CWP-1027

Errors incurred in lossy compression of seismic data

Abdul Hafiz S Issah[1,2] & Eileen R Martin[1,2,3,4]
[1] *Center for Wave Phenomena,* [2] *Department of Applied Math and Statistics,* [3] *Department of Geophysics,* [4] *Hydrologic Science and Engineering Program*
Colorado School of Mines, Golden CO 80401
email: aissah@mines.edu

ABSTRACT
New technologies such as low-cost nodes and distributed acoustic sensing are making it easier to continuously collect broadband, high-density seismic monitoring data. To reduce the time to move data from the field to computing centers, reduce archival requirements, and speed up interactive data analysis and visualization, we are motivated to investigate the use of lossy compression on seismic array data. In particular, there is a need to not just quantify the errors in the raw data, but also the characteristics of the spectra of these errors, and the extent to which these errors propagate into detectability and pick times of microseismic events. We compare three types of lossy compression: sparse thresholded wavelet compression, zfp compression and low-rank singular value decomposition compression. We apply these techniques to compare compression schemes on two publicly available datasets: an urban dark fiber DAS experiment, and a surface DAS array above a geothermal field. We find that depending on the level of compression needed, and the importance of preserving large versus small seismic events, different compression schemes are preferable.

Key words: data compression, wavelets, zfp, SVD, efficiency, storage

1 INTRODUCTION

Innovations in instrumentation have given rise to a variety of ways of gathering seismic data. One such innovation is Distributed Acoustic Sensing (DAS) which allows the recording of high-frequency data at close sensor spacing. An example was recorded at PoroTomo Natural Laboratory. This dataset was recorded with an 8700m horizontal cable and a 363 m vertical cable in a borehole at 1.021m spacing deployed to study the occurrence of seismic activities around Brady's Hot Spring geothermal site and produced about 1GB of data every 30 seconds totaling several terabytes (Coleman, 2016). This experiment along with eight others produced about 750TB of DAS data from 2015 to 2020, which is a much faster rate of data acquisition than traditional or nodal arrays and inhibits scientists' ability to quickly access, analyze and visualize these new data sources (Lindsey and Martin, 2021). This highlights the need to quantitatively compare the various options for reducing data storage and data movement during processing. A potential solution would be to compress seismic data. Compression of seismic data can be achieved by the transformation of data into a sparse representation so that fewer data points are needed to capture the important parts of the data. Efforts in seismic data compression have been skewed toward active seismic data and have been mainly centered around wavelet decomposition and the discrete cosine transform in the past (Averbuch et al., 2001), (Bosman and Reiter, 1993). These methods provide lossy compression, that is, they introduce some noise in the data at significant compression rates(Donoho et al., 1999). More recently, the advancement of machine learning has given rise to more efforts toward using autoencoders for the compression of active seismic data(Valentine and Trampert, 2012).

In this paper, we provide a preliminary assessment of the suitability of three methods - wavelet decomposition, low-rank approximation, and zfp floating point compression - for compressing seismic monitoring data. We choose these methods based on the ease of providing analytical error bounds for the errors introduced at various levels of compression. We give a brief overview of these methods, their comparison in different metrics for assessing the integrity of reconstructed data, and the effect of compression on event detection.

2 THEORETICAL AND COMPUTATIONAL BACKGROUND

Here, we provide an introduction to three types of compression which are frequently applied to reduce the size of spatiotemporal scientific data. ****Hafiz, update this text****

WAVELET COMPRESSION

Given a time domain signal, f(t) and using wavelet with mother wavelet, ψ and father wavelet, ϕ , the discrete wavelet transform can be computed as

$$f = \sum_{j=L+1}^{J} \sum_{n=0}^{2^{-J}-1} d_j[n]\psi_{j,n} + \sum_{n=0}^{2^{-J}-1} a_j[n]\phi_{J,n} \tag{1}$$

Where the inner products, $d_j[n] = \langle f, \psi_{j,n} \rangle$ and $a_j[n] = \langle f, \phi_{J,n} \rangle$ are the detail and approximation wavelet coefficients respectively; $\psi_{j,n}$ and $\phi_{j,n}$ are scaled and translated versions of the mother and father wavelets respectively and are defined as

$$\psi_{j,n}(x) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{x-n}{2^j}\right) \qquad \text{and} \qquad \phi_{j,n}(x) = \frac{1}{\sqrt{2^j}}\phi\left(\frac{x-n}{2^j}\right)$$

This representation is sparse with isolated cones of high wavelet coefficient amplitude around and leading to areas of high amplitude in time (Mallat, 2009). This property allows the application of wavelet transform for de-noising (Donoho, 1995) and allows the preservation of events when wavelet transform is used for compression Villasenor et al. (1996). We compress the data by doing a soft-thresholding of the wavelet coefficients. For a pre-determined threshold, T, we can construct an approximation of our signal

$$\tilde{f} = \sum_{j=L+1}^{J} \sum_{n=0}^{2^{-J}-1} d_j[n]\psi_{j,n} + \sum_{n=0}^{2^{-J}-1} a_j[n]\phi_{J,n} - \left(\sum_{j=L+1}^{J} \sum_{n=0}^{2^{-J}-1} \tilde{d}_j[n]\psi_{j,n} + \sum_{n=0}^{2^{-J}-1} \tilde{a}_j[n]\phi_{J,n} \right) \tag{2}$$

$$\tilde{a}_j[n] = \begin{cases} a_j[n], & a_j[n] \leq T \\ T, & a_j[n] > T \end{cases} \qquad \tilde{d}_j[n] = \begin{cases} d_j[n], & d_j[n] \leq T \\ T, & d_j[n] > T \end{cases}$$

The error in approximation, E is then

$$E = \sum_{j=L+1}^{J} \sum_{n=0}^{2^{-J}-1} \tilde{d}_j[n]\psi_{j,n} + \sum_{n=0}^{2^{-J}-1} \tilde{a}_j[n]\phi_{J,n} \tag{3}$$

If a big enough T is chosen, the wavelet representation of this approximation has few non-zero coefficients that can be encoded to achieve compression. Hence the threshold determines the amount of compression and the associated error in the approximation.

To exploit redundancies present in two-dimensional datasets, wavelet compression can be performed using two-dimensional wavelet decomposition (Villasenor et al., 1996). In this approach, wavelets derived from one-dimensional wavelets are employed to decompose the data in both dimensions. This technique allows for the efficient representation and compression of the data, taking advantage of redundancies in both dimensions.

Zfp Floating point compression

Zfp uses processes such as block transforms and embedded coding, commonly used in image compression, to perform compression that abides by the more stringent requirements of floating point scientific data (Lindstrom, 2014). This process is outlined in detail in Lindstrom (2014), so we briefly outline the process here. For d-dimensional data, the data is sectioned into blocks of 4^d values, which are assumed to be approximately continuous within any block. Each block is compressed separately via the following steps, and each step may introduce some error:

(i) Convert floating point values in the 4^d block to scaled integers with a common exponent.

(ii) Perform a block transform to introduce some sparsity into the integer representation.

(iii) Encode the numbers in the sparse representation only; This encoding takes sparsity into consideration and only uses as many bits as required to save non-zero entries. The process also allows for specified bits to allocate for each block (fixed-rate), the number of binary exponents to encode (fixed precision), and the maximum error allowed for each floating point value (fixed accuracy).

Truncated SVD Compression

Singular Value Decomposition (SVD) can be used to decompose data represented in matrix form into three matrices; matrices of left singular vectors, right singular vectors, and singular values, where the singular values represent the amount of variation in the data explained by each singular vector. SVD has been widely used in various scientific fields (e.g. signal processing, image compression, data mining, and machine learning) due to its ability to identify and capture important patterns and structures that capture the highest possible amount of variability in the data. The application of SVD to seismic data matrices obtained by organizing data collected through Distributed Acoustic Sensing (DAS) as channels by samples offers a unique opportunity to decouple the channels and samples and enable efficient processing (Martin, 2019).

