Unsupervised Deep Clustering of Seismic Data: Monitoring the Ross Ice Shelf, Antarctica

William F. Jenkins II[1], Peter Gerstoft[1], Michael J. Bianco[1], Peter D. Bromirski[1]

[1]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA

Key Points:

- Deep embedded clustering (DEC) reveals what types of dominant signals are detected, not just when or where they are detected.
- DEC can be adapted to various kinds of data sets, enabling rapid exploration of "big data" in seismology.
- Paired with environmental data, deep embedded clustering could provide insights into the causes of seismicity.

Corresponding author: W. F. Jenkins II, wjenkins@ucsd.edu

Abstract

Advances in machine learning (ML) techniques and computational capacity have yielded state-of-the-art methodologies for processing, sorting, and analyzing large seismic data sets. In this work, we consider an application of ML for automatically identifying dominant types of impulsive seismicity contained in observations from a 34-station broadband seismic array deployed on the Ross Ice Shelf (RIS), Antarctica from 2014 to 2017. The RIS seismic data contain signals and noise generated by many glaciological processes that are useful for monitoring the integrity and dynamics of ice shelves. Deep embedded clustering (DEC) was employed to efficiently investigate these signals. DEC automatically groups these signals into hypothetical classes without the need for manual labeling, allowing for comparison of their signal characteristics and spatial and temporal distribution with potential source mechanisms. The DEC algorithm uses spectrograms as input and encodes their salient features into a 9-feature representation. Encoding is performed with an autoencoder, a type of deep neural network that is trained iteratively and seeks to reconstruct the input spectrograms from the encoded representation. Eight classes of dominant seismic signals were identified and compared with environmental data such as temperature, wind speed, tides, and sea ice concentration. The greatest seismicity levels occurred at the RIS front during the 2016 El Niño summer, and near grounding zones near the front throughout the deployment. We demonstrate the spatial and temporal association of certain classes of seismicity with seasonal changes at the RIS front, and with tidally driven seismicity at Roosevelt Island.

Plain Language Summary

Using a machine learning technique called deep embedded clustering (DEC), we demonstrate the ability to automatically identify different types of impulsive seismic signals. The DEC algorithm encodes spectrograms into simplified representations and separates the representations into distinct clusters of signal types. The DEC technique was applied to seismic data recorded on the Ross Ice Shelf, Antarctica from 2014 to 2017. In addition to knowing when and where signals are detected, DEC enables users to determine the signal characteristics. Paired with environmental data, DEC can be used to identify whether certain environmental factors are associated with particular classes of seismicity.

1 Introduction

Ice sheets and ice shelves in West Antarctica are experiencing rapid change. Between 2003 and 2019, the West Antarctic Ice Sheet (WAIS) experienced a net ice loss of 169 billion tons per year, contributing 7.5 mm to sea level rise (Smith et al., 2020). Warming oceans are enhancing basal melting of ice shelves that reduces the buttressing of grounded ice sheets (De Angelis & Skvarca, 2003; Thoma et al., 2008; Pritchard et al., 2012; Paolo et al., 2015), leading to increased discharge of ice into the ocean and raising sea level (Scambos et al., 2004; Dupont & Alley, 2005; Rignot et al., 2014; Fürst et al., 2016). With West Antarctica alone containing a sea level rise potential of 5.6 m (Smith et al., 2020), monitoring the loss of ice shelves plays a critical role in anticipating future sea level rise and associated societal impacts on coastlines and the environment. Increased seismic activity, such as icequakes resulting from fracturing, can give indications of changes in iceberg calving rates and the integrity of ice shelves and are observable using glacial seismology methods (Aster & Winberry, 2017). However, the prevalence of extensive, continuously recording seismic observing systems has led to an abundance of data which is becoming increasingly difficult to analyze using conventional signal processing. At the same time, advances in computing capabilities and machine learning algorithms have enabled more efficient, data-driven approaches to study natural processes and phenomena. To analyze large seismic data sets more efficiently, we adapt contemporary machine learning techniques to augment existing signal processing and data analysis techniques.

Seismology is a data-intensive field with well-developed signal processing and analytical methods. The recent introduction of machine learning techniques has led to the development of complementary tools that give seismologists novel approaches to traditional analyses, such as earthquake detection and early warning, phase picking, ground-motion prediction, tomography, and geodesy (Kong et al., 2019; Bianco et al., 2019; Johnson et al., 2019). In this study we present an extension of *clustering* (Mousavi et al., 2016; Snover et al., 2020), a form of unsupervised machine learning used to discover classes of similar signals within a data set (Bishop, 2006; Holtzman et al., 2018; Johnson et al., 2020), and which is commonly used as an exploratory tool for large, unlabeled data sets.

To test the applicability of clustering groups of similar signals for monitoring ice shelves, we focus specifically on the Ross Ice Shelf (RIS), Antarctica, where a 34-station passive seismic array was deployed from November 2014 to January 2017 to observe the response of the RIS to ocean gravity wave impacts and investigate the structural dynamics of the ice shelf (Bromirski et al., 2015). The array, shown in Figure 1, continuously recorded long- and short-period seismic signals that exhibited seasonal and spatial variations related to the shelf's coupling to the ocean, atmosphere, and crust (Baker et al., 2019). Signals and ambient noise of interest on the RIS include tidally-driven stick-slip seismicity at Whillans Ice Stream (Bindschadler, King, et al., 2003; Bindschadler, Vornberger, et al., 2003; D. A. Wiens et al., 2008); basal micro-earthquakes and tremor (Barcheck et al., 2018); tidally and thermally driven rift fractures (Olinger et al., 2019); diurnal seismicity associated with subsurface melting (MacAyeal et al., 2019); wind-generated resonance in the ice (Chaput et al., 2018); flexural and plate waves generated by ocean swell, infragravity waves, and tsunami (Bromirski & Stephen, 2012; Bromirski et al., 2017; Chen et al., 2018); regional and teleseismic earthquakes (Baker et al., 2020); and icequakes generated by ocean gravity waves (Chen et al., 2019). Ambient seismic noise, which can be used to estimate the RIS structure (Diez et al., 2016), also contains spectra from ocean gravity waves, whose dispersion can be used to identify their source distance and origin (Bromirski et al., 2015; Hell et al., 2019).

The seismic data recorded on the RIS are diverse and encompass numerous source mechanisms with a wide range of spatiotemporal variability. In this study, we apply an unsupervised clustering methodology to the RIS array sesimic data to identify classes of seismic events with similar temporal and spectral characteristics. The occurrences and distributions of these signal classes provide information on glaciological processes affecting ice shelf evolution.

2 Background

Grouping seismic signals with similar characteristics (clustering) allows investigation of spatiotemporal variability associated with glaciological processes that result from environmental forcing.

2.1 Clustering

There are numerous methods to cluster data, (Aggarwal & Reddy, 2014), many of which have been adapted for use in seismology and geophysics (Kong et al., 2019). Hi-

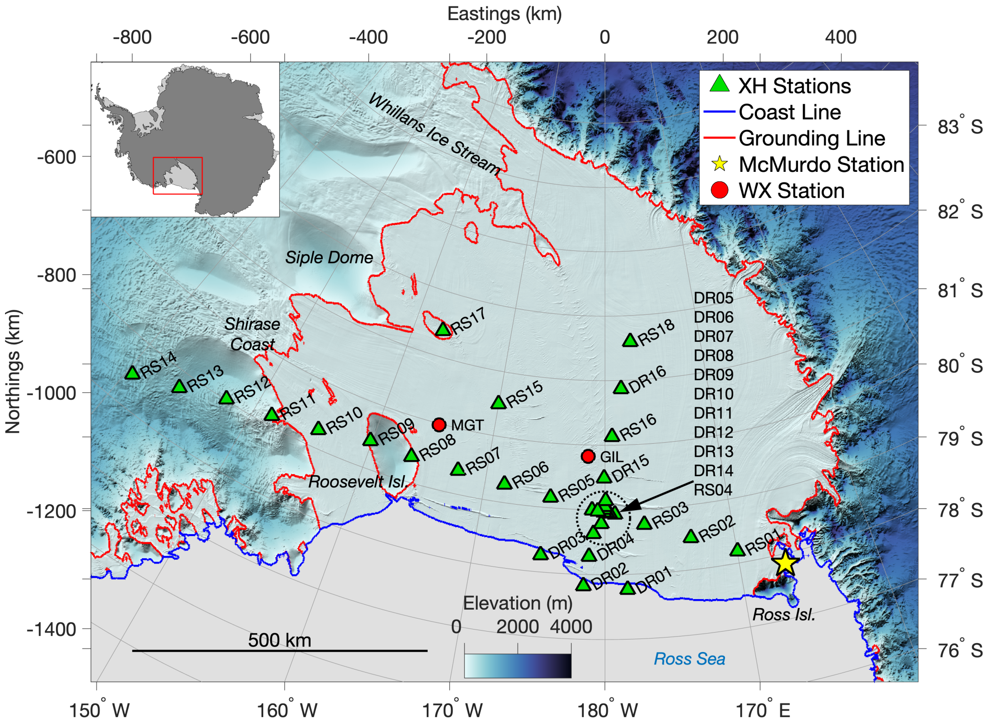

Figure 1. The passive broadband seismic array deployed from November 2014 to January 2017 consisted of 34 seismic stations and was deployed as part of the Ross Ice Shelf Dynamic Response to Wave-Induced Vibrations Project (Bromirski et al., 2015). RIS surface elevation, ice and water layer thicknesses, and grounding and coast lines were obtained from Bedmachine (Morlighem et al., 2017; Greene et al., 2017).

erarchical clustering has been used by Mousavi et al. (2016) to automatically discriminate between shallow and deep earthquakes, and by Trugman and Shearer (2017) to more precisely localize earthquakes. Graphical clustering has been used to localize sources in a dense seismic array by Riahi and Gerstoft (2017), and by Telesca and Chelidze (2018) to cluster seismic events in time. Distance-based clustering, like the popular k-means algorithm, (MacQueen, 1967; Hartigan & Wong, 1979) has been used by Chamarczuk et al. (2020) to cluster seismicity based on features extracted from seismic data. Perol et al. (2018) used k-means to define probabilistic earthquake locations as part of their convolutional neural network (CNN) detection and localization technique. A novel approach was presented by Seydoux et al. (2020), who detect and cluster seismic signals and background noise with the use of a deep scattering neural network and a Guassian mixture model.

Not all clustering methods involve machine learning. Template matching, in which a matched filter is constructed from a template waveform, is used to scan through continuous recordings to locate similar signals (Gibbons & Ringdal, 2006; Beaucé et al., 2018; Chamberlain et al., 2018). Yoon et al. (2015) and Bergen and Beroza (2018) present computationally efficient techniques in which seismic data are encoded using locally sensitive hashing, allowing similar signals to be identified within a hash table. Hotovec-Ellis and Jeffries (2016) developed an approach that uses correlation-based similarity search to automatically detect and cluster repeating volcanic seismicity in continuous data. Cole (2020) adopted the method of Hotovec-Ellis and Jeffries (2016) to cluster RIS array data at stations RS09, RS10, and RS11 in order to characterize tidal forcing of seismicity at these stations.

2.2 Dimensionality

Data are considered high-dimensional when many features are required to represent or describe the data. Seismic data represented as time series, spectrograms, scalograms, or energy envelopes can contain thousands of features (e.g., discrete samples in a time series, or bins in a spectrogram). Clustering performed directly on such input data (Aggarwal & Reddy, 2014) is vulnerable to the "curse of dimensionality" (Bellman, 1961), i.e., as the dimensionality of the input data increases, the number of data points required to maintain sufficient sampling density increases exponentially. A further consideration is that clustering error metrics can give less meaningful results as dimensionality increases.

As high-dimensional data are difficult to cluster (Aggarwal et al., 2001; Steinbach et al., 2004), dimensionality reduction remains a major focus of development (Yang et al., 2017). To address the issue of dimensionality, it is desirable to transform the input data to a lower-dimensional representation described by fewer, more salient features. Reddy et al. (2012) use principal component analysis to compress seismic data to maximize feature variance, while Bianco and Gerstoft (2018) and Bianco et al. (2019) reduce dimensionality using adaptive dictionary learning for high-resolution seismic tomography.

