SANDIA REPORT
SAND2023-11183
Printed August 2023

Sandia National Laboratories

Clustering Acoustic Background Noise in the Stratosphere Using Machine Learning

Ashley Chacon and Sarah Albert

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico
87185 and Livermore,
California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@osti.gov
Online ordering:	http://www.osti.gov/scitech

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.gov
Online order:	https://classic.ntis.gov/help/order-methods/

ABSTRACT

Infrasound, characterized by low-frequency sound inaudible to humans (<20 Hz), emanates from natural and anthropogenic sources. Its efficacy for monitoring phenomena necessitates robust sensing networks. Traditional ground-based infrasound sensors have limitations due to atmospheric dynamics and noise interference. Balloon-bore sensors have emerged as an alternative, offering reduced noise and improved capabilities. This study bridges clustering algorithms with balloon-borne infrasound data, a domain yet to be explored. Employing K-Means, DBSCAN, and GMM algorithms on normalized and reshaped data and only normalized data from a New Zealand-based NASA balloon flight, insights into background noise at stratospheric altitudes were revealed. Despite challenges arising from distinguishing signals amid unique background noise, this research provides vital reference material for noise analysis and calibration. Beyond infrasound event capture, the dataset enriches comprehension of background noise characteristics in the southern hemisphere.

ACKNOWLEDGEMENTS

This work was sponsored by the United States Department of Defense, Defense Threat Reduction Agency (DTRA) under the Nuclear Arms Control Technology (NACT) program.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

CONTENTS

LIST OF FIGURES

ACRONYMS AND TERMS

Acronym/Term	Definition
DBSCAN	Density-based Spatial Clustering of Applications with Noise
GMM	Gaussian Mixture Models
GUI	Graphical User Interface

1. INTRODUCTION

Infrasound is low frequency sound below the range of human hearing (< 20 Hz). It is generated by both natural and anthropogenic sources including chemical and nuclear explosions, rocket launches, volcanoes, earthquakes, and ocean processes. Infrasound travels efficiently through the atmosphere, making it useful for monitoring and characterizing sources of interest. Infrasound sensors have traditionally been ground-based (see Burlacu et al, 2010; Christie & Campus, 2010; Alaska Volcano Observatory, 2016; Marty, 2019; Bondár et al., 2022). Unfortunately, these ground-based networks are at the mercy of atmospheric winds and source characteristics. They can only record distant sources when the stratospheric winds are favorable and/or the signal can be transmitted through the thermosphere. Ground-based sensors are also often plagued by wind noise (Raspet et al., 2019). More recently, balloon-borne infrasound sensors have gained attention due to lower background noise over their ground-based counterparts. Balloon-borne infrasound sensors have shown utility in recording ocean processes (Bowman, 2016; Bowman & Lees, 2017), lightning (Lamb et al., 2018), earthquakes (Brissaud et al., 2021; Garcia et al., 2022;), chemical explosions (Bowman & Albert, 2018; Young et al., 2018; Bowman et al., 2021; Silber et al., 2023), and the Hunga Tonga volcanic eruption (Podglajen et al., 2022).

Clustering algorithms play a prominent role in understanding complex datasets by grouping data points into clusters based on shared characteristics. In unsupervised machine learning, these algorithms are essential tools as they enable the exploration of data structures without prior knowledge of class labels. Some of the more commonly used clustering algorithms include K-Means, Density-based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixture Models (GMM). A number of previous studies have used clustering methods on infrasound data, focusing on characterizing volcanic activity (Cannata et al., 2011; Watson, 2020; Witsil et al., 2020) and battlefield acoustics (Fields et al., 2021). However, clustering methods have yet to be applied to infrasound data recorded on balloon-borne sensors. Therefore, we present a novel study analyzing single-day records from a balloon-borne infrasound dataset. We present results from clustering algorithms, including K-Means, DBSCAN, and GMM and compare them against each single-day output to ascertain the emergence of any variations or patterns across several days of data.

This page left blank

2. METHODS

2.1. Balloon Experiment

On May 16, 2016 a NASA Ultra Long Duration Balloon was launched out of Wanaka, New Zealand. An infrasound payload was attached to the gondola of the balloon and recorded data for 19.5 days. The balloon flew for a total of 46 days, circumnavigating the southern hemisphere (Bowman & Lees, 2018). The attached infrasound payload consisted of three InfraBSU infrasound microphones (Marcillo et al., 2012) and an Omnirecs Datacube digitizer which recorded data at 200 samples per second at a gain of 64. Data was converted to a single channel using the method outlined in Bowman & Lees (2018). The balloon maintained a float altitude of ~33 km; the full trajectory is shown in Figure 1. A detailed description of the experiment can be found in Bowman & Lees (2018).