Although constructing the full SVD can be expensive, randomized SVD is a powerful and efficient algorithm for computing partial SVD of large-scale matrices by using randomized projection methods to quickly approximate the dominant singular vectors. One of the primary benefits of randomized SVD is its ability efficiently decompose large matrices storing high-dimensional data (Halko et al., 2011). This makes it an attractive tool for a wide range of scientific applications. Given data stored a matrix $D \in \mathbb{R}^{N_c \times N_t}$, where N_c is the number of channels, r is the rank, and N_t is the number of time samples, to achieve compression we:

(i) Construct a low-rank (rank, r) approximation, $D_r = U\Sigma V$ using randomized SVD. Where $U \in \mathbb{R}^{N_c \times r}$, $\Sigma \in \mathbb{R}^{r \times r}$, $V \in \mathbb{R}^{r \times N_t}$

(ii) Combine U and Σ into a $(N_c \times r)$ matrix.

(iii) Storing $(N_c \times r)$ and $(r \times N_t)$ matrices provides a compression factor of $\frac{N_c * N_t}{r*(N_t+N_c)}$

3 DATA USED FOR TESTING

We use two publicly available DAS datasets for testing and comparing these compression techniques' effect on the characteristics of compressed data and the errors incurred in event detection workflows. In particular, we use one urban dark fiber dataset, the FORESEE data, and one geothermal monitoring dataset from the Brady Hot Springs site. Here, we provide a brief overview of both datasets.

THE FORESEE DATA

For the norm of noise and frequency preservation experiments, we used data from the FORESEE-urban project which was continuously recorded between April 2019 and Oct 2021 (Zhu et al., 2021). The data examples in Figure 1 illustrate two instances from the dataset. The top image shows some recordings of thunder quakes captured array-wide at around 10 seconds and from about 40 to 50 seconds. In the bottom image, noise is observed between approximately 2000 m to 2500 m and around 3000 m. This noise may potentially be attributed to the presence of vehicle or foot traffic. This dataset contains recordings of signals like these and others from natural and anthropogenic sources (Zhu et al., 2021).

The Brady Hot Springs Data

The data used in our computational experiments for event detection was recorded in 2016 at the Brady Hot Springs site in Nevada as part of an investigation into the feasibility of using Distributed Acoustic Sensing (DAS) for cost-effective monitoring of geothermal reservoirs. The data consists of approximately 8km of fiber optic cable deployed horizontally in a shallow trench, with 1m channel spacing resulting in around 8000 channels recording at 1000 samples per second (Coleman, 2016). Figure 2 shows the map of the Brady Hot Springs geothermal field's horizontal DAS channels, color-coded by channel number and an example of continuously recorded by this array. The dataset recorded microseismic activities that have been studied and cataloged((Li and Zhan, 2018)). We take advantage of the catalog to identify events and compare detectability in compressed data.

4 COMPUTATIONAL EXPERIMENTS

When seismic data is reduced using lossy compression techniques, errors are expected to occur in the compressed data. These errors can significantly affect the quality of the data and the results obtained from seismic processing workflows. These errors can cause distortions in the seismic waveforms and may propagate through the seismic processing workflow, affecting subsequent steps such as imaging and inversion, which can lead to errors in the estimation of seismic properties, imaging, and inversion. Therefore, it is crucial to carefully evaluate the effects of lossy compression on seismic data and the results obtained from processing workflows.

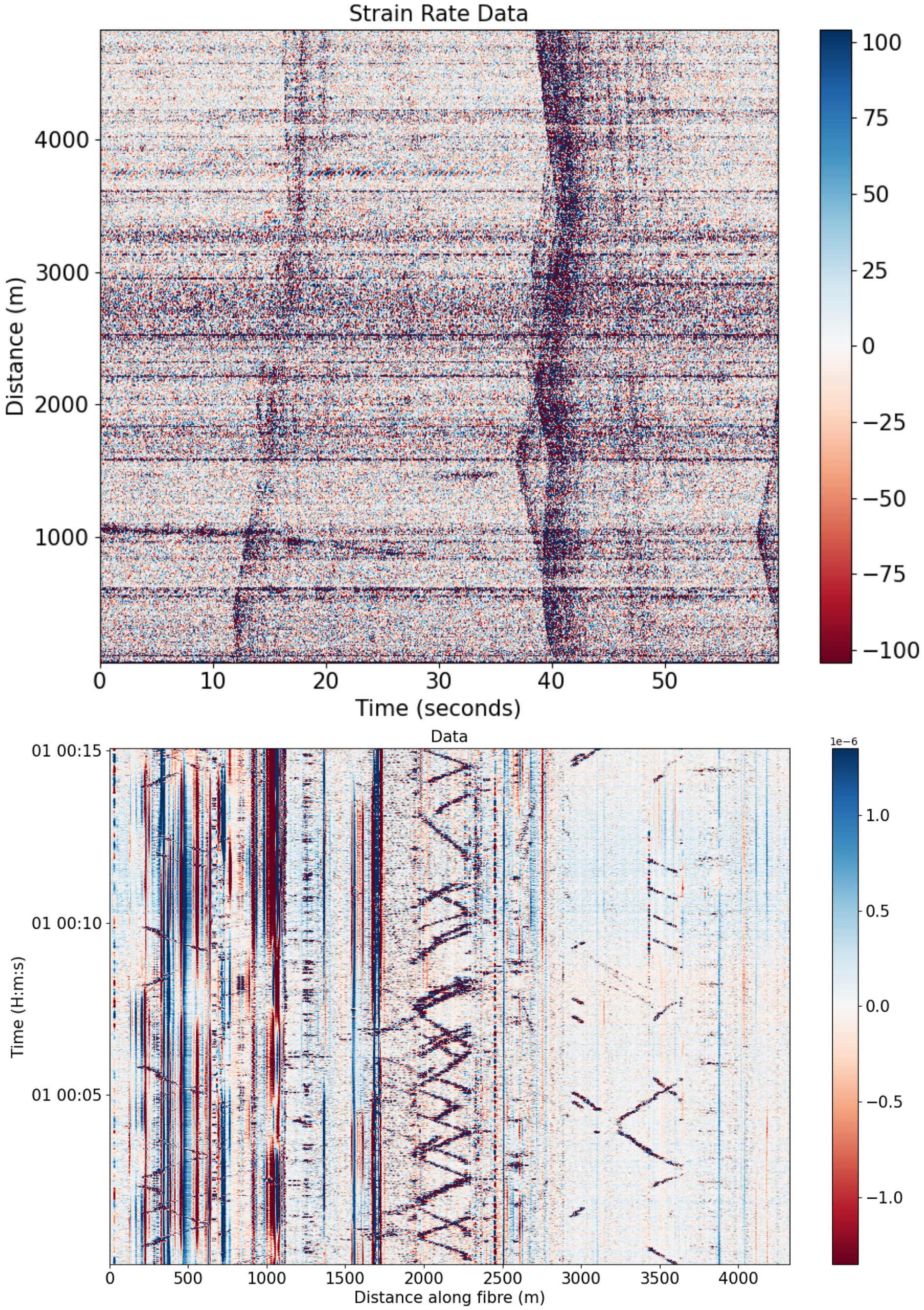

Figure 1. Data from the FORESEE-urban project. (Top) An example of 60 seconds of continuous data recorded at 03:33:35 on 04/15/2019. (Bottom) 20-minute data used for frequency preservation experiment recorded from UTC 00:00:04 to 00:20:04 on 08/01/2019

Our experiments were designed to explore how the errors from varying compression types and ratios propagate into (i) the level of noise in the data, (ii) the distribution of error across frequency ranges, and (iii) errors in key metrics for microseismic event detection.