The approach to reducing dimensionality in this study employs an autoencoder, a model whose output aims to reproduce its input via a series of non-linear transformations employing a deep neural network (DNN) (Hinton, 2006; Murphy, 2012; Yang et al., 2017). These non-linear transformations provide greater capacity in dimension reduction, and can better model data with low-dimensional representations than, for example, principal component analysis (PCA) (Goodfellow et al., 2016). The autoencoder first encodes input data such as an image—in our case, a spectrogram—into a latent feature vector. Next, the autoencoder decodes the latent features and reconstructs the original image. Since the autoencoder provides a non-linear transformation of the data, it must be trained using gradient descent. In this iterative training, the error between the input and output is minimized. In doing so, the salient features of the data are learned by the network weights. With the dimensionality of the input data reduced in the latent feature space, clustering algorithms can be applied to the data's latent feature space.

2.3 Clustering in Reduced Dimensions

A method that has shown improvement over traditional clustering techniques was developed by Xie et al. (2016), whose *deep embedded clustering* (DEC) uses the latent feature space as input to an adaptive clustering algorithm. DEC consists of two processes: first, an autoencoder is trained to represent the data's salient features; then the encoding layers and clustering layer are jointly optimized. Yang et al. (2017) extend the approach in DEC by jointly optimizing the clustering step with training the entire autoencoder, not just the encoder layers.

Additional variations of DEC have been proposed: Xie et al. (2016) used a stacked de-noising autoencoder (Vincent et al., 2010), and (Min et al., 2018) employed autoencoders composed of CNN layers and other architectures. More recently, Chazan et al.

(2019) developed a novel approach in which joint clustering is performed with a mixture of autoencoders, each representing a cluster. Mousavi et al. (2019) used DEC to predict whether seismic detections were local or teleseismic, and Snover et al. (2020) demonstrated DEC's ability to cluster anthropogenically generated seismic noise.

In this study, we implement DEC on RIS seismic data collected from December 2014 to November 2016, identifying several different classes of signals. Additionally, we demonstrate the utility of DEC as an exploratory tool for large, real-world seismic data sets by associating the clustering results with observed environmental factors.

3 Ross Ice Shelf Seismic Array and Data

Each station in the RIS seismic array consisted of 3-component Nanometrics Trillium 120 PHQ seismometers emplaced 1 m below the surface of the ice, powered by solar panels during the austral summers, and lithium-ion batteries during the austral winters. Two subarrays comprised the array. The larger subarray consisted of 18 stations spaced approximately 80 km apart (prefix RS), primarily oriented parallel to the RIS front. The RS stations sampled short-period orthogonal components of ground velocity at a sampling rate of 100 Hz, except for two stations that sampled at 200 Hz. The smaller subarray consisted of 16 stations (prefix DR) arranged approximately orthogonal to the icefront along the international date line, sampling ground velocity with a sampling rate of 200 Hz. For this study, we were primarily interested in the detection and classification of icequakes and local/regional earthquakes, using only vertical component observations. Representative types of signals detected are shown in Figure 2.

Seismic data from each station were processed in 24-hour segments. Data were linearly de-trended, tapered with a Hann window, and decimated to a sampling rate of 50 Hz. Instrument responses for all stations were removed, giving acceleration in m/s^2. A bandpass filter with cutoff frequencies at 3 and 20 Hz was applied to remove long-period signals originating from tides, tsunamis, infragravity waves, ocean swell, and teleseisms. An event detection algorithm, the Z-detector (Swindell & Snell, 1977; Withers et al., 1998), was used to detect impulsive signals, particularly icequakes and local earthquakes, with a sliding window of 3 s. The detector was applied to data from each station between 2 December 2014 and 20 November 2016 for a total of 719 days of array data, yielding 427,798 detections.

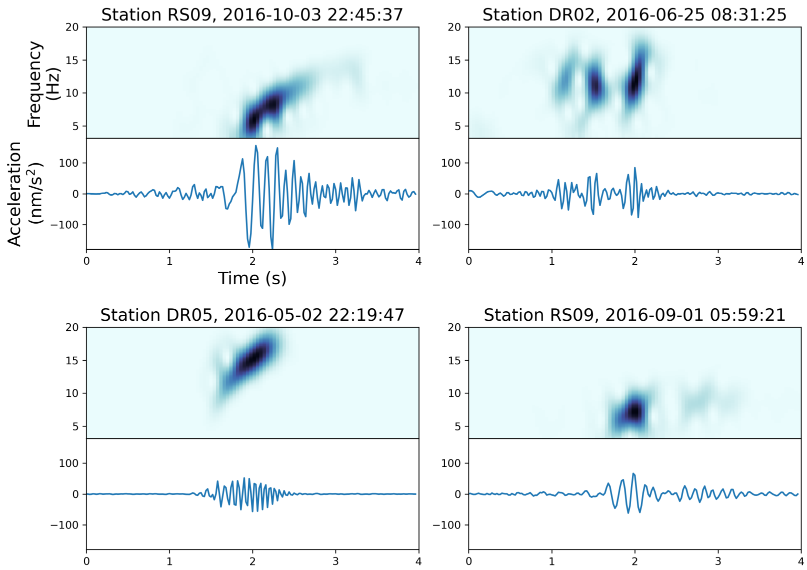

Figure 2. Seismic signals detected on the Ross Ice Shelf were diverse with variation in time, space, and source mechanism. Shown are examples of acceleration response seismograms and their respective normalized spectrograms spanning the 3-20 Hz band that were typical for the data set. The normalized spectrograms were used as input to the deep embedded clustering (DEC) model.

Upon detection, a 4 s trace centered on the spectral peak of each triggered event was saved for processing. For each seismic trace saved, a spectrogram was computed using the short-time Fourier transform with a 0.4 s Kaiser window, NFFT=256, and 90% overlap. Spectrograms contained one channel of amplitude information, 87 frequency bins, and 100 time bins for a total of 8,700 features per spectrogram. Finally, for each spectrogram, the spectral levels were centered about the mean spectral value and normalized to the interval $[-1, 1]$.

4 Deep Embedded Clustering Model

The objective of the DEC model, shown in Figure 3, is to encode the input data—in this case, spectrograms of seismic signals—into a layer containing latent (lower-dimensional) features, called the *embedded* layer, to which a clustering algorithm is applied. The outcome of the clustering performance is then used to refine both the autoencoder model and the clustering layer in an effort to obtain more accurate latent space embeddings while improving clustering performance. In the implementation that follows, the 8,700 features

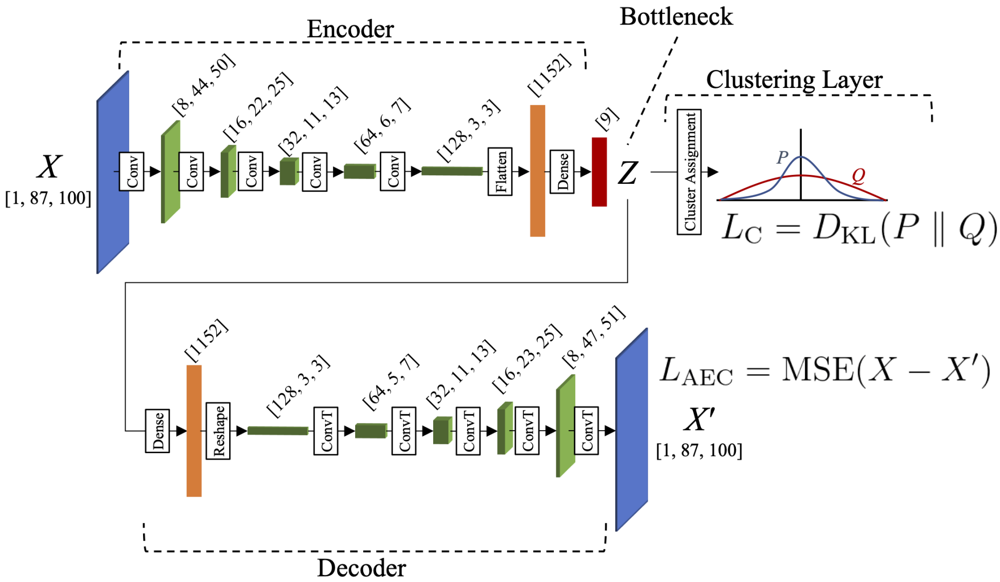

Figure 3. The deep embedded clustering model uses a convolutional autoencoder that encodes the data space X into the latent feature space Z, and a decoder that recovers the original input X from Z. The mean squared error (MSE) between the input X and the reconstruction X' is used as the autoencoder loss function. The latent feature space Z lies at the bottleneck between the encoder and decoder, providing the input to the clustering layer, which separately outputs a loss function. The two loss functions are combined and used to train the parameters that map $X \to Z \to X'$.

of an unencoded spectrogram are reduced to a latent feature space of just 9 embedded features with the use of a convolutional autoencoder, a type of DNN composed of convolutional layers and their transposes.

4.1 Dimensionality Reduction with a Convolutional Autoencoder

Autoencoders provide a useful means of data approximation using a lower-dimensional representation via a sequence of non-linear transformations. The autoencoder model consists of three components: an *encoder*, a *bottleneck*, and a *decoder* (Murphy, 2012). First, the encoder maps input data from a data space X into a latent feature space Z, which is contained within the bottleneck of the model. Next, the decoder attempts to reconstruct X from Z. This process is performed iteratively with the objective of minimizing the error between X and the decoder output, X'. In minimizing the error, the autoencoder learns the salient features of X and accurately encodes them in Z, thus reducing the dimensionality of the clustering task.

Consider a data set of spectrograms $\mathcal{D} = \{\boldsymbol{x}_i \in X^M\}_{i=1}^N$, where $\boldsymbol{x}_i$ is a vector representation of the i^{th} spectrogram in a data set containing N spectrograms, and the number of features in $\boldsymbol{x}_i$, M, is the spectrogram size (the product of the number of frequency bins and time bins). In the encoder stage, the mapping of X to Z is described by $f_\theta : X \to Z$, where θ are parameters that are learned through iterative model training. The decoder stage is a mirror operation of the encoder and seeks to map the latent feature space Z to the reconstruction X' by $g_\theta : Z \to X'$. The overall mapping of the autoencoder can be described as $F_\theta : X \to Z \to X'$, where $F_\theta = g_\theta \circ f_\theta$. Input spectrograms $\boldsymbol{x}_i$ map to their corresponding latent feature vectors by $\boldsymbol{z}_i = f_\theta(\boldsymbol{x}_i) \in Z^D$, where D is the number of embedded features, and to their reconstructions by $\boldsymbol{x}'_i = F_\theta(\boldsymbol{x}_i) \in X'$.