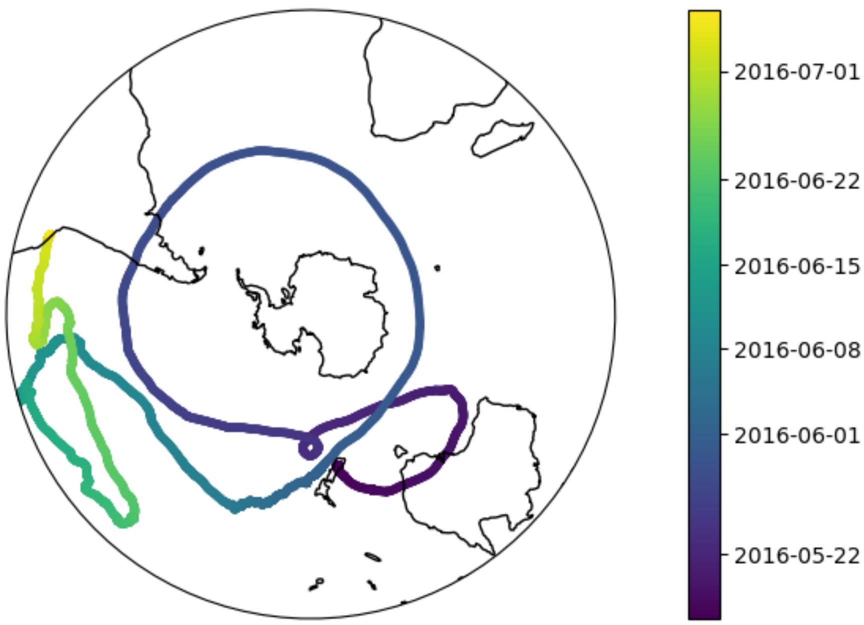

Figure 1. Trajectory of the NASA Ultra Long Duration Balloon that was launched out of Wanaka, New Zealand on May 16, 2016. The balloon carried an infrasound payload that recorded data for 19.5 days.

2.2. Data Curation

In any time series analysis, informative data processing is essential to unveiling meaningful insights from recorded signals. Prior to analysis, the Wanaka dataset underwent a series of preprocessing steps. This section outlines the filter method, generation of spectrograms for visual representation, and normalization of data.

2.2.1. Preprocessing

The ObsPy library is a widely used tool in seismology, and provides a range of functionalities for signal processing and analysis. In the first step of preprocessing, the Wanaka dataset was detrended and lowpass filtered below 20 Hz using ObsPy. This filter was chosen to attenuate high-frequency noise and unwanted data. The 20 Hz cutoff frequency for the lowpass filter was chosen based on characteristics of the dataset and the frequency band for infrasound (< 20 Hz).

2.2.2. Generating Spectrograms

Spectrograms provide a comprehensive view of data, aiding in the identification of short-term and/or long-term patterns. After applying the lowpass filter, the pre-processed data was transformed into spectrograms by computing the Short-Time Fourier Transform (STFT), which involved dividing the data into hour-long segments and calculating the Fourier Transform for each. We used a Tukey window with shape parameter of 0.25 and a segment length of 625.

2.2.3. Normalization

Normalization is an essential step in data processing to ensure different datasets can be compared and analyzed on a common scale. In the context of the Wanaka dataset, a simple manual min-max normalization technique is utilized. Manual min-max scaling is a linear transformation technique that maps data values to a specific range, between 0 and 1. Min-max scaling was chosen over other normalization techniques due to its several advantages, including simplicity and transparency. This scaling technique is particularly useful when the distribution of the data varies widely and when the data range is known to be consistent across different subsets of the dataset.

2.3. Clustering Algorithms

2.3.1. Clustering Methods

In the initial phase of preprocessing, the data was reshaped to align with the input expectations of the following clustering algorithms: K-Means, DBSCAN, and GMM. This step was pivotal in ensuring that the algorithms could accurately operate on the data's dimensions. The initial dataset included both spectrograms and date/time information. After experimenting with reshaping of the initial data, the date/time dimension was removed from the dataset for analysis. This step was essential in mitigating issues related to varying scales and magnitudes of features, thereby allowing the clustering algorithms to perform optimally on comparable scales.