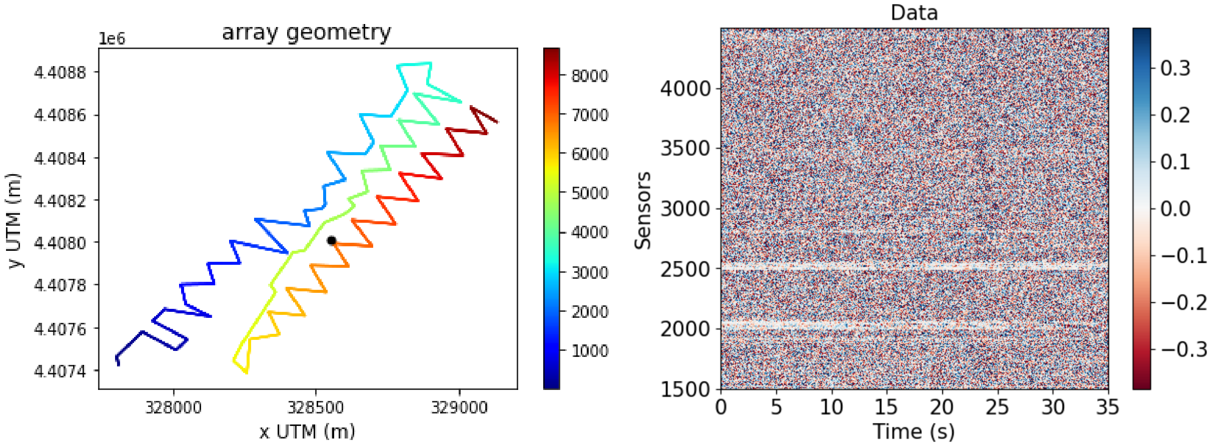

Figure 2. (Left) Map of the Brady Hot Springs geothermal field's horizontal DAS channels, color-coded by channel number. (Right) An example of 35 seconds of continuous data recorded at Brady Hot Springs at UTC 08:39:13 on 03/14/2016.

NORM OF NOISE

One objective when using lossy compression is to minimize the amount of noise introduced into the data during the compression process. To investigate the level of noise introduced by different compression schemes at various compression rates, we conducted experiments on ten days of data from the FORESEE-urban project recorded from UTC 23:51:35, 04/09/2019 to UTC 00:07:35, 04/21/2019. We selected this particular dataset because it contains a diverse range of recorded signals that we wish to preserve when compressing passive seismic data.

For each compression scheme the workflow for comparison is:

(i) Begin with data $D \in \mathbb{R}^{N_c \times N_t}$
(ii) Compress D
(iii) Reconstruct compressed data $\tilde{D} \in \mathbb{R}^{N_c \times N_t}$
(iv) Define the noise as the difference between the reconstructed and original data
(v) Compute the normalized Frobenius norm error defined as $\|D - \tilde{D}\|_F / \|D\|_F$

The results of our experiment are presented in Figure 3. Among the three modes of zfp compression, fixed accuracy mode performs better than fixed precision and fixed rate, which produce comparable noise. This can be attributed to the fixed accuracy mode encoding as many bit planes as necessary to achieve a specific absolute error, while the other modes encode a fixed amount of bit planes irrespective of the level of improvement they provide. Consequently, at the same rate or precision, compressed data may have high or low noise, depending on compressibility, but no improvement in the compression rate. In contrast, fixed accuracy mode produces the best compression achievable at a specific absolute noise cost for every data. The only drawback to this mode is that it requires knowledge of the range of values in the data to provide an absolute error tolerance, making it unsuitable for on-the-fly compression.

When comparing 2D and 1D wavelet compression, it is observed that 2D wavelet compression results in lower norm of noise. This improved performance could be attributed to the additional dimension available for compression, which allows for more effective utilization of redundancy and better preservation of signal integrity (Villasenor et al., 1996).

Regarding how the different compression schemes compare, we found that the three modes of zfp compression introduce the least noise up to about a 15X compression rate, followed by SVD and 2D/1D wavelet compression in the same range. However, at higher compression rates, zfp compression begins to incur errors at a higher rate, with the exception of fixed accuracy mode. The errors incurred by the wavelet transform earlier may be attributed to its denoising quality. At higher compression rates, there may be less noise, and compression may be depleting the real signals to a lesser extent. Since zfp is not tailored to seismic data, it incurs more errors at high compression rates, as expected for a lossy compression scheme. SVD produces errors at a steady rate. Since a higher level of compression means removing a higher amount of the singular values, the steady rate of growth in errors may be explained by a close to uniform distribution of singular values of the seismic data used. This is an area that may be explored in future studies.

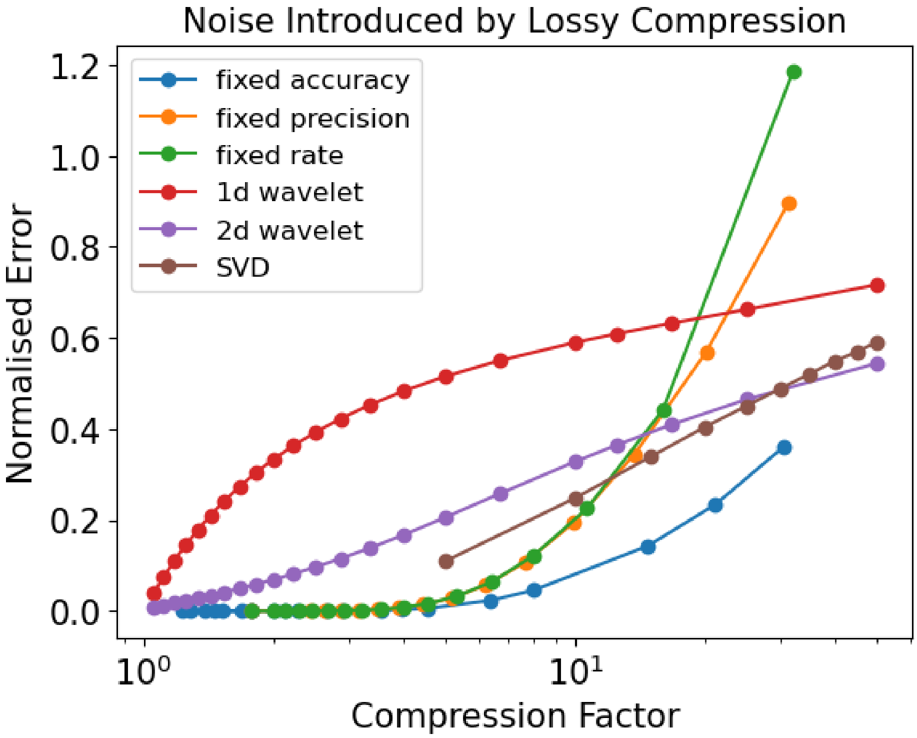

Figure 3. Frobenius norm of noise introduced in data at various levels of compression for compression with wavelet decomposition, singular value decomposition, and zfp floating point compression.

Differences in Frequency Content

In order to assess the extent to which the frequency content of the data is preserved following lossy compression, a methodology was employed whereby the proportion of the power spectrum retained at each frequency was calculated, then averaged across frequencies. The step-by-step process is as follows for each file containing a window of data in time:

(i) Obtain $\tilde{D}$ previously defined by compression and subsequent decompression of the data, D
(ii) Identification of the power spectrum, $P(\tilde{D})$ of $\tilde{D}$ in pre-determined time windows, T_w (5 seconds for this experiment) such that $P(\tilde{D}) \in N_c \times N_t/2 \times N_t/T_w$
(iii) Division of the power spectrum at each frequency by the original power spectrum ie $P_{\tilde{D}/D} = \frac{P(\tilde{D})}{P(D)}$

The outcome was a three-dimensional data cube for $P_{\tilde{D}/D}$ that captured information across time windows of size N_t/T_w, frequencies of size $N_t/2$, and channels, N_c. Weighted averaging was then performed along each dimension to investigate the patterns of frequency preservation across the various dimensions.

The data obtained by averaging the ratios of the compressed to original energy across the various frequencies can be represented as a two-dimensional dataset (time by channels), revealing the variations in frequency preservation. As depicted in Figure 4, the patterns observed in this dataset closely resemble the trends in the data shown in Figure 1. In the case of wavelet and low-rank compression, there is a general reduction in energy that can be attributed to the thresholding operation used in both compression methods. Additionally, wavelet decomposition exhibits better preservation of small events, while low-rank compression tends to preserve mostly high-amplitude events. On the other hand, zfp compression leads to a general increase in energy but preserves the energy around the identified events to levels close to pre-compression levels. This may be due to the spurious frequencies introduced at multiples of a quarter of the sampling rate as depicted in Figure 5. This effect of this might be attenuated when there are other strong frequencies corresponding to events and exaggerated when there aren't many other strong frequencies in noisy parts. In contrast, such false frequencies are not observed in wavelet or low-rank compression.

