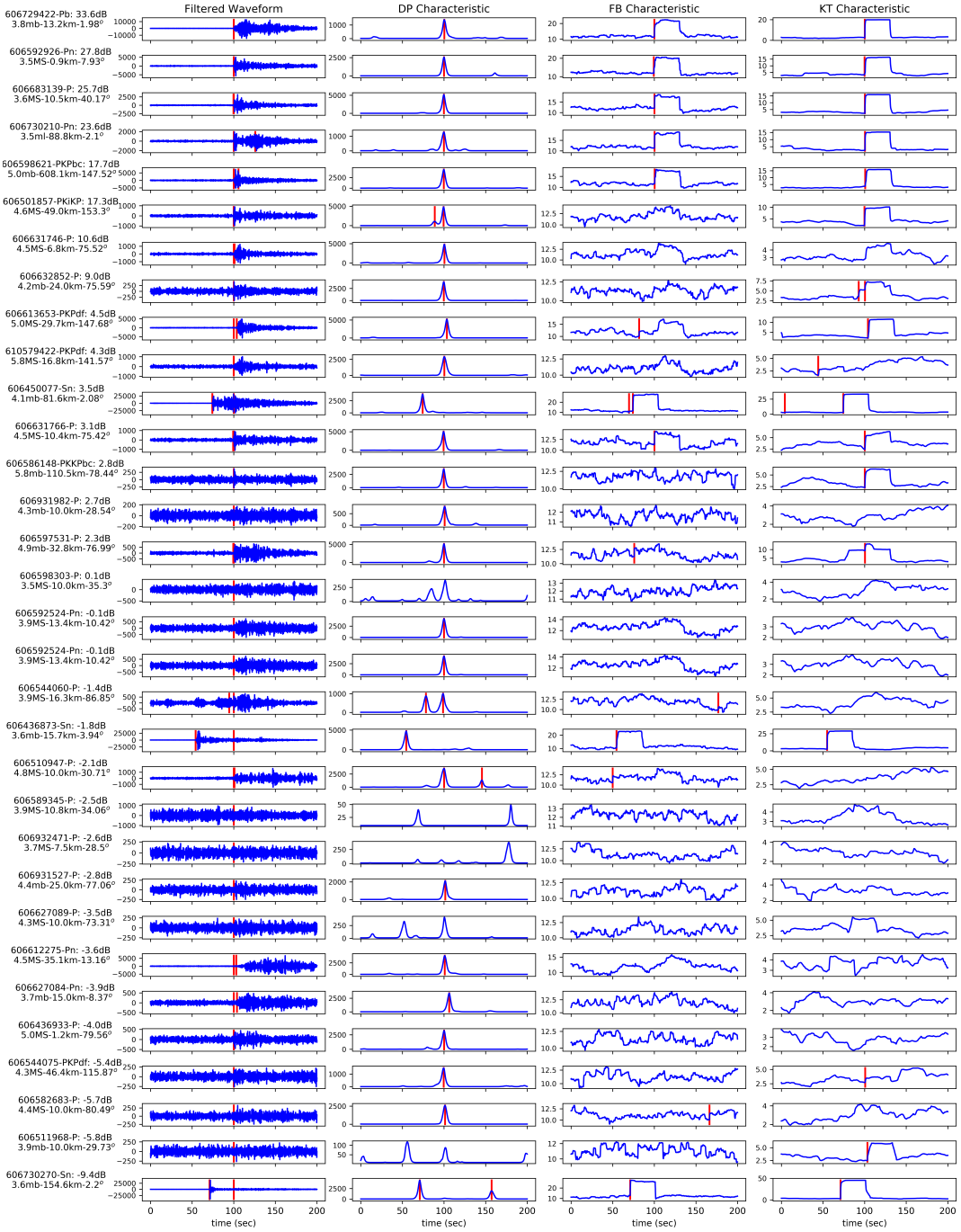

Figure 40. BURAR Waveform Analysis: Presented here are 32 randomly-selected events from the BURAR Array test set. For each event, the time-series waveform is shown at left (bandpass filtered between 1 and 4 Hz), annotated with the array-beam cataloged arrivals in red. The next three columns demonstrate the characteristic function for DeepPick, FBPicker and KTPicker respectively, annotated with any predicted arrivals in red. The signals are sorted in descending SNR levels to demonstrate increasingly difficult detection problems.