As the autoencoder is composed of convolutional layers and their transposes, F_θ is a nonlinear mapping that must be appropriately parameterized. This is accomplished by iteratively learning the parameters θ in order to minimize the error between the input and reconstructed data. The mean squared error (MSE) between an input spectrogram with M features and its reconstruction, defined as

$$\ell(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{M} \sum_{m=1}^{M} (x_m - x'_m)^2, \tag{1}$$

is averaged over the N samples in the data set to obtain the autoencoder loss function:

$$L_{\text{AEC}} = \frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{x}_i, \boldsymbol{x}'_i). \tag{2}$$

Performing this calculation over the entire data set at once is computationally expensive, memory intensive, and can lead to poor convergence. Instead, the loss is calculated in mini-batch subsets of the data space. For each mini-batch loss, the gradient of the loss function with respect to the weights comprising θ is computed, and the weights updated. When all mini-batches have been processed in this way, the next training epoch begins and the process is repeated. This mini-batch procedure, called stochastic gradient descent (Goodfellow et al., 2016), effectively adds noise to the gradient from sampling mini-batches, which ensures that the model does not converge on local minima. In parallel with autoencoder training, a subset of the data is used to validate the model's performance after each training epoch. Training is performed until a specified maximum num-

Table 1. *Convolutional Autoencoder Architecture*

Layer Name	Type	Input Shape	Filters	Activation	Output Shape	**Trainable Parameters**
Input	-	-	-	-	[1, 87, 100]	-
Conv1	Convolution	[1, 87, 100]	8	ReLU	[8, 44, 50]	80
Conv2	Convolution	[8, 44, 50]	16	ReLU	[16, 22, 25]	1,168
Conv3	Convolution	[16, 22, 25]	32	ReLU	[32, 11, 13]	4,640
Conv4	Convolution	[32, 11, 13]	64	ReLU	[64, 6, 7]	18,496
Conv5	Convolution	[64, 6, 7]	128	ReLU	[128, 3, 3]	73,856
Flat	Flatten	[128, 3, 3]	-	-	[1,152]	0
Encoded	Fully Connected	[1,152]	-	ReLU	[9]	10,377
FC	Fully Connected	[9]	-	ReLU	[1,152]	11,520
Reshape	Reshape	[1,152]	-	-	[128, 3, 3]	0
ConvT1	Transposed Conv	[128, 3, 3]	64	ReLU	[64, 5, 7]	73,792
ConvT2	Transposed Conv	[64, 5, 7]	32	ReLU	[32, 11, 13]	18,464
ConvT3	Transposed Conv	[32, 11, 13]	16	ReLU	[16, 23, 25]	4,624
ConvT4	Transposed Conv	[16, 23, 25]	8	ReLU	[8, 47, 51]	1,160
Decoded	Transposed Conv	[8, 47, 51]	1	Linear	[1, 95, 101]	73
Output	Crop	[1, 95, 101]	-	-	[1, 87, 100]	-
					Total	**218,250**

ber of epochs is reached, or if no improvement in the validation data's MSE is observed after ten epochs.

The design choice of autoencoder architecture can be informed by *a priori* knowledge of a data set and its features, as well as practical considerations such as computational resources available. Our DNN architecture, detailed in Table 1, was designed to be computationally efficient, simple to construct, and robust enough to learn salient features from a noisy seismic data set. The model's encoder is composed of five, two-dimensional convolutional layers with rectified linear unit (ReLU) activation functions after each layer. The kernel shape of each convolution is [3, 3], with a stride of two pixels. The number of filters doubles at each layer. The output of the final convolution is flattened and linearly transformed with a dense layer into the encoded layer, which contains the embedded latent feature space. For decoding, the inverse of the encoder operations are performed

Table 2. *Deep Embedded Clustering Model Parameters with Selected Hyper-parameters*

N	N_{train}	N_{val}	Learning Rate	Batch Size	Clusters	λ
427,798	100,000	25,000	0.001	1024	8	0.05

through linear transformation, reshaping, and transpose convolutions. In total, θ contains 218,250 trainable parameters under this architecture.

During training, the trainable parameters are optimized using the Adaptive Moment Estimation (Adam) algorithm (Kingma & Ba, 2017), which propagates the loss after each batch iteration backwards through the model. In this configuration, there are two principal hyper-parameters to address. First is the learning rate, which controls the step size used to step down the gradient of the loss. A learning rate which is too small will cause the model to converge extremely slowly or settle into a local minimum, while one that is too large may converge slowly or not at all. The second hyper-parameter is the batch size, which sets the number of spectrograms to be passed through the model at one time. The optimal configuration is found through hyper-parameter tuning, a process in which the autoencoder is trained with the various hyper-parameter permutations, with the hyper-parameters that yield the best results kept.

Autoencoder training was performed using spectrograms randomly selected without replacement. Of the selected spectrograms, 80% were used for training and 20% for validation. The number of spectrograms used and the optimal hyper-parameters are listed in Table 2. As seen in Figure 4a, training and validation losses fall off exponentially with each training epoch. To prevent the autoencoder from over-fitting, training is stopped early if the validation loss has not decreased in 10 epochs. The effectiveness of the autoencoder's ability to reconstruct the input spectrogram is illustrated in Figure 5. Because the convolutional layers and their transposes consist of down-sampling and up-sampling operations, some loss of resolution in time and frequency is expected. Nevertheless, the structure of the spectrogram is largely preserved, with the salient information of the input encoded into the latent feature space.

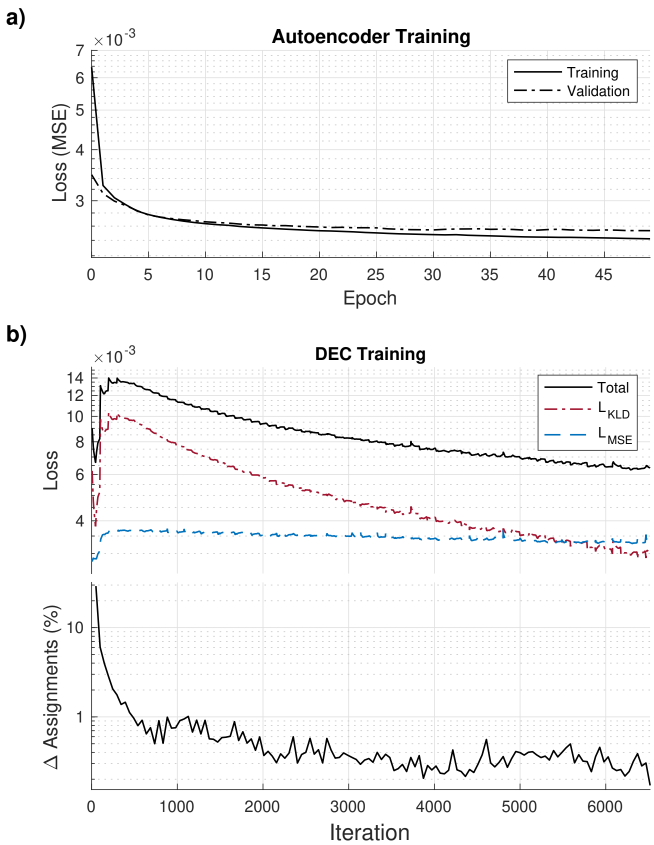

Figure 4. (a) Training and validation losses during autoencoder training. To avoid over-fitting the model, training was stopped at 49 epochs when the validation error began to increase. (b) In the upper plot, loss curves are shown for deep embedded clustering (DEC). In the lower plot, the percentage of samples which undergo class reassignment at each update interval is shown; training is stopped once the percent change is less than 0.2%

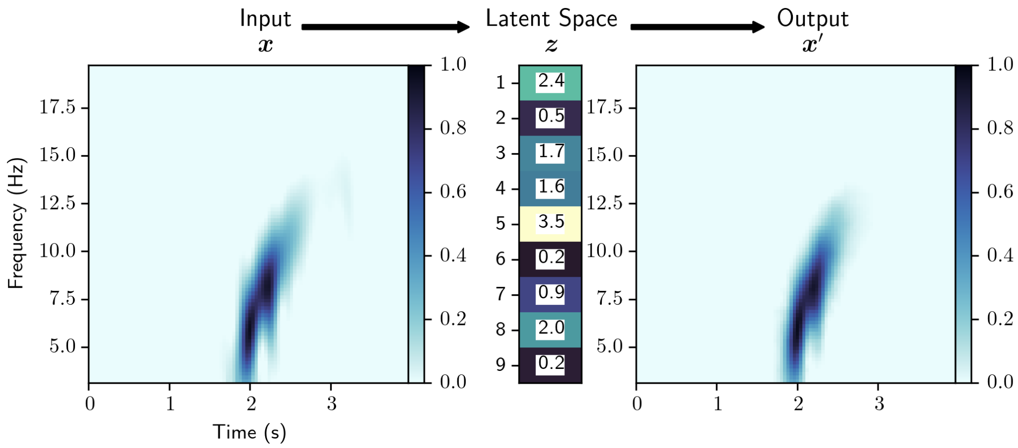

Figure 5. A trained autoencoder takes an input spectrogram $\boldsymbol{x}$, encodes it into an embedded, 9-dimensional latent feature vector $\boldsymbol{z}$, then reconstructs the input as $\boldsymbol{x}'$. The autoencoder preserves features correlated within a given cluster and discards the remaining signal. This can help with signal identification.

4.2 Clustering Layer

In the DEC framework, clustering is performed in the latent feature space, Z, with the goal of finding K distinct classes of signals within the data. We assume that the data form clusters which are separable in Z space, and that these clusters coalesce around unique locations $\{\boldsymbol{\mu}_j \in Z\}_{j=1}^{K}$, i.e., centroids around which other similar signals may be found. The Euclidean distance between a centroid and a latent feature vector is given by

$$d_{i,j} = \|\boldsymbol{z}_i - \boldsymbol{\mu}_j\|_2. \tag{3}$$

The distance metric $d_{i,j}$ is a measure of the similarity between features indexed by i and j, based on their Euclidean distance in the latent space.

Centroids are initialized with the k-means clustering algorithm (Hartigan & Wong, 1979), which randomly seeds the centroids and then iteratively updates their positions by minimizing the inertia, defined as the sum of the Euclidean distance between the centroids and the data points within each cluster:

$$\sum_{i=1}^{n} \min_{\boldsymbol{\mu}_j \in k} (d_{i,j}). \tag{4}$$

Because the k-means solution is guaranteed to converge on a local minimum, the algorithm is sensitive to the random seeding in its initialization. Two steps are taken to mitigate this. First, to improve the random seeding procedure and to speed up convergence, an improved initialization scheme, k-means++ (Arthur & Vassilvitskii, 2007), is used instead of purely random seeding. Second, the k-means algorithm is repeated 100 times, and the solution that yields the minimum inertia is used to intialize the centroids.

With the centroids initialized by k-means, DEC seeks to further improve clustering by using the difference between the embedded spectrograms and the cluster centroids as a loss function that provides feedback to the model parameters. Because the input data is unlabeled, a self-supervised method is required. We implement the method developed by Xie et al. (2016), who, drawing from the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm (van der Maaten & Hinton, 2008), propose measuring the difference between a t-Student's distribution kernel of the embedded spectrograms and an auxiliary target distribution. A simplified Student's t-distribution is used to measure the similarity between an embedded spectrogram, $\boldsymbol{z}_i$, and the cluster centroids $\boldsymbol{\mu}_j$:

$$q_{ij} = \frac{(1+ \parallel \boldsymbol{z}_i - \boldsymbol{\mu}_j \parallel^2)^{-1}}{\sum_j (1+ \parallel \boldsymbol{z}_i - \boldsymbol{\mu}_j \parallel^2)^{-1}}. \tag{5}$$

Equation (5) results in a set of soft class assignments, i.e., the probability that embedded spectrogram i will be assigned to class j, which are used to compute the auxiliary target distribution, p. The form of p is designed to improve clustering performance, emphasize embeddings with high-confidence assignments, and normalize each cluster centroid's contribution to the loss function so that large clusters minimally distort Z (Xie et al., 2016):

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})}. \tag{6}$$

The dissimilarity between the distributions given by equations (5) and (6) is measured using the Kullback-Leibler divergence (Kullback & Leibler, 1951). From the divergence the clustering layer's loss function is obtained:

$$L_{\mathrm{C}} = D_{\mathrm{KL}}(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{7}$$

In DEC, the clustering layer is attached to the trained autoencoder's bottleneck, as shown in Figure 3. During training of the DEC model, the loss functions from equations (1) and (7) are combined into a total loss function,

$$L = L_{\mathrm{AEC}} + \lambda L_{\mathrm{C}}, \tag{8}$$

where λ is a hyper-parameter that balances the contributions of the two losses, since they are of differing magnitudes. λ must be tuned: if it is too large, the clustering loss will cause model instability and lead to distortion of the latent space, in which case the latent space will no longer represent the salient features of the data. If λ is too small, the effect on clustering performance will be minimal. We found that $\lambda = 0.05$ yielded optimal performance for model training and clustering.

Two constituent processes occur simultaneously during DEC model training. First, the full loss from equation (8) is backpropagated through the DEC model parameters, which include the autoencoder as well as the cluster centroids. Second, to account for the cluster centroids changing as training progresses, the distributions q_{ij} and p_{ij} are updated at intervals. The update interval is a hyper-parameter that must be tuned. Through hyper-parameter tuning, an update interval of twice per training epoch was found to be optimal for clustering performance, minimizing DEC loss, and training within a reasonable time frame. Training is stopped after the number of samples changing assignments after every update interval reaches less than 0.2% of the total number of samples.