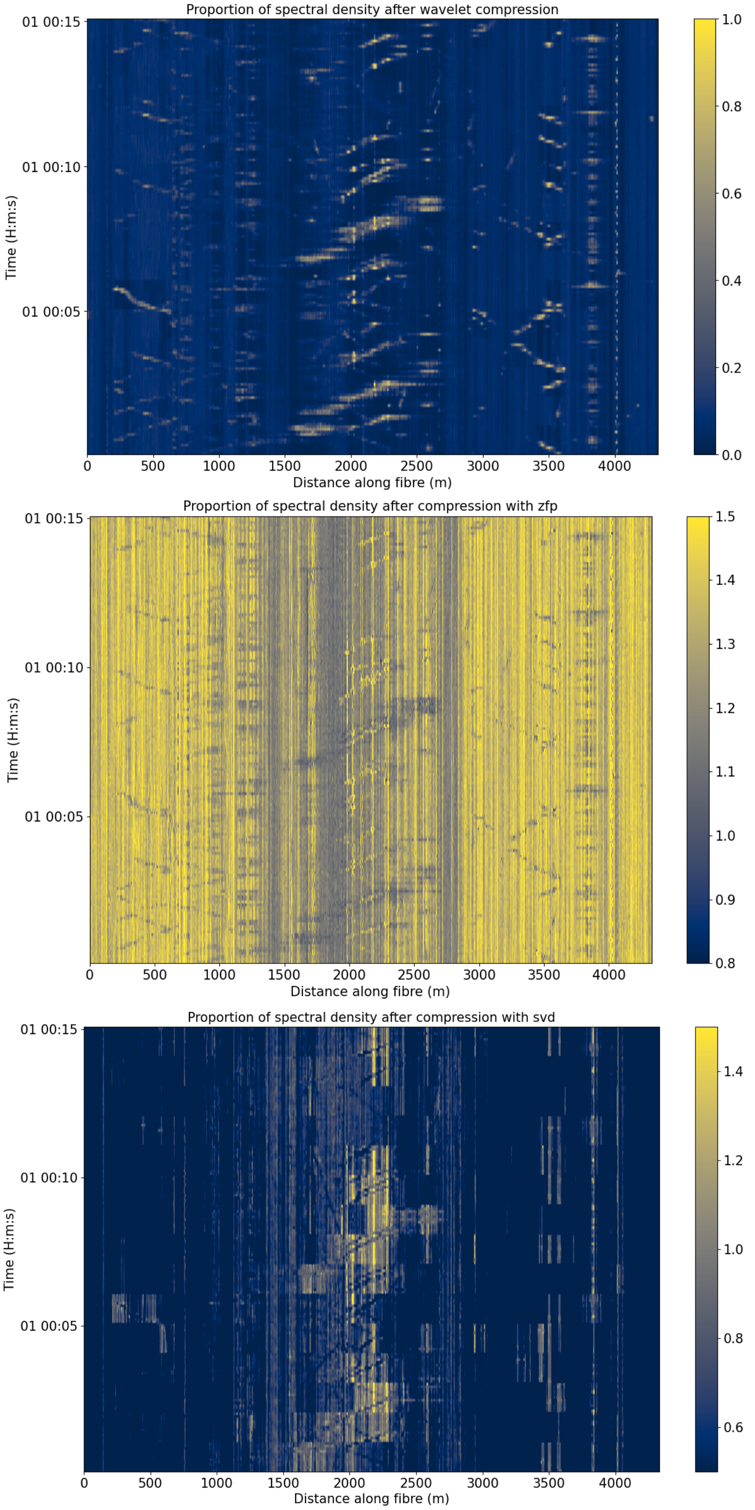

Figure 4. Frequency preservation with respect to compression for (top) wavelet compression, (middle)zfp compression, and (bottom)SVD compression. These images are $N_t/T_w \times N_c$ data in time and channel respectively obtained by averaging $P_{\tilde{D}/D}$ along the frequency axis

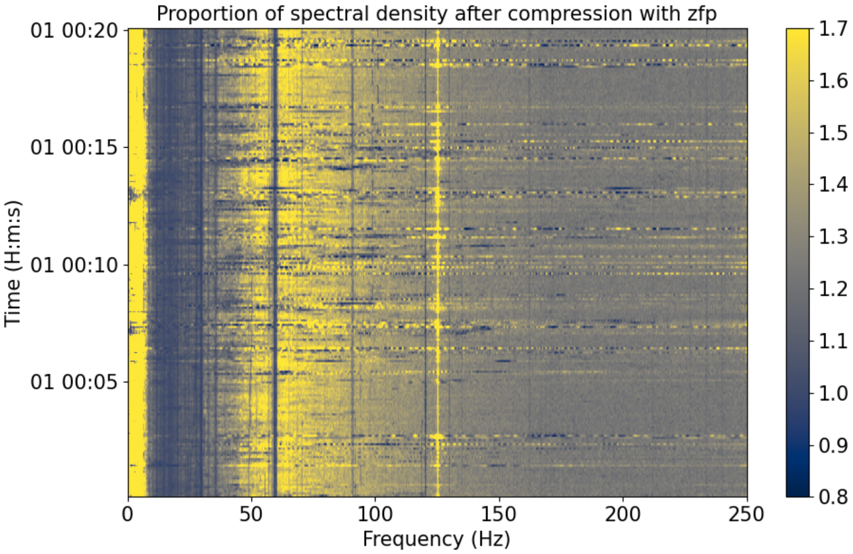

Figure 5. Frequency preservation with respect to compression for zfp compression. This image is constructed from $N_t/T_w \times N_t/2$ data in time and frequency respectively obtained by averaging $P_{\tilde{D}/D}$ along the channel axis

Changes in Template Matching Event Detection

Template matching enables the detection of similar events by comparing known seismic events, referred to as templates, with continuously recorded seismic data. The underlying principle of template matching involves calculating the cross-correlation coefficient of the template waveform and recorded seismic data at different time intervals. This correlation coefficient is normalized to account for variations in signal amplitude and noise levels and provides a quantitative measure of similarity between the template and the recorded data. The significance of template matching lies in its ability to enhance the detection capability for microseismic events. By applying cross-correlation analysis, even weak events that may be buried within the background noise can be identified and accurately located (Gibbons and Ringdal, 2006).

To evaluate the impact of noise introduced by different lossy compression schemes on event detection, we conducted a series of experiments using a template matching workflow. In particular, we investigate two questions:

(i) To what extent is the array-wide detection significance of varying size events impacted?

(ii) Are there biases or increases in the variability of event times picked across the array as higher compression ratios are used for any types of compression?

To test microseismic event detection via template matching, we use the Brady Hot Springs data. This dataset has been previously studied, and various microseismic activities have been cataloged providing a baseline catalog for comparison (Li and Zhan, 2018). Our goal in this study is to apply a similar workflow as outlined in Li and Zhan (2018) and compare the performance of event detection at various compression rates for 1D wavelet, zfp, and SVD compression. By analyzing the results of these experiments, we can gain valuable insights into the trade-off between compression rates and event detection performance for the compression schemes considered.

To provide a detailed view of the template matching process, we calculated the normalized cross-correlation of each channel's recording of the template event in Figure 6 with its recording of the noise pictured in Figure 2. The resulting array-wide normalized cross-correlation is shown in Figure 7.

Typically template matching would be carried out on small collections of point sensors, with large amplitude normalized cross-correlations being considered a possible event, and potentially a voting system among sensors to ensure multiple sensors detected the same possible event. With thousands of sensors, we use a simple metric: calculating the array-wide average of the normalized cross-correlations at each time, then calculating the envelope of the resulting time series. Despite time-delays in the original event as it moves across the array, if a similar event occurs in the same location, the large normalized cross-correlations across the array are expected to occur at the same time lag. Thus, averaging across all sensors does not cause destructive interference.