2.3.2. Graphical User Interface (GUI)

To facilitate a deeper understanding of the clustering algorithm's output, a graphical user interface (GUI) was developed to provide interactive data visualization of data points in a given clustering algorithm plot. This GUI was crafted to enable users to gain insight to the nature of individual data points in a plot by hovering over them. Upon hovering over a specific data point, the GUI dynamically generates and displays the corresponding spectrogram. This visualization tool bridges the gap between abstract cluster assignments and underlying infrasound events, enhancing the interpretability of the clustering algorithm's results.

3. RESULTS AND DISCUSSION

First, we focused on a single day if data collected from the Wanaka region. This dataset was subjected to clustering using K-Means, DBSCAN, and GMM algorithms with two types of inputs: (1) normalized and reshaped data and (2) only normalized data. While utilizing normalized and reshaped data for clustering, a notable observation emerged. The results indicated significant presence of noise within data, which appeared to be clustered with signals that could potentially be interpreted as relevant infrasound events. The clustering algorithms struggled to distinctly separate noise from the signal-bearing data points, resulting in cluster assignments that were less coherent. This led us to believe that only normalizing the data may prove more useful.

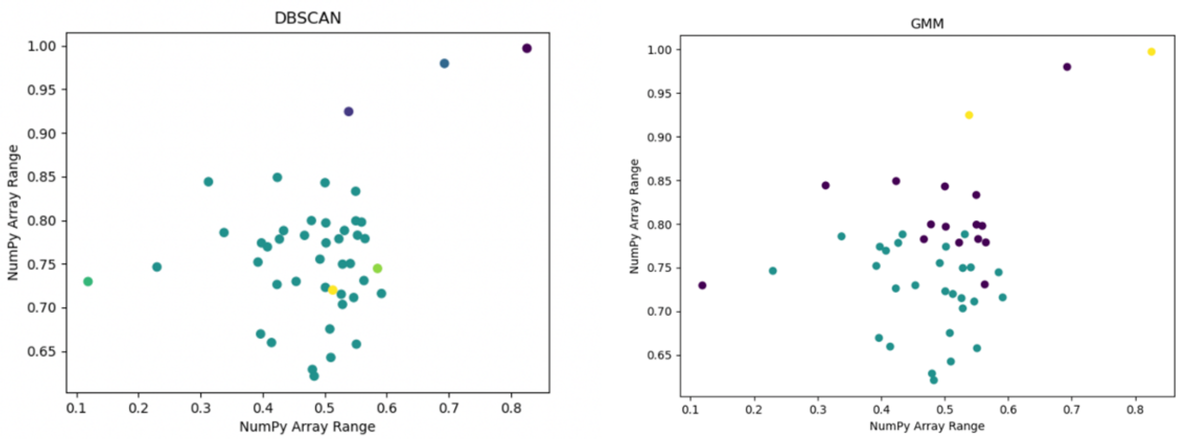

Figure 2. Data visualization of DBSCAN and GMM models on Julian day 138, which is the first day applied to clustering algorithms.

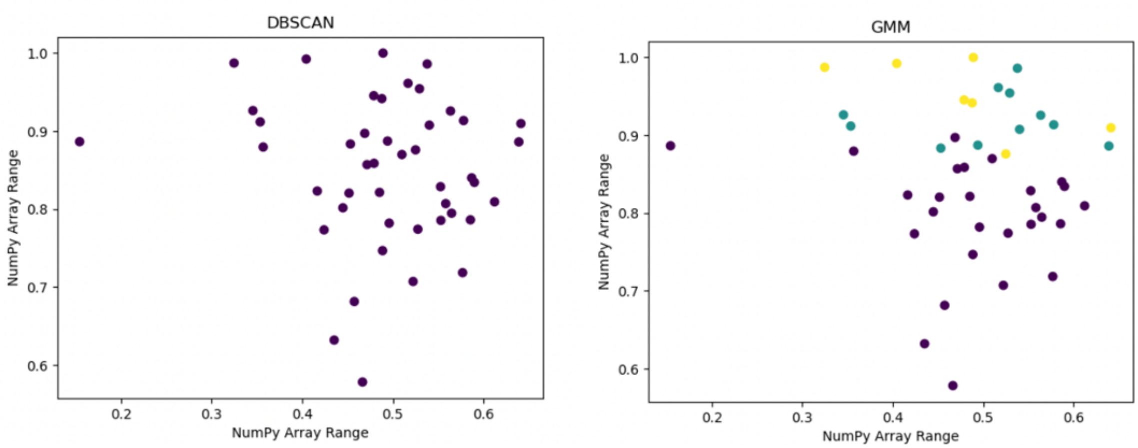

Figure 3. Data visualization of DBSCAN and GMM models on Julian day 156, which is the last day applied to clustering algorithms.