4.3 Selecting Optimal Number of Clusters

Determining the optimal number of clusters, K, is a major challenge in unsupervised machine learning. Although there are statistical methods available for choosing the optimal number of clusters (Rousseeuw, 1987; Tibshirani et al., 2001), in this study we treat K as a hyper-parameter, iterating the DEC workflow over a range of values for K and evaluating the results to choose the best value. Results are evaluated both quantitatively and qualitatively. Quantitative evaluation is performed for each class by examining cumulative distribution functions and probability density functions as functions of distance to each class centroid, $d_{i,j}$ (equation (3)). The qualitative approach is to visually inspect the similarity of the latent feature vectors $\boldsymbol{z}_i$ to their respective class centroids $\boldsymbol{\mu}_j$, and to see if the spectrograms and seismograms assigned to each class likewise exhibit similarity. In general, the formation of two or more similar classes may indicate that too many classes were initialized, and the data can be grouped into a single class in post-processing. Too much variance among the spectrograms within a class may indicate the need for one or more additional classes. We found that $K = 8$ was the optimal number of classes for the full RIS data set.

5 Results

5.1 Clustering Performance

Clustering with DEC results in two distinct phases: first, the k-means algorithm sets the initial centroids, but the latent data are left unmodified. Second, during DEC, centroids are further refined while the latent data are moved much closer to their respective centroids, with some data reassigned to different classes altogether.

The performance of DEC is qualitatively checked by comparing centroids to their respective assigned latent data samples. Results for the training data set are shown in Figure 6. Each class j is represented by the columns in Figure 6, with each centroid $\boldsymbol{\mu}_j$ and its reconstruction $g_\theta(\boldsymbol{\mu}_j)$ plotted along the top row. Although the centroid is not a member of the data set, because the centroid represents the salient features of its class, its reconstruction is expected to resemble the spectrograms $\boldsymbol{x}_i$ assigned to its class. Subsequent rows show the latent feature vectors $\boldsymbol{z}_i$, spectrograms $\boldsymbol{x}_i$, and associated seismograms of the data samples assigned to the respective classes.

For each class, latent feature vectors $\boldsymbol{z}_i$ exhibit similar values to the class centroid $\boldsymbol{\mu}_j$, indicating that DEC has successfully grouped similar latent data samples into a class, and that the centroid is representative of the data in its class. The spectrograms in each class are likewise similar to each other and to the centroid reconstruction $g_\theta(\boldsymbol{\mu}_j)$, confirming that the latent features embedded in the centroids are representative of the spectrograms in the class. Finally, the similarity in the latent space and the time-frequency domain extends to the time domain, where seismograms in each class are similar to one another.

In addition to checking the efficacy of the clustering, visual examination of the results in Figure 6 permits an indication of whether or not an appropriate number of clusters was chosen. For example, classes 4 and 8 exhibit similar characteristics in time and frequency, distinct from each other primarily in peak frequency. If such distinctions are not useful or if similarities are redundant, classes can be combined in post-processing. If too few clusters are selected, classes may contain widely differing signals, indicating the need to increase the number of clusters.

To determine to what extent DEC further improves clustering over k-means clustering, t-SNE is used to visualize the 9-dimensional latent space in two dimensions (van

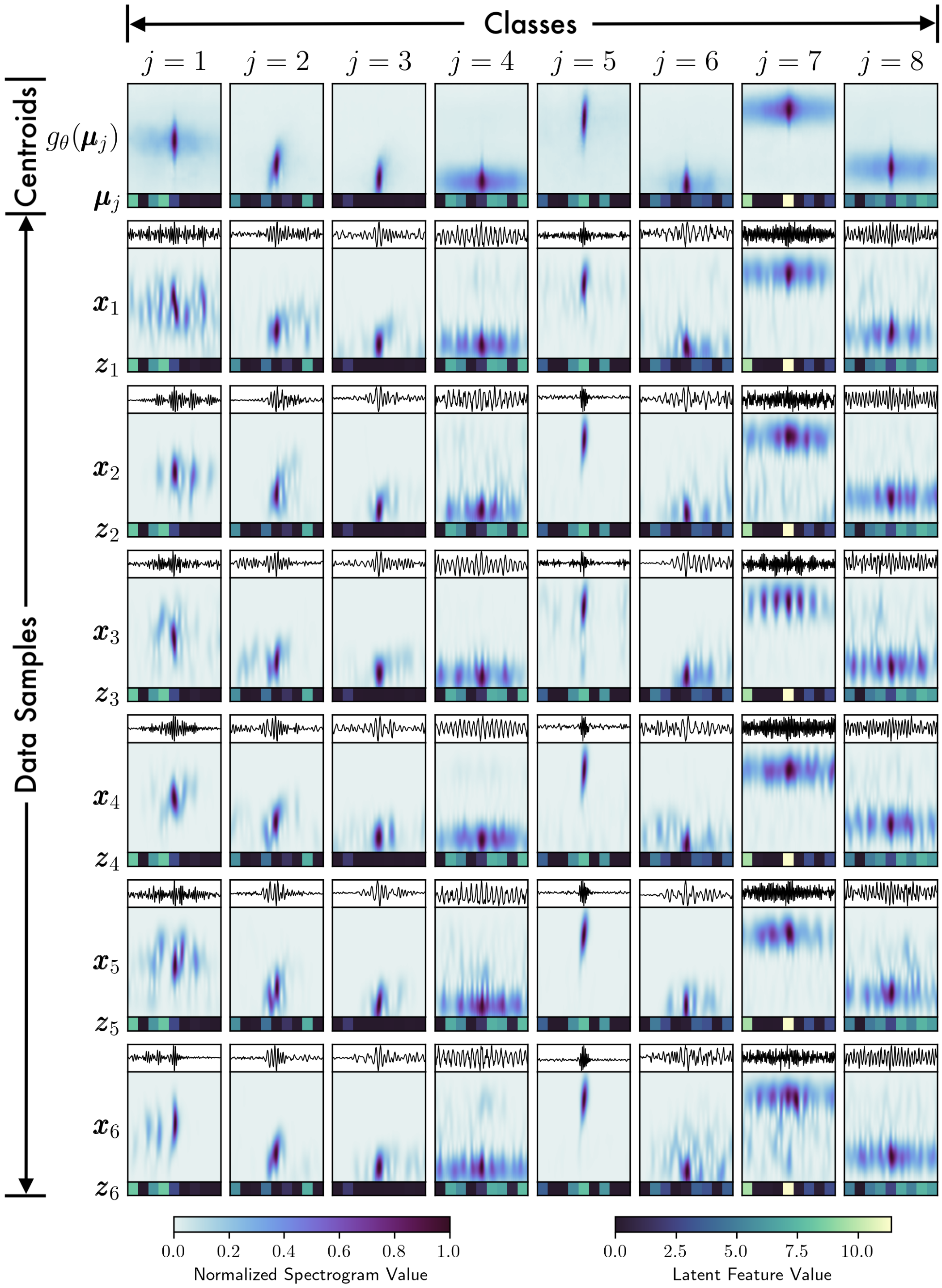

Figure 6. Within a given class j, the cluster centroids $\boldsymbol{\mu}_j$ are similar to the latent feature data $\boldsymbol{z}_i$. Though the centroids are not members of the data set, their reconstructions $g_\theta(\boldsymbol{\mu}_j)$ exhibit similar characteristics to the spectrograms $\boldsymbol{x}_i$ assigned to each class. Seismograms plotted above each spectrogram also exhibit similarity within each class.

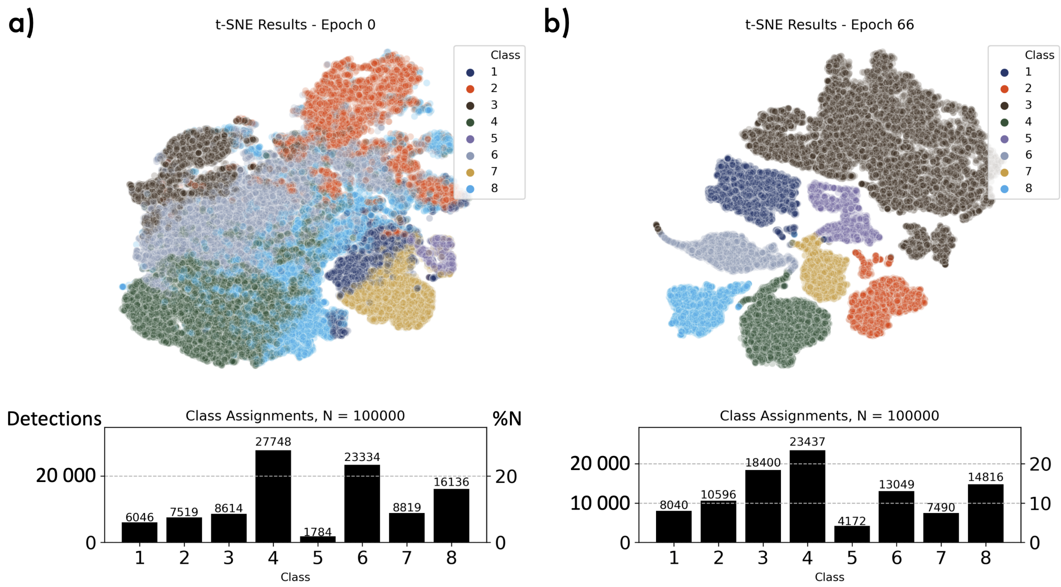

Figure 7. Visualization of the 9-dimensional latent data space is shown in two dimensions using the t-Student Stochastic Neighbor Embedding (t-SNE) plot. (a) Clustering and label assignment is performed with k-means clustering, with relatively poor separation within the data and overlapping classes. (b) After DEC, clusters are well separated and contain nearly homogeneous class members. Class histograms (a) before and (b) after DEC reveal the extent to which DEC reassigns latent data samples to different classes.

der Maaten & Hinton, 2008). t-SNE can illuminate possible clusters within data in an unsupervised manner by displaying data in geometrically separated clusters. In Figure 7, t-SNE results of the latent feature space clustered with k-means show that the data are largely contiguous with few exceptions. Applying the labels assigned by k-means to the data points shows that, while there is some geometric separation between the clusters, the embedding is characterized by overlapping and dispersed class members, indicating poor separation in the latent space and potentially incorrect assignment of samples to classes. Contrast this with Figure 7, in which t-SNE results at the conclusion of DEC show both geometric separation as well as homogeneous class assignments.