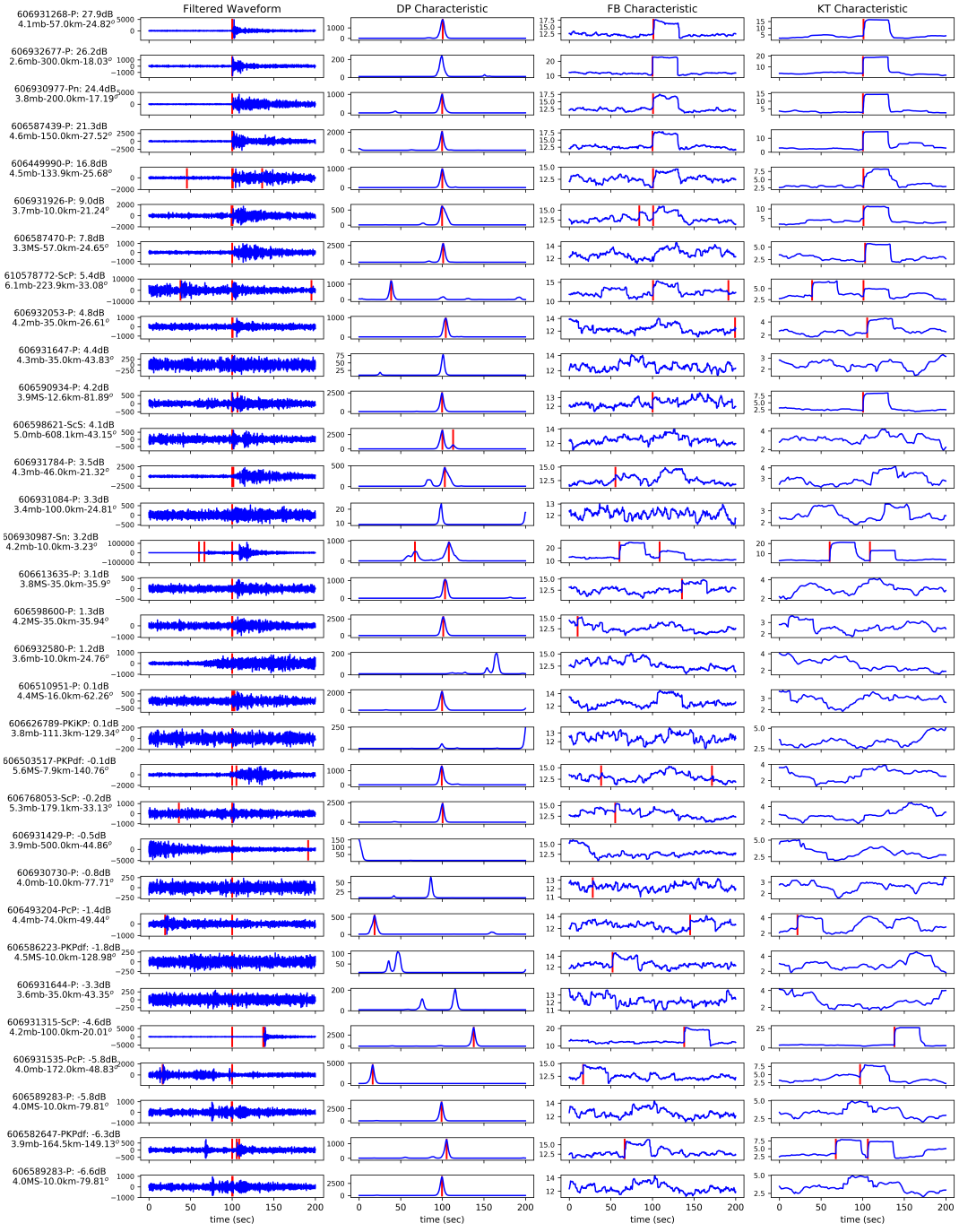

Figure 41. ASAR Waveform Analysis: Presented here are 32 randomly-selected events from the ASAR Array test set. For each event, the time-series waveform is shown at left (bandpass filtered between 1 and 4 Hz), annotated with the array-beam cataloged arrivals in red. The next three columns demonstrate the characteristic function for DeepPick, FBPicker and KTPicker respectively, annotated with any predicted arrivals in red. The signals are sorted in descending SNR levels to demonstrate increasingly difficult detection problems.

Bibliography

1. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015. [Online]. Available: https://www.tensorflow.org/

2. R. V. Allen, "Automatic earthquake recognition and timing from single traces," *Bulletin of the Seismological Society of America*, vol. 68, no. 5, pp. 1521–1532, 10 1978.

3. K. S. Anant and F. U. Dowla, "Wavelet transform methods for phase identification in three-component seismograms," *Bulletin of the Seismological Society of America*, vol. 87, no. 6, pp. 1598–1612, 12 1997.