Interestingly, the application of the same clustering algorithms to only normalized data displayed a similar trend. The results still exhibited a considerable amount of noise, suggesting that the normalization process did not entirely alleviate the issue of noise contamination within the clusters. The abundance of noise in the clustering algorithm results warrants a comprehensive discussion, particularly in the context of the data collection methodology. As the infrasound data was collected from an airborne recording system in the stratosphere, this methodological approach presents unique challenges in differentiating between relevant signals and background noise, something that has never been done at these altitudes.

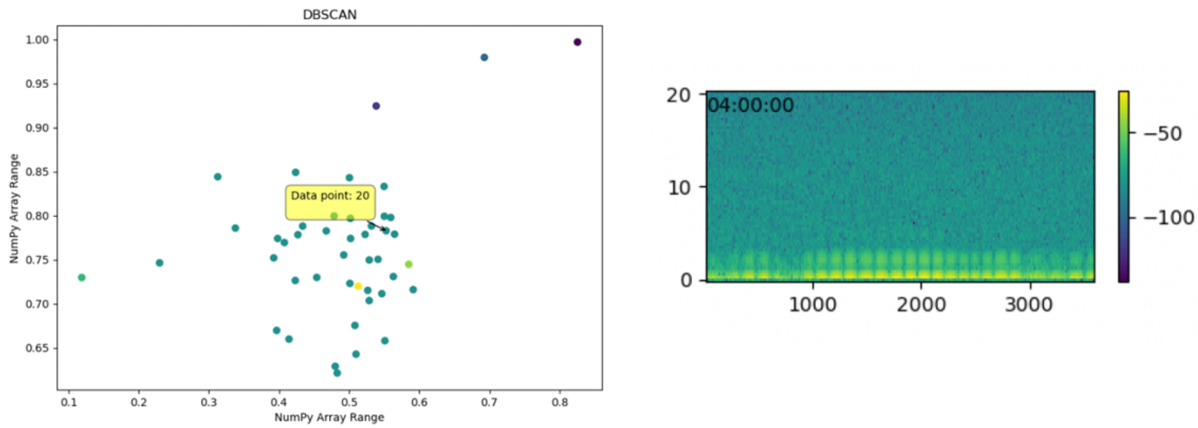

Figure 4. DBSCAN plot and corresponding spectrogram for Julian day 138.

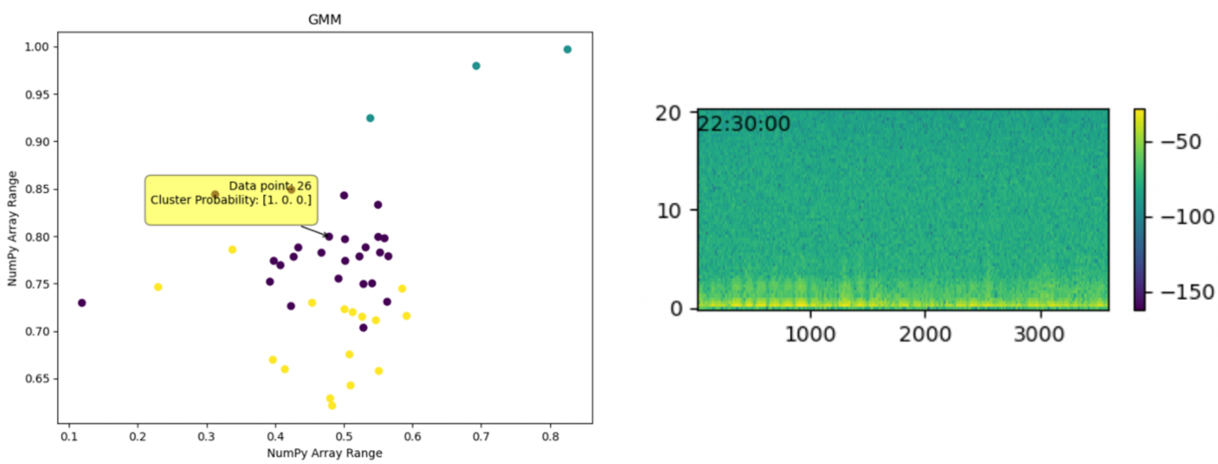

Figure 5. GMM plot and corresponding spectrogram for Julian day 138. Cluster probability displays which cluster the data point is assigned to. In this case, out of the three clusters, data point 26 belongs to cluster 1.