While t-SNE offers an intuitively visual way to look for clusters in data, results are sometimes difficult to interpret and are impossible to reproduce exactly due to the inherent randomness of the algorithm. Running t-SNE iteratively and with the same random seed can mitigate these limitations, but examination of the effects of DEC on the densities of the clusters provides a more concrete visualization. In Figure 8a, the cumu-

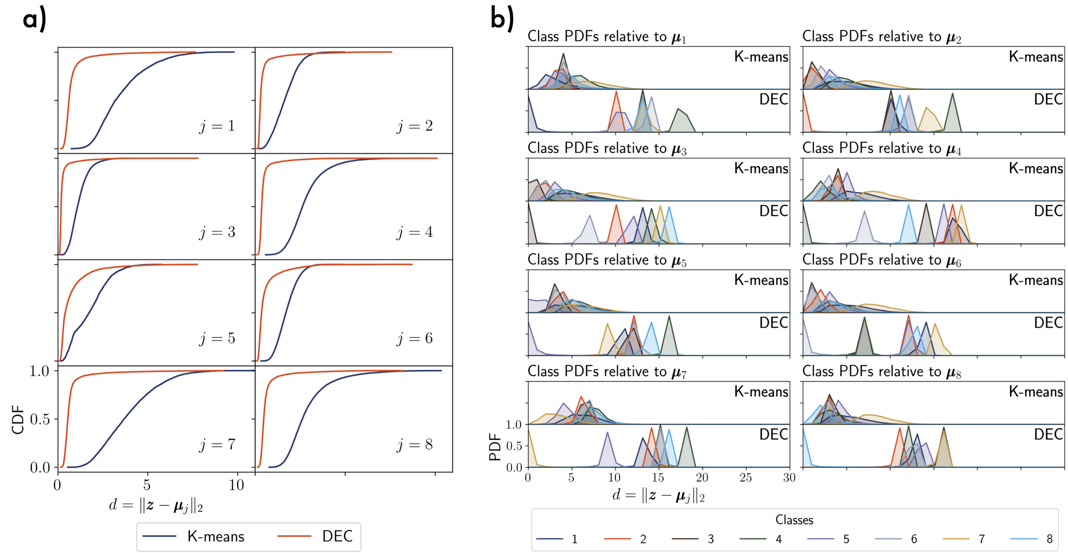

Figure 8. (a) Within each class, deep embedded clustering (DEC) reduces the mean distance of the assigned latent data to the centroid. The variance of the distance also decreases. As a result, the cumulative distribution functions shift to the left and have a steeper slope. (b) The effects of DEC are also evident for each class probability density function (PDF) with respect to the distance from the centroids. In addition to moving the assigned class members closer to the centroid, DEC also increases the distance to the other class centroids and PDFs. The total effect is to separate the latent data samples of one class from the other classes, resulting in better clustering performance than k-means clustering.

lative distribution functions (CDF) for each class are shown as functions of distance to the centroid (equation (3)). For each class, the latent data move substantially closer to their assigned centroid, as evidenced by the decreased mean and variance of the CDF. Of interest to the ability for DEC to distinguish between clusters is the relation of each cluster to the others. In Figure 8b, the probability density functions (PDF) of all clusters are shown as functions of distance to each centroid. Before DEC, though k-means clustering results in the PDF of each class being closest to its centroid, there is significant overlap with other clusters, and the clusters themselves are not particularly dense. After DEC, the PDF of each class is closer to its centroid, denser, and farther removed from the other clusters. Thus, DEC effectively separates each cluster from the others, allowing for better distinction between clusters in the latent space.

The effects of DEC become readily apparent when the latent feature vectors are stacked and sorted according to their distance from each centroid, as shown in Figure 9. By sorting the latent space by sample index i such that $d_{i+1,j} > d_{i,j}$, cluster separation can be visualized directly in the latent space. Before DEC, centroids are initialized with the k-means algorithm without modification to the latent data. Closest to each class centroid, the latent feature vectors are similar in appearance to the centroid, but transition continuously to different patterns as the sorted index i increases. The contrast with the latent space after DEC is stark: because DEC moves latent data assigned to a particular class closer to the centroid, the effect is that the latent feature vectors take on similar values, and therefore appearance, to the centroid. The result is that the latent space appears more sharply segmented after DEC, with the samples closest to the centroid of nearly uniform appearance to the centroid itself. For reference, the relative location of the other class centroids are marked with white vertical lines. Before DEC, the latent feature vectors belonging to the other classes are not readily apparent, whereas after DEC, most of the other centroid locations are associated with their distinctive latent feature vectors.

5.2 DEC Methodology Considerations

One of the key strengths of DEC is its employment of an autoencoder to reduce the dimensionality of the input data to obtain more effective clustering performance. By reducing the dimensionality of the data space, the complexity of the clustering problem is similarly decreased and the distance metrics gain relevance. The ability of the autoencoder to quickly learn the salient features of the data and embed them into the latent space makes the technique adaptable to new data sets. While the autoencoder design choice for this study was sufficiently robust, autoencoder design presents opportunities for further experimentation and improvement. Design variables that could be altered in the DNN architecture include the number of layers, dimensions of the latent feature space, activation function types, incorporation of max-pooling and drop-out layers, and filter size, depth, and stride.

A second key strength of DEC is that clustering improvement and model optimization occur simultaneously. The outcome is denser, more separated clusters. This is a desirable effect in distance-based clustering, but it introduces a vulnerability: the success of the results may depend on the quality of the initial centroids. The k-means algorithm

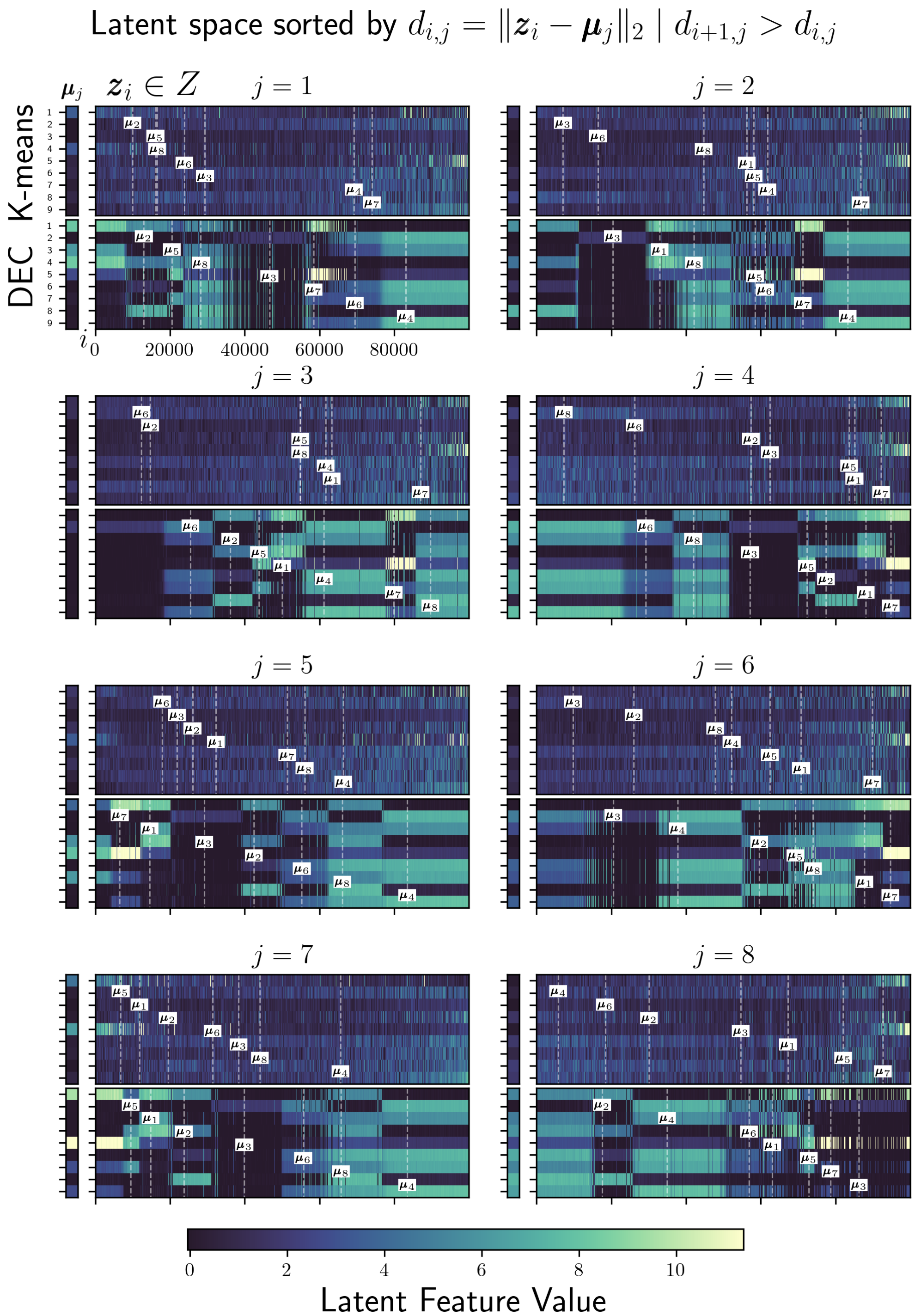

Figure 9. For each class j, latent data samples $\boldsymbol{z}_i$ are shown stacked according to their distance $\|\boldsymbol{z}_i - \boldsymbol{\mu}_j\|$ from the centroid $\boldsymbol{\mu}_j$ (shown to the left). Distance of the other cluster centroids relative to the selected class j are indicated with vertical dotted lines. Deep embedded clustering (DEC) brings assigned data $\boldsymbol{z}_i$ closer to the class centroid, resulting in homogeneity among the latent feature vectors assigned to that class.

assumes that all classes identified have approximately equal sample size and variance, yet we cannot expect naturally occurring seismic sources to exhibit such statistics, particularly given the highly diverse nature of the data in time and space. However, Xie et al. (2016) demonstrated that DEC is robust against imbalances in class sample sizes, including size disparities up to a factor of 10. Further refinement of the centroid initialization step could be accomplished with a clustering algorithm that takes into account differing sample sizes and variances; for example, Gaussian mixture models are capable of characterizing clusters that are elongated and of varying size. Alternatively, an algorithm like k-medians, which can use the "Manhattan" distance $d_{i,j} = \|\boldsymbol{z}_i - \boldsymbol{\mu}_j\|_1$ to optimize centroids, could be used to overcome the inherent challenges of clustering in high-dimensional space. However, we assess that further refinement of centroid initialization is not necessary to achieve a workflow capable of performing satisfactory data exploration.

The flexibility afforded by DEC extends not only to model design, but also to data pre- and post-processing. Whereas model design is largely concerned with *how* the salient features are learned, data pre-processing is concerned with *what* is supplied to the model. This information is dependent on the choice of signal processing parameters, particularly signal duration, filter cutoff frequencies, and seismic event detection algorithm. Additionally, various data transforms commonly used to characterize seismic waveforms can be used as input to DEC (Mousavi et al., 2016). In our case, we used spectrograms, but other transforms, such as continuous wavelet transform scalograms, could just as easily be used as input to the DEC model. In post-processing, redundant or similar results can be combined.

6 Discussion: Glaciological Implications

The full RIS array data set contains 427,798 seismic detections. A summary of the data set statistics and class characteristics (Table 3) shows the total number of detections for each class, as well as the percentage of detections occurring in the austral summers (January, February, and March) versus the austral winters (June, July and August). Classes 1, 4, and 7 have pronounced differences (more than 10%) between the number of detections occurring in the summers versus the winters, while differences for classes 2, 3, 5, and 8 are less pronounced (between 5% and 10%). Class 6 appears to have little difference (less than 5%) between austral summers and winters. Inter-annual com-

Table 3. *Austral Summer (January-February-March) and Winter (June-July-August) Detection Statistics, Average Peak Frequencies, and Amplitude Characteristics for Each Signal Class over the Entire Seismic Array*

	Detections				Amplitude (accel., nm/s^2)			
Class	N	%N Summer (JFM) Total | 2015 | 2016	%N Winter (JJA) Total | 2015 | 2016	Mean Peak Freq (Hz)	Mean	Median	Standard Deviation	Maximum
1	34,919	19 | 9 | 10	31 | 10 | 21	10.5	53	5	400	26,780
2	45,079	31 | 15 | 16	23 | 11 | 12	7.2	150	46	992	69,410
3	78,861	32 | 17 | 15	24 | 14 | 11	5.3	187	38	6,706	1,632,000
4	100,009	19 | 10 | 9	31 | 17 | 13	4.8	13	6	129	23,650
5	18,268	18 | 7 | 11	25 | 6 | 18	14.8	409	15	5,684	461,200
6	55,633	30 | 17 | 13	27 | 16 | 11	4.4	54	8	315	25,166
7	32,276	11 | 7 | 4	43 | 27 | 16	16.1	7	4	22	2,709
8	62,753	23 | 10 | 13	29 | 10 | 19	6.9	18	4	334	41,923

parisons for each season show that classes 1, 2, 5, and 8 experienced an increase in activity in the 2016 austral summer over the 2015 austral summer, with classes 5 and 8 exhibiting the largest changes. Classes 5 and 8 also increase by factors of three and two, respectively, in the 2016 austral winter over the 2015 austral winter. These trends can be investigated in more detail from Figure 10a, where the frequencies of detections shown as a function of station and month exhibit spatiotemporal patterns that may reveal associations between environmental forcing and seismicity. Clustering enables these patterns to be further explored by class and month (Figure 10b), and by class and station (Figure 10c).