4. D. N. Anderson, D. K. Fagan, M. A. Tinker, G. D. Kraft, and K. D. Hutchenson, "A Mathematical Statistics Formulation of the Teleseismic Explosion Identification Problem with Multiple Discriminants," *Bulletin of the Seismological Society of America*, vol. 97, no. 5, pp. 1730–1741, 10 2007. [Online]. Available: http://dx.doi.org/10.1785/0120060052

5. P. D. Anderson, "Machine learning approach to identification of seismic events," Ph.D. dissertation, Air Force Institute of Technology, 2018.

6. M. Baer and U. Kradolfer, "An automatic phase picker for local and teleseismic events," *Bulletin of the Seismological Society of America*, vol. 77, no. 4, pp. 1437–1445, 8 1987.

7. S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv e-prints*, vol. abs/1803.0, 3 2018. [Online]. Available: http://arxiv.org/abs/1803.01271

8. G. Beall, D. Brown, J. Carter, M. Fisk, J. Hanson, H. Israelsson, R. Jenkins, D. Jespen, C. Katz, R. LeBras, W. Nagy, D. Wahl, J. Wang, and X. Yang, *IDC Processing of Seismic, Hydroacoustic, and Infrasonic Data*. Vienna, Austria: Science Applications International Corporation (SAIC), 1999.

9. E. Beaucé, W. B. Frank, and A. Romanenko, "Fast Matched Filter (FMF): An Efficient Seismic Matched-Filter Search for Both CPU and GPU Architectures," *Seismological Research Letters*, vol. 89, no. 1, pp. 165–172, 12 2017. [Online]. Available: https://dx.doi.org/10.1785/0220170181

10. M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 6 2003. [Online]. Available: https://doi.org/10.1162/089976603321780317

11. K. J. Bergen and G. C. Beroza, "Earthquake Fingerprints: Extracting Waveform Features for Similarity-Based Earthquake Detection," *Pure and Applied Geophysics*, 2018. [Online]. Available: http://link.springer.com/10.1007/s00024-018-1995-6

12. B. J. Bernstein, "Eclipsed by Hiroshima and Nagasaki: Early Thinking about Tactical Nuclear Weapons," *International Security*, vol. 15, no. 4, pp. 149–173, 1991. [Online]. Available: http://www.jstor.org/stable/2539014

13. M. Beyreuther, R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann, "ObsPy: A Python Toolbox for Seismology," *Seismological Research Letters*, vol. 81, no. 3, pp. 530–533, 5 2010. [Online]. Available: http://dx.doi.org/10.1785/gssrl.81.3.530

14. D. Bobrov, I. Kitov, and L. Zerbo, "Perspectives of Cross-Correlation in Seismic Monitoring at the International Data Centre," *Pure and Applied Geophysics*, vol. 171, no. 3-5, pp. 439–468, 2014. [Online]. Available: https://doi.org/10.1007/s00024-012-0626-x

15. G. Bomford, *Geodesy*. New York, NY: Oxford University Press, 1980.

16. P. Bormann and IASPEI, *New Manual of Seismological Observatory Practice (NMSOP-2)*, 2nd ed. Potsdam, DE: GFZ German Research Centre for Geosciences, 1 2012, vol. 2 Volumes.

17. S. R. Bratt and T. C. Bache, "Locating events with a sparse network of regional arrays," *Bulletin of the Seismological Society of America*, vol. 78, no. 2, pp. 780–798, 4 1988.

18. C. J. C. Burges, J. C. Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 165–174, 2003.

19. R. Busby, R. Woodward, K. Hafner, F. Vernon, and A. Frassetto, "The Design and Implementation of EarthScope's USArray Transportable Array in the Conterminous United States and Southern Canada," Earth Scope, Tech. Rep., 2018.

20. J. C. Pechmann and H. Kanamori, "Waveforms and spectra of preshocks and aftershocks of the 1979 Imperial Valley, California, Earthquake: evidence for fault hetergeneity," *Journal of Geophysical Research*, vol. 871, pp. 10 579–10 598, 12 1982.

21. J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

22. C. Chen and A. A. Holland, "PhasePApy: A Robust Pure Python Package for Automatic Identification of Seismic Phases," *Seismological Research Letters*, vol. 87, no. 6, pp. 1384–1396, 8 2016. [Online]. Available: http://dx.doi.org/10.1785/0220160019

23. Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based Classification: Concepts and Algorithms," *J. Mach. Learn. Res.*, vol. 10, pp. 747–776, 6 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1577069.1577096

24. Y. Chen, W. Gong, C. Chen, and W. Li, "Learning Orientation-Estimation Convolutional Neural Network for Building Detection in Optical Remote Sensing Image," *2018 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2018*, pp. 1–12, 2019.