As the airborne sensor continued to collect infrasound data from the atmosphere, an analysis of each clustering output of the 19.5-day dataset indicated a changing distribution pattern, which can be seen in Figures 2 and 3. When observing each day's DBSCAN and GMM output, noticeable infrasound events are present above the background noise. Example spectrograms are shown in Figures 6 a-f and were generated from data collected at the same time over several days, specifically between the hours of 21:00 and 23:30. For all days, most of the recorded energy falls below ~5 Hz. However, on Julian days 142, 150, and 153 there are obvious broadband signals, suggesting infrasound events were recorded. However, it is important to note that the majority of spectrograms resemble those from Julian days 138, 147, and 156, which do not contain signals. Clustering algorithms typically struggle with data inequality such as this.

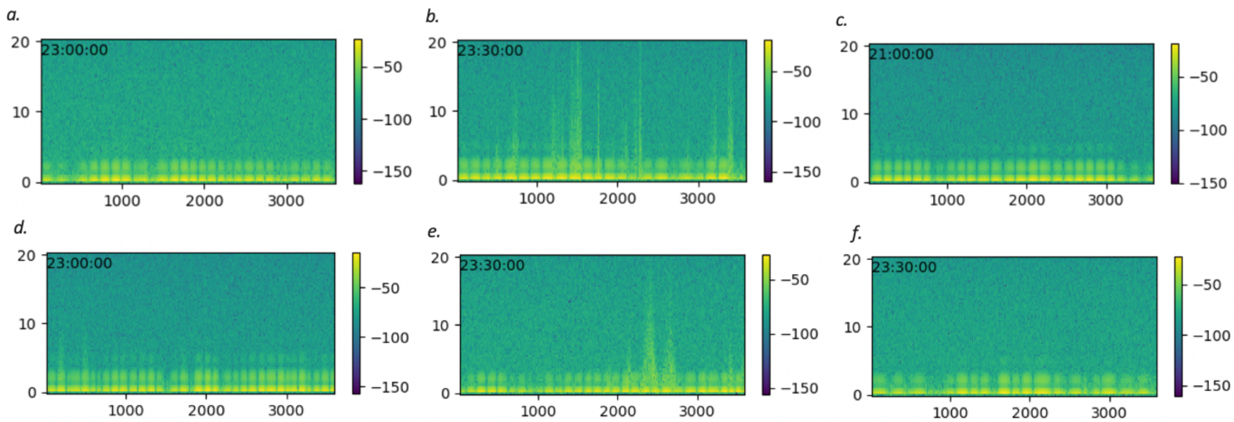

Figure 6. Generated spectrograms for Julian days 138, 142, 147, 150, 153, and 156 respectively.

The presence of significant noise in the clustering results poses challenges in accurately identifying and isolating genuine infrasound events. Clustering algorithms rely on patterns and relationships within data points to form distinct clusters. Additional investigations into the origin and characteristics of the noise detected in the dataset are necessary. Identifying the sources of noise and understanding their behavior can aid in developing tailored preprocessing methods. However, the dataset's richness in background noise serves as a reference point for noise analysis and calibration. As the dataset encompasses a diverse range of noise sources, it can aid researchers in developing noise reduction techniques, enhancing the quality of infrasound data collected from similar altitudes and regions. Despite these challenges, this dataset contributes to the broader understanding of infrasound recorded at stratospheric altitudes in the southern hemisphere. While capturing events may not have been the primary outcome, the dataset offers a glimpse into the background noise characteristics unique to this region. By acknowledging and leveraging this information, researchers can enhance their comprehension of infrasound-generating processes on a global scale.

This page left blank

4. CONCLUSIONS

We employed K-Means, DBSCAN, and GMM clustering algorithms on both reshaped and normalized data and only normalized data from a NASA Ultra Long Duration Balloon flight out of Wanaka, New Zealand. When clustering reshaped data, the outcomes were dominated by the presence of noise within the dataset, assigning both noise and potential events of interest to clusters. Interestingly, applying the same clustering algorithms to normalized data yielded similar results. Given that the infrasound data was sourced from an airborne recording system at stratospheric altitudes, and little is known about background noise in this region, there are unique challenges inherent to distinguishing between relevant signals and background noise. This work provides insight into background noise at these altitudes and holds potential as a reference for noise analysis and calibration. While infrasound event capture was not the primary outcome, this dataset allows a glimpse into the distinctive background noise and characteristics specific to stratospheric altitudes in the southern hemisphere.