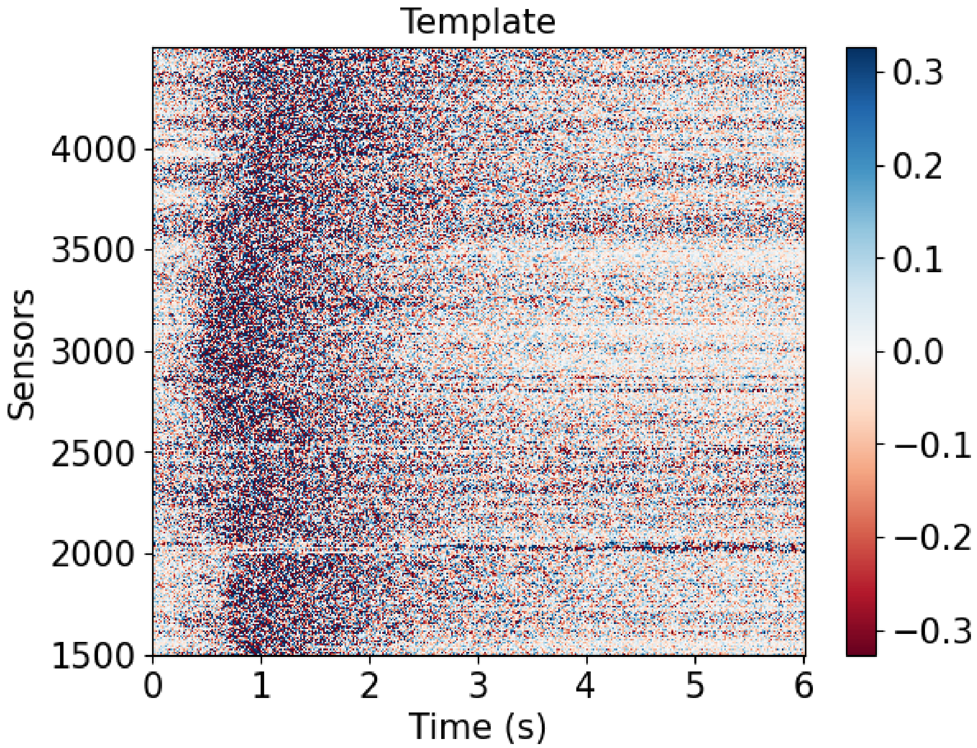

Figure 6. An example template event recorded at Brady Hot Springs at 8:39:05.24 on 03/14/2016.

The average envelope for each type of compression at multiple compression levels is shown in Figure 8. Three events are detected, even though these events were barely distinguishable in the raw data. We refer to the first event (between 1 and 2 seconds) as event 1, which is the mid-sized event, the second event (between 2 and 3 seconds) as event 2, which is the small event, and the third event (between 22 and 23 seconds) as event 3, which is the largest event. We see that for all three types of compression, the three events appear to be largely distinguishable from the background noise level. 1D wavelet compression maintains a more constant peak amplitude across compression ratios (1x original, 5x, 10x, 20x, 50x, and 100x) than do Zfp and SVD compression, although there is a small amount of amplitude loss at higher compression ratios (noticeable at 20x, 50x, 100x). The zfp compression shows some amplitude loss in the events, which particularly makes it difficult to distinguish event 2, but event 1 and event 3 are clearly above the noise level at all compression levels (1x original, 10x, 22x, 34x, 45x). The SVD compression leads to significant energy loss, particularly at the higher compression ratios (e.g. the peak of events 1 and 3 for 100x compressed data are less than half their original amplitudes).

As in Li and Zhan (2018), we use the detection significance of each event pick to provide a single array-wide value to quantitatively compare the preservation of each event when using various compression schemes and ratios. The detection significance is defined as $\frac{CC_i - M}{MAD}$ for a value CC_i on the array-wide average of the normalized cross-correlations at each time, CC. M is the median of CC and MAD is the median absolute deviation defined as $median(|CC - M|)$. We set a detection significance minimum threshold of 9, which was set to provide a reasonable bound on false detections Li and Zhan (2018). The detection significance for each of the three events across all compression schemes and compression ratios is shown in Figure 9. Notice that all three compression schemes preserve events 1 and 3 (the mid-sized and large events) as picked events above the threshold, even at high compression ratios (e.g. 45x and 50x for all, and 100x for wavelet and SVD). Note that the SVD compressed data have the largest drop in detection significance relative to 1D compressed and zfp compressed data. At 100x SVD compression, mid-sized event 1 is barely above the detection significance threshold. Even in the original uncompressed data, the small event 2 starts out very close to the detection threshold. With SVD compression at 20x and higher compression, and with 1D wavelet compression at the 100x level, event 2 drops substantially below the threshold for detection significance. It is not surprising that a barely-detectable event in the raw data could be lost in some of the highly compressed data. Although SVD compressed events all have a drop in detection significance, the 1D wavelet compression and zfp compression of the large event 3 data actually show an uptick in detection sig-

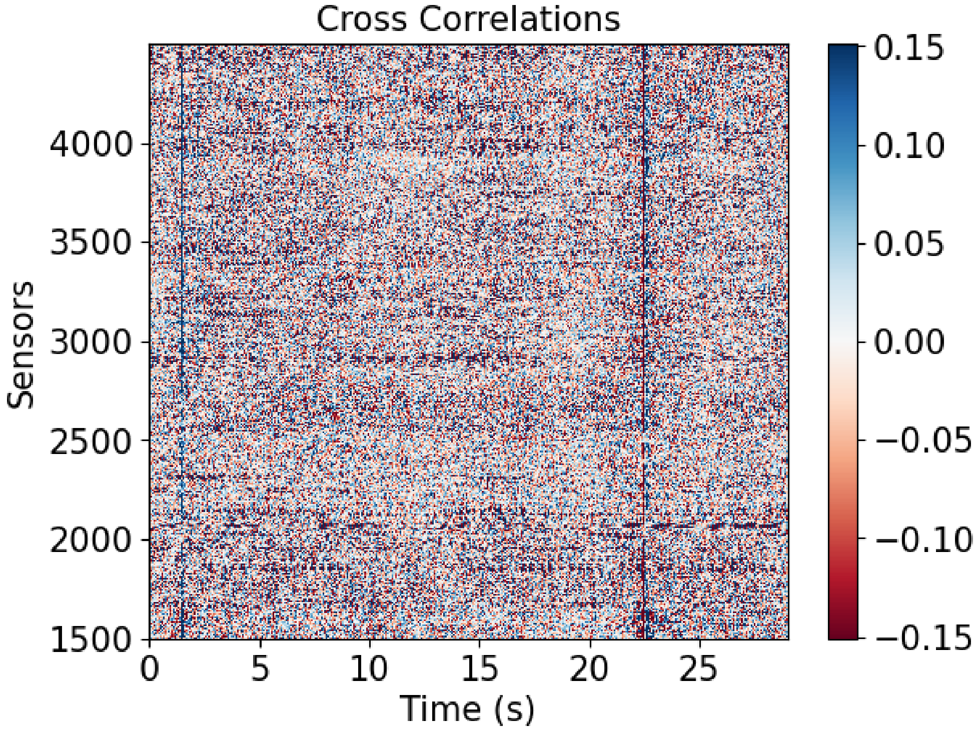

Figure 7. The array-wide template matching results showing the normalized cross-correlation of each channel's continuous recording with its template event recording. Vertical lines indicate many channels with high similarities at a particular time.

nificance at high compression ratios, likely indicating some denoising occurring to further emphasize this largest event during the compression process.

Figure 10 shows the event detected using the template matching workflow discussed on the original data. When compressed data is used for all the events in this catalog we see similar trends to the one explained by the small-scale case. The results of this is summarised in Figure 11 showing the variation in detection significance with compression. In these plots, we expect a trend with a slope of 1 in the situation where there aren't any changes in detection significance. For wavelet compression (top image Figure 11), we have a lot of points close to this trend even at high compression rates although points with smaller detection significance show a slight decrease while points with high detection significance show some increase indicating some denoising effect. This leads to smaller events eventually being missed at higher compression rates. SVD compression (bottom image Figure 11)shows a similar but shows more reduction in detection significance for mid and small size events. The image for zfp (middle Figure 11) shows less predictability; some small events increase in detection significance even though most events are observed to be reducing in detection significance.