25. F. Chollet and others, "Keras," 2015. [Online]. Available: https://keras.io

26. S. Chopra, R. Hadsell, and Y. LeCun, "Learning a Similarity Metric Discriminatively, with Application to Face Verification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 539–546, 2005.

27. CTBTO, "CTBT: Ending Nuclear Explosions," CTBTO, Vienna, Austria, Tech. Rep., 2017. [Online]. Available: https://www.ctbto.org/fileadmin/user_upload/public_information/2019/CTBT_FactSheet_English_Feb_2019.pdf

28. J. Dickey, B. Borghetti, and W. Junek, "Improving Regional and Teleseismic Detection for Single-Trace Waveforms Using a Deep Temporal Convolutional Neural Network Trained with an Array-Beam Catalog," *Sensors*, vol. 19, no. 3, 2019. [Online]. Available: https://doi.org/10.3390/s19030597

29. J. Dickey, B. Borghetti, W. Junek, and R. Martin, "Beyond Correlation: A Path-Invariant Measure for Seismogram Similarity," *Seismological Research Letters*, vol. 91, no. 1, pp. 356–369, 2019. [Online]. Available: http://arxiv.org/abs/1904.07936

30. S. J. Diehl and J. C. Moltz, *Nuclear Weapons and Nonproliferation: A Reference Handbook*, ser. Contemporary world issues. ABC-CLIO, 2008.

31. D. A. Dodge and W. R. Walter, "Initial Global Seismic Cross-Correlation Results: Implications for Empirical Signal Detectors," *Bulletin of the Seismological Society of America*, vol. 105, no. 1, pp. 240–256, 1 2015. [Online]. Available: http://dx.doi.org/10.1785/0120140166

32. P. S. Dysart and J. J. Pulli, "Spectral study of regional earthquakes and chemical explosions recorded at the NORESS array," Center for Seismic Studies, Tech. Rep., 1987.

33. A. Frankel, "Precursors to a magnitude 4.8 earthquake in the Virgin Islands: Spatial clustering of small earthquakes, anomalous focal mechanisms, and earthquake doublets," *Bulletin of the Seismological Society of America*, vol. 72, no. 4, pp. 1277–1294, 8 1982.

34. W. F. Freiberger, "An Approximate Method in Singal Detection," *Quarterly of Applied Mathematics*, vol. 20, no. 4, pp. 373–378, 1963. [Online]. Available: http://www.jstor.org/stable/43636443

35. G. B. Giannakis and M. K. Tsatsanis, "Time-domain tests for Gaussianity and time-reversibility," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3460–3472, 1994.

36. ——, "Signal detection and classification using matched filtering and higher order statistics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 7, pp. 1284–1296, 1990.

37. S. J. Gibbons and F. Ringdal, "The detection of low magnitude seismic events using array-based waveform correlation," *Geophysical Journal International*, vol. 165, no. 1, pp. 149–166, 4 2006. [Online]. Available: https://dx.doi.org/10.1111/j.1365-246X.2006.02865.x

38. T. Goforth and E. Herrin, "An automatic seismic signal detection algorithm based on the Walsh transform," *Bulletin of the Seismological Society of America*, vol. 71, no. 4, pp. 1351–1360, 8 1981.

39. S. Greenhalgh, D. Sollberger, C. Schmelzbach, and M. Rutty, *Single-station polarization analysis applied to seismic wavefields: A tutorial*, 1st ed. Elsevier Inc., 2018, vol. 59. [Online]. Available: http://dx.doi.org/10.1016/bs.agph.2018.09.002

40. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70. JMLR, 2017, pp. 1321–1330.

41. R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1735–1742, 2006. [Online]. Available: http://ieeexplore.ieee.org/document/1640964/

42. K. Hara, R. Vemulapalli, and R. Chellappa, "Designing Deep Convolutional Neural Networks for Continuous Object Orientation Estimation," *arXiv e-prints*, vol. abs/1702.0, 2017. [Online]. Available: http://arxiv.org/abs/1702.01499

43. D. B. Harris, "A waveform correlation method for identifying quarry explosions," *Bulletin of the Seismological Society of America*, vol. 81, no. 6, pp. 2395–2418, 12 1991.

44. ——, "Subspace Detectors: Theory," Lawrence Livermore National Laboratory (LLNL), Livermore, CA, Tech. Rep., 7 2006. [Online]. Available: http://www.osti.gov/servlets/purl/900081-cxsqw2/

45. ——, "Comparison of the direction estimation performance of high-frequency seismic arrays and three-component stations," *Bulletin of the Seismological Society of America*, vol. 80, no. 6B, pp. 1951–1968, 1990.