REFERENCES

[1] Bowman, D.C. and Lees, J.M., 2017. A comparison of the ocean microbarom recorded on the ground and in the stratosphere. *Journal of Geophysical Research: Atmospheres*, *122*(18), pp.9773-9782.

[2] Bowman, D.C. and Albert, S.A., 2018. Acoustic event location and background noise characterization on a free flying infrasound sensor network in the stratosphere. *Geophysical Journal International*, *213*(3), pp.1524-1535.

[3] Bowman, D.C. and Krishnamoorthy, S., 2021. Infrasound from a buried chemical explosion recorded on a balloon in the lower stratosphere. *Geophysical Research Letters*, *48*(21), p.e2021GL094861.

[4] Brissaud, Q., Krishnamoorthy, S., Jackson, J.M., Bowman, D.C., Komjathy, A., Cutts, J.A., Zhan, Z., Pauken, M.T., Izraelevitz, J.S. and Walsh, G.J., 2021. The first detection of an earthquake from a balloon using its acoustic signature. *Geophysical Research Letters*, *48*(12), p.e2021GL093013.

[5] Cannata, A., Montalto, P., Aliotta, M., Cassisi, C., Pulvirenti, A., Privitera, E. and Patanè, D., 2011. Clustering and classification of infrasonic events at Mount Etna using pattern recognition techniques. Geophysical Journal International, 185(1), pp.253-264.

[6] Fields, M.P., Bennett, H. and Scoggins, R., 2021, April. Machine learning for source classification utilizing infrasound data. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III (Vol. 11746, pp. 677-686). SPIE.

[7] Garcia, R.F., Klotz, A., Hertzog, A., Martin, R., Gérier, S., Kassarian, E., Bordereau, J., Venel, S. and Mimoun, D., 2022. Infrasound from large earthquakes recorded on a network of balloons in the stratosphere. *Geophysical Research Letters*, *49*(15), p.e2022GL098844.

[8] Lamb, O.D., Lees, J.M. and Bowman, D.C., 2018. Detecting lightning infrasound using a high-altitude balloon. *Geophysical Research Letters*, *45*(14), pp.7176-7183.

[9] Marcillo, O., Johnson, J. B., & Hart, D., 2012. Implementation, characterization, and evaluation of an inexpensive low-power low-noise infrasound sensor based on a micromachined differential pressure transducer and a mechanical filter. *Journal of Atmospheric and Oceanic Technology*, 29, 1275–1284. https://doi.org/10.1175/JTECH-D-11-00101.1

[10] Podglajen, A., Le Pichon, A., Garcia, R.F., Gérier, S., Millet, C., Bedka, K., Khlopenkov, K., Khaykin, S. and Hertzog, A., 2022. Stratospheric balloon observations of Infrasound waves from the 15 January 2022 Hunga eruption, Tonga. *Geophysical Research Letters*, *49*(19), p.e2022GL100833.

[11] Silber, E.A., Bowman, D.C. and Ronac Giannone, M., 2023. Detection of the Large Surface Explosion Coupling Experiment by a Sparse Network of Balloon-Borne Infrasound Sensors. *Remote Sensing*, *15*(2), p.542.

[12] Watson, L.M., 2020. Using unsupervised machine learning to identify changes in eruptive behavior at Mount Etna, Italy. Journal of Volcanology and Geothermal Research, 405, p.107042.

[13] Witsil, A.J. and Johnson, J.B., 2020. Analyzing continuous infrasound from Stromboli volcano, Italy using unsupervised machine learning. Computers & Geosciences, 140, p.104494.

[14] Young, E.F., Bowman, D.C., Lees, J.M., Klein, V., Arrowsmith, S.J. and Ballard, C., 2018. Explosion-generated infrasound recorded on ground and airborne microbarometers at regional distances. *Seismological Research Letters*, *89*(4), pp.1497-1506.

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
Sarah Albert	6752	salber@sandia.gov
Daniel Gonzales	6752	dgonza2@sandia.gov
Technical Library	1911	sanddocs@sandia.gov

Email—External

Name	Company Email Address	Company Name
Ashley Chacon	achacon14@miners.utep.edu	The University of Texas at El Paso (UTEP)

This page left blank

This page left blank

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.