To address the question of whether event pick times are reliable, we need to quantify the distribution of how each event's pick times change throughout the array of compressed data when compared to the event pick times on the original data. In particular, we need to know if the median pick time shift is staying very close to 0 (indicating there is no substantial array-wide bias) and compare the rate at which the minimum, Q1, Q3, and maximum values are spreading apart for higher compression ratios. For each type of compression, for each compression ratio, we created a box and whiskers plot for the distribution of each event's changes in pick times across all channels, which is shown in Figure 12. For each type of compression, these boxplots are overlaid for all three events (color-coded) so that the spread for a small, mid-sized, and large event can be compared easily.

There is not an apparent median bias for any of the events or compression types with a compression ratio less than 100x. At 100x there appears to be a slightly (< 0.05 second) late median of picks in SVD compression, and a positive skew distribution (based on investigation of min, max, Q1 and Q3) to the picks in 1D wavelet compression at 100x, although the median of the wavelet distribution appears to be unbiased. We see that for all types of compression, the extremes and quartiles spread out more with higher compression ratios. In the range of 5x - 15x compression, all three compression schemes perform reasonably well, with 5x 1D wavelet, 14x zfp, 5x and 10x SVD compression all yielding distributions for all three events that are completely contained

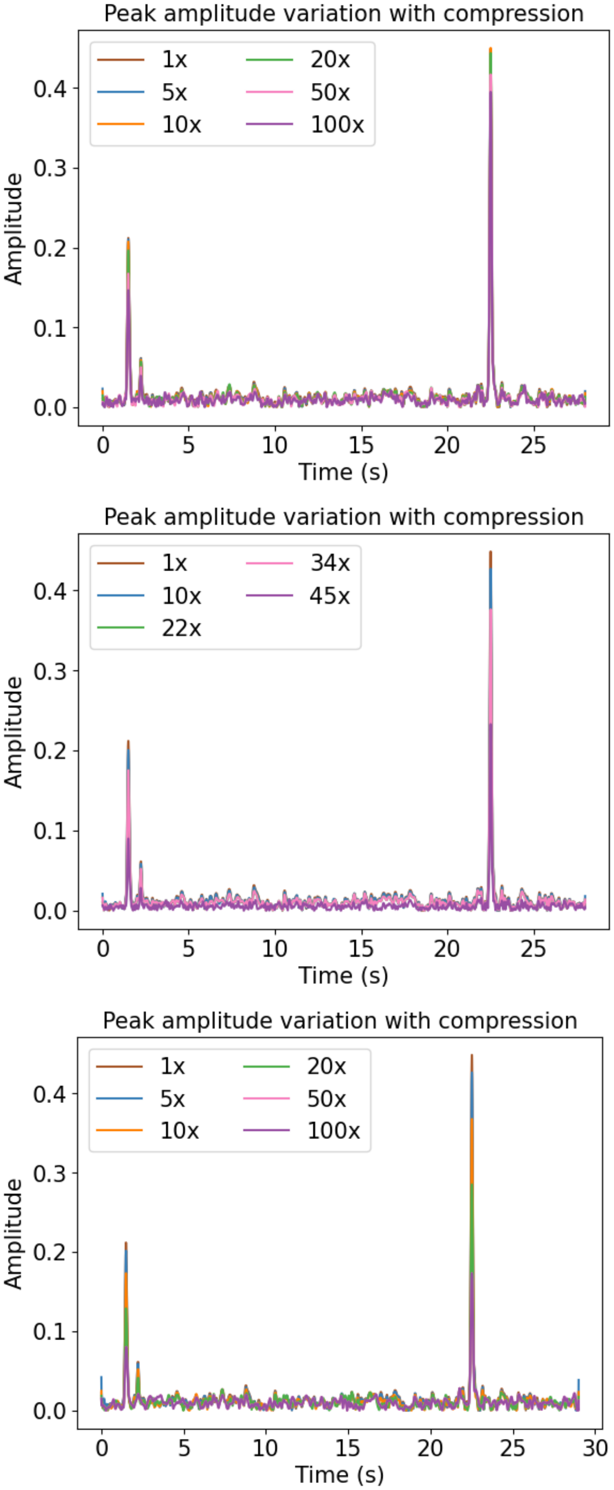

Figure 8. The envelope of the average normalized cross-correlations in Figure 7 during three events shows that the event picks are largely similar for various levels of compression. This was tested with (top) 1D wavelet compressed data, (middle) zfp compressed data, and (bottom) SVD compressed data.

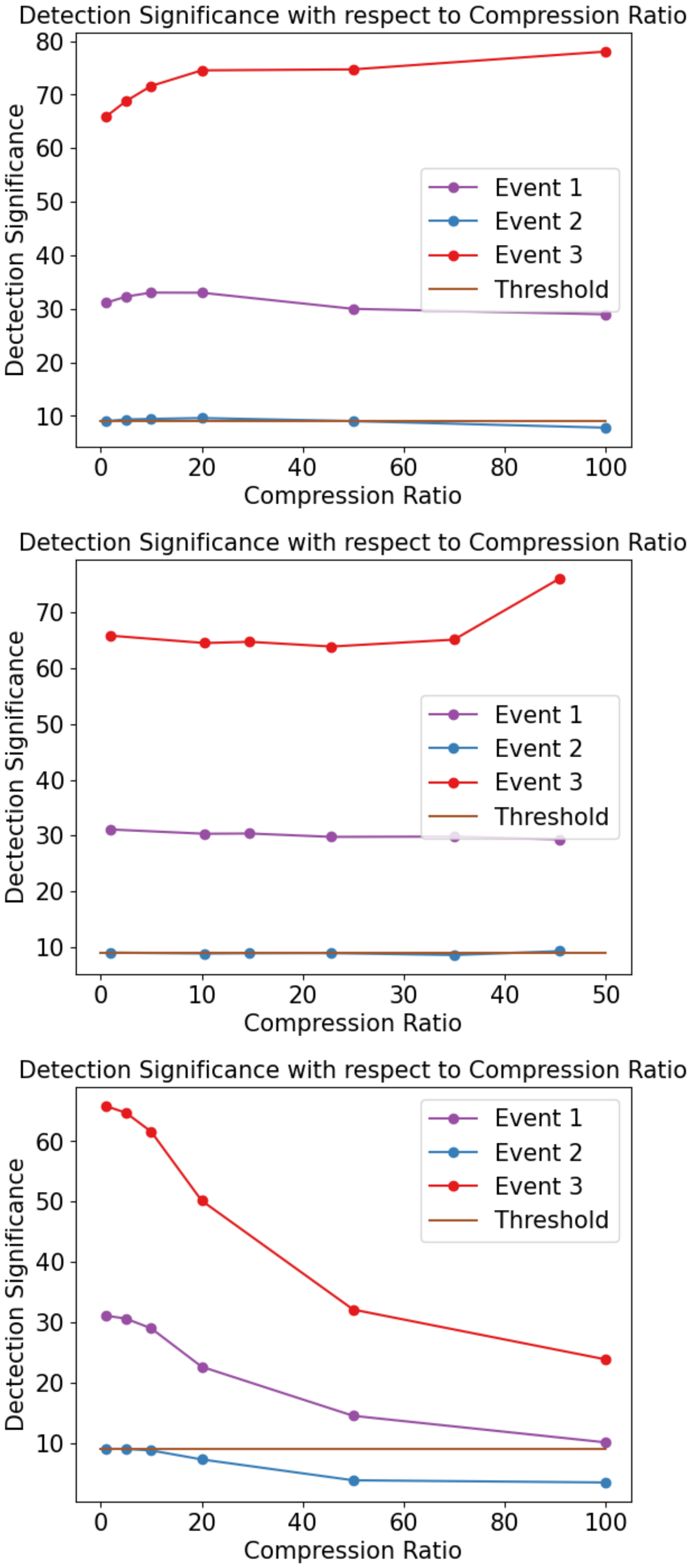

Figure 9. Change in detection significance with compression rate for (top) 1D wavelet compressed data, (middle) zfp compressed data, (bottom) low-rank SVD compressed data.