From Figure 10a, certain patterns are readily apparent, such as increased seismic detections during the austral summer months at stations DR01, DR02, and DR03. These three stations were located approximately 2 km from the ice front and detected seismicity associated with ocean gravity waves impacting at the shelf front that cause fracturing (icequakes) and calving (Chen et al., 2019). Furthermore, seismicity at these stations during the 2016 austral summer is higher than the same period in 2015. The remaining DR stations and stations RS01 through RS07 exhibit the opposite pattern: austral

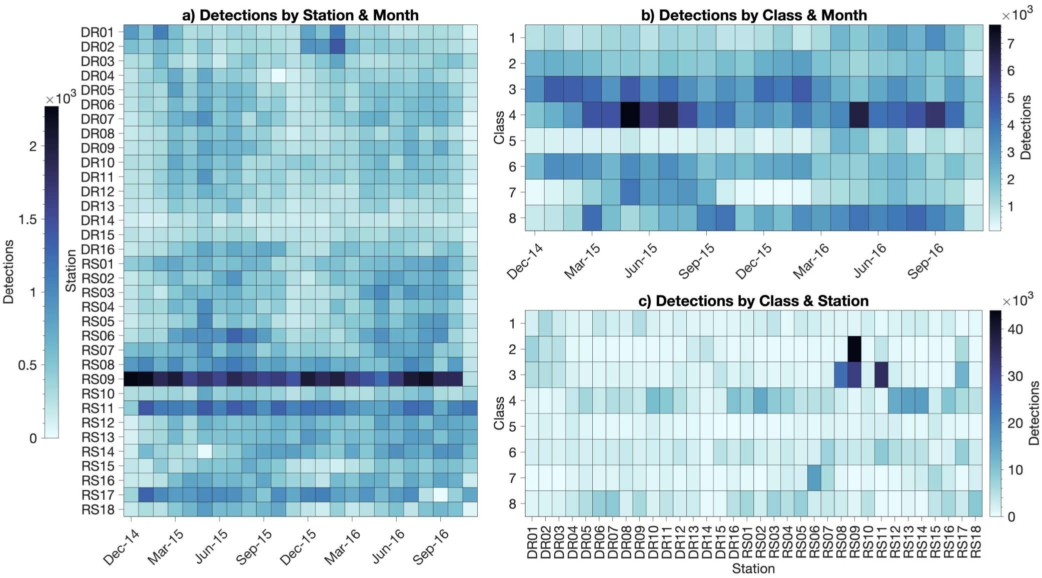

Figure 10. (a) The frequency of detections comprising the Ross Ice Shelf data set is shown by station and month. Deep embedded clustering (DEC) provides a further breakdown by (b) class and month for all stations, and (c) class and station.

summers are relatively quiet, with increased detection frequencies in the austral winters. The four most active stations were located near grounding zones: station RS09 (41,615 detections) on the eastern flank of Roosevelt Island; station RS11 (25,884 detections) on the Shirase Coast; station RS08 (18,655 detections) on the western flank of Roosevelt Island; and station RS17 (18,653 detections) on Steershead Ice Rise. All of these stations exhibited persistent seismicity throughout the two deployment years, with the exception of station RS17, which was offline for several weeks from August to September 2016.

Some classes of signal detections exhibit temporal patterns that are visible in Figure 10b. Classes 1 and 5 have increased detection frequencies in the austral winter of 2016, while classes 2 and 3 have increased detections in the austral summers. Class 6 also exhibits increased detections in the austral summers, though detections in the austral winter of 2015 are also high. Meanwhile, classes 4, 7, and 8 exhibit low seismicity during the austral summers. A further dimension to the analysis is shown in Figure 10c, which shows the distribution of classes by station. The most frequently occurring class in the data set is class 4, which from Figure 6 resembles seismic tremor, and occurs across the array with peak activity in the austral winters. Class 3 is the prominent signal type at stations near grounding zones (RS08, RS09, RS11, and RS17), as is class 2 at RS09.

An important caveat for the detection statistics shown in Table 3 and Figure 10 arises from the physics governing seismic propagation. For a given amplitude, low frequency seismic energy propagates farther than high frequency seismic energy. We thus expect the seismometers in the RIS array to detect low-frequency signals from farther away than high-frequency signals, skewing the total detections in favor of the lower-frequency detections. Factoring in signal amplitude also affects the range at which seismic energy is detected. From Table 3, class 7 has an average spectral peak at 16.1 Hz, the highest of the classes, with a total of 32,276 detections, the second lowest of the classes. Similarly, class 5 has the second-highest average spectral peak at 14.8 Hz, with the least amount of detections among the classes. These two classes are nevertheless distinct from each other in amplitude and waveform type: from Table 3, class 7 has a mean amplitude of 7.0 nm/s^2, while class 5 has a mean amplitude of 408.5 nm/s^2; and from Figure 6, class 7 consists of continuous signals while class 5 consists of high-frequency impulses likely resulting from fracturing. Detection statistics are further affected by signal-to-noise ratios at the seismometers and by limitations of the automated seismic event detector, such as the inability to separate signals from different classes that are received simultaneously.

Though the sources of uncertainty in the detection statistics are nontrivial, with a proper understanding of these limitations and when paired with environmental data, the clustering results can nevertheless be used to analyze the association of potential seismic source mechanisms that may be related to ice shelf dynamics. In the following sections, we provide vignettes using stations DR02 and RS09 to demonstrate the utility of DEC in exploring data and identifying potential causes of seismicity when examined in conjunction with environmental data.

6.1 Seasonal seismicity at the RIS front

Approximately 2 km from the RIS front on Nascent Iceberg, station DR02 exhibits a seasonal pattern of seismicity associated with changes in air temperature and sea ice concentration in the Ross Sea. During the austral summer, sea ice melts and the concentration (Figure 11a) decreases from nearly 100% to approximately 25%, permitting ocean gravity waves to impinge directly on the ice shelf front. Additionally, warmer air

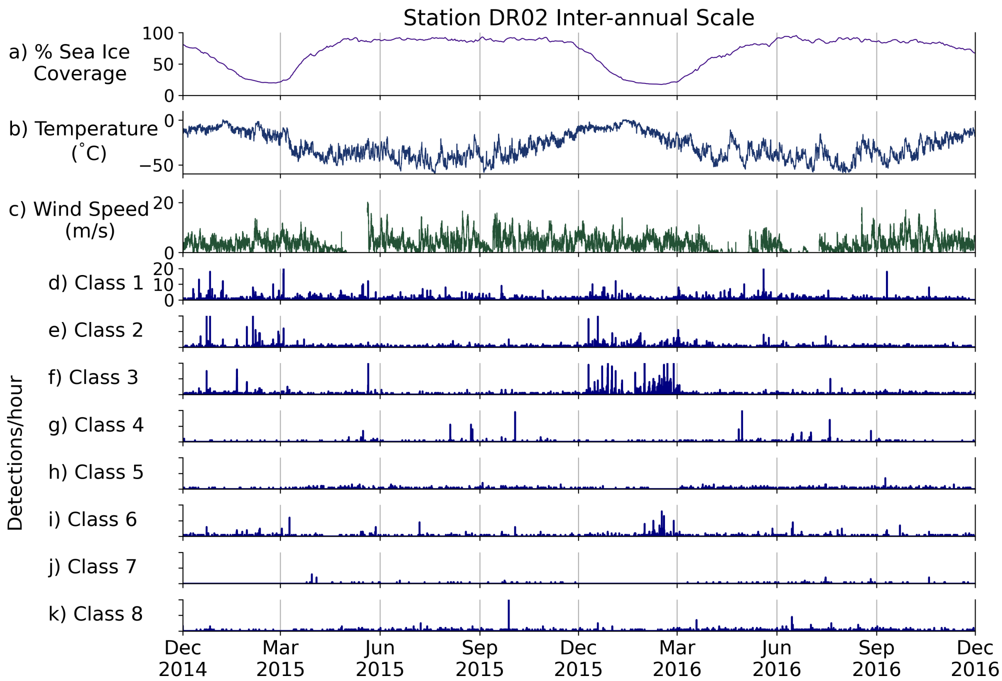

Figure 11. Two years of (a) sea ice coverage on the Ross Sea, (b) temperature and (c) wind speed at Gill automated weather station (approximately 223 km south of DR02), and (d-k) icequake detection statistics for each signal class. Classes 2, 3, and 6 exhibit increased seismicity during the austral summers. Sea ice concentration from NSIDC (Cavalieri et al., 1996, updated yearly); weather station data from AMRC, SSEC, UW–Madison.

temperatures (Figure 11b) accelerate calving and ice shelf structural failure at the front, processes which generate seismic activity (Chen et al., 2019). During the austral winter, sea ice coverage on the Ross Sea reaches nearly 100%, damping higher frequency ocean gravity waves such as swell.

Increased levels of seismicity are observed for classes 2, 3, and 6 at DR02 (Figure 11e,f,i) during the austral summers. Classes 3 and 6 are especially active during the 2016 austral summer, when strong El Niño conditions led to anomalously persistent high temperatures across West Antarctica (Nicolas et al., 2017) and ocean-ice shelf interactions were likely enhanced. Patterns similar to the seismicity at DR02 were observed at stations DR01 and DR03, also located near the RIS front, and can be seen in the total detections by station and month in Figure 10a. Widespread surface melt on the RIS was observed between 10-21 January 2016 (Nicolas et al., 2017; Chaput et al., 2018), which can lead to hydrofracture and contribute to ice shelf disintegration (Hubbard et al., 2016; Alley et al., 2018).

Although class 1 has elevated activity during the summers, it maintains activity throughout the winter months, suggesting that gravity wave activity is not the dominant forcing. The persistence of class 1 signals, which often consist of impulse trains, suggests they may be caused by icequakes resulting from the motion of the ice shelf itself, as the ice flow velocity in the vicinity of station DR02 is among the highest observed on the RIS (Klein et al., 2020). Classes 4, 5, 7, and 8 (Figure 11g,h,j,k) are more active during the coldest periods of the year, suggesting that these signals may be associated with extremely cold temperatures or strong wind events. Cold-weather enhanced seismicity has been documented at a rift approximately 140 km south of the ice front Olinger et al. (2019). Across all classes, discrete instances of high seismicity occur that do not correspond to environmental forcing. Such instances may indicate the occurrence of fracturing ice (icequakes) or events associated with crevasse expansion.

6.2 Diurnal seismicity on Roosevelt Island

Station RS09 on the eastern flank of Roosevelt Island experienced the most detections on the array, comprising 9.7% of detections in the full data set. In Figure 12, potential environmental sources of seismicity are compared to the seismicity of each class. Temperature and wind speed (Figure 12a,b) were recorded at a nearby automated weather station, Margaret, 122 km southwest of RS09. Tides (Figure 12c) were realized from the CATS2008 model (Padman et al., 2002) and calculated at station RS10, which is on floating ice and approximates the tidal signal in the basin between Roosevelt Island and the Shirase Coast. Seismicity for classes 2, 3, and 6 (Figure 12d,e,i) dominate the detection results and are active throughout the year, with classes 2 and 3 comprising 52.8% and 38.0% of the detections, respectively. Classes 1, 4, 5, 7, and 8 (Figure 12d,g,h,j,k) are comparatively sparse, with seismicity limited to what appear to be discrete events that could be associated with large fracture or crevasse events.