46. A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv e-prints*, vol. abs/1703.0, 2017.

47. M. J. Hinich, "Frequency–wavenumber array processing," *The Journal of the Acoustical Society of America*, vol. 69, no. 3, pp. 732–737, 3 1981. [Online]. Available: https://doi.org/10.1121/1.385572

48. E. Hoffer and N. Ailon, "Deep metric learning using triplet network," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9370, no. 1271, pp. 84–92, 2015.

49. L. Hutchings and F. Wu, "Empirical Green's Functions from small earthquakes: A waveform study of locally recorded aftershocks of the 1971 San Fernando Earthquake," *Journal of Geophysical Research*, vol. 95, pp. 1187–1214, 2 1990.

50. H. Israelsson, "Correlation of waveforms from closely spaced regional events," *Bulletin of the Seismological Society of America*, vol. 80, no. 6B, pp. 2177–2193, 12 1990.

51. P. Jain, B. Kulis, I. S. Dhillon, and K. Grauman, "Online metric learning and fast similarity search," *Advances Neural Information Processing Systems*, pp. 1–8, 2008. [Online]. Available: http://papers.nips.cc/paper/3446-online-metric-learning-and-fast-similarity-search.pdf

52. P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon, "Metric and Kernel Learning Using a Linear Transformation," *Journal of Machine Learning Research*, vol. 13, 10 2009.

53. S. R. Jammalamadaka and A. SenGupta, *Topics in Circular Statistics*. WORLD SCIENTIFIC, 4 2001, vol. Volume 5. [Online]. Available: https://doi.org/10.1142/4031

54. D. Jang and C. D. Yoo, "Fingerprint matching based on distance metric learning," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1529–1532.

55. W. N. Junek, T. F. Vandemark, T. R. Saults, D. B. Harris, D. A. Dodge, S. Matlagh, G. A. Ichinose, A. Poffenberger, and R. C. Kemerait, "Integration of Empirical Signal Detectors into the Detection and Feature Extraction Application at the United States National Data Center," in *CTBTO Science and Technology Meeting*, Vienna, Austria, 2013, p. 116.

56. W. N. Junek, T. Kværna, M. Pirli, J. Schweitzer, D. B. Harris, D. A. Dodge, and M. T. Woods, "Inferring Aftershock Sequence Properties and Tectonic Structure Using Empirical Signal Detectors," *Pure and Applied Geophysics*, vol. 172, no. 2, pp. 359–373, 2 2014. [Online]. Available: https://doi.org/10.1007/s00024-014-0938-0

57. A. Jurkevics, "Polarization analysis of three-component array data," *Bulletin of the Seismological Society of America*, vol. 78, no. 5, pp. 1725–1743, 10 1988.

58. H. Kanamori and M. Ishida, "The foreshock activity of the 1971 San Fernando earthquake, California," *Bulletin of the Seismological Society of America*, vol. 68, no. 5, pp. 1265–1279, 10 1978.

59. C. F. Karney, "Algorithms for geodesics," *Journal of Geodesy*, vol. 87, no. 1, pp. 43–55, 2013.

60. C. Kobryn, J. Wang, J. W. Given, T. C. Bache, S. R. Bratt, and R. M. Fung, "The Intelligent Monitoring System," *Bulletin of the Seismological Society of America*, vol. 80, no. 6B, pp. 1833–1851, 12 1990.

61. G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, vol. 2, 2015.

62. Q. Kong, D. T. Trugman, Z. E. Ross, M. J. Bianco, P. Gerstoft, and B. J. Meade, "Machine Learning in Seismology: Turning Data into Insights," *Seismological Research Letters*, vol. 90, no. 1, pp. 3–14, 11 2018. [Online]. Available: https://dx.doi.org/10.1785/0220180259

63. J. Kortström, M. Uski, and T. Tiira, "Automatic classification of seismic events within a regional seismograph network," *Computers and Geosciences*, vol. 87, pp. 22–30, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.cageo.2015.11.006

64. V. B. G. Kumar, G. Carneiro, and I. Reid, "Learning Local Image Descriptors with Deep Siamese and Triplet Convolutional Networks by Minimizing Global Loss Functions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5385–5394.

65. L. Küperkoch, T. Meier, J. Lee, W. Friederich, and E. W. Group, "Automated determination of P-phase arrival times at regional and local distances using higher order statistics," *Geophysical Journal International*, vol. 181, no. 2, pp. 1159–1170, 5 2010. [Online]. Available: http://dx.doi.org/10.1111/j.1365-246X.2010.04570.x

66. L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, "Learning by Tracking: Siamese CNN for Robust Target Association," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 418–425, 2016.

67. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

68. Z. Li, M.-A. Meier, E. Hauksson, Z. Zhan, and J. Andrews, "Machine Learning Seismic Wave Discrimination: Application to Earthquake Early Warning," *Geophysical Research Letters*, vol. 45, no. 10, pp. 4773–4779, 5 2018. [Online]. Available: https://doi.org/10.1029/2018GL077870

69. A. B. Lockman and R. M. Allen, "Single-station earthquake characterization for early warning," *Bulletin of the Seismological Society of America*, vol. 95, no. 6, pp. 2029–2039, 2005.

70. A. Lomax, C. Satriano, and M. Vassallo, "Automatic Picker Developments and Optimization: FilterPicker—a Robust, Broadband Picker for Real-Time Seismic Monitoring and Earthquake Early Warning," *Seismological Research Letters*, vol. 83, no. 3, pp. 531–540, 5 2012. [Online]. Available: http://dx.doi.org/10.1785/gssrl.83.3.531

71. L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

72. I. W. McBrearty, A. A. Delorey, and P. A. Johnson, "Pairwise Association of Seismic Arrivals with Convolutional Neural Networks," *Seismological Research Letters*, vol. 90, no. 2, 1 2019. [Online]. Available: https://dx.doi.org/10.1785/0220180326

73. J. E. Medalia, *Comprehensive Nuclear-Test-Ban Treaty: Background and Current Developments*, ser. CRS report for Congress. CreateSpace Independent Publishing Platform, 2015.

74. M.-A. Meier, "Advancing Real-Time Seismic Risk Mitigation: Probabilistic Earthquake Early Warning and Physics Based Earthquake Triggering Models," Ph.D. dissertation, ETH Zurich, 2015. [Online]. Available: dx.doi.org/20.500.11850/155496

75. Y. Motoya and K. Abe, "Waveform Similarity among Foreshocks and Aftershocks of the October 18, 1981, Eniwa, Hokkaido, Earthquake," in *Practical Approaches to Earthquake Prediction and Warning*, C. Kisslinger and T. Rikitake, Eds. Dordrecht: Springer Netherlands, 1985, pp. 627–636. [Online]. Available: https://doi.org/10.1007/978-94-017-2738-9_24

76. M. Nakano, D. Sugiyama, T. Hori, T. Kuwatani, and S. Tsuboi, "Discrimination of Seismic Signals from Earthquakes and Tectonic Tremor by Applying a Convolutional Neural Network to Running Spectral Images," *Seismological Research Letters*, vol. 90, no. 2A, pp. 530–538, 1 2019. [Online]. Available: https://doi.org/10.1785/0220180279

77. S. Noda, S. Yamamoto, S. Sato, N. Iwata, M. Korenaga, and K. Ashiya, "Improvement of back-azimuth estimation in real-time by using a single station record," *Earth, Planets and Space*, vol. 64, no. 3, pp. 305–308, 2012.

78. L. H. Ochoa Gutierrez, "Machine Learning Fast estimation of earthquake Back-azimuth using a single seismological station." *AGU Fall Meeting Abstracts*, 12 2018.

79. C. Panagiotakis, E. Kokinou, and F. Vallianatos, "Automatic P-Phase Picking Based on Local-Maxima Distribution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 8, pp. 2280–2287, 2008.

80. R. G. Pearce, I. Kitov, I. Data, C. Division, C. Provisional, and T. Secretariat, "Mitigation of IDC waveform analysts ' increasing workload," in *CTBTO Science and Technology Meeting*, Vienna, Austria, 2011.

81. T. Perol, M. Gharbi, and M. Denolle, "Convolutional Neural Network for Earthquake Detection and Location," *Science Advances*, vol. 4, no. 2, p. e1700578, 2 2018. [Online]. Available: dx.doi.org/10.1126/sciadv.1700578

82. N. Reimers and I. Gurevych, "Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging," *CoRR*, 7 2017. [Online]. Available: http://arxiv.org/abs/1707.09861

83. F. Ringdal and E. S. Husebye, "Application of arrays in the detection, location, and identification of seismic events," *Bulletin of the Seismological Society of America*, vol. 72, no. 6B, pp. S201–S224, 12 1982.

84. A. T. Ringler, D. C. Wilson, T. Storm, B. Marshall, C. R. Hutt, and A. A. Holland, "Noise Reduction in Long-Period Seismograms by Way of Array Summing," *Bulletin of the Seismological Society of America*, vol. 106, no. 5, pp. 1991–1997, 8 2016. [Online]. Available: http://dx.doi.org/10.1785/0120160129

85. Z. E. Ross, M. J. Bianco, P. Gerstoft, B. J. Meade, Q. Kong, and D. T. Trugman, "Machine Learning in Seismology: Turning Data into Insights," *Seismological Research Letters*, vol. 90, no. 1, pp. 3–14, 2018.