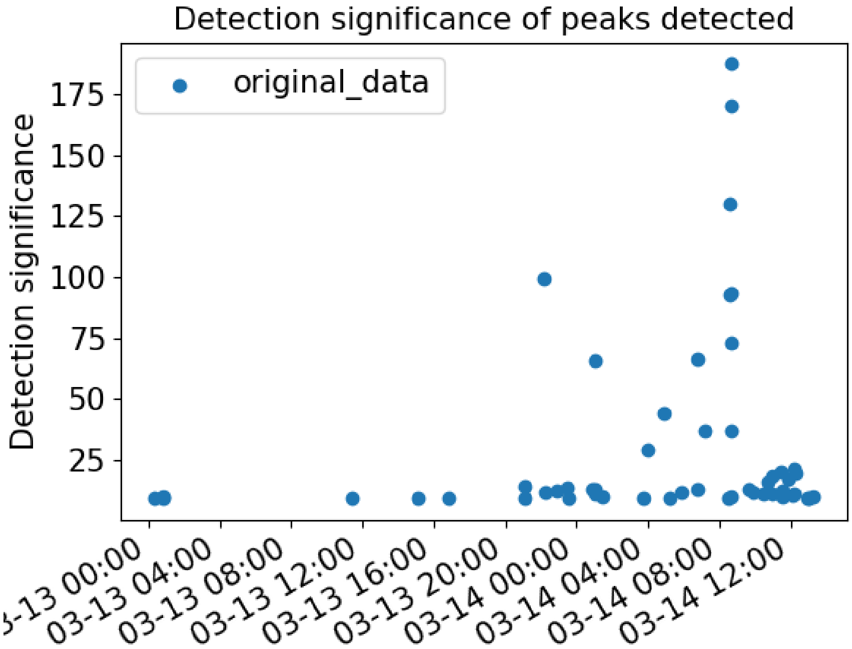

Figure 10. Catalog of events detected by template matching using the uncompressed data.

within a ±0.1 second shift. Among these low-ratio compression schemes, 1D 10x wavelet compression does perform the worst with the largest spread for all three events, and the small event 2 distribution only having its inner quartile range (Q1-Q3) contained within ±0.1 seconds, while its extremes have errors bounded by 0.3 seconds. In the range 20x - 35x, we see that zfp at 22x has the only distributions for all 3 events that are completely contained within ±0.1 seconds, although both 1D wavelet and SVD perform similarly well at 20x compression for the large event 3. In the range 45-50x, we see that 1D wavelet at 50x and zfp at 45x have very compact distributions for the large event 3, while SVD at 50x has a substantial spread in event pick shifts. All three compression schemes in this compression ratio range perform similarly for event 1 and event 2: with extremes that exceed ±0.4 seconds, and interquartile ranges that are bounded within ±0.2 seconds. Zfp did not allow higher compression ratios, but 1D wavelet and SVD compression were tested at the 100x ratio. 1D wavelet compression outperformed SVD compression on the large event 3 in the sense of providing a more compact distribution of time shifts, although 1D wavelet's distribution of pick time shifts is skewed to have larger positive shifts. Both schemes performed similarly at the 100x level on event 1 and event 2, with extreme shifts approximately around ±0.6 seconds.

5 DISCUSSION

Overall, we see that while zfp has the lowest data errors at lower compression ratios, wavelet compression (especially in 2D) has lower errors at higher compression ratios, and low-rank SVD has an error growth that sits in-between zfp and wavelets. Wavelets strongly improve broadband representation of strong events over quiet noise, and SVD tends to have better broadband representation of louder signals, but zfp tends to more evenly distribute errors in the frequency content across loud and quiet events. Using a microseismicity dataset, we could see that after feeding compressed data through a template matching workflow, all types of compression could preserve the events at smaller compression ratios. SVD compression tended to have the largest drop in detection significance at high compression ratios, although it still preserved the detection significance of a mid-sized and large event even at 100x compression. The unbiased picks with increasing variability due to higher compression ratios from Figure 12 suggests the opportunity to design a postprocessing scheme that promotes spatial coherency in event picks across the array. In this way, more reliable picks can be used from any highly compressed data, particularly as an input to the event location or for tomographic imaging using microseismic events. This analysis workflow was extended to a 36-hour period recording with 52 events of varying detection significance, and we found that wavelet compression preserved the detection significance better than zfp and SVD compression

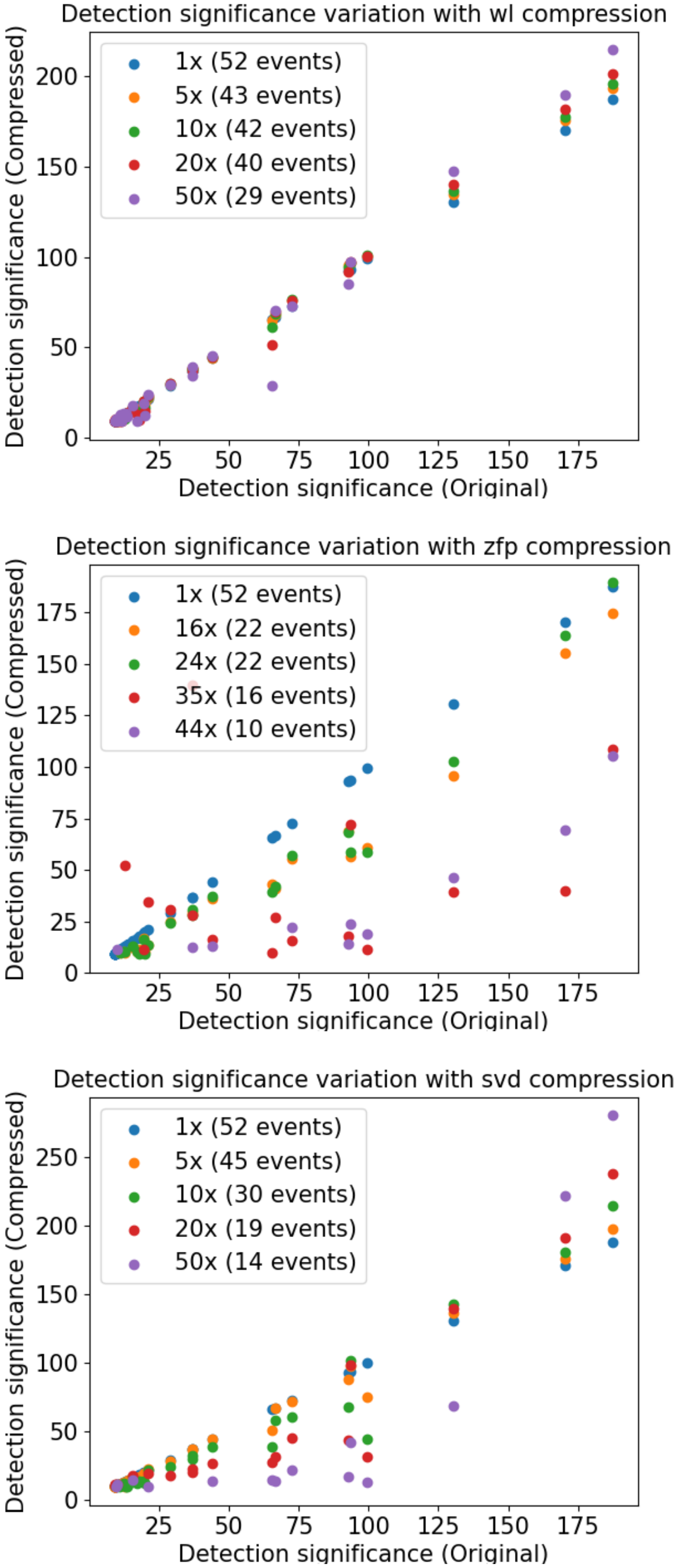

Figure 11. Trends in detection significance and events detected as compression rate is increased for (top) wavelet, (middle) zfp and (bottom) SVD compressions.