Of particular interest at station RS09 is evidence of seismicity corresponding with the diurnal tide. On the inter-annual timescale, class 6 exhibits a periodic modulation of seismicity which tends to correspond with spring tides. To examine the data on a finer

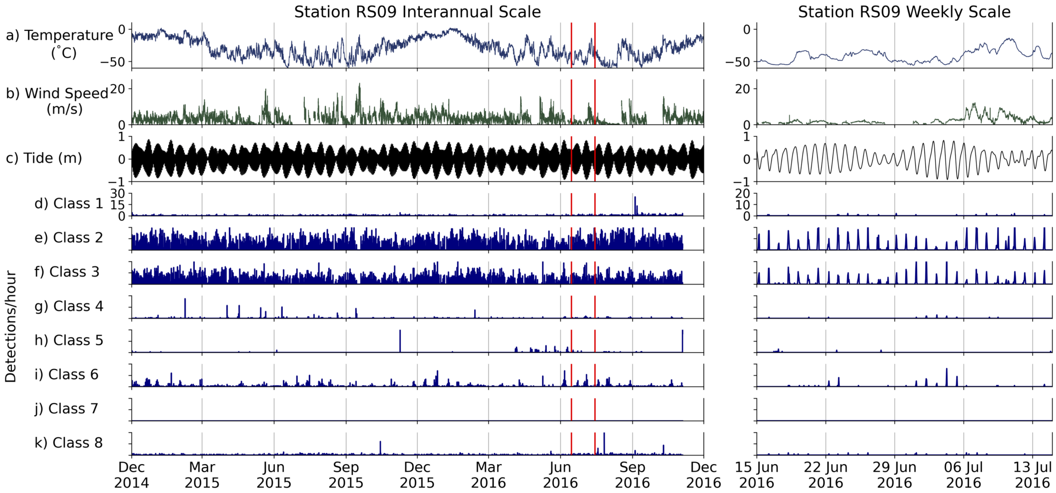

Figure 12. Two years of (a) temperature and (b) wind speed at Margaret automated weather station (MGT, approximately 122 km southwest of RS09, Figure 1), c) model-derived tides calculated at station DR10, and (d-k) icequake detection statistics for each signal class. Interannual timescale is shown at left with vertical red lines indicating the subset weekly time-scale at right. The diurnal tidal signal corresponds to seismicity for classes 2, 3, and 6. Tidal model from (Padman et al., 2002); weather station data from AMRC, SSEC, UW–Madison.

timescale, data are shown between 15 June 2016 and 15 July 2016. This weekly timescale shows that classes 2 and 3, the dominant signal classes, correspond to diurnal tides. Even some relatively non-active classes (1, 4, and 8) show signs of diurnal seismicity. These results are consistent with a previous study that found more than 95% of detections at RS09 were from tidally induced swarms of icequakes that occur throughout the year (Cole, 2020).

Other stations located in grounding zones exhibit similar patterns of seismicity, though to a lesser extent than RS09. Station RS11, located east of RS09 across an expanse of floating ice, exhibits similar patterns of seismicity to RS09. This suggests that ice shelf seismicity at Roosevelt Island and the Shirase Coast grounding zones are associated with similar ice shelf processes. RS08, on the western flank of the RIS, and RS17, at Steershead Ice Rise, also exhibit diurnal seismicity, suggesting a dynamic diurnal process common to the grounding zones and corresponding to the tidal cycle. Class 3 signals are the most common among the four stations in the grounding zone. With a mean peak frequency of 5.3 Hz and a mean amplitude of 187 nm/s^2, these signals are among the strongest detected on the array.

7 Conclusions

Deep embedded clustering (DEC) is an effective way to explore large seismic data sets, particularly in its ability to identify dominant types of seismicity. The results provided by DEC, when contextualized with non-seismic environmental data, can assist in the identification or correlation of seismic source mechanisms, as demonstrated with the RIS environmental data. Additionally, DEC can be readily tailored to investigate different aspects of the same or new data sets. Combined with its effectiveness at clustering seismic detections, this flexibility suggests that DEC can be incorporated into existing seismic workflows in order to speed up exploratory data analysis.

Application of DEC to the Ross Ice Shelf array data set identified eight classes of impulsive signals, with linkage of three of the classes to tidal variability near grounding zones. Additionally, stations near the front showed increased icequake activity during the 2016 El Niño austral summer. The highest seismicity was observed at grounding zones, particularly along the eastern flank of Roosevelt Island.

As seismic data sets grow ever larger, novel machine learning techniques will be necessary to enable researchers to fully utilize this data. DEC has the potential to become an important tool for exploring these large data sets, and to complement other machine learning-based tools as well as conventional signal processing approaches. The incorporation of such tools will enable more thorough and timely geophysical data analysis, thus improving the response of geophysical research to the needs of society in a rapidly changing earth.

Acknowledgments

This work was funded by the Office of Naval Research through the National Defense Science and Engineering Graduate Fellowship Program, and by National Science Foundation grant PLR 1246151. Seismic data from network XH (D. Wiens & Bromirski, 2014) were downloaded through IRIS Web Services (`https://service.iris.edu/irisws/`). Seismic data were processed using Obspy software (Beyreuther et al., 2010). Figures were generated in MATLAB (`https://www.mathworks.com`) and with Matplotlib (`https://`

matplotlib.org). The DEC model was produced using PyTorch (https://pytorch.org). Antarctica elevation data, grounding line, and coast line were obtained from Bedmachine (Morlighem et al., 2017) and plotted using Antarctic Mapping Tools for MATLAB (Greene et al., 2017). Surface temperatures were obtained from AMRC, SSEC, University of Wisconsin–Madison (https://amrc.ssec.wisc.edu). Tide data were generated by the CATS2008 model (Padman et al., 2002). Ross Sea ice coverage was obtained from NASA NSIDC (Cavalieri et al., 1996, updated yearly). Code for this workflow is available at https://github.com/NeptuneProjects/RISClusterPT.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In G. Goos, J. Hartmanis, J. van Leeuwen, J. Van den Bussche, & V. Vianu (Eds.), *Database Theory — ICDT 2001* (Vol. 1973, pp. 420–434). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/3-540-44503-X_27

Aggarwal, C. C., & Reddy, C. K. (Eds.). (2014). *Data clustering: Algorithms and applications*. Boca Raton: Chapman and Hall/CRC.

Alley, K., Scambos, T., Miller, J., Long, D., & MacFerrin, M. (2018, June). Quantifying vulnerability of Antarctic ice shelves to hydrofracture using microwave scattering properties. *Remote Sensing of Environment*, *210*, 297–306. doi: 10.1016/j.rse.2018.03.025

Arthur, D., & Vassilvitskii, S. (2007). K-Means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035). USA: Society for Industrial and Applied Mathematics.

Aster, R. C., & Winberry, J. P. (2017, December). Glacial seismology. *Reports on Progress in Physics*, *80*(12), 126801. doi: 10.1088/1361-6633/aa8473

Baker, M. G., Aster, R. C., Anthony, R. E., Chaput, J., Wiens, D. A., Nyblade, A., ... Stephen, R. A. (2019, December). Seasonal and spatial variations in the ocean-coupled ambient wavefield of the Ross Ice Shelf. *Journal of Glaciology*, *65*(254), 912–925. doi: 10.1017/jog.2019.64

Baker, M. G., Aster, R. C., Wiens, D. A., Nyblade, A., Bromirski, P. D., Gerstoft, P., & Stephen, R. A. (2020, October). Teleseismic earthquake wavefields observed on the Ross Ice Shelf. *Journal of Glaciology*, 1–17. doi: 10.1017/jog.2020.83

Barcheck, C. G., Tulaczyk, S., Schwartz, S. Y., Walter, J. I., & Winberry, J. P. (2018, March). Implications of basal micro-earthquakes and tremor for ice stream mechanics: Stick-slip basal sliding and till erosion. *Earth and Planetary Science Letters*, *486*, 54–60. doi: 10.1016/j.epsl.2017.12.046

Beaucé, E., Frank, W. B., & Romanenko, A. (2018, January). Fast Matched Filter (FMF): An Efficient Seismic Matched-Filter Search for Both CPU and GPU Architectures. *Seismological Research Letters*, *89*(1), 165–172. doi: 10.1785/0220170181

Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Rand Corporation.

Bergen, K. J., & Beroza, G. C. (2018, June). Detecting earthquakes over a seismic network using single-station similarity measures. *Geophysical Journal International*, *213*(3), 1984–1998. doi: 10.1093/gji/ggy100

Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., & Wassermann, J. (2010, May). ObsPy: A Python Toolbox for Seismology. *Seismological Research Letters*, *81*(3), 530–533. doi: 10.1785/gssrl.81.3.530

Bianco, M. J., & Gerstoft, P. (2018, December). Travel Time Tomography With Adaptive Dictionaries. *IEEE Transactions on Computational Imaging*, *4*(4), 499–511. doi: 10.1109/TCI.2018.2862644

Bianco, M. J., Gerstoft, P., Olsen, K. B., & Lin, F.-C. (2019, December). High-resolution seismic tomography of Long Beach, CA using machine learning. *Scientific Reports*, *9*(1), 14987. doi: 10.1038/s41598-019-50381-z

Bindschadler, R. A., King, M. A., Alley, R. B., Anandakrishnan, S., & Padman, L. (2003, August). Tidally Controlled Stick-Slip Discharge of a West Antarctic Ice Stream. *Science*, *301*(5636), 1087–1089. doi: 10.1126/science.1087231

Bindschadler, R. A., Vornberger, P. L., King, M. A., & Padman, L. (2003). Tidally driven stick–slip motion in the mouth of Whillans Ice Stream, Antarctica. *Annals of Glaciology*, *36*, 263–272. doi: 10.3189/172756403781816284

Bishop, C. (2006). *Pattern Recognition and Machine Learning* (First ed.). Springer-Verlag New York.

Bromirski, P. D., Chen, Z., Stephen, R. A., Gerstoft, P., Arcas, D., Diez, A., ... Nyblade, A. (2017, July). Tsunami and infragravity waves impacting A ntarctic ice shelves. *Journal of Geophysical Research: Oceans*, *122*(7), 5786–5801. doi: 10.1002/2017JC012913

Bromirski, P. D., Diez, A., Gerstoft, P., Stephen, R. A., Bolmer, T., Wiens, D. A., … Nyblade, A. (2015, September). Ross ice shelf vibrations. *Geophysical Research Letters*, *42*(18), 7589–7597. doi: 10.1002/2015GL065284

Bromirski, P. D., & Stephen, R. A. (2012). Response of the Ross Ice Shelf, Antarctica, to ocean gravity-wave forcing. *Annals of Glaciology*, *53*(60), 163–172. doi: 10.3189/2012AoG60A058

Cavalieri, D. J., Parkinson, C. L., Gloersen, P., & Zwally, H. J. (1996, updated yearly). *Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data, Version 1.* Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center.

Chamarczuk, M., Nishitsuji, Y., Malinowski, M., & Draganov, D. (2020, January). Unsupervised Learning Used in Automatic Detection and Classification of Ambient-Noise Recordings from a Large-N Array. *Seismological Research Letters*, *91*(1), 370–389. doi: 10.1785/0220190063

Chamberlain, C. J., Hopp, C. J., Boese, C. M., Warren-Smith, E., Chambers, D., Chu, S. X., … Townend, J. (2018, January). EQcorrscan: Repeating and Near-Repeating Earthquake Detection and Analysis in Python. *Seismological Research Letters*, *89*(1), 173–181. doi: 10.1785/0220170151

Chaput, J., Aster, R. C., McGrath, D., Baker, M., Anthony, R. E., Gerstoft, P., … Stevens, L. A. (2018, October). Near-Surface Environmentally Forced Changes in the Ross Ice Shelf Observed With Ambient Seismic Noise. *Geophysical Research Letters*, *45*(20). doi: 10.1029/2018GL079665

Chazan, S. E., Gannot, S., & Goldberger, J. (2019, March). Deep Clustering Based on a Mixture of Autoencoders. *arXiv:1812.06535 [cs, stat]*.

Chen, Z., Bromirski, P. D., Gerstoft, P., Stephen, R. A., Lee, W. S., Yun, S., … Nyblade, A. A. (2019, August). Ross Ice Shelf Icequakes Associated With Ocean Gravity Wave Activity. *Geophysical Research Letters*, *46*(15), 8893–8902. doi: 10.1029/2019GL084123

Chen, Z., Bromirski, P. D., Gerstoft, P., Stephen, R. A., Wiens, D. A., Aster, R. C., & Nyblade, A. A. (2018, October). Ocean-excited plate waves in the Ross and

Pine Island Glacier ice shelves. *Journal of Glaciology*, *64*(247), 730–744. doi: 10.1017/jog.2018.66

Cole, H. M. (2020). *Tidally Induced Seismicity at the Grounded Margins of the Ross Ice Shelf, Antarctica* (Master's Thesis). Colorado State University, Fort Collins, Colorado.