86. Z. E. Ross, M. A. Meier, and E. Hauksson, "P Wave Arrival Picking and First-Motion Polarity Determination With Deep Learning," *Journal of Geophysical Research: Solid Earth*, vol. 123, no. 6, pp. 5120–5129, 4 2018. [Online]. Available: http://arxiv.org/abs/1804.08804

87. Z. E. Ross, M. Meier, E. Hauksson, and T. H. Heaton, "Generalized Seismic Phase Detection with Deep Learning," *Bulletin of the Seismological Society of America*, vol. 108, no. 5A, pp. 2894–2901, 10 2018.

88. S. Rost and C. Thomas, "Array Seismology: Methods and Applications," *Reviews of Geophysics*, vol. 40, no. 3, pp. 2–27, 12 2002. [Online]. Available: https://doi.org/10.1029/2000RG000100

89. A. E. Ruano, G. Madureira, O. Barros, H. R. Khosravani, M. G. Ruano, and P. M. Ferreira, "Seismic detection using support vector machines," *Neurocomputing*, vol. 135, pp. 273–283, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2013.12.020

90. C. D. Saragiotis, L. J. Hadjileontiadis, and S. M. Panas, "PAI-S/K: A robust automatic seismic P phase arrival identification scheme," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 6, pp. 1395–1404, 2002.

91. D. Sarkar, R. Bali, T. Ghosh, and O. S. L. Platform, *Hands-on Transfer Learning With Python*, 1st ed. Packt Publishing, 2018.

92. D. P. Schaff, W. Y. Kim, and P. G. Richards, "Seismological Constraints on Proposed Low-Yield Nuclear Testing in Particular Regions and Time Periods in the Past, with Comments on "Radionuclide Evidence for Low-Yield Nuclear Testing in North Korea in April/May 2010" by Lars-Erik De Geer," *Science and Global Security*, vol. 20, no. 2-3, pp. 155–171, 2012.

93. F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *CVPR*, 3 2015.

94. H. Schulte-Theis and M. Joswig, "Master-event correlations of weak local earthquakes by dynamic waveform matching," *Geophysical Journal International*, vol. 113, no. 3, pp. 562–574, 6 1993. [Online]. Available: https://dx.doi.org/10.1111/j.1365-246X.1993.tb04652.x

95. T. J. Sereno and G. B. Patnaik, "Initial Wave-Type Identification with Neural Networks and its Contribution to Automated Processing in IMS Version 3.0," Science Applications International Corporation, San Diego, CA, Tech. Rep., 12 1993.

96. P. Sidiropoulos, "N-sphere chord length distribution," *ArXiv e-prints*, 2014. [Online]. Available: http://arxiv.org/abs/1411.5639

97. R. Sleeman and T. van Eck, "Robust automatic P-phase picking: an on-line implementation in the analysis of broadband seismogram recordings," *Physics of the Earth and Planetary Interiors*, vol. 113, no. 1, pp. 265–275, 1999. [Online]. Available: https://doi.org/10.1016/S0031-9201(99)00007-2

98. W. Stauder and A. Ryall, "Spatial distribution and source mechanism of microearthquakes in Central Nevada," *Bulletin of the Seismological Society of America*, vol. 57, no. 6, pp. 1317–1345, 12 1967.

99. D. A. Strickland, "Scientists as Negotiators: The 1958 Geneva Conference of Experts," *Midwest Journal of Political Science*, vol. 8, no. 4, p. 372, 11 1964. [Online]. Available: https://www.jstor.org/stable/2108688

100. B. W. Stump, M. A. Hedlin, D. C. Pearson, and V. Hsu, "Characterization of mining explosions at regional distances: Implications with the international monitoring system," *Reviews of Geophysics*, vol. 40, no. 4, pp. 2–21, 2002.

101. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 9 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

102. K. A. Sverdrup and T. H. Jordan, "Teleseismic location techniques and their application to earthquake clusters in the South-Central Pacific," *Bulletin of the Seismological Society of America*, vol. 71, no. 4, pp. 1105–1130, 8 1981.

103. R. Tibi, C. Young, A. Gonzales, S. Ballard, and A. Encarnacao, "Rapid and Robust Cross-Correlation-Based Seismic Signal Identification Using an Approximate Nearest Neighbor Method," *Bulletin of the Seismological Society of America*, vol. 107, no. 4, pp. 1954–1968, 7 2017.

104. H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory.* Melbourne, FL, USA: Krieger Publishing Co., Inc., 2 1968.

105. B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

106. J. C. VanDecar and R. S. Crosson, "Determination of teleseismic relative phase arrival times using multi-channel cross-correlation and least squares," *Bulletin of the Seismological Society of America*, vol. 80, no. 1, pp. 150–169, 2 1990.