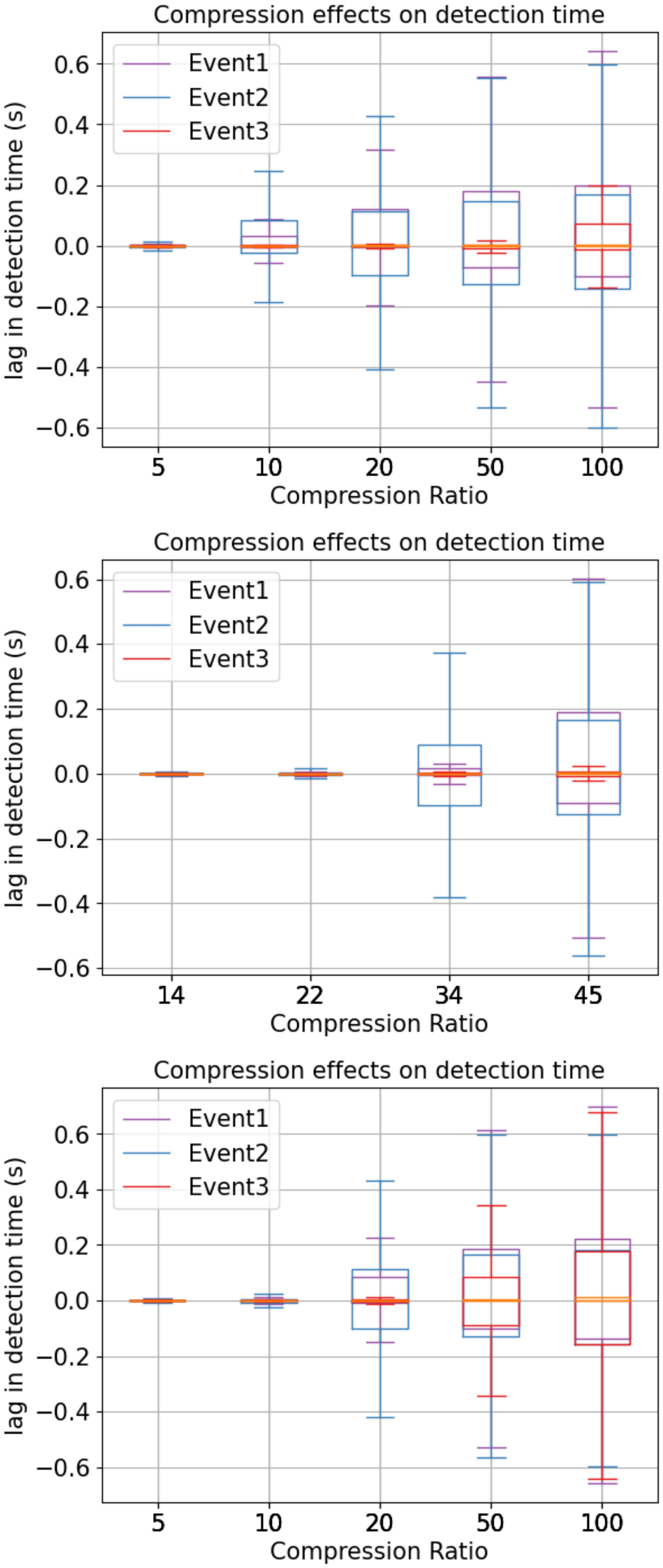

Figure 12. Boxplots show the distribution across all channels in the array of picked event times from template matching applied to the (top) 1D wavelet compressed data, (middle) zfp compressed data, and (bottom) low-rank SVD compressed data. These are shown at various compression ratios for three events.

at similar compression ratios, and tended to increase detection significance for larger events and higher compression ratios (i.e. emphasizing and denoising large events). Zfp compression typically led to a reduction in detection significance across all event sizes. SVD compression tends to reduce detection significance for smaller to mid-sized events and tended to increase detection significance for larger events, with more decrease/increase in significance for higher compression levels. Ultimately, SVD and wavelet representations have been integrated into a larger number of analysis workflows that can operate on data directly in their compressed representation, which may lead us to prefer these in some contexts, but zfp is being increasingly used in scientific computing, so algorithms incorporating zfp may be on the horizon in the coming years.

6 CONCLUSIONS

New technologies to continuously collect high-resolution seismic data for long periods of time are pushing us to consider lossy compression as a means of reducing data movement time, archival requirements, and processing/visualization time (particularly when interactive workflows are desirable). In this report, we compare the benefits and drawbacks of wavelet compression, zfp compression, and SVD compression at compression ratios ranging between 5x and 100x. These are tested on two public datasets: an urban dark fiber experiment as well as a geothermal field microseismicity monitoring experiment. We see that different compression schemes have the lowest errors at low compression rates versus high compression rates, and compare these errors as they propagate through an entire template matching microseismicity detection workflow.

7 ACKNOWLEDGEMENTS

We thank the National Science Foundation Grant 2046387, the Air Force Research Laboratory Subaward 62681767-227888 through Stanford University, and the sponsor companies of the Center for Wave Phenomena (CWP), whose support made this research possible. We thank the Virginia Tech ARC and Colorado School of Mines CIARC for computing resources. We also thank students and faculty of CWP and the Martin Group for helpful discussions throughout this work.

REFERENCES

Averbuch, A., F. Meyer, J.-O. Stromberg, R. Coifman, and A. Vassiliou, 2001, Low bit-rate efficient compression for seismic data: IEEE Transactions on Image Processing, **10**, 1801–1814.

Bosman, C., and E. Reiter, 1993, Seismic data compression using wavelet transforms: SEG Technical Program Expanded Abstracts 1993, Society of Exploration Geophysicists, 1261–1264.

Coleman, T., 2016, Brady's Geothermal Field - Metadata for DTS and DAS Surveys. (Type: dataset).

Donoho, D., 1995, De-noising by soft-thresholding: IEEE Transactions on Information Theory, **41**, 613–627.

Donoho, P. L., R. A. Ergas, and R. S. Polzer, 1999, Development of seismic data compression methods for reliable, low-noise, performance: SEG Technical Program Expanded Abstracts 1999, Society of Exploration Geophysicists, 1903–1906.

Gibbons, S., and F. Ringdal, 2006, The detection of low magnitude seismic events using array-based waveform correlation: Geophysical Journal International, **165**, 149–166.

Halko, N., P.-G. Martinsson, and J. Tropp, 2011, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions: SIAM Review, **53**, 217–288.

Li, Z., and Z. Zhan, 2018, Pushing the limit of earthquake detection with distributed acoustic sensing and template matching: a case study at the Brady geothermal field: Geophysical Journal International, **215**, 1583–1593.

Lindsey, N. J., and E. R. Martin, 2021, Fiber-Optic Seismology: Annual Review of Earth and Planetary Sciences, **49**, 309–336.

Lindstrom, P., 2014, Fixed-Rate Compressed Floating-Point Arrays: IEEE Transactions on Visualization and Computer Graphics, **20**, 2674–2683.

Mallat, S. G., 2009, A wavelet tour of signal processing: the sparse way, 3rd ed ed.: Elsevier/Academic Press.

Martin, E. R., 2019, A scalable algorithm for cross-correlations of compressed ambient seismic noise: Presented at the SEG Technical Program Expanded Abstracts 2019, Society of Exploration Geophysicists.

Valentine, A. P., and J. Trampert, 2012, Data space reduction, quality assessment and searching of seismograms: autoencoder networks for waveform data: Geophysical Journal International, **189**, 1183–1202.

Villasenor, J., R. Ergas, and P. Donoho, 1996, Seismic data compression using high-dimensional wavelet transforms: Proceedings of Data Compression Conference - DCC '96, IEEE Comput. Soc. Press, 396–405.

Zhu, T., J. Shen, and E. R. Martin, 2021, Sensing earth and environment dynamics by telecommunication fiber-optic sensors: an urban experiment in pennsylvania, usa: Solid Earth, **12**, 219–235.