De Angelis, H., & Skvarca, P. (2003, March). Glacier Surge After Ice Shelf Collapse. *Science*, *299*(5612), 1560–1562. doi: 10.1126/science.1077987

Diez, A., Bromirski, P., Gerstoft, P., Stephen, R., Anthony, R., Aster, R., ... Wiens, D. (2016, May). Ice shelf structure derived from dispersion curve analysis of ambient seismic noise, Ross Ice Shelf, Antarctica. *Geophysical Journal International*, *205*(2), 785–795. doi: 10.1093/gji/ggw036

Dupont, T. K., & Alley, R. B. (2005). Assessment of the importance of ice-shelf buttressing to ice-sheet flow. *Geophysical Research Letters*, *32*(4). doi: 10.1029/2004GL022024

Fürst, J. J., Durand, G., Gillet-Chaulet, F., Tavard, L., Rankl, M., Braun, M., & Gagliardini, O. (2016, May). The safety band of Antarctic ice shelves. *Nature Climate Change*, *6*(5), 479–482. doi: 10.1038/nclimate2912

Gibbons, S. J., & Ringdal, F. (2006, April). The detection of low magnitude seismic events using array-based waveform correlation. *Geophysical Journal International*, *165*(1), 149–166. doi: 10.1111/j.1365-246X.2006.02865.x

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Greene, C. A., Gwyther, D. E., & Blankenship, D. D. (2017, July). Antarctic Mapping Tools for Matlab. *Computers & Geosciences*, *104*, 151–157. doi: 10.1016/j.cageo.2016.08.003

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. doi: 10.2307/2346830

Hell, M. C., Cornelle, B. D., Gille, S. T., Miller, A. J., & Bromirski, P. D. (2019, November). Identifying Ocean Swell Generation Events from Ross Ice Shelf Seismic Data. *Journal of Atmospheric and Oceanic Technology*, *36*(11), 2171–2189. doi: 10.1175/JTECH-D-19-0093.1

Hinton, G. E. (2006, July). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507. doi: 10.1126/science.1127647

Holtzman, B. K., Paté, A., Paisley, J., Waldhauser, F., & Repetto, D. (2018, May). Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field. *Science Advances*, *4*(5), eaao2929. doi: 10.1126/sciadv.aao2929

Hotovec-Ellis, A. J., & Jeffries, C. (2016, April). *Near Real-time Detection, Clustering, and Analysis of Repeating Earthquakes: Application to Mount St. Helens and Redoubt Volcanoes* [Invited]. Reno, NV, USA.

Hubbard, B., Luckman, A., Ashmore, D. W., Bevan, S., Kulessa, B., Kuipers Munneke, P., . . . Rutt, I. (2016, September). Massive subsurface ice formed by refreezing of ice-shelf melt ponds. *Nature Communications*, *7*(1), 11897. doi: 10.1038/ncomms11897

Johnson, C. W., Ben-Zion, Y., Meng, H., & Vernon, F. (2020, August). Identifying Different Classes of Seismic Noise Signals Using Unsupervised Learning. *Geophysical Research Letters*, *47*(15). doi: 10.1029/2020GL088353

Johnson, C. W., Meng, H., Vernon, F., & Ben-Zion, Y. (2019, August). Characteristics of Ground Motion Generated by Wind Interaction With Trees, Structures, and Other Surface Obstacles. *Journal of Geophysical Research: Solid Earth*, *124*(8), 8519–8539. doi: 10.1029/2018JB017151

Kingma, D. P., & Ba, J. (2017, January). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*.

Klein, E., Mosbeux, C., Bromirski, P. D., Padman, L., Bock, Y., Springer, S. R., & Fricker, H. A. (2020, October). Annual cycle in flow of Ross Ice Shelf, Antarctica: Contribution of variable basal melting. *Journal of Glaciology*, *66*(259), 861–875. doi: 10.1017/jog.2020.61

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2019, January). Machine Learning in Seismology: Turning Data into Insights. *Seismological Research Letters*, *90*(1), 3–14. doi: 10.1785/0220180259

Kullback, S., & Leibler, R. A. (1951, March). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. doi: 10.1214/aoms/1177729694

MacAyeal, D. R., Banwell, A. F., Okal, E. A., Lin, J., Willis, I. C., Goodsell, B., & MacDonald, G. J. (2019, September). Diurnal seismicity cycle linked to subsurface melting on an ice shelf. *Annals of Glaciology*, *60*(79), 137–157. doi: 10.1017/aog.2018.29

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 1: Statistics* (pp. 281–297). Berkeley, Calif.: University of California Press.

Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., & Long, J. (2018). A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture. *IEEE Access*, *6*, 39501–39514. doi: 10.1109/ACCESS.2018.2855437

Morlighem, M., Williams, C. N., Rignot, E., An, L., Arndt, J. E., Bamber, J. L., … Zinglersen, K. B. (2017, November). BedMachine v3: Complete Bed Topography and Ocean Bathymetry Mapping of Greenland From Multibeam Echo Sounding Combined With Mass Conservation. *Geophysical Research Letters*, *44*(21). doi: 10.1002/2017GL074954

Mousavi, S. M., Horton, S. P., Langston, C. A., & Samei, B. (2016, October). Seismic features and automatic discrimination of deep and shallow induced-microearthquakes using neural network and logistic regression. *Geophysical Journal International*, *207*(1), 29–46. doi: 10.1093/gji/ggw258

Mousavi, S. M., Zhu, W., Ellsworth, W., & Beroza, G. (2019, November). Unsupervised Clustering of Seismic Signals Using Deep Convolutional Autoencoders. *IEEE Geoscience and Remote Sensing Letters*, *16*(11), 1693–1697. doi: 10.1109/LGRS.2019.2909218

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective.* Cambridge, MA: MIT Press.

Nicolas, J. P., Vogelmann, A. M., Scott, R. C., Wilson, A. B., Cadeddu, M. P., Bromwich, D. H., … Wille, J. D. (2017, August). January 2016 extensive summer melt in West Antarctica favoured by strong El Niño. *Nature Communications*, *8*(1), 15799. doi: 10.1038/ncomms15799

Olinger, S. D., Lipovsky, B. P., Wiens, D. A., Aster, R. C., Bromirski, P. D., Chen, Z., … Stephen, R. A. (2019, June). Tidal and Thermal Stresses Drive Seismicity Along a Major Ross Ice Shelf Rift. *Geophysical Research Letters*, *46*(12), 6644–6652. doi: 10.1029/2019GL082842

Padman, L., Fricker, H. A., Coleman, R., Howard, S., & Erofeeva, L. (2002). A new tide model for the Antarctic ice shelves and seas. *Annals of Glaciology*, *34*, 247–254. doi: 10.3189/172756402781817752

Paolo, F. S., Fricker, H. A., & Padman, L. (2015). Volume loss from Antarctic ice shelves is accelerating. *Science*, *348*(6232), 327–331. doi: 10.1126/science.aaa0940

Perol, T., Gharbi, M., & Denolle, M. (2018, February). Convolutional neural network for earthquake detection and location. *Science Advances*, *4*(2), e1700578. doi: 10.1126/sciadv.1700578

Pritchard, H. D., Ligtenberg, S. R. M., Fricker, H. A., Vaughan, D. G., van den Broeke, M. R., & Padman, L. (2012, April). Antarctic ice-sheet loss driven by basal melting of ice shelves. *Nature*, *484*(7395), 502–505. doi: 10.1038/nature10968

Reddy, T. A., Devi, K. R., & Gangashetty, S. V. (2012, March). Nonlinear principal component analysis for seismic data compression. In *2012 1st International Conference on Recent Advances in Information Technology (RAIT)* (pp. 927–932). Dhanbad, India: IEEE. doi: 10.1109/RAIT.2012.6194558

Riahi, N., & Gerstoft, P. (2017, March). Using graph clustering to locate sources within a dense sensor array. *Signal Processing*, *132*, 110–120. doi: 10.1016/j.sigpro.2016.10.001

Rignot, E., Mouginot, J., Morlighem, M., Seroussi, H., & Scheuchl, B. (2014). Widespread, rapid grounding line retreat of Pine Island, Thwaites, Smith, and Kohler glaciers, West Antarctica, from 1992 to 2011. *Geophysical Research Letters*, *41*(10), 3502–3509. doi: 10.1002/2014GL060140

Rousseeuw, P. J. (1987, November). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. doi: 10.1016/0377-0427(87)90125-7

Scambos, T. A., Bohlander, J. A., Shuman, C. A., & Skvarca, P. (2004). Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica. *Geophysical Research Letters*, *31*(18). doi: 10.1029/2004GL020670

Seydoux, L., Balestriero, R., Poli, P., de Hoop, M., Campillo, M., & Baraniuk, R. (2020, December). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature Communications*, *11*(1), 3972. doi: 10.1038/s41467-020-17841-x

Smith, B., Fricker, H. A., Gardner, A. S., Medley, B., Nilsson, J., Paolo, F. S., . . .

Zwally, H. J. (2020, June). Pervasive ice sheet mass loss reflects competing ocean and atmosphere processes. *Science*, *368*(6496), 1239–1242. doi: 10.1126/science.aaz5845

Snover, D., Johnson, C. W., Bianco, M. J., & Gerstoft, P. (2020). Deep clustering to identify sources of urban seismic noise in Long Beach, CA. *Geophysical Research Letters*, 33.

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The Challenges of Clustering High Dimensional Data. In L. T. Wille (Ed.), *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition* (pp. 273–309). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-662-08968-2_16

Swindell, H., & Snell, N. (1977). *Station processor automatic signal detection system, phase I: Final Report, Station Processor Software Development, Texas Instruments Report No. ALEX (01)-FR-77-01.* (Tech. Rep.). Texas Instruments.

Telesca, L., & Chelidze, T. (2018, November). Visibility Graph Analysis of Seismicity around Enguri High Arch Dam, Caucasus. *Bulletin of the Seismological Society of America*, *108*(5B), 3141–3147. doi: 10.1785/0120170370

Thoma, M., Jenkins, A., Holland, D., & Jacobs, S. (2008). Modelling circumpolar deep water intrusions on the amundsen sea continental shelf, antarctica. *Geophysical Research Letters*, *35*(18). doi: 10.1029/2008GL034939

Tibshirani, R., Walther, G., & Hastie, T. (2001, May). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. doi: 10.1111/1467-9868.00293

Trugman, D. T., & Shearer, P. M. (2017, March). GrowClust: A Hierarchical Clustering Algorithm for Relative Earthquake Relocation, with Application to the Spanish Springs and Sheldon, Nevada, Earthquake Sequences. *Seismological Research Letters*, *88*(2A), 379–391. doi: 10.1785/0220160188

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010, December). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning*

Research, 38.

Wiens, D., & Bromirski, P. (2014). *Collaborative Research: Dynamic Response of the Ross Ice Shelf to Wave-Induced Vibrations, and Collaborative Research: Mantle Structure and Dynamics of the Ross Sea from a Passive Seismic Deployment on the Ross Ice Shelf.* International Federation of Digital Seismograph Networks.

Wiens, D. A., Anandakrishnan, S., Winberry, J. P., & King, M. A. (2008, June). Simultaneous teleseismic and geodetic observations of the stick–slip motion of an Antarctic ice stream. *Nature*, *453*(7196), 770–774. doi: 10.1038/nature06990

Withers, M., Aster, R., Young, C., Beiriger, J., Harris, M., Moore, S., & Trujillo, J. (1998, February). A Comparison of Select Trigger Algorithms for Automated Global Seismic Phase and Event Detection. *Bulletin of the Seismological Society of America*, *88*(1), 12.

Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised Deep Embedding for Clustering Analysis. *Proceedings of the 33rd international conference on machine learning*, 10.

Yang, B., Fu, X., Sidiropoulos, N. D., & Hong, M. (2017, June). Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering. *arXiv:1610.04794 [cs]*.

Yoon, C. E., O'Reilly, O., Bergen, K. J., & Beroza, G. C. (2015, December). Earthquake detection through computationally efficient similarity search. *Science Advances*, *1*(11), e1501057. doi: 10.1126/sciadv.1501057