107. F. Waldhauser and D. P. Schaff, "Large-scale relocation of two decades of Northern California seismicity using cross-correlation and double-difference methods," *Journal of Geophysical Research: Solid Earth*, vol. 113, no. 8, 8 2008.

108. J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning Fine-grained Image Similarity with Deep Ranking," *arXiv e-prints*, 4 2014. [Online]. Available: http://arxiv.org/abs/1404.4661

109. J. Wang, "Adaptive training of neural networks for automatic seismic phase identification," *Pure and Applied Geophysics*, vol. 159, no. 5, pp. 1021–1041, 2002. [Online]. Available: https://dx.doi.org/10.1007/s00024-002-8671-5

110. Z. Wang, W. Li, Y. Kao, D. Zou, Q. Wang, M. Ahn, and S. Hong, "HCR-NET: A hybrid of classification and regression network for object pose estimation," *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1014–1020, 2018. [Online]. Available: https://dl.acm.org/doi/10.5555/3304415.3304559

111. E. P. Xing, M. I. Jordan, S. Russell, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems (NIPS)*, vol. 15, no. 2, pp. 505–512, 2002. [Online]. Available: https://dl.acm.org/doi/10.5555/2968618.2968683

112. Z. Yang, R. Salakhutdinov, and W. W. Cohen, "Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks," *CoRR*, 3 2017. [Online]. Available: http://arxiv.org/abs/1703.06345

113. C. E. Yoon, O. O'Reilly, K. J. Bergen, and G. C. Beroza, "Earthquake detection through computationally efficient similarity search," *Science advances*, vol. 1, no. 11, pp. e1 501 057–e1 501 057, 12 2015. [Online]. Available: https://dx.doi.org/10.1126/sciadv.1501057

114. C. Yu, V. Vavryčuk, P. Adamová, and M. Bohnhoff, "Moment Tensors of Induced Microearthquakes in The Geysers Geothermal Reservoir From Broadband Seismic Recordings: Implications for Faulting Regime, Stress Tensor, and Fluid Pressure," *Journal of Geophysical Research: Solid Earth*, vol. 123, no. 10, pp. 8748–8766, 10 2018. [Online]. Available: https://doi.org/10.1029/2018JB016251

115. S. K. Yung and L. T. Ikelle, "An Example of Seismic Time Picking By Third-Order Bicoherence," *Geophysics*, vol. 62, no. 6, pp. 1947–1952, 5 1997. [Online]. Available: http://library.seg.org/doi/10.1190/1.1444295

116. M. Zhang and L. Wen, "An effective method for small event detection: match and locate (MnL)," *Geophysical Journal International*, vol. 200, no. 3, pp. 1523–1537, 2 2015. [Online]. Available: https://dx.doi.org/10.1093/gji/ggu466

Vita

Joshua Dickey received the M.Sc. degree in electrical engineering from the University of South Florida, in 2007. He is currently pursuing a Ph.D. degree at the Air Force Institute of Technology, where he is involved in the development of advanced signal processing algorithms for the detection and characterization of seismic signals. His research interests include deep learning and remote sensing, particularly in the field of geoscience.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
09-07-2020	Doctoral Dissertation	Sept 2017 — July 2020

4. TITLE AND SUBTITLE

Neural Network Models for Nuclear Treaty Monitoring: Enhancing the Seismic Signal Pipeline with Deep Temporal Convolution

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Joshua T. Dickey

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENG-DS-20-J-004

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Technical Applications Center
Director Systems Development
David C. Merker, SES
1020 S. Patrick Dr. Bldg 10989
PAFB FL 32925-3516
david.merker@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

AFTAC/SD

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Seismic signal processing at the IDC is critical to global security, facilitating the detection and identification of covert nuclear tests in near-real time. This dissertation details three research studies providing substantial enhancements to this pipeline. Study 1 focuses on signal detection, employing a TCN architecture directly against raw real-time data streams and effecting a 4 dB increase in detector sensitivity over the latest operational methods. Study 2 focuses on both event association and source discrimination, utilizing a TCN-based triplet network to extract source-specific features from three-component seismograms, and providing both a complimentary validation measure for event association and a one-shot classifier for template-based source discrimination. Finally, Study 3 focuses on event localization, and employs a TCN architecture against three-component seismograms in order to confidently predict backazimuth angle and provide a three-fold increase in usable picks over traditional polarization analysis.

15. SUBJECT TERMS

Deep Learning; Seismology; Nuclear Treaty Monitoring

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Brett Borghetti, AFIT/ENG
U	U	U	UU	149	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x4612; brett.borghetti@afit